Studies in Big Data 49

Himansu Das
Rabindra K. Barik
Harishchandra Dubey
Diptendu Sinha Roy *Editors*

Cloud Computing for Geospatial Big Data Analytics

Intelligent Edge, Fog and Mist Computing

Springer

Studies in Big Data

Volume 49

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Big Data” (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and others. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/11970

Himansu Das · Rabindra K. Barik
Harishchandra Dubey · Diptendu Sinha Roy
Editors

Cloud Computing for Geospatial Big Data Analytics

Intelligent Edge, Fog and Mist Computing

Springer

Editors
Himansu Das
School of Computer Engineering
Kalinga Institute of Industrial Technology (KIIT)
Bhubaneswar, Odisha, India

Rabindra K. Barik
School of Computer Application
Kalinga Institute of Industrial Technology (KIIT)
Bhubaneswar, Odisha, India

Harishchandra Dubey
Center for Robust Speech Systems
University of Texas at Dallas
Richardson, TX, USA

Diptendu Sinha Roy
Department of Computer Science and Engineering
National Institute of Technology, Meghalaya
Shillong, Meghalaya, India

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-030-03358-3 ISBN 978-3-030-03359-0 (eBook)
https://doi.org/10.1007/978-3-030-03359-0

Library of Congress Control Number: 2018960234

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Current advances in the field of cloud and fog computing approaches for geospatial database mapping and geospatial web services are increasing the demand for better and quick integrated geospatial information and knowledge to support variety of users and their applications. Geospatial web services are the different varieties of web services for extraction of information from images with a simple hyperlinked image like Google Map, open street maps, etc. The main challenges are to visualize, integrate, discover, and manipulate geospatial data at the appropriate level to users in an environment that comprises of thick, thin, and mobile environments.

Big data analytics with the help of cloud computing is one of the emerging areas for processing, analysis, and transmission of any data particularly geospatial big data. The cloud computing is also one of the paradigm where cloud infrastructures help to increase the throughput and reduce latency for assisting at the edge of the client. As cloud computing based geospatial information and geospatial web services become easier to retrieve or access by the global community of users. It increases the challenges to focus the provision of geospatial information towards the context of applications and users. The plan is to provide the correct and right information at right time to the right people for the right decision-making to the variety of people across the globe. This book discusses the emergence of cloud computing for analytics in big data from various geospatial applications. With the integration of different location-based sensors and mobile devices, it gives the new paradigm of Internet of Things (IoT). Recently, the propagation of various low-cost sensors reduced costs of cloud computing infrastructure resources and increasing accessibility of machine learning and deep learning platforms have pooled to boost the applications and advancements in the area of geospatial computing and IoT.

The main objective of this edited book is to cover the state-of-the-art reference, advancement made, as well as prompting future directions on both the theories and applications in cloud computing and its applications in several diversified fields. It aims to provide an intellectual forum for researchers in Academia, Scientists, and Engineers from a wide range of application areas to present their latest research findings in cloud computing, geospatial data, and big data analytics and related areas and to identify future challenges in this novel combination of research areas.

The interpretation of geospatial big data is used for highly automated approaches relying on new machine learning and deep learning approaches in cloud environments. Extraction of useful information at large scale from heterogeneous, geo-referenced data is the major research topic in geospatial research area. This means it requires building on cloud computing based architectures, geospatial web services, geospatial web semantics and ontology, geospatial data visualization, and geospatial data analytics for geospatial big data. This book will focus on the recent trends and challenges for employing cloud computing for geospatial big data analytics in Intelligent Edge, Fog, and Mist Computing.

To achieve the objectives, this book includes 13 chapters contributed by promising authors.

In Chapter "Big Data Scientific Workflows in the Cloud: Challenges and Future Prospects", Khan et al. highlighted open research problems and give insights into the specific issues like workflow scheduling, execution, and deployment of big data scientific workflows in a multisite cloud environment. In Chapter "Trust Model Based Scheduling of Stochastic Workflows in Cloud and Fog Computing", authors focus on assuring trusted environment in the cloud and how this model ensures the user's requests is serviced with enough security.

In Chapter "Trust-Based Access Control in Cloud Computing Using Machine Learning", Khilar et al. illustrate the trust-based approach based on machine learning approach that predicts the trust values of the user and resources in the cloud environment. Chapter "Cloud Security Ontology (CSO)" addresses cloud security ontology and its strength and totality compared to prior ontologies. In Chapter "Cloud Based Supply Chain Networks—Principles and Practices", author describes the parameters of the characteristics of cloud computing in supply chain network for the purpose of modeling and analyzing the information flow. It will enable the decision maker to derive the necessary results by suitably incorporating the factors in the analysis of cloud supply chain network. Chapter "Parallel Computation of a MMDBM Algorithm on GPU Mining with Big Data" presents the performance of fast classifier method and radix algorithm to relate the processing time of Mixed Mode data Based Miner (MMDBM), SLIQ CPU with GPU computing, and computed acceleration ratio (Speed-up) time. In Chapter "Data Analytics of IoT Enabled Smart Energy Meter in Smart Cities", smart energy meters' data analytics framework is addressed by employing latest data processing techniques or tools along with gamification approach for enhancing consumers' engagement. Benefits of smart energy meter's analytics are also discussed for motivating consumers, utilities, and stakeholders. In Chapter "A New and Secure Intrusion Detecting System for Detection of Anomalies Within the Big Data", Gupta et al. present the intrusion detection system for detection of anomalies for large-scale environments.

In Chapter "Geospatial Big Data, Analytics and IoT: Challenges, Applications and Potential", Kashyap et al. describe the challenges of geospatial big data and its applications in different diversified fields. Chapter "Geocloud4GI: Cloud SDI Model for Geographical Indications Information Infrastructure Network" discussed Cloud SDI representation named as Geocloud4GI for giving out investigation and

dispensation of geospatial facts particularly for registered Geographical Indications (GIs) in India. The primary purpose of Geocloud4GI framework is to assimilate the entire registered GIs' information and related locations such as statewise and yearwise registered in India. Chapter "The Role of Geospatial Technology with IoT for Precision Agriculture" combines the geospatial technology with IoT for precision to monitor and predict the critical parameters such as water quality, soil condition, ambient temperature and moisture, irrigation, and fertilizer for improving the crop production. It also describes geospatial and IoT in smart farming, and the prediction of the amount of fertilizer, weeds, and irrigation will be accurate and it helps the farmers in making decisions related to all the requirements in terms of control and supply.

In Chapter "Design Thinking on Geo Spatial Climate for Thermal Conditioning: Application of Big Data Through Intelligent Technology", Das et al. explored the design aspect of thermal insulation capacity of rooftops through application of big data and intelligent technologies. It investigates the testing of construction of materials through different kinds of material mix. Chapter "Hyperspectral Remote Sensing Images and Supervised Feature Extraction" describes the few supervised feature extraction techniques for hyperspectral images, i.e., prototype space feature extraction (PSFE), modified Fisher's linear discriminant analysis (MFLDA), maximum margin criteria (MMC) based, and partitioned MMC based methods are explained.

Topics presented in each chapter of this book are unique to this book and are based on unpublished work of contributed authors. In editing this book, we attempted to bring into the discussion all the new trends and experiments that have made cloud computing for geospatial big data analytics. We believe the book is ready to serve as a reference for larger audience such as system architects, practitioners, developers, and researchers.

Bhubaneswar, India — Himansu Das
Bhubaneswar, India — Rabindra K. Barik
Richardson, USA — Harishchandra Dubey
Shillong, India — Diptendu Sinha Roy

Acknowledgements

The making of this edited book was like a journey that we had undertaken for several months. We wish to express our heartfelt gratitude to our families, friends, colleagues, and well-wishers for their constant support throughout this journey. We express our gratitude to all the chapter contributors, who allowed us to quote their remarks and work in this book. In particular, we would like to acknowledge the hard work of authors and their cooperation during the revisions of their chapters. We would also like to acknowledge the valuable comments of the reviewers which have enabled us to select these chapters out of the so many chapters we received and also improve the quality of the chapters. We wish to acknowledge and appreciate the Springer team for their continuous support throughout the entire process of publication. Our gratitude is extended to the readers, who gave us their trust, and we hope this work guides and inspires them.

Contents

Big Data Scientific Workflows in the Cloud: Challenges and Future Prospects

Samiya Khan, Syed Arshad Ali, Nabeela Hasan, Kashish Ara Shakil and Mansaf Alam

Abstract The concept of workflows was implemented to mitigate the complexities involved in tasks related to scientific computing and business analytics. With time, they have found applications in many diverse fields and domains. Handling big data has given rise to many other issues like growing computing complexity, increasing data size, provisioning of resources and the need for such systems to enable working together of heterogeneous systems. As a result, traditional systems are deemed obsolete for this purpose. To meet the variable resource requirements, cloud has emerged as an ostensible solution. Execution and deployment of big data scientific workflows in the cloud is an area that requires research attention before a synergistic model for the same can be presented. This paper identifies open research problems associated with this domain, giving insights on specific issues like workflow scheduling and execution and deployment of big data scientific workflows in a multi-site cloud environment.

Keywords Scientific workflows · SWfMS · Big data · Cloud computing
Workflow scheduling · Multisite cloud

S. Khan (✉) · S. A. Ali · N. Hasan · M. Alam
Department of Computer Science, Jamia Millia Islamia, New Delhi, India
e-mail: samiyashaukat@yahoo.com

S. A. Ali
e-mail: arshad158931@st.jmi.ac.in

N. Hasan
e-mail: hasan.nabeela28@gmail.com

M. Alam
e-mail: malam2@jmi.ac.in

K. A. Shakil
Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India
e-mail: shakilkashish@yahoo.co.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_1

1 Introduction

Scientific and business analytics entails several applications that require the use of scientific workflows to mitigate the complexities involved. In fact, in fields like astronomy, social science, bioinformatics and neurosciences, in addition to several others, scientific workflow management systems are found to be effective, so much so that they are irreplaceable in their realm of usage [1–3].

Similar to the traditional data management systems and computing infrastructures that have been proven insufficient for the challenges posed by big data analytics; similarly, traditional scientific workflows are also unable to mitigate the big data challenges posed by the growing scale and computational complexity of analytics tasks.

Data is being generated in this world at an alarming rate in view of the ever-increasing popularity of social networks like Facebook, eBay and Google+, amongst many others. Most of the data being generated is already on the cloud. The big data available in the world today is expected to rise to 44 ZB by the year 2020, recording a 10 times rise from 2012 [4]. The fundamental challenge in the management of this data is its storage and processing, as the present-day systems cannot support the same.

In addition to the above-mentioned, processing of data makes use of complex, computing intensive algorithms. This requires systems that can handle the processing requirement of data mining algorithms [5, 6], making high performance computing the second requirement of big data analytics that any system claiming to be an effective solution needs to fulfill.

Cloud computing has also found applications in many fields including optimization and inter-disciplinary purposes [7, 8]. As a result, it is an apparent solution to the big data problem [9]. Firstly, the cloud provides an operative storage solution for the huge data storage problem. The cloud adopts pay-as-you-go model, which is primarily why it is able to offer a cost-effective solution to the problem. Besides this, cloud computing allows user to get the hardware required without the need to buy it, giving a scalable solution to the computing hardware requirement.

These are the reasons why most enterprises and the scientific community have chosen to adopt the cloud for big data management and analysis. The advent of concepts like cyber-foraging systems [10], fog computing [11], edge computing and mist computing [12], apart from many others, has facilitated the process. Moreover, the availability of effective, efficient and open-source ecosystems like Hadoop [13] has facilitated the adoption process immensely. The complexity of the problem further intensifies when cloud-based big data analytics requires the execution of scientific workflows, which are data intensive in nature.

There are three main facets of this increasing complexity. Firstly, the types and sources of data are diverse. Secondly, the whole concept of using distributed computing for data processing is based on moving code to the data, which is not always possible because of compatibility issues and proprietorship of the code. Lastly, it is not possible to keep all the data throughout the lifecycle of the execution of a

workflow. Therefore, redundant data needs to be removed. Several workflows have been proposed for running data intensive workflows, which shall be discussed in the following sections.

Evidently, executing data intensive workflows shall require handling of large datasets. Therefore, the most obvious solution to ease the execution process is to use parallelization. Cloud is a great solution to solve the need for unlimited resources in data intensive workflow execution [14]. However, this is faced with several challenges. Performance and cost optimization efforts in this direction are focused towards improving the scheduling algorithm [15], which still remains an area of research interest. Parallelization may be implemented at the single site cloud-level or a multi-site cloud-level. Lately, multi-site cloud parallelization has gained immense research attention. Moreover, the use of workflow partitioning techniques for efficient multi-site cloud parallelization is also being explored.

Existing literature in this area is premature and leaves a lot of scope for future research. The motivation behind this research work is to examine existing approaches and the systems that have been employed to implement and execute big data scientific workflows in the cloud to identify open research problems in this field. In addition to several others, workflow scheduling and execution and deployment of Scientific Workflow Management Systems (SWfMS) in multisite cloud are identified as two key areas where future research in this field can be centered. Further investigation has been performed to identify specific research efforts in these areas. Besides this, the challenges and opportunities with respect to scientific workflows, evolution of SWfMS in the edge computing respect and their deployment have also been discussed.

The organization of the rest of this chapter has been described below. The first section introduces scientific workflows, giving insights into fundamental definitions and concepts related to the topic. As part of this section, a brief comparison of the different scientific management systems and frameworks that support cloud-based execution has also been presented. The next section throws light on how and where the cloud paradigm fit into the scientific workflow concept for big data analytics. The section that follows elaborates on the challenges and issues that arise in bringing together this synergistic approach.

The next section of the paper explains existing systems and research gaps that exist and can be worked upon. This paper identifies Workflow Scheduling algorithms and scientific workflows in the multisite cloud as two key areas of potential future work in this field, which have been discussed in detail. Besides this, it also discusses the challenges and opportunities related to scientific workflows in the era of edge computing. The paper concludes with a remark on existing challenges and future research direction.

2 Background

At the basic level, a workflow can be explained as a logical sequence of activities or data processing tasks, which works on the basis of predefined rules. The fundamental usage of a workflow is to automate any process. Typically, there are two types of workflows—(1) scientific workflows and (2) business workflows. Scientific workflows find applications in the field of scientific computing for automating scientific experiments and processes [16] while the latter is used for automating business processes.

There are several ways in which scientific workflows are represented. The most commonly used representations are Directed Acrylic Graphs (DAG) [17] and Directed Cyclic Graphs (DCG) [14]. The two fundamental entities that need to be described with respect to a scientific workflow are activities and tasks. An activity is a logical step that needs to be performed as part of a scientific workflow [18]. On the other hand, a task is an instance of an activity [19]. Therefore, a task represents the execution of an activity.

The transition of a workflow from its initiation to its completion is termed as the scientific workflow lifecycle. Moreover, the Scientific Workflow Management System (SWfMS) performs initiation and management of workflow execution. Görlach et al. [20] proposed that a scientific workflow is divided into four phases namely deployment phase, modeling phase, execution phase and monitoring phase. There are several scientific workflows that are used as solutions to field-specific and generic problems. Liu et al. [14] compared Swift [21], Pegasus [22], Taverna [23], Kepler [24], Galaxy [25], Chiron [26], Askalon [27], Triana [28] and WS-PGRADE/gUSE [29], of which the first eight are typical Scientific Workflow Management Systems and WS-PGRADE/gUSE is a gateway framework. These systems are briefly discussed and compared in this section.

2.1 Pegasus

This SWfMS is being used by multiple disciplines, which include earthquake science, climate modeling, bioinformatics, astronomy and genome analysis, to name a few. Some of the key features of Pegasus [30] include—

- Portability across infrastructures like grid and cloud
- Scalability
- Optimized scheduling algorithms
- Provenance data support
- Support for data transfer
- Fault tolerance

It is important to mention that the support for data transfer makes Pegasus [30] a good choice for data-intensive applications. Moreover, provenance data support

makes the process of debugging much simpler. Finally, the availability of a package for Pegasus in Debian repository and a detailed user guide are also some crucial features of this SWfMS.

Pegasus [22] consists of five main components, which include monitoring component, remote execution engine, job scheduler, local execution engine and mapper. The user provides an abstract workflow, which is converted into an executable workflow by the mapper. It is the responsibility of the local execution engine to analyze the workflow, along with the sub-workflows, for dependencies. On the basis of this analysis, the fragments of the workflow are submitted to the execution engines.

The execution of the workflow fragments on the remote execution engines need to be scheduled, which is done by the job scheduler. On the other hand, the remote execution engine independently manages the execution of workflow fragments and the monitoring component monitors the whole process of execution. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  The user is allowed to provide an abstract workflow in the form of DAX or DAG in a XML file, which can be generated using the APIs provided by Pegasus [31]. Besides this, it also provides a lightweight web dashboard for monitoring the execution of the workflow.
- User Services Layer
  In this layer, workflow monitoring and provenance management are supported. Pegasus makes use of Pegasus/Wings framework [32] for provenance management and Stampede infrastructure [33, 34] for monitoring. In order to perform these functions, the data is collected from the logs.
- WEP Generation Layer
  As part of this layer, reduction of abstract workflow is performed on the basis of the intermediate data from previous executions. It is the task of the job scheduler to perform site execution and it uses standard algorithms like round robin, random and min-min for this purpose. Besides this, the job scheduler may also take into account factors like data and computation significance or location of data.
- Infrastructure Layer
  The scientific workflows are executed using clouds and grids.

2.2 Swift

Swift [21], like Pegasus, has multi-disciplinary applications. Some of these applications include economics, astronomy and neuroscience. This SWfMS is based on GriPhyN Virtual Data System (VDS), which was created for expression, execution and tracking of workflow results. Moreover, the salient tasks of this system are data management, task management and program optimization and scheduling.

The execution of data intensive workflows in Swift [21] entails five functional phases namely provisioning, provenance management, scheduling, and execution

and program specification. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  This SWfMS allows the user to specify workflows using SwiftScript and XDTM. While the former allows definition of sequential and parallel procedures, the latter performs mapping of data to physical resources.
- User Services Layer
  As part of this layer, users can access provenance data.
- WEP Generation Layer
  In the WEP generation layer, abstract WEPs for each site are generated.
- WEP Execution Layer
  In order to schedule abstract WEPs, the Karajan workflow execution engine is utilized. Some of the main functions performed by this workflow execution engine include task initiation, grid services access, task submission, data transfer and task scheduling [35].
- Infrastructure Layer
  Each of the execution sites possesses a dynamic resource provisioner, which provides access to computing resources like cloud, grid and cluster. In order to manage data staging, task allocation and facilitate communication for execution, Coasters [36] are used.

2.3 Kepler

This SWfMS is a part of the Kepler [24] project and is specifically built upon Ptolemy II system. The most important feature of Kepler [37] is its ability to allow workflows to make use of different execution models. Moreover, the integration of a graphical workbench adds to the power of the system by leaps and bounds. Some of the popular applications of this SWfMS include data management, oceanography and biological process simulation. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  In Kepler, each of the workflow activities like statistical operations and signal processing is associated with different actors to define the complete workflow.
- User Services Layer
  Actors like provenance recorder [16] are used for provenance data management.
- WEP Generation Layer
  A component named director is used for handling the workflow in this layer. Kepler supports different directors for different execution models.
- WEP Execution Layer
  Depending upon the director chosen in the WEP generation layer, static or dynamic scheduling may be used [27, 38]. The fault tolerance feature is provided using three mechanisms namely forward recovery, check pointing and watchdog process.

- Infrastructure Layer
  In order to provide data access, OpenDB connection actor is provided. Moreover, for biological and ecological datasets, EML Data Source actor may be used. The compatibility of Kepler with Cloud is established through Kepler EC2 actors [39].

2.4 *Taverna*

This SWfMS is open source and a component of the myGrid project [23]. It was originally created to work on biological experiments. Some of its application areas include chemistry, bioinformatics and astronomy. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  Taverna provides a GUI to allow the user to provide a DAG representation of the workflow [37].
- User Services Layer
  In order to monitor the workflow, Taverna makes use of a state machine [40]. The provenance data is collected from the remotely invoked web services and local execution information [41].
- WEP Generation Layer
  Complex parts of the workflow are identified and simplified for parallelization and design simplification [42]. In this manner, optimization of workflow structure is done. The generation of WEP is performed after checking for availability of the required services.
- WEP Execution Layer
  Task execution is left to the grid and web services.
- Infrastructure Layer
  Computing resources are provided by the cloud or grid.

2.5 *Chiron*

This SWfMS uses the database approach to execute the scientific workflow, in parallel [26]. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  The data and workflow activities need to be expressed in the form of algebraic expressions.
- User Services Layer
  The provenance data is collected with the help of the algebraic approach. Besides this, features like workflow steering and monitoring are also supported.

- WEP Generation Layer
  In this layer, the workflow is represented as a conceptual model in a XML file. Workflow schedules are optimized by differentiation between blocking activities. Besides this, this SWfMS supports parallelism of different types including pipeline, data, independent and hybrid parallelism.
- WEP Execution Layer
  The input data, database information and scheduling method is specified in the execution module file. Dynamic scheduling is used for executing tasks. PROV-Wf [43] provenance model is used for collection of provenance data, execution data and light domain data. MPJ [27], which is similar to MPI message passing system, is used for executing tasks.
- Infrastructure Layer
  Data storage is done using database and shared-disk file system. Moreover, compatibility with the cloud is established via the extension, Scicumulus [44, 45].

2.6 Galaxy

This SWfMS is web-based and was specifically developed to support genomics research [25]. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  The workflow can be expressed using a web-based GUI, which can be installed on private or public server.
- User Services Layer
  Data can be uploaded from user's personal computer. In addition to this, workflow information, including provenance data, can be shared on a public website.
- WEP Generation Layer
  In order to implement workflow parallelization, dependencies between different activities are managed by Galaxy.
- WEP Execution Layer
  Dynamic scheduling is used for dispatching executable tasks. On the other hand, in order to execute tasks, Gridway is used.
- Infrastructure Layer
  For achieving storage provisioning and dynamic computing, Globus [46] and CloudMan [47] are used. CloudMan middleware is incorporated for adapting Galaxy for the cloud.

2.7 Triana

The GEO 600 project10 developed Triana as a tool for data analysis [28, 48]. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  As part of the presentation layer, Triana provides a GUI.
- User Services Layer
  Stampede monitoring infrastructure [48] is implemented as part of this layer for monitoring of workflows.
- WEP Generation Layer
  The different data processing systems are realized using different components. The concept of components is similar to that of actors used in Kepler.
- WEP Execution Layer
  GAT (Grid Application Toolkit) is used for development of grid-oriented components while GAP (Grid Application Prototype) is used as an interface for interaction between service-oriented networks.
- Infrastructure Layer
  Computing resources can be used from the cloud. In order to run scientific workflows on the cloud, communication between virtual machines needs to be established. For this purpose, RabbitMQ12 11, a message broker platform, is used.

2.8 Askalon

Askalon [27, 49] is a scientific workflow management system that was originally developed for the grid environment. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  A GUI is provided, which uses UML (unified modeling language) for expressing workflows.
- User Services Layer
  Users can monitor the workflows online. Moreover, Askalon also offers dynamic workflow steering to handle unforeseen execution environments and dynamically occurring exceptions [50].
- WEP Generation Layer
  Optimization of workflow representations is done in Askalon with the help of loops.
- WEP Execution Layer
  Askalon makes use of static as well as hybrid scheduling. In addition, fault tolerance is provided by the system using an execution engine.

- Infrastructure Layer
  Availability of resources and deployment of executable tasks on the same are managed by a resource manager. Dynamic creation of virtual machines can be done for execution of scientific workflows in the cloud [51]. Cost estimation [51] and cost-effective dynamic task and resource scheduling are also offered in the cloud. Moreover, execution on the federated multi-site cloud is also known [52].

2.9 WS-PGRADE/gUSE

WS-PGRADE/gUSE [29, 53] is a gateway framework that has found widespread applications in fields like seismology, biology, neuroscience and astronomy. In addition, it is also used in teaching, research and commercial applications. Users are allowed to design scientific workflows using a web portal. The user services are provided using Grid and Cloud User Support Environment (gUSE). Lastly, access to the grid and cloud are provided using a web-based application called DCI Bridge [54]. The breakdown of the different layers of the workflow execution process is as follows:

- Presentation Layer
  A web-based interface can be used for defining workflows. This gateway framework can also be used to define parameter sweep workflows and meta-workflows.
- User Services Layer
  There is an inbuilt repository that allows sharing of information between users of the SWfMS. A workflow template is provided that can be modified to adjust the parameters for other workflows. The gUSE services allow the users to monitor these workflows. It is important to mention that this SWfMS does not support provenance data management.
- WEP Generation Layer
  In this layer, a workflow is represented in XML. On the basis of the workflow structure, data and independent parallelism are supported [19]. Scheduling of tasks is performed by DCI Bridge in a dynamic manner. On the other hand, tasks are executed by web services, which are in turn enabled using web containers.

2.10 Comparison of Different Scientific Workflows

All the systems and frameworks discussed in this paper support dynamic scheduling, independent parallelism and cloud-based scientific workflow execution. A summary of the feature-based comparison of these systems and frameworks is given in Table 1.

Table 1 Comparison of SWFMS and workflows

SWfMS/Feature	Workflow structure	Static scheduling	Information sharing	User interface	Special feature	Applications
Pegasus [22]	DAG	✓	×	Textual	High performance and scalable	Used for executing workflows for astronomy, biology etc.
Swift [21]	DAG and DCG	×	×	Textual	High performance and scalable	Used for executing workflows for astronomy, biology etc.
Taverna [23]	DAG	×	✓	GUI	Desktop-based GUI	Used for executing workflows for astronomy, biology etc.
Kepler [24]	DAG and DCG	✓	×	GUI	Desktop-based GUI	Used for executing workflows for astronomy, biology etc.
Galaxy [25]	DAG and DCG	×	✓	GUI (accessible from web)	Web-based system for genomic research	Used for executing bioinformatics workflows only
Chiron [26]	DAG and DCG	×	×	Textual	Workflow parallelization based on algebraic approach	Large scale scientific experiments
Askalon [27]	DAG and DCG	×	✓	GUI (desktop and Web)	Allows execution of scientific workflows on a multi-site cloud environment	Scientific applications
Triana [28]	DAG and DCG	×	×	GUI	Can use P2P services	E-Science applications
WS-PGRADE/gUSE [29]	DAG	✓	✓	GUI	Allows execution of scientific workflows in distributed computing infrastructures	Used for executing workflows for astronomy, biology, seismology and neuroscience

3 Scientific Workflows in the Cloud

Workflows are described as complex graphs, unfolding the concurrent tasks that any concerned application may include. Evidently, any data analytics task will require data access, processing and visualization. Therefore, workflow tasks must address these components of the system effectively and efficiently.

An obvious approach for using workflows in the cloud environment is to refactor the existing scientific workflows according to the Cloud Computing paradigm. One of the first works that proved the viability and effectiveness of the cloud for running scientific workflows was Keahey and Freeman [55]. As part of this research, a Nimbus Cloudkit was developed, which was made available to the scientific community for satisfying their infrastructure and resource needs.

Vöckler et al. [56] implemented a cloud-based scientific workflow application for processing astronomical data. As part of this project, Pegasus workflow management system was deployed on multiple and different clouds to test the viability of the system for the application stated. The findings of this experiment included a conclusion that user experience was not affected by the underlying differences in the different clouds used and users were able to accomplish basic tasks with ease, indicating that the management overheads of the system were ignorable.

Although, this seems like an obvious option, it is challenging in consideration of the high complexity of scientific workflows. Developers will have to invest a huge amount of effort and time to implement the application logic and mitigate the issues involved in the integration of cloud with workflow logic.

A more feasible approach is the integration of SWfMS into the cloud environment. In this way, traditional workflows will not have to be refactored, addressing the challenges associated with the same, and they can still be used to process cloud data. This concept has given rise to Cloud Workflow platform, which is provisioned to the users as a service.

There are several advantages of using this approach. Some of the fundamental benefits include scalability, flexible resource allocation, easier deployment of applications and a better return on investment from the organization's point of view. However, what this approach also adds to this list is an increased overhead, but this facet can be ignored keeping in mind the multi-fold benefits of scientific workflow management systems that this approach allows researchers to leverage.

Typically, workflow-based data mining on the cloud makes use of the service-oriented approach. According to Talia [57], there are three advantages of using this approach namely, execution scalability, distributed task interoperability and a flexible programming model. The scalable model of execution provided by this approach helps in considerably reducing the completion time of the task. There are many frameworks and architectures that have been proposed in this regard. They have been discussed below.

Lin et al. [58] proposed one of the first architectures given for Scientific Workflow Management Systems. It introduced a four-layer architecture with workflow management layer, presentation layer, operational layer and task management layer,

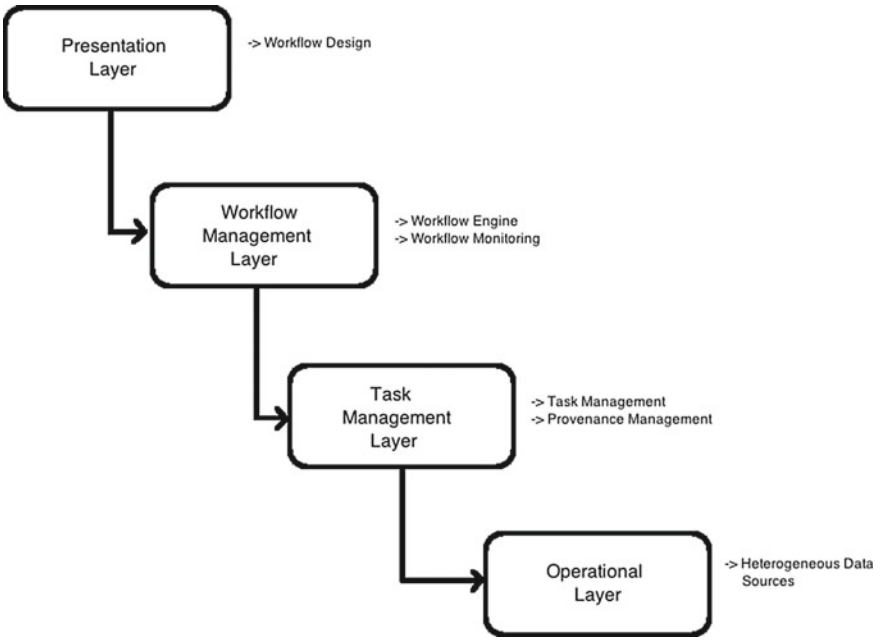


Fig. 1 Reference architecture by Lin et al. [58]

being the four layers, considering visualization, analytical engine and data acquisition. However, this architecture was a base architecture that had no provisions for management of security issues. Figure 1 illustrates the architecture and the components that it includes.

Zhao et al. [59] proposed a service framework for integration of Eucalyptus and OpenNebula with Swift Workflow Management System that addresses these challenges. Figure 2 illustrates the basic layout of this framework. They have also presented an implementation based on this service framework and makes use of the Montage Image Mosaic Workflow.

However, porting of other traditional workflows like VIEW and Taverna is yet to be explored in addition to investigating the possibility of automatic deployment of workflow applications in the virtual cluster. This framework provides the background for designing efficient SWfMSs. The functional architecture given by Liu et al. [14] is shown in Fig. 3. The user services layer caters for user functionality while the presentation layer is the GUI that shall be presented to the user for giving instructions to the system.

The WEP generation layer generates the workflow execution plan (WEP). This layer is used by the system to interpret the user instructions and determine the flow of execution of the workflow. The generated WEP is given to the WEP execution layer for execution. The last layer is the infrastructure layer. It delivers the required computing and storage resources.

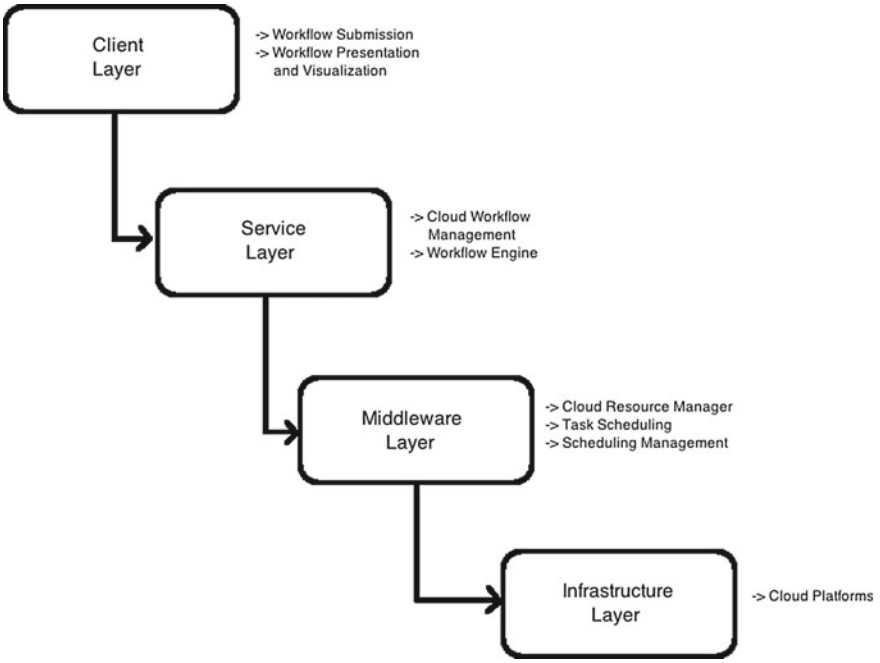


Fig. 2 The service framework by Zhao et al. [59]

Li et al. [15] proposed scientific workflow management system architecture for manufacturing big data analytics. The architecture is based on the three architectures previously mentioned and divides the system into four layers namely, infrastructure, management, service and application layers. This system is implemented as a secondary system over Kepler, an existing system.

4 Challenges and Opportunities

Traditionally, scientific workflows are implemented on grids [37] or clusters, workstations and supercomputers. This implementation faces several limitations and obstacles, the most profound of which are scalability and computing complexity. Apart from these, several other issues like resource provisioning have also been known [60]. This section discusses some of the most prominent challenges facing the development and use of scientific workflows for big data analytics in the cloud environment.

The input that goes into scientific workflows and the output that comes out of the same are data objects that are distributed in nature. The type, size and complexity of these data objects shall vary. There is a rapid increase in the data coming from

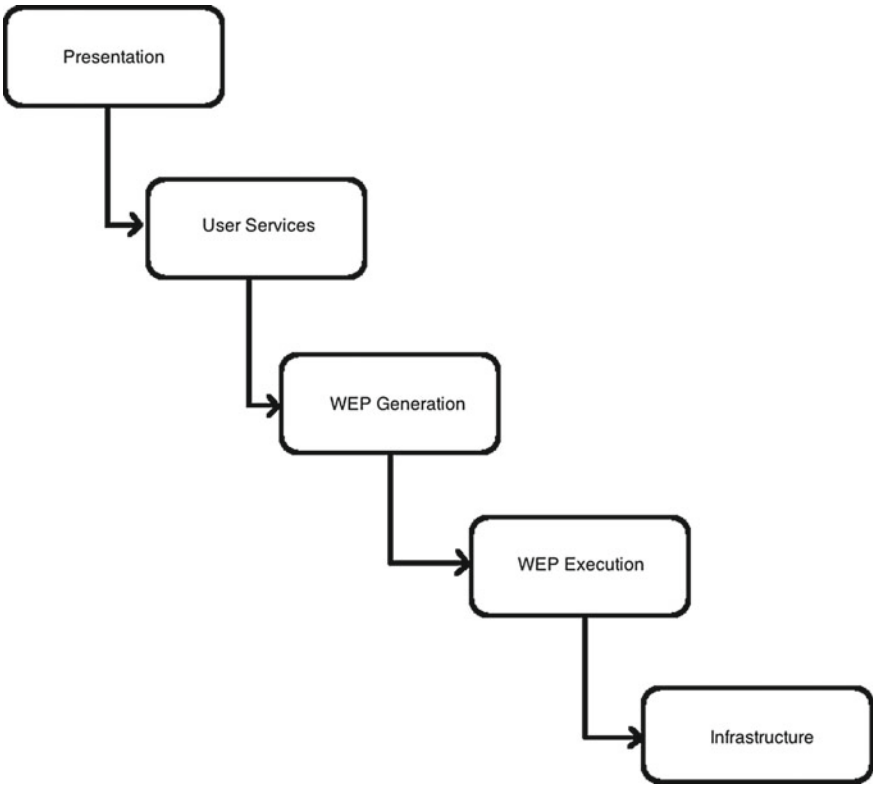


Fig. 3 SWfMS architecture by Liu et al. [14]

the different sources of scientific computing like sensors, experiments and networks, giving rise to 'data deluge' [61].

According to this concept, the whole theory of scientific research is changing. While, earlier, a question was asked and data was collected that can answer the question, today, we have heaps of data and retrospectively, what we are looking for is the question that this data heap can answer. Evidently, the scale and size of data are huge. In order to handle the growing complexity of data storage and processing, several thousands of computation nodes may be used. This will not only make the operation feasible, but it shall also make it efficient.

Research has always been a collaborative work and as the world is converging after the advent of Internet, geographical boundaries and distances are no longer a limitation. However, scientific projects running at varied geographical locations add to the complexity of using workflow management systems in view of the fact that the execution environments used by different organizations may be different.

Therefore, the interactions between the execution environment on the host system and the workflow management system need to be smoothened for resource management, security and load balance, besides others [59]. When it comes to heterogeneous

execution of scientific workflows, the varied computational ability and performance of systems may also impact the traditional scientific workflow execution. In addition to the above-mentioned, there are two main areas of possible future research namely, workflow scheduling and execution and deployment of scientific workflows on the cloud. Besides this, the growing popularity of edge computing has given rise to many integration challenges. These challenges and opportunities for future research have been discussed in this section.

4.1 Workflow Scheduling

The process of allocating resources to a scientific workflow for operation in terms of storage and computing resources is termed as resource provisioning. Resource allocation in cooperative cloud environments is a hot topic of research [62]. Typically, a scientific workflow is allocated resources as and when it is deployed. Moreover, these resources are fixed prior execution. Evidently, the scale of the scientific workflow is limited by resource allocation. At another level, the scale of scientific workflow is also limited by the total resource pool size. In order to allow smooth and scalable execution of scientific workflows, efficient dynamic resource provisioning [63, 64] shall be required.

There are several other challenges in this category of research. One of the fundamental challenges that need to be clearly addressed in this context is how a resource can be best represented for scientific workflows [14]. Every workflow has a set of supported tools and resource types. Identifying whether a resource type or tool is compatible with a workflow is also a daunting research task. Recent work in this field has been focused towards automated provisioning and developing algorithms for effective resource provisioning. One such system designed for the cloud to manage provisioning of resources for the workflow is the wrangler system [65].

Another critical issue in management systems is to devise an efficient scheduling algorithm. The objective of a scheduling algorithm is mapping of tasks and resources. In the present context, this should be distributed and heterogeneous in nature. Therefore, there is a need for deterministic workflows that require task parameters and resource configuration or availability. However, the workflow must not need an input on where these resources are located [66]. In view of this, workflows that require an abstract mathematical model as input are chosen. The workflow is represented in the form of a DAG. Here, nodes represent tasks and edges are indicative of the relationships between tasks [64]. Several heuristic and meta-heuristic algorithms are present to solve this NP Complete problem keeping in mind the resource QoS and execution cost [9, 63, 67, 68].

Execution of scientific workflows can be performed on Hadoop YARN. Hi-WAY [69] gives an engine for scientific workflow execution. In other words, it provides an application master that can control scientific workflow task execution on top of YARN. However, some fundamental shortcomings exist in this engine. One such shortcoming comes from the fact that the containers allocated for task execution are

uniform for all applications and tasks, which leaves room for optimization. Therefore, customized container allocation can considerably improve the performance of the engine.

Several challenges are involved in adopting the cloud environment for scheduling workflow applications due to resource heterogeneity and on-demand services offered by cloud service providers using pay-per-use model. It is the main challenging issue in IaaS (Infrastructure-as-a-Service) clouds. In recent years, various scalable and dynamic algorithms were proposed. In order to facilitate adaptation to changes in workload and environment, enhancements were made to existing algorithms.

Mapping of the distributed resources while satisfying user's quality of service (QoS) parameters become a tedious job. Many heuristic and meta-heuristic techniques are applied by the researchers for scheduling workflow applications to generate near-optimal solution for such problems, but as workflow scheduling is an NP-complete problem, meta-heuristic approaches are more preferred options. Most researchers focused on developing nature inspired meta-heuristic algorithms [70] like ACO (Ant Colony Optimization) [71], GA (Genetic Algorithm) [72], PSO (Particle Swarm Optimization) [73, 74], and SA (Simulated Annealing) [75], for solving multi-objective workflow scheduling problem.

Filgueira et al. [76] presented a Workflows-as-a-Service (WaaS) model for data intensive applications. As part of this model, containers are used to deploy stream-based workflow applications in the cloud, which makes the approach particularly easy. On similar lines, Esteves and Veiga [77] middleware, models like Skyport [78] and architecture presented by Wang et al. [79] allow management of multiple workflows' execution in the cloud. In order to develop a well-defined VM sharing model, algorithms must be tailor-made for WaaS platform. In other words, they must be scalable, efficient in auto-scaling and capable of making decisions quickly, which potentially lead to lower costs and higher profit. Elastic Resource Provisioning and Scheduling (EPSM) [80] was proposed in response to these requirements, which includes containers for addressing issues like resource utilization inefficiencies, minimization of the overall costing involved in resources' rental, addressing and adhering to the deadline constraints of workflows and producing higher quality schedules.

Enhancements to a meta-heuristic solution like Shuffled Frog Leaping Algorithm (SFLA) were proposed. An Augmented Shuffled Frog Leaping Algorithm (ASFLA) [81] is an enhancement to an SFLA algorithm, which aims at reducing the total execution time, in addition to a considerable decrease in specified deadlines for the workflow execution. It has proven to outperform PSO and SFLA. Another algorithm proposed to optimize workflow schedule length, also known as make span, is Discrete Binary Cat Swarm Optimization (DBCSO) [82]. It is an enhanced version of Cat Swarm Optimization (CSO), which simulates and uses cat behavior [82, 83] to solve an optimization problem. The discrete version of CSO, which is also called DBCSO, is adopted for obtaining solutions for binary knapsack problem [84] and travelling salesperson problem [85] by giving lesser make span as compared to standard PSO and binary PSO because of its seeking mode.

Hybridization of various popular meta-heuristic algorithms has also proven to be good in terms of performance. For example, Hybrid Particle Swarm Optimization (HPSO) [86], which is a hybrid of Budget and Deadline-constrained heterogeneous Earliest Finish Time (BDHEFT) algorithm and multi-objective PSO, tries to optimize the makespan and cost under the deadline and budget constraints. Another algorithm, which is a hybridization of Heterogeneous Earliest Finish Time (HEFT) and Gravitational Search Algorithm (GSA) that schedule length ratio (SLR), considers monetary cost ratio (MCR) as the metrics for comparison of performance of the proposed with the existing algorithms. Algorithms that are hybridized versions of meta-heuristic approaches are known to perform better than the base algorithms [87–91]. It is because of these reasons that this research problem has gained immense attention, particularly in the area in cloud computing. Many other options for resource selection can be explored that can affect the performance of the algorithm.

Verma and Kaushal [92] proposed an MP (Max Percentages) algorithm, which resembles Min-Min, Sufferage and Max-Min algorithms and was tested to outperform classic algorithms both in terms of load balancing and time of completion. The limitation of Min-Min algorithm is that it does not guarantee load balancing, particularly in a situation where the use of some resources for computing is more advantageous than others [93, 94]. However, it does have a good impact on the workflow's total completion time. On the other hand, Max-Min algorithm [95] overcomes this limitation of Min-Min algorithm providing a good total completion time and load balance, but it is not the preferred algorithm in cases where number of tasks taking a longer time outnumber the number of tasks that are comparatively shorter.

The Sufferage [96] algorithm is a heuristic algorithm that is known to be better than Max-Min and Min-Min when it comes to handling resource variations. However, it is not deemed appropriate for data intensive applications in which reuse of files is common [97]. A hybrid algorithm, called MP algorithm [97], uses Min-Min and Sufferage algorithms for minimization of total completion time and considers heterogeneous resources and all the information about the workflow to balance load. Therefore, this algorithm combines the advantages of the three base algorithms to provide the best optimized solution to the problem. Moreover, in order to optimize performance, this algorithm makes use of the intrinsic correlation between tasks and resources.

These algorithms can be extended by inclusion of various parameters such as options for advance reservation and pre-emptive jobs. Future work in this field may also include addition of more clouds for distribution of workload and improvement in performance. Presently, email alerts are provided as a service. This functionality can be extended to support storage of online data and provide a synchronization mechanism. Many other options for resource selection can be explored that can further improve the performance of the algorithm. Table 2 makes an elaborate comparison of these workflow scheduling algorithms.

Table 2 Comparison of workflow scheduling algorithms

Algorithm	Modified version	Scheduling factors	Tools	Key findings
Particle swarm optimization (PSO)	Standard PSO [98]	Cost optimization, time, resource utilization	Amazon EC2	Minimization of cost, makespan, QoS increment
	Discrete PSO [99]	Cost, time, and security	CloudSim	Reducing cost while considering security
	Hybrid PSO [92]	Cost, makespan and energy consumption	Amazon EC2	Better performance and minimization of costs within a given execution time
Cat swarm optimization (CSO)	Standard CSO [100]	Self-position consideration (SPC), seeking range of the selected dimension (SRD), mixture ratio (MR), counts of dimension to change (CDC), seeking memory pool (SMP), random value (r1) and constant (c1)	WorkflowSim	Load distribution and minimization of cost
	Discrete binary CSO [82]	Seeking range of the selected dimension (SRD), self-position considering (SPC), counts of dimension to change (CDC) and seeking memory pool (SMP)	MATLAB	Minimizing makespan and minimization of cost and CPU idle time
Shuffled frog leaping algorithm (SFLA)	Standard SFLA [101]	Number n of frogs in a memeplex, number m of memeplexes, number N of evolution or infection steps in a memeplex between two successive shuffling, number q of frogs in a submemeplex and the maximum step size Smax allowed during an evolutionary step	Evolver	Effective solution for combinatorial optimization problems, a highly efficient computing performance and good global search capability
	Augmented SFLA [81]	Memetic iterations, no of memeplex, and periodicity (n)	A customized simulator created in Java	The overall execution cost for the considered workflow is reduced. Moreover, this algorithm outperforms its counterparts in minimizing execution cost and adhering to the scheduled deadlines
Hybrid of min-max, min-min and sufferage algorithm	Max-percentages (MP) algorithm [15]	Crossover probability, population size, fitness, replacing mutation probability, stopping condition, selection scheme and initial individuals	Java	This algorithm outperforms classic algorithms in improving load balancing level and total completion time

4.2 Execution and Deployment of Scientific Workflows in Cloud

Most of the research work being performed in this field is related to scientific workflow management systems' deployment in the cloud environment and their execution in the multisite cloud. A relevant work in this area provides architecture for efficient execution of scientific workflow management systems using the distributed approach [102]. The described architecture has shown significant cost reduction and can be treated as a good base architecture, which can be improved using dynamic scheduling, distributed provenance management, use of multisite spark and better data transfer techniques.

The master-slave architecture can be changed to peer-to-peer architecture for improving the system. A significant attempt in this direction has incorporated multi-objective scheduling [67] in the architecture. It has been proposed that location awareness can bring about a noteworthy improvement in data transfer issues and performance [103].

Existing literature indicates research work on provenance management, metadata management [104] and big data management [105] in the multisite environment. In order to effortlessly amalgamate the SWfMSs in the cloud environment, it is important to include cloud configuration parameters and resource descriptions to provenance data. Ahmad [106] describes how execution reproducibility is affected in the light of cloud-aware provenance.

Effective management of metadata and big data generated in the SWfMS can bring considerable improvements in the performance of the system. It is shown that the use of distributed approach for metadata management betters the system performance by as much as 50% [103]. This approach can be extended for heterogeneous multisite environment.

Evidently, the data distributed across geographical locations suffers from management issues like cost-related trade-offs, low latency and high throughput. Challenges are all the more magnified considering the volume of big data. An implementation on Azure cloud presents a uniform data management system. This system allows spreading of data amongst geographically separated locations. However, the existing system uses per-site registration of metadata. It is proposed that a hierarchical system of global nature must replace the existing metadata registration system.

While working in the multisite cloud environment, it is important to mention that the included clouds may be heterogeneous in nature and the individual cloud providers are yet to provide interoperability. To explore the management and deployment of workflows over heterogeneous clouds, a broker-based framework was proposed by Jrad et al. [107], which allows automatic selection of target clouds. Kozlowsky et al. [108] provided an internal architecture to enable compatibility for workflow management systems by resolving the DCI interoperability issues prevalent at the middleware level.

Some other research in the area has been targeted towards creating specific applications for domains like astronomy [2], geo-data analysis [1] and bioinformatics [3],

in addition to several others. Talia [57] demonstrated the use of Data Mining Cloud framework and established the effectiveness and linear scalability of this framework in bioinformatics, network intrusion and genomics. However, this framework was not tested for big data and complex data mining.

4.3 The Edge Computing Perspective

With the increasing use of devices like sensor-based wearable systems, tablets and smart-phones, more applications are drifting towards an architecture that places one server at the center and process the data generated by all these devices using the same. However, there is an inevitably increasing demand for computational infrastructure and communication, which puts quality of service into questionable domain.

The concept of edge computing [109] has been developed to push a section of the computational ability to the edge of the network. In this manner, the computational power tapped in the edge nodes like switches, routers and stations can be effectively brought to use. It will be most appropriate to state that edge computing is a method that is used for optimization of cloud computing systems or applications by taking the services or data of the system away from the core or center to the edges or logical extremes of the Internet.

Theoretically, it is possible to facilitate edge computing on the many nodes that exist between the edge device and cloud. These nodes include switches, gateways, routers and base stations. However, base stations may not be an appropriate choice for this purpose, as they possess digital signal processors or DSPs, which are not designed for general purpose computing. Moreover, they are designed for customized workloads and whether or not it is possible to run additional workloads on them is not known. It is possible to upgrade the edge nodes' resources in order to enable them for general purpose computing [109]. Besides this, it is also possible to replace DSPs with general-purpose processors to enable computing, but this is expected to call for a huge investment.

Recently, there have been some efforts by many commercial vendors towards using software solutions for realization of the Edge Computing concept. An example of such efforts include Cisco's IOx16, which provides an execution environment that can be run on service routers that are integrated in its system [110]. Evidently, any software solution created in this regard is specific to the hardware and cannot be expected to work well in heterogeneous environments. This brings us to one of the most crucial challenges in this domain. There is a need to develop software solutions that are portable and can be used across different environments.

Many techniques have been developed for facilitation of task partitioning at varied geographic locations [111, 112] in view of the changing requirements of distributed computing settings [113, 114]. Usually, task partitioning needs to be explicitly expressed. However, when the computational tasks are offloaded to edge nodes, the biggest challenge posed is to allow the system to partition tasks automatically without the need to define the location or capabilities of the edge nodes. Moreover,

there is an inherent requirement from the system to offer flexibility in defining a computational pipeline, which also requires the need for development of schedulers that shall allow deployment of partitioned tasks onto the edge nodes.

Workflows are typically employed in scenarios where the resources from outside the cloud need to be used. For instance, if the input data needs to be taken from a private database and processed on a public cloud, then a workflow is the most appropriate option available. As previously mentioned, toolkits and software frameworks for data-intensive workflows have gathered enormous research attention. The addition of edge nodes in the scenario poses a challenge and can be seen as an opportunity for research in the development of toolkits and software frameworks that can allow effective inclusion of edge nodes, configured for general-purpose computing, in the distributed computing scenario.

It is also important to note that scientific workflows are majorly focused towards fields like astronomy and bioinformatics. On the other hand, the use cases for edge analytics are different and dedicated towards user-driven applications. Therefore, traditional workflows may not be appropriate for expressing applications associated with edge analytics. The fundamental requirements of the programming model configured for exploiting the competences of edge computing include support for data and task-level parallelism, in addition to execution of workloads on several hierarchical stages of hardware.

In order to implement such a programming model, the concerned programming language must consider the resource capacity of the workflow and heterogeneous elements in the hardware. The system may be faced with instances where the edge nodes may be vendor-specific and any framework attempting to support such a workflow must be able to account for the same. This requirement increases the complexity of the model, manifold.

5 Conclusion

Scientific Workflow Management Systems use a viable approach for integrating scientific workflows with cloud-based big data analytics. Some of the main advantages of using this approach are better scalability, higher flexibility and easier deployment. In view of these benefits, several frameworks, architectures and scheduling algorithms have been proposed.

However, research on scientific workflow management systems is still in its infancy. This chapter compares the different frameworks and architectures proposed and in use for exploiting the power of scientific workflows in the cloud environment. It explores the possibility of using these systems for big data analytics. While many systems have been designed and implemented, security issues remain unaddressed in all these solutions.

In addition, there are many open research areas like possibility of automatic deployment of workflow applications in the virtual cluster and the ever-growing need to improve the performance of workflow scheduling algorithms. Moreover,

optimization of scheduling algorithms, multi-site cloud execution of workflows and tailoring the existing systems to help them unlock the power and capacity of edge computing need attention in future research related to this field.

References

1. Gao, S., Li, L., Goodchild, M.F.: A scalable geoprocessing workflow for big geo-data analysis and optimized geospatial feature conflation based on Hadoop. In: CyberGIS All Hands Meeting (CyberGIS AHM'13) (2013)
2. Hoffa, C., Mehta, G., Freeman, T., Deelman, E., Keahey, K., Berriman, B., Good, J.: On the use of cloud computing for scientific workflows. In: IEEE Fourth International Conference on eScience, 2008, eScience'08, pp. 640–645. IEEE (2008)
3. Kashyap, H., Ahmed, H.A., Hoque, N., Roy, S., Bhattacharyya, D.K.: Big data analytics in bioinformatics: a machine learning perspective. arXiv:1506.05101 (2015)
4. IDC. EMC Digital Universe with Research & Analysis. EMC.com. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed 12 March 2018
5. Das, H., Naik, B., Behera, H.S.: Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In: Progress in Computing, Analytics and Networking, pp. 539–549. Springer, Singapore (2018)
6. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of intrusion detection using data mining techniques. In: Progress in Computing, Analytics and Networking, pp. 753–764. Springer, Singapore (2018)
7. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud Computing for Optimization: Foundations, Applications, and Challenges, vol. 39. Springer (2018)
8. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (eds.): Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2017, vol. 710. Springer (2018)
9. Khan, S., Shakil, K.A., Alam, M.: Cloud-based big data analytics—a survey of current research and future directions. In: Big Data Analytics, pp. 595–604. Springer, Singapore (2018)
10. Panigrahi, C.R., Tiwary, M., Pati, B., Das, H.: Big data and cyber foraging: future scope and challenges. In: Techniques and Environments for Big Data Analysis, pp. 75–100. Springer, Cham (2016)
11. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., Mankodiya, K.: Fog assisted cloud computing in era of Big Data and Internet-of-Things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
12. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Das, H.: Mist-gis: optimizing geospatial DATA analysis using mist computing. In: Progress in Computing, Analytics and Networking, pp. 733–742. Springer, Singapore (2018)
13. Reddy, K.H.K., Das, H., Roy, D.S.: A Data Aware Scheme for Scheduling Big-Data Applications with SAVANNA Hadoop. Futures of Network. CRC Press (2017)
14. Liu, J., Pacitti, E., Valduriez, P., Mattoso, M.: A survey of data-intensive scientific workflow management. J. Grid Comput. **13**(4), 457–493 (2015)
15. Li, X., Song, J., Huang, B.: A scientific workflow management system architecture and its scheduling based on cloud service platform for manufacturing big data analytics. Int. J. Adv. Manuf. Technol. **84**(1–4), 119–131 (2016)
16. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-Science: an overview of workflow system features and capabilities. Future Gener. Comput. Syst. **25**(5), 528–540 (2009)
17. John, S., Mohamed, M.: A network performance aware QoS based workflow scheduling for grid services. Int. Arab J. Inf. Technol. (2016)

18. Bux, M., Leser, U.: Parallelization in scientific workflow management systems. arXiv:1303.7195 (2013)
19. Chen, W., Deelman, E.: Partitioning and scheduling workflows across multiple sites with storage constraints. In: International Conference on Parallel Processing and Applied Mathematics, pp. 11–20. Springer, Berlin, Heidelberg (2011)
20. Görlach, K., Sonntag, M., Karastoyanova, D., Leymann, F., Reiter, M.: Conventional workflow technology for scientific simulation. In: Guide to e-Science, pp. 323–352. Springer, London (2011)
21. Zhao, Y., Hategan, M., Clifford, B., Foster, I., Laszewski, G.V., Nefedova, V., Raicu, I., Stef-Praun, T., Wilde, M.: Swift: fast, reliable, loosely coupled parallel computation. In: 2007 IEEE Congress on Services, pp. 199–206. IEEE (2007)
22. Deelman, E., Vahi, K., Juve, G., Rynge, M., Callaghan, S., Maechling, P.J., Mayani, R., et al.: Pegasus, a workflow management system for science automation. Future Gener. Comput. Syst. **46**, 17–35 (2015)
23. Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T., Goble, C.: Taverna, reloaded. In: International Conference on Scientific and Statistical Database Management, pp. 471–481. Springer, Berlin, Heidelberg (2010)
24. Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., Mock, S.: Kepler: an extensible system for design and execution of scientific workflows. In: 16th International Conference on Scientific and Statistical Database Management, 2004. Proceedings, pp. 423–424. IEEE (2004)
25. Goecks, J., Nekrutenko, A., Taylor, J.: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. **11**(8), R86 (2010)
26. Ogasawara, E., Dias, J., Oliveira, D., Porto, F., Valduriez, P., Mattoso, M.: An algebraic approach for data-centric scientific workflows. Proc. VLDB Endow. **4**(12), 1328–1339 (2011)
27. Fahringer, T., Prodan, R., Duan, R., Hofer, J., Nadeem, F., Nerieri, F., Podlipnig, S., et al.: Askalon: a development and grid computing environment for scientific workflows. In: Workflows for e-Science, pp. 450–471. Springer, London (2007)
28. Curcin, V., Ghanem, M.: Scientific workflow systems-can one size fit all? In: Cairo International Biomedical Engineering Conference, 2008, CIBEC 2008, pp. 1–9. IEEE (2008)
29. Kacsuk, P., Farkas, Z., Kozlovszky, M., Hermann, G., Balasko, A., Karoczkai, K., Marton, I.: WS-PGRADE/gUSE generic DCI gateway framework for a large variety of user communities. J. Grid Comput. **10**(4), 601–630 (2012)
30. Yildiz, U., Guabtni, A., Ngu, A.H.: Business versus scientific workflows: a comparative study. In: 2009 World Conference on In Services-I, pp. 340–343. IEEE (2009)
31. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. J. Internet Serv. Appl. **1**(1), 7–18 (2010)
32. Altintas, I., Barney, O., Jaeger-Frank, E.: Provenance collection support in the Kepler scientific workflow system. In: International Provenance and Annotation Workshop, pp. 118–132. Springer, Berlin, Heidelberg (2006)
33. Ganga, K., Karthik, S.: A fault tolerant approach in scientific workflow systems based on cloud computing. In: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), pp. 387–390. IEEE (2013)
34. Ostermann, S., Prodan, R., Fahringer, T.: Extending grids with cloud resource management for scientific computing. In: 10th IEEE/ACM International Conference on Grid Computing, 2009, pp. 42–49. IEEE (2009)
35. Sarkhel, P., Das, H., Vashishtha, L.K.: Task-scheduling algorithms in cloud environment. In: Computational Intelligence in Data Mining, pp. 553–562. Springer, Singapore (2017)
36. De AR Gonçalves, J.C., de Oliveira, D., Ocaña, K.A., Ogasawara, E., Mattoso, M.: Using domain-specific data to enhance scientific workflow steering queries. In: International Provenance and Annotation Workshop, pp. 152–167. Springer, Berlin, Heidelberg (2012)
37. Yu, J., Buyya, R.: A taxonomy of workflow management systems for grid computing. J. Grid Comput. **3**(3–4), 171–200 (2005)

38. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the Kepler system. Concurr. Comput. Pract. Exp. **18**(10), 1039–1065 (2006)
39. Wang, J., Altintas, I.: Early cloud experiences with the Kepler scientific workflow system. Procedia Comput. Sci. **9**, 1630–1634 (2012)
40. Kim, J., Deelman, E., Gil, Y., Mehta, G., Ratnakar, V.: Provenance trails in the wings/Pegasus system. Concurr. Comput. Pract. Exp. **20**(5), 587–597 (2008)
41. Mangala, N., Ch, J., Shashi, S., Subrata, C.: Galaxy workflow integration on Garuda grid. In: IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 194–196 (2012)
42. Mattoso, M., Werner, C., Travassos, G.H., Braganholo, V., Ogasawara, E., Oliveira, D., Cruz, S., Martinho, W., Murta, L.: Towards supporting the life cycle of large scale scientific experiments. Int. J. Bus. Process Integr. Manag. **5**(1), 79–92 (2010)
43. Terstyanszky, G., Kukla, T., Kiss, T., Kacsuk, P., Balaskó, Á., Farkas, Z.: Enabling scientific workflow sharing through coarse-grained interoperability. Future Gener. Comput. Syst. **37**, 46–59 (2014)
44. Kacsuk, P.: Science Gateways for Distributed Computing Infrastructures. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-11268-8_10
45. Bergmann, R., Gil, Y.: Retrieval of semantic workflows with knowledge intensive similarity measures. In: International Conference on Case-Based Reasoning, pp. 17–31. Springer, Berlin, Heidelberg (2011)
46. Liu, B., Sotomayor, B., Madduri, R., Chard, K., Foster, I.: Deploying bioinformatics workflows on clouds with galaxy and Globus provision. In: 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, pp. 1087–1095 (2012)
47. Buyya, R., Yeo, C.S., Venugopal, S.: Market-oriented cloud computing: vision, hype, and reality for delivering it services as computing utilities. In: 10th IEEE International Conference on High Performance Computing and Communications, pp. 5–13 (2008)
48. Vahi, K., Harvey, I., Samak, T., Gunter, D., Evans, K., Rogers, D., Taylor, I., Goode, M., Silva, F., Al-Shkarchi, E., Mehta, G.: A general approach to real-time workflow monitoring. In: High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion, pp. 108–118 (2012)
49. Yuan, D., Cui, L., Liu, X.: Cloud data management for scientific workflows: research issues, methodologies, and state-of-the-art. In: 2014 10th International Conference on Semantics, Knowledge and Grids (SKG), pp. 21–28 (2014)
50. Oinn, T., Li, P., Kell, D.B., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turi, D., Zhao, J.: Taverna/my Grid: aligning a workflow system with the life sciences community. In: Workflows for e-Science, pp. 300–319. Springer, London (2007)
51. Kozlovszky, M., Karóczkai, K., Márton, I., Kacsuk, P., Gottdank, T.: DCI bridge: executing Ws-pgrade workflows in distributed computing infrastructures. In: Science Gateways for Distributed Computing Infrastructures, pp. 51–67. Springer, Cham (2014)
52. Litzkow, M.J., Livny, M., Mutka, M.W.: Condor—a hunter of idle workstations. In: Distributed Computing Systems, 8th International Conference on Semantics, Knowledge and Grids (SKG), pp. 104–111 (1988)
53. Brandic, I., Dustdar, S.: Grid vs Cloud—a technology comparison. IT-Inf. Technol. Methoden und innovative Anwendungen der Informatik und Informationstechnik **53**(4), 173–179 (2011)
54. Ramakrishnan, A., Singh, G., Zhao, H., Deelman, E., Sakellariou, R., Vahi, K., Blackburn, K., Meyers, D., Samidi, M.: Scheduling data-intensive workflows onto storage-constrained distributed resources. In: Seventh IEEE International Symposium on Cluster Computing and the Grid, 2007, pp. 401–409. IEEE (2007)
55. Keahey, K., Freeman, T.: Contextualization: providing one-click virtual clusters. In: IEEE Fourth International Conference on eScience, 2008, eScience'08, pp. 301–308. IEEE (2008)
56. Vöckler, J.S., Juve, G., Deelman, E., Rynge, M., Berriman, B.: Experiences using cloud computing for a scientific workflow application. In: Proceedings of the 2nd International Workshop on Scientific Cloud Computing, pp. 15–24. ACM (2011)

57. Talia, D.: Clouds for Scalable Big Data Analytics. IEEE Computer Society. http://scholar.google.co.in/scholar_url?url=http://xa.yimg.com/kq/groups/16253916/1476905727/name/06515548.pdf&hl=en&sa=X&scisig=AAGBfm12aY-Nbu37oZYRuEqeqsdslzKfBQ&nossl=1&oi=scholarr&ved=0CCYQgAMoADAAahUKEwi3k4Hymv7GAhUHUKYKHdToBCM. Accessed 16 March 2018
58. Lin, C., Lu, S., Fei, X., Chebotko, A., Pai, D., Lai, D., Fotouhi, F., Hua, J.: A reference architecture for scientific workflow management systems and the VIEW SOA solution. IEEE Trans. Serv. Comput. **2**(1), 79–92 (2009)
59. Zhao, Y., Li, Y., Lu, S., Raicu, I., Lin, C.: Devising a cloud scientific workflow platform for big data. In: 2014 IEEE World Congress on Services (SERVICES), pp. 393–401. IEEE (2014)
60. Juve, G., Deelman, E.: Scientific workflows in the cloud. In: Grids, Clouds and Virtualization, pp. 71–91. Springer, London (2011)
61. Bell, G., Hey, T., Szalay, A.: Beyond the data deluge. Science **323**(5919), 1297–1298 (2009)
62. Das, H., Jena, A.K., Badajena, J.C., Pradhan, C., Barik, R.K.: Resource allocation in cooperative cloud environments. In: Progress in Computing, Analytics and Networking, pp. 825–841. Springer, Singapore (2018)
63. Malawski, M., Juve, G., Deelman, E., Nabrzyski, J.: Algorithms for cost-and deadline-constrained provisioning for scientific workflow ensembles in IaaS clouds. Future Gener. Comput. Syst. **48**, 1–18 (2015)
64. Kwok, Y.K., Ahmad, I.: Dynamic critical-path scheduling: an effective technique for allocating task graphs to multiprocessors. IEEE Trans. Parallel Distrib. Syst. **7**(5), 506–521 (1996)
65. Juve, G., Deelman, E.: Wrangler: virtual cluster provisioning for the cloud. In: Proceedings of the 20th International Symposium on High Performance Distributed Computing, pp. 277–278. ACM (2011)
66. Barolli, L., Chen, X., Xhafa, F.: Advances on cloud services and cloud computing. Concurr. Comput. Pract. Exp. **27**(8), 1985–1987 (2015)
67. Ali, S.A., Alam, M.: A relative study of task scheduling algorithms in cloud computing environment. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 105–111. IEEE (2016)
68. Rodriguez, M.A., Buyya, R.: Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds. IEEE Trans. Cloud Comput. **2**(2), 222–235 (2014)
69. Bux, M., Brandt, J., Witt, C., Dowling, J., Leser, U.: Hi-WAY: execution of scientific workflows on Hadoop YARN. In: Proceedings of the 20th International Conference on Extending Database Technology (EDBT), Venice, Italy (2017)
70. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)
71. Ritchie, G., Levine, J.: A fast, effective local search for scheduling independent jobs in heterogeneous computing environments (2003)
72. Falzon, G., Li, M.: Enhancing genetic algorithms for dependent job scheduling in grid computing environments. J. Supercomput. **62**(1), 290–314 (2012)
73. Grosan, C., Abraham, A., Helvik, B.: Multiobjective evolutionary algorithms for scheduling jobs on computational grids. In: International Conference on Applied Computing, pp. 459–463 (2007)
74. Das, H., Jena, A.K., Nayak, J., Naik, B., Behera, H.S.: A novel PSO based back propagation learning-MLP (PSO-BP-MLP) for classification. In: Computational Intelligence in Data Mining, vol. 2, pp. 461–471. Springer, New Delhi (2015)
75. Gamal, A., Hamam, Y.: Task allocation for maximizing reliability of distributed systems: a simulated annealing approach. J. Parallel Distrib. Comput. **66**(10), 1259–1266 (2006)
76. Filgueira, R., Ferreira da Silva, R., Krause, A., Deelman, E., Atkinson, M.: Asterism: Pegasus and dispel4py hybrid workflows for data-intensive science. In: 2016 Seventh International Workshop on Data-Intensive Computing in the Clouds (DataCloud), pp. 1–8. IEEE (2016)
77. Esteves, S., Veiga, L.: WaaS: workflow-as-a-service for the cloud with scheduling of continuous and data-intensive workflows. Comput. J. **59**(3), 371–383 (2016)

78. Gerlach, W., Tang, W., Keegan, K., Harrison, T., Wilke, A., Bischof, J., Dsouza, M., et al.: Skyport-container-based execution environment management for multi-cloud scientific workflows. In: 2014 5th International Workshop on Data-Intensive Computing in the Clouds (DataCloud), pp. 25–32. IEEE (2014)
79. Wang, J., Korambath, P., Altintas, I., Davis, J., Crawl, D.: Workflow as a service in the cloud: architecture and scheduling algorithms. Procedia Comput. Sci. **29**, 546–556 (2014)
80. Rodriguez, M.A., Buyya, R.: Scheduling dynamic workloads in multi-tenant scientific workflow as a service platforms. Future Gener. Comput. Syst. **79**, 739–750 (2018)
81. Kaur, P., Mehta, S.: Resource provisioning and work flow scheduling in clouds using augmented shuffled frog leaping algorithm. J. Parallel Distrib. Comput. **101**, 41–50 (2017)
82. Chu, S., Tsai, P., Pan, J.: Cat swarm optimization. In: Pacific Rim International Conference on Artificial Intelligence, pp. 854–858. Springer, Berlin, Heidelberg (2006)
83. Chu, S., Tsai, P.: Computational intelligence based on the behavior of cats. Int. J. Innov. Comput. Inf. Control **3**(1), 163–173 (2007)
84. Sharafi, Y., Khanesar, M.A., Teshnehlab, M.: Discrete binary cat swarm optimization algorithm. In: 2013 3rd International Conference on Computer, Control & Communication (IC4), pp. 1–6. IEEE (2013)
85. Tsai, P.W., Pan, J.S., Chen, S.M., Liao, B.Y., Hao, S.P.: Parallel cat swarm optimization. In: 2008 International Conference on Machine Learning and Cybernetics, vol. 6, pp. 3328–3333. IEEE (2008)
86. Verma, A., Kaushal, S.: Cost-time efficient scheduling plan for executing workflows in the cloud. J. Grid Comput. **13**(4), 495–506 (2015)
87. Ahmad, S.G., Liew, C.S., Munir, E.U., Ang, T.F., Khan, S.U.: A hybrid genetic algorithm for optimization of scheduling workflow applications in heterogeneous computing systems. J. Parallel Distrib. Comput. **87**, 80–90 (2016)
88. Tao, F., Feng, Y., Zhang, L., Liao, T.W.: CLPS-GA: A case library and Pareto solution-based hybrid genetic algorithm for energy-aware cloud service scheduling. Appl. Soft Comput. **19**, 264–279 (2014)
89. Kar, I., Parida, R.R., Das, H.: Energy aware scheduling using genetic algorithm in cloud data centers. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3545–3550. IEEE (2016)
90. Kar, I., Das, H.: Energy aware task scheduling using genetic algorithm in cloud datacentres. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 106–111 (2016)
91. Sahoo, A.K., Das, H.: Energy efficient scheduling using DVFS technique in cloud datacenters. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 59–66 (2016)
92. Verma, A., Kaushal, S.: A hybrid multi-objective particle swarm optimization for scientific workflow scheduling. Parallel Comput. **62**, 1–19 (2017)
93. Ezzatti, P., Pedemonte, M., Martín, A.: An efficient implementation of the Min-Min heuristic. Comput. Oper. Res. **40**(11), 2670–2676 (2013)
94. He, X., Sun, X., Laszewski, G.V.: QoS guided min-min heuristic for grid task scheduling. J. Comput. Sci. Technol. **18**(4), 442–451 (2003)
95. Singh, M., Suri, P.K.: QPS Max-Min<> Min-Min: a QoS based predictive Max-Min, Min-Min switcher algorithm for job scheduling in a grid. Inf. Technol. J. **7**(8), 1176–1181 (2008)
96. Tabak, E.K., Cambazoglu, B.B., Aykanat, C.: Improving the performance of independent task assignment heuristics minmin, maxmin and sufferage. IEEE Trans. Parallel Distrib. Syst. **25**(5), 1244–1256 (2014)
97. Casanova, H., Legrand, A., Zagorodnov, D., Berman, F.: Heuristics for scheduling parameter sweep applications in grid environments. In: 9th Heterogeneous Computing Workshop, 2000 (HCW 2000) Proceedings, pp. 349–363. IEEE (2000)
98. Chen, W., Zhang, J.: A set-based discrete PSO for cloud workflow scheduling with user-defined QoS constraints. In: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 773–778. IEEE (2012)
99. Jianfang, C., Junjie, C., Qingshan, Z.: An optimized scheduling algorithm on a cloud workflow using a discrete particle swarm. Cybern. Inf. Technol. **14**(1), 25–39 (2014)

100. Bahrami, M., Bozorg-Haddad, O., Chu, X.: Cat swarm optimization (CSO) algorithm. In: Advanced Optimization by Nature-Inspired Algorithms, pp. 9–18. Springer, Singapore (2018)
101. Eusuff, M., Lansey, K., Pasha, F.: Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. Eng. Optim. **38**(2), 129–154 (2006)
102. Liu, J.: Multisite management of scientific workflows in the cloud. Distributed, parallel, and cluster computing. Ph.D. dissertation, Universite de Montpellier (2016)
103. Liu, J., Pacitti, E., Valduriez, P., Oliveira, D., Mattoso, M.: Scientific workflow execution with multiple objectives in multisite clouds. In: BDA: Bases de Données Avancées (2016)
104. Pineda-Morales, L., Liu, J., Costan, A., Pacitti, E., Antoniu, G., Valduriez, P., Mattoso, M.: Managing hot metadata for scientific workflows on multisite clouds. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 390–397. IEEE (2016)
105. Tudoran, R., Costan, A., Antoniu, G.: Overflow: multi-site aware big data management for scientific workflows on clouds. IEEE Trans. Cloud Comput. **4**(1), 76–89 (2016)
106. Ahmad, M.K.H.: Scientific workflow execution reproducibility using cloud-aware provenance. Ph.D. dissertation, University of the West of England (UWE) (2016)
107. Jrad, F., Tao, J., Streit, A.: A broker-based framework for multi-cloud workflows. In: Proceedings of the 2013 International Workshop on Multi-cloud Applications and Federated Clouds, pp. 61–68. ACM (2013)
108. Kozlowszky, M., Karóczkai, K., Marton, A., Balasko, A., Marosi, A., Kacsuk, P.: Enabling generic distributed computing infrastructure compatibility for workflow management systems. Comput. Sci. **13**(3), 61 (2012)
109. Varghese, B., Wang, N., Barbhuiya, S., Kilpatrick, P., Nikolopoulos, D.S.: Challenges and opportunities in edge computing (2016). arXiv:1609.01967
110. Meurisch, C., Seeliger, A., Schmidt, B., Schweizer, I., Kaup, F., Mühlhäuser, M.: Upgrading wireless home routers for enabling large-scale deployment of cloudlets. In: International Conference on Mobile Computing, Applications, and Services, pp. 12–29. Springer, Cham (2015)
111. Chen, W., Deelman, E.: Integration of workflow partitioning and resource provisioning. In: Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID 2012), pp. 764–768 (2012)
112. Tang, W., Jenkins, J., Meyer, F., Ross, R., Kettimuthu, R., Winkler, L., Yang, X., Lehman, T., Desai, N.: Data-aware resource scheduling for multicloud workflows: a fine-grained simulation approach. In: 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom), pp. 887–892 (2014)
113. Yin, D., Kosar, T.: A data-aware workflow scheduling algorithm for heterogeneous distributed systems. In: International Conference on High Performance Computing and Simulation (HPCS), 2011, pp. 114–120. IEEE (2011)
114. Ghafarian, T., Javadi, B.: Cloud-aware data intensive workflow scheduling on volunteer computing systems. Future Gener. Comput. Syst. **51**, 87–97 (2015)

Trust Model Based Scheduling of Stochastic Workflows in Cloud and Fog Computing

J. Angela Jennifa Sujana, M. Geethanjali, R. Venitta Raj and T. Revathi

Abstract The Cloud computing is a lucrative, challenging and beneficial technology in the IT world. The emergence of Internet of Things (IoT) has made cloud computing to be combined with fog computing, in order to avoid latency. These technologies have daring challenges. This chapter focuses on two major challenges, namely security and scheduling of user requests. The security is met by our proposed trust model which includes both direct trust and reputation relationship. This chapter initially, focuses on assuring trusted environment in the cloud. Then a trust model for cloud cum fog environment is proposed. The new trust model would ensure that the user's requests are serviced with enough security guaranteed level based on the Service Level Agreement (SLA) negotiated with the cloud provider. Based on the trust value computed, the user's requests are scheduled to the appropriate resource by applying the Trust based Stochastic Scheduling (TSS) algorithm. The trust based stochastic scheduling minimizes makespan of the schedule is achieved for a secured cloud environment

Keywords Trust model · Stochastic scheduling · Service level agreement Cloud computing

1 Introduction

Now-a-days, information technology based industries have started using cloud computing and it has a great impact in the way we use software and other computing resources. Cloud computing can be used to host any services, which the user wants to access through the internet. The cloud gives many benefits like pay as we use, anywhere access and rapid elasticity. Hence using cloud for executing any scientific workflows will be beneficial. Focusing on this we visualize two main challenges in

J. Angela Jennifa Sujana (✉) · M. Geethanjali · R. Venitta Raj · T. Revathi
Department of Information Technology, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India
e-mail: ang_jenefa@mepcoeng.ac.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*, Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_2

hosting workflows in cloud. They are scheduling the tasks in a workflow to suitable VMs in cloud and security ensured.

Many scientific and real time applications have evolved with the invent of IoT. These scientific workflows have now extended its usage to fog computing also [1, 2]. To host these applications in Cloud or Cloud cum Fog environment, then the security provided by the cloud provider or the trust on fog nodes is a major concern to the users [3, 4]. This issue arises due to the multitenancy i.e., a variety of users who are availing the cloud services will be provided with the resources from the same datacenter at the cloud provider side. The fog nodes also have wide access to all the edge devices, such as sensors, smart phones, vehicles etc., Hence, there is a possibility of security breach or any type of vulnerability may happen in the cloud and fog environment. In this context, security plays a vital role. Examples of security sensitive applications are storing and processing sensitive data, electronic transaction etc.

Thus in this chapter we focus on the problem of effectively hosting workflows in cloud with a good schedule and in a good secured environment. We also present a trust model for fog computing. Specifically, this work focuses on Infrastructure as a Service (IaaS) type of service. In IaaS the user will be allocated Virtual Machines (VM), which is created on top of the physical host servers running in the datacenter of the cloud provider.

The scheduling policy adapted by a cloud provider plays a vital role, since cloud computing is based on pay as you go and on demand provisioning. Owing to its significance many research works were carried out on scheduling. Scheduling is the method by which any workflow can be mapped with the appropriate resource. The best schedule has to provide an optimized resource allocation, such that the makespan of the workflow execution is minimized [5–7]. For the benefit of both cloud provider and cloud user, it becomes necessary to give due consideration to resource allocation in the cloud environment. To provide a break through, one has to adapt any heuristic method that satisfies the need of both the user and the provider. Here our objective is to maximize the trust and minimize the makespan of the schedule.

To impart security to the cloud users we focus on implementing a trust model. Normally, research works concerning security have concentrated on authentication, integrity, confidentiality [8–10] and access control mechanism [11, 12]. Nevertheless, the cloud provider must ensure the cloud user that their infrastructure (in case of IaaS) is safe and secure. Hence we focus on designing a trust model which will assure that the allocated resource (i.e., the virtual machine) is a trusted one. In other words, the trust model will ensure that the resources allocated to the user are maintained in trusted zones, which will cater to distinct security guaranteed levels. The Security Guaranteed Level (SGL) is the level which specifies the percentage of security assured for the VM. This is described in Sect. 3.2. In general, the cloud user will sign a Service Level Agreement (SLA) with the cloud provider regarding the quality of service expected by the user. Based on the differentiated SLA the service will be provided and the users will be charged accordingly. The quality of service requirements will differ from customer to customer depending upon their applications. Depending upon the SLA, the cost of the cloud service would vary. The security requirement

(i.e., security demand) is also specified as part of the SLA. The proposed trust model makes use of this security demand and ensure that the user is given the best secured resources.

This chapter aims at the integration of security and scheduling which is much demanding and more challenging. Few researchers have done considerable work in this aspect [4, 8–10, 13–15]. But considering a workflow with stochastic behaviour and dynamic environment is lacking. Thus this chapter proposes a stochastic scheduling approach, which considers the user demand from the SLA signed between the user and the provider. Less works is focused on integrating the trust model with workflow scheduling in cloud. In our present work we integrate both security and scheduling for every cloud users' request. Security can be achieved by our trust model and our scheduling algorithm aims at reducing the schedule length i.e., makespan. We focus on scheduling tasks which has stochastic process time. The novelty is the stochastic top levels for schedule planning on the dynamic systems along with the trust value. Though the presence of an additional trust model adds additional overhead that accounts for additional time complexity, the overall performance is optimized even in the presence of trust model.

The present work aims at using random variables with uniform distribution for representing the stochastic behaviour of the tasks. The prioritization is done based on the stochastic top level. The workflow scheduling is integrated with trust model.

2 Related Work

Workflows are normally modelled as a Directed Acyclic Graph (DAG) to represent the various tasks in the workflow and the precedence constraint relation that exists between them. Scheduling the tasks in a DAG is a NP—hard problem [16]. Most of the research work on scheduling concentrates on two types of system, namely homogeneous or heterogeneous. Since cloud computing is of heterogeneous type, our scheduling algorithm has to focus on heterogeneous resources. In addition to that workflow scheduling algorithms can be classified into two main groups, as heuristic based and metaheuristic based algorithms [17, 18]. The former can be further classified into a variety of categories such as list-scheduling algorithms, clustering algorithms, duplication-based algorithm [19]. This work focuses on heuristic based scheduling and more specifically list scheduling. List scheduling has been considered in most of the previous works [8, 9, 15, 19–27]. List scheduling, which works based on the principle of priority has been applied to both homogeneous and heterogeneous systems. The Dynamic Level Scheduling (DLS) developed by Sih and Lee [22] computes the availability of each resource and allow a task to be scheduled to a currently busy resource, which was not done in the previous work by El-Rewini and Lewis [28] i.e., the Mapping Heuristic algorithm. Topcuoglu et al. proposed the Heterogeneous Earliest Finish Time (HEFT) algorithm [19] that is highly competitive and capable of generating a good schedule compared to other scheduling algorithms with a lower time complexity. The HEFT algorithm has two major phases: a task

prioritizing phase and a resource selection phase. The main impact in the improvement of HEFT algorithm is due to the use of Earliest Finish Time (EFT) rather than Earliest Start Time (EST). Ilavarasan et al. developed High-Performance Task Scheduling (HPS) [29]. This algorithm uses three phases, namely, level sorting, task prioritization and processor selection. Priority is computed and assigned to each task using the attributes Down Link Cost (DLC), Up Link Cost (ULC), and Link Cost (LC) of the task. The processor that gives the minimum EFT for a task is selected to execute that task.

We focus on scheduling with the stochastic tasks. Bertsekas and Castanon [30] proposed heuristics based rollout algorithms which are derived from stochastic dynamic programming algorithm. Shmoys and Sozio [31] focused on 2-stage stochastic problem. Though, the 2-stage approximation method can give solution to single processor problem, it cannot give solution when working in multi-processor systems. Besides, to calculate the probable makespan, Gourgand et al. [32] developed a recursive method based on a Markov chain. Later on, for stochastic scheduling problem, Megow et al. formulated online scheduling algorithms and with guaranteed performance measures [33].

Precedence constraints among tasks play an important role in several real-world workflow scheduling problems. So, it is essential to consider the precedence constraints between tasks in stochastic scheduling. Skutella and Uetz [34] modelled the precedence constrained stochastic tasks as a Directed Acyclic Graph (DAG). They proposed methods with first constant-factor approximation. To reduce the estimated value of the total weighted completion time, these methods are derived by combining linear programming relaxation and delayed list scheduling algorithm. SHEFT (Stochastic HEFT) method [35] is the modified version of HEFT algorithm which includes the stochastic behaviour of tasks in all the three stages and the performance was found to be superior to the existing scheduling algorithm. The Dynamic Level Scheduling (DLS) algorithm [22, 36] utilizes a parameter called dynamic level (DL), which is the variance between the stable level of a task and its earliest execution start time. The task-processor pair which offers the highest value of DL is chosen for execution in each step of scheduling process. The Stochastic Dynamic Level Scheduling algorithm (SDLS) [23] is based on DLS method which incorporates the response from stochastic environment and the task-processor pair is found with respect to the maximum value of stochastic dynamic level of the tasks. In all these works, researchers have tried up with different factors for prioritization. In the present work we focus on using random variables with uniform distribution for representing the stochastic behaviour of the tasks. The prioritization is done based on the stochastic top level. The workflow scheduling is integrated with trust model.

Tao and Xiao have proposed to combine security model along with scheduling for real time applications [8]. They have investigated the problem of scheduling a set of independent real-time tasks with various security requirements. A security overhead model that can be used to reasonably measure security overheads incurred by the security-critical tasks is devised. Next, they propose security-aware real-time heuristic strategy for clusters (SAREC), which incorporate the earliest deadline first (EDF) scheduling policy. Similarly, Tang et al. [9] have proposed a security-driven

scheduling architecture that can dynamically measure the trust level of each node in the system by using differential equations. They introduce task priority rank to estimate security overhead of such security-critical tasks. In another similar work, Wei Wang et al. have considered Bayesian Trust model for providing security to cloud computing along with scheduling [14]. We design a trust model which will adopt to the cloud environment and measure the honesty of the user and the provider. In literature many trust models have been proposed, such as Eigen Trust model [37], Bayesian Model [14, 38], Peer Trust [39], Power Trust [40] etc. But these trust models are derived for a specific environment. In the current work, we design a trust model for cloud environment in particular based on probability distribution function. This trust model is then merged with scheduling.

The main contributions of this work are

- Trust model for the cloud users to satisfy the security demand of the users.
- Optimized stochastic workflows scheduling which minimizes the makespan with increased speedup.
- Inclusion of variance with mean in the stochastic top level, which leads to better results.
- Design of new trust model for cloud cum fog environment.

3 Proposed Trust Model

3.1 System Architecture

The system architecture of the proposed Trust based Stochastic Scheduling is depicted in Fig. 1. The cloud provider has a pool of resources which is normally provisioned to the users as virtual machines. Generally, the virtual machines are maintained in different *availability zones*, to service the user to accomplish their corresponding service quality requirements [41, 42]. Normally, the formation of the availability zones is based on different characteristics of the underlying hardware.

In the present work, we consider that the provider maintains the virtual machines in different zones based on the hardware grades and security services. When the user comes with a request for resources from the provider for executing their workflow, they have to negotiate with the provider for the SLA. Based on this SLA, which the user selected, the *Security demand* will be decided. Then the trust model will calculate the trust value for the user's request. Both the trust model and scheduling model will get the needed information from the Cloud Information Service as depicted in Fig. 1. Then the stochastic model is applied to the workflow, which calculates the stochastic top level. Then the trust value and the stochastic bottom level are combined to get the trust based stochastic top level. This is further used in the scheduling process for the best VM selection for executing the workflow.

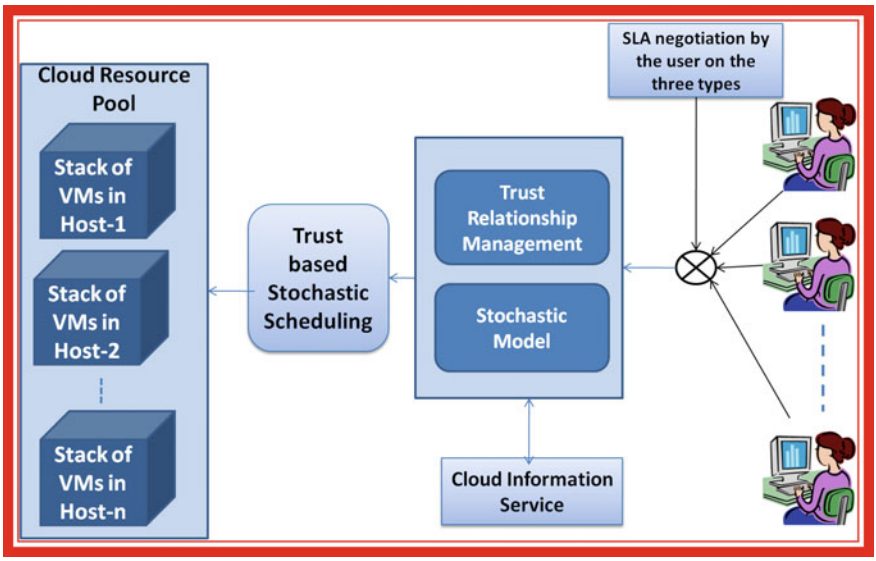


Fig. 1 System architecture

Table 1 List of VM types used in the experiment

Type	vCPU	Memory (GB)	Storage (GB)	MIPS	Cost
t2.small	1	2	1 × 160	1000	$0.06
t2.medium	2	4	2 × 240	2000	$0.12
m3.medium	1	3.75	1 × 400	2000	$0.10
m3.large	2	7.5	2 × 320	4000	$0.18
m3.xlarge	4	15	4 × 320	8000	$0.24

In cloud environment, each cloud provider offers several VM configurations, often referred to as instance types. An instance type could be defined in terms of hardware metrics such as main memory, CPU (number of cores and clock frequency), available storage space, and price per hour. In the present work, we have used VM configuration as given in Table 1, which are in par with Amazon EC2 Instances. Cloud providers provide virtual machines as instances and these instances are available in varied types such as small, medium, large, xlarge and xxlarge. These VM are allocated on demand by the user.

We define three types of Service Level Agreements (SLAs). They are Gold, Silver and Bronze based on different levels of services and security zones. Gold SLA is the highest sophisticated one and Bronze the least sophisticated SLA. Based on the category of the SLA selected by the user, the privileges and the services will differ. The parameters we have considered in the SLA are security level, performance, availability, backup and service initialization time as shown in Fig. 2.

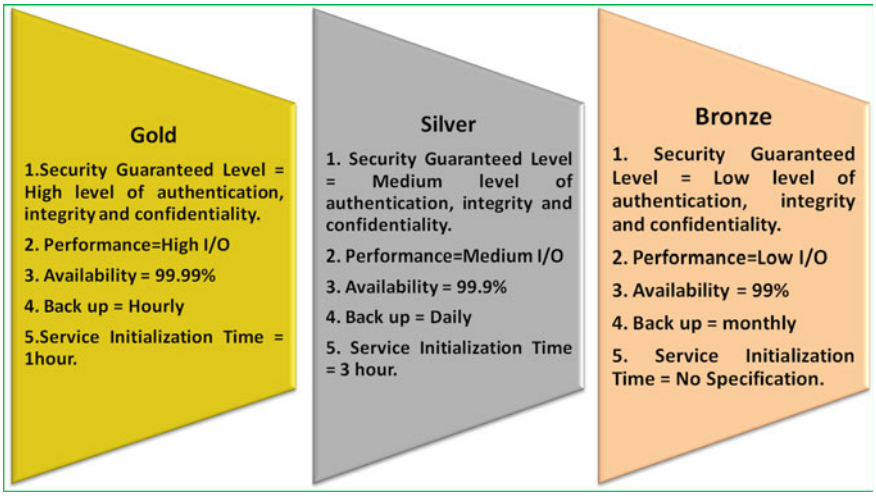


Fig. 2 Sample service level agreement provided by a cloud provider

Each type has a fixed limit for each parameters and the SLA negotiation will be done with a flexibility of values within the limits. The security parameter focuses on the three main attributes namely, authentication, integrity and confidentiality. As part of the SLA the user has to specify their security demand. Normally this is got from the user as a percentage value or in a predefined range associated with a category. Hence, the user is free to select a security zone from Gold, Silver and Bronze. The user can specify their Security Demand (SD) as percentage value for authentication, integrity and confidentiality. It is represented as sd_a, sd_c *and* sd_i respectively. If the user gives only one single value, then the same value is considered for all the three attributes. Otherwise the user is free to give their preference for security in terms of authentication, integrity and confidentiality separately. The Security demand is defined as an array list for each task in the workflow and it is calculated as given in Eq. (1).

3.2 Service Level Agreement (SLA)

$$SD_{u_i} = Avg(sd_a + sd_c + sd_i) \tag{1}$$

The idea is to define a Security Guaranteed Level (SGL) for each virtual machine running in the host servers of the datacenter. This SGL is framed from three main parameters authentication, integrity and confidentiality. The SGL is thus defined in terms of three attributes as sgl_a, sgl_c and sgl_i representing the authentication, integrity and confidentiality respectively and it is defined in Eq. (2). The values for these three attributes will be in the interval as given in Table 2 based on uniform

Table 2 Interval limit of security guaranteed level for different SLAs

Zone type	SGL value
Gold	[0.9, 1]
Silver	[0.5, 0.89]
Bronze	[0.1, 0.49]

Table 3 Hash functions for integrity [8]

Hash function	Security level (SL)	Performance (KB/ms)
MD4	0.18	23.90
MD5	0.26	17.09
RIPEMD	0.36	12.00
RIPEMD-128	0.45	9.73
SHA-1	0.63	6.88
RIPEMD-160	0.77	5.69
Tiger	1.00	4.36

distributions. To achieve confidentiality one has to use cryptographic algorithms and for integrity good hash functions has to be used. Similarly, for authentication robust authentication methods are to be used. The required cryptographic algorithms for confidentiality, hash functions for integrity and authentication methods are adapted from the work done by Tao and Xiao [8, 15].

The SGL for the three zones are termed as high, medium and low and the values will be in the fixed interval as given in Table 2 based on uniform distribution. The security guaranteed levels are assured by implementing the corresponding method as shown in Table 3. For example, Table 3 shows that the usage of SHA-1 will assure a security guaranteed level of 0.63 for integrity. In this way we will be able to find a suitable security mechanism for each SGL value. The security levels for cryptographic algorithms and authentication methods are also given as tabulated value in the work done by Tao and Xiao [8, 15]. The security guaranteed level of a VM is computed by the average of the security guaranteed level for authentication, confidentiality and integrity

$$SGL_{VM_j} = Avg\{sgl_a, sgl_c, sgl_i\} \tag{2}$$

4 Trust Model for Cloud

The trust model consists of two main factors, namely direct trust and reputation trust. The Direct trust is used to represent the direct relationship between the cloud provider and the cloud consumer i.e. cloud user. Direct trust accounts to the trustworthiness of the resource provided by the cloud provider. It mainly ensures that sufficient security mechanisms are adopted at the physical host level in the datacenter and also

identifies any unethical event executed by the user. This is necessary because of the multi-tenancy property of the cloud environment. Any user using a VM on a physical host may try to hack or snoop into another user's VM. Hence, we focus on providing physical security, which will be accounted in the trust model. To ensure physical security, security methods as described in the previous section are used. To know whether the user is a trusted user or not, any malicious event from the VM of the user and any security concerning events like VM Theft, VM Escape, Hyper Jacking, Data leakage, Denial of Service (DoS) attack are recorded. Thus the direct trust is the component which is calculated by the users' direct experience in the cloud.

The second factor, reputation trust is designed by the recommendations by other existing users in the cloud. The reputation trust is calculated by the feedback vote entrusted by the user on using any service. Here, the users can vote the services provided by the cloud provider. A reputation system is also called as collaborative sanctioning systems. Reputation systems are already being used in successful commercial online shopping applications. This will be constituted for the reputation of the cloud user. The reputation is also influenced by the negative factor like false recommendations. This may be done in order to defame any provider or user. The sybil attack has been studied and the defense was proposed in the work by Yu et al. [7].

We represent the direct relationship and the recommendation relationship using a discrete distribution function. Let $\mathrm{U} = \{u_1, u_2, \ldots, u_n\}$ represents the set of n users who consume the cloud service for executing their workflows and $\mathrm{VM} = \{vm_1, vm_2, \ldots, vm_p\}$ represents the set of available resources i.e. the stack of VMs in different physical hosts, with the cloud provider. The resource represents the virtual machine provided by the cloud provider. When the user makes a request for executing the workflow, suitable virtual machines from the pool of resources are to be assigned. The best virtual machines have to be selected for executing the workflow.

In the present work, we have added the trust model which will evaluate the user's trust value by considering the direct trust relationship and the reputation of user, which was obtained from the existing users. However, we give less weightage to the reputation of the user, because the existing users may downplay in order to slur the other one. The individual user behaviour also may vary from time to time. There is a possibility of deterioration in their truthfulness. To address the above said factors, we have included the decay factor in the trust relationship.

4.1 Trust Relationship Management for Cloud Environment

The trust relationship management involves the trust value calculation for the direct interaction between the user and the cloud provider and reputation of the user. The trust value is represented by $TR_{i,j}^{SLA_k}$ and the computation is given in Eq. (3). It denotes the trust value for the ith user on jth VM with kth SLA zone. The direct interaction (i.e. direct relationship) is represented by the function $D(u_i, vm_j, SLA_k)$, where u_i represents the ith user, vm_j represents the jth VM and SLA_k represents the kth SLA

Table 4 Trust point value for each zone

SLA zone	Gold	Silver	Bronze
tp value	0.02	0.05	0.08

zone opted by the user from the set of SLAs provided by the cloud provider. The indirect trust or the reputation trust is represented by the function $F(u_i, vm_j)$. The terms w_D and $w_{\mathrm{Re}p}$ represent the weight factor of direct relationship and reputation. The weight for direct relationship is given more weightage than reputation in order to avoid the sybil type of attacks. Normally, having a higher weight factor for w_D is advisable to avoid the reputation due to faulty recommendation.

$$TR_{i,j}^{SLA_k} = w_D \cdot D(u_i, vm_j, SLA_k) + w_{\mathrm{Re}p} \cdot F(u_i, vm_j)$$
$$\text{where} \quad w_D + w_{\mathrm{Re}p} = 1, \quad w_D > w_{\mathrm{Re}p} \tag{3}$$

The trust relationship is handled by the event recording mechanism. The event recording mechanism traces all the interactions between the user and the provider and also all the external events of the VM used by the user. The events and interactions are categorized into two. They are trusted events and untrusted events. Events like VM Theft, VM Escape, Hyper Jacking, Data leakage, Denial of Service (DoS) attack are recorded as untrusted events. Any normal executions are considered as trusted events. The event recording mechanism will track the events from the logs and award the trust points to the user. The event recorder will count the number of trusted events as ***tc***. This ***tc*** represents the truthful and successful interaction between the user u_i and the provider on using the VM vm_j. The number of untrusted event count is termed as ***uc***. This ***uc*** represents the untruthful or misbehaved interactions between the user u_i and the provider on using the VM vm_j. The total number of interactions is recorded by the count variable ***count***. This count variable must be equal to the sum of ***tc*** and ***uc***. Thus ***count*** = ***tc*** + ***uc***.

The trust points for each interactions is termed as ***tp.*** The trust points value varies for each zone. The trust point ***tp*** value for each zone is given in Table 4. We introduce the H factor which will represent the untrustworthiness of user. When $H = \phi$, it represents that the user is trustworthy and can have a positive credit to his trust value in direct relationship. The term ***tp*** represents the bonus increment value for the trustworthy behaviour of user. The Trust value will be slowly incremented for each trustworthy transaction and the *tp* values will be decided based on the SLA Zone values as given in Table 4. Similarly for each untrustworthy transaction the trust value will be divided by a factor of 2^{uc}. For example, the penalty for first untrustworthy transaction is ½ and for the second is ¼ and so on. This keeps on decrementing the trust value. We design the direct relatioship $D(u_i, vm_j, SLA_k)$ as given in Eq. (4).

$$D(u_i, vm_j, SLA_k) = tp_{cur} = \begin{cases} tp.e^{\lambda(SGL_{vm_j} - SD_{u_i})}, count = 0 \\ (tp_{cur} + tp).e^{\lambda(SGL_{vm_j} - SD_{u_i})}, count > 0 \,\&\, H = \phi \\ \left(\frac{tp_{cur}}{2^{uc}}\right) tp.e^{\lambda(SGL_{vm_j} - SD_{u_i})}, count > 0 \,\&\, H \neq \phi \end{cases} \quad (4)$$

The direct relationship between user and provider is also influenced by environmental factors. In a longer run, the impact of environmental factor is not negligible and it is likely to increase exponentially. Hence we introduce a constant factor λ which will be fixed by the cloud provider based on the feedback from the event recording mechanism.

The reputation of the jth VM by ith user is denoted by $F\left(u_i, vm_j\right)$ as given in Eq. (5). The users can give a feedback on the VM used by them. This feedback is represented as the vote for each interaction. So when a user is about to choose a VM, this reputation factor will mark the trustworthiness of that VM for other users.

$$F(u_i, vm_j) = \sqrt{\frac{1}{count - 1} \sum_{l=1}^{count} (v_l - \bar{v}_l)^2 \cdot (1 - e^{-x})} \quad (5)$$

where v_l represents the rating given by the other existing users. The rating i.e. the vote value for the user is fixed in the 5-level scale. Level-1 vote credits for a low level and level-5 for the maximum. $\bar{v}_i$ denote the mean value of votes. The ***count*** value is used here since the user will vote for the service they got from the VM. We assume that after each interaction the user will submit his rating as the feedback. If any user hasn't submitted their rating, then it will be left as blank and also for new users. To address the faulty recommendation problem, the decay factor x is included whose value is set to 0.55 [39]. The decay factor is accounted in the recommendation as the exponential value $1 - e^{-x}$. The Cloud Information Service (CIS) is responsible for maintaining all the values.

4.2 Trust Relationship Management for Cloud Cum Fog Environment

The recent scientific and real time applications like smart chips, smart processor, smart city, smart health monitoring etc., have extended the use of cloud and the new technology fog computing has evolved. The current trend is hosting these scientific and real time applications in the cloud cum fog environment. The IoT devices or any smart devices that are used in these applications are called as the edge devices. Since the volume of the edge devices is more, the need for fog computing has aroused.

Also the usage of the cloud resources and services alone for workflow applications results in latency. This is due to the internet usage for the communication of all the data from and to the edge devices. The purpose of fog computing is to introduce processing nodes at the near affinity to the edge devices, so that some local processing, pre-processing can be done at the fog nodes. These scientific workflows have now extended its usage to fog computing also. Grouping these fog nodes based on regions is one approach in implementation. Here we use the region based grouping of fog nodes.

The Architecture of the trust model for cloud cum fog environment is given in Fig. 3. The architecture consists of three main layers, namely cloud layer, fog layer and edge layer. Here we host the Trust Manager (TM) in the cloud, which will be acting like the master in the trust model. This TM is responsible for computing the trust values. The trust relationship management architecture for cloud cum fog environment makes use of a Trust Manager (TM) in the cloud layer and Trust Supervisor (TS) in each region of the fog layer. The fog layer is divided into regions and they are represented as $R = \{R_1, R_2, \ldots, R_n\}$. Each region has many fog nodes and they are represented as $FN = \{FN_1, FN_2, \ldots, FN_n\}$. Each fog node is associated with a zone tag indicating the security zone it belongs to and the trust value. The trust manager plays the role of the manager by coordinating all the trust supervisors in each region. It computes the trust value of all the fog nodes by making use of the details in the local trust table and local trust event recorder. The role of the trust supervisor is to maintain the Local Trust Table (LTT) and Local Trust Event Recordings (LTER). The TS will act as the designated trusted node in the region. The fields of the LTT are IP address, Zone Tag, direct trust value, reputation trust value and trust Value. Here the IP address is used to identify any fog node uniquely. The IP address field holds the *ipaddress* of the fog (computing) nodes. The Zone Tag corresponds to the three zones Gold, Silver and Bronze as {G}, {S} and {B} respectively.

The LTER records the doubtful events like unauthorized access, side-channel attack, spoofing IP address, reporting wrong data etc. Based on the number events the trust values will be updated and it is as described in Sect. 4.1.

The trust value is represented by TV_{FN_i,R_j} and the computation is given in Eq. (6). Here FN_i denotes the ith fog node and R_j represents the jth region in the fog layer. The direct trust value (i.e. direct relationship) is represented by the function $D(FN_i, R_j, zt_k)$, where FN_i denotes the ith fog node and R_j represents the jth region and zt_k represents the kth zone opted by the user from the three zones. The indirect trust or the reputation trust is represented by the function $F(FN_i, R_j)$. The terms w_D and $w_{\mathrm{Re}p}$ represent the weight factor of direct relationship and reputation.

$$\begin{aligned} TV_{FN_i,R_j} &= w_D \cdot D(FN_i, R_j, zt_k) + w_{\mathrm{Re}p} \cdot F(FN_i, R_j) \\ \text{where} \quad & w_D + w_{\mathrm{Re}p} = 1, \quad w_D > w_{\mathrm{Re}p} \end{aligned} \tag{6}$$

To compute the direct trust value we have the LTER, which adopts the same method of calculating the ***tc***, ***uc and count*** as mention in Sect. 4.1. The trust point for each interaction is termed as ***tp.*** The values are calculated as per values in Table 4.

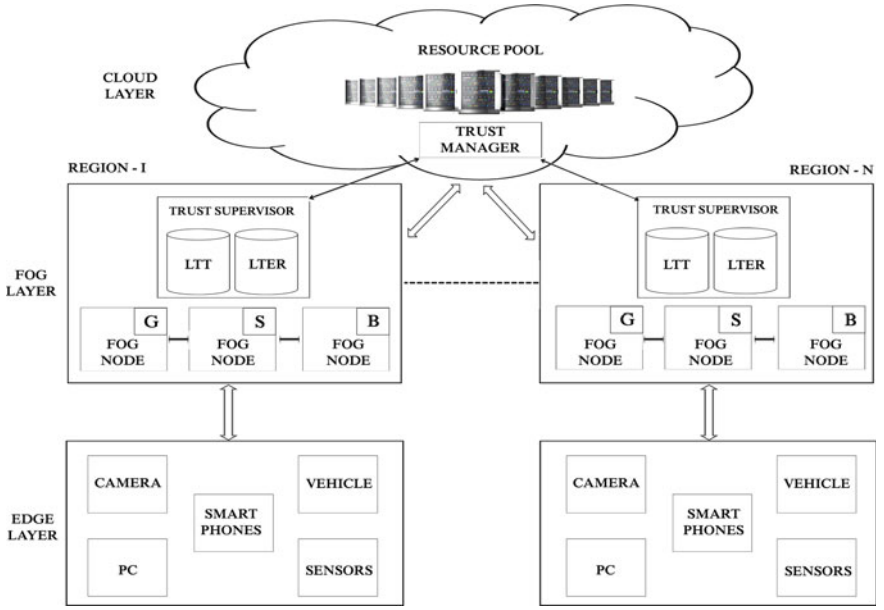


Fig. 3 The architecture of the trust model for cloud cum fog environment

To incorporate the concept of using three zones in each region we include the weight factor for the zones. They are represented as w_{zt_k}. The zt_k corresponds to {G}, {S} and {B}. Their weight values are as follows, $w_{zt_G} = 5$, $w_{zt_S} = 3$, $w_{zt_B} = 1$. The direct trust value and reputation trust value are computed by the Eqs. (7) and (8) respectively. The term ∂ in Eq. 8 is the decay factor for the indirect trust.

$$D(FN_i, R_j, zt_k) = \begin{cases} tp \cdot w_{zt_k}, count = 0 \\ (tp_{cur} + tp) \cdot w_{zt_k}, count > 0 \,\&\, H = \phi \\ \left(\frac{tp_{cur}}{2^{uc}}\right) tp \cdot w_{zt_k}, count > 0 \,\&\, H \neq \phi \end{cases} \tag{7}$$

$$F(FN_i, R_j) = \sqrt{\frac{1}{count - 1} \sum_{l=1}^{count} (v_l - \bar{v}_l)^2 \cdot \partial} \tag{8}$$

Now the edge devices will communicate with the fog nodes. The edge devices will first contact the trust supervisor of a particular region for knowing the trust values. Then based on the trust value and the zone needed it can select the fog node for its application processing. In this way the trustworthiness of the cloud cum fog environment is ensured.

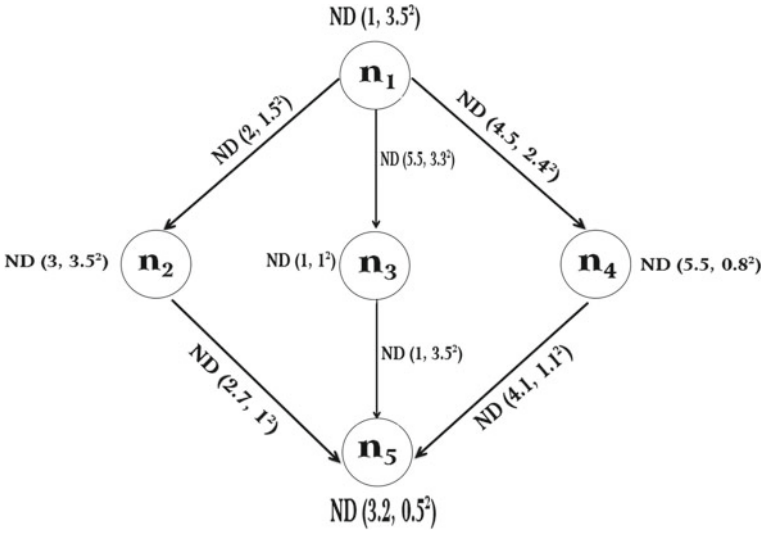


Fig. 4 A stochastic job application sample

4.3 Stochastic Model

In the present work, we consider the scheduling of precedence constrained stochastic tasks which are modelled as a Directed Acyclic Graph (DAG), $G = (N, E)$ [21, 34],

where $N = \{n_1, n_2, n_3, \ldots, n_i\}$ is the group of n precedence constrained stochastic tasks. These tasks are to be processed on any of the accessible virtual machines. $E \in N \times N$ is the set of edges, which represents the inter-task dependencies between tasks. For instance, edge $e_{i,j} \in E$ indicates precedence constraint which means the task n_i should complete its execution before the task n_j. The task with no predecessors is called entry task, n_{entry} and the task with no successors is called exit task, n_{exit}. The parent tasks of the task n_i is denoted as $par(n_i)$ and the child tasks of a task n_i is represented as $child(n_i)$.

Figure 4 shows a stochastic job application with 5 tasks. Each node represents the task execution time $r(n_i)$ and each edge represents the inter-task communication time $r(e_{i,j})$. Generally stochastic tasks will have dynamic changes in their task execution time and inter-task communication time. To make the problem simpler and representable, we consider the normal distribution of these values. In this sample, the normal distribution of task execution time and inter-task communication time is illustrated as $ND(\mu, \sigma^2)$, where μ and σ^2 are the mean and variance of normal distribution. It is assumed that task execution time $r(n_i)$ and inter-task communication time $r(e_{i,j})$ are random variables of normal probability distribution function since the model is stochastically independent. Let $VM = \{vm_1, vm_2, \ldots, vm_p\}$ be the p number of virtual machines in cloud environment. The computational capacity of pth virtual machine is represented as $c(vm_p)$.

4.4 Scheduling Attributes

In this scheduling of stochastic tasks, the following scheduling attributes are defined and used throughout this article. $STime(n_i, vm_p)$ represents the earliest execution start time of task n_i on Virtual machine vm_p. The execution time of task n_i on virtual machine vm_p is termed in Eq. (9).

$$ETime(n_i, vm_p) = \frac{r(n_i)}{c(vm_p)}. \tag{9}$$

Since the computational capacity of virtual machine varies from one to other in cloud systems, the execution times of tasks on virtual machines are too varying. The earliest execution completion time $CTime(n_i, vm_p)$ of task n_i on virtual machine vm_p is computed as given in Eq. (10).

$$\begin{aligned} CTime(n_i, vm_p) &= STime(n_i, vm_p) + ETime(n_i, vm_p) \\ &= STime(n_i, vm_p) + \frac{r(n_i)}{c(vm_p)}. \end{aligned} \tag{10}$$

Since the processing time $r(n_i)$ involves normal distribution with random variable, the execution time and earliest completion time are random as well. The task n_i that is executed on the virtual machine is termed as $proc(n_i)$.

$$sch = \{proc(n_1), proc(n_2), \ldots, proc(n_n)\} \tag{11}$$

Equation (11) denotes the entire schedule of application. *Sch* gives the list of processors task that are assigned to tasks list. The finish time of virtual machine vm_p is found by Eq. (12)

$$FTime(vm_p) = \max_{proc(n_i)=vm_p} \{CTime(n_i, vm_p)\} \tag{12}$$

The Data Ready Time of task n_i on virtual machine vm_p is denoted by $DRTime(n_i, vm_p)$ and is computed using Eq. (13)

$$DRTime(n_i, vm_p) = \max_{n_{j \in par(n_i), proc(n_j) \neq vmp}} \{C(e_{j,i})\} \tag{13}$$

where $par(n_i)$ is the function which returns the immediate successors of task n_i and $C(e_{j,i})$ denotes the communication finish time of edge $e_{j,i}$ and is computed by Eq. (14).

$$C(e_{j,i}) = CTime\left(n_j, vm_{proc(n_j)}\right) + r(e_{i,j}) \tag{14}$$

If n_i is an entry task, i.e., $par(n_i) = \emptyset$, $DRTime(n_i, vm_p) = 0$, for all $vm_p \in VM$. The start time of task n_i on virtual machine vm_p is restricted by virtual machine $vm'_p s$ finish time $FTime(vm_p)$ and edges of n_i. It is defined in Eq. (15)

$$\begin{aligned} STime(n_i, vm_p) &= \max\{DRTime(n_i, vm_p), FTime(vm_p)\} \\ &\forall n_i \in N \quad \text{and} \quad vm_p \in VM, \end{aligned} \tag{15}$$

Assuming $STime\left(n_{entry}, vm_{proc(n_{entry})}\right) = 0$, the scheduling process begins and the makespan is computed by Eq. (16). Since our workflow has precedence constrained tasks, the completion time of the exit task will be the makespan of the schedule. The exit task is the last task to be executed and all the other tasks will be executed beforehand.

$$makespan = CTime(n_{exit}, vm_{proc(n_{exit})}) \tag{16}$$

A normal random variable, X is denoted as $X \sim ND(\mu, \sigma^2)$. Hence, the basic operations on normal random variable

1. If X_i is normally distributed with expected mean μ_i and variance σ_i^2 then $X = \sum_{i=1}^{n} X_i$ is also normally distributed with mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$. I.e.,

$$X \sim ND\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_2'\right). \tag{17}$$

2. The maximum value of a set of normal random variables is not a normal random variable. The maximum value has to be found in Eqs. (12), (13) and (15). But the maximum value from the set of values violates the property of normal random variable.

Finding the maximum value among the tasks is needed for prioritization. Clark [43] has established a technique to recursively estimate the estimated value and variance of the maximum value of a determinate set of normally distributed random variables. This is used for finding the maximum value among the tasks. The expected value of max{$C(e_{1,n})$, $C(e_{2,n})$} with $\rho_{1,2} = 0$ are computed by using Clark's first equation which is given in Eq. (18)

$$\begin{aligned} &E\left[MAX\left\{C(e_{1,n}), C(e_{2,n})\right\}\right] \\ &= E\left[C(e_{1,n})\right]\Phi(\xi_{1,2}) + E\left[C(e_{2,n})\right]\Phi(-\xi_{1,2}) + \varepsilon_{1,2}\psi(\xi_{1,2}) \\ &= E\left[C(e_{2,n})\right] + \left(E\left[C(e_{1,n})\right] - E\left[C(e_{2,n})\right]\right)\Phi(\xi_{1,2}) + \varepsilon_{1,2}\psi(\xi_{1,2}) \end{aligned} \tag{18}$$

where

$$\varepsilon_{1,2} = \sqrt{Var\left[C(e_{1,n})\right] + Var\left[C(e_{2,n})\right] - 2\rho_{1,2}\chi_{1,2}}$$
$$= \sqrt{Var\left[C(e_{1,n})\right] + Var\left[C(e_{2,n})\right]} \quad Since\ \rho_{1,2} = 0\ and,$$
$$\xi_{1,2} = \frac{E\left[C(e_{1,n})\right] + Var\left[C(e_{2,n})\right]}{\varepsilon_{1,2}}, \quad and$$
$$\chi_{1,2} = \sqrt{Var\left[C(e_{1,n})\right] Var\left[C(e_{2,n})\right]}, and$$
$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}, \quad and$$
$$\Phi(x) = \int_{-\infty}^{x} \psi(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt.$$

The variance of max{$C(e_{1,n})$, $C(e_{2,n})$} is given by Clark's second equation in Eq. (19) as follows

$$\begin{aligned} Var[DRT(v_n)] &= Var\left[MAX\left\{C(e_{1,n}), C(e_{2,n})\right\}\right] \\ &= \left(E^2\left[C(e_{1,n})\right] + Var\left[C(e_{1,n})\right]\right)\Phi(\varepsilon_{1,2}) \\ &\quad + \left(E^2\left[C(e_{2,n})\right] + Var\left[C(e_{2,n})\right]\right)\Phi(-\varepsilon_{1,2}) \\ &\quad + \left(E\left[C(e_{1,n})\right] + E\left[C(e_{2,n})\right]\right)\varepsilon_{1,2}\psi(\varepsilon_{1,2}) \\ &\quad - E^2\left[MAX\left\{C(e_{1,n}), C(e_{2,n})\right\}\right] \end{aligned} \tag{19}$$

In a similar way, Clark's equations can be used recursively to determine the expected value and variance of maximum value from the set of values.

5 Stochastic Scheduling

Stochastic modelling is used to solve real-world problems in which unpredictable events are present. It involves probability distributions to model the problems. Depending on the situation, we have to choose the parameters that can be represented as a random variable. Because of stochastic behaviour, this model allows random deviation in one or more inputs over time and finds the estimation of potential outcomes in order to guess what will happen under disparate cases. In the present work we focus on representing the task execution time $r(n_i)$ and the inter-task communication time $r\left(e_{i,j}\right)$ as random variables.

The user will submit the workflow G, which has normal distribution based task execution time $r(n_i)$ and the inter-task communication time $r\left(e_{i,j}\right)$. The stochastic behaviour of the tasks is to be handled and the cloud providers have to perform dynamic allocation by finding the apt priority level of the task. Each virtual machine has different computational capacity. Similarly, the tasks from the users have different execution time. The Providers' motto is to increase their system throughput and efficiency. The user's aim is to pay less for the VM usage. So the providers have to find the optimum Task-VM pair according to complement task's execution time and virtual machines' computational capacity in order to satisfy the user as per the SLA agreement. This Task-VM pair should minimize the makespan and satisfy the user's security demand.

5.1 Stochastic Top Level (STL)

The real-world problem modelled as a workflow has precedence constrained stochastic tasks. Finding the priority of the tasks is the major step in scheduling strategy. The prioritization of the tasks in the workflow has to be done with utmost care for stochastic tasks. This is due to the dependency of the task execution time $r(n_i)$ and the inter-task communication time $r\left(e_{i,j}\right)$ on the random variables. Generally scheduling algorithms use top_level or bottom_level as the key factor for prioritization. In the present work we focus on using top_level as the factor for prioritization. The primary factor to be considered is the varied execution time of the same task on different VM instances. To resolve the issue of different execution time of the same task on different VM instances many strategies have been tried and it was concluded that using the average value gives the best result in the research work in [24]. Hence we use average computation capacity of the VM in our Stochastic Top Level (STL) calculation and it is defined in Eq. (20). The average computational capacity of all virtual machines is defined as,

$$\overline{c(VM)} = \frac{1}{p} \sum_{i=1}^{p} c(vm_i) \tag{20}$$

Algorithm StLevel(G)

Input: A Workflow G = (N,E) of precedence constrained Stochastic Tasks, N-Set of task nodes, E-set of edges

Output: Stochastic Top Level for Task Set N, STL[N]

1. Initialize Task array T[|N|] from G
2. $T_0[|N|] = TopologicalSort(N, |N|)$;
3. Task n_{entry}=$T_0[0]$;
4. Compute $STL(n_{entry})$ using Eq. (22)
5. While $(index \leq T_0.length)$
6. {
7. Task $n_x = T_0[index]$;
8. For each parent n_i of n_x apply Eq. (17) to calculate the expected value and variance of $(STL(n_x)) = r(e_{x,i}) + STL(n_i)$;
9. If the task n_x has more parents then
10. For all other parent task n_i
11. Apply Eqs. (18) and (19) to calculate the expected value and variance of $STL(n_x) = \max\{STL(n_x), r(e_{x,i}) + STL(n_i)\}$;
12. End if
13. Construct the approximate normal distribution of $STL(n_x)$ using the expected value and variance of $STL(n_x)$
14. Apply Eq. (17) to compute $STL(n_x) = STL(n_x) + \frac{r(n_x)}{c(vm)}$;
15. index++;
16. }

Another main factor is that the tasks considered are stochastic tasks, whose execution time and inter-task communication time depend on random variables. The STL will consider the completion time of the tasks and tend to schedule the tasks which can find a free VM with the security demand satisfied earlier. In stochastic scheduling algorithm it is critical to calculate STL since task processing time and inter-task communication time are unpredictable in nature. The computation procedure of the STL is defined in the **StLevel Algorithm**. The StLevel algorithm returns the stochastic top level of all the tasks.

The stochastic top level of task n_x is the arbitrary distance of a longest path from the task n_x to entry task, and recursively defined as in Eq. (21),

$$STL(n_x) = max_{n_i \in par(n_x)} \left\{ STL(n_i) + r(e_{i,x}) + \frac{r(n_i)}{c(vm)} \right\} \tag{21}$$

The stochastic top level of entry task is defined as,

$$STL(n_{entry}) = \frac{r(n_{entry})}{c(vm)} \tag{22}$$

The StLevel algorithm gets the workflow G as the input and returns the STL value as the output. At first the task set N in the workflow G will be sorted in the topological order. Initially the STL for the entry task is calculated. The while loop defined from line 5–16 calculates the STL values for all the remaining tasks recursively. To estimate the relative priority among the child tasks of a single parent the mean and variance are used. The STL values for all tasks are used to find the precedence constrained tasks among the set of tasks.

5.2 *Trust Based Stochastic Scheduling (TSS)*

This section describes the Trust based Stochastic Scheduling (TSS) algorithm which is in accordance with DLS [22, 44] and SDLS [23] algorithms. The Trust based Stochastic Scheduling (TSS) algorithm is based on Trust based Stochastic Dynamic Level (TSDL) which is defined in Eq. (24). The TSDL is the estimated dynamic priority level based on trust and STL values with different tasks and virtual machines. In the computation of TSDL we have to consider the variation in the computation capacities of the virtual machines for the same task so that the best virtual machine is selected for tasks. To consider the variation in the computation capacities among the VMs, a factor δ is used, which will represent the relevance of the VM vm_p for the task n_x as an integer value. It is defined as $\delta(n_x, vm_p)$ as given in Eq. (23).

$$\delta(n_x, vm_p) = \frac{r(n_x)}{\overline{c(vm)}} - \frac{r(n_x)}{c(vm_p)} \tag{23}$$

If the value of $\delta(n_x, vm_p)$ is greater than zero, then it means that the VM vm_p will execute the task n_x faster. A positive increase in the value of $\delta(n_x, vm_p)$ indicates that the virtual machine vm_p is faster than other processors to execute the task n_x. If the value of $\delta(n_x, vm_p)$ is lesser than zero, then it means that the VM vm_p will execute the task n_x slower. A negative increase in the value of $\delta(n_x, vm_p)$ indicates that the virtual machine vm_p is slower than other processors to execute the task n_x. Since $STL(n_x)$ indicates precedence constrained level for task execution, this factor is required for computing the TSDL value in Eq. (24).

$$TSDL(n_x, vm_p) = \left\{ STL(n_x, vm_p) * TR_{i,j}^{SLA_k} \right\} - STime(n_x, vm_p) + \delta(n_x, vm_p) \tag{24}$$

The virtual machine vm_p with large $TSDL(n_x, vm_p)$ is allocated to the task n_x. The TSS algorithm finds the optimum Task-VM pair and it is given in TSS Algorithm.

Algorithm TSS(G)
Input: A Workflow G = (N,E) of precedence constrained Stochastic Tasks, N-Set of task nodes, E-set of edges
Output: Entire Schedule $sch = \{proc(n_1), proc(n_2), \ldots., proc(n_n)\}$.

1. Compute STL of each task using StLevel Algorithm
2. Initialize Stack ST(|N|);
3. ST.Push(n_{entry});
4. While (!ST.isEmpty())
5. {
6. For each task n_i in ST
7. {
8. For each VM type
9. {
10. Use Eq. (24) to calculate $TSDL(n_x, vm_p)$;
11. }
12. }
13. Find optimal Task-VM pair (n_x, vm_p) whose $TSDL(n_x, vm_p)$ is stochastically lesser than the TSDL of all other Tasks-VM pairs
14. Task n_i = ST.Pop();
15. Assign task n_i to virtual machine vm_p, i.e., $proc(n_i) = vm_p$;
16. Task [] L= free child tasks of n_i;
17. ST.push(L) $\forall\ task\ in\ L$;
18. Update the earliest execution start time of task on virtual machine;
19. }

6 Results and Discussion

The experiments were performed with precedence constrained stochastic tasks which is modeled as a DAG using random graph generator, where the task computation times and inter-task communication times among the tasks are normally distributed. The algorithms are also tested with Epigenomic and Montage, which are real world workflow for the performance analysis of the proposed TSS algorithm. The simulations are done using CloudSim simulator.

6.1 Analysis with Random Graph Generator

The experimental analysis with random graph generator is discussed. There are several parameters used by the random graph generator to define a DAG such as size, levels, link density, maximum and minimum expected values of task processing times and inter-task communication times. The expected value and variance of each task processing time on every processor are uniform random variables in the intervals $[T\mu_{min}, T\mu_{max}]$ and $[T\sigma_{min}, T\sigma_{max}]$ respectively. The expected value and variance of the communication time on each edge are uniform random variables in the intervals $[E\mu_{min}, E\mu_{max}]$ and $[E\sigma_{min}, E\sigma_{max}]$ respectively.

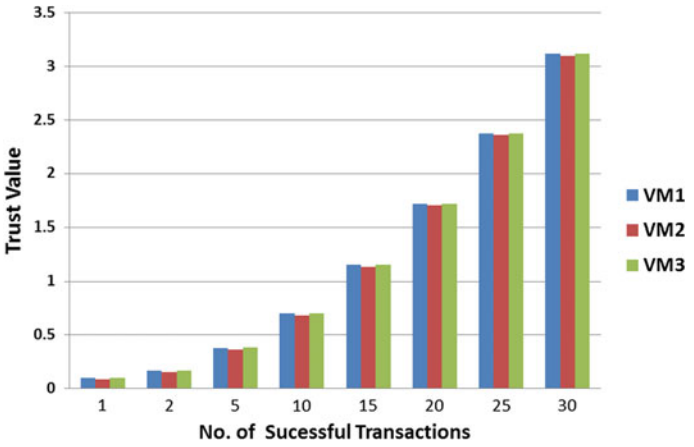


Fig. 5 Trust value calculated for three VMs with 20 users

The Communication to Computation Ratio (CCR) encodes the complexity of the computation of a task depending on the number of elements in the workflow application. Different CCR values are used in the experiments. The range of values used in the simulation is 0.1, 0.2, 0.5, 1 and 2 for CCR.

In order to differentiate the virtual machines' computational capacity, each virtual machine has been created by varying MIPS rate and PesNumber according to Table 1.

Figure 5 shows the average trust value calculated for some 20 users with different SLA and with three VM types for some 30 successful transactions. It can be noted that there is a steady increase in the trust value, which proves the efficiency of this trust model.

A scatter plot is used to depict the direct trust values calculated as per Eq. (7) for different arrival rate of the user jobs as shown in Fig. 6. The results indicate that the trust model works well when the arrival rate $\lambda = 0.5$.

Makespan and speedup are the parameters used for performance analysis. The makespan (or schedule length) is defined as the completion time of the exit task n_{exit} as defined in Eq. (16).

The speedup is computed by dividing the sequential execution time (i.e., the cumulative execution time) by the parallel execution time (i.e., the makespan of the output schedule) as shown in Eq. (25).

$$speedup = \frac{\sum_{n_i \in N} ETime(n_i)}{makespan} \tag{25}$$

where $ETime(n_i)$ is the execution time of task n_i.

The sequential execution time is computed by assigning all stochastic tasks to a single processor that minimizes the cumulative of the computation times. If the sum

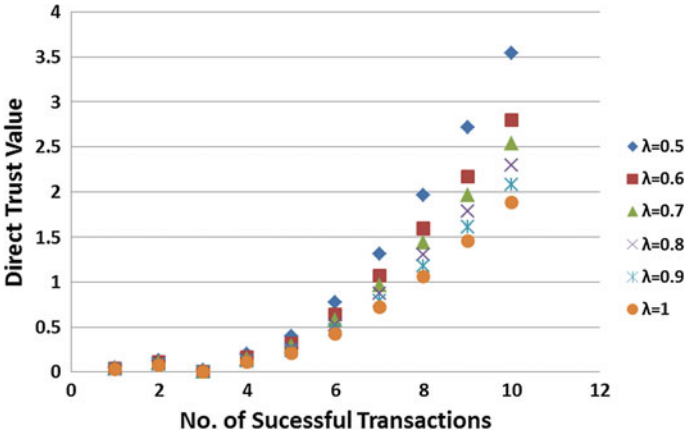
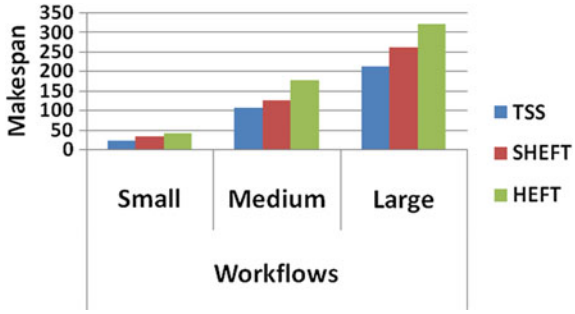


Fig. 6 Scatter plot for trust values with different arrival rates

Fig. 7 Makespan obtained for small, medium and large size of workflows



of the computational times is maximized, it results in higher speedup, but ends up with the same ranking of the scheduling algorithms.

The Trust based Stochastic Scheduling (TSS) algorithm is compared with the HEFT and SHEFT scheduling algorithms in terms of makespan and speedup. First, an analysis based on the performance metrics, makespan and speedup is done, by comparing the proposed mechanism with the existing scheduling algorithms with different number of virtual machines having different security guaranteed levels as specified in Table 3 and [8]. Different sizes of workflows are tested. These workflows are termed as small, medium and large based on the number of tasks in the workflow. Workflows with 25, 50 and 100 tasks are generated and they are designated as small, medium and large workflows respectively. The results are shown in Figs. 7 and 8. It is clearly observed that the proposed TSS algorithm is better than the existing HEFT and SHEFT scheduling algorithm in terms of makespan and speedup.

Figure 9 is the plot of the makespan for a set of 30 tasks and by varying the number of virtual machines. From the plot it is very clear that the proposed TSS algorithm consistently gives good results.

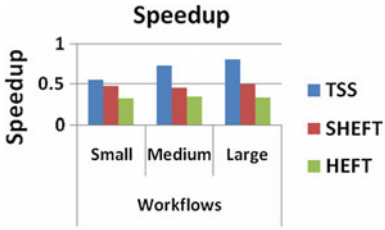


Fig. 8 Speedup obtained for small, medium and large size of workflows

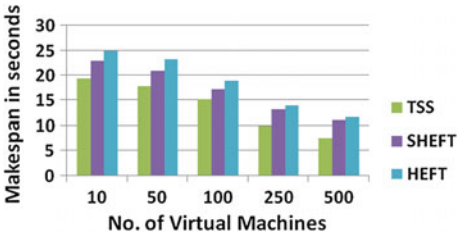


Fig. 9 Makespan obtained by varying the number of virtual machines

This is because the TSS algorithm takes the expected value and variance of the task processing times and inter-task communication times for the stochastic scheduling problem. While the other existing scheduling algorithms take only the expected values of task processing time and inter-task communication time that are less suitable for stochastic task scheduling problem. Moreover, if the number of processors increases, the makespan of the schedule decreases and speedup increases. The results of TSS algorithm are better than the existing algorithms in terms of makespan and speedup. This is due to the fact that TSS algorithm can process the stochastic tasks with normally distributed task processing time and inter-task communication times using Clark's equations, finding the optimized schedule. Also the TSS algorithm considers trust level and stochastic top level.

7 Conclusion

This chapter have focused on two major issues in cloud and fog computing namely security and scheduling. It also introduced a trust model for cloud cum fog environment. The present work has integrated security demand of a cloud user with an efficient task scheduling for a workflow. A novel trust model based stochastic scheduling algorithm is introduced. The inclusion of variance of the stochastic tasks in the prioritization phase using STL helps in achieving optimized schedule. Though the inclusion of the trust model is an additional overhead it doesn't affect the speedup of the TSS algorithm. The proposed TSS method has been implemented and compared with the existing HEFT and SHEFT algorithms. It is found that the designed TSS gives better results.

The trust model proposed in the work can be further improved by extending the untrusted event recording mechanism to a broader range than for individual regions. Also in future this work could be extended to multicloud environment. Besides, energy awareness may be included as another objective under concern.

References

1. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Das, H.: Mistgis: optimizing geospatial data analysis using mist computing. In: Progress in Computing, Analytics and Networking, pp. 733–742. Springer, Singapore (2018)
2. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., Mankodiya, K.: Fog Assisted Cloud Computing in Era of Big Data and Internet-of-Things: Systems, Architectures, and Applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
3. Wang, T., et al.: A novel trust mechanism based on fog computing in sensor cloud system. Future Gener. Comput. Syst. (2018). https://doi.org/10.1016/j.future.2018.05.049
4. Nitti, M., Girau, R., Atzori, L.: Trustworthiness management in the social Internet of Things. IEEE Trans. Knowl. Data Eng. **26**(5) (2014)
5. Durillo, J.J., Prodan, R.: Multi-objective workflow scheduling in Amazon EC2. Clust. Comput. **17**(2), 169–189 (2014)
6. Malawski, M., Figiela, K., Nabrzyski, J.: Cost minimization for computational applications on hybrid cloud infrastructures. Future Gener. Comput. Syst. **29**(7), 1786–1794
7. Yu, H., Kaminsky, M., Gibbons, P.B., Flax-man, A.D.: SybilGuard: defending against sybil attacks via social networks. IEEE/ACM Trans. Netw. **16**(3), 576 589 (2008)
8. Xie, T., Qin, X.: Scheduling security-critical real-time applications on clusters. IEEE Trans. Comput. **55**(7) (2006)
9. Tang, X., Li, K., Zeng, Z., Veeravalli, B.: A novel security-driven scheduling algorithm for precedence-constrained tasks in heterogeneous distributed systems. IEEE Trans. Comput. **60**(7), 1017–1029 (2011)
10. Xie, T., Qin, X.: Performance evaluation of a new scheduling algorithm for distributed systems with security heterogeneity. J. Parallel Distrib. Comput. **67**, 1067–1081 (2007)
11. Jia, C., Xie, L., Gan, X.C., Liu, W., Han, Z.: A trust and reputation model considering overall peer consulting distribution. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **42**(1), 164–177 (2012)
12. Zhang, P., Zhou, M., Fortino, G.: Security and trust issues in fog computing: a survey. Future Gener. Comput. Syst. **88**, 16–27 (2018)
13. Al-Kahtani, M.A., Sandhu, R.: Induced Role Hierarchies with Attribute-Based RBAC, SACMAT03, June 2–3, Como, Italy. ACM 1-58113-681-1/03/0006 (2003)
14. Wang, W., Zeng, G., Tang, D., Yao, J.: Cloud-DLS: dynamic trusted scheduling for cloud computing. Expert Syst. Appl. Elsevier **39**, 23212329 (2012)
15. Tao Xie and Xiao Qin: Security-aware resource allocation for real-time parallel jobs on homogeneous and heterogeneous clusters. IEEE Trans. Parallel Distrib. Syst. **19**(5), 682–697 (2008)
16. Gary, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Co., San Francisco, CA (1979)
17. Kar, I., Parida, R.R., Das, H.: Energy aware scheduling using genetic algorithm in cloud data centers. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3545–3550. IEEE (2016)
18. Kar, I., Das, H.: Energy aware task scheduling using genetic algorithm in cloud datacentres. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 106–111 (2016)
19. Topcuoglu, H., Hariri, S., Wu, M.-Y.: Performance-effective and low complexity task scheduling for heterogeneous computing. IEEE Trans. Parallel Distrib. Syst. **13**(3), 260–274 (2002)

20. Arabnejad, H., Barbosa, J.G.: List scheduling algorithm for heterogeneous systems by an optimistic cost table. IEEE Trans. Parallel Distrib. Syst. **25**(3), 682–694 (2014)
21. Tang, X., Li, K., Liao, G., Li, R.: List scheduling with duplication for heterogeneous computing systems. J. Parallel Distrib. Comput. Elsevier **70**, 323–329 (2010)
22. Sih, G.C., Lee, E.A.: A compile-time scheduling heuristic for interconnection-constrained heterogeneous machine architectures. IEEE Trans. Parallel Distrib. Syst. **4**(2), 175–187 (1993)
23. Li, K., Tang, X., Veeravalli, B.: Scheduling precedence constrained stochastic tasks on heterogeneous cluster systems. IEEE Trans. Comput. **63**(99), 191–204 (2013)
24. Zhao, H., Sakellariou, R.: An experimental investigation into the rank function of the heterogeneous earliest finish time scheduling algorithm. In: Proceedings of 9th International Euro-Par Conference, vol. 2790, pp. 189–194. Springer (2003)
25. Das, H., Jena, A.K., Badajena, J.C., Pradhan, C., Barik, R.K.: Resource allocation in cooperative cloud environments. In: Progress in Computing, Analytics and Networking, pp. 825–841. Springer, Singapore (2018)
26. Nayak, J., Naik, B., Jena, A. K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)
27. Sarkhel, P., Das, H., Vashishtha, L.K.: Task-scheduling algorithms in cloud environment. In: Computational Intelligence in Data Mining, pp. 553–562. Springer, Singapore (2017)
28. El-Rewini, H., Lewis, T.G.: Scheduling parallel program tasks onto arbitrary target machines. J. Parallel Distrib. Comput. **9**(2), 138–153 (1990)
29. Ilavarasan, E., Thambidurai, P., Mahilmannan, R.: High Performance Task Scheduling Algorithm for Heterogeneous Computing System, Distributed and Parallel Computing, Springer LNCS, vol. 3719, pp. 193–203 (2005)
30. Bertsekas, D.P., Castanon, D.A.: Rollout algorithms for stochastic scheduling problems. J. Heuristics **5**(1), 89–108 (1999)
31. Shmoys, D.B., Sozio, M.: Approximation algorithms for 2-stage stochastic scheduling problems. In: Lecture Notes in Computer Science, vol. 4513, pp. 145–157. Springer (2007)
32. Gourgand, M., Grangeon, N., Norre, S.: A contribution to the stochastic flow shop scheduling problem. Eur. J. Oper. Res. **151**(2), 415433 (2003)
33. Megow, N., Uetz, M., Vredeveld, T.: Models and algorithms for stochastic online scheduling. Math. Oper. Res. **31**(3), 513525 (2006)
34. Skutella, M., Uetz, M.: Stochastic machine scheduling with precedence constraints. SIAM J. Comput. **34**(4), 788802 (2005)
35. Tang, X., Li, K., Liao, G., Fang, K., Wu, F.: A stochastic scheduling algorithm for precedence constrained tasks on grid. Future Gener. Comput. Syst. **27**(8), 1083–1091 (2011)
36. Canon, L.C., Jeannot, E.: Evaluation and optimization of the robustness of DAG schedules in heterogeneous environments. IEEE Trans. Parallel Distrib. Syst. **21**(4), 532–546 (2010)
37. Kamvar, S., Schlosser, M., Garcia-Molina, H.: The Eigen trust algorithm for reputation management in P2P networks. In: Proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, pp. 640651 (2003)
38. Nielsen, M., Krukow, K., Sassone, V.: A Bayesian model for event-based trust. Electron. Notes Theor. Comput. Sci. **172**(1), 499–521 (2007)
39. Xiong, L., Liu, L.: Peer trust: supporting reputation-based trust for peer-to-peer electronic communities. IEEE Trans. Knowl. Data Eng. **16**(7), 843–857 (2004)
40. Zhou, R., Hwang, K.: Power trust: a robust and scalable reputation system for trusted peer-to-peer computing. IEEE Trans. Parallel Distrib. Syst. **18**(4), 460–473 (2007)
41. http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html. Accessed July 2016
42. http://docs.openstack.org/developer/nova/aggregates.html. Accessed July 2016
43. Clark, C.: The greatest of a finite set of random variables. Oper. Res. **9**(2), 145–162 (1961)
44. Kwok, K.Y.-K., Ahmed, I.: Static scheduling algorithms for allocating directed task graphs to multiprocessors. ACM Comput. Surv. **31**(4), 406–471 (1999)

Trust-Based Access Control in Cloud Computing Using Machine Learning

Pabitr Mohan Khilar, Vijay Chaudhari and Rakesh Ranjan Swain

Abstract Cloud computing is a distributed computing environment which hosts dedicated computing resources accessed anytime from anywhere. This brings many advantages such as flexibility of data access, data omnipresence, and elasticity [1–7]. As there is no control of data owner over the data, this brings security threats. Providing a secure cloud environment from the malicious user is one of the important and challenging tasks among scientific and business user community. Over the time, various control access models have been proposed for secure access in the cloud environment such as cryptographic-based access model, identity-based access control model and trust-based access control model. The users and cloud resources should be trusted before accessing the cloud. It is observed that the existing access control models mainly overlook the user behavior and scalability of the trust management system. We have considered the trust-based approach which provides access to the user in the cloud by their trust value computed based on the past accesses and behavior. We consider important parameters such as user behavior, bogus request, unauthorized request, forbidden request and specification of range. We proposed a trust evaluation strategy based on the machine learning approach predicting the trust values of user and resources. The machine learning techniques such as K-Nearest neighbor, decision tree, logistic regression and naive Bays are considered as the important strategies to evaluate the trust management system in our proposed work. We implemented our proposed machine learning method in jupyter notebook simulator tool. We found better result in terms of efficiency, prediction time and error rate which is presented in the result section of this chapter.

P. M. Khilar (✉) · V. Chaudhari · R. R. Swain
Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela 769008, Odisha, India
e-mail: pmkhilar@nitrkl.ac.in

V. Chaudhari
e-mail: 216cs3173@nitrkl.ac.in

R. R. Swain
e-mail: 514cs1006@nitrkl.ac.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_3

Keywords Cloud computing · Machine learning · Access control
Cryptography based access model

1 Introduction

Cloud computing is a paradigm that shares computing and storage infrastructure over a scalable network of resources [9, 10]. In the modern world, data are scattered in different data centers and applications are run in remote servers. The cloud technology brings the scattered data and the remote applications to user laptop in a virtual form. The main idea is to make computing and storage infrastructure available for cloud users irrespective of time and location. In order to commercialize the cloud technology, cloud users need to have the trust that the resource providers complete the submitted job as per the service level agreement (SLA) so that the information of the processed data is secured. Researchers believes that the biggest cloud computing issue is trust. Trust plays an important role in all commercial cloud environments and trust management is an integral part of commercial aspects of cloud technology. Cloud infrastructure supports three types of service delivery models such as Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). The cloud service providers offer infrastructure, platform and software to the users in an economical and trustworthy manner. Trust becomes a complex issue in the cloud computing arena. Companies like Google and Amazon have implemented reputation based trust management system and it helps the users to locate the trustworthy resource providers for doing e-business transactions in a secure and confident manner. E-bay has a built-in centralized model of trust. There are several trust frameworks which are studied in cloud environments.

Cloud computing provides a flexible solution to the on-line execution system for scalability applications. Fog computing provides a distributed solution to accomplish the elasticity, and scalability of effective information sharing system by reducing the computational cost. The fog computing features are same as cloud computing with extra attributes such as location awareness and edge data centric computing for big data processing. Edge computing provides a real time critical data processing in locally to reducing the traffic in the centralized storage of cloud. The IoT device data are collected and processed in the edge of the network before sent to the cloud storage. It is an optimization method of cloud computing with low latency, low transmission cost, reducing traffic, and reducing bottlenecks failure. Mist computing provides a lightweight service in network with small microchips and micro-controllers. It works with fog computing over the cloud platform. Mist computing contains local decision making data processing, which extends significant solution between centralized cloud computing and decentralized edge or fog computing. It is highly robust in nature.

In this chapter, we present the security related issues including various access control techniques present in the literature and a proposed access control model based on the trust value of the users and resources for cloud computing environment which prevent malicious users from accessing the resources available in cloud [6].

The basic information security requirements are confidentiality, integrity, availability, non-repudiation. Out of them the top most threats are denial of service (Dos) attack which is a security threat against availability [8]. We proposed a trust-based access control model using machine learning against denial of service (DoS) attack. In which, access to the user will be given based on the satisfaction of trust value. We compared our proposed method with the traditional methods that is able to fulfill the cloud computing security requirements.

2 Related Work

Cloud computing is the new technology which provides various kind of services to its users. Today users have high needs for good network bandwidth, very high computational power, high-speed memory devices and finally a huge space to store his huge amount of private data. The threats are also growing as the technology enhancing. There are various threats available against cloud computing, but researchers believes that DoS attack is the top most attack against the infrastructure of any service. Further we found that the DoS attack rate is growing year by year. DoS attack is a malicious attempt to make unavailable the resource, in which there is an involvement of an eve or attacker or a network of computer called as Bot.

If the security of cloud breach then resources of the cloud will definitely be affected and then the service-layer-agreement (SLA) between user and cloud-service-provider is violated. In fact, the service quality is reduced due to security breach in cloud computing environment. Therefore, protecting every users data and their computation is utmost necessary in recent years due to increasing dependence in the cloud by many users. Cloud-service-providers also have to protect their resources so that Quality-Of-Service can be maintained. Our main goal is to give access only those users which are not malicious and additionally select only those resources which are capable of providing good Quality-of-service to its customer.

The traditional methods are not able to fulfill the cloud computing security requirements. Therefore access-control-methods was developed to fulfill the security requirements of cloud computing. Various researchers have proposed different access-control-models for cloud. The access control methods such as MAC, DAC, attribute-based-access-control and role-based-access-models have been proposed by many authors in literature [14]. In Role-based-access-control-model (RBAC) role and permission are bond together. However, the RBAC model does not consider the user behavior in the cloud. Some researchers proposed other access-control-model called trust-based-access [15–22]. The existing techniques do not consider malicious activity done by various users seriously in cloud environment in order to give access in the cloud. Therefore we consider users behavior to provide secure access in cloud environment.

3 Objective of the Work

The objective of this work is given as follows:

(i) **Classification of Authorization Problem**: Design a trust-based-access-control-model using machine learning and deep learning which authorizes the user to give access the resources based on the trust value.
(ii) **Classification of Service Layer Agreement Problem**: Evaluating the trust value of cloud resources before giving access to the user.
(iii) **Control Access to Resources**: On the basis of calculated trust values, the users are prioritized. Based on the priority, the users get access to the demanded resources.

4 Proposed Work

Cloud computing is a new business model, which provides different services to the users and works on pay-per-use policy. The main advantage of using the cloud is that you can scale up/down your computation resources easily on-demands and can access your resources any time anywhere. It is very cost effective because you dont need to buy physical resources and you have to pay as-per-use basis. Most of the companys works on small projects, if they buy physical resources then there is loss as compared to profit from the project. If organizations want to buy physical infrastructure, then they have to install, maintain servers, networks etc., and hire manpower for this, then their profit is low. If they join cloud then they dont have to worry about these kind of Cloud resources. Cloud provides various kind of services such as Platform-as-a-service (PaaS), Infrastructure-as-a-service (IaaS) and Software-as-a-service (SaaS) [4].

Since all the data are stored and computation is performed on the cloud servers, users have not any kind of control over his data and computation. If any attack will happen in the cloud, than his private data and computation will be lost. Therefore security of the users computation and data is the responsibility of cloud-service-provider (CSP). This is mentioned in service-layer-agreement (SLA). The cloud-service-provider have to secure cloud form all the attacks and threats thats why various access-control-methods are applied to secure cloud environment. Before accessing the cloud services both cloud-service-provider and the user should trust each other. Now before giving access of service to the user, first access-control-mechanism will check the trust value of the user, and is found is above the threshold then it provide access to its service and the same behavior is applied upon cloud-service-provider. User trust value is based on the behavior parameters, CSPs trust value is also based on their service records, and the opinion they got from users. Access-control-methods are used to stop malicious users to access the cloud services.

Thus our main task is to improve the security of the cloud, so that various attacks and threats can be stopped. If any malicious user attacks the system then cloud service

is affected and cloud is unable to give Quality-of-service to its customer, then the Service-layer-agreement break and users trust on the cloud-service-provider loose. Therefore cloud-service-provider should provide access only to the authenticated users. To check the authenticity of the user, CSP checks their trust values. If user trust value is greater than the threshold value, then the user is considered as an authentic user. It is noted that the trust value of the users depend upon their behavior parameters in the cloud.

4.1 Proposed Model

We proposed a trust-based-access-control-model consisting various sub-modules using machine learning shown in Fig. 1. If any user wanted to access any cloud service then his request will pass through several modules before completing the authorization process. All the services and resources are protected through this proposed model in cloud computing environment. Resource catalog contains several resources which have the different reputation and trust value in the cloud environ-

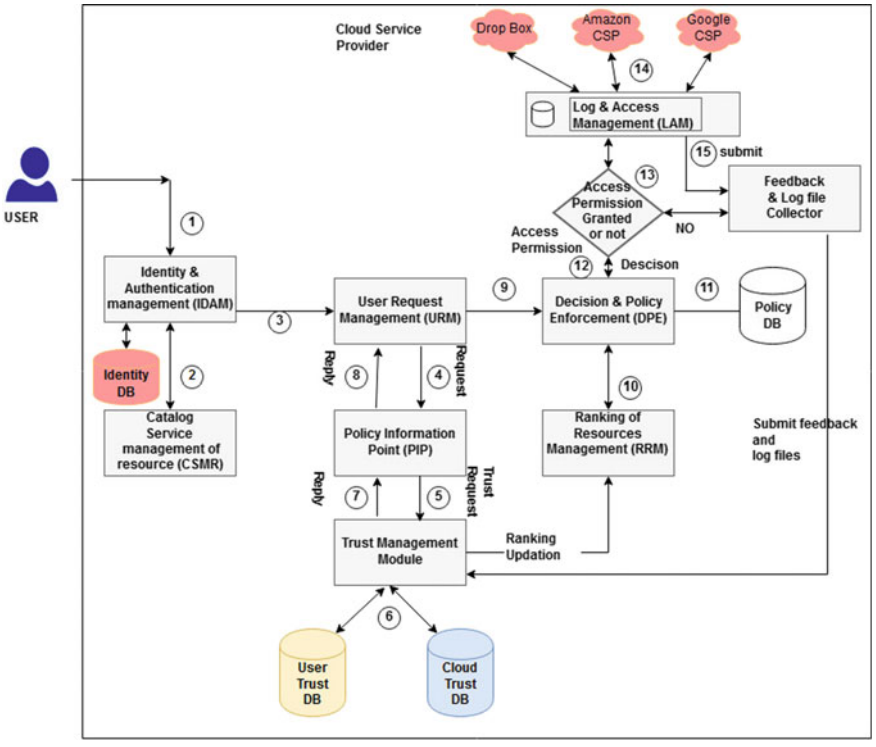


Fig. 1 Proposed architecture

ment. When a user requests a resource our proposed model checks whether the requesting user is a trusted user or not. If the user is found as a trusted user then the best resource for the requesting user is provided from the mutual trust relationship and his computational needs.

We describe the functions of each sub-module of the proposed architecture as follows.

(i) **Identity and Authentication Management (IDAM)**: This is the module responsible for storing users authentication and identity information such as username and password. It acts as an interface between users and the system. Registration of new users, password change or update everything is handled by this module. When login request comes then from the identity database of the users, it matches all the credentials and decided if access is granted or refused. Requested service chosen from the catalog with user id forward to the URM for further authentication purpose by this module.

(ii) **Catalog service Management of Resource (CSMR)**: Catalog service Management of resource (CSRM) updates records of all the services offered by cloud-service-provider (CSP). When a user enters into the system then from the catalog he can choose services offered by CSPs.

(iii) **User Request Management (URM)**: This module gets users trust from the PIP module and then it forward all the information in request vector (ReqInfo [userId, UserTrust, requested-resourceId]) to the Decision and Policy Enforcement module.

(iv) **Policy Information Point (PIP)**: This module is responsible for providing user trust value when he get the request from URM module.

(v) **Decision and Policy Enforcement (DPE)**: In access-control-model DPE is a very important component, which decide whether requested resource should be given to the user no not. When DPE receives ReqInfo[..] vector from URM then it fetch policy details from the policy database. It then compare all the users details with the security policies of the cloud-service-provider. If the user and the requested resource passes all the security policies threshold value, then access is given to user else this activity of the user stores in feedback and log file collector module. This module provides access based on mutual trust (it is defined as user and resource both should be trusted).

(vi) **Log and Access Management (LAM)**: This is the module which is responsible for trusted user secure access to the cloud services. Users activities and his behavior stores in Log files. After execution of the request, it submit his feedback to the Trust management module. Log file contains the activities of the users so this information is used to update the trust value of the users.

(vii) **Ranking of Resources Management (RRM)**: This module contains the list of all the resources according. All the resources are ranked according to their capabilities and performance which updates time to time when RRM got the trust reputation and trust values from the TMM. Performance of the resources is based on their trust value and reputation in the cloud. There are different

cloud-service-providers who provide a similar type of services to its users. Best resource should be given to the users which are handle by RRM module.

(viii) **Trust Management Module (TTM)**: Trust management module is accountable for handling the real-time value of user trust based on trust evidence of user. This evidence is based on user behavior in the cloud and used to evaluate trust value of the user (trustworthiness). This is the module which takes Log file (shows every interaction of the user based on access of cloud service) and based on the behavior trust evidence, it recalculate the value of the trust. It also calculates cloud resource reputation and trust value and then send those trust values to the RRM module for resource ranking purpose.

4.2 Authorization Process

When a user wants to access any cloud service he has to submit all his requirement details to the CSP such as Processor, computing power, software, networking speed, security and kind of operating system. There is no negotiation between the user and cloud-service-provider about providing these services and financial charge. The negotiation is an agreement which is called as service-layer-agreement (SLA). The following tasks describe the step by step functions about the whole authentication process of the user request for access of a required service as given below.

(i) **Step 1**: IDAM receives and check the Login credentials from the user. It check the login credential and if those credentials are found correct, then the user can proceed further to access cloud services.

(ii) **Step 2**: The user chooses his required service from the catalog service management of resource (CSMR). This module then sending user request to the IDAM.

(iii) **Step 3**: IDAM collect user requirement from CSMR module and then form a request info vector, which contains the user id and request service information. Then IDAM forward this request info vector to the user request management module (URM).

(iv) **Step 4**: User request module (URM) send user id to the PIP for trust value and pip forward it to TMM. TMM forward user trust to policy information point (PIP).

(v) **Step 5**: Now URM adds user trust value to the request info vector and then forward this to the Decision and policy enforcement (DPE) module.

(vi) **Step 6**: Finally, DPE collects all the information from URM. DPE compares received vector information with the required security policies access from the policy database.

(vii) **Step 7**: RRM module select the most trusted and suited resource according to the job and then send it to the DPE.

It provides secure access to the user so that the user can access his requested service safely. It also contain log file which forward to the TMM module later. TMM

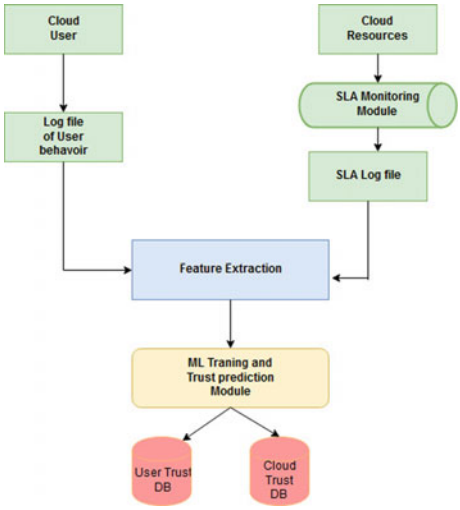


Fig. 2 Architecture of trust management module

module process the log file of the user behavior and SLA. Update user trust DB and cloud trust DB respectively. Our proposed model is able to achieve secure access based on mutual trust.

4.3 Trust Management Module (TMM)

Trust Management module is the most crucial part of trust-based-access-control-model shown in Fig. 2. It is used for calculation of trust for both cloud-service-provider and user. It is composed of various sub-modules. One thing has to be noticed here is that when it receives the log files either cloud users or cloud-service-provider, users have to calculate trust value quickly, so that malicious user can be found soon before they can harm the system and likely to be best cloud resource can be provided to the user quickly through updating the current trust value of resource.

This model is not biased towards any cloud-service-provider or user too. It can be observed that the TMM contains several phases for calculating trust value. We explain the way the trust value is calculated based on the activities and functionalities of all the phases involved.

(i) **Cloud user**: every user has to register before accessing the cloud service. Identity database contains all the authorization details of all the users. When a user requests a resource from cloud server then his authentication and authorization have to be checked by various component and if he passes then he can access the cloud service.
(ii) **Cloud resources**: Any hardware or software part can be cloud resource. All cloud resources access based on their reputation and trust value in the cloud.

Before giving access to the resource, users authorization has to be checked properly and all service-layer-agreement should be fulfilled.

(iii) **Log file of user behavior**: Log file create when Log and access Management module (LAM) record all activities performed by the user in the cloud. This log file captures real user behavior through which user trust value is predicted by machine learning algorithms.

(iv) **SLA monitoring module**: SLA monitoring module captures the real-time behavior of cloud resources. This also captures the performance of the cloud resources.

(v) **Feature extraction module**: All the log files feeds into this module. This module captures all the related parameters and their corresponding values from the log files. Those value passes to the ML training and trust Prediction module.

(vi) **ML training and test prediction module**: this module takes all the parameters and data from the feature extraction module. Feed this data to our machine learning model and train the models. Now our trained model predict the trust value of the cloud users.

4.4 Trust Evaluation Parameters

To calculate trust value of cloud users and cloud-service-providers, we use generic trust evaluation parameters. In order to calculate user trust value, the parameters such as user behavior is considered. Whereas for computing trust value for various resources, SLA parameters and public opinion are used.

4.4.1 User Behavior Parameters

As it is mentioned earlier the user trust value is calculated based on user behavior parameters. Various attackers or malicious users perform several malicious activities to steal the private data and computation. When a malicious user attacks the system then services get affect and cloud-service-provider is not be able to fulfill the service-layer-agreement. Monitor module monitors all the interactions of the users and stores those interaction info. into the user-behavior-database. To evaluate the trust value of users, several parameters have been taken into considerations. To ease the description, we specify the following notations and their meaning in the Table 1.

1. **Bogus Request Rate (BRR)**:
   Bad requests or the bogus requests are such types of requests in which syntax use in a malformed way due to this, Cloud server is not be able to understand this type of requests. When an attacker or the malicious user launch a denial-of-service (DoS) attack into the system then he sends a lot of packet request in short period of time. DoS attack leads to resource exhaustion and bandwidth depletion that consumes most of the the bandwidth. In this attack, the availability of cloud resource affects. Bogus request rate is calculated by Eq. 1.

Table 1 Notations and their meaning

Sl. no.	Notation	Meaning
1	BR	Bad request
2	UR	Unauthorized request
3	FR	Forbidden request
4	NF	Not found request
5	MNAR	Method not allowed request
6	RNS	Range not specified
7	TR	Total request
8	UT	User trust
9	AUT	Average user trust

$$BRR = BR/TR \tag{1}$$

2. **Unauthorized Request Rate (URR)**:
Unauthorized requests mean that users are trying to access those resources (R k) for which he has not proper authorization. This type of requests indicates that the malicious user is trying to steal/modify data or computation. These illegal operations can be via malicious code in the programs or through some malicious computation. Unauthorized request rate (URR) is calculated by Eq. 2.

$$URR = UR/TR \tag{2}$$

3. **Forbidden Request Rate (FRR)**:
Forbidden request encounter when a user trying to access some file or data after successfully logged in into the system but fails due to not having proper authority to access that file/data. It shows the user is trusted to some level of degree but is nt allowed to access that file or resource (R k). Malicious users use these type of information to leak the data. Forbidden request rate is defined in Eq. 3.

$$FRR = FR/TR \tag{3}$$

4. **Not Found Request Rate (NFRR)**:
Not Found request encounter by the users, when the server doesnt find anything on the requested location. This happens when a website has been moved to another server recently but DNS still points to the old location. Attackers use this to, increase the bounce rate of the cloud resource (R k). Not Found request rate is defined in Eq. 4.

$$NFRR = NFR/TR \tag{4}$$

5. **Method Not Allowed Request Rate (MNARR)**:
Method Not Allowed request (MNAR) is the type of the response when user forward such type of request to the web server with an HTTP method that due to its configuration, it is not allowed. This type of response comes when a particular

Table 2 Description of notations

Sl. no.	Notation	Meaning
1	SJ	Total submitted job for period of time T
2	ACP	Total accepted tasks for period of time T
3	TC	Total completed task in time period T
4	ART	Average response time
5	BD	Bandwidth of the network
6	UPF	Positive feedback given by the user
7	UNF	Negative feedback given by the user
8	RCP	Correct opinion received by other services
9	RNP	Negative opinion received by other services
10	GCP	Correct opinion given to other services by resource
11	GNP	Negative opinion given to other services

HTTP method is a ban on a particular web resource (R k) from the owner due to security concerns. MNAR rate (MNARR) can be calculated in Eq. 5.

$$MNARR = MNAR/TR \tag{5}$$

6. **Range Not Satisfied Request Rate (RNFRR)**:
Range Not Satisfied (RNF) response received when the client has asked for the range of the file which lies beyond the file size. The purpose of this kind of request is to create unnecessary traffic into the network. RNFR rate is calculated in the Eq. 6.

$$RNFRR = RNF/TR \tag{6}$$

4.4.2 SLA Parameters

SLA parameters are part of the service-layer-agreement between cloud user and cloud service provider. These are the parameters we use to calculate the resources trust value. In the cloud if security do not be properly managed by cloud-service-provider then the service Quality is reduced and that CSP is less trustworthy. The notations and their meaning are shown in Table 2.

There are various parameters involved in resource trust value calculation and defined as:

(i) **Resource Availability (RAV)**:
Availability of any cloud service can be defined in terms of resource accessibility. If any malicious user or attacker attack the system then services of the cloud at that time are inaccessible. Resource availability is defined as total number of accepted task from all the submitted task. RAV is calculated by Eq. 7.

$$RAV = ACP/SJ \tag{7}$$

(ii) **User Trust Satisfaction (UTS)**:
User trust satisfaction can be expressed as total number of successful tasks has been executed form accepted task in a time period T by a resource. User Trust Satisfaction is calculated in the Eq. 8.

$$UTS = TC/ACP \tag{8}$$

User Trust Satisfaction is also called reliability. Reliability of a system can be affected by several reasons like a failure of a job due to the scalability restrictions, due to time restriction or network failure.

(iii) **Service Quality (SQ)**:
This parameter is basically a collection of various parameters such as average response time and efficiency. The system convert average response time to between 0 and 1 based on the threshold value.

(iv) **Bandwidth of Network**:
Bandwidth (BD) can be defined as, in a fixed time interval how much data we can transmit using the network. The system converts its value into a probable value between 0 and 1 using threshold value.

(v) **Opinion**:
Opinion plays an import part in evaluating trust value of cloud services. Evaluation using opinion model shows us significant improvement from the existing trust model. We have considered the opinion from cloud users and also the opinion from the other service providers. UPF and UNF represent the positive feedback and negative feedback given by the user to the resource R_k. RCP, RNP represents all the correct and incorrect opinion R_k is received from other services. GCP and GNP represent all the correct and incorrect opinion R_k is given to the other services.

4.5 Trust Evaluation Strategy

When a time interval end then we calculate user trust value and cloud trust value based on their interaction with the cloud environment. Several interactions is there in this time interval. In our first step, we calculate user trust value in current time intervals by using his behavior or interaction parameters after that we calculate average trust value of the user from the current trust value and from the previous time intervals average trust value. Let t_n denotes the time interval of current window and $t_n - 1$ denotes the previous time interval window.

How to measure the trust value of each cloud user is an incumbent research problem in cloud computing. AS day by day cloud computing is growing very fast, securing cloud through trust is becoming more and more promising and has attracted a lot of researchers attention.

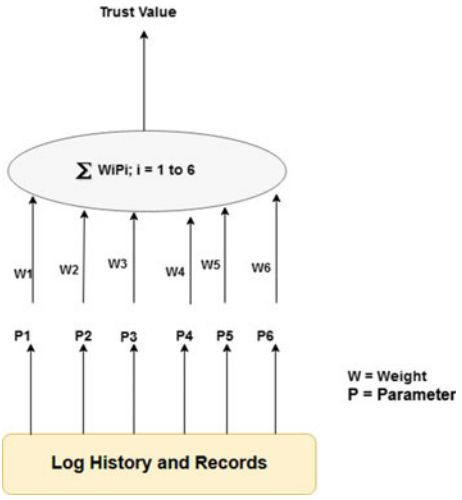


Fig. 3 Calculation of user trust value

Now from the Eq. 11 average user trust value (AUT) and user trust value (UT) will be calculated.

We have used Apache server Log file for data of the parameters. Figure 3 shows that every parameter is combined with weight factors and their strength value determines by their contribution factor. Some parameters are more dominant then other parameters. Since parameter P_1, P_2, P_3, P_4 determines the DOS attack, unauthorized access, information leak and to increase the bounce rate so that page ranking of that resource reduced respectively are more crucial modeling characteristics. To analyze the user behavior, their calculation is required and must be included for all users trust value. We have to quantify all the parameters further according to their contribution to determining user trust value. Thus trust value calculation for users is shown by Eq. 9:

$$\begin{aligned} T_{neq} =& W_1 \times BRR + W_2 \times URR + W_3 \times FRR + W_4 \times NFRR \\ &+ W_5 \times MARR + W_6 \times RNFR \end{aligned} \tag{9}$$

$$UT = 1 - T_{neg} \tag{10}$$

Figure 4 shows us the time window diagram for trust value calculation.

$$AUT = \alpha \times (UT)t_n + (1 - \alpha) \times (AUT)t_n - 1 \tag{11}$$

where $W_1 + W_2 + W_3 + W_4 + W_5 + W_6 = 1$.

Now Eq. 9 represents the negative trust value of a cloud user, while Eq. 10 is used to calculate the positive trust value of the cloud user. Finally from Eq. 11, we calculate average trust value. Average user trust value is calculated by the positive trust value at t_n time window and average trust value at $t_n - 1$ time window. The average trust value of past behavior plays a significant role of calculating the current

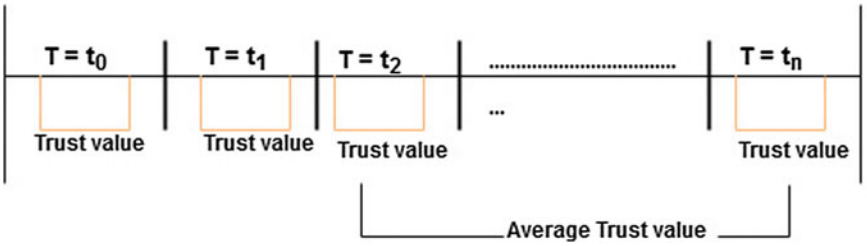


Fig. 4 Time window diagram for calculating average trust

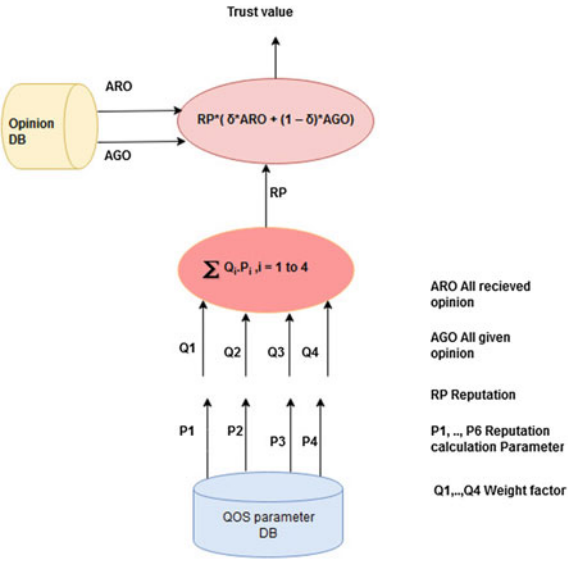


Fig. 5 Trust value calculation of cloud resource

average trust value of the user. In our second step, we calculate reputation and trust value of the resources R_k. The calculation of reputation and trust value of the resource shown in Fig. 5.

The formula for calculation of Reputation of cloud resource is shown by Eq. 12:

$$RP = Q_1 \times RAV + Q_2 \times UTS + Q_3 \times SQ + Q_4 \times BD \tag{12}$$

The formula for calculation of Trust value of cloud resource is shown below:

$$CT = RP \times (\delta \times (UPF + UNF + RCP + RNP) + (1 - \delta) \times (GCP + GNP)) \tag{13}$$

where $Q_1 + Q_2 + Q_3 + Q_4 = 1$. Where RP and CT represents the reputation of cloud resource and trust value of cloud resource respectively. Q_1, Q_2, Q_3 and Q_4 are the weight parameters of the RAV, UTS, SQ, and BD respectively. δ is the weight parameters which is associated with opinions. UPF, UNF, RCP, and RNP are the opinions

received by the resource R_k. GCP and GNP are the opinions given to other services by resource R_k.

5 Implementation Work and Results

We have implemented our proposed strategy using machine learning algorithms. Using our proposed model, we calculate trust value for both cloud user and cloud-service-provider. We have classified users and resources into different classes based on their trust values.

5.1 Experimental Setup

We used Jupiter notebook as codding editor and different libraries for building machine learning models such as Sckitlearn and Tensorflow. For calculation of trust, we have included different weight factors which represents the priority of parameters in security requirement. The different probabilities values of weight factors for SLA parameters and user behavior parameters is considered to calculate trust.

5.2 Result and Analysis

5.2.1 Trust Value of Different Type of Users

We have calculated average trust value of all cloud users in our dataset. Each users trust value is calculated from present time interval with the average trust value of past time window. The new average trust value with weight factors shown in Table 3. On the basis of the average trust value, we are classifying users into 4 categories such as malicious, moderate, high, and very high trust.

Table 3 Weight for user behavior parameters

Weight parameters	Probability value
W_1	0.2
W_2	0.2
W_3	0.2
W_4	0.2
W_5	0.1
W_6	0.1

Table 4 Weights for Resource Behaviour Parameters

Weight parameters	Probability value
Q_1	0.3
Q_2	0.4
Q_3	0.1
Q_4	0.2

We assume user trust threshold value is below 0.6 for the malicious user, between 0.6 and 0.8 for the moderate user, between 0.8 and 0.9 for the high trust user and above for very highly trusted users. Similarly we have shown SLA weight parameters value shown in Table 4.

5.3 Performance Evaluation of Cloud Users Using Access Model Based on Machine Learning

We have considered 9000 different users and based on trust value we have divided them into four classes. The threshold value below 0.6 is for malicious user, between 0.6 and 0.8 for moderate trusted user, for high trusted user it lies between 0.8 and 0.9 and above 0.9 it is for very high trusted user. We have taken 70–30 ratio of dataset for training and testing purpose.

5.3.1 Classification Accuracy of Machine Learning Algorithms for User

Figure 6 shows comparison between NTBAC [22] and proposed method after applying different classifiers. On applying KNN, Nearest Centroid, Gaussian NB, Decision Tree, Linear SVC, Logistic Regression, Ridge, and MLP classifiers we are getting 20, 16.48, 26.51, 23.46, 29.08, 23.7, 12.48, and 27.07% higher accuracy than the NTBAC model. On average we are getting 22.34% higher accuracy by using our proposed method rather than NTBAC method. Increase in accuracy is due to increased in number of parameters used in proposed model.

5.3.2 Performance Parameters Comparison of Machine Learning Algorithms

Table 5 shows comparison of Time, Mean Absolute Error, Root Mean Absolute Error, Precision, Recall, and F1-score between NTBAC and proposed method.

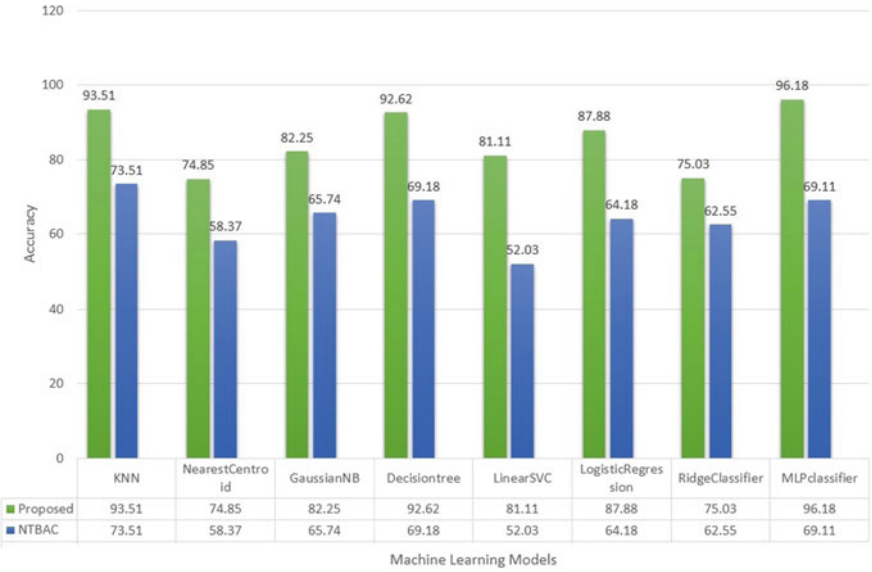


Fig. 6 Efficiency comparison between NTBAC and proposed method

Table 5 Comparison for ML and proposed algorithm performance parameters

ML algorithms/ parameters	Time (s)		MAE (%)		RMAE (%)		Precision		Recall		F1-Score	
	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop
KNN	0.12	0.99	28.92	6.48	58.15	25.45	0.75	0.94	0.74	0.94	0.74	0.94
Nearest centroid	0.26	0.005	54.7	25.29	89.91	50.59	0.62	0.77	0.58	0.75	0.58	0.75
Gaussian NB	0.009	0.0089	42.85	17.81	77.48	42.38	0.71	0.85	0.66	0.82	0.65	0.83
Decision tree	0.109	0.069	34.7	7.37	65.4	27.14	0.70	0.93	0.69	0.93	0.69	0.93
Linear SVC	1.859	1.83	49.07	19.33	71.62	44.96	0.45	0.81	0.52	0.81	0.40	0.80
Logistic regression	0.385	1.15	39.89	12.11	69.31	34.8	0.65	0.89	0.64	0.88	0.63	0.88
Ridge classifier	0.250	0.06	40.18	27.67	67.57	58.65	0.65	0.76	0.63	0.76	0.61	0.73

5.3.3 Classification Result

On the basis of trust value users are classified into four types such as malicious, moderate, high, and very high trust as shown in Table 6.

Table 6 Classification outcome for various ML techniques

ML algorithms/ users class	KNN		NC		GNB		DT		Linear SVCF		Logistic regression	
	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop
Class 0	515	663	552	745	269	595	723	679	324	687	322	680
Class 1	1169	1050	711	870	1039	941	991	1049	1932	1236	1255	948
Class 2	712	672	981	788	1046	860	674	618	95	427	791	711
Class 3	304	315	456	297	346	304	312	354	349	350	332	361

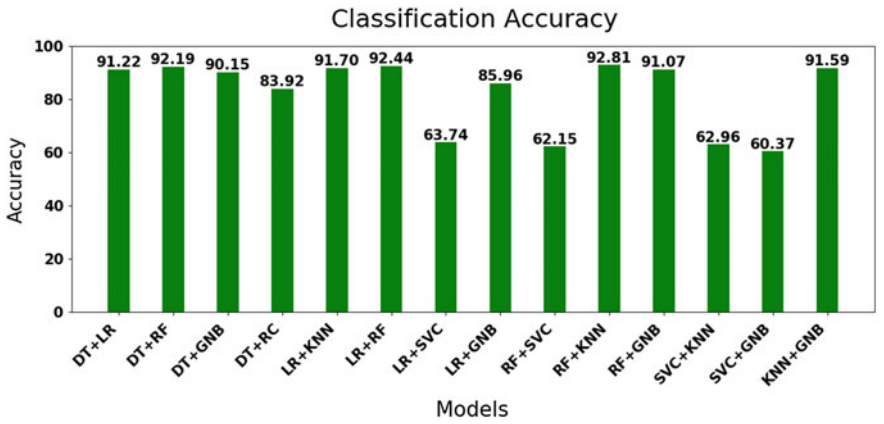


Fig. 7 Comparison of accuracy between various ensemble models

5.4 Performance Evaluation of Cloud Users Using Ensemble Machine Learning Algorithms

5.4.1 Accuracy of Ensemble Machine Learning Algorithms for Cloud User

Figure 7 shows accuracy comparison between ensemble machine learning model. Ensemble model of random forest and K-nearest neighbor perform better than that of other models. Highest accuracy we achieved is 92.81% using RF+KNN ensemble model.

5.4.2 Comparison of Performance Parameters of Ensemble Machine Learning Algorithms for Cloud User

Table 7 shows the comparison of Time, Mean Absolute Error, Root Mean Absolute Error, Precision, Recall, and F1-Score between various ensemble models of Machine Learning.

Table 7 Performance evaluation of ensemble models based on various parameters

ML algorithms/ performance parameters	Time (s)	Precision	Recall	F1-Score	MAE	RMSE
DT+LR	0.342	0.91	0.91	0.91	0.087	0.29
DT+RF	0.23	0.92	0.92	0.92	0.078	0.27
DT+GNB	0.093	0.90	0.90	0.90	0.098	0.31
DT+RC	0.45	0.84	0.84	0.84	0.185	0.48
LR+KNN	1.15	0.92	0.92	0.92	0.082	0.28
LR+RF	1.15	0.93	0.92	0.92	0.075	0.27
LR+SVC	10.9	0.45	0.64	0.51	0.497	0.87
LR+GNB	0.99	0.87	0.86	0.86	0.14	0.37
RF+SVC	9.75	0.45	0.62	0.50	0.51	0.88
RF+KNN	0.29	0.93	0.93	0.93	0.071	0.268
RF+GNB	0.201	0.91	0.91	0.91	0.089	0.29
SVC+KNN	9.71	0.45	0.63	0.51	0.50	0.87
SVC+GNB	9.5	0.44	0.60	0.49	0.53	0.89
KNN+GNB	0.125	0.92	0.92	0.92	0.084	0.289

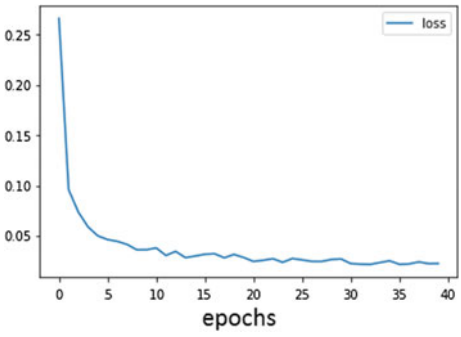


Fig. 8 Model with 40 neurons

5.4.3 Performance Evaluation of Cloud Users Using Keras Neural Network

Figures 8, 9, and 10 represent the graph of loss function by applying neural network. In Fig. 8 one hidden layer with 40 neurons, in Fig. 9 one hidden layer with 50 neurons, and Fig.10 three hidden layers with 40, 50, and 70 neurons has been consider. We are achieving efficiency 97.2, 97.38, and 98.46 respectively.

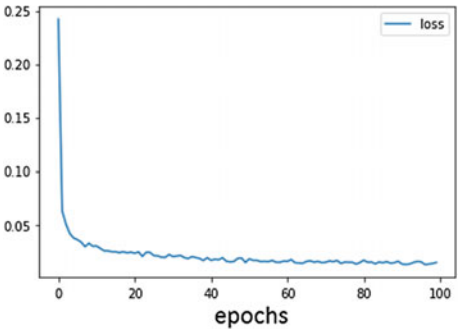


Fig. 9 Model with 50 neurons

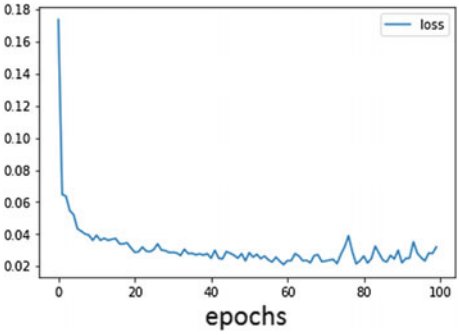


Fig. 10 Model with three layers having 40, 50, and 70 neurons

5.5 Performance Evaluation of Cloud Resource Using Machine Learning Algorithms

We have taken 3000 different Resources and based on trust value, we have divided them into 7 classes. Trust value below 0.5 belongs to Class 0, between 0.5 and 0.65 belongs to Class 1, Class 2 consists of trust value between 0.65 and 0.70, Class 3 trust value ranges between 0.70 and 0.75, Class 4 trust value between 0.75 and 0.80, Class 5 have trust value between 0.80 and 0.90, Class 6 trust value between 0.90 and 0.95 and greater than 0.95 belong to Class 7. We have taken 70–30 ration of the dataset for training and testing purpose.

5.5.1 Accuracy of Machine Learning Algorithms for Resources

We have shown graph of accuracy for cloud resources from Fig.11. We are calculating accuracy of cloud resources using ensemble machine learning models. We observed the accuracy better than existing method. The more accurate results we obtained by Decision tree, Support vector machine, logistic regression and MLP classifier machine learning approach over TMQoS [11–13].

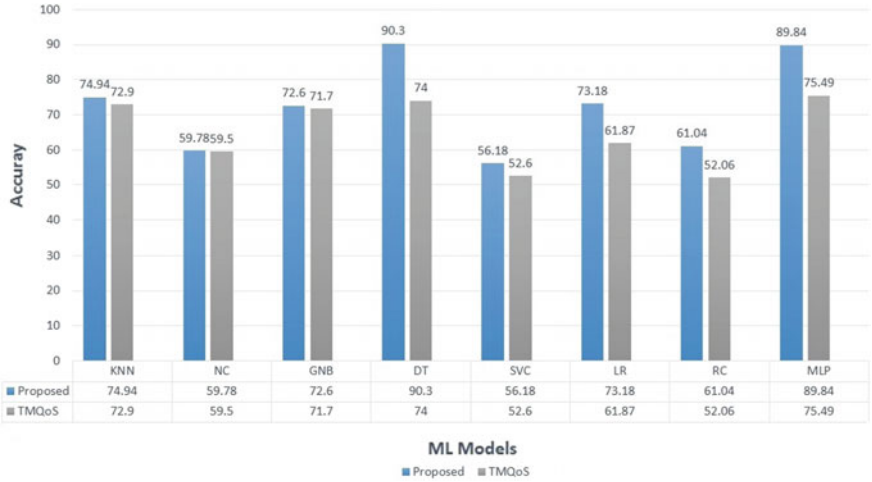


Fig. 11 Efficiency comparison between traditional and proposed method using machine learning

Table 8 Comparison for proposed versus traditional methods based on various parameters

ML algorithms/ parameters	Time (s)		MAE (%)		RMAE (%)		Precision		Recall		F1-Score	
	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop	Trad	Prop
KNN	1.07	0.23	27.3	25.1	52.8	50.4	0.72	0.74	0.73	0.75	0.72	0.74
Nearest centroid	0.15	0.10	44.3	43.9	72.4	71.7	0.62	0.63	0.60	0.60	0.59	0.57
Gaussian NB	0.10	0.016	29.4	27.5	56.3	52.8	0.71	0.71	0.72	0.73	0.68	0.70
Decision tree	0.40	0.35	26.5	9.6	51.4	31.1	0.74	0.90	0.74	0.90	0.74	0.90
Linear SVC	14.11	22.7	48.1	35.7	76.2	62.3	0.45	0.48	0.53	0.56	0.44	0.47
Logistic regression	3.6	6.7	44.6	28.8	76.8	55.7	0.53	0.75	0.62	0.73	0.55	0.70
Ridge classifier	0.03	0.98	74.3	43.3	84.5	72.9	0.42	0.51	0.52	0.61	0.40	0.53

5.5.2 Various Performance Parameter Comparison of Machine Learning Algorithms

In Table 8, we have shown comparison between traditional and proposed model based on parameters such as Time, MAE, RMSE, Precision, Recall, and F1-Score.

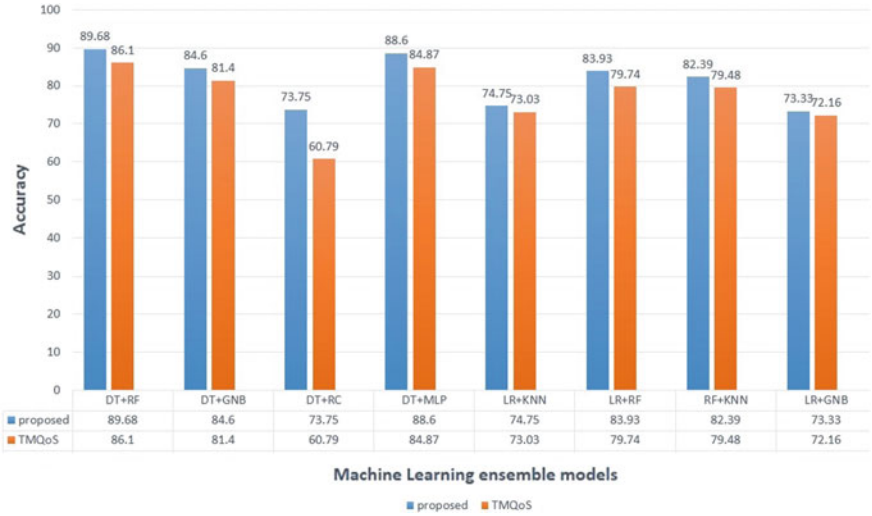


Fig. 12 Efficiency comparison between traditional and proposed method

5.6 Performance Evaluation of Cloud Resource Using Ensemble Machine Learning Algorithms

5.6.1 Classification Accuracy of Ensemble Machine Learning Algorithms for Cloud Resource

In Fig. 12, we have shown comparison of efficiency between traditional and proposed different method using machine learning algorithms. The better result we observed and our proposed method is more efficient than existing method.

5.6.2 Various Performance Parameter of Ensemble Machine Learning Algorithms for Cloud Resource

The comparison of Time, Mean Absolute Error, Root Mean Absolute Error, Precision, Recall, and F1-score between various ensemble models of Machine Learning as shown in Table 9.

5.6.3 Performance Evaluation of Cloud Resource Using Keras Neural Network

In Figs. 13, 14 and 15, we have shown the graph of loss function by applying neural network. In the first graph, we have taken 3 layer input, output and one hidden layer

Table 9 Performance evaluation of ensemble models based on various parameters

ML algorithms/ performance parameters	Time (s)	Precision	Recall	F1-Score	MAE (%)	RMSE (%)
DT+LR	4.32	0.84	0.85	0.84	15.23	39.22
DT+RF	0.23	0.84	0.81	0.80	19.96	47.21
DT+GNB	0.25	0.71	0.74	0.69	29.2	60.1
DT+RC	7.86	0.73	0.73	0.72	27.25	52.75
LR+KNN	6.75	0.84	0.84	0.83	16.13	40.36
LR+RF	37.64	0.75	0.74	0.73	25.50	50.77
LR+SVC	6.92	0.71	0.73	0.71	27.85	55.01
LR+GNB	31.77	0.72	0.73	0.72	27.17	52.53
RF+SVC	1.46	0.79	0.79	0.79	20.70	45.93
RF+KNN	0.54	0.83	0.81	0.80	20.15	47.40
RF+GNB	0.54	0.83	0.80	0.81	20.15	47.40
SVC+KNN	19.71	0.45	0.63	0.51	50.12	87.3
SVC+GNB	31.93	0.74	0.73	0.72	27.85	54.96
KNN+GNB	1.18	0.73	0.72	0.71	29.09	56.26

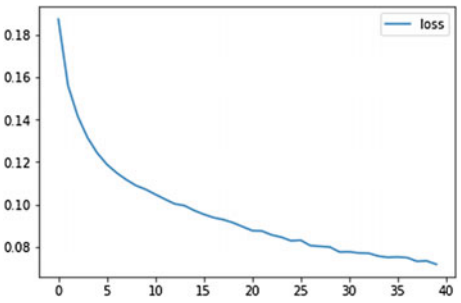


Fig. 13 Model with 40 neurons

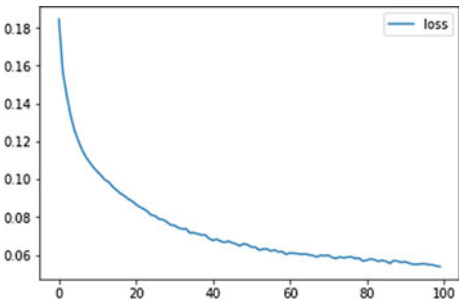


Fig. 14 Model with 50 neurons

with 40 neurons, in second also 3 layer but 50 neurons and finally, in the third graph, we have taken input, output and three hidden layers with 40, 50, and 70 neurons. The efficiency of the respective layers is 98.7, 97.11, and 97.32.

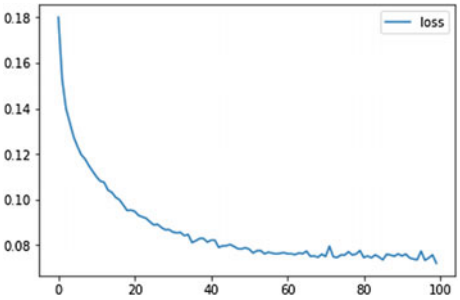


Fig. 15 Model with three hidden layer with 40, 50, and 70 neurons respectively

6 Conclusion

Trust-based access control model is an efficient technique for security in cloud computing systems. We proposed a trust-based access control model using machine learning technique. The main purpose of our model is to give access to an authorized user in the cloud and select the trusted resource for his computation. Both the user and cloud resource are evaluated based on their trust values. We have also categorized users and cloud resources into various classes. Different application gives access based on different level of trust. Our machine learning approach is capable of dealing with a huge number of activity logs within very less time which makes our model really fast. Finally, we compare our proposed model with existing model. We observed that our proposed model perform well as compared to existing model. In future, we can combine our model with the role-based access-control model for cryptography-based-access-control model to achieve more secure cloud environment. We can also consider more security parameters in the cloud by using different techniques and methods. Deep learning models such as RNN and CNN can be used to improve accuracy further.

References

1. Tianfield, H.: Cloud computing architectures. In: 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1394–1399 (2011)
2. Dillon, T., Wu, C., Chang, E.: Cloud computing: issues and challenges. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA). IEEE, pp. 27–33 (2010)
3. Krutz, R. L., Vines, R. D.: Cloud Security: A Comprehensive Guide to Secure Cloud Computing. Wiley Publishing (2010)
4. Gong, C., Liu, J., Zhang, Q., Chen, H., Gong, Z: The characteristics of cloud computing. In: 2010 39th International Conference on Parallel Processing Workshops (ICPPW). IEEE, pp. 275–279 (2010)
5. Jamshidi, P., Ahmad, A., Pahl, C.: Cloud migration research: a systematic review. IEEE Trans. Cloud. Comp. **1**(2), 142–157 (2013)

6. Xiao, Z., Xiao, Y.: Security and privacy in cloud computing. IEEE Commun. Surv. Tutor. **15**(2), 843–859 (2013)
7. Ristenpart, T., Tromer, E., Shacham, H., Savage, S: Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 199–212. ACM (2009)
8. Kandula, S., Katabi, D., Jacob, M., Berger, A.: Botz-4-sale: surviving organized ddos attacks that mimic flash crowds. In: Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation, vol. 2, pp. 287–300. USENIX Association (2005)
9. Yaar, A., Perrig, A., Song, D: Fit: fast internet traceback. In: Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2005, vol. 2. IEEE, pp. 1395–1406 (2005)
10. Ateniese, G., Di Pietro, R., Mancini, L.V., Tsudik, G.: Scalable and efficient provable data possession. In: Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks, p. 9. ACM (2008)
11. Wang, C., Ren, K., Wang, J.: Secure and practical outsourcing of linear programming in cloud computing. In: Proceedings IEEE INFOCOM 2011, pp. 820–828. IEEE (2011)
12. Hamlen, K., Kantarcioglu, M., Khan, L., Thuraisingham, B.: Security issues for cloud computing. In: Optimizing Information Security and Advancing Privacy Assurance: New Technologies, vol. 150 (2012)
13. Takabi, H., Joshi, J.B., Ahn, G.-J.: Security and privacy challenges in cloud computing environments. IEEE Secur. Priv. **8**(6), 24–31 (2010)
14. Bai, Q.-h., Zheng, Y.: Study on the access control model. In: Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), 2011, vol. 1, pp. 830–834. IEEE (2011)
15. Langaliya, C., Aluvalu, R.: Enhancing cloud security through access control models: a survey. Intern. J. Comp. Appl. **112**(7) (2015)
16. Jaeger, T., Prakash, A.: Implementation of a discretionary access control model for script-based systems. In: Proceedings of Eighth IEEE Computer Security Foundations Workshop, 1995, pp. 70–84. IEEE (1995)
17. Kuhn, D.R., Coyne, E.J., Weil, T.R.: Adding attributes to role-based access control. Computer **43**(6), 79–81 (2010)
18. Hur, J., Noh, D.K.: Attribute-based access control with efficient revocation in data outsourcing systems. IEEE Trans. Parallel. Distrib. Syst. **22**(7), 1214–1221 (2011)
19. Lin, G., Wang, D., Bie, Y., Lei, M.: Mtbac: a mutual trust based access control model in cloud computing. China. Commun. **11**(4), 154–162 (2014)
20. Gholami, A., Arani, M.G.: A trust model based on quality of service in cloud computing environment. Intern. J. Data. Theory. Appl. **8**(5), 161–170 (2015)
21. Mell, P., Grance, T., et al.: The NIST definition of cloud computing. https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf (2011)
22. Behera, P.K., Khilar, P.M: A novel trust based access control model for cloud environment. In: Proceedings of the International Conference on Signal, Networks, Computing, and Systems, pp. 285–295. Springer (2017)

Cloud Security Ontology (CSO)

Vaishali Singh and S. K. Pandey

Abstract Research study in cloud computing technology reveals the realization of security importance within its versatile areas. The present security concerns related to issues and challenges have observed a slow cloud computing adoption rate. For enhancing cloud computing security aspects, expertise from various security domains are adopting an ontology based security framework. Lack of visibility in cloud computing system creates numerous cloud security issues, which requires high-level collaboration among the security entities. To sustain collaborative framework security consistency, an innovative approach is required. For this, Cloud Security Ontology (CSO) using Protege software with OWL/XML language is proposed together with OWL-based security ontology, which includes cloud security requirement, cloud vulnerability, cloud threat, cloud risk, control and their relationship with their subclasses. CSO is proposed as a potency of cloud architecture to deal with the challenges related to security goals, favorable realization of security in cloud system, appropriate scheduling and understanding of upcoming threats, risks, vulnerabilities and their possible countermeasures. The proposed CSO was depicted by measuring its strength and totality compared to prior ontologies. Furthermore, a proscribed testing with end-users was performed to estimate its usability.

Keywords Cloud computing · Cloud Security Ontology (CSO)
Security ontology · Cloud architecture · Protégé and OWL · Threats

V. Singh (✉)
Department of Computer Science and Engineering, Jagannath University, Jaipur, India
e-mail: vaishalisingh@stxaviersjaipur.org

V. Singh
Department of Computer Science, St. Xavier's College, Jaipur, India

S. K. Pandey
E-Governance, DeitY, Ministry of Electronics & Information Technology, Government of India, New Delhi, India
e-mail: santo.panday@yahoo.co.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_4

1 Introduction

Computing servers are the centralized standards for data analysing, processing and storing data in all over the world. For more advancement, smart cities and industries are shifting towards Cloud and Internet of Things (IoT) with new computing technologies like fog, edge and mist computing [1, 2]. It has been forecasted that world by 2022 will be utilizing 17.6 billions of (IoT) devices with 9 billion tele-subscriptions, having 90% broadband, eightfold increase of telecommunication traffic while transferring data over the network [3]. Thus, the network paradigm is shifting to cloud and other technologies (fog, edge and mist) for better scalability and fast computing power between the devices and cloud but with security issues [3, 4].

Presently Cloud technology is the succeeding phase in Internet evolution. It has provided a multiplicity of on-demand services along with computing assets or resources (e.g., services, networks, applications servers and storage) to be accessed over the Web through the cloud service end user with minimal service provider interaction [5]. The cloud enhances collaboration, tool and location independence, agility, application programming interface, virtualization, scalability, multi-tenancy, elasticity, availability, and provides prospective for cost diminution [5].

IoT and Cloud has a balancing relationship where large amount of information is generated by IoT and Cloud provides pathway for the information from sender to destination. With the IoT, a large amount of information is been generated but consequently, Cloud based companies have started facing challenges for big data management and related to security issues. New technologies like fog, edge and mist has emerged to overcome the data management problems evolved from Cloud and are complementary to each other [6]. Fog computing is a method used for framing a gateway for gathering and managing all computing capabilities between the sensors and the cloud server in one connection with separate computing power and data storage for multiple sensors [7].

Another application optimizing method is edge computing where a small part of the application is used within its own services and data from many central nodes termed as core to different edges of the Internet. Mist computing supports these device sensors and edges on various network to get sufficient computing power using microcontrollers and microchips embedded on the device.

However, cloud, fog, edge and mist computing all have their own advantages and disadvantages [8]. But focusing on cloud a dark side is faced by smart cities and industries like response time issues, unstable usages in dual offline/online paradigm, underestimated bandwidth capacity, data corpulence, high power consumption, security and privacy challenges. Security is considered as a foremost cause for cloud industries to turn into new computing technologies. The present research study focuses on Cloud future which is considered as a non-foggy domain.

Despite of unlimited benefits in the cloud, consumers are still disinclined to introduce business over cloud system because of considerable barriers to adoption. The most important barriers to adoption are security, critically significant aspects of which have been collected from the available information of diverse

agencies in prior studies [9]. Cloud computing adoptions are vulnerable by indistinguishable security concerns that influence mutually the provider and the user of cloud [9].

Security refers to procedures and standards to provide information assurance. The logical and security physical concerns across the service models (software, platform and infrastructure) and delivery model (public, private or hybrid) are addressed by cloud system security [10]. Cloud security methodology is the holistic approach towards finding security countermeasures towards some service framework, at some cloud user stage [10].

Ontology is one aspect of that holistic approach for analyzing security countermeasures [11]. The undefined security terminology and lack of knowledge and invisibility in levels of domains, assets, and threats of security can be overcome by the improvement of security ontology, used to set up the relationship among the entities [12]. For this, the Cloud Security Ontology (CSO) using Protégé software with OWL/XML language is proposed together with OWL-based security ontology [13].

Aside from this preface on background facts, the following paper is organized viz.: Sect. 2-'Review of Literature'; Sect. 3-'CSO Development'; Sect. 4-'Related Work' of cloud ontology; Sect. 5-'Ontology driven CSO development'; Sect. 6-'Evaluation', Sect. 8-'Conclusion and Future Work' are reported.

2 Review of Literature

The security is concerned as the foremost issues among the cloud stakeholders. Every concern brings miscellaneous effects on distinct assets while examining the issues related to security of Cloud [11, 12]. Even though after a vast study in various dimensions of Cloud, one fails to understand the requirements of security, which results in low adoption rate [11, 12]. A new approach or service is required to better understand the security domain and end-users needs. There are special security measures used. One of them is ontological based approach for security which identify, analyze and elicit the security countermeasures among the entities [11, 12].

But explaining an ontological significance is taken as a complex task for the scientific society and research communities. There are prior research work done to study the security ontologies in special classified domains [11, 12]. The prior study has examined related works of security ontologies for focusing on expanding generalized base for the growth of cloud [11, 12]. The security ontologies that were focused were grouped into three major categories: generalized security ontologies, specific security ontologies and miscellaneous security ontologies [11, 12].

1. The **generalized security ontologies** aimed to cover security features, which had formed explicit domain terminology for dissimilar stakeholders. This category of ontology pays attention on the security development and contribution to knowledge database with general logical perceptive without human intervention

[11]. Some of the generalized security ontologies were *cloud computing security taxonomies* [14], *ontology-based Security* [15] *and ontology-based multi-agent model based on information security system* [16].

2. The **specialized security ontologies** focused on a range of computational models having variables from general terminologies related to security requirements application based security, network, risk and web services etc. These ontologies were alienated into five sub categories with respect to special aspects of security. Some of the specialized security ontologies were [11]:

   (1) **Web Services (WS) and Web Ontology Language (OWL) based Security Ontologies**—(*OWL-based ontology* [17], *Ontological structure for information security domain knowledge* [18], *Security Attack Ontology* [19], *OWL-DL Ontology* [20], *Modeling Enterprise Level Security Ontology* [21]);
   (2) **Network Security Ontologies**—(*Security Taxonomy of Internet Security* [22], *Ontology based Model for Security Assessment* [23], *Ontology-based Unified Problem Solving Method Development Language* (*UPML*) [24], *Security Toolbox*: *Attacks and Countermeasures* (*STAC*) *Ontology* [25], *Network Attack Ontology* [26], *Ontology-based Attack Model* [23]);
   (3) **Security Requirements related Ontologies**—(*Ontologies for Security Requirements* [27], *Extended Ontology for Security Requirements* [28], *Modelling Reusable Security Requirements based Ontology* [29], *Security based Ontology for Adaptive Mapping of Security Standards* [30], *Security and Domain Ontologies for Security Requirements Analysis* [31], *Ontology based Information Security Requirements Engineering* [32]);
   (4) **Risk-based Security Ontologies**—(*Security Ontologies*: *Improving Quantitative Risk Analysis* [33], *SemanticLIFE* [34], *Ontology for Industrial Risk Analysis* [35]);
   (5) **Application based Security Ontologies**—(*Security Ontology to Context-Aware Alert Analysis* [36], *Security Ontology for Mobile Applications* [37], *Security Ontology for Mobile Agents Protection* [38], *NRL* (*Naval Research Laboratory*) *Security Ontology* [39], *Ontology based on e-health applications* [40], *Ontology Based Interoperation Service* (*OBIS*) [41]).

3. **Miscellaneous Security Ontologies** [11]—There are numerous ontologies which cannot be sited in any of the aforementioned categories; thus such types of ontologies are placed in miscellaneous category. Some of the specialized security ontologies were: (*Specification Means Ontology* (*SMO*) [42], *Information Security Measuring Ontology* (*ISMO*) [43], *Security Asset-Vulnerability Ontology* (*SAVO*)-(*Security Attack Ontology* (*SAO*), *Security Defence Ontology* (*SDO*), *Security Algorithm-Standard Ontology* (*SASO*), *Security Function Ontology*(*SFO*)) [44], *Vulnerability-Centric Modeling Ontology* [45], *Cyber Ontology* [46], *Utility Ontologies* [46], *Security Toolbox*: *Attacks and Countermeasures* (*STAC*) *Ontology* [25], *Ontological approach toward cyber security in Cloud Computing* [47], *Ontology in Cloud Computing* [48], *Ontology-based access*

control model: *cloud security policy* [49], *Cloud Ontology* [50], *Security Ontology Driven Multi Agent System Architecture*: *Cloud Data Storage* [51]).

The prior related research have specific and different ontologies for the providers to assess the security, but silently, the end-users of cloud services are facing problems of deficiency in security sphere due to no specification of security attributes like threat risk, vulnerability and countermeasure in cloud [11, 12]. For this reason, Cloud needs to develop a Security Ontology based on all the domain which covers all the attributes of security like risk, vulnerability, threat and countermeasure and which is easy for end user to understand [11, 12].

Cloud Security Ontologies will help in improving the security attributes as a whole by analyzing the concepts and creating taxonomies, of each attributes and finding their interrelations with appropriate countermeasures. CSO will enhance the levels of cloud security for detecting appropriate mitigation technique [11, 12].

3 Cloud Security Ontology Development

The operational framework of entity-relationship (E-R) model of a definite knowledge domain is known as Ontology. To define security terminologies and to remove issues and challenges between the providers and users, one can approach for security ontology framework. These security ontologies have assets, threats, vulnerabilities and mitigation techniques as their basic four security components for standard building block of risk analysis.

Each of these basic four components majorly represents security, through individual ontologies for classification and definition of a specific domain vocabulary of entity and their relationship. The National Institute of Standards and Technology (NIST) Special Publication 800-12 has framed the security relationship model, which is the conceptual structure of developing security ontology.

3.1 Ontology: Definition and Role

Today the semantic Wide Web network, comprises of the Web pages, clips, audio, images, media objects and the objects with organizations, workforce, events and locations. A formal act of representing a collection of concepts and their inter-relationship within a domain is well-known as Ontology. A vocabulary is created to develop the domain that consists of the kind of objects needed to be used, their properties and relationships [52].

Ontology describes instances and objects individually, with the collections of concepts of class set and their attributes like characteristics, aspects, properties, features, restrictions and rules, parameters, relations, function terms, and axioms.

Ontology first acquires domain knowledge through identification and collection of expert views [52]. Then concept structures need to be defined all along among the domain allied properties and relationships. When ontology is developed it is verified, ensured and finally committed within its planned location [52].

To encode ontology different languages are used. For e.g.

- DARPA Agent Markup Language (DAML)
- Semantic Web Rule Language (SWRL)
- Ontology Inference Layer
- Web Ontology Language (OWL)
- Ontology Interchange Language (OWL).

Ontology editors are considered to create and manipulate the applications [53]. These editors articulate ontologies in one of many ontology languages.

Some of the editors are as follows [53].

- Protégé (free, open source ontology editor)
- Onto Edit
- Knowledge acquisition system
- DERI Ontology Management Environment (DOME)
- RDF knowledge bases
- Knoodl (Community-Oriented Development of OWL-based ontologies).

3.2 CSO Architecture: Definition and Role

Researchers have investigated the state-of-the-art security domain ontologies in prior surveys. Several ontology prototypes and their security approach and relationships within a domain are resultant from security standards with basic concepts such as attacks or threats, vulnerabilities prospects [54]. CSO has gathered a large number of related terms as shown in Fig. 1.

CSO defines relationships between the following concepts:

- Cloud Security Threats
- Cloud Assets
- Cloud Risk
- Cloud Vulnerability
- Cloud Security Requirements
- Control.

The main subclasses are Cloud Security Threats and Cloud Security Requirements used to state Control. The CSO depicts that the Assets *belongto* resources, which *useSW* and *runOn* Hardware and Software. Cloud Security Threats *threaten* Assets, which *hasorigin* Threat Origin and *hassource* Threat Source. Cloud Security Threats *haveprobability* of Cloud Risk and are *exploitedby* Cloud Vulnerability. Cloud Vulnerability *hasseverity* Severity Measures and are *vulnerableon* Assets. Thus Assets

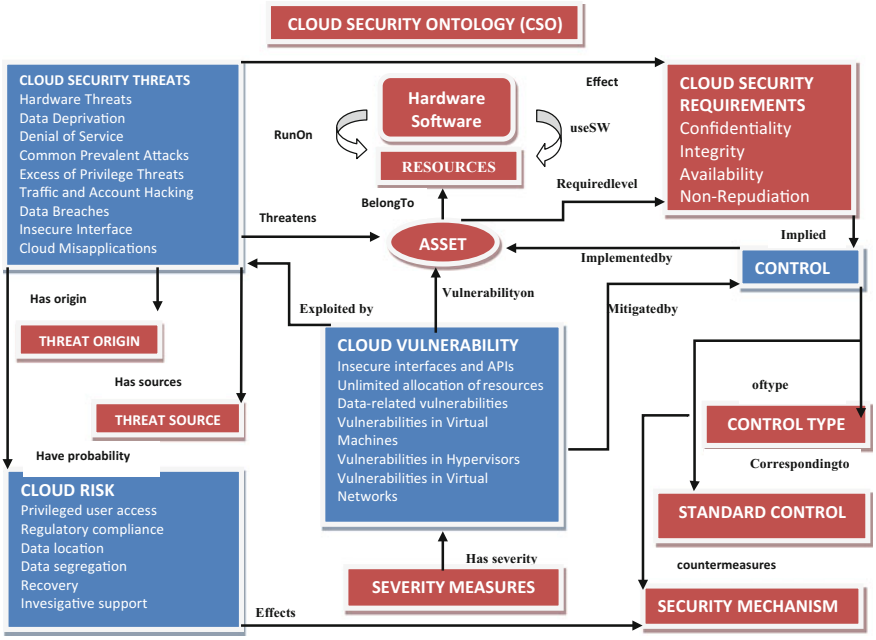


Fig. 1 CSO architecture

require Cloud Security Requirements *implied* Control. Cloud Security Threats *effect* Cloud Security Requirements directly. Control is *oftype* Control Type *correspondsto* Standard Control having *countermeansures* Security Mechanisms. Cloud Risk *effects* Security Mechanisms. Cloud Vulnerability is *mitigatedby* Control and *implementedon* Assets.

3.3 Security Requirements: Cloud Computing

3.3.1 Confidentiality Integrity and Availability (CIA)

A diligent effort on security requirements offered by the cloud providers is partially justified with the facts that cloud is secure [55]. The main pillars of security aspects are Confidentiality, Integrity, and Availability. In an organization the expertise team works on processes such as software installation, data analysis, database creation, transportation and access mechanism based on CIA criteria [55].

- ***Confidentiality***: Multi layered approaches are adopted by cloud providers through security such as encryption, role-based user access control, certificates and intrusion detection systems etc. Large indefinite numbers of confidentiality threats are found in multi-tenancy and multitasking in cloud [56]. Cloud data confidentiality

is associated with authentication of users. Thefts create problems while accessing control mechanism while establishing the user's individuality in the information system. Hence, confidentiality is a key feature in data security to maintain and manage personal data of users. Cloud providers are liable for adopting protected method for ensuring confidentiality [56]. Cloud is associated to various legal challenges due to data storage at different locations, which leads to risk of confidentiality loss. Cloud providers are dealing with personal facts and figures of the clients should ensure important confidentiality protection [56].

- ***Integrity***: Quality and accuracy are inherent necessities of users, assets, processes and organizations. In information technology, assets are processed and transferred to various remote area/s. Therefore, the processed data requests to be operated by the authorized party in an authorized manner [57]. The organization needs to eliminate the inappropriate utilization of information and services and requires reducing the risk of loss using integral security mechanism. Integrity provides enhanced transparency for depth and strength of control. This in depth analysis exposes and identifies the entity that has modified the assets and is potentially harming the integrity [57]. Authorization is one mechanism, which ensures access control of users and resources of the organization. The cloud environment must provide specified entities that are authorized to interact with the assets. Cloud providers need to make certain about data integrity and accuracy of assets in the cloud system [57].
- ***Availability***: Availability gets pretentious by technical issues such as not proper functioning of computer and communication device, accidental or deliberate natural phenomena [58]. Availability is confirmed by severely maintaining and repairing hardware when essential, providing a definite assess of non-redundancy and failover, ample communication bandwidth and preventing the incidence of bottlenecks [58]. The implementation of availability ensures backup power systems emergency, updated system upgrades and security against malicious practices [58].

3.4 Non-repudiation (NR)

Non-repudiation is another security requirement in the present scenario. In case of entities, disputes and network faults while exchanging confidential information among sender and receiver, secure point to point communication is required. But applications in existent world have multiple entities [59]. Non-repudiation ensures acceptance of responsibility of submitting or receiving the message. Non-repudiation can be established by digital signature, message transfer agents, timestamps and protocols used in data transfer [59]. For developing a secure cloud environment it's a present necessitate for understanding the various attributes of non-repudiation that affects cloud and have derivative from several area/s wherever it is well-practiced [59]. Some are depicted in Fig. 2.

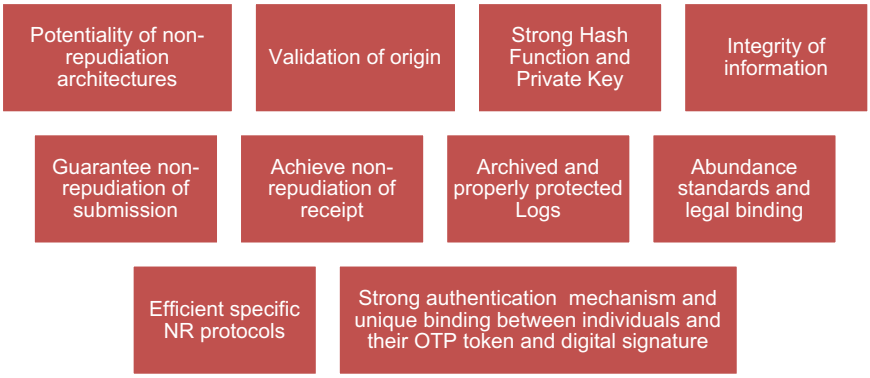


Fig. 2 Non-repudiation attributes

4 Related Work

Prior research studies have been conducted in literature on security related ontology. To develop ontology for security, it is valuable to realize the significance of ontology and some prior security ontology works. Out of that there is security ontology of CDS (Cloud Data Storage security) was based on OWL [60]. This was based on security objectives, frameworks, subsystems components, assets, risks, vulnerabilities, threats and their relations in cloud [60]. This ontology was proposed as Multi-Agent Systems Architecture by Protégé software. It focused on the concerns related to the challenges of security goals (confidentiality, availability, and integrity and correctness assurance) for ensuring the security concerns of CDS [60]. The interior parts of security concept were depending on the instances and subclasses, which provide the domain vocabulary of security. This ontology was based on five main steps:

- Domain (particular environment),
- Purpose (anticipated outcome that is intended),
- Scope setting (state of the environment where a situation exists),
- Imperative expressions acquisition, classes and class hierarchy conceptualization,
- Creation of instances.

This ontology created a prototype model using the Protégé developed on the 3 main steps and was tested to determine the efficiency of security mechanism.

5 Ontology Driven (CSO) Development

The proposed CSO explains the inter-relationship among Cloud Security Threats, Cloud Assets, Cloud Risk, Cloud Vulnerability, Cloud Security Requirements and Control. It explains how the precautionary components are positioned and how they

communicate with security architecture. These components help to maintain the security attributes (Non-repudiation (NR) and Confidentiality, Integrity and Availability (CIA)).During development of ontology certain steps are performed:

5.1 Determination of the Domain and Scope of the CSO

Scope and domain are the majority vital factors in ontology creation. Ontology sets words used for experts sharing information in a domain. Ontology creation needs comprehension of ontology related queries [61, 62]. The main objectives being evaluate, analyze, select, and categorize security ontologies, as a scale study together with security requirement attributes. Ontology-design process determines basic concepts such as domain to be covered in the respective ontology, the place where it will be used and the user maintaining the ontology [61, 62]. The ontology is used by service customer and cloud provider to recognize the requirements of security and their countermeasures. Security is a problem dealing with comprehension of the domain and systems operation [61, 62]. Experts' evaluation has concluded the deficiency of specificity of the types of attacks and threats with the existing domain problem. This can be addressed by proper metaphors of classes and concepts with their relationships by sighting the security needs for a detailed domain [61, 62].

5.2 Consider Reusing Accessible Ontologies

As discussed in the above section, provisions from the security database information, security data and a distinct information security data file, discrete standard security library have been applied in the CSO. The efficiency of CSO lies with the fine points on the associations between user's under studied ontologies and the Service Provider's in cloud sub-ontologies. These linkages among the Service Provider's are traversed and specific actions triggered.

5.3 Enumerate Imperative Keywords of Ontology

Useful Keywords of CSO development are the security nouns explaining it's domain area, cloud security classes and subclasses.

5.4 *Define the Properties of Classes*

All imperative terms and conditions from protégé principle are listed and then conceptualized into an abstract or general idea and relations surrounded by concepts to describe associated classes and their hierarchy. Building OWL CSO ontology is iterative process.

1. WebProtégé—ontology development domain for the Internet users [63]. It makes it simple to develop, upload, transform and contribute to ontologies for viewing and editing. This completely supports the recent OWL 2 Web Ontology Language [63]. The extremely configurable user interface develops the ideal environment for beginners and experts. The most significant factors include sharing and permissions, threaded notes and discussions and email notifications [63]. This document explains and specifies XML presentation syntax for OWL, which is definite as a dialect comparable to OWL Abstract Syntax [63].
2. Classes, properties and individuals are designed by an editor. The editor "Protege-OWL" is available on its web page for download [64]. Protege is a developed using Java program, for that reason it requests installation of 1.5 version or advance version of Java Runtime Environment to run [64].
3. Next, a reasoner is implemented to make sure whether the expected ontology was designed and its assertions are reliable [65].
4. For better perceptive of the ontology, visualization tools are used [66].

    a. OWLViz is one the visualization tool, which provides ontological class hierarchies in web language for envisioning and navigation, permitting evaluation of the assert and inferred class pecking order [67]. OWLViz saves mutually the asserted perspective and inferred views. These views are existing graphics formats like JPG, PNG and SVG [67]. As shown in Fig. 3.
    b. Another visualization plug-in and graphical tool is OntoGraf. This has develop into a standard measurement of the Protege-OWL editor [68]. As shown in Fig. 4.

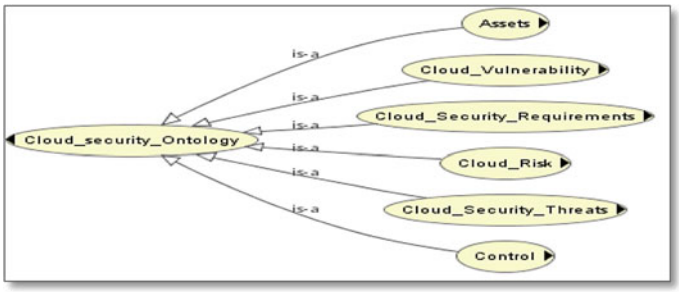


Fig. 3 Visualization (OWLViz) of CSO

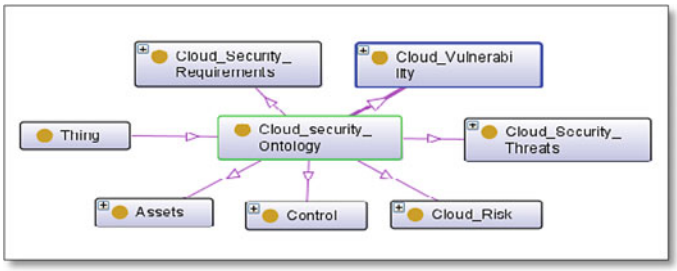


Fig. 4 Visualization (OntoGraf) of CSO

5.5 Building the CSO

This segment will explain how to assemble a easy CSO to get comfortable with Protege-OWL editor [54].

Define the classes and hierarchy associated with individual groups
Cloud security term is related with the high class security properties that depict the outline of cloud issues. Formally, ontology is an interlinked concepts within hierarchy of classes related with its features. Given Fig. 3 Visualization (OWLViz) of CSO and Fig. 4 Visualization (OntoGraf) of CSO spotlights on the most important classes associated with CSO domain and the interlinks surrounded by them.

Four main classes have been defined in this ontology. Editor's user interface has many multiple tabs that contains several views and customized layout. Some predefined protégé tabs are functional in developing ontologies. The instance of two protégé tabs are:

- **Classes tab** is required to develop class hierarchy [69, 70].
- **Object Properties tab** is required to create object properties and allocate security domains and its ranges [69, 70].

The OWL demonstration of these classes is as follows:

A. **Cloud Assets class**: This class represents the interdependency of assets in cloud as made known in figure. Assets can be stated as organization's data, software (S/w) and hardware (H/w) that are used in cloud system activities [71]. As shown in Fig. 5.
B. **Cloud Vulnerability class**: This class represents the weakness as vulnerabilities of cloud. The probability in which, assets cannot refuse to comply with the action of a threat representative is identified as vulnerability [72]. As shown in Fig. 6.

Five key factors of cloud limitation: [73].

- ***Performance***—There was a latest trouble with Twitter ("Fail Whale") and Steve Jobs' awkwardness at the set of network connections outage at the

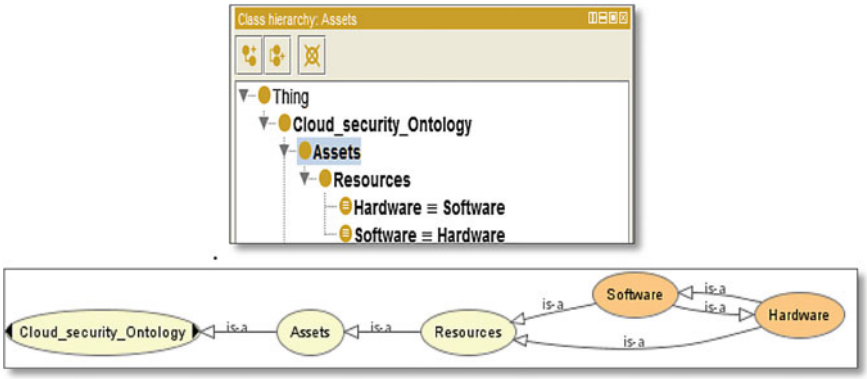


Fig. 5 Assets class of CSO (OWLViz)

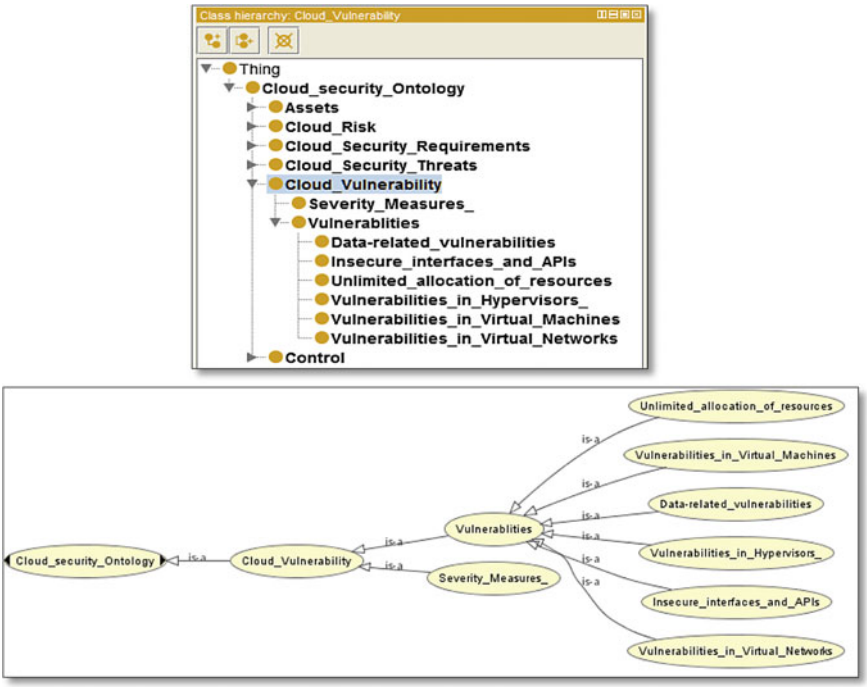


Fig. 6 Cloud vulnerability class of CSO

preface of the innovative iPhone don't precisely express fuzzy approach concerning the Web (Internet) and network performance while dealing with cloud services [73].

- ***Return-On-Investment***—The purpose of short term Return-On-Investment has been driven from the initial stages of cloud computing [73].

- ***Market churn***—Inevitably the introduction of dot-com in cloud market may lead to crash [73].
- ***Privacy***—Security and privacy both are like siblings but both are very separate issues and challenges [73]. Advocates have recently woken up in cloud computing business and accountability for information is a foremost one [73].
- ***Security***—Security is still a major issue in cloud technology. There is no sense in signing the contract if vendor's security architecture is not known. Necessary awareness of risks and vulnerabilities are required for both the customer and service provider.

C. **Cloud Security Requirements class**: This class focuses on the security requirements of cloud. The most researched key factors of security requirements are: Non-repudiation (NR) and Confidentiality, Integrity and Availability (CIA) [74]. A cloud security need explains that should be focused in use by cloud customers to manage and evaluate the cloud security of their environment with the goal specific aim of modifying risks and deploying a suitable stage of support [74]. As shown in Fig. 7.

Key stages follows [74]:

- Make certain effective processes under governance.
- Audit operational.
- Business processes.
- Manage identities, people and roles.

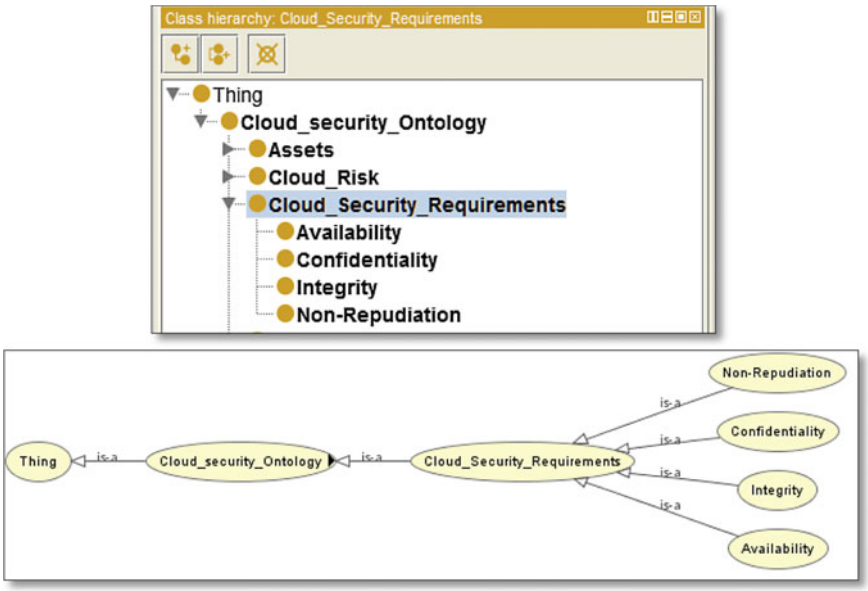


Fig. 7 Cloud vulnerability class of CSO

- Make certain protection of information.
- Implement privacy strategies.
- Evaluate cloud security requirements.
- Guarantee inter-networking connections in a secured cloud.
- Estimate security control within hardware infrastructure.
- Recognize the security requirements.

D. **Cloud Risk class**: This class focuses on the risks component of cloud where the assets vulnerabilities are exploited by the threats, thereby causing destruction to a company [75, 76]. Virtualization holds a huge level of threats posed by the hardware machines with their unique characteristic of exploiting the target on the main virtual server and the guest on the virtual server [75, 76]. As shown in Fig. 8.

Authorization, Authentication, and Access Control (AAA): mechanism is critical, but a lot depends on process as well [75, 76]. For example: Authentication to develop single-sign-on (SSO) on using shared name spaces provides a better productivity but is more susceptible to risks [75, 76]. Availability: Risk like redundancy and fault tolerance is experienced by the customers of the public cloud. Each service claims

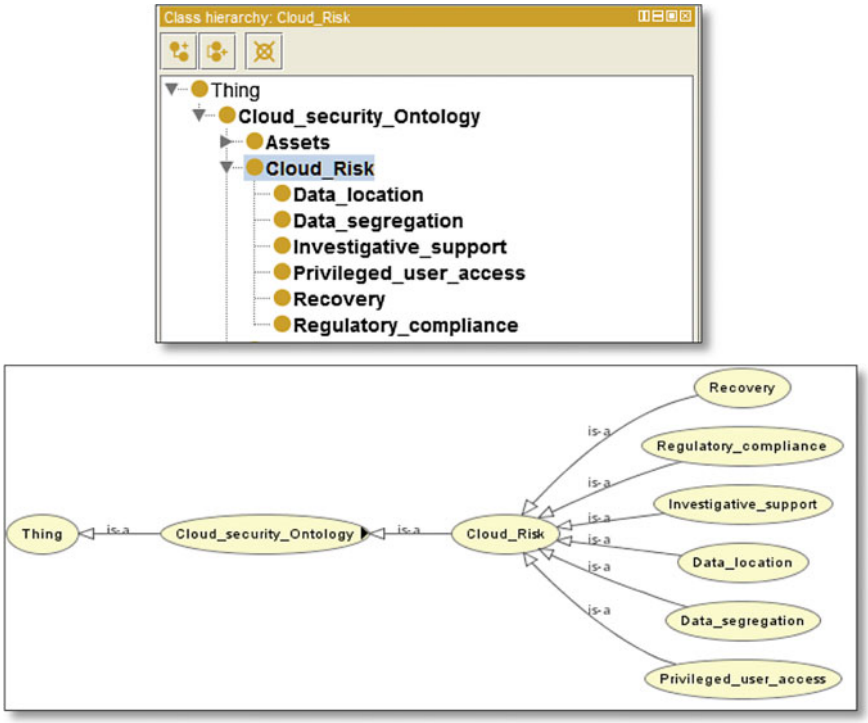


Fig. 8 Cloud risk class of CSO

for the availability and unique fault tolerance [75, 76]. Ownership: Cloud customers experience this type of risk where multiple owners of the data and contracts are explicitly stated at different data storages for only providers [75, 76].

E. **Cloud Security Threats class**: This class reflects the threat components to cloud security [10]. As shown in Fig. 9.

Subclass of cloud threats represents the identified and classified threat categories [10]. In the previous paper, we provided an indication of the major security threats under various categories, which may serve as a primary step towards a development of CSO ontology in the relevant area/s of threats [10]. As shown in Fig. 10.

To make navigation interactive between the OWL ontology, OntoGraf visualization tool is used that contains various layouts automatically organizing the configuration of ontology. Different interrelationships are maintained by subclass, object, domain/range object properties, and equivalence. Node types and relationships are classified to assist to create the view as desired.

F. **Control class**: A countermeasure is a method that is functional to prevent and reduce potential threats to operating systems, servers, networks, and information systems [77]. Furthermost divided into following subclasses as shown in Fig. 11.

- Standard_Control—Various state security's standards numerous information security standard supports good practices and well defined frame works in proper analyzed structure for managing system design of information security controls.
- Control_Type—An organizational concept used to illustrate the category of control and the nature of the tools and techniques associated with its implementation. Control types include technical, policy, and procedural.
- Security_Mechanisms—This is a procedural method used to form a tool for enforcing security strategies and polices. Cloud Security provides services for implementation of these policies through security mechanisms.

6. Object properties are the hierarchical display which exactly shows the controlling units as same as in class hierarchy views. Figure shows Object properties. As shown in Fig. 12.

5.6 OWL-DL

OWL-DL presents the superclass-subclass relationships which can be processed by designing and developing in a reasoner. The DL Query tab provides influential and user-friendly feature for penetrating a categorized ontology. The query language assisted by the plugin is deployed on the Manchester OWL syntax, an accessible language rules for OWL DL that is established on accumulating all data asset with reference to a rigorous property and class into a solitary assemble, called a frame. As shown in Fig. 13.

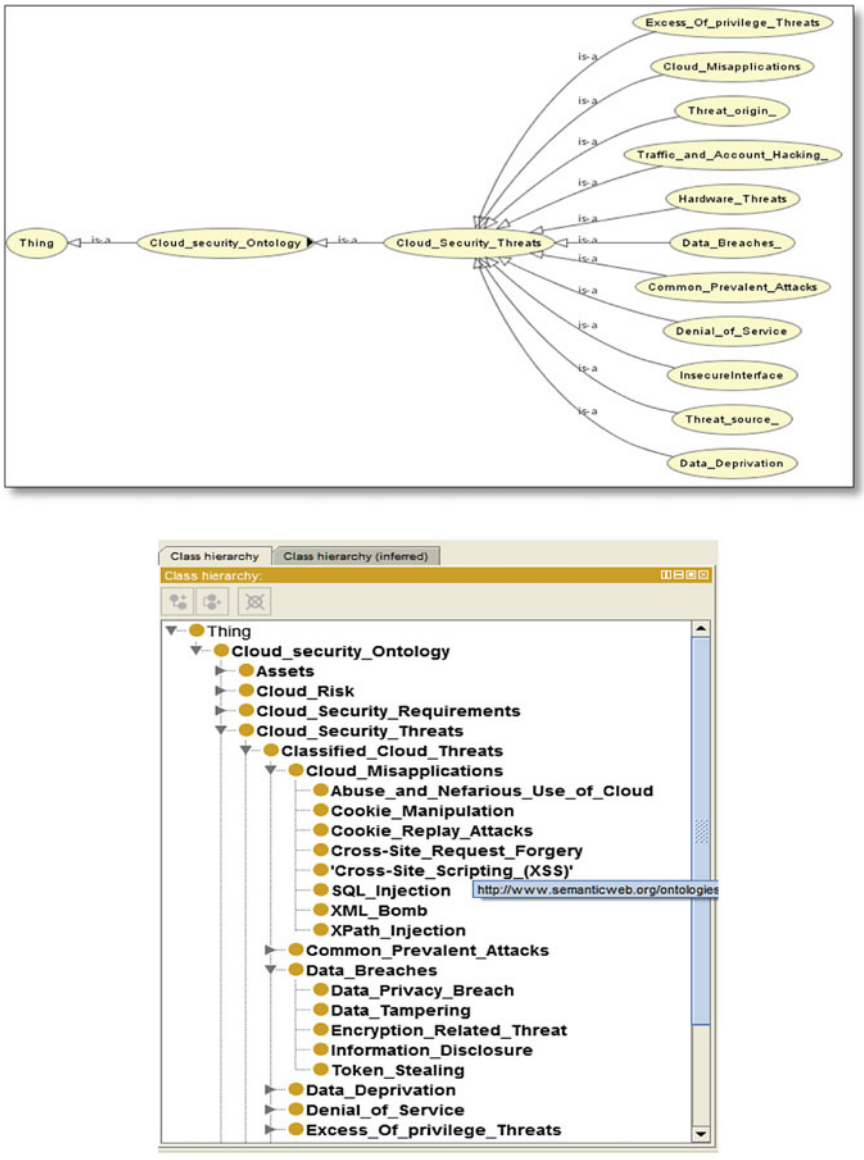


Fig. 9 Cloud threats class of CSO

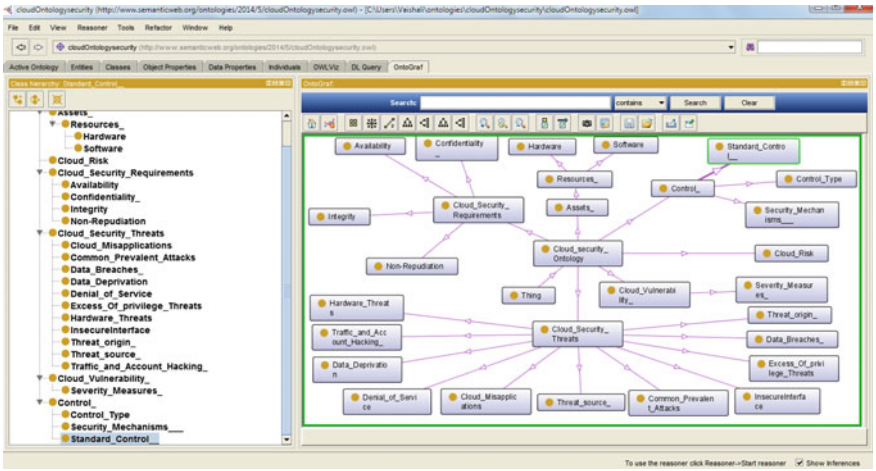


Fig. 10 Cloud threats class of CSO using OntoGraf

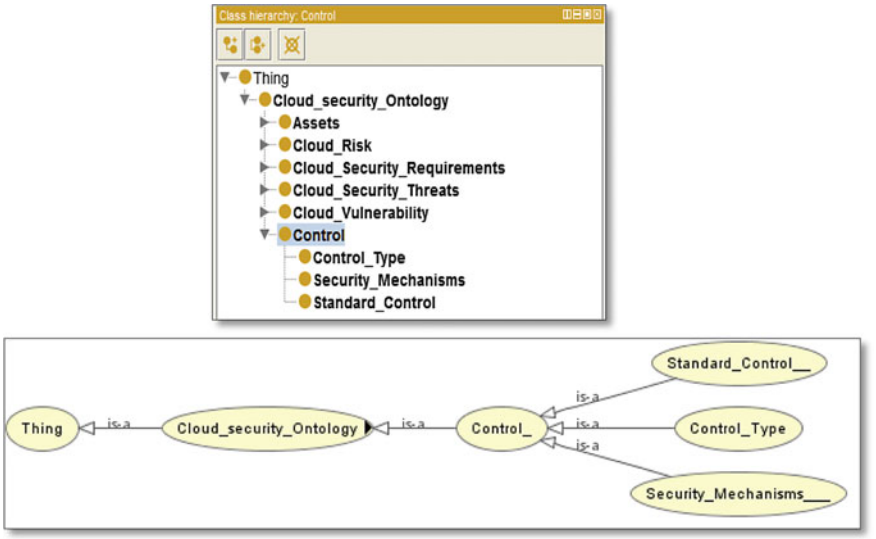


Fig. 11 Cloud control class of CSO

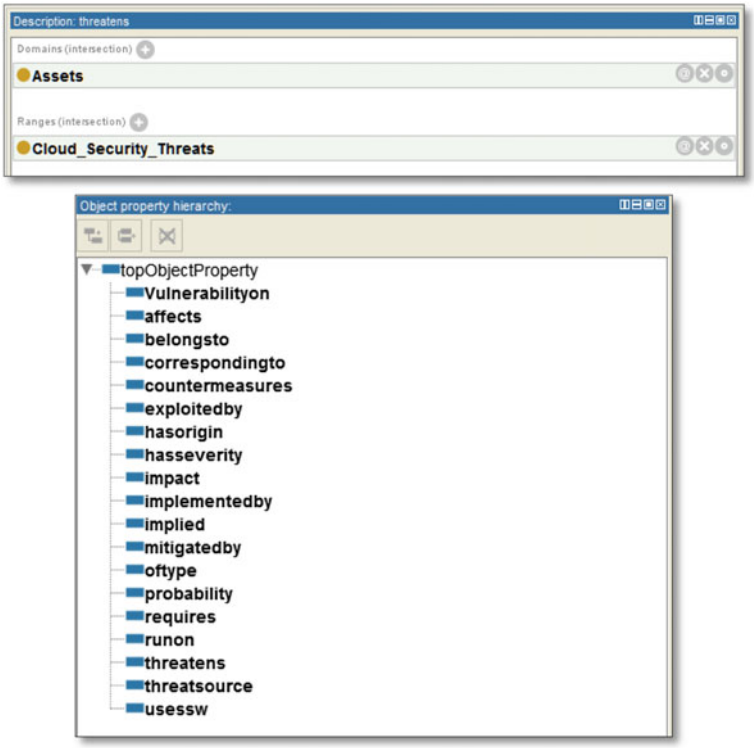


Fig. 12 Object property hierarchy of CSO

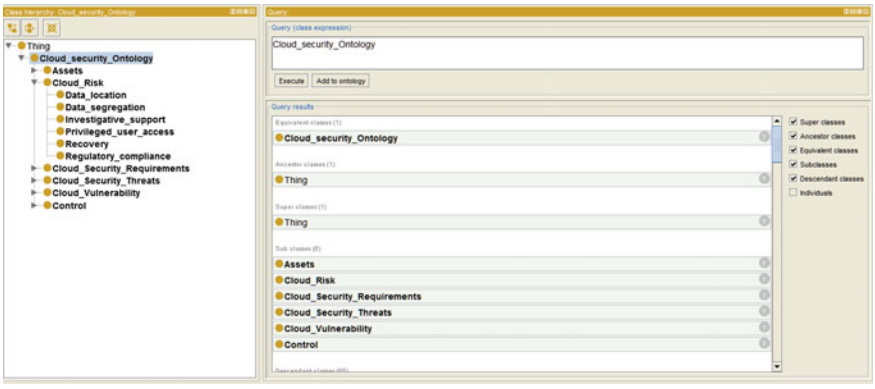


Fig. 13 OWL-DL query of CSO

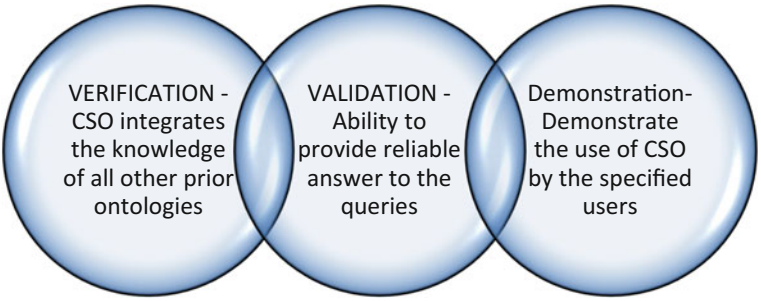


Fig. 14 Criterion to estimate the CSO

6 Evaluation

The aim is to develop the CSO having high level concept and inter-relationships for increasing the reusability of countermeasures and security requirements.CSO has been evaluated on the following criteria as shown in Fig. 14.

6.1 Verification

The first criteria verifies that the CSO is more complete in knowledge of elements and their relationships than the prior literature covered in related work. To accomplish this task, the previous research focused on "Revisiting Security Ontologies", which is a significant revision of security ontologies. The learning of these accessible security ontologies has tried to examine 'how every attribute of assets, threats, security objectives, vulnerabilities and countermeasures are enclosed within the aspects of ontology' [11].

In addition, the study has confirmed whether the projected security ontologies can be used for defining the CSO through the decisive results [11]. Later on, comparative study of Cloud Security Ontologies through the research has presented a concise discussion on few major ontologies [12]. A relative study was also able to use many attributes that were notorious based on the well-known practices with comparable studies in security area [12]. Applying all the weaknesses and strengths of each prior research study covered in the literature, results to demonstrate that the CSO is a complete set of concepts and its relationship with respect to all other security ontologies.

6.2 Validation

The CSO undergoes the validation process using a set of questions and resulting to reliable answers using its terminology for specific domain manually using instantiating the concept of the core ontology.

For which the DL query language is used based on class-object expressions supported by Manchester OWL syntax and built-in reasoner or classifier (HermiT), to execute the active ontology. The section list a set of questions that a security expert is going to handle during the requirement phase of the CSO. Each question deal with the DL Queries, which are more powerful feature for searching the categories according to shared characteristics in ontology.

For instance the set of questions as follows:

a. **Asset Identification analysis**

- Classify the Assets?

Example of a DL query to find the super classes, Ancestor classes, Equivalent class, SubClasses. As shown in Fig. 15.

b. **Cloud security Threat analysis**

- What subclass threat belongs to which super class of Threat? As shown in Fig. 16.

c. **What are the key factors of security requirements in Cloud Computing?** As shown in Fig. 17.

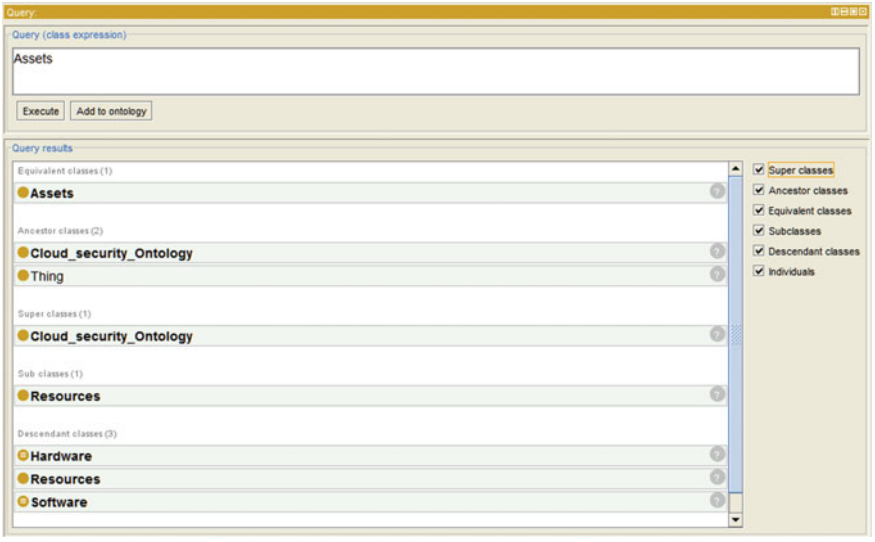


Fig. 15 Classify the assets

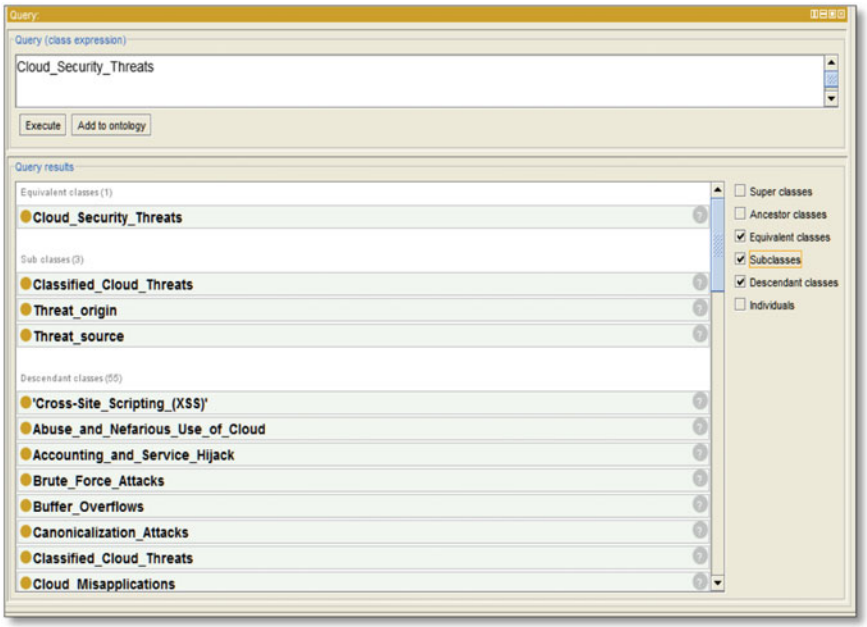


Fig. 16 Cloud security threat analysis

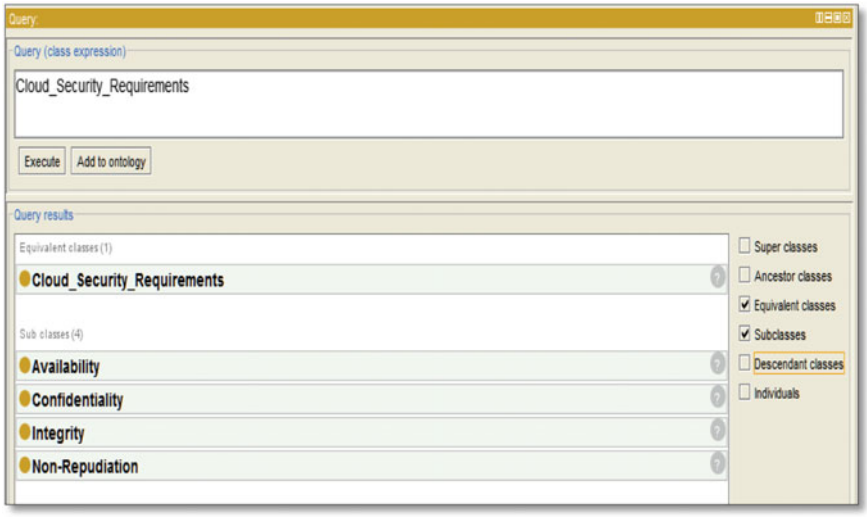


Fig. 17 Key factors of security requirements

This section has demonstrated how the CSO could increase the awareness of the client in regards to the cloud security requirements in each domain (threat, risk,

countermeasures). The practice of gathering the requirements of a system for users and customers will bring the valuable knowledge and ideas to the SRE (Security Requirement Engineering).

6.3 Demonstration

To demonstrate the usefulness of the CSO, a set of tests and experiments were applied on users. The design and analysis method was used to proceed with the procedure of the test. A group of respondents from academics, research organizations and industries were contacted through calls and emails (LinkedIn, Research Institute, Associations, Industrialists, Laboratories etc.) on the basis of their profile and job position. The test presented CSO Architecture with its concept and inter-relations. The architecture was demonstrated using protégé software. A meeting was held to deal with CSO manipulation, skillfully or efficiently by the respondents. Test ended with the respondents by filling a questions set. The results were summed up on the basis of questions set in graphical tabular form. As shown in Fig. 18.

The degree of agreement was set under a scale (1–5) on the basis of questionnaire. (5-Completely Agree, 4-Agree, 3-Neither agree nor disagree, 2-Disagree, 1-Completely disagree).

The respondents were asked to evaluate the CSO under the scale 1–5 through these 6 questions in the area of Key Factors:

I. ***Key Factor 1***: ***Determine concept and inter-relationship***

(a) How strong is the concept and inter-relationship of the CSO?
(b) Does the CSO increase the awareness in the area of security domain?

II. ***Key Factor 2***: ***New Components Discovery***

(c) Does the CSO helpful in discovering new components?

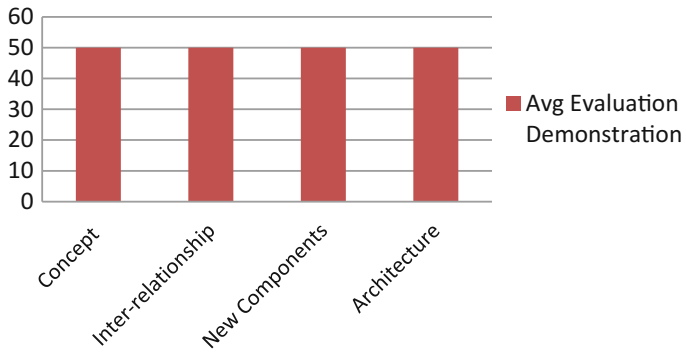


Fig. 18 Demonstration

(d) Does the CSO used for a specific domain area?

III. ***Key Factor 3***: ***Demonstrate Architecture*** (***Conceptual Structure***)

(a) Do you find the architecture of the CSO easy to use and understand?
(b) Does the CSO helps in creating a model for an organization?

7 Evaluated Result

After going through the feedback from different experts, it was observed that majority of respondents admired the interactive environment of CSO. Most of the respondents had found that the CSO entails the main key concepts after undergoing and determining the concept and inter-relationship of CSO with the new components discovery feedback resulted that CSO has helped in discovering new innovative elements for the experts working in security domain area, in view of the fact that it is difficult to bear in brainpower hundreds of risks, threats, vulnerabilities, and security requirements specifications with their countermeasures.

Among the optimistic qualitative feedback outputs of the questionnaire that were provided by respondents had mentioned: "*The CSO justifies the relationship between cloud security risks*, *threats*, *vulnerabilities*". One respondent mentioned that: "*The researchers working in ontological domain area will be definitely benefited by this ontology and more measures will be undertaken to increase the robustness of cloud security*". That was a motivating tip that can be improved in the upcoming by providing new method to modernize automatically the experts of the security ontology.

The next sequence of questions were mainly dedicated to the subsequent levels of the research development and their answers represent an imperative contribution for future work. It was presented to respondents who did not know it before. One respondent mentioned that "*This ontology has laid emphasis on Non-Repudiation which is also a very important factor in cloud Security Requirement besides CIA*". A common answer was: "*Yes this ontology is domain specific but should be adaptable to the emerging new cloud security threats*." The conversation with respondents that followed the prior questionnaire shows that, although the CSO has the main concepts, and inter-relations, this is still not sufficient for consumers to construct cloud security framework with it. More strategies are required, not for the security ontology itself but also for the procedure of using it for threats necessities.

One participant mentioned, "*Furthermore key concepts are required for the applications to diverse specific domains*". "*This ontology can be used as a base for developing a framework for implementing cloud security issues with the help of use cases*". The ontology can be used in diverse application with a number of additional associations with domain experts, consulting documents and records. On the present research study, new methods are in process to formulate the procedure in routine by means of the core CSO with diverse spheres ontologies.

8 Conclusion and Future Work

This comprehensive study for sustaining collaborative framework security consistency is based on the aspects of CSO using Protégé software with OWL/XML language is proposed together with OWL-based security ontology, which includes cloud security requirements, cloud security threats, cloud vulnerability, cloud risk, control and their relationship. Various Security requirements such as Confidentiality, Integrity, Availability (CIA) and Non-repudiation (NR) for finding future countermeasures have also been highlighted. CSO is purposely proposed to present essential security ethics for guidance to vendors and assistance to customers in analyzing the entire security threats of a provider. For a holistic tackling of security aspects in the cloud architecture, CSO is imperative.

In spite of collaborative frameworks, there is an imperative necessitate to put effort additionally in the security area/s to approach with the new ideas associated to the innovative counter measures. Thus, the future work aims to the center of attention on the classified recurring threats and finding their possible security requirements with their countermeasures, which will be inculcated in the CSO for further research to standardize security prospects, cloud nomenclature and terms, with security measures implemented. Future of Cloud will never be foggy environment, but still now more organizations are approaching to Internet of Things [78, 79]. Except security issues of individual threats, adoption of additional capabilities has imparted more limitations and drawbacks into cloud. Present cloud architecture does not fit the requirements of projects due to the issues of centralized nature. The network central nodes have only the right to store and process data.

The future study can be on developing new architectures for distributing computing power consistently around the network using techniques like fog computing, mist computing and edge computing. These computing technologies are extensive adaptation of cloud computing [80]. But like cloud, the fog, mist and edge computing are also more popular for their numerous issues related to risks and vulnerabilities [81]. The issues and challenges need to be resolved at the initial stage in order to provide secure platform for the users. The countermeasures for environment and network issues are still not appropriate for these new computing technologies as these are still in the evolving phase to provide security like cloud. The improvements in cloud computing will help us to implement the IoT using these all new technologies (fog computing, mist computing and edge computing) [82]. The present smart cities and industries needs fast methods for organizing and managing their heterogeneous resources. Future work can be on developing different types of security ontologies for these methods used in different computing technologies (Cloud Computing, fog computing, mist computing and edge computing) [82].

References

1. Malik, Y.: Internet of Things Bringing Fog, Edge & Mist Computing. https://medium.com/@YogeshMalik/fogcomputing-edge-computing-mist-computing-cloudcomputing-fluid-computing-ed965617d8f3. 21 Sept 2017
2. M destrian, Fog, edge, cloud and mist computing. https://intellinium.io/fog-edge-cloud-and-mist-computing/, 25 Jan 2017
3. Mahesa, R.: How cloud, fog, and mist computing can work together. https://developer.ibm.com/dwblog/2018/cloud-fog-mist-edge-computing-iot/
4. Martin, M.J.: Cloud, Fog, and now, Mist Computing. https://www.linkedin.com/pulse/cloud-computing-fog-now-mist-martin-ma-mba-med-gdm-scpm-pmp, 24 Dec 2015
5. Singh, V., Pandey, S.K.: Research in cloud security: problems and prospects. Int. J. Comput. Sci. Eng. Inf. Tech. Res. (IJCSEITR) **3**(3), 305–314 (2013)
6. Mohan, N., Kangasharju, J.: Edge-Fog Cloud: A Distributed Cloud for Internet of Things Computations. https://arxiv.org/pdf/1702.06335.pdf, 7 Mar 2017
7. Barika, R.K., Dubeyb, A.C., Tripathic, A., Pratikd, T., Sasane, S., Lenkad, R.K., Dubey, H., Mankodiya, K., Kumar, V.: Mist data: leveraging mist computing for secure and scalable architecture for smart and connected health. Sci. Dir. Procedia Comput. Sci. **125**, 647–653 (2018)
8. Barik, R.K., et al.: Fog assisted cloud computing in era of big data and internet-of-things: systems, architectures, and applications. In: Mishra, B., Das, H., Dehuri, S., Jagadev, A. (eds.) Cloud Computing for Optimization: Foundations, Applications, and Challenges. Studies in Big Data, vol. 39. Springer, Cham (2018)
9. Singh, V., Pandey, S.K.: Revisiting cloud security issues and challenges. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(7), 1–10 (2013)
10. Singh, V., Pandey, S.K.: Cloud security related threats. Int. J. Sci. Eng. Res. **4**(9), 2571 (2013)
11. Singh, V., Pandey, S.K.: Revisiting security ontologies. Int. J. Comput. Sci. **11**(6), 150–159 (2014)
12. Singh, V., Pandey, S.K.: A comparative study of cloud security ontologies. In: 2014 IEEE 3rd International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), pp. 797–803 (2014)
13. De Novoa, S.R.: PROTEGE. https://cwi.unik.no/images/e/ef/UNIK4710-Protege_Presentation.pdf
14. Srinivasan, M.K., Sarukesi, K., Rodrigues, P., SaiManoj, M., Revathy, P.: State-of-the-art Cloud Computing security taxonomies: a classification of security challenges in the present Cloud Computing environment. http://dl.acm.org/citation.cfm?id=2345474. Accessed 11 Feb 2014
15. Tsoumas, B., Gritzalis, D.: Towards an ontology based security management. In: AINA 2TH International Conference Advanced Information Networking Applications, vol. 1, pp. 985–992 (2006)
16. Gorodetski, I., Popyack, L.J., Kotenko, I.V., Skormin, V.A.: Ontology-based multi-agent model of an information security system. In: Proceedings of 7th International Workshop. RSFD-GrC'99, 9–11 Nov 1999, pp. 528–532
17. Herzog, A., Shahmehri, N., Duma, C.: An ontology of information security. Int. J. Inf. Secur. Priv. **1**(4), 23 (2007)
18. Fenz, S., Ekelhart, A.: Formalizing information security knowledge. In: 4th International Symposium on Information Computer and Communications Security (ASIACCS ″09), pp. 183–194 (2009)
19. Vorobiev, A., Han, J.: Security attack ontology for web services. In: Second International Conference on Semantics Knowledge and Grid, p. 42 (2006)
20. Denker, G.L., Kagal, T.: Finin.: security in the semantic web using OWL. Inf. Secur. Techn. Report **10**(1), 51–58 (2005)
21. Fenz, S.: Ontology-based generation of IT security metrics. In: SAC ″10 Proceeding of the 2010 ACM Symposium on Applied Computing, pp. 1833–1839

22. Abbas, A., El Saddik, A., Miri, A.: A state of the art security taxonomy of internet security: threats and countermeasures. GESTS Int. Trans. Comput. Sci. Enegry **19**(1)
23. Gao, J., Zhang, B., Chen, X., Luo, Z.: Ontology-based model of network and computer attacks for security assessment. J. Shanghai Jiaotong Univ. (Sci.) **18**(5), 554–562
24. Liu, F.H., Lee, W.T.: Constructing enterprise information network security risk management mechanism by ontology. Tamkang J. Sci. Eng. **13**(1), 79–87 (2010)
25. Gyrard, A., Bonnet, C., Boudaoud, K.: The STAC (Security Toolbox: Attacks and Countermeasures) ontology. ACM Companion, 13–17 May 2013
26. van Heerden, P.R., Irwin, B., Burke, I.D.: Classifying network attack scenarios using an ontology. http://eprints.ru.ac.za/4170/1/Classifying%20Network.pdf
27. Souag, A., Salinesi, C., Comyn-Wattiau, I.: Ontologies for security requirements: a literature survey and classification. In: Proceedings CAiSE 2012 International Workshop, vol. 2012, 25–26 June 2012, pp. 61–69
28. Massacci, F., Mylopoulos, J., Paci, F., Tun, T.T., Yu, Y.: An extended ontology for security requirements. http://securitylab.disi.unitn.it/lib/exe/fetch.php?media=wisse-cameraready-paper7.pdf. Accessed 14 Feb 2014
29. Velasco, J.L., Valencia-García, R., Fernández-Breis, J.T., Toval, A.: Modelling reusable security requirements based on an ontology framework. http://ws.acs.org.au/jrpit/JRPITVolumes/JRPIT41/JRPIT41.2.119.pdf. Accessed 15 Feb 2014
30. Ramanauskaite, S., Olifer, D., Goranin, N., Čenys, A.: Security ontology for adaptive mapping of security standards. Int. J. Comput. Commun. Control (IJCCC), **8**(6) (2013)
31. Souag, A., Salinesi, C., Wattiau, I., Mouratidis, H.: Using security and domain ontologies for security requirements analysis. In: 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops (COMPSACW), 22–26 July 2013, pp. 101–107
32. Chikh, A., Abulaish, M., Nabi, S.I., Alghathbar, K.: An ontology based information security requirements engineering framework. http://www.abulaish.com/uploads/STA11B.pdf. Accessed 16 Feb 2014
33. Ekelhart, A., Fenz, S., Neubauer, T.: Security ontologies: improving quantitative risk analysis. In: 40th Annual Hawaii International Conference on System Sciences HICSS, pp. 156a (2007)
34. Ahmed, M., Anjomshoaa, A., Nguyen, T.M., Min Tjoa, A.: Towards an ontology based organizational risk assessment in collaborative environments using the Semanticlife. http://publik.tuwien.ac.at/files/pub-inf_4730.pdf. Accessed 17 Feb 2014
35. Assali, A.A., Lenne, D., Debray, B.: Ontology development for industrial risk analysis. In: IEEE International Conference on Information & Communication Technologies: from Theory to Applications (ICTTA 2008), Damascus, Syria, Apr 2008
36. Xu, H., Xiao, D., Wu, Z.: Application of security ontology to context-aware alert analysis. In: Eighth IEEE/ACIS International Conference on Computer and Information Science ICIS 2009, 1–3 June 2009, pp. 171–176
37. Beji, S., El Kadhi, N.: Security ontology proposal for mobile applications. In: Tenth International Conference on Mobile Data Management: Systems MDM '09, Services and Middleware, 18–20 May 2009, pp. 580–587
38. Hacini, S., Lekhchine, R.: Security ontology for mobile agents protection. Int. J. Comput. Theor. Eng. **4**(3) (2012)
39. Kim, A., Luo, J., Kang, M.: Security ontology for annotating resources. Research Lab, NRL Memorandum Report
40. Dritsas, S., Gymnopoulos, L., Karyda, M., Balopoulos, T., Kokolakis, S., Lambrinoudakis, C., Katsikas, S.: A knowledge-based approach to security requirements for e-health applications. http://www.ejeta.org/specialOct06-issue/ejeta-special06oct-4.pdf. Accessed 18 Feb 2014
41. Ciuciu, I., Clearhout, B., Schilders, L., Meersman, R. Ontology-based matching of security attributes for personal data access in ehealth. In: OTM 2011, Part II, LNCS 7045, pp. 605–616. Springer, Berlin Heidelberg (2011)
42. Bialas, A.: Ontology-based security problem definition and solution for the common criteria compliant development process. In: 2009 Fourth International Conference on Dependability of Computer Systems IEEE Computer Society

43. Evesti, A., Savola, R., Ovaska, E., Kuusijärvi, J.: The design, instantiation, and usage of information security measuring ontology (2011)
44. Vorobiev, A., Han, J., Bekmamedova, N.: An ontology framework for managing security attacks and defences in component based software systems. In: 19th Australian Conference on Software Engineering IEEE Computer Society (2008)
45. Elahi, G., Yu, E., Zannone, N.: A modeling ontology for integrating vulnerabilities into security requirements conceptual foundations? In: Proceedings of 28th International Conference on Conceptual Modeling, vol. 5829, 9–12 Nov 2009. Springer, Berlin Heidelberg, pp. 99–114
46. Obrsta, L., Chaseb, P., Markeloffa, R.: An ontology of the cyber security domain. http://ceurws.org/Vol966/STIDS2012_T06_ObrstEtAl_CyberOntology.pdf. Accessed 19 Feb 2014
47. Takahashi, T., Kadobayashi, Y., Fujiwara, H.: Ontological approach toward cyber security in Cloud Computing. In: Proceedings of the 3rd International Conference on Security of Information and Networks, 07 Sept 2010, pp. 100–109
48. Subramani, K., Rajagopal, P.D.P., Sundaramoorthi, S.: Ontology in Cloud Computing. http://cloudontology.wikispaces.asu.edu/Use+of+Ontology+in+Cloud+Computing#UseOfOntologyInCloudComputing-DesignofSecuritySystem. Accessed 19 Feb 2014
49. Choi, C., Choi, J., Kim, P.: Ontology based access control model for security policy reasoning in cloud computing. J. Supercomput. **67**(3), 711–722
50. Kamalakannan, E., Prabhakaran, B., Arvind, K.S.: A study on security and ontology in cloud computing. Int. J. Adv. Res. Comput. Commun. Eng. **2**(10) (2013)
51. Talib, A.M., Atan, R., Abdullah, R., Murad, M.A.M.: Security ontology driven multi agent system architecture for cloud data storage security ontology development. IJCSNS Int. J. Comput. Sci. Netw. Secur. **12**(5) (2012)
52. Bermejo, J.: A Simplified Guide to Create an Ontology, AS Lab R-2007–004 v 0.1 Draft, 22 May 2007
53. OWL Web Ontology Language, W3C Recommendation, 10 Feb 2004
54. Fenz, S. Security Ontology. http://stefan.fenz.at/research/securityontology/
55. Goel, A., Goel, S.: Security issues in cloud computing. Int. J. Appl. Innov. Eng. Manage. **1**(4) (2012)
56. Kauba, C., Mayer, S.: Data confidentiality and privacy in cloud computing, 14 July 2013
57. Al-Saiyd, N.A., Sail. N. Data integrity in cloud computing security. J. Theor. Appl. Inf. Technol. **58**(3) (2013)
58. Ahuja, S.P., Mani, S.: Availability of services in the era of cloud computing. Netw. Commun. Technol. **1**(1) (2012)
59. Feng, J., Chen, Y., Ku, W.S., Liu, P.: Analysis of integrity vulnerabilities and a non repudiation protocol for cloud data storage platforms. In: 39th International Conference on Parallel Processing Workshops (ICPPW). IEEE, 13–19 Sept 2010
60. Talib, A.M., Atan, R., Abdullah, R., Murad, M.A.A.: Security ontology driven multi agent system architecture for cloud data storage security: ontology development. IJCSNS **12**(5) (2012)
61. Boyce, S., Pahl, C.: Developing domain ontologies for course content. Edu. Technol. Soc. **10**(3), 275–288 (2007)
62. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontolog. Stanford University, Stanford, CA, 94305
63. Tudorache, T., Vendetti, J., Noy, N.F.: Web-Protege: A Lightweight OWL Ontology Editor for ´the Web. Stanford Center for Biomedical Informatics Research, Stanford University, CA, US
64. Saadati, S., Denker, G.: An OWL-S Editor Tutorial Version 1.1, SRI International Menlo Park, CA 94025. http://owlseditor.semwebcentral.org/documents/tutorial.pdf
65. Using DL reasoners in Protege-OWL, Protégé. http://protegewiki.stanford.edu/wiki/Using_Reasoners
66. Sivakumar, R., Arivoli, P.V.: Ontology visualization protégé tools –a review. Int. J. Adv. Inf. Technol. (IJAIT) **1**(4) (2011)
67. Horridge, M.: OWLViz. http://protegewiki.stanford.edu/wiki/OWLViz
68. Falconer, S.: OntoGraf. http://protegewiki.stanford.edu/wiki/OntoGraf

69. Basic Editing with Protégé, Open Semantic Framework. http://wiki.opensemanticframework.org/index.php/Basic_Editing_with_Prot%C3%A9g%C3%A9
70. Introduction to Ontologies with Protégé, Trac Powered. https://wiki.csc.calpoly.edu/OntologyTutorial/wiki/IntroductionToOntologiesWithProtege
71. Risner, J.: CMRP AssetPoint, Cloud Computing for Maintenance and Asset Management
72. Marinescu, D.C.: Cloud computing: cloud vulnerabilities. TechNet Mag. http://technet.microsoft.com/en-us/magazine/dn271884.aspx (2013)
73. Soat, J.: The cloud's five biggest weaknesses, information week. http://www.informationweek.com/cloud/the-clouds-five-biggest-weaknesses/d/d-id/1089865? (2010)
74. Revalla, M., Gupta, A., Bhase, V.: Proceeding of the International Conference on cloud Security management, Center for Information Assurance and Cybersecurity University of Washington, Seattle, USA, pp. 111, 17–18 Oct 2013
75. A Coalfire Perspective Top 10 Risks in the Cloud by BalajiPalanisamy, VCP, QSA, Coalfire March. http://www.coalfire.com/medialib/assets/PDFs/Perspectives/Coalfire-Top-10-Risks-in-the-Cloud.pdf (2012)
76. Grimes, R.A.: The 5 cloud risks you have to stop ignoring, Infoworld. http://www.infoworld.com/article/2614369/security/the-5-cloud-risks-you-have-to-stop-ignoring.html, 19 Mar 2013
77. Kaur, N., Kama, K.P.: Attacks and their countermeasures in cloud computing. Discovery **15**(9) (2014)
78. Moy Chatterjee, J.: Fog computing: beginning of a new era in cloud computing. Int. Res. J. Eng. Technol. (IRJET), **4**(05), p. 735 (2017)
79. Preden, J.S., Tammemäe, K., Jantsch, A., Leier, M., Riid, A., Calis, E., Wien, T.: The benefits of self-awareness and attention in fog and mist computing. http://www.cs.rug.nl/~roe/courses/isc/FogMistComputing.pdf
80. Iorga, M., Feldman, L., Barton, R., Martin, M.J., Goren, N., Mahmoudi, C.: Fog Computing Conceptual Model, NIST. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-325.pdf, 14 Mar 2018
81. Aazam, M., Khaled, S.Z., Harras, A.: Offloading in fog computing for IoT: review, enabling technologies, and research opportunities. Future Gener. Comput. Syst. **87**, 278–289 (2018)
82. Dhiman, E.P.: Fog computing-shifting from cloud to fog. Int. J. Eng. Sci. Comput. http://ijesc.org/upload/4993917066f6f913eac87adaa8e9dc3c.FOG%20COMPUTING-Shifting%20from%20Cloud%20to%20Fog.pdf, Mar 2017

Cloud Based Supply Chain Networks—Principles and Practices

Anil B. Gowda and K. N. Subramanya

Abstract Cloud computing services are growing and developing at a rapid pace. The growth of cloud services is taking place in various forms that are suitable for various applications. Cloud computing as a computing service caters to the needs of various applications by providing a distributed environment for computing services. The services are useful in managing various engineering and management processes in the supply chain network (SCN). A supply chain network is encapsulated with challenges related to capital costs, operational costs, timely availability of information, management information system and overhead costs. Cloud computing services offer services that helps in minimizing the problems related to the challenges faced and thereby increases the productivity of the supply chain network. The supply chain firms can benefit from cloud services by way of reduced capital expenditure and reduced operational costs. Supply chain firms can be decentralized units which can utilize fog computing in which data and applications are distributed logically between data points and cloud. Also, emerging concepts of mist computing can be helpful wherein the system architecture pushes the processes nearer to the data source in a supply chain network. The expenditure mainly is because of the operational cost which can be minimized by suitable deployment of cloud computing service. The cloud services are available today as metered services thereby the operational cost can be reduced by regulating the usage. Cloud computing services from Microsoft, Google, and Amazon are already available at reasonable rates and many services providers are entering the market to pose more competition which is healthy for users from reduced pricing. It can be expected that with the developments in cloud services, the service charges will decrease which will be a great advantage to the supply chain network firms and other related firms. A supply chain network enabled with cloud computing services is referred to as the Cloud Supply Chain Network (CSCN). In this paper, the benefits and opportunities of cloud supply chain

A. B. Gowda (✉)
Jain University, Bangalore, India
e-mail: anilblr2008@gmail.com

K. N. Subramanya
RV College of Engineering, Mysore Road, RV Vidyaniketan Post, Bangalore, India
e-mail: subramanyakn@rvce.edu.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_5

network along with the challenges faced by the firms are discussed. Firms while planning to adopt cloud services to enhance their supply chain network processes look for budgeting plans which can be determined using functions as discussed in this paper. The perspectives, principles and practices in cloud supply chain network are detailed. It also describes how one can consider the parameters of the characteristics of cloud computing in supply chain network for the purpose of modeling and analyzing the information flow. A framework of the design factors of cloud supply chain is explained which will enable the decision maker to derive the necessary results by suitably incorporating the factors in the analysis of cloud supply chain network.

Keywords Cloud supply chain networks · Cloud characteristics
Fog and mist computing · Cloud benefits

1 Introduction

The Information Technology has evolved rapidly over the last few decades. Applications for various purposes also have been emerging and growing time to time. The application evolution over decades lead to the development of applications today which are available on cloud. The applications are fast emerging on cloud platform. The cloud market is growing at a fast pace worldwide. The cloud services market in India is anticipated to grow at a rapid rate. The U.S. National Institute of Standards and Technology (NIST) has a set of working definitions that separate cloud computing into service models and deployment models [18]. The NIST model originally did not require a cloud to use virtualization to pool resources, nor did it absolutely require that a cloud support multi-tenancy in the earliest definitions of cloud computing. Multi-tenancy is the sharing of resources among two or more clients. The cloud computing networks use virtualization and supports multi-tenancy. The NIST has laid down various definitions related to cloud computing, cloud platforms, cloud services and various other related recommendations that can be considered for designing a cloud-based service. NIST defines Cloud Computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

1.1 Cloud Computing and Supply Chain Network—A Perspective

Cloud computing refers to various computing facilities to suit applications and services that operate on a distributed network by using resources that are located in remote places but available to the user as virtual resources. The virtual resources are

accessed by common Internet protocols and networking standards. Virtual resources are enormous in size. Details of the physical system to be ported on the cloud are specified and abstracted by the user. Not all applications are benefited by cloud. Applications with issues related to cost, latency, transaction control, security and licenses can benefit from cloud. Cloud computing takes the technology, services and application that are similar to those on the internet and turns them into a self-service utility. Cloud refers to the concept of Abstraction and Virtualization.[1] Abstraction means deriving the application details of system implementation from users and developers for the purpose of running the application physical systems not known to the user or developer. Virtualization means providing virtual platform of resources for computing services and sharing resources. Systems and storage can be provisioned as needed from a centralized infrastructure. Cloud computing is available as deployment model and service model. Deployment models refers to the location and management of the cloud infrastructure. Services models consists of the particular types of services that can be accessed on a cloud computing platform.

A Supply Chain Network (SCN) is a network that consists of all partners involved, directly or indirectly, in fulfilling a customer request. The supply chain includes the manufacturers, the suppliers, transporters, warehouses, retailers, and customers. Within each organization, the supply chain includes functions that are involved in receiving and fulfilling a customer request. The customer request could be taken as one of the basis to decide on new product development, marketing, distribution, finance and customer service. Information technology applications and systems are essential to run various types of businesses and enterprises [7]. IT applications help in governing and optimizing operations in the retail, manufacturing and other industries. Optimizing Return-on-Investment (ROI) in the IT area is essential for a business performance. In a supply chain management there are various stages. Each stage is a set of supply chain process. All processes in the chain connected with various supply chain partners is a network referred to as supply chain network (SCN). Broadly a SCN involves stages that include the following:

- Customers
- Retailers
- Wholesalers
- Manufacturers
- Suppliers

Information flow takes plays in two directions. Flow from customers to suppliers and from suppliers to customers passing through various stages in the SCN. Understanding the information flow through various processes in a supply chain can be cyclic or sequential. The knowledge of cyclic or sequential flow is essential in order to analyze the flow in the supply chain network. The flow of information upstream can lead to flow of product or service downstream and these flows can be analyzed through various process mechanism. In a cyclic process, the supply chain is considered to be divided into number of cycles that exists between stages of supply chain. Various

[1]Cloud Computing Bible—Barrie Sosinsky, Wiley India Pvt. Ltd., reprint 2012.

cyclic processes that can be considered are the customer order, manufacturing, and procurement. In a sequential process, the push and pull exerted on the supply chain at every stage is considered to analyze the flow upstream and downstream.

A pull process is triggered by a customer order whereas a push process is triggered in anticipation of a customer orders. The push process is a reactive process and the push is a speculative process. Pull process is based on known customer demand. Throughout the processes, the goal of the buyer is to ensure product availability and to achieve economies of scale in ordering. The processes can be divided into three main categories of management. The first category is the management of customer relationship processes, the second category is the management of firm processes and the third category is the management of supplier relationship processes. Some of the processes that are observed in the three categories that are related to the information flow are the information pertaining to the details of the product, information about funds required to generate, receive and fulfill a customer request, information about generated customer demand, facilitating orders, placement and tracking of orders etc. The information gathered on planning of internal production and storage capacity, preparation of supply and demand plans, and fulfillment of actual orders is critical in supply chain network decision.

1.2 Cloud Based Supply Chain Network (CSCN) and Its Benefits

A proper integration of technology like a cloud computing service after careful evaluation of the services and investments in the technology in a planned manner can help supply chain firms to become highly competitive. There is a need to assess various cloud computing services that are suitable for adoption in supply chain networks and determine their contributions in a collaborative manner. Though many direct and indirect benefits can be realized from cloud computing services for supply chain networks, not all the services are necessary. Every process in the supply chain may or may not require all the services and hence the key cloud services have to be identified. Some of the benefits offered by cloud computing services for supply chain networks are indicated in Table 1.

A supply in a supply chain network refers to supply moving from suppliers to manufacturers to distributors to retailers to customers along a chain. It is essential to visualize the flow of information pertaining to the information flow of various critical points like the forecasted demand of items, product features expected from the company, size of order to be fulfilled by the different nodes of the supply chain of the company, information on the financial aspects, flow of funds, information about the stocks at different levels of the supply chain network, product status in terms of production stage, completed production, packaging details, rejected items, items in transport, items in queue, items delivered, logistics information and many more. Also, it is critical to interlink the suppliers to the manufacturers, manufacturers to

Table 1 Benefits of using cloud computing services in SCN

#	Factor	Description
1	Metered services	The use of cloud system resources can be metered, verified, audited and reported to the customer which enables a supply chain network process to be very cost effective
2	Pooling and sharing of resources	A cloud service provider creates resources that are pooled together supporting supply chain network partners enabling them for multi-tenant usage and share resources
3	Upgrading systems	Cloud computing service system is centralized and can be easily upgraded offering better facilities to their supply chain network partners
4	Third party services	A cloud computing deployment lets one supply chain firm to manage their requirements by letting the services to another competent part to accomplish their task
5	Scalability	Resources can be increased or decreased easily with cloud services that helps supply chain networks to manage resource efficiently
6	Quality of service	The Quality of Service for supply chain network can be assured by the vendor or a partner rendering the service
7	Convenient utilization	Depending upon the supply chain network service offered, firms can minimize hardware or software licensing costs
8	Service on demand	A supply chain network partner can avail computer resources from cloud service provider without interacting with cloud service provider agent.
9	High speed broadband N/W	Access to resources in the cloud is available over the network catering to the supply chain network process loads that demand high speed broad band services
10	Reduced costs	cloud networks operate at higher efficiencies and with greater utilization of resources in the form of sharing and collaboration and hence significant cost reductions are often encountered which is beneficial to supply chain network

the wholesalers, wholesalers to the retailers and retailers to the consumers along both directions in the supply chain. Supply chains are actually networks and hence are also referred to as Supply Network or Supply Web. Integration of the Cloud technologies to handle the information flow and managing volumes of data that moves upstream and downstream in the supply chain is very essential. Integrating cloud computing services with Supply Chain Networks for deriving an efficient supply chain network is referred to as the Cloud Supply Chain Networks (CSCN). A supply chain network firm always looks for significant benefits by way of integrating Information Technology so that they can stay competitive in business. There is always a pressure from the competitors for pricing various products and services.

Like a Supply Chain, a typical Cloud Supply Chain Network (CSCN) may involve cloud interventions at different stages that include the Customers, Retailers, Wholesalers, Manufacturers and Suppliers and additionally the cloud service provider. The cloud service provider can be considered as one of the suppliers in the CSCN.

Whenever an item is supplied, it is based on the customer demand which causes the demand information flow upstream through various stages. The flow of items takes place starting from the supplier going through various stages and finally the customer along with relevant information. The entire flow is governed by the actual demand or is generated through a forecast which is also based on actual orders from the customers from the past orders. The demand information is then communicated upstream from customer level to supplier level.

1.3 Principles and Practices in CSCN

The Information Technology growth has led to development of various principles which over years have been evolving to cater to the needs of supply chain network. The requirements of managing the supply chain network processes to be enabled with cloud computing services necessitates that certain methodologies be adopted in building the cloud supply chain network. Factors that govern the performance of supply chain network need to be understood in order to incorporate the cloud service. A system of CSCN can be analyzed keeping certain performance drivers and metrics in mind. A framework that encapsulates the required drivers and metrics can be built to study the impact of one characteristic over other characteristics. Based on the platform of supply chain application, appropriate cloud service characteristics are factored and detailed study on the influence of each characteristic helps in arriving at decision for the integration of cloud computing service in supply chain network.

Cloud Computing refers to applications and services that run on a distributed network using virtualized resources and accessed by common Internet protocols and networking standards. It is distinguished by the notation that resources are virtual and limitless and that details of the physical systems on which software runs are abstracted from the user. Two different classes of cloud computing services are those that are based on deployment models and those based on the service models. Deployment models tells where the cloud is located and for what purpose. Private, public, community, and hybrid clouds are deployment models that can be considered for building up the IT resources for a SCN firm. Service models describe the type of service that the service provider is offering. The service models for SCN firms could be one or more of Software as a Service, Platform as a Service and Infrastructure as a Service. These models are considered on the lines of the recommendations of NIST. Service models built on one another define what a vendor must manage and what the client's responsibility is. The massive scale of cloud computing systems was enabled by the popularization of the Internet and the growth of some large service companies. Cloud computing makes the long-held dream of utility computing possible with a pay as you go, infinitely scalable, universally available system. Not all applications benefit from deployment in the cloud. Issues with latency, transaction control, and in particular security and regulatory compliance are of particular concern. These models can be adopted to suit the requirements of the SCN firm. A brief detail of the suitability of these models for SCN firm are listed below.

- SCN Public Cloud—The public cloud infrastructure is available for public. Supply chain firms can categorize it for a large industry group. The SCN firms can get associated with the cloud service providers and design a collaboration model to suit their business needs.
- SCN Private Cloud—The private cloud infrastructure is operated for the exclusive use of firms that have supply chain network enabled on cloud services. The cloud services may be managed by the SC firms or a third party who supports the SC firms. Private clouds may be either on-or off- premises.
- SCN Hybrid Cloud—A hybrid cloud combines multiple clouds belonging to firms that are managing supply chains. Clouds retain their unique identities and are bound together as a unit. A hybrid cloud can offer standardized access to supply chain data and supply chain applications, as well as support portability of SC applications.
- SCN Infrastructure as a Service (SCNIaaS)—Infrastructure provides virtual machines, virtual storage, virtual infrastructure and other hardware assets as resources for managing various SCN business processes.
- SCN Platform as a Service (SCNPaaS)—PaaS provides virtual machines, operating systems, applications services, development frameworks, transaction, and control structures that can be provided with customized services to suit the requirements of SCN firms.
- SCN Software as a Service (SCNSaaS)—SaaS is a complete operating environment with applications, management, and the user interface that can help manage the SCN processes.

A trade off can be drawn depending upon the requirements of supply chain network firm. Firms can decide on the adoption of cloud computing services for managing the processes of a supply chain network.

The ability to access pooled resources on a pay-as-you go basis provides a number of system characteristics that completely alter the economics of information technology infrastructures and allows new types of access and business models for user applications in particular supply chain management. A supply chain firm enabled with cloud computing services can improve supply chain performance in terms of responsiveness and efficiency.

Various supply chain components to be considered for designing and practical implementation of cloud computing services in a supply chain network firm are indicated in Table 2.

The main cloud computing service factors contributing to the performance of better supply chain network are considered in the form of scalability, price flexibility, better collaboration services, multi-tenancy, decision support, reduced hardware cost, reduced software cost and enhanced efficiency. These factors essentially help in driving various supply chain processes in a supply chain network. Resources necessary for the supply chain processes can be better utilized. These factors are considered as the key drivers in the supply chain network. Each of these drivers can be assessed for its applicability in a supply chain network by using various optimization and analytical methods. The three important concepts that can be considered to govern

Table 2 Components for cloud based SCN

#	Components	Design consideration
1	Facility	Facilities are the actual physical locations in the SCN supported by cloud computing services where the product is stored, assembled or fabricated. Decisions regarding the role, location, capacity and flexibility of facilities have a significant impact on the supply chains performance. The firm's production can significantly get affected based on the cloud computing services available. Variables like capacity of the firm to perform its intended functions, centralizing or decentralizing activities, number of such activities depends on available cloud services
2	Transportation	Transportation in SCN refers to moving items from point to point in the supply chain network. Shipment information can become more exhaustive and quicker when resorted to cloud computing services. Based on the information on shipment which can rapidly be provisioned with a cloud service, a firm can decide whether transportation from a supply source will be direct to the demand point or will go through intermediate consolidation points and the choice of transportation mode. Important variables that can be considered to analyze the transportation aspect are average inbound transportation cost, incoming shipment size, inbound transportation cost per shipment, outbound transportation cost, outbound shipment size, inbound and transportation cost per shipment
3	Pricing	Pricing determines how much a SCN firm will charge for goods and services that it makes available in the supply chain enabled with cloud computing services. Factors that can be considered in pricing are the economies of scale, low price and high price in a day, and comparing fixed pricing strategy with varying price especially when the processes are supported on cloud computing facility. Variables are the profit margin, variable cost per unit, average sale price and order size which can get significantly impacted by the cloud computing facility in SCN
4	Inventory	Inventory includes all raw materials, work in process, and finished goods within a SCN supported with cloud services. Changing inventory policies can dramatically alter the cloud enabled supply chains efficiency and resposiveness. To make inventory more responsive and more efficient, supply chain firms must make decisions on average amount of inventory, safety inventory held in case demand exceeds expectations, seasonal inventory to counter predictable variability in demand since these parameters can get significantly influenced by cloud services
5	Sourcing	Sourcing refers to the choice of a vendor. Vendor is one who will deliver a product or service to the cloud-based supply chain network. The strategic decision is to determine what function a firm performs and what function the firm outsources. The most significant sourcing decision for a cloud-based supply chain firm is whether to perform a task in-house or outsource it to a third party. Variables that can be considered are average purchase prices, quantity lead time, reliability and quality, number of suppliers, service providers for managing the cloud supply chain processes
6	Information	Information based on cloud enabled services is potentially the biggest driver of performance in the cloud supply chain network because it directly affects each component in a SCN. Important parts of information characteristics are the dynamics of push versus pull, sharing of information, forecasting, and technology used which can get greater boost with the support of cloud computing platform. Coordination occurs when all stages of a cloud supply chain work towards the objective of maximizing total CSCN profitability based on information sharing. Cloud supported applications and techniques can be used to forecast sales or market conditions. Variables that can be considered for quick flow of information may occurs frequently because of inaccurate forecasts, updations, errors in forecast, seasonal variations, variation in plan from actual when using a cloud service in supply chain management

the implementation of cloud computing services in a supply chain network firms are the CSCN Systems concept, CSCN Cost concept and CSCN Operations concept.

(a) CSCN Systems Concept: For a CSCN, the systems concept emphasizes interdependence not only between functions within an organization but also among multiple organizations that collectively deliver products and services to the customers. A suitable deployment of cloud can help minimize the losses and wastages of resources in a SCN firm. The impact of using cloud services on the SC functions should be considered in building a system for efficient management of CSCN.
(b) CSCN Cost Concept: Value delivered to the customer can be maximized by SCN firms only if the total cost incurred by all the links in the supply chain network serving the customer is minimized which can be made possible by CSCN. A cloud model in the supply chain process can help in achieving cost reduction. The concept emphasizes the need for intercompany coordination, cooperation and collaboration using cloud services. Cloud computing services can be considered in all activities from design and development to manufacture and distribution in order to minimize the total cost and thereby maximize the value delivered to the customer.
(c) CSCN Operations concept: The operations concept involved in a CSCN helps the decision makers to explore the possibilities of choosing the right set of alternatives or combination of alternatives in fulfilling a cloud supply chain objective, in a manner that the overall cost of the CSCN operations is minimized.

Components of supply chain network and the three concepts play a critical role in a supply chain network firms strategic decision-making process. It is possible to enable different business models by tactically considering the variables in the design of CSCN components to be enabled with cloud computing services. It is possible for the vendors providing cloud computing services to appropriately implement services for supply chain network using the service models like IaaS, PaaS and SaaS. Practically it is convenient if a sequential process is followed to arrive at the decision. Firms can customize the processes for decision making for cloud adoption or for cloud based SCN decision to manage the SCN of the firm. A flow of the information in a CSCN starting with the anticipation of demand to the stage of arriving at a final strategic decision is indicated in Fig. 1.

The industry is growing at a rapid rate. With most of the suppliers having set a system for delivery the supply chain network is under the ever-increasing stress of time reduction and cost reduction. Added to this the supply chain firms face the problem of inventory. It is a situation of tradeoff between time and cost. The services of IT in the form of cloud computing services are becoming a necessity. The awareness levels are increasing about its benefits and cost reduction along with time reduction. The ever-growing demand for IT services is now getting shaped in the form of cloud computing framework that can be a boon to all such supply chain firms. However, the benefits can be derived more efficiently if a mechanism to analyze such cloud services and its adoption can be taken up on a case to case basis.

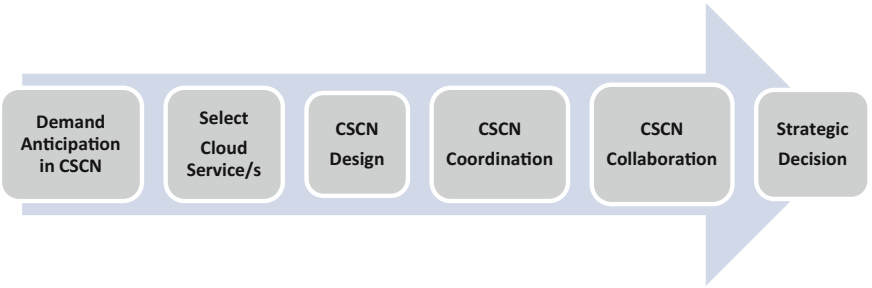


Fig. 1 Processes to arrive at a decision in CSCN

1.4 Forecasting in a Cloud Based Supply Chain Network

Another important aspect of managing the CSCN is the accuracy with which the demand is forecasted. Various methods of forecasting can be used to arrive at a demand. But the accuracy of the forecasts is uncertain. A deviation in the forecasted value can escalate the cost of the unit supply. The cost per unit gets magnified as one moves up the supply chain in the CSCN. Demand forecasts is the backbone of all major decision in the supply chain planning. A cloud-based supply chain will get a relief either in the form of reduced infrastructure cost and reduction in variability in demand at various nodes of supply chain through ha proper implementation of cloud services at each node the supply chain network. All push processes in the CSCN are based on the anticipated customer demand, whereas all pull processes are based on the response to customer demand. A cloud service will help in bridging the gap between the customer and the various players of the CSCN. Errors if any in the forecasted figures can be minimized greatly through constant interaction with every member connected through a cloud application service. With the narrowing of the deviations in the forecasts, predictions become closer to reality and thus decision making by supply chain managers become easy. Various aspects of forecasts should be considered in decision making in the cloud enabled supply chain network. Forecasts are always inaccurate and should thus include both expected value of the forecast and a measure of forecast error. Long term forecasts are usually less accurate than short-term forecasts. As the information moves up the SCN the forecast figures tend to get more and more distorted.

Customer demand is influenced by a variety of factors and can be predicted with a certain degree of accuracy. In arriving at the forecast figures various factors can be identified especially when the supply chain is supported on a cloud-based platform. Through proper supply chain collaboration of partners and vendors connected through a cloud based virtual network it is possible to share the resources and information on customer demand. The variations leading to inaccurate forecasts can be mitigated by adopting a proper IT service which can be enabled through collaborating firms on a cloud platform. The common factors considered in forecasting can all be view from cloud service perspective in order to achieve an efficient forecast

result. Factors that can be considered in conjunction with cloud service to explore possibility of a better forecast are Past demand, Lead time, advertising or marketing efforts, economics, discounts and competitor's strategy. Some of the forecasting methods that are commonly used are qualitative forecast, time series forecast, causal forecast and simulation. A company may find it difficult to decide which method is most appropriate for forecasting.

In a CSCN, the purpose of forecasting method is to predict the demand. In its most general form the systematic demand data contains a cyclical component, a trend component and a seasonal factor. These components in a CSCN can be analyzed and irregular components if any can be determined. All other variations in the time series data can be attributed to the irregular component which may arise out of uncertainties in market or unforeseen circumstance like a natural calamity. De-seasoned demand information can be found out and used for various applications in a supply chain firm. Over years it is seen that the food grain industry is suffering with the problem of grains getting spoilt resulting in losses to the industry. A good forecasting package provides forecasts across a wide range of products that are updated in real time by incorporating any new demand information. Quick analysis and rapid provisioning of the demand information using cloud services in a SCN is possible that will enable minimize the losses. Much of the progress in areas such as collaborative planning is due to IT innovations like cloud computing and other techniques that allow the exchange and incorporation of forecasts between supply chain enterprises. It contains tools to perform what-if analysis regarding the impact of potential changes in prices on demand.

1.5 Fog and Mist Computing Integration with CSCN

It is important to develop a framework that helps a manager understand how information is utilized by various segments of IT within supply chain. This is possible today with the help of various types of cloud services. The use of information in the supply chain has increasingly been enabled by the support of cloud computing services and applications developed on cloud platform to support enterprise resource management. The evolution of cloud-based enterprise solution provides insights not only into the future of cloud systems, but also into what lies in the successful development of cloud supply chain network processes. From an enterprises perspective all processes within its supply chain can be categorized into three main areas—processes focused downstream, processes focused internally and processes focused upstream. This has led to the realization of four categorical processes namely Customer Relationship Management, Internal Supply Chain Management, Supplier Relationship Management and Transaction Management. Each category can be considered as a decentralized unit initially to build the cloud applications for integration into the overall system during implementation. Fog computing which is conceptualized as decentralized platform for computing various needs from applications to storage to data processing in a distributed environment specially located between the data

source and cloud computing services [22]. It is expected that with the introduction of fog computing at decentralized units of the supply chain networks, the overall efficiency can improve. Some of the computing services in Fog computing are the smart developments in vehicles, shopping malls, and customized networks.

Managers can consider several general ideas when they are planning regarding cloud supply chain and select a cloud service that addresses the company's key success factors. Adopt incremental steps to ensure success of cloud-based supply chain projects. Management must consider the depth to which a cloud computing system deals with the firm's key success factors. There is a trade-off between the ease of implementing a cloud-based system and the level of complexity involved in the cloud supply chain networks and this can be done more efficiently with the help of Fog computing. If there are trends that indicate insignificant characteristics that will become crucial in future, then managers need to make sure that the choice of appropriate cloud computing service is made. Information Technology using cloud computing services for managing supply chain processes can be highly beneficial to the firm if the technology can be well planned and harnessed. A cloud computing service is designed for various applications and supply chain firms can derive the benefits by utilizing the cloud services in a planned manner. Firms can resort to a mechanism of identifying what type of cloud computing service would be beneficial to the firm and think in terms of resorting to Fog computing. Brainstorming over the key aspects for the firm's requirement can help in identifying the necessary utilities using cloud computing services. Considering the virtualization and abstraction as the two key features in the designing of a cloud service, one can even ideate the usage of Mist computing [21]. A Mist computing is a concept that can be used in supply chain network to push the computing power of cloud in modular form to the edge of the process network very close to the data source. Mist can be a way of deploying packets of computing resources at various nodes of the supply chain network. Mist can be thought of as a localized system of computing units that can work independently at the node level before being integrated into the main cloud servers.

To create shared pools of resources for managing the SCN processes, understanding of the different cloud computing service models is necessary. Only after careful assessment of the key factors, it is possible to derive an efficient CSCN. The key to creating a pool of hardware and software necessary for managing the supply chain processes is to provide an abstraction mechanism. This will enable mapping of the logical address to a physical resource as desired. Cloud computing networks use a set of techniques to create virtual servers, virtual storage, virtual networks and virtual applications. When supply chain process requests increase in volume the load of such requests can slow down the SCN processes. Cloud computing service offer a mechanism of load balancing to deal with such service request volumes and balance the processing load of SCN processes. Under such situations load balancing through cloud service helps in routing the SCN information to virtual machines that offer high performance. Virtualization assigns a logical name for a physical resource and then provides a pointer to that physical resource when a request is made in a

SCN. Virtualization provides a means to manage resources efficiently because the mapping of virtual resources to physical resources can handle various changes in the SCN processes.

To enable supply chain network processes with cloud computing services, the computing resources must be highly configurable and flexible. One can define the features in software and hardware that enable flexibility to suit the desired supply chain process requirements. The cloud computing connectivity for supply chain firms can be configured in the form of Physical to Virtual, Virtual to Virtual, Virtual to Physical or Physical to Physical and very conveniently integrate with the decentralized fog and mist computing services. The connectivity may also include the component of cloud and datacenter. Virtualization enables cloud supply chain service interface to the clients, scalability of IT resources, sharing of computing services and metering facility for usage of cloud services. A fog, mist and cloud computing services can be realized to work in tandem with each other to bring out the best results of performance with one or more combinations of the related processes in a supply chain network [21, 22].

1.6 Coordination and Collaboration in Cloud Supply Chain Network

Supply chain performance can be enhanced by proper planning of the coordination of various functions internally within the firm. The coordination can improve if all stages of the supply chain take actions together with a common goal which can lead to increased total supply chain profits. This can be achieved effectively with the support of cloud computing services. A well-established good IT infrastructure is necessary for effective collaboration. Coordination and collaboration can suffer if each function or the firm works independently with a holistic goal. In such cases of coordination and collaboration the effect of stock piling can be reduced by considering suitable predictive models for demand prediction. Prediction at various levels of collaboration can be simulated in advance to understand how bull whip effect can be minimized [11, 26]. The benefits of such an integrated cloud supply chain process are:

(i) Reduction in the manufacturing cost.
(ii) Reduction in the Inventory cost and minimizing the space requirement for warehouses.
(iii) Decrease in the replenishment lead time.
(iv) Minimize the transportation cost.
(v) Minimize labor cost for shipping and receiving.
(vi) Proper planning of stock level is possible.
(vii) Effective utilization of resources is possible.

However, there are certain obstacles in cloud supply chain Coordination and Collaboration which are:

(i) Information Processing- Information related to forecast based on orders and not customer demand can be inaccurate. Any variability in customer demand is magnified as orders move up the supply chain to manufacturers and suppliers. A small change in customer demand becomes magnified as it moves up the supply chain in the form of customer orders causing Bullwhip effect in the SCN.
(ii) Inadequate Information—Inadequate information shared between stages of the supply chain magnifies the information distortion.
(iii) Placement of orders—Inappropriate information that causes an increase in variability in orders placed.
(iv) High quantity ordered—When a firm places orders that are much larger than what is demanded, the variability of orders is magnified up the supply chain.
(v) Larger lead time—Information distortion is magnified if t lead times between stages are long.
(vi) Pricing—pricing policies for a product can cause an increase in variability of orders placed.

There are various factors that contribute to the escalation of the cost in a supply chain network. Some of the factors that add to the cost are the material cost, transportation cost, labor cost, cost due to delay in supply, cost of storage, rental charges, technology cost, infrastructure cost, manufacturing cost, administrative cost and so on. A cloud based SCN connects all partners through the internet by means of cloud services. The partners can benefit from cloud connectivity in the form of lower investment costs and efficient flow of information at a lower cost. A broad band connectivity will allow the partners to avail high end technology benefits even without the need to possess the license on site. The benefits can be transferred through a pay as per use facility. Technology and specialized services not available in one place can be accessed and availed at a nominal cost from locations where such facilities may not be available through the cloud platform. It is also possible to consider the introduction of fog and mist computing at various levels of the collaboration nodes and achieve data synchronization using appropriate Fog or Mist services. Fog, Mist and Cloud services can be considered as complementary computing systems [21, 22].

In the current paper, various benefits offered by cloud computing services and opportunities foreseen for cloud supply chain network are detailed. In the following sections, a background of the research problem is explained. In the next section, details of literature survey are provided indicating various research works carried out in the area of cloud computing and supply chain networks. Various research objectives pertaining to the research problem are listed followed by sections containing analysis details. Out of various services available in cloud computing, the ones that are commonly useful for managing the supply chain processes are listed out and importance of each in the context of supply chain network is highlighted. Keeping in view various perspectives and practices in cloud supply chain network, factors are identified using factor analysis method. The factor analysis helps in determining the key cloud computing characteristics that will be significant for a supply chain network. Various levels of hierarchy exist in supply chain firms. Based on which, typical

hierarchy is considered as test case to model the information flow for decision support in a cloud supply chain network. The analysis done on CSCN as multiple criteria multiple level hierarchy framework using Analytic Hierarchy Process is explained. The results obtained from the analysis are discussed in the context of cloud supply chain network. The final section emphasizes various practical conclusions and highlight scope for further research which is followed by acknowledgement and a list of various references.

2 Background of the Research Problem

Cloud supply chain network is not free from challenges. As technologies evolve it is also prone to various challenges that need to be understood before considering adoption of such technologies. Cloud computing services for supply chain network needs to be assessed before its adoption since it can suffer because of factors that can hinder in its service. Some of the challenges are:

- Lack of information on which cloud services to consider for SCN
- Identification of appropriate cloud service factors to enhance the SCN
- Inadequate information for taking strategic decisions on investments and budgeting in SCN
- Uncertainty about the performances of information models
- Although many cloud computing applications are very capable, it is not certain as to what extent the contributions exist in a cloud supply chain network.

In addition, certain challenges faced by supply chain firms in the cloud supply chain network are impact of cost pressure, market volatility, margin reduction and shorter product life cycles.

Hence it is essential for supply chain firms to think aggressively on resorting to newer technologies to thwart the challenges and prepare themselves to be fit in the competitive environment. The success of a supply chain network rests in the ability of the supply chain firm to integrate information technology like cloud computing services into a supply chain network process. Integration of such technologies can offer benefits and provide competitive advantage to the supply chain firm over other firms.

Though the industry is full of challenges, there is however a positive belief that technology will offer better and better advantages slowly over a period of time. Firms should understand the incremental benefit and start adopting the technology as quickly as possible for a better cause despite the hardships one has to go through in migrating over to the new system. Cloud computing is one such system which is still growing and can offer tremendous cost benefits to the suppliers, manufacturers and also the end users. Players of the supply chain networks enabled with cloud computing services should have a good understanding of the growing standards of cloud adoption in business and be prepared for collaboration and related services.

The standardization of services will only help in implementing a global system which will be consistent with the local standards and yet meet the global requirements.

3 Review of Literature

The application of cloud computing in supply chain is innovative and has opened a new research field allowing developments for future enterprises IT solutions [12]. The solution is however not complete without optimization. Various concepts have been considered and algorithms have been developed inspired by nature for optimization and their uses in solving problems of cloud computing [19]. A Supply Chain Network (SCN) is two or more partners linked and enabled for flow of goods, information, and funds whereas a Cloud Supply Chain Network (CSCN) is two or more partners linked by the provision of cloud services and enabled for efficient flow of information leading to an efficient flow of items, funds and information pertaining to it. It is important to consider that in a CSCN the sharing of information is not the only thing leading to costs but, also the management and restructuring of services, information and finance for an optimization of the SCN which can be enabled effectively through a cloud service connectivity. This calls for suitable modeling which are driven by data, obtained at various levels of the CSCN and scaled up for bigger requirements. Various models that are used for decision making are driven by large volume of data. The outcome of a model serves as a basis to help managers take optimum decisions. Importance is given to analyze the necessity for optimization of resources since any application has the potential to scale up rapidly or shrink depending on the situation [5]. A detailed study of various applications, challenges and foundation has been done. Developments in the business world and information technology have enabled decision makers to have a sharper focus on various issues. Globalization of the supply chain activities in many companies has led to adoption of information technology. Research work on cloud migration is in early stage of maturity and therefore identifies the necessity for migration framework to initiate the process and improve migration to cloud [13].

It is also possible to consider edge assisted cloud computing and its relation to the emerging domain of Fog-of-things (FoT) which provides local computation close to clients or cloud by employing low power embedded computers [22]. Over years, various information technologies have been developed and today the world looks at cloud computing techniques as one of the major developments in IT that can help decision making. With the emergence of cloud technologies, the supply chain network management and the underlying strategy and operations can benefit a lot in terms of cost minimization and reduced operation time. The features available through cloud implementation in an organization to manage the supply chain become an essential component of modeling a CSCN. This is essential for efficient and smooth manufacturing of various products and distribution of the products. CSCN can be used similarly to derive the two competitive advantages: cost leadership and differentiation. Cloud supply chain profitability is the total profit to be shared across all supply

chain stages and intermediaries. The higher the cloud supply chain profitability, the more successful is the cloud supply chain. In years to come, it is pertinent that global software development will benefit from the cloud's infrastructure, platform, and provision of software as a service feature [23]. Cloud supply chain success could be measured in terms of cloud supply chain profitability and not in terms of the profits at an individual stage. In order to achieve higher profitability, it is necessary to reduce the costs involved at various stages of the supply chain and cloud is one such option available to the firms. The industry analysts have provided following interesting facts and figures, which has triggered the authors to take up this exploratory work. Cloud services will be essential tools for addressing the biggest business demands of IT like the speed, cost, scale, rich variety of solutions, which are essential ingredients of CSCN.

MistGIS framework for mining analytics from geospatial big data has been considered which will assist the fog and cloud computing [21]. A prototype supply chain management system protects the confidentiality of private data while rapidly adapting to changing business needs as and when needed dynamically [8]. Of the companies using cloud computing services to enable accessibility from anywhere; majority are specifically using cloud computing services for software as a service. Cloud offers various prospects to the managers for managing applications on the SCN. However, it also comes with certain challenges. One has to consider these factors before considering cloud computing as a support for their applications. The contribution to the analytical study of real time information sharing based on Cloud Computing services, using a simulation model to calculate the expected benefits of a Cloud Computing in a supply chain network is very effective in analyzing various scenarios [16].

Identifying the increasing potential for cloud computing solutions for the supply chain networks and the need to improve and extend business processes outside the four walls through collaborations with trading partners is paramount. For all such purposes, cloud computing offers a cost-efficient opportunity to conduct business activities [10]. The adoption of cloud computing for supply chain system and represent supply chains as a set of service offerings and customer demand as a service request is an essential process in considering cloud supply chain [14]. An approach to the design of discrete event simulation experiments aimed at performance analysis using case study on cloud computing is necessary in order to understand the behavior of various cloud characteristics [20]. Emerging supply chains and supply networks need to be analyzed in detail with different cases integrated with cloud computing services. Various processes involved in the cloud supply chain have to be examined and analyzed as in [17].

Various categories of Application Programming Interface (API) namely Ordinary Programming, Deployment, Cloud services, Image and Infrastructure Management and Internal Interfaces which can be effectively considered in the modeling of a cloud supply chain network should be considered in designing a cloud supply chain network [1]. Another concern in supply chain network is the storage of perishable items like grains. There can be significant benefits for managing such storages using cloud services after critically examining the storage scenario. The spoilage rates in

India's grain supply chains have been estimated to be 25–30%. There is a scope of utilizing cloud computing services that may help the firms to reduce the spoilage rate and bring the full benefit of IT to their small merchants [25]. Cloud computing can be seen as a supply chain integrated computer that delivers and refines computing power towards its customers whereby supply chains can be optimized by actively managing the processes [9]. Various scheduling algorithms are useful to manage the cloud computing resources for which a simulation technique can be considered for comparison and enable firms to decide adapt cloud environment in work places [24]. Forecasting demand is an important step in supply chain network for which simulation analysis on cloud workload prediction can be done to evaluate the accuracy of future workload prediction by relating it to the demand estimated [6].

It is also important to minimize the effect of bullwhip effect in a supply chain network. The benefits of using cloud computing services can be analyzed using simulation approach to create specific supply chain and quantify the bullwhip effect. Thereafter computations can be done using an adaptive network based fuzzy system to quantify and reduce the bullwhip effect in a multi-product, multi-stage supply chain inventory model [11]. The Bullwhip-Effect affects the efficiency of traditional supply chains and therefore there is a need to identify how it applies to the world of Cloud Computing [15]. It is possible to consider economical cloud computing solution to minimize operational cost and used DVFS (Dynamic Voltage and Frequency Scaling) for cloud computing datacenters as in [2]. Various approaches to analyze the cloud computing benefits in supply chain networks are essential in order to carefully arrive at the relationships between different cloud computing services. The analytical methods that can be considered to determine key factors out of many factors and deriving a relationship from multi-criterion two level hierarchy in a decision system using AHP helps in strategic decision of investments budgeting and resource allocation in supply chain network firm [3, 4].

3.1 Research Gap

A CSCN that links the customer to the supplier through a chain of retailer, whole seller and manufacturer is driven by the players of sourcing, material manager, distributors, and logistic manager. The key to an efficient information flow in the supply chain network is the understanding of various characteristics of cloud supply chain network which is not available in the literature. The supply chain firms have been bringing about stiff competition in the market in order to stay ahead in the race. The main objective of the supply chain has been to maximize the profit at every stage of the supply chain. To achieve this, it is necessary to minimize the cost at every stage. Analysis of the cloud computing benefits and information flow modeling for a supply chain network is not addressed in the literature. Also, the influence of one cloud characteristics over the other that is necessary for strategic decision in supply chain network is not addressed. Hence a thorough research is done to bring out the details

that will enable SCN firms to take strategic decisions on implementing cloud services for managing their supply chain networks.

4 Research Objectives

For the purpose of analyzing the problem defined for this paper, various objectives are stated as given below:

i. To understand the important characteristics of cloud computing services suitable for supply chain network.
ii. To elaborate on the Principles and Practices for cloud supply chain network.
iii. To understand the analytics involved in identifying key characteristics.
iv. To bring out the contribution of key characteristics of cloud in cloud supply chain network.

5 Methodology, Analysis and Results

Cloud computing model offers a promise of cost savings with increased IT agility. Cloud industry is growing quickly and vendors are investing significant amount of money to develop 'Software-as-a-service' (SaaS) and harness the benefits of cloud computing in various business activities. It is essential to determine the key cloud computing characteristics for supply chain network and evaluate the importance of each characteristics.

A three-step methodology provides a detailed insight into the various characteristics of cloud supply chain network. The steps are:

(i) Gathering information from survey
(ii) Identification of key characteristics using Factor Analysis
(iii) Multi-criteria decision analysis using AHP

A survey method helps in finding out quickly the benefits of cloud computing services for supply chain. The benefits are variables. In a factor analysis of the benefits of cloud to suit the firm's needs is important to decide the set of factors to be adopted for the CSCN. This depends on the application of cloud needed to manage the supply chain network. Many companies have already adopted Cloud SaaS and many more companies are in the process of adopting it. Companies have realized that cloud computing can offer large benefits to the supply chain network. As a case of a small and medium enterprise, cloud services for managing supply chain network various cloud benefits are considered. In a particular case, certain cloud benefits were identified through a survey. The influence of various cloud characteristics on designing a cloud supply chain network was analyzed and key factors identified.

Various factors that can be considered for assessment of adopting cloud for supply chain networks are Collaboration, Pricing, Investment, Convenience, Efficiency, Scalability, and Launching new services, Cost of both hardware and software, Multitenancy, Decision support, Utility and Operational cost. The key factors vary from company to company depending on the nature of their services. A firm can list out their requirements and identify key factors from the list of indicated factors. Not all factors contribute equally to the supply chain network. A detailed analysis on factoring the benefits of the cloud services for supply chain network is done for a selected set of factors.

5.1 Factor Analysis

For the purpose of identifying key factors and evaluating the factors, a typical practical case is considered with selected CSCN factors namely Process Efficiency (F1), Pricing (F2), Investment (F3), Convenience (F4), Collaboration (F5), Scalability (F6), and Software Cost (F7). These factors representing the cloud computing services available as benefits to the supply chain network are listed out based on the responses obtained from the respondents with adequate know how of the domain. Additional factors can also be considered on the same lines for factor analysis. For the purpose of analyzing a practical case, a sample set is considered with respect to the seven factors listed above. From the set, key factors are identified along with the details of total variance and contributions of key factors to the supply chain network [3].

From the data obtained on the seven factors, correlation coefficient is calculated between every pair of factors. The correlation coefficient can be obtained by using statistical analysis on gathered information from the stakeholders of a supply chain firm. Important results in the form of tables obtained from factor analysis are shown in the following paragraphs. Sample correlation table from statistical analysis is shown in Table 3 with respect to the seven factors considered above.

In the present case, 2 levels of factor loading are considered. More loadings can be considered on a case to case basis. This methodology enables the firms to take

Table 3 Sample correlation coefficient for 7 factors

Factor	F1	F2	F3	F4	F5	F6	F7
F1	1.000	-0.833	0.289	0.255	0.272	−0.498	0.156
F2	−0.833	1.000	0.000	0.000	0.091	0.498	0.156
F3	0.289	0.000	1.000	0.627	0.288	0.048	0.629
F4	0.255	0.000	0.627	1.000	0.224	0.395	0.305
F5	0.272	0.091	0.288	0.224	1.000	−0.080	0.725
F6	-0.498	0.498	0.048	0.395	−0.080	1.000	0.198
F7	0.156	0.156	0.629	0.305	0.725	0.198	1.000

Table 4 Results after 2 levels of factor loading

Factor	Loading-1	Loading-2	Communality
F1	**0.705**	0.571	0.8
F2	−0.630	**−0.709**	0.9
F3	**0.492**	−0.491	0.5
F4	**0.590**	−0.510	0.6
F5	0.617	**−0.655**	0.8
F6	−0.614	**−0.720**	0.9
F7	0.721	−0.775	1.1

decision on the set of services out of many to be adopted for the firms and thereby minimize the cost of implementing cloud services to manage supply chain networks. On factoring at various levels, and inspecting the communality values, key factors can be identified.

In order to understand the factors more critically, a Factor Analysis method is adopted to determine the key factors out of many that will help firms to optimize and plan their resources. Factoring helps in finding out groups of variables with similar characteristic that a firm will benefit from. The results of factoring after 2 levels is shown in Table 4.

After two loadings, it is found that key factors are F2, F5 and F6 corresponding to Pricing, Collaboration and Scalability.

More combinations of cloud services can be taken up for grouping with the help of factor analysis as and when more cloud services and applications get developed.

In addition to factor analysis, a detailed analysis on multiple criterions with respect to the key factors is done with the help of Analytic Hierarchy Process (AHP). In the AHP analysis, pair wise comparison of the key factors can be done and the weighted scores of each factor are computed. Influence of one factor on the other factors and the proportion of contribution of each factor can be determined. This will help firms to get clear idea about the budgeting of their resources commensurate to the contribution of each factor.

5.2 Multicriteria AHP Analysis

From the list of key factors identified, the next step is to map the key factors with hierarchy levels of the supply chain network firm. Each level is assigned with the factors and categories encompassing the major components of a CSCN. A study by Prof. Thomas L. Saaty on multiple criterions decision making using Analytic Hierarchy Process (AHP) is helpful to focus on important aspects of a problem especially when different opinions are considered with respect to decision alternative. The three factors identified above and three stakeholder categories of supply chain firm is considered for the purpose of determining the weightages.

Table 5 S-C-P consistency result from AHP

	S	C	P		
S	1.00	1.00	0.11		
C	1.00	1.00	0.17		
P	9.00	6.00	1.00		
Sum	11.00	8.00	1.28	SH1 scores	Consistency measure
Normalized comparison score					
S	0.091	0.125	0.087	**0.101**	3.006
C	0.091	0.125	0.130	**0.115**	3.006
P	0.818	0.750	0.783	**0.784**	3.044
			Eigen value		3.018
			Consistency ratio (CR)		0.02

Table 6 Final AHP Scores for S-C-P

	S	C	P	Factor weight
SH1	0.101	0.115	0.784	0.087
SH2	0.065	0.199	0.735	0.128
SH3	0.067	0.794	0.139	0.785
Proportion-score	0.070	0.659	0.271	1.000

Categories: Stakeholder-1 (SH-1), Stakeholder-2 (SH-2), Stakeholder-3 (SH-3)
Factors: Scalability (S), Collaboration (C), Pricing (P)

Pair wise comparison is done between the two key factors using AHP and checked for consistency as indicated in Table 5 for stakeholder-1.

Similar computations for other stakeholders are done and overall key factor weightages are obtained. The final score model is shown in Table 6, which indicates the proportion of each key factor of Cloud Supply Chain Network.

From the final scores, it is seen that out of three key factors identified from factor analysis, from the Stakeholders point of view the Collaboration factor contributes more (65.9%) followed by Pricing factor (27.1%) and then Scalability factor (7.0%).

6 Conclusions and Scope for Further Research

From the above analysis, various results are obtained. The conclusions are drawn from the results in the context of the research problem stated above. It is also necessary to see that there is a scope for further research in this domain of cloud supply chain networks. Various conclusions and the further scope of research are discussed below:

(i) From the practical case considered with respect to CSCN factors namely Process Efficiency (F1), Pricing (F2), Investment (F3), Convenience (F4), Collabora-

tion (F5), Scalability (F6), and Software Cost (F7), factor analysis resulted in three key factors. The key factors are F2, F5 and F6 corresponding to Pricing, Collaboration and Scalability. It indicates that the firm can focus on these three factors out of many factors for building up their CSCN processes. Once it is done, firms can then adopt similar method by excluding these two factors and analyzing other factors and again identify key factors. In this manner a firm can implement the CSCN services stage by stage. This will enable firm to plan their budget and other resources and deploy resources in stages corresponding to these key factors. Subsequently firms can focus on investing on other factors depending on the stages of implementation. Hence firms can avoid unnecessary investments on factors that are not key factors.

(ii) The three factors identified above and three stakeholder categories of supply chain firm is considered for the purpose of determining the weightages. Three categories considered in the analysis were Stakeholder-1 (SH-1), Stakeholder-2 (SH-2), Stakeholder-3 (SH-3). Three key factors obtained from factor analysis were Scalability (S), Collaboration (C), Pricing (P). AHP analysis was done and the results were found to be in good agreement with the results obtained from BPMSG a decision tool used for pair wise comparison and setting priorities. This is a very useful information for firms planning to budget their resources. From the analysis results it is found that from Stakeholders point of view the Collaboration factor contributes more (65.9%) followed by Pricing factor (27.1%) and then Scalability factor (7.0%). These proportions vary from firm to firm and can be evaluated by considering the actual details of the SCN firm and their stakeholders. A beta function f(β) based on the proportions is useful for making various decision in supply chain network.

$$f(\beta) = 0.07\,\beta_1 + 0.659\,\beta_2 + 0.271\,\beta_3$$

In the above function, β1 corresponds to Scalability, β2 corresponds to Collaboration and β3 corresponds to Pricing factors respectively. The beta function can be used by the SCN firms in arriving at strategic decisions pertaining to investments, budgeting, resource allocations and many more. With more and more developments taking place in the cloud computing services, it is possible to upgrade the information by incorporating the factors in the analysis and redesigning the hierarchy and deriving the revised proportion.

(iii) From the methodology discussed in the above sections, it is also possible to map the supply chain network nodes with relevant SCN firm and its hierarchy to analyze the Bullwhip effect. There is a scope to analyze the effect using a beta function derived as explained in the above section. A beta function will enable SCN firms to identify critical proportions of the cloud computing characteristics in SCN that will minimize the effect and thereby increase the performance efficiency of the SCN firms.

(iv) Over years, various information technologies have been developed and today the world looks at cloud computing techniques as one of the major develop-

ments in IT that can facilitate quick decision making. As requirements of various firms changes over a period of time it offers tremendous scope for research. There is a scope to design information flow models for such requirements. A model for information flow is a construct of the important factors representing various important benefits that a cloud computing technology will offer. Information flow models are helpful for generating quick information that will benefit the supply chain network in increasing its efficiency.

(v) Another important consideration for research is the collaboration and coordination problems in CSCN. Various nodes of a collaboration network can be encapsulated in the information flow model and analysis can be done to determine the key contributors of CSCN. The results can help achieve better performances in collaborating SCN firms.

(vi) With the emergence of newer concepts of Fog and Mist computing services, it gives rise to opportunities to do a research. Fog and Mist computing services at decentralized branches of the SCN can be researched to check if the efficiency improves.

There are many factors that influence the decision making in a supply chain leading to successful operation of various activities in a supply chain. Hence it is important that the information is available instantaneously for making quick decision. Quick decisions help in fulfilling customer demands quickly. With the existing vendors catering to the regular supplies, new partners offering value-added services need to be integrated. This also causes a change in the business relationship from long term contracts to small short-term contracts. This leads to an event driven management for dynamically creating and delivering individualized products and services. Developments in the business world and information technology have enabled decision makers to have a sharper focus on various issues. Globalization of the supply chain network activities in many companies has led to adoption of cloud computing services.

Acknowledgements The authors extend their heartfelt gratitude to the Editor, Prof Himansu Das for providing an opportunity to write this paper. The authors thank him for giving valuable suggestions and comments to improve the content of this paper.

References

1. Ahronovitz, M., Dustin Amrhein et al.: Cloud Computing Use Cases, White Paper Produced by Use Case Discussion Group, Version 4.0, pp. 1–67 (2010)
2. Sahoo, A.K., Das, H.: Energy efficient scheduling using DVFS technique in cloud datacenters. Int. J. Comput. Sci. Informat. Technol. Res. **4**(1), 59–66 (2016)
3. Gowda, A.B., Subramanya, K.N.: The Influence of variables on designing a cloud supply chain network: a factor analysis approach. IUP J. Supply Chain Manag. **XII**(3), 35–49 (2015)
4. Gowda, A.B., Subramanya, K.N.: An analysis of the benefits of cloud services for supply chain using analytic hierarchy process. IUP J. Comput. Sci. **9**(4), 31–45 (2015)
5. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud Computing for Optimization: Foundations, Applications, and Challenges, vol. 39. Springer (2018)

6. Calheiros, R.N., Masoumi, E., Ranjan, R., Buyya, R.: Workload prediction using ARIMA model and its impact on cloud applications QoS. IEEE Trans. Cloud Comput. **3**(4), 449–458 (2015)
7. Ferguson, D.F., Hadar, E.: Optimizing the IT business supply chain utilizing cloud computing. In: 8th International Conference and Expo on Emerging Technologies for a Smarter World (CEWIT). pp. 1–6. IEEE Conference Publications (2011)
8. Kerschbaum, F., Schroepfer, A., Zilli, A., et.al.: Secure collaborative supply chain management. IEEE Comput. Soc. 38–43 (2011)
9. Fischer, F., Turner, F.: Business Operations—Systems Perspectives in Global Organizations [DDBA-8110-7], pp. 1–18. Walden University (2009)
10. Gribben, G.: BAPCO eProcurement, Cloud Computing the Future of ICT Delivery, pp. 1–18 (2010)
11. Ghaffari, M., Javadian, N., Narimani, F.: An improvement approach for bullwhip effect in multi-product and multi stage supply chains. In: 6th International Conference on Information Systems Logistics and Supply Chain, 1–4 June 2016, pp. 1–8 (2016)
12. Erbes, J., Reza, H., Nezhad, M., Graupner, S.: HP Laboratories, From IT Providers to IT Service Brokers: The Future of Enterprise IT in the Cloud World, GS (2012)
13. Jamshidi, P., Ahmad, A., Pahl, C.: Cloud migration research: a systematic review. IEEE Trans. Cloud Comput. **1**(2), 142–157 (2013)
14. Leukel, J.; Kirn, S., Schlegel, T.: Supply chain as a service: a cloud perspective on supply chain systems, abstract. IEEE Syst. J. 16–27 (2011)
15. Lindner, M., McDonald, F., McLarnon, B., Robinson, P.: Towards automated business-driven indication and mitigation of VM sprawl in Cloud supply chains, May 2011, abstract. In: IEEE International Symposium on Integrated Network Management, pp. 1062–1065 (2011)
16. YiPeng, L.: The Impact of "Cloud Computing"—based Information sharing on supply chain, management of E-commerce and E-Government (ICMeCG) abstract. In: Fifth International Conference, pp 173–175. IEEE Conference Publications
17. Maik Lindner, M. Lindner, F.G. Marquez, C. Chapman, S. Clayman, D. Henriksson, and E. Elmroth. (2010), The Cloud Supply Chain: A Frame work for Information, Monitoring, Accounting and Billing, 2nd International ICST Conference on Cloud Computing, in press, Springer Verlag, pp. 5
18. National Institute of Standards and Technology: US Department of Commerce, Special Publication 800–145, The NIST Definition of Cloud Computing, Recommendations of NIST—Peter Mell and Timothy Grance, Computer Security Division, NIST, Baithersburg, MD 20899–8930, p. 2 (2011)
19. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)
20. Pereira, L.A., Mamani, E.L.C., Santana, M.J., Santana, R.H.C., Northon Nobile, P., Monaco, F.J.: Non stationary simulation of computer systems and dynamic performance evaluation: a concern based approach an case study on cloud computing. In: 27th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) on IEEE Conference, 17–21 Oct 2015, pp. 130–137 (2015)
21. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Mankodiya, K., Kumar, V., Das, H.: MistGIS: Optimizing Geospatial Data Analysis Using Mist Computing, Progress in Computing, Analytics and Networking, pp. 733–742., Springer, Singapore (2018)
22. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., Lenka, R.K., Mishra, B.S.P., Das, H., Mankodiya, K.: Fog Assisted Cloud Computing in Era of Big Data and Internet -of-Things: Systems, Architectures and Applications, Cloud Computing for Optimization: Foundations, Applications and Challenges, pp. 367–394. Springer, Cham (2018)
23. Hashmi, S.I., Clerc, V., et al.: Using the Cloud to Facilitate Global Software Development Challenges, GS (2011)
24. Santra, S., Dey, H., Majumdar, S., Jha, G.S.: New simulation toolkit for comparison of scheduling algorithm on cloud computing. In: International Conference on Control, Instrumentation, Communication (2014)

25. Tsao, H.J., Parikh, S., Ghosh, A.S., Pal, R., Ranalkar, M., Tarapore, H., Venkatsubramanyan, S.: Streamlining grain supply chains of India: Cloud computing and distributed hubbing for wholesale-retail logistics. In: IEEE International Conference on Service Operations and Logistics and Informatics (SOLI), pp. 252–257 (2010)
26. Ying, X., Zhou, L.: Measuring Bull whip effect in a single echelon supply chain using fuzzy approach. Int. J. Innovat. Manag. Technol. **3**(5), 494–498 (2012)

Parallel Computation of a MMDBM Algorithm on GPU Mining with Big Data

S. Sivakumar, S. Vidyanandini, Soumya Ranjan Nayak and S. Sundar

Abstract Big data is the collection of data sets which are large and complex in nature. It contains structured and unstructured types of data. For example, Financial Services, Retail, Manufacturing, Healthcare, Social network (Twitter, Fackbook, Linkedin and Google), Digital pictures and Videos. To extract useful data from big data, several classifiers like SLIQ, SPRINT, MMDBM are used. Among this one of the fast classifier is the Mixed Mode data Based Miner (MMDBM) using Graphical Processor Unit (GPU) mining. This classifier describes the outline of parallel computing with high performance, using radix algorithm for multicore GPUs, by taking a program presented by Compute Unified Device Architecture (CUDA). The classifier can deal with both categorical and numerical attributes in a simple manner. The classification method handles big data with huge number of attributes by taking it from the medical data base. This can be parallelized on GPU to get high-speed and better performance than CPU-Radix sort algorithm. We proposed the parallelized Radix sort algorithm on GPU computing using CUDA platform developed by NVIDIA Corporation. In this chapter, we discuss the performance of fast classifier method and radix algorithm to relate the processing time of MMDBM, SLIQ CPU with GPU

S. Sivakumar · S. R. Nayak (✉)
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur 522502, India
e-mail: nayak.soumya17@gmail.com

S. Sivakumar
e-mail: sivaiit7@gmail.com

S. Vidyanandini
Department of Mathematics, SRM Institute of Science and Technology, Chennai 603203, India
e-mail: vidhyanandhini.maths@gmail.com

S. Sundar
Department of Mathematics, Indian Institute of Technology Madras, Chennai 600036, India
e-mail: slnt@iitm.ac.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*, Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_6

computing and computed acceleration ratio (Speed-up) time. Also, The classifiers [SLIQ, SPRINT, MMDBM] are evaluated and compared with CPU and GPU. GPU provides quick and accurate results with least processing time and supports real time applications.

Keywords Classification · GPU mining · Decision Tree · Radix sort

1 Introduction

NVIDIA corporation has developed a unique software called Compute Unified Device Architecture (CUDA) [1] which enables developers to straight with the GPU and run programs on them. In this way productively using the upside of parallelization. CUDA is certifiably not another programming language, rather an expansion to C with GPU-specific orders, alternatives and activities. Programs written in CUDA are compiled by *nvcc* compiler and can be run just on NVIDIA's GPU's [2]. A CUDA program can be kept running on a few number of processors core and just the number of processor requirements are to be known to the run time system. The attributes of CUDA program comprises of two sections: the principle part of the program which executes serially on the CPU (host), and the kernal, called as the main program, is executed in parallel on the GPU (device). Host work lie printf cannot be called from the kernal. Graphic Processing Unit (GPU) accessible on product video adapters has developed into exceptionally parallel, multithreaded, many-core processor. These GPUs have good computational power and additionally high memory transfer speed that can be exploited by general purpose in the high-performance applications. These programmable GPU are otherwise called general purpose graphic processing units (GPGPU, from now onward we will utilize term the GPU). GPU is expert in compute-intensive, exceedingly parallel computation simply like graphics rendering is finished. GPU depends on SIMD architectural model and used by data-parallel programming model [1, 2]. As of late, the GPU on graphics processing unit has turned out to be well known be-cause for the high parallelization and powerful computing ability of oat point. A few archives show the computing intensity of GPU, which would now be able to incomprehensibly surpass a CPU [3, 4]. NVIDIA Corporation, market leader in GPU market, presented a general purpose parallel computing structure in November 2006, to bridge the computing capacities of their high-end GPUs. CUDA depends on another model of parallel programming and direction set of design,it uses the parallel computing engine in GPUs to work out numerous computational issues that is complex in a more effective manner than compared to CPU.

CUDA accompanies a software background that allows developers to utilize C programming which is an high-level language. Different languages, for example, FORTRAN, C++, OpenGL, and DirectX will be supported in future. Data mining incorporates different advances, for example, classification, association rule mining, and clustering.

Classification is one of the vital problem in the field of Data Mining and Knowledge Management that has been examined broadly studied over the years [5, 6]. In this algorithm, we have given a record or the input data, called the trail database, in which each record contains a few characteristics (attribute). An attribute can deal with either a categorical or a numerical attribute. In ordered domain if the value of an attribute is ordered, then the attribute value is known as numerical attribute (e.g., Weight,Age, Sports, Sleep and Drink). If the vales are of an attribute is an unordered domain, then the attribute known as categorical attribute (e.g., Sex, BP) [7, 8].

Among the categorical attributes if one attribute is assigned as a class attribute, then it is known as class names to which each record belongs. The goal of classification is to observe the datainput and to build up an exact description or model for every class utilizing the current data features.After building the class model, it can be utilized.

The role of classification is to utilize the training data set to build a model of the class name with the main goal, that it can be utilized to arrange new data whose class names are obscure [9–11]. We show a parallelized decision tree algorithm and GPU-Radix sort calculation in view of CUDA. Additionally, we will talk about the efficiency of parallel Radix sort method on GPU and a study on the comparison of computational acceleration ratio (Speed-up) and productivity of GPU with CPU for the fast classifier algorithm [12–14]. GPU—radix sort provides fast and accurate outcomes with less processing time.

2 Basic Definitions: Big data

Big data is a group of data sets which are complex and large that becomes complicate to process while using it in available database. The objection contains capturing, storage, searching, sharing, curation, analysis, visualization and classification.

The leaning to big data sets is due to the extra data obtainable from analysis of a single large set of related data, as related to distinct smaller sets with available total amount of data, allowing relationships to be found in spotting business trends, prevent diseases, concluding quality of research, link permissible citations, defining real-time roadway traffic conditions and combat crime (Wikipedia).

2.1 Characteristics of Big Data

Big Data can be defined by the four Vs: **Velocity, Volume, Variety, and Value**. This shows whether to include Big Data to our data architecture is shown in the Fig. 1.

Volume

In past, collected data by Organizations from various resources like social media, business transactions and machine related data from sensor created problem in

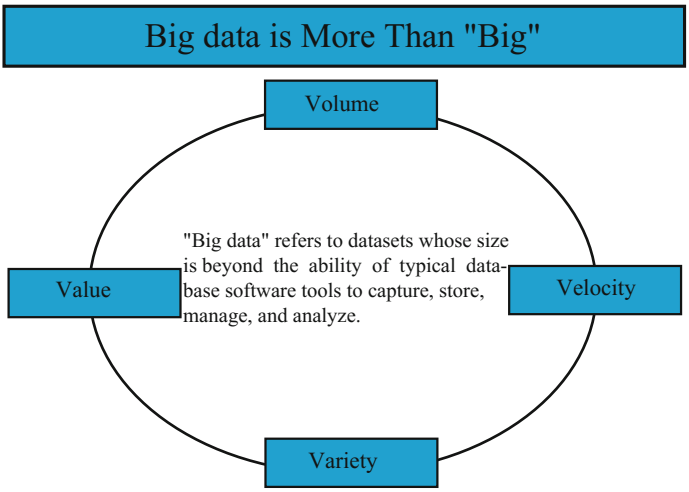


Fig. 1 Characteristics of big data

storage level. But current technologies like Hadoop reduced burden by storing large amount of data [15].

Velocity
Big data relates with real-time challenges. In context, the processing speed and the data generated to meet different challenges and the demand are focused only on the growth and development of data.

Variety
There are many natures of data which is available in different formats. That makes people who examine the result and efficiently use the resulting insight. Big data includes images, text, video, audio and completes the remaining pieces using data fusion.

Value
Data has real value which should be discovered. There are measurable and analytical techniques to derive the value from data by learning from the sentiment. To make a applicable over by location or In recognizing piece of equipments about to fail.

3 Solutions and Challenges of Big data

1. Location of Big Data Sources-Commonly Big Data are put away in different areas.
2. Volume of the Big Data-size of the Big Data develops persistently.
3. Hardware assets RAM quantity.

4. Privacy-Medical information, bank exchanges.
5. Having domain knowledge.
6. Getting significant data.

Solutions

1. Parallel processing programming (ex: GPU Computing).
2. A productive stage for computing won't have concentrated information storage rather than that platform will be appropriated in big scale storage.
3. Regulating access to the information.

4 Data Mining Techniques

This chapter is mainly based on the three methods, namely SRINT, SLIQ and MMDBM. Classification method known as SPRINT that eradicate all memory limits that confine Decision Tree (DT) techniques, which shows that this algorithm is scalable and quick [10, 16].

SLIQ algorithm is based on a DT algorithm, it deals together with categorical and numerical data sets.This algorithm builds a tree that is an accurate tree. It utilizes an enhanced sorting system to diminish the cost of assessing numeric attributes in tree building phase.The above method is coordinated with a breadth-first tree developing procedure to improve classification of disk-resident data sets. SLIQ utilizes a quick method to identify splits for attributes,which is categorical in nature [8, 15, 17]. On the other hand,decision tree algorithm is used for taking care of categorical and numerical data in huge datasets known as (MMDBM). This technique utilized, deals huge data set with substantial arrangement of many attributes or data, and obtaining effective all the numeric attribute the mid points are presented. This is incompletely separated into 2 areas, first one predictive classifier with gives a point by point portrayal of our method and second one is OOD (Object Oriented Design), which gives the implementation of OOD in our method and depiction of created front-end for this method [7, 18].

In this chapter, the author contrasts their method with familiar SLIQ, SPRINT and MMDBM method. We outline that decision tree algorithm SPRINT, SLIQ and MMDBM, which have accomplished compactness, best accuracy and efficiency for big data set [7, 8, 10].

4.1 Sorting Algorithms on GPU

Sorting algorithm is a calculation of building block of basic significance and is a standout among the most generally considered algorithmic problems. All algorithms depend on the accessibility of effective sorting schedules as a reason for their own proficiency. sorting itself is of more significance in applications running from database

system to computer graphics, and several different methods can be easily expressed in terms of sorting algorithm. In this manner vital to give effective arranging schedules for any platform, and architectures. Computer predicts that, there is a proceeding which needs to analyse effective sorting method on system architectures [17, 19].

The significance of sorting has prompt the plan of efficient parallel sorting algorithms to be executed in different parallel architectures. Sorting algorithms are naturally parallel as they are formalized as far as physically parallel comparator devices. Algorithms related to sorting are singularly attractive on platforms where the expanding of reliant data is highly impossible, because of the links between correlations, which are stable regardless of the input. Alternative normal way to deal with parallel sorting algorithm is to separate input sequence into pieces, which can be sorted individually by the accessible processors. The sorted continuance should then be combined to deliver the outcome.

4.2 Radix Sort

Radix method is one of the powerful sorting algorithms [20–22]. This algorithm is fast especially for a huge problem size. It is frequently as possible utilized among the most productive for sorting algorithm in small keys. This sorting methodology expects the keys are denoted as d-digit numbers with respect to $radix - r$ notation. On binary PCs, it is most likely going to accept that, the radix sort $r = 2b$ and the keys are an integral multiple of b bits in length. The sorting method itself includes of d passes, it considers the ith digits of the keys all together from least to most significant digit [23, 24]. In every pass, the input flow is sorted with the value to digit i of the keys, it is essential that this sort be stable (i.e., it conserves equal digits with comparative sequence of keys).The radix method utilized inside every process of sort is generally a bucket sort or counting sort [20, 25, 26].

In each pass, every key can be mapped with one of r buckets. To calculate the output basis where the element ought to be written, can also be referred as the rank of the element, we should basically compute the quantity of elements already in the current bucket plus lower numbered buckets. After each element rank calculation, sorting method is finalized by passing the elements into the array output in the location controlled by their ranks. A parallel algorithm of radix sort is given underneath.

5 MMDBM Algorithm

Input: The attributes A is containing n number attributes $A = \{a_1, a_2, \ldots, a_n\}$ in parallel

Output: Count the node value and construction of the decision tree.

Data value were developed randomly in database.

Transfer the data from Device (GPU) to Host (CPU) (cudaMemcpy), dispatch the value in arrays.
Copy to GPU and radix sort for sorting the random all the attributes from data base inside the GPU.
Get the split point value of each attribute.
a_i is the attribute name and v_i is the split point value of each attribute,
For i = 1 to n // n is the number of the attributes node
    If<condition > Then C // C is class count
        This algorithm is using multiple If statements
        IF ($a_1 \leq v_1$) AND ($a_2 \leq v_2$) AND,...,AND ($a_n \leq v_n$) THEN C.
            IF $a_i \leq v_i$ is true goto left side node traverse up to N number of the node THEN count the class values
            C = C + 1;
        else
            IF $a_i \leq v_i$ is false goto right side node traverse up to N number of the node.
            Count the class value or traversal node that already exist then update the appropriate class count value.
                C = C + 1;
            else
                Missing count value and update the class count value.
            M = M + 1;
            Missing class count values, if the same data exist then update the appropriate class values.
        End IF
    End IF
End For

6 Proposed Method

Today processing time of parallel mining is minimized by GPU mining in classification with less time. This method comes from energy consumption and significant power [27]. In this chapter, we propose a universal parallel mining method for development of DT algorithm on GPU mining. The reasons we preferred MMDBM decision tree method is by utilizing split point to apply on each node using radix sort algorithm. Finally, it counts the class values. The attribute list can be constructed in two models. The first model is transferring the available data to its matching list. Second model completes the data development. After constructing two models, the attribute lists must be sorted. There are well-defined CUDA library known as CUDPP. It has two inputs in sorting method, the first method is key array and the second method is value array that sort two 1-D arrays.

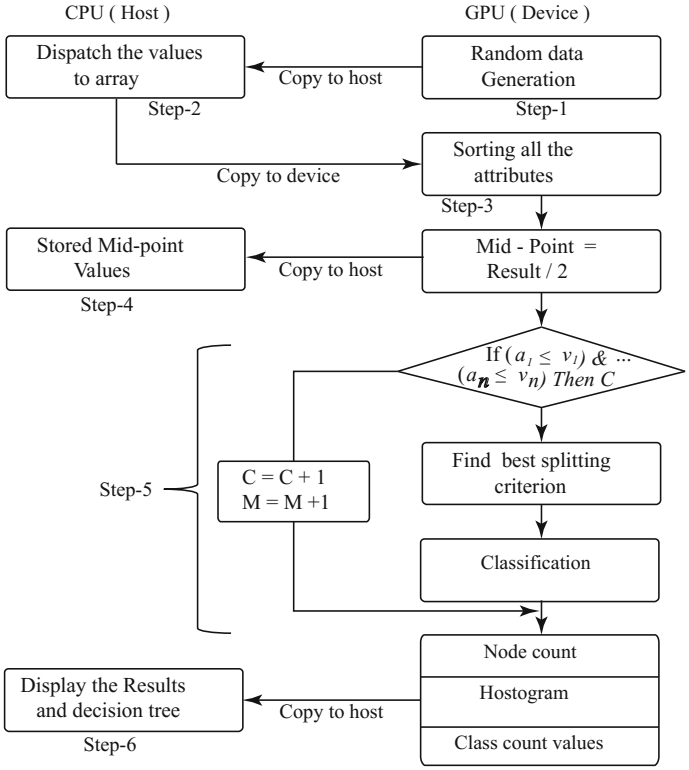


Fig. 2 Design for fast classifier mining algorithm

After key array sort, the significant array element value could change the situation as per its relating key element, called key value pair sorting.The key value pair is uniquely supported by sorting method of CUDPP, however we have two values for one (attribute for key is value field, rid and class values).

For every two values, we change the CUDPP method into one key. To get the higher performance. The kernel includes a grid of scalar threads. All thread has one unique identifier (thread ID) that partition the work within the threads.

In a grid, threads are organized as blocks denoted as cooperative thread arrays (CTAs), where a solitary CTA thread approaches a successive high-speed memory called the shared memory. We altered the sorting method from the cooperative thread arrays level to public interface level [3, 28].

The accompanying advances delineate the principle implementation step in our execution. Further, Fig.2 demonstrates the execution part done on the host and on the device, individually, and the data exchanges that occur among the host and the device (GPU).

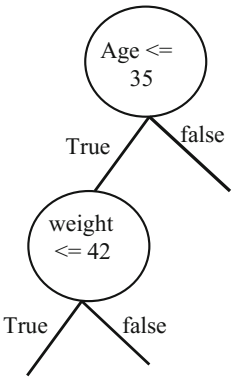


Fig. 3 Splitting point for node

Step by Step Process in Algorithm
The data has generated randomly in GPU device, it is transferred to the CPU host and stored in an array (step 1 and step 2). An array values are copied to GPU and all the numeric attributes values are sorted using radix sort algorithm (step 3). The sorting data values are transferred to CPU host and mid-point value of all the numeric attributes (step 4) is obtained. The mid value checks the condition of a record in each attribute value, if true condition goto left node and if condition is false goto right node. Figure 3 demonstrate splitting point of node.

Final class values are counted, and we calculate the histogram of node level and the condition is checked from top to bottom is called distributes (step5) (refer the tree). The final result has transferred to from GPU to CPU and calculate the acceleration ratio time of GPU mining.

6.1 Implementation of Algorithm for GPU

The pre-sorting process is completed and the mid-point value of each attributes is found.

$$\text{IF (age} \leq \text{mid-point) Then C.}$$

Age is the attributes variable name and values are in data sets table. If the condition is either true or false, the corresponding node data is passed to an- other node pointed as leaf from the corresponding parent node, which is called split point [7, 29].

6.2 Finding Split Points Code for GPU

By our algorithm, we find the mid-point by sorting the array elements and storing the middle element of the sorted array. There by we made only one pass in finding

the midpoint. Once the midpoint is calculated, the arrays are passed to the GPU classification function along with the midpoints to classify the records. The GPU code for finding the mid-points is given below.

```
Radix_sort<<< 1, SO >>> (d_values,t_values,1,d_split, d_e, d_f, d_t);
cudaMemcpy (sag, d_values, size, cudaMemcpyDeviceToHost);
printf(" \n\n MidPoint of the AGE is:\t");
MA=sag[SO /2];
printf("%d\n",MA);
cudaMemcpy (ad_values, wt, size, cudaMemcpyHostToDevice);
Radix_sort<<< 1, SO >>> (ad_values,at_values,1,ad_split,ad_e,ad_f, ad_t);
cudaMemcpy(swt, ar_values, size, cudaMemcpyDeviceToHost);
printf("\n\n MidPoint of the WEIGHT is: \t");
MW=swt[SO /2];
printf("%d \n", MW);
Radix_sort<<< 1, SO >>> (sld_values,slt_values,1,sld_split,sld_e, sld_f, sld_t);
cudaMemcpy (swt, slr_values, size, cudaMemcpyDeviceToHost);
printf(" \n\n MidPoint of the SLEEP is:\t");
MW=swt[SO/2];
printf("%d \n",MW);
Radix_sort<<< 1, SO >>>(spd_values,spt_values,1,spd_split,spd_e, spd_f, spd_t);
cudaMemcpy(spt, spr_values, size, cudaMemcpyDeviceToHost);
printf(" \n\n MidPoint of the SPORTS is:\t");
MW=swt[SO/2];
printf("%d\n",MW);
Radix_sort<<< 1, SO >>>(drd_values,drt_values,1,drd_split,drd_e, drd_f, drd_t);
cudaMemcpy(drt, drr_values, size, cudaMemcpyDeviceToHost);
printf(" \n\n MidPoint of the DRINKING HABBIT is:\t");
MW=swt[SO/2];
printf("%d\n",MW);
```

6.3 Best Split Point

The split has been applied to every node. After getting the mid- point values and scanning the attributes of the all records from connected data sets. The node can be classified using IF ($x_1 \leq v_1$) AND ($x_2 \leq v_2$) AND,…, AND ($x_n \leq v_n$) THEN C (class value) rule [35]. We have used two types of attributes, one is numeric and another one is categorical (male is 1 and female is 0). The splitting method used by the node be determined by the type of attribute, each attribute in the node creates two child nodes that are attached to the parent node, at each split point the histogram is calculated.

This process is based on n number of attributes. Finally we count the class value based on the class attribute (high BP, normal BP and low BP), already the exist-

Table 1 Scalability of the acceleration ratio times

Radix-sort with MMDBM algorithm	Number of records in seconds							
	100 × 100	200 × 100	300 × 100	400 × 100	500 × 100	600 × 100	800 × 100	1000 × 100
Generate random	0.090	0.210	0.320	0.410	0.510	0.650	0.850	1.20
Sorting	0.030	0.030	0.040	0.048	0.058	0.080	0.091	0.140
Classification time	0.510	1.010	1.530	2.040	2.590	3.080	4.120	5.21
CPU time	0.630	1.250	1.830	2.490	3.680	3.810	5.061	6.55
GPU time	0.510	1.010	1.530	2.040	2.590	3.060	4.920	5.180
Acceleration	1.235	1.237	1.196	1.2200	1.420	1.245	1.020	1.260

ing class count value is updated according to the appropriate class count value by $C = C + 1$. The process is completed in all the attributes, which is called distribution. Thus we calculate the histogram of each distribution. (refer the distribution of node count).

6.4 Acceleration Ratio for GPU

GPU performance GPU execution to examine the acceleration execution, an acceleration ratio (speed-up) γ is define by $\gamma = \frac{t_{CPU}}{t_{GPU}}$ where the complete processing time on the CPU, t_{CPU} contains the time at execution of the loop,the aggregate sum of the processing time on the GPU, t_{GPU} incorporates extra time of copying data between Device and Host in the interest of fairness [12, 30].
CPU Time = Sorting Time + Classification Time + Generate the random Values.
GPU Time = Data transfer from Host to Device and Device to Host.
Acceleration Ratio = CPU computation Time/GPU computation Time.
We calculated CPU, GPU and Acceleration ratio times (Table 1).

7 Experimental and Comparison

To test the effectiveness of our fast classifier algorithm, it has been implemented with GPUs radix sort. Now we consider medical database for blood pressure (BP), where data mining techniques are applied. Test has been carried out to evaluate and generate the random values, sorting, classification, CPU and GPU total processing times. The medical database for blood pressure where the radix sort is used to predict high risk, with every person having low BP, normal BP,high BP based on the class value.

Table 2 Distribution of the data collection

Classified node and travelled pat	Class count values				
	Low BP	High BP	Normal BP	Missing count	Total
Dist 1:1-3-6-13-27-55-111	394	450	391	17	1,252
Dist 2:1-3-6-13-26-52-104	390	421	381	19	1,211
Dist 3:1-2-5-11-22-44-89	373	392	353	18	1,136
Dist 4:1-2-4-9-18-36-72	393	472	372	22	1,259
Dist 5:1-2-4-8-16-32-64	318	456	302	7	1,083
Dist 6:1-3-7-15-30-60-120	411	435	396	12	1,254
Dist 7:1-2-4-9-19-39-79	259	482	401	15	1,157
Dist 8:1-2-4-8-17-35-70	447	493	427	11	1,378
Dist 9:1-2-5-11-23-47-94	487	448	450	13	1,398
Dist 10:1-3-7-15-31-62-124	421	493	401	9	1,324
Dist 11:1-3-6-12-25-50-100	452	418	432	17	1,319
Dist 12:1-2-4-9-19-39-78	302	482	352	15	1,151
Dist 13:1-3-7-15-31-63-126	415	470	405	13	1,303
Dist 14:1-3-6-13-27-54-108	496	428	476	14	1,414
Dist 15:1-2-4-8-16-33-67	427	393	444	20	1,284
Dist 16:1-2-4-9-19-38-77	482	377	462	13	1,334
Dist 17:1-2-5-11-23-46-92	397	482	395	6	1,280
Dist 18:1-3-6-12-24-48-96	417	481	407	18	1,323
Dist 19:1-2-5-10-20-40-81	418	423	427	23	1,291
Dist 20:1-2-5-10-21-43-86	391	401	391	7	1,190
Dist 21:1-3-7-14-28-56-112	305	482	305	18	1,110
Dist 22:1-1-3-7-14-29-59-118	490	491	421	9	1,411
Dist 23:1-3-6-12-25-51-102	293	487	388	5	1,173
Dist 24:1-2-4-9-18-37-75	292	441	225	7	965

7.1 Medical Data Set for BP

This algorithm is utilized as a part of two categorical attributes Sex and BP, Sex is male or female and BP is high BP, low BP and normal BP and five numerical properties, Age is describing the age of a man (years), Person weight (kilo grams), Sports activity of a man (1–10), Man consider Sleep on an average (0–24) and Drink is extent of drinking of a man. We have classified all attributes and obtained 24 distributions of the nodes and travelling path from thirty thousand records and checked the quantity of patients with high BP, low BP and normal BP. Every distribution is produced by various IF-Then rule. The above rule can be obtained gradually on predicted rules of the total class count, every distribution of the node count and travelled path (refer the Table 2).

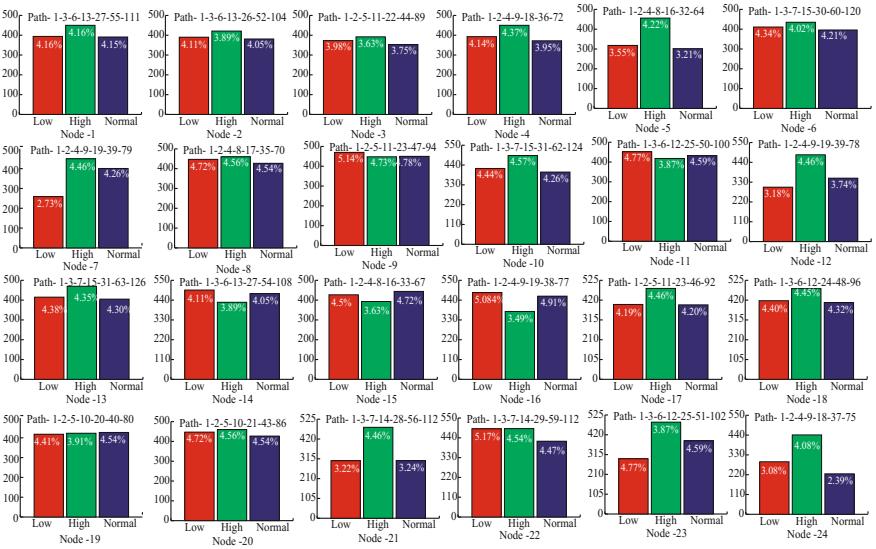


Fig. 4 Distribution of the node count values

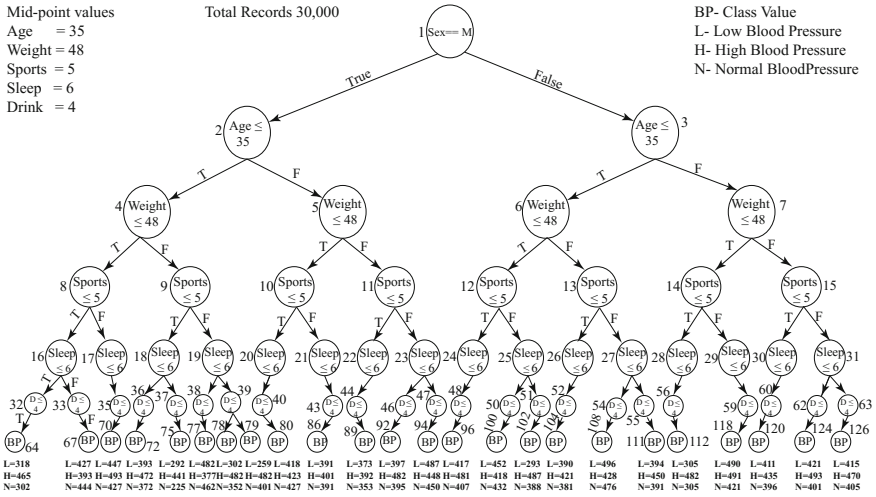


Fig. 5 Classification tree

The above classification tree is constructed to classify all the attributes to get the distribution of the 24 node count values and the travelling path from 30,000 records (refer the Fig.4). For the above given method, a big data set was created to check the high accuracy with minimum processing time. Figure 5 represent the classification tree for the medical data set. The supervised learning testicles were done with various amount of data provided to the program ranging from 20,000 to 10,00,000

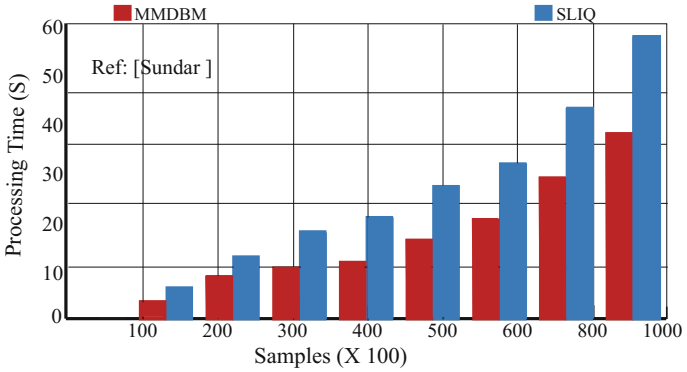


Fig. 6 Scalability of processing time in MMDBM and SLIQ using Java

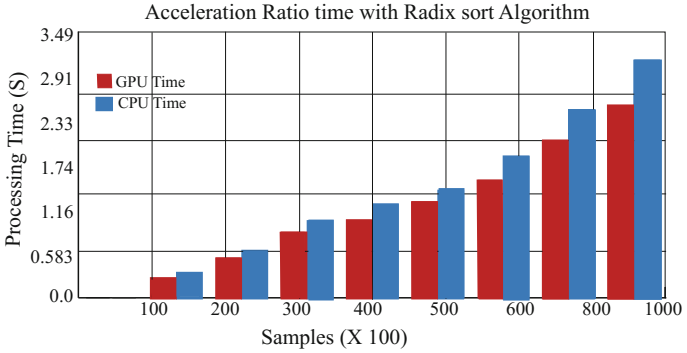


Fig. 7 Scalability of processing time in CPU and GPU

and classification is completed. With the increase in number of records. The above algorithm improves in prediction of accuracy. Figure 2 demonstrates the prediction rules achieved from the database. Figure 4 demonstrates the class distribution at every one of the nodes count values. Figure 6 Processing time Scalability in MMDBM and SLIQ using Java The results found in this paper is compared with the processing time of (SLIQ and MMDBM) Fig. 6 [7], which is mention above. when SLIQ and MMDBM (Fig. 6) classifier is compared with fast classifier mining algorithm, we get comparable results in the low error least processing time. However the error with SLIQ and MMDBM remained constant at larger values. In GPU, the classification time is larger than the SLIQ and MMDBM by almost 80%. The processing times for classification is shown in the Fig. 7. This test was done on a ubuntu 12.0 along with CUDA variant 5.0, The equipment platform comprises of an Intel Core i7-245M CPU, 2.50 GHz and 6 GB RAM. Summarized GPU features utilized as a part of the experiments. The GPU utilized is a NVIDIA GT525M card with 4095 MB.

8 Application

1. Healthcare associations can accomplish better insight into disease trends and patient medications.
2. Public sector agencies can get scam and different threats in real-time.
3. Applications of Multimedia data.
   - Face-To-Face class room Courses.
   - E-Learning Courses/Distance Education online courses.
   - Videos and Photos from social network.
4. Recommended framework.
5. Integration and mining of Bio data from different sources in Biological system by NSF (National Science Foundation).
6. Classifying the Big data stream in run time, by Australian Research board.

9 Conclusion

A Fast Classifier Algorithm has been programmed to radix sort algorithm on many core GPUs and has been tested using medical databases. This method can handle huge number of data set and attributes. The GPUs-radix sort algorithm gives magnificent scalability with the medical data sets that has been taken for analysis and exploring. The above problem was taken into consideration and tested for accuracy and the code has been provided. We discussed an efficient parallel radix sort algorithm on GPU and a study on the comparison of computational acceleration ratio (Speed-up) and efficiency of CPU with GPU for the fast classifier. A case study is used to compare the classifier with an existing CPU-Radix sort classification techniques and GPU-Radix sort provides quick and accurate results with less processing time and supports real time applications. Therefore, GPU mining proves its superiority over SLIQ and MMDBM classifier.

References

1. NVIDIA Corporation.: NVIDIA CUDA Programming Guild, 3.2 edn. (2010)
2. NVIDIA Corporation.: NVIDIA CUDA Best Practices Guild, 3.2 edn. (2010)
3. Chiu, C.C., Luo, G.H., Yuan, S.M.: A decision tree using CUDA GPUs, iiWAS '11. In: Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, pp. 399–402
4. Nayak, J., Naik, B., Jena, A. K., Barik, R. K., & Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–2. Springer, Cham (2018)
5. Shapiro, G.P., Frawley, W.J..: Knowledge Discovery in Databases. AAAI/MIT Press (1991)
6. Breiman, L. et al.: Wadswort, Classification and Regression Trees, Belmont (1984)

7. Sundar, S., Srikanth, D., Shanmugam, M.S.: A new predictive classifier for improved performance in data mining: object oriented design and implementation. In: Proceedings of the International Conference on Industrial Mathematics, pp. 491–514. IIT Bombay, Narosa, (2006)
8. Mehta, M., Agrawal, R., Rissanen, J.: SLIQ: a fast scalable classifier for data mining. In: Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology, pp. 18–32 (1996)
9. Agarwal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of International Conference Very Large Data Bases, pp. 487–499 (1994)
10. Shafer, C.J., Agrawal, R., Mehta, M.: SPRINT: a scalable parallel classifier for data mining. In: Proceedings of the 22th International Conference on Very Large Data Bases, pp. 544–555 (1996)
11. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Das, H.: Mistgis: optimizing geospatial data analysis using mist computing. In: Progress in Computing. Analytics and Networking, pp. 733–742. Springer, Singapore (2018)
12. Panchatcharam, M., Sundar, S., Vetrivel, V., Klar, A., Tiwari, S.: GPU computing for meshfree particle method. Int. J. Numer. Anal. Model. Ser. B **4**, 394–412 (2013)
13. Panigrahi, C.R., Tiwary, M., Pati, B., Das, H.: Big data and cyber foraging: future scope and challenges. In: Techniques and Environments for Big Data Analysis, pp. 75–100. Springer, Cham (2016)
14. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud Computing for Optimization: Foundations, Applications, and Challenges, vol. 39. Springer (2018)
15. Reddy, K.H.K., Das, H., Roy, D.S.: A Data Aware Scheme for Scheduling Big-Data Applications with SAVANNA Hadoop. Futures of Network. CRC Press (2017)
16. Das, H., Naik, B., Behera, H.S.: Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach. Progress in Computing. Analytics and Networking, pp. 539–549. Springer, Singapore (2018)
17. Sarkar, J.L., Panigrahi, C.R., Pati, B., Das, H.: A novel approach for real-time data management in wireless sensor networks. In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, pp. 599–607. Springer, New Delhi (2016)
18. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., Mankodiya, K.: Fog assisted cloud computing in Era of big data and internet-of-things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations. Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
19. Kar, I., Parida, R.R., Das, H.: Energy aware scheduling using genetic algorithm in cloud data centers. In International Conference on IEEE Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3545–3550), Mar 2016
20. Satish, N., Harris, M., Garland, M.: Designing efficient sorting algorithms for many core GPUs. In: Proceedings of IEEE International Symposium on Parallel & Distributed Processing (2009)
21. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of Intrusion Detection Using Data Mining Techniques. Progress in Computing. Analytics and Networking, pp. 753–764. Springer, Singapore (2018)
22. Das, H., Jena, A. K., Nayak, J., Naik, B., Behera, H.S.: A novel PSO based back propagation learning-MLP (PSO-BP-MLP) for classification. In: Computational Intelligence in Data Mining-Volume 2, pp. 461–471. Springer, New Delhi (2015)
23. Dusseau, A.C., Culler, D.E., Schauser, K.E., Martin, R.P.: Fast parallel sorting under LogP: experience with the CM-5. IEEE Trans. Parallel Distrib. Syst. **7**(8), 791–805 (1996)
24. Grand, S.L.: In: Nguyen, H. (ed.) Broad-Phase Collision Detection with CUDA, in GPU Gems 3. Addison-Wesley Professional, ch. 32, pp. 697–721 (2007)
25. Coremen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithm, 2nd edn. MIT Press (2001)
26. Zagha, M., Blelloch, G.E.: Radix sort for vector multiprocessors. In: Proceedings of ACM/IEEE Conference on supercomputing, pp. 712–721 (1991)
27. Nasridinov, A., Lee, Y., Park, Y.-H.: Decision tree construction on GPU: ubiquitous parallel computing approach. Computing **96**, 403–413 (2014)

28. Harris, M.: CUDPP:CUDA Data-Parallel Primitives Library 1.1.1, NVIDIA, UCDAVIS,29 (2010). http://code.google.com/p/cudpp/
29. AKGÖEK, Ö.: A rule induction algorithm for knowledge discovery and classification. Turk. J. Electr. Eng. Comput. Sci. **21**, 1223–1241 (2013)
30. Sundar, S., Panchatcharam, S.: Finite pointset method for 2D dam-break problem with GPU acceleration. Int. J. Appl. Math. **25**, 547–557 (2012)
31. Michile, D., Spiegelhalter, D.J., Taylor, C.C.: Machine Learning Neural and Statistical Classification. Ellis Horwood (1994)
32. Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., Swami, A.: An interval classifier for database mining application. In: Proceedings of the VLDB Conference, pp. 560–573 (1992)
33. Nasridinov, A., Lee, Y., Park, Y.-H.: Decision tree construction on GPU: ubiquitous parallel computing approach. Computing **96**, 403–413 (2014)
34. Sarkhel, P., Das, H., Vashishtha, L.K.: Task-scheduling algorithms in cloud environment. In: Computational Intelligence in Data Mining, pp. 553–562. Springer, Singapore (2017)
35. Sivakumar, S., Nayak, S.R., Vidyanandini, S., Palai, G.: An empirical study of supervised learning methods for breast cancer diseases. Int. J. Light Electron Opt. **175**, 105–114 (2018)

Data Analytics of IoT Enabled Smart Energy Meter in Smart Cities

Kiran Ahuja and Arun Khosla

Abstract In current energy production and distribution system, a smart energy meter has been a significant conceptual paradigm. There is a dire requirement to make energy usage more efficient and effective due to limited nonrenewable energy resources and renewable energies (REs) available at high cost. It creates a critical environment for future economic developments and social improvements such as smart cities. In recent years, numbers of smart meters are being installed in residential areas and other sites of smart cities. Smart meters are capable to provide numerous informative recordings of electricity consumption along with accurate processing of billing, Automated Meter Reading (AMR) data processing, detection of energy theft and early warning of blackouts, fast detection of turbulences in energy supply, real time pricing updates, and Demand Response (DR) system for energy saving and efficient usage of energy generated. To take full benefit of smart metering intelligence, numbers of technical issues are required to be addressed. The major concern is to work with very large volume of data. There is a need to develop efficient data fusion and integration techniques. Numerous big data integration and analytics engines are required, which can perform tasks such as outage management, asset management and fault detection especially in case of DR system, customer segmenting, load forecasting and targeting. Data analytic approaches transform volume of data into actionable information for consumers, utilities and authorities. Although numerous analytical algorithms are available, which can process huge volume of data, but many of these are not capable to complete task sufficiently. A few research procedures have been accounted for on investigating streaming data, but still needs a lot of work to be done in making these commercially reasonable. In this chapter, smart energy meters' data analytics framework is proposed by employing latest data processing techniques/tools along with gamification approach for enhancing consumers' engagement. Benefits of smart energy meter's analytics are also discussed

K. Ahuja (✉) · A. Khosla
Department of Electronics and Communication Engineering, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India
e-mail: askahuja2002@gmail.com

A. Khosla
e-mail: khoslaak@nitj.ac.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_7

for motivating consumers, utilities and stakeholders. Researchers, utilities, authorities can take benefits from proposed algorithm by planning their future action with supplementary participation of real time consumers due to gamification approach. By gamification, consumers' engagement improves and it alters their less sustainable behavior on a voluntary basis.

Keywords Smart meter · ICT · IoT · Gamification · Data analytic tools

1 Introduction

Traditional networks are connected with telecommunication technologies, Information and Communication Technologies (ICT) to make services more flexible, sustainable and efficient in smart cities to facilitate residents. The diverse components of smart city are smart health care, smart energy and smart transportation etc. These make cities smarter and more efficient. ICT is a transformation key from traditional cities to smart cities. Smart energy is one of the prerequisite components of smart cities. ICT along with traditional energy system make it smart energy. The various components of smart energy are smart grid, smart infrastructure, and smart meters with an appropriate utilization of ICT. The basic need of a smart energy system is to acquire the information smartly via smart infrastructure and accumulate the energy usage. Optimized consumption of energy is the instant need of moment due to carbon footprints, greenhouse gas emissions and global warming etc. The efficient utilization of smart metering, storage system and management will provide optimal energy consumption. Currently, smart meters are installed in number of residential areas as shown in Fig. 1 and other premises. If these are utilized properly, these can provide affluent data for analytics by recording electricity consumption. Smart meters make easier reading, billing and data processing. These are capable to detect energy losses in terms of fraud and cautioning of power outages before time, quick detection of disturbances in supply. Installation of smart meters makes possible real-time pricing schemes, and demand-response for efficient usage of energy generated and saving.

Smart meters make capable users to review their electricity usage timely by acquiring data. The basic components required for smart metering are shown in Fig. 2. When consumers are being aware about their electricity consumption, then they will improve savings and make energy efficient system. These activities are together known as demand end management, which directly benefits the consumers. The acquired data provides facility to understand consumer's profile, needs and behavior for expected outcomes (such as Outage Management (OM), Peak Load Management (PLM), and Power Quality Management (PQM)). Better consumer awareness is required for reducing energy consumption, which will directly affect the need of auxiliary power plants and generated greenhouse gases. By restricting and dropping electricity use in the midst of peak hours are cutting down the need of peaker plants which in general make higher carbon emissions.

Fig. 1 Smart energy meter deployed in residential area [20]

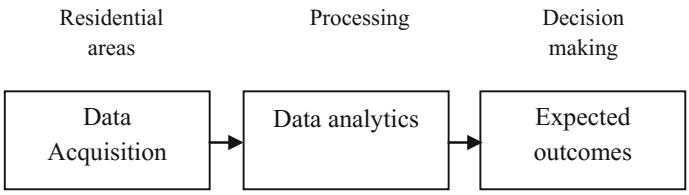


Fig. 2 Basic components required for smart meter data analytics

Load control aspect of smart meters makes enable switching (On/Off) of individual appliances as per need [1]. This feature is helpful to customers when the price of power is high and distributors can take its benefit when a network is near to its upper limit. The data procured from smart meters will be vital for market demands, forecasting load, planning operations, abrupt changes and disruptions by performing appropriate data analytics. There are numerous tools and algorithms are available for data analytics [2–5]. In this chapter, we propose a data analytics process in addition with gamification to change the behavior of user to achieve the targeted outcomes (such as energy efficiency, reduction of carbon footprints, change in unethical behavior of consumers etc.). In Sect. 2, background related to data analytic in the field of electricity/smart grids (SGs) is discussed. Proposed framework is described in Sect. 3. Comprehensive discussion of gamification approach is provided in Sect. 4. Conclusions are drawn in Sect. 5.

2 Related Work

The extensive deployment of smart meters will have serious privacy implications as these can unintentionally disclose thorough information about domestic activi-

ties. Researchers in [6] showed that even lacking of advance familiarity of domestic activities, it is feasible to extract consumption patterns from smart meters by employing statistical techniques. The patterns revealed a range of routine information which indirectly ceases the security. There is a need to design and implement privacy enhancing smart meter system which can allow utility to achieve their metering goals without any compromise with the privacy of customers [7]. However, in [8], authors focused on the importance of smart meters in load shifting to off-peak hours to take advantage of smart tariff. Ultimately, it reduces cost to customers for working during off-peak hours thus improving the reliability of the network.

Authors in [9] built a data signature database for smart meters data management by using different time resolution datasets. Data signature scan be employed to identify anomalies in different applications such as consumer energy consumptions, communication networks, grid operation, and control. Data signatures require data correlation among various data sources from transmission and distribution power grids in future. Researchers worked on the improvement of energy efficiency by customer's segmentation dependent upon consumption patterns. Energy customers' segmentation must be proper for better results. Samples were dispersed due to segmentation and created poor correlation among the parameters of electricity consumption. The choice of tariff type (non-flat tariff, dual-tariff) helped in deduction of the energy consumption. Their approach showed limited improvement due to lack of data sets and limited analysis period [10]. Authors in [11] described a predictive model constructed from data collected by smart meters for electricity theft detection. Number of sources of error and noise in measurement motivated them to apply statistical estimation procedure detection. Their methodology excellently distinguished between theft and no theft case. But accuracy for smaller amount of theft is still a challenge to resolve. Researchers in [12] identified problems of data privacy and security in context of smart metering. To ensure the consumers data privacy and protection, some initiatives were suggested such as guidelines for data services provided to consumers, consumer's control over data release from their end and protocols design requirement for data access at consumer end.

In [13, 14], several new techniques and methodologies were proposed, depending upon short-term energy interval data analysis for enhancing building's energy performance and operation. There are number of developments still possible for robust mass market implementation. Author in [15], presented an architecture for smart homes to address the four main challenges i.e. provision of low-overhead data collection, energy usage characteristics of modern devices, tracking of real time known devices' behavior and automatic detection of unknown devices. Here, smart energy meters were employed for efficient data collection and analysis to realize the behavior of household devices. For practical implementation, a set of models was constructed for accurate dealing with real-world devices than existing models. The modeling was employed to track the behavior of specific devices as well as unknown devices in smart home outlets. But in future, there is a need to optimize the smart home devices efficiency and their performance. Researchers in [16], proposed a fine-tuned predictive model for loss calculation in distributive network branch for power theft detection. The proposed predictive models were tested on distributed power circuits

and these showed better results as compared to theft detection by using actual smart meters data analysis. Researchers in [17] proposed a set of data mining algorithms for analysis of energy consumption patterns. They revealed household characteristics and offered attractive visualization of patterns. They aimed to explore algorithmic approaches for mining usage patterns and utilized for consumption forecasting and development of energy management strategies in future.

Researchers in [6] implemented a hybrid architecture comprises of Spark, Hive and PostgreSQL/MADlib to update smart meter data analysis. The architecture divided into various layers (acquisition layer, processing layer and analytics layer) and supported diverse at a processing units and analytics algorithm. Real-time data streams, batch analytics and OLTP (On-Line Transaction Processing) operations for social-economic were handled by hybrid architecture. For evaluation, benchmark work was conducted and verified its effectiveness. In future work, they planned to work on analytic layer and processing layer for enhancement of the performance by including more data types such as IoT, water, gas, and heat consumption etc.

Researchers designed a mechanism for acquiring data for customer usage through an open source data management platform i.e. Open Smart Energy Gateway (OpenSEG), to empower data management of smart meters' data. The designed architecture effectively worked with the ZigBee. It reduced cyber-security risks and provided secure data directly from smart meters to customers in real time. It stored 48 h of current consumption data in a circular cache in XML format and automated data transfer to utility. The designed system was used for homes, residential areas and commercial buildings in California [18]. The data processing method defined in [19] was envisioned for daily, monthly or annual consumption analysis and load profiles generation. The two-phased process employed for identifying suspected data and then addressed the suspected data with a gap filling process.

Authors presented an inclusive study of smart metering and data analytics for smart electricity meter [20]. They established a framework by utilizing analytics tools for fulfilling the stakeholders' requirements. The framework identified the smart metering limitations and wide range of data analytic tools for major activities of SG and smart metering. Furthermore, they concluded that processes and work flows for real time diagnosis are still in need to be designed in future. Researchers in [21] also designed and implemented a performance benchmark for smart meter data analytics tasks by employing five numeric computing platforms (Matlab, PostgreSQL/MADlib, System C, Hive and Spark/Spark Streaming). They proposed offline feature extraction as well as online anomaly detection framework. They generated large number of synthetic datasets from a small real data seed due to privacy issues in real time data collection of IoT enabled smart energy meter. Five platforms were compared on the basis of multicore machine performance and application development effort. Siemens devised highly scalable EnergyIP Analytics tool for utilities and power grid operators to handle big data generated by SGs. Complex data pattern scan be analyzed for energy theft, identification of overloaded or vulnerable devices and load forecasts at different levels of distribution grid dependent upon finely granulated meter data [22, 23].

Authors in [24] comprehensively reviewed the data analysis methods to assist the Intelligent Energy Networks (IENs). The data analysis focused on consumer clustering, forecasting demands, energy usage monitoring, pricing, generation optimization, and diagnostics. For data analysis, typical methods were applied such as linear regression, Support Vector Machine (SVM) and neural networks. Authors in [25] provided data management technologies and big data solutions for smart grids along with implementable tools and technical needs. Authors investigated smart meter data granularity while worried about privacy perseveration [7]. They suggested that household sorting quality can be modified by artifact development employing detailed survey.

The development of IENs requires advanced data analysis methods, tools and technologies which significantly improve the performance of smart energy networks. To meet the challenge of smart energy system's big data analysis, development of new framework is required for effective assistance in the development of IoT enabled smart energy system. In following section, a new framework is proposed along with the benefits of data analysis, outlined for encouragement of consumers as well utilities.

3 Proposed Criteria of Data Analytics of Smart Energy Meter

In our purposed framework as shown in Fig. 3, the consumer is equipped with smart energy meter which collects data on timely basis. Smart energy meters are embedded systems with controllers to manage the metering process, display unit, and communication module. These are comprised of electronic hardware and software to acquire data or observe data at desired time intervals along with time stamping. The data transmission's timing is decided by utilities, as data can be transmitted hourly, weekly, monthly etc. It is tough to transmit data after every second, so to track the electricity consumption accurately; hour-based data is stored in data logger and then transmitted through communication networks. Various architectures and topologies options are available for communication in smart energy metering. The smart meters are capable to transmit the data via available communication networks (such as Power Line Communications (PLC), Broadband over Power Line (BPL), Radio Frequency, cellular (2G/3G/4G/5G), and Wireless Local Area Network (WLAN) etc. For proper data transfer, Always Best Connected (ABC) network or ubiquitous communication network is required. Smart energy metering is featured with two-way communication system, it helps utility to control load devices as well as meters from distant for smooth processing. Two-way communication system ensures issuance of command/price signal from the utility to end consumers. Power is required to transmit and receive signals; this constraint is evident while choosing the best network for ubiquitous communication. A highly reliable communication network is desirable for transferring the high volume of data, as number of consumers with smart meters

is involved at different locations. Selection of a suitable communication network is a meticulous process because in this process a number of key factors are involved (i.e. huge amount of data transfer, confidentiality of sensitive data, authenticity and precision in communication data, cost effectiveness and restriction in accessing data). At the service provider as well as consumer end, a smart meter communicates consumption. Display unit shows data consumption to the consumers and make them aware about their energy usage. On another end, service provider's pricing information controls load devices to regulate user's assigned load limits and directives.

The data is collected from groups of smart meters in local data concentrators (i.e. Data Concentrator Unit (DCU)/Distributive Fog Computing (DFC) devices). Two main challenges come up from data management point of view while monitoring smart power networks: First one is real-time data acquisition, which is being resolved by deploying smart energy meters for reading consumption data over short time periods and second one is big data processing. To address these issues, distributive fog computing [26] (also known as edge computing) concept have been employed. The DFC devices worked as local servers with feeble performance. These devices are composed of distributed computing system such as personal, private and enterprise cloud. Location awareness and low latency are the two prominent characteristics of DFC.

Subsequently, data is sent to data storage devices (data servers) using a backhaul channel to cloud where data storing devices, servers, and processing amenities along with managing and billing applications reside. The collected data consists of critical personal information so the storage amenities need to be disaster proof and requisite back up plans for unexpected scenarios [27]. The cost hike is occurred with such provisions. Virtualization and cloud computing solve these issues smartly [28, 29]. Virtualization permits accessible resources fusion to get better efficiency and good investment returns, though it needs extra technology and complexity. Cloud computing enables virtual resources access at different locations, yet it brings serious concern for data security [30, 31]. Cloud computing has constraints in terms of various regulations and laws applied for data collection from diverse locations. Cloud computing on the other hand, reduces the cost of special purpose data centers as it utilizes the competence of diverse service providers [32]. Timing synchronization is significant for consistent transmission of data to cloud or other centralized systems for data analysis and bill management. It is a serious issue in case of wireless communication network.

The main components required for data analytics are data center infrastructure, servers for data handling, storage system, data base system, virtualization systems for efficient utilization of discrete storage devices and computing resources. The basic purpose is to utilize all available data from various resources, link together with accessible data analysis, data mining techniques, and infer valuable information for decision makings. Data analysis provides information in the form of consumption patterns, thermal sensitivity, daily profiles, classification of consumers, anomaly detection etc. Here, data analysis can be done by employing data analytic tools as described below to provide the information to the utility in a useful form.

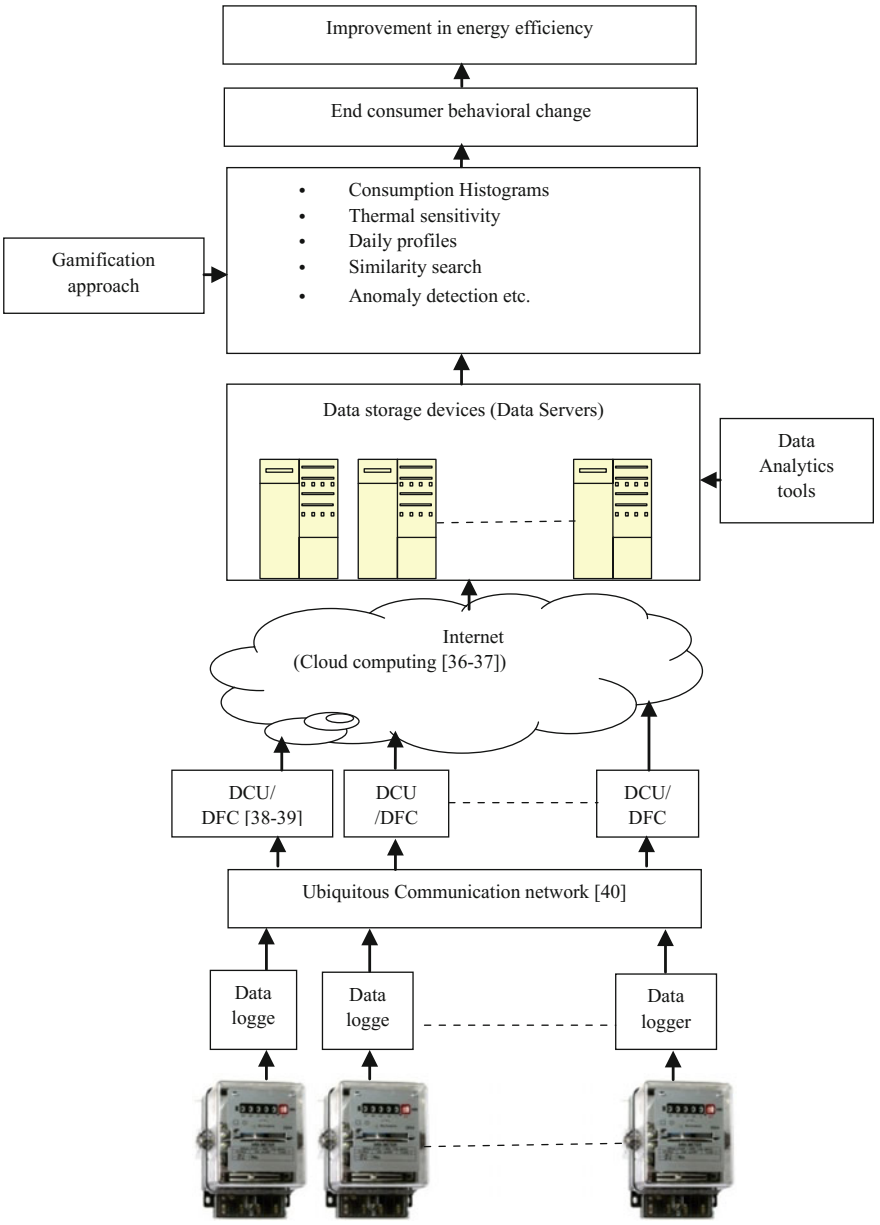


Fig. 3 IoT enabled smart energy meters

3.1 Data Analytics Tools for Smart Energy Metering

There are numerous mathematical tools (i.e. machine learning, deep learning, data-mining [1–33]) and statistical techniques (such as Self-Organizing Maps (SOMs [34]), Support Vector Machines (SVMs), Principle Component Analysis (PCA), and Fuzzy Logic (FL)) exist for smart metering [35]. Data analytics tries to realize suitable patterns or associations in a collection of data. The main goal of data analysis is to realize unknown associations among the data, especially when the data originates from different databases. The analysis utilizes advanced statistical methods such as clustering, classification, regression, forecast, anomaly detection, sense cyberattacks and demand response supply system.

Unsupervised learning is widely utilized for clustering of high-dimensional data vectors. It has ability to summarize the input data space and envision outcomes for interpretation. It enables the pictorial inspection of probable patterns and structures of data which can be directly employed for clustering and produces useful information. In smart energy metering, SOM, real time processing tools, batch processing tools, hybrid processing tools and customer analytics tools etc. have been generally employed for exception capture and profiling [36, 37].

Supervised learning is mainly employed for pattern recognition for classification and regression analysis. SVM [38], Multi-Dimensional Scaling (MDS) [39], Grade Correspondence Analysis (GCA) modelling, Periodic Auto Regression (PAR) algorithm [21] etc. are examples of data points' mapping in space for separating the categories with clear cut gap margins. These have been utilized for appliance-type recognition from smart meter data collection [40, 41] and for electricity theft detection also [42, 43].

Various mathematical tools are required for data aggregation [44], data reduction technique for consumption analysis [45] and detection of anomalies caused by malicious modification of data network [46] PCA, Collective Contextual Anomaly Detection using Sliding Window (CCAD-SW), Ensemble Anomaly Detection (EAD) [47], ARIMA and adaptive Artificial Neural Network (ANN) [48] are tools which generate orthogonal linear transformation, and translate data to a new coordinate system in such a way that the highest variance by any projected data becomes first coordinate and the second greatest variance becomes second coordinate etc. [49].

To improve the clustering consistency and cyber-attacks detection, FL has been employed. It makes an automatic decision-making platform for SGs. It is a tool of reasoning which approximates the values rather than crisp values. In [34, 42, 50] FL has improved the reliability of smart meters clustering, which is required to handle scalability issues. Intelligence in SGs and integration of technologies make the system open to cyberattacks. In [51], a FL-based technique was detecting cyber-attacks efficiently. These are some of the examples that show how existing techniques can be deployed.

Other mathematical tools such as Hidden Markov model and Bayesian techniques are utilized in various smart-metering applications, e.g. load disaggregation [52], appliance recognition [53] and power demand analysis [54]. In a broader range,

numerous methods are tailored and applied for smart metering for more benefits and efficiency. Table 1 points out the numerous methods and technologies employed for various data analytic processes.

3.2 Benefits of Data Analytics of Smart Energy Meter

Smart meters data analytics can optimize, manage and address peak demand issues for smart grids. Utilities/Stakeholders can better understand consumers' consumption patterns for providing personalized services to consumers. Moreover, consumers can better understand their own consumption ratings, and help themselves by saving energy. In the following points, smart energy meters data analytics' benefits are highlighted [6]:

- Consumption analysis and pattern discovery facilitated both consumers as well as utilities to take energy efficient decisions.
- Segmentation of consumers as per consumption and load profiles, promotes the most suitable energy-savings plans to a targeted segment.
- Induction of energy rebate or energy efficient programs.
- Forecasting of individual customer's energy consumption (e.g. daily, weekly, and monthly) is possible.
- Consumers can compare and correct the service providers' performance.
- Unpaid energy bills can be identified for revenue protection and may be reduced in future.
- Power quality monitoring (phase, voltage, current, active and reactive power, power factor) and its improvement.
- Smart home energy data management becomes easier.
- Anomaly detection possible.
- Equipment load management.
- Time-based pricing possible.
- Failure and outage notification.
- Remote command operations (turn on/off) by two-way communication with other intelligent devices.
- Energy theft detection.
- Load limiting for Demand Response (DR) purposes.
- Environmental conditions improvement by reducing carbon emissions via efficient power consumption.
- Feedback service allows sending alert and awareness messages and comparative statement with respect to pre-set time interval.

Table 1 Data analytic tools for smart energy metering

S.no.	Process	Tool
1.	Clustering [56–69], exception capture [70], profiling [38]	• Self-organizing maps (SOMs) [36] • Hierarchical cluster analysis [63] • Agglomerative clustering algorithm based on Ward's method [71] • C-means [72] • Real time processing tools [73] • Batch processing tools [73] • Hybrid processing tools [73] • Customer analytics tools (Hadoop, Storm, Flink, Spark) [73, 74] • Chicco [75]
2.	Recognize patterns, classification [10, 15, 16, 18, 57, 76, 77], regression [78–80], electricity theft detection [43, 44]	• Support vector machines (SVMs) [39] • Lance–Williams algorithms [81] • Multi-dimensional scaling (MDS) [40] • Grade correspondence analysis (GCA) [82] • Grade Stat tool [83] • Sequential pattern discovery using equivalence classes (SPADE) algorithm [84] • Non-technical and technical loss model [11, 85, 86] • Periodic auto regression (PAR) algorithm [6] • Naive Bayes algorithm [6] • Off-the-shelf classifier [7] • Decision tree classifier [7] • NP-complete [7] • Greedy approximation algorithms [52, 87] • Random forest [88] • AdaBoost [89] • Beckel's algorithm [90–92]
3.	Data aggregation [45], consumption analysis [20, 1, 15, 46, 60, 93–98], detection of anomalies [38, 50, 99–117]	• Principle component analysis (PCA) [50] • ARIMA and adaptive artificial neural network (ANN) [49] • K-nearest neighborhood (KNN) [102] • PARX [118] • Log-normal distribution function [118] • Statistical-based [119] • Nearest neighbor-based [119] • Cluster-based [119] • Classification-based [119] • Spectral decomposition-based techniques [119] • Unsupervised contextual and collective detection approach [120] • Stacked sparse autoencoder [121]

(continued)

Table 1 (continued)

S.no.	Process	Tool
4.	Improve the reliability of clustering, detecting cyber-attacks [12, 122–128]	• Fuzzy logic (FL) [34, 50–52] • Analysis process designer (APD) [129] • Data mining (DM) workbench [129] • Knowledge discovery in databases (KDD) [129] • SAPNet weaver business intelligence (SAP BI) [129] • Trusted platform module (TPM) [128–134]
5.	Load disaggregation [48], appliance identification [49], supply demand analysis [53, 135]	• Non-intrusive load monitoring (NILM) [15, 16, 136, 137] • Bayesian and hidden Markov model techniques [138] • Autoregressive integrated moving average (ARIMA) model [139] • Engineering algorithms [140] • Hourly simulation modeling [140] • Billing data analysis [140] • Interval meter data analysis [140] • End-use metered data analysis [140] • Statistically adjusted engineering (SAE) billing analysis [140] • NOSQL demand response automation [141]

4 Gamification Approach

The smart metering has a potential to assist humanities for collective goal of dropping energy demand and accepting energy-efficient lifestyles. Smart meters deployment and data analytics ultimately require the adaptation and action by society and communities to ensure success. By deploying smart meters, energy-related behavioral changes and enhancement in the energy efficiency is only possible by active engagement of consumers, especially in the household domain. To create a relation with the end-consumers and awake awareness among them require a motivational and knowledge enhancing system as shown in Fig. 4. Gamification is such a technique which activates natural wish for competition, accomplishment, rank, learning and self-expression. By gamification, participants' engagement improves and alters their less sustainable behavior on a voluntary basis.

In the proposed framework, consumers are act as players. The consumers remain motivated by offering rewards in lieu of their respective energy related tasks/achievements. Consumers gain points dependent upon their desirable behavior, less frequent usage of high electricity consuming devices and other activities. The game's objective is to change the behavior of consumers towards electricity usage and saving energy for improving efficiency.

For motivational purpose or enhancing consumers' engagement as well as knowledge, following criteria are needed to follow:

Fig. 4 Represents cloud gamification process [55]

- The current status and points must be displayed on a website of game designed for smart energy system, Facebook, Twitter etc.
- Compare consumers' real energy usage with neighbors and friends for rendering the participation.
- The game must consist of valuable information about energy saving methods, hints for energy behavior, and recommendations for energy efficient devices usage.
- To increase energy efficiency, reward points can be presented for every kWh saved; in case peak demand hours load shifting reward, points can be added for every shifted of kWh.
- Rewards can be categorized into personal, material and competitive rewards to encourage consumers. Penalty points can also be added for caution participation of consumers.
- Goals and levels in game are required to add time to time for provisioning of continuous motivation to consumers.
- To communicate the outcomes and winners of the game, utilities/companies can use additional communication media like publications of the utilities, local newspapers, newsletter etc. and a newsletter can also be used for reminding the consumers about future game events, new goals or levels.
- Promotion of the game must be organized by the utility/company accountable for the smart metering.
- The energy utilities must encourage billing by simply tweeting or using a link, which can take the consumer to the payment gateway directly.
- Deliver personalized experiences at moments that matters via gamification.

- For children, there must be an extra fun section in which they can competent to play energy related games and full emphasis on the importance of energy savings schemes (e.g. sustainability, environmental conservation, climate change, etc.) and also offer opportunities and recommendations so that children can save energy.
- Search partners/sponsors for game advertisement and extra inputs for rewards/prices offer to consumers.
- Game rules need to be flexible for consumers, so that these didn't over burden them.
- Consumers in the game utilize private data from smart meters so they must be conscious about data security and privacy.

Gamification techniques are capable to make effectively motivational behavior change, when these are implemented on everyday life routine. In addition, with gamification, social comparisons among electricity consumers effectively trigger motivation. Thus, both competitive and collaborative perspectives work together with social contribution for attaining the collective goal of improving energy efficiency. Environmental concerns are affected by this approach and it significantly reduces carbon emissions. Behavioral changes also educate and guide consumers towards appropriate actions and motivate engagement in energy-saving practices.

5 Conclusion

With the extensively employment of smart meters in smart cities, production of considerable volume of data, offering the opportunity for utilities/companies to improve end-consumer's services, lowering the cost, enhancing energy efficiency from consumers' point of view, significant amount of reduction in the bills and energy saving. Smart meters data analytics is a tedious task, it involves data acquisition, preprocessing, analysis and visualization. In this chapter, we proposed a new framework for data analytics of IoT enabled smart energy meter. The approach consists of various data analytic tools for analysis and gamification technique for consumer behavioral change towards electricity utilization. The proposed approach's data analysis focused on consumers clustering, forecasting of energy consumption, dynamic energy pricing, monitoring, energy generation optimization dependent upon demand response criteria, cyber-attacks and anomaly detection tools. Correspondingly, exemplary methods such as support vector machine, self-organizing maps, linear regression/prediction, and artificial neural networks etc., can be broadly employed. The benefits of data analysis of smart energy meters are outlined to encourage consumers and utilities for analysis. Game design techniques and mechanics integrated with smart meters to facilitate utilities for enhancing awareness among consumers, promotion of new offers and announcement of future plans. Gamification technique creates playful environment for consumers, attracts young generation and helps in raising energy policies and management issues. Data usage privacy and security rules are need to be perfectly designed by gamification approach to avoid ambiguity or

cyber-attacks. Consumers save energy and improve energy efficiency by behavioral change triggered by gamification approach indirectly reduces the carbon emissions, which is the emergent requirement of upcoming green smart cities.

Acknowledgements This work is supported by funding from Science and Engineering Research Board, New Delhi, India under File no. PDF/2016/001246.

References

1. Pathak, V.: Meter data acquisition system (MDAS) implementation challenges in India's R-APDRP. Smart Energy Meter. Int. (2) (2013)
2. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (eds.): Progress in computing, analytics and networking. In: Proceedings of ICCAN 2017, vol. 710. Springer (2018)
3. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Das, H.: Mistgis: Optimizing geospatial data analysis using mist computing. In: Progress in Computing, Analytics and Networking, pp. 733–742. Springer, Singapore (2018)
4. Das, H., Roy, D.S.: A grid computing service for power system monitoring. Int. J. Comput. Appl. **62**(20) (2013)
5. Das, H., Jena, A.K., Rath, P.K., Muduli, B., Das, S.R.: Grid computing-based performance analysis of power system: a graph theoretic approach. In: Intelligent Computing, Communication and Devices, pp. 259–266. Springer, New Delhi (2015)
6. Liu, X., Nielsen, P.S.: Streamlining smart meter data analytics. In: Proceedings of the 10th Conference on Sustainable Development of Energy, Water and Environment Systems, International Centre for Sustainable Development of Energy, Water and Environment Systems (2015)
7. Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., Irwin, D.: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, pp. 61–66. Zurich, Switzerland, Nov 02–02, 2010
8. Vijayapriya, P., Bapna, G., Kothari, D.P.: Smart tariff for smart meters in smart grid. Int. J. Eng. Technol. **2**(5), 310–315 (2010)
9. Aljamea, M.M. et al.: Smart meter data analysis. In: Proceedings of the International Conference on Internet of things and Cloud Computing. ACM (2016)
10. Pombeiro, H., Pina, A., Silva, C.: Analyzing residential electricity consumption patterns based on consumer's segmentation. In: Proceedings of the First International Workshop on Information Technology for Energy Applications. Lisbon, Portugal (2012)
11. Nikovski, D.N. et al.: Smart meter data analysis for power theft detection. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, Heidelberg (2013)
12. Ząbkowski, T., Gajowniczek, K.: Smart metering and data privacy issues. Inf. Syst. Manage. **2**(3), 239–249 (2013)
13. Jalori, S.: Leveraging smart meter data through advanced analytics: applications to building energy efficiency. Arizona State University (2013)
14. Das, H., Panda, G.S., Muduli, B., Rath, P.K.: The complex network analysis of power grid: a case study of the West Bengal power network. In: Intelligent Computing, Networking, and Informatics, pp. 17–29. Springer, New Delhi (2014)
15. Barker, S.K.: Model-driven analytics of energy meter data in smart homes (2014)
16. Sahoo, S. et al.: Electricity theft detection using smart meter data. In: Innovative Smart Grid Technologies Conference (ISGT), 2015 by IEEE Power & Energy Society (2015)
17. Gajowniczek, K., Ząbkowski, T.: Data mining techniques for detecting household characteristics based on smart meter data. Energies **8**(7), 7407–7427 (2015)
18. Janie et al.: Design of an open smart energy gateway for smart meter data management (2015)

19. Fowler, K.M. et al.: Simplified processing method for meter data analysis, No. PNNL–24331. Pacific Northwest National Laboratory (PNNL), Richland, WA (United States) (2015)
20. Alahakoon, D., Yu, X.: Smart electricity meter data intelligence for future energy systems: a survey. IEEE Trans. Industr. Inf. **12**(1), 425–436 (2016)
21. Liu, X. et al.: Smart meter data analytics: systems, algorithms, and benchmarking. ACM Trans. Datab. Syst. (TODS) **42**(1), 2 (2016)
22. http://www.Smartgrid.gov/the_Smart_grid#Smart_home. Accessed 15 June 2017
23. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of intrusion detection using data mining techniques. In: Progress in Computing, Analytics and Networking, pp. 753–764. Springer, Singapore (2018)
24. Siemens, A.G.: Siemens expands data analysis tool for smart metering by adding big data option. In: E-world Energy and Water. Essen, Germany, 16–18 Feb 2016
25. Ma, Z. et al.: The role of data analysis in the development of intelligent energy networks. arXiv:1705.11132 (2017)
26. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., Mankodiya, K.: Fog assisted cloud computing in era of big data and internet-of-things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
27. Das, H., Jena, A.K., Badajena, J.C., Pradhan, C., Barik, R.K.: Resource allocation in cooperative cloud environments. In: Progress in Computing, Analytics and Networking, pp. 825–841. Springer, Singapore (2018)
28. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud computing for optimization: foundations, applications, and challenges, vol. 39. Springer (2018)
29. Kar, I., Parida, R.R., Das, H.: Energy aware scheduling using genetic algorithm in cloud data centers. In: Proceedings of IEEE International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3545–3550, March 2016
30. Panigrahi, C.R., Tiwary, M., Pati, B., Das, H.: Big data and cyber foraging: future scope and challenges. In: Techniques and Environments for Big Data Analysis, pp. 75–100. Springer, Cham (2016)
31. Sahoo, A.K., Das, H.: Energy efficient scheduling using DVFS technique in cloud datacenters. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 59–66 (2016)
32. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)
33. Kar, I., Das, H.: Energy aware task scheduling using genetic algorithm in cloud data centres. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 106–111 (2016)
34. Baqui, N.M.: Fuzzy decision model for the smart grid. M.S. thesis, Dept. Comp. Sci., North Dakota State Univ. Agriculture Appl. Sci., Fargo, ND, USA (2012)
35. Deign, J., Salazar, C.M.: Data management and analytics for utilities, FC Business Intelligence Ltd. (2013)
36. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
37. Nagi, J., Yap, K.S., Tiong, S.K., Ahmed, S.K.: Electrical power load forecasting using hybrid self-organizing maps and support vector machines. In: Proceedings of 2nd International Power Engineering Optimization Conference (PEOCO '08), pp. 51–56, June 2008
38. De Silva, D., Yu, X., Alahakoon, D., Holmes, G.: A data mining framework for electricity consumption analysis from meter data. IEEE Trans. Ind. Informat. **7**(3), 399–407 (2011)
39. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273 (1995)
40. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**, 325–328 (1966)
41. Mittelsdorf, M., Huwel, A., Klingenberg, T., Sonnenschein, M.: Submeter based training of multi-class support vector machines for appliance recognition in home electricity consumption data. In: Proceedings Smart Greens, pp. 151–158 (2013)

42. Moro, J.Z., Duarte, L.F.C., Ferreira, E.C., Dias, J.A.S.: A home appliance recognition system using the approach of measuring power consumption and power factor on the electrical panel, based on energy meter ICs. Circuits Syst. **4**, 245–251 (2013)
43. McLaughlin, S., Holbert, B., Zonouz, S., Berthier, R.: AMIDS: amulti-sensor energy theft detection framework for advanced metering infrastructures. In: Proceedings 3rd IEEE International Conference Smart Grid Communications (Smart Grid Comm '12), pp. 354–359 (2012)
44. Anas, M., Javaid, N., Mahmood, A., Raza, S.M., Qasim, U., Khan, Z.A.: Minimizing electricity theft using smartmeters in AMI. In: Proceedings 7th International Conference P2P, Parallel Grid Cloud Internet Computing (3PGCIC '12), pp. 176–182 (2012)
45. Li, D., Aung, Z., Williams, J., Sanchez, A.: Efficient authentication scheme for data aggregation in smart grid with fault tolerance and fault diagnosis. In: Proceedings of 3rd IEEE PES International Conference Innovative Smart Grid Technologies (ISGT '12), pp. 1–8 (2012)
46. Abreu, J., Azevedo, I., Pereira, F.: A contribution to better understand the demand for electricity in the residential sector. In: Proceedings of European Council for an Energy Efficient Economy (ECEEE '11) Summer Study, pp. 1739–1750 (2011)
47. Anwar, A., Mahmood, A.N.: Cyber security of smart grid infrastructure: the state of the art in intrusion prevention and detection. CRC Press/Taylor & Francis, Boca Raton, FL (2014)
48. Araya, D.B. et al.: An ensemble learning framework for anomaly detection in building energy consumption. Energy Build. **144**, 191–206 (2017)
49. De Nadai, M., van Someren, M.: Short-term anomaly detection in gas consumption through ARIMA and artificial neural network forecast. In: IEEE Workshop on Environmental, Energy and Structural Monitoring Systems, pp. 250–255. IEEE Press, New York (2015)
50. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. **2**, 433–459 (2010)
51. Mirmojarabian, S.A.: Reliability computation of clustered smart meters using fuzzy logic. In: Proceedings Iranian Conference of Fuzzy Systems, pp. 1–6 (2013)
52. Ahmad, S., Baig, Z.: Fuzzy-based optimization for effective detection of smart grid cyber-attacks. Int. J. Smart Grid Clean Energy **1**(1), 15–21 (2012)
53. Chahine, K., Drissi, K., Pasquier, C., Kerroum, K., Faure, C., Jouannet, T., Michou, M.: Electric load disaggregation in smart metering using a novel feature extraction method and supervised classification. Energy Proc. **6**, 627–632 (2011)
54. Lukaszewski, R., Liszewski, K., Winiecki, W.: Methods of electrical appliances identification in systems monitoring electrical energy consumption. In: Proceedings of 7th IEEE International Conference Intelligent Data Acquisition on Advanced Computer System Technology Applications, pp. 1–14 (2013)
55. Guideline: Gamification—Making Energy Fun. www.smartgrid-engagement-toolkit.eu
56. Kamgarpoury, M., Tembine, H.: A Bayesian mean field game approach to supply demand analysis of the smart grid. In: Proceedings of 1st IEEE International Black Sea Conference on Communication Networks, pp. 196–200 (2013)
57. Abreu, J.M., Camara, F.P., Ferrao, P.: Using pattern recognition to identify habitual behavior in residential electricity consumption. Energy Build. **49**, 479–487 (2012)
58. Albert, A., Gebru, T., Ku, J., Kwac, J., Leskovec, J., Rajagopal, R.: Drivers of variability in energy consumption. In: Proceedings of ECML-PKDD DARE Workshop on Energy Analytics (2013)
59. Albert, A., Rajagopal, R.: Smart meter driven segmentation: what your consumption says about you. IEEE Trans. Power Syst. **4**(28) (2013)
60. Ardakanian, O., Koochakzadeh, N., Singh, R.P., Golab, L., Keshav, S.: Computing electricity consumption profiles from household smart meter data. In: Proceedings of EnDM Workshop on Energy Data Management, pp. 140–147 (2014)
61. Chicco, G., Napoli, R., Piglione, F.: Comparisons among clustering techniques for electricity customer classification. IEEE Trans. Power Syst. **21**(2), 933–940 (2006)
62. Espinoza, M., Joye, C., Belmans, R., DeMoor, B.: Short-term Load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series. IEEE Trans. Power Syst. **20**(3), 1622–1630 (2005)

63. Figueiredo, V., Rodrigues, F., Vale, Z., Gouveia, J.: An electric energy consumer characterization framework based on data mining techniques. IEEE Trans. Power Syst. **20**(2), 596–602 (2005)
64. Ghofrani, M., Hassanzadeh, M., Etezadi-Amoli, M., Fadali, M.: Smart meter based short-term load forecasting for residential customers. In: North American Power Symposium (NAPS) (2011)
65. Mattern, F., Staake, T., Weiss, M.: ICT for green—how computers can help us to conserve energy. In: Proceedings of e-Energy, pp. 1–10 (2010)
66. Todorovic, M., Tai, J.: Buildings energy sustainability and health research via inter disciplinarity and harmony. Energy Build. **47**, 12–18 (2012)
67. Gottwalt, S., Ketter, W., Block, C., Collins, J., Weinhardt, C.: Demand side management—a simulation of household behavior under variable prices, vol. 39, no. 12, pp. 8163–8174 (2011)
68. Lai, J., Yik, F.: An analytical method to evaluate facility management services for residential buildings, vol. 46, no. 1, pp. 165–175 (2011)
69. Adnan, R., Setan, H., Mohamad, M.N.: Multiple outliers detection procedures in linear regression. Matematika **19**, 29–45 (2003)
70. Jakkula, V., Cook, D.: Outlier detection in smart environment structured power datasets. In: 6th International Conference on Intelligent Environments, pp. 29–33. 2010. IEEE Press, New York (2010)
71. Zhang, T., Zhang, G., Lu, J., Feng, X., Yang, W.: A new index and classification approach for load pattern analysis of large electricity customers. IEEE Trans. Power Syst. **27**(1), 153–160 (2012)
72. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**, 236–244 (1963)
73. Daki, H. et al.: Big data management in smart grid: concepts, requirements and implementation. J. Big Data **4**(1), 13 (2017)
74. Reddy, K.H.K., Das, H., Roy, D.S.: A data aware scheme for scheduling big-data applications with SAVANNA hadoop. In: Futures of Network. CRC Press (2017)
75. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press: Berkeley, CA, USA (1967)
76. Chicco, G.: Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy **42**(1), 68–80 (2012)
77. Martani, C., Lee, D., Robinson, P., Britter, R., Carlo, Ratti C.: ENERNET: studying the dynamic relationship between building occupancy and energy consumption. Energy Build. **47**, 584–591 (2012)
78. Figueiredo, J., Sá, J.: A SCADA system for energy management in intelligent buildings. Energy Build. **49**, 85–98 (2012)
79. Lee, A.H., Fung, W.K.: Confirmation of multiple outliers in generalized linear and nonlinear regressions. J. Comput. Stat. Data Anal. **25**(1), 55–65 (1997)
80. Magld, K.W.: Features extraction based on linear regression technique. J. Comput. Sci. **8**(5), 701–704 (2012)
81. Zhang, Y., Chen, W., Black, J.: Anomaly detection in premise energy consumption data. In: Power and Energy Society General Meeting, pp. 1–8. IEEE Press, New York (2011)
82. Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies hierarchical systems. Comput. J. **9**, 373–380 (1967)
83. Szczesny, W.: On the performance of a discriminant function. J. Classif. **8**, 201–215 (1991)
84. GradeStat—Program for Grade Data Analysis. http://www.gradestatipipan.waw.pl
85. Zaki, M.J.: Spade: an efficient algorithm for mining frequent sequences. Mach. Learn. **42**, 31–60 (2001)
86. Nagi, J., Yap, K.S., Tiong, S.K., Ahmed, S.K., Mohamad, M.: Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans. Power Del. **25**(2), 1162–1171 (2010)

87. Depuru, S.S.S.R.: Modeling, detection, and prevention of electricity theft for enhanced performance and security of power grid, The University of Toledo, Aug 2012
88. Murphy, K.P.: Machine learning: a probabilistic perspective. The MIT Press (2012)
89. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
90. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997)
91. Hopf, K., Sodenkamp, M., Kozlovskiy, I., Staake, T.: Feature extraction and filtering for household classification based on smart electricity meter data. In: Computer Science-Research and Development, pp. 141–148. Springer, Berlin, Heidelberg, Zürich (2014)
92. Sodenkamp, M., Hopf, K., Staake, T.: Using supervised machine learning to explore energy consumption data in private sector housing. In: Handbook of Research on Organizational Transformations through Big Data Analytics, pp. 320–333 (2014)
93. Sodenkamp, M., Kozlovskiy, I., Staake, T.: Supervised classification with interdependent variables to support targeted energy efficiency measures in the residential sector. Decis. Anal. **3** (2016)
94. Ardakanian, O., Koochakzadeh, N., Singh, R.P., Golab, L., Keshav, S.: Computing electricity consumption profiles from household smart meter data. In: EDBT/ICDT Workshops, vol. 14, pp. 140–147 (2014)
95. Karjalainen, S.: Consumer preferences for feedback on household electricity consumption. Energy Build. **43**, 458–467 (2011)
96. Sütterlin, B., Brunner, T.A., Siegrist, M.: Who puts the most energy into energy conservation? A segmentation of energy consumers based on energy-related behavioral characteristics. Energy Policy **39**(12), 8137–8152 (2011)
97. Rasanen, T., Voukantsis, D., Niska, H., Karatzas, K., Kolehmainen, M.: Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl. Energy **87**(11), 3538–3545 (2010)
98. Jain, R.K., Smith, K.M., Culligan, P.J., Taylor, J.E.: Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl. Energy **123**, 168–178 (2014)
99. Tsekouras, G., Hatziargyriou, N., Dialynas, E.: Two-stage pattern recognition of load curves for classification of electricity customers. IEEE Trans. Power Syst. **22**(3), 1120–1128 (2007)
100. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 15 (2009)
101. Cheng, B., Longo, S., Cirillo, F., Bauer, M., Kovacs, E.: Building a big data platform for smart cities: experience and lessons from Santander. In: IEEE International Congress on Big Data, pp. 592–599. IEEE Press, New York (2015)
102. Chou, J.S., Telaga, A.S.: Real-time detection of anomalous power consumption. Renew. Sustain. Energy Rev. **33**, 400–411 (2014)
103. Janetzko, H., Stoffel, F., Mittelstdt, S., Keim, D.A.: Anomaly detection for visual analytics of power consumption data. Comput. Graph. **38**, 27–37 (2014)
104. Kiran, M., Murphy, P., Monga, I., Dugan, J., Baveja, S.S.: Lambda architecture for cost-effective batch and speed big data processing. In: IEEE International Conference on Big Data, pp. 2785–2792. IEEE Press, New York (2015)
105. Schneider, M., Ertel, W., Ramos, F.: Expected similarity estimation for large-scale batch and streaming anomaly detection. arXiv:1601.06602 (2016)
106. Janetzko, H., Stoffel, F., Mittelstädt, S., Keim, D.A.: Computers and graphics anomaly detection for visual analytics of power consumption data. Comput. Graph. **38**, 1–11 (2013)
107. Wrinch, M., El-Fouly, T.H.M., Wong, S.: Anomaly detection of building systems using energy demand frequency domain analysis. In: 2012 IEEE Power and Energy Society General Meeting, pp. 1–6 (2012)
108. Hill, D.J., Minsker, B.S.: Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model Softw. **25**(9), 1014–1022 (2010)

109. Fontugne, R., Ortiz, J., Tremblay, N., Borgnat, P., Flandrin, P., Fukuda, K., Culler, D., Esaki, H.: Strip, bind, and search: "a method for identifying abnormal energy consumption in buildings". In: 12th International Conference on Information Processing in Sensor Networks, pp. 129–140 (2013)
110. Arjunan, P., Khadilkar, H.D., Ganu, T., Charbiwala, Z.M., Singh, A., Singh, P.: Multi-user energy consumption monitoring and anomaly detection with partial context information. In: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, pp. 35–44. ACM (2015)
111. Zorita, A.L., Fernández-Temprano, M.A., García-Escudero, L.-A., Duque-Perez, O.: A statistical modeling approach to detect anomalies in energetic efficiency of buildings. Energy Build. **110**, 377–386 (2016)
112. Pe˜na, M., Biscarri, F., Guerrero, J.I., Monedero, I., Léon, C.: Rule-based system to detect energy efficiency anomalies in smart buildings: a data mining approach. Expert Syst. Appl. **56**, 242–255 (2016)
113. Capozzoli, A., Lauro, F., Khan, I.: Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst. Appl. **42**(9), 4324–4338 (2015)
114. Hayes, M.A., Capretz, M.A.: Contextual anomaly detection framework for big sensor data. J. Big Data **2**(1), 1–22 (2015)
115. Zhao, Z., Mehrotra, K.G., Mohan, C.K.: Ensemble algorithms for unsupervised anomaly detection. In: Current Approaches in Applied Artificial Intelligence, pp. 514–525. Springer (2015)
116. Amozegar, M., Khorasani, K.: An ensemble of dynamic neural network identifiers for fault detection and isolation of gas turbine engines. Neural Netw. **76**, 106–121 (2016)
117. Aburomman, A.A., Reaz, M.B.I.: A novel SVM-KNN-PSO ensemble method for intrusion detection system. Appl. Soft Comput. **38**, 360–372 (2016)
118. Brown, M., Barrington-Leigh, C., Brown, Z.: Kernel regression for real-time building energy analysis. J. Build. Perf. Simul. **5**(4), 263–276 (2011)
119. Liu, X., Nielsen, P.S.: Regression-based online anomaly detection for smart grid data. arXiv: 1606.05781 (2016)
120. Chakrabarti, A., Marwah, M., Arlitt, M.: Robust anomaly detection for large-scale sensor data. In: Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments. ACM (2016)
121. Rossi, B. et al.: Anomaly detection in smart grid data: an experience report. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) (2016)
122. Yuan, Y., Jia, K.: A distributed anomaly detection method of operation energy consumption using smart meter data. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP) (2015)
123. NISTIR 7628: Guidelines for Smart Grid Cyber Security Requirements. http://csrc.nist.gov/publications/nistir/ir7628/introduction-to-nistir-7628.Pdf
124. Pallotti, E., Mangiatordi, F.: Smart grid cyber security requirements. In: 10th International Conference on Environment and Electrical Engineering (EEEIC), pp. 1–4 (2011)
125. Lu, Z., Lu, X., Wang, W., Wang, C.: Review and evaluation of security threats on the communication networks in the smart grid. In: Military Communication Conference, 2010—MILCOM 2010, pp. 1830–1835 (2010)
126. Cleveland, F.: Cyber security issues for advanced metering infrastructure (AMI). In: Power and Energy Society General Meeting—Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE, pp. 1–5, July 2008
127. Berthier, R., Sanders, W., Khurana, H.: Intrusion detection for advanced metering infrastructures: requirements and architectural directions. In: First IEEE International Conference on Smart Grid Communications (Smart Grid Comm), Oct 2010, pp. 350–355 (2010)
128. Berthier, R., Sanders, W.: Specification-based intrusion detection for advanced metering infrastructures. In: IEEE 17th Pacific Rim International Symposium on Dependable Computing (PRDC), Dec 2011, pp. 184–193 (2011)

129. Flath, D.W.I.C., Nicolay, D.W.I.D., Conte, T., van Dinther, C., Filipova-Neumann, L.: Cluster analysis of smart metering data, an implementation in practice. Business Inf. Syst. Eng. **1** (2012)
130. Kush, N., Foo, E., Ahmed, E., Ahmed, I., Clark, A.: Gap analysis of intrusion detection in smart grids. In Valli, C., ed.: 2nd International Cyber Resilience Conference, SECAU—Security Research Centre (Aug 2011), pp. 38–46
131. McLaughlin, S., Podkuiko, D., McDaniel, P.: Energy theft in the advanced metering infrastructure. In: Proceedings of the 4th International Conference on Critical Information Infrastructures Security. CRITIS '09, pp. 176–187. Springer (2010)
132. McLaughlin, S., Podkuiko, D., Miadzvezhanka, S., Delozier, A., McDaniel, P.: Multi-vendor penetration testing in the advanced metering infrastructure. In: Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10, pp. 107–116. ACM (2010)
133. Kadurek, P., Blom, J., Cobben, J., Kling, W.: Theft detection and smart metering practices and expectations in the Netherlands. In: Innovative Smart Grid Technologies Conference Europe (ISGT Europe), 2010 IEEE PES, pp. 1–6 (2010)
134. Kroneis, H., Marsoner, H., Noormofidi, T.: Method for calibration of a measurement apparatus. U.S. Patent No. 5,185, 263, 9 Feb 1993
135. Electricity peak demand consumption management, Analytics case study by Deloitte (2014)
136. Birt, B.J., Newsham, G.R., Beausoleil-Morrison, I., Armstrong, M.M., Saldanha, N., Rowlands, I.H.: Disaggregating categories of electrical energy end-use from whole-house hourly data. Energy Build. **50**, 93–102 (2012)
137. Hart, G.W.: Nonintrusive appliance load monitoring. Proc. IEEE **80**, 1870–1891 (1992)
138. Ozoh, P., Apperley, M.: Simulating electricity consumption pattern for household appliances using demand side strategies: a review. In: Proceedings of the 15th New Zealand Conference on Human-Computer Interaction. ACM (2015)
139. Vadda, P., Seelam, S.M.: Smart metering for smart electricity consumption. Master Thesis, Electrical Engineering, School of Computing, Blekinge Institute of Technology, 37179 Karlskrona, Sweden, May 2013
140. Stern, F.: Peak demand and time-differentiated energy savings cross-cutting protocols. National Renewable Energy Laboratory (NREL) (2013)
141. Taube, B., Bienert, R.: Advanced data management and analytics for automated demand response (ADR) based on NoSQL (2012)

A New and Secure Intrusion Detecting System for Detection of Anomalies Within the Big Data

Amara S. A. L. G. Gopal Gupta, G. Syam Prasad and Soumya Ranjan Nayak

Abstract With the rapid growth of various technologies the level for the security has even become quite challenging and for the recognition frameworks in anomaly, several methods and methodology and actions region unit created to follow novel attacks on the frameworks or systems. Detection frameworks in anomaly upheld predefined set of instructions and protocols. It's hard to mandate all strategies, to beat this countless machine learning plans and downside unit existing. Unique issue is Keyed Intrusion Detection System namely kids that are completely relying on key privacy and procedure used to produce the key. All through this algorithmic program, intruder only ready to recoup or improve key by communicating with the Intrusion Detection System and perspective the tip result after it and by abuse this theme can't prepared to meet security norms. In this way supported learning we'd quite recently like the topic that can assist us with providing extra security on Data Storage. To reduce the attack risk, a dynamic key theory is bestowed and analyzed we've an inclination to face live about to planned theme for extra security that is ready to be secure delicate information of fluctuated domains like in consideration area enduring associated information like contact points of interest and antiquity.

Keywords Intrusion detection system · Feature selection · Machine learning Detection frame work in anomaly · Attacks

A. S. A. L. G. G. Gupta · G. S. Prasad · S. R. Nayak (✉)
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522502, Guntur, Andhra Pradesh, India
e-mail: Nayak.Soumya@kluniversity.in

A. S. A. L. G. G. Gupta
e-mail: Amara_Gupta@kluniversity.in

G. S. Prasad
e-mail: Syamprasad.Gudapati@kluniversity.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_8

1 Introduction

With the rapid increase use of wireless network [1–4] services and applications, security becomes a primary concern. From security view, data integrity and confidentiality are major issues for information systems. As of late utilization (recent years) of web has been improved awesomely. The majority of persons utilized cloud [5–13] numerous (avoid wasting) and lots of it utilize web to transmit their information. Here remains likelihood that the information could be hack and see defrauded (victimized). On behalf of higher assurance after such unapproved or unapproved peers, varied finding in Anomaly identification proof designs are displayed in continuous year. Security drawback separated into dual gatherings in that malicious is the first one and elective is non-vindictive movement. Cryptography is the way of securing information by making sure that the data can be understood only by the authenticated person. Once you encode your file, you can configure the file to be encrypted by using Dynamic Key Generation and Dynamic Key distribution policy and asset delivery policy [14]. The information encryption, decryption process become one of the most commonly means. A malicious assault or attack is an undertaking to explicitly misuse or favorable position of some individual's Computer, paying little heed to whether over system faults, shared building, elective varieties of shared planning and phishing. Might be finished the goal of taking specific material, (for example, in social network) or to downsize the common sense of an objective system. Hateful Code chiefly Pelt in Message, site, genuine urls and etc. as Associate in illustration or example Viruses, Worms, Phishing, Spam, Trojan Horse. Non-malignant (Malicious) intruders happen because of safety arrangements are poor, it handles the susceptibilities change, and missteps or mistakes toward require put. There are two essential strategies in cryptography symmetric and asymmetric cryptography technique. Symmetric cryptography—In this method, an entity can transmit data over a transmission channel by using a single key for both encryption and decryption.

Asymmetric cryptography—In asymmetric method key, mainly two keys is used for encryption and decryption. One key will be utilized for encryption technique, and another key will be utilized for decryption [15]. An intrusion detection system might be a tool or software package request, which overlooks the framework, or system that exercise the malicious activities or advent violations and processes reports to a management station. IDS region unit reachable in an extremely shift of "flavors" and approach the independent of sleuthing distrustful activity from various perspectives.

There are 2 detection systems mostly networking based intrusion detection systems and Host Based detection (intrusion) systems security system in Network is a NIDS that specialize in the assaults which originated from the at intervals of the users in approved or network. A few frameworks may imagine stopping interruption yet be that as it may this can be often neither required nor expected of a watching system [16]. In order to evade detection the intruder makes some arrangements that effect forever. In expressions of system safety, the evasion attack implies that bypass a flaw in (an exceptionally very) security framework that allows an intruder to maintain a strategic distance from security and its mechanism to induce framework or system

network access therefore on attack, or various form of malware, deliver an exploit, whereas not detection evasions area unit usually accustomed counter system-based interruption recognition and obstacle frameworks yet can likewise be acclimated by-pass firewalls. To Smash a System Security in a network device, some additional targets of avoidance are interpreting it in-successful to succeeding focused on attacks which are specified. A small amount of detection schemes area unit introduced in a decade ago with monitor from such avoidance attacks. In this paper, our motive is to avoid attacks at regular intervals. A one within the entire theme to avoid attacks which in a evasion is Intrusion (keyed) Detection System. These systems first time introduced at DIMVA'10 by Drazenovic at and Mrdovic. The mainly important a part of attacks (network or system) take place at the applying layer, analyzation of payload in a packet is important for finding out. Sadly, packets with malicious may even be organize to ancient weight, if the detection technique in anomaly is believed to avoid detecting.

Traditional payload model is essential needy or depend. For each execution of the strategy a secret key acts different and it is reserved secret. So model of ancient payload is secret though public detection technique in order to prevent attacks.

2 Literature and Background Work

The Machine learning has been utilized as a part of huge region of security related assignments like system interruption discovery and spam separating, malware and, to recognize amongst malevolent and justify sample tests is not flippant issue. Antagonistic learning research not exclusively been immediate the issue of break down security of common learning Algorithms to deliberately focused on attacks, yet in addition that of plan learning algorithms with update security. A certain presumption at the back of machine learning and pattern recognition algorithms are that training information or test information are delivering from the same, potentially not known. The attacks are greatly effective, demonstrating that it is sensibly simple for an attacker to recoup the key in any of the two algorithms which is utilized as a part of this intrusion detection framework [17–34].

2.1 Antagonist Study and Avoidance

In Security connected issues or tasks like network interruption discovery and malware and separating the spam Machine Learning [35–37] has wide utilized to perceive among st vindictive and genuine examples is significant issue, Dalvi explorer indistinguishable issue in [38] consequently evasion will be ordered. In any case, these issues are especially troublesome for machine learning calculations because of the nearness of adaptive adversary and intellectual who will sensibly control the input information to minimize the execution of the recognition framework (system)

which violate the essential assumption of information inactive, training and test information follow a parallel (although generally unidentified) division. Research in antagonist study has not solely been addressing to the issue of addressing safety of flow learning Algorithms to de-liberately focused on attacks, anyway likewise that of learning calculations and algorithms with enhanced security. Toward for protecting avoidance attacks, express data of various types of antagonist information exploitation takes remained integrated into study. By considering some algorithms example like game-theoretical. A hidden statement behind pattern recognition and ancient machine learning algorithms is that instruction and test data are drained from a similar, in all probability anonymous, division. This statement is still feasible to be offended in antagonist settings, while attackers could suspiciously organize the participation file to downgrade the performance of a network. The observations of Meek and Lowd are that the enemy needn't mold the classifier plainly and individual invention instance of bottom attacker cost as within the setting in Dalvi et al. For antagonist classifier reverse engineering (ACER) problem they formalize a notion of turnaround engineering. Given an aggressor cost operate function; they examine the complexness of decision at least attacker cost occurrence that the classifier marks as negative. They imagine no information of general data, although the antagonist will understand the feature area and additionally should have one helpful example and one unhelpful example. In this a learning problem named ACER delivers a method of qualify however tough toward exercise requests to turn around contrive a classifier from specific assumption class using a specific quality space. They show that linear classifier is ACRE learnable with linear few alternative minor restrictions and attacker cost functions. In this ACRE-learnable is a classifier present a question algorithm is a polynomial that discoveries an attacker cost lowest undesirable example.

2.2 Preventing Strategies on Evasion

Dalvi et al. [39] Initiate an alternate set of attacks which is a polymorphic, known as integration (mix) attacks, that might effectively avoid network anomaly which is on a IDS frequency (byte) based through suspiciously identical the information of the attack instance which is mutated to the expected profiles. The predictable polymorphic combinations which are integrated attacks will be analysis as a subclass of the attack which called mimicry attack. In this author capture a reasonable method to the problem and properly define the phases and step by step process needed towards embrace obtainable such attacks. They not exclusively explain that such attacks are achievable but furthermore investigate the stability of avoidance under totally dissimilar conditions. By using a byte frequency-based anomaly IDS and PAYL it will present a complete and detailed method. For monitor the packet payload for anomalies, several application anomaly IDS are considered. G. Fumera, F. Roli and B. Biggio experiments and shows the results (which are systematical) supported and consequent the methodical structure, which shows data hiding to the antagonist through the minimization of the judgment utility, will look up the inflexi-

bility classifier avoidance. Kruegel et al. outlined four absolutely unusual models, in particular, length, character dissemination, probabilistic illustrative phonetics, and token discoverer, for the introduction of HTTP assaults. PAYL, arranged the records by Stolfo and Wang which is meaningful recurrence of event of each byte contained by the payload of a regular bundle or packet. A different framework is made for each port and bundle length. In their original edge work, the creators educated an enhanced depiction regarding PAYL that processes different shapes designed for each port. By the maximum point of the movement, group is achieved to diminish the quantity of outlines. The main problem of the structure is that they are responsibility not consider a difficult attacker, who could know the IDS operation at the objective and aggressively try to avoid it. Data Security is supposed to be providing in each organization as it is the major necessity for every excellent work. Author contemplate a technique consisting in hiding from everything data around the adversary toward the enemy complete the presentation of a little randomness within the assessment function and target a completion of this approach throughout a classifier system which is a multiple. The implementation of this process applies to all those areas where the data security is desperately needed to its best quality.

2.3 A Secured Intrusion Detection System

Barreno et al. [40] planned Intrusion Detection System (Keyed) within which key (secret) plays key role. Anomaly detector in network or system reviews payloads in a packet.

The projected procedure takes three essential phases for key completion which are discussed in following subsections.

(1) Instruction Method

In this method payload separated into words, the sequence of byte which called as words placed among delimiters.
From this any unique 2-byte assign to secret set. These sets are once more classified into frequency count and straight words.

(2) Detection Method

In this mode (detection) abnormality score get counted according to word rate of recurrence count (frequency).

(3) Selection of Key

The Key got chosen when its score and checking its recognition quality. For Generation of new key 3 steps want to repeat every time (Fig. 1).

Fig. 1 Generation of new key

2.4 KEY Recovery Attacks

Attacker takes smooth the progress of several requests to urge supplementary data associated to secret key. The Research Analysis of Arturo Ribagorda Juan Benjamin Ramos E. Tapiador, Agustin Orfila, expressions that in Induction Detection method attacker basically intelligent to act through it and process the response of the collaboration attacker and their attacks arranged the protected data, with the help of these attack, it makes 256 exact queries to Detection System 256 with every cautious key component and one ending query to see that set parallels to the key.

3 Existing System

The Antagonist classifier turns around engineering downside (ACRE) because the work out of learning adequate data a couple of classifier to build attacks, moderately than craving for higher methods. The main question of calculating higher ways towards alters an attack so it avoids recognition by a classifier named as Bayes. In present system, the construction of the matter fundamentally in hypothetic (game) terms, wherever every alteration in occurrence is higher and self-made recognition and evasion consume numerable utilities to the classifier and also the antagonist, severally. The setting utilized in thought an antagonist with crammed with data of the classifier to be evaded. Shortly when, on the other hand evasion is done once such data is unattainable.

Here uses a oracle association as unlimited antagonistic model: Likewise, a classifier is ACRE k-learnable if the value isn't minimum however enclosed by k. Thus, result thought is to search out example with a functional scope of question for avoid discovery. If there exists a calculation that finds an insignificant cost case avoid discovery use exclusively poly-ostensibly a few inquiries then a classifier is asserted to be ACRE learnable. Intended for approximately open issues and contests connected with the evasion classifier downside. The role of machine learning in security application Supplementary some extra works have reentered, with thorough stress on anomaly appreciation. Among the outcomes given, it's confirmation that direct classifiers with continuous alternatives are ACRE k-learnable for straight value works subsequently, these classifiers not proper for adversarial situations and won't be utilized. Later work by sums up these outcomes to arched instigating classifiers, seeing that it's regularly not important to pivot design the decision limit to assemble con-

cealed cases of close insignificant cost. The disadvantages of existing systems are as follows:

- Malicious Node expends extra vitality
- Doesn't get together with security models.

4 Proposed Work

Our strikes are to an excellent degree working on, showing that it's rationally clear to recover the key for a guilty party in any of the settings inspected. We tend to assume that such a nonappearance of security reveals that designs like kids shouldn't envision key-recuperation attacks. The attacks here showed could be thwarted by displaying distinctive spontaneous security methods the structure, on behalf of example, compelling the maximum outrageous size of words and payloads, which containing such cost as demand parts. Presently we have combat/fought that obstacle against such strikes is essential to any classifier that undertakings to disappoint evading by relying on a riddle bit of information. We have given swap on this and extra open demand in the trust of including further research about there. Along these lines, our proposal for possible arrangements is to create choices considering liberal measures rather than detailed solutions. Our point is redesign intrusion detection system and meets all properties that add the security with the objective and is able to prepare towards anchor store information in vast information bases.

4.1 System Structural Design

- Node Creation
- Node Routing
- Recovery Attacks (KEY)
- Key Anomaly Detection
- Antagonist Models Revisited (Fig. 2).

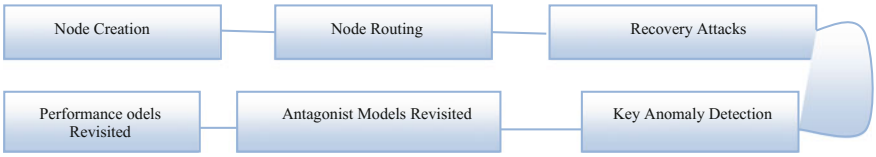


Fig. 2 System design

4.1.1 Node Creation

A structure Node (center point) is relationship points that can get, make, store or send data along appropriated compose routes. Every framework or system hub whether it's an endpoint for data transmissions or a redistribution point, has either a modified or fabricated capacity to see, process and forward transmissions to other framework hubs. Network hubs appeared with the utilization of dispersed systems and packet exchanging. Nodes play out a variety of functions. A system hub is a gadget that plays out a particular function. Every hub needs a MAC address for each system interface card (NIC). The individual computers on the fringe of the network, those that don't likewise interface different systems, and those that frequently associate temporarily to at least one cloud are called end hubs. Normally, inside the distributed computing build, the individual client/client computer that interfaces into one very much oversaw cloud is called an end hub. Since these computers are a piece of the system yet un-managed by the cloud's host, they introduce noteworthy risks to the whole cloud. Node is the tool that we will be using to build our server (Fig. 3).

4.1.2 Node Routing

Routing is the way toward choosing a way for activity in a system, or between or over numerous systems. Extensively, it is performed in numerous sorts of systems, including circuit-exchanged systems, for example, people in general exchanged phone arrange (PSTN), and PC systems, for example, the Internet. General useful PCs likewise forward bundles and perform routing, although they have no exceptionally enhanced equipment for the task. The routing process often organizes sending based on directing tables, which preserve up a highest of the courses to dissimilar system goals. Routing tables might be indicated by a director, learned by observing system activity or worked with the help of routing conventions.

Here a remote framework is made. Every one of the center points are randomly sent in the framework area. This framework describes the relationship of collaborating programs in an application. Centers are doled out with adaptability. The server segment gives a capacity or administration to one or numerous systems, which start demands for such services. Servers are characterized by the administrations they provide. Here able to do or adjusted for diverting effortlessly starting with one then onto

Fig. 3 Node creation

the next of different hubs and so on. Hubs center points' transportability is set center move beginning with one position then onto the following. Source and destination are portrayed. Data traded from source center point to end center (Fig. 4).

4.1.3 Recovery Attacks (KEY)

The risk of the misusing the data is not only during its transmission but also at its source address before the data is transmitted. Here in this framework, job doesn't suitable fit confidential in light of the way that our standard target isn't to strike the knowledge count itself, but instead to recuperate one quantity of secret information. The key distribution accesses the key storage and gets accesses to all the keys on their levels of encryption, along these lines, might be essential to effectively dispatch an evasion attack. KIDS Anomaly Detection System is the one important issue begins from the nonappearance of broadly perceived contradicting models giving a right depiction of the intruder's goals and, his capacities one such model for secure machine learning and discussed distinctive general strike classes. The keys need to be stored to maintain the record of keys divided among the users. Key organization concerns keys at the customer level, either between customers or frameworks. This is instead of key booking; key arranging ordinarily suggests within treatment of the key material within the activity.

4.1.4 Key Anomaly Detection

The uncertainties made about the aggressor's capacities are essential to genuinely separate the security of any arrangement, yet some of them may well be improbable for a few applications. Obstinately known with the center of attention talked worried over is that they got to set up doubtlessly described and induced badly arranged models for secure machine learning calculations.

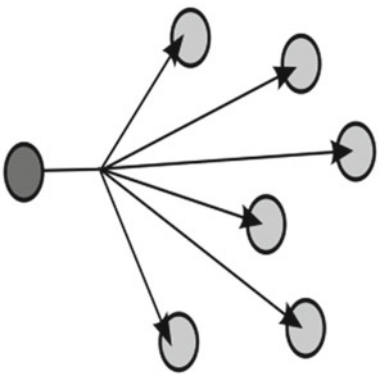


Fig. 4 Node routing

4.1.5 Antagonist Models Revisited

There occurs a inquiry calculation of polynomial that finds a most minimal assailant cost negative case. Encryption plays a significant role in not only providing security but also creating a sophisticated environment for an intruder to decrypt the code easily. The data encryption life cycle changes its form dynamically from one phase to another in a cyclic fashion (Fig. 5, 6 and 7).

In this a couple of similarities in cryptography with Chosen normal text Attacks (CPA). This theory has remained finished by various mechanisms in protected chain learning.

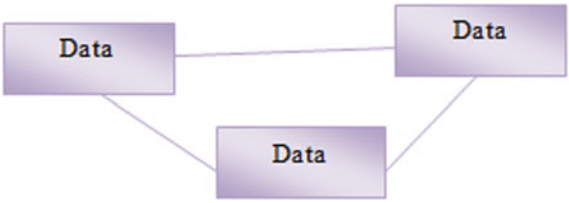


Fig. 5 Data encryption cycle

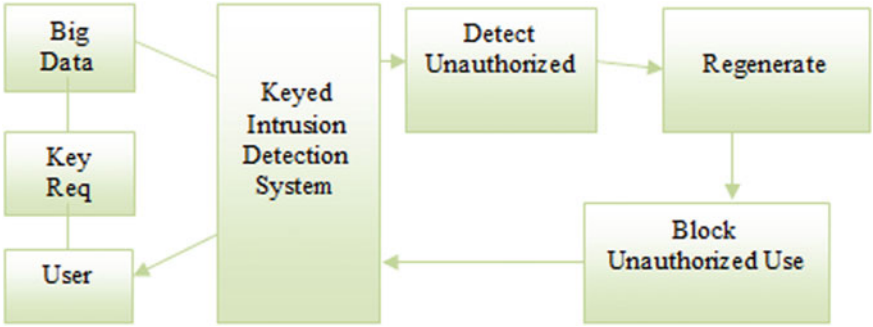


Fig. 6 Overall system design

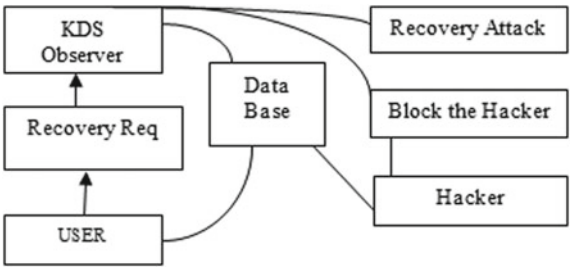


Fig. 7 Block diagram of recovery attack

5 Performance Analysis

In this paper we have measure the performance analysis by means of secret key generation through KDC and generated key in terms of both data presented in Fig. 8. Another performance analysis we have taken care by means key storage in the database which are used for accessing the data into the various levels which is represented on Fig. 9.

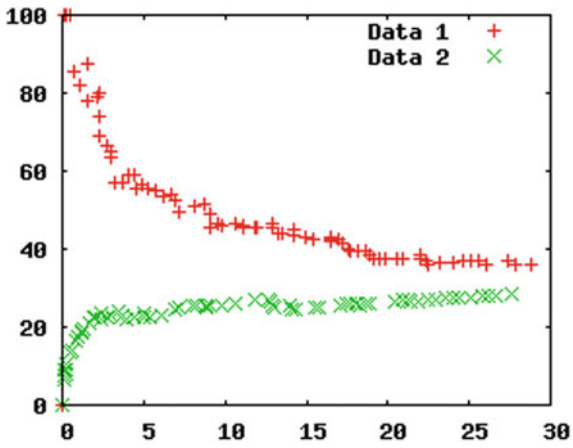


Fig. 8 Secret key generation through KDC

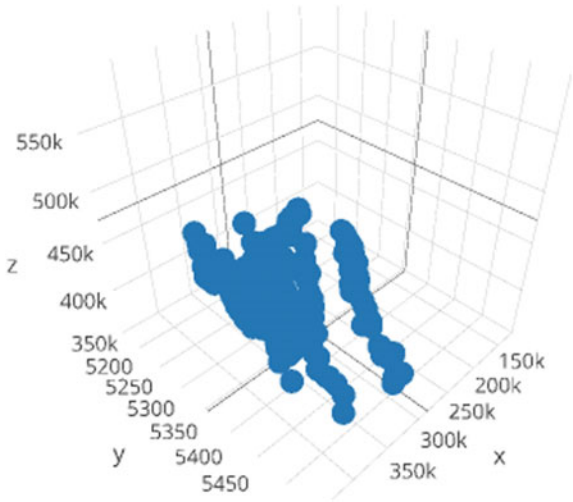


Fig. 9 Key storage access in various levels

6 Conclusion

We have investigated the quality of Intrusion Detection System against key-recuperation assaults and have showed key-recovery assaults as per ill-disposed settings, contingent upon the criticism given by KIDS to testing inquiries. A conclusive objective is to keep away from the structure, and we have as of late expected that knowing the key is key to make an attack that avoids recognizable proof or, in any occasion, that out and out supports the system. It remainders to be realized whether a keyed classifier, for example, Intrusion Detection System can be essentially avoided without plainly improving the key. Investigation in this paper shows sensibly on behalf of attacker to recoup the key. Our concentration in this effort has been on recuperating the key through capable techniques, displaying that the course of action procedure spills data about it that can be utilized by a hacker.

Acknowledgements The authors would like to thank the management of Koneru Lakshmaiah Education Foundation (Deemed to be University) for their support throughout the completion of this project discussions and comments.

References

1. Mishra, B.B., Dehuri, S., Panigrahi, B.K., Nayak, A.K., Mishra, B.S.P., Das, H.: Computational intelligence in sensor networks. In: Studies in Computational Intelligence, vol. 776. Springer (2018)
2. Sarkar, J.L., Panigrahi, C.R., Pati, B., Das, H.: A novel approach for real-time data management in wireless sensor networks. In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, pp. 599–607. Springer, New Delhi (2016)
3. Das, H., Naik, B., Pati, B., Panigrahi, C.R.: A survey on virtual sensor networks framework. Int. J. Grid Distrib. Comput. **7**(5), 121–130 (2014)
4. Panigrahi, C.R., Sarkar, J.L., Pati, B., Das, H.: S2S: a novel approach for source to sink node communication in wireless sensor networks. In: International Conference on Mining Intelligence and Knowledge Exploration, pp. 406–414. Springer, Cham, Dec 2015
5. Das, H., Jena, A.K., Badajena, J.C., Pradhan, C., Barik, R.K.: Resource allocation in cooperative cloud environments. In: Progress in Computing, Analytics and Networking, pp. 825–841. Springer, Singapore (2018)
6. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud computing for optimization: foundations, applications, and challenges, vol. 39. Springer (2018)
7. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (eds.): Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2017, vol. 710. Springer (2018)
8. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)
9. Reddy, K.H.K., Das, H., Roy, D.S.: A data aware scheme for scheduling big-data applications with SAVANNA Hadoop. In: Futures of Network, CRC Press (2017)
10. Sarkhel, P., Das, H., Vashishtha, L.K.: Task-scheduling algorithms in cloud environment. In: Computational Intelligence in Data Mining, pp. 553–562. Springer, Singapore (2017)
11. Panigrahi, C.R., Tiwary, M., Pati, B., Das, H.: Big data and cyber foraging: future scope and challenges. In: Techniques and Environments for Big Data Analysis, pp. 75–100. Springer, Cham (2016)

12. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R.K., Pratik, T., Sharma, S., Das, H. et al.: Mistgis: optimizing geospatial data analysis using mist computing. In: Progress in Computing, Analytics and Networking, pp. 733–742. Springer, Singapore (2018)
13. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., Mankodiya, K. et al.: Fog assisted cloud computing in era of big data and internet-of-things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
14. Eason, G., Noble, B., Sneddon, I.N.: On certain integrals of Lipschitz-Hankel type involving products of Bessel functions. Phil. Trans. Roy. Soc. London. A **247**, 529–551 (1955)
15. Yuping, Z., Xinghui, W.: Research and realization of multi-level encryption method for database. In: Advanced Computer Control, Proceeding of ICACC Conference, vol. 3, pp. 1–4 (2010)
16. Tapiador, J.E., Orfila, A., Ribagorda, A., Ramos, B.: Key recovery attacks on KIDS, a keyed anomaly detection system. IEEE Trans. Depend. Sec. Comp. **12**, 312–325 (2015)
17. Nath, A., Bhowmik, S., Basu, D., Bose, A., Chatterjee, S.: bit level multi way feedback encryption standard version-2 (BLMWFES-2). In: Advanced Communication Control and Computing Technologies. Proceeding of ICACCCT, pp. 1702–1707 (2014)
18. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. J. Mach. Learn. Res. **20**, 97–112 (2011)
19. Bose, A., Basu, D., Chatterjee, S., Nath, A., Bhowmik, S.: Bit level multi way feedback encryption standard Ver-1(BLMWFES-1). In: Communication Systems and Network Technologies. Proceedings of CSNT, pp. 793–799 (2014)
20. Gates, C., Taylo C.: Challenging the anomaly detection paradigm: a provocative discussion. In: Proceedings New Security Paradigms Workshop (NSPW), pp. 21–29 (2006)
21. Mohammadi, M.S., Bafghi, A.G.: A dynamic, zero-message broadcast encryption scheme based on secure multiparty computation. In: Information Security and Crytology. Procedding of ISC, pp. 12–17 (2012)
22. Fu, B., Lin, J., Duan, G., Analysis of multi-biometric encryption at feature-level fusion. In: Intelligent Control and Automation. Proceedings of WCICA, pp. 4563–4567 (2012)
23. Liu, T., Liu, Y., Mao, Y., Sun, Y., Guan, X., Gong, W., Xiao, S.: A dynamic secret-based encryption scheme for smart grid wireless communication. IEEE Trans. Smart Grid. **5**, 1175–1182 (2014)
24. Fogla, P., Sharif, M., Perdisci, R., Kolesnikov, O.M., Lee, W.: Polymorphic blending attack. In: Proceedings of the 15th USENIX Security Symposium (Security '06) (2006)
25. Gmira, F., Hraoui, S., Saaidi, A., Oulidi, A.J., Satori, K.: Securing the architecture of the JPEG compression by an dynamic encryption. In: Intelligent Systems and Computer Vision. Proceeding of ISCV, pp. 1–6 (2015)
26. Mande, P.J., Chunchure, B.: Key-recovery attacks prevention in keyed anomaly detection system. Int. J. Innov. Res. Comp. Eng. **3**, 12973–12975 (2015)
27. Yu, P.H., Pooch, U.W.: d-key dynamic encryption—a security enhancement protocol for Mobile Ad Hoc network. In: Ubiquitous and Future Networks. Proceeding of ICUFN, pp. 183–188 (2009)
28. Liu, T., Guo, H.,: Dynamic encryption key design and its security evaluation for memory data protection in embedded systems. In: IT Convergence and Security. Proceeding of ICITCS, pp. 1–4 (2014)
29. Kolcz, A., Teo, C.H.: Feature weighting for improved classifier robustness. In: Proceeding of Email and Anti-Spam CEAS, pp. 1–8 (2009)
30. Fu, B., Lin, J., Duan, G.: Analysis of multi-biometric encryption at feature-level fusion. In: Intelligent Control and Automation. Proceeding of WCICA, pp. 4563–4567 (2014)
31. Zhou, J., Cao, Z., Dong, X., Lin, X.: TR-MABE: White-box traceable and revocable multi-authority attribute-based encryption and its applications to multi-level privacy-preserving e-healthcare cloud computing systems. In: Computer Communications. Proceeding of INFOCOM, pp. 2398–2406 (2015)

32. Feng, L., Piao, H., Ling, H., Jin, B.: A network disk encryption with dynamic encryption key. In: Computer Design and Applications. Proceeding of ICCDA, vol. 5, pp. 614–616 (2010)
33. Zhang, H.Y., Zhou, Q.S., Tao, Y.: A kind of dynamic encryption algorithm and its application. In: Circuits, Communications and System. Proceeding of PACCS, vol. 2, pp. 15–18 (2010)
34. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of intrusion detection using data mining techniques. In: Progress in Computing, Analytics and Networking, pp. 753–764. Springer, Singapore (2018)
35. Das, H., Naik, B., Behera, H.S.: Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In: Progress in Computing, Analytics and Networking, pp. 539–549. Springer, Singapore (2018)
36. Das, H., Jena, A.K., Nayak, J., Naik, B., Behera, H.S.: A novel PSO based back propagation learning-MLP (PSO-BP-MLP) for classification. In: Computational Intelligence in Data Mining, vol. 2, pp. 461–471. Springer, New Delhi (2015)
37. Pradhan, C., Das, H., Naik, B., Dey, N.: Handbook of Research on Information Security in Biomedical Signal Processing, pp. 1–414. IGI Global, Hershey, PA (2018). https://doi.org/10.4018/978-1-5225-5152-2
38. Charan, G.S., Kumar, S.S.V.N., Karthikeyan, B., Vaithiyanathan, V., Lakshmi, K.D.: A novel LSB based image steganography with multi-level encryption. In: Innovations in Information, Embedded and Communication Systems. Proceeding of ICIIECS, pp. 1–5 (2015)
39. Dalvi, N., Domingos, P., Mausam., Sanghai, S., Verma, D.: Adversarial classification. In: Knowledge Discovery and Data Mining, Proceeding of ACM SIGKDD Conference, vol. 4, pp. 99–108 (2004)
40. Barreno, M., Nelson, B., Joseph, A.D., Tygar, J.D.: The security of machine learning. Mach. Learn. **81**, 121–148 (2010)

Geospatial Big Data, Analytics and IoT: Challenges, Applications and Potential

Ramgopal Kashyap

Abstract Machine learning gives to great degree critical instruments for astute geo-and ecological information investigation, handling and representation. This chapter introduces a review of provincial arrangement of ecological information, record of consistent natural and contamination information, including the utilization of programmed calculations, enhancement of checking systems. Machine learning calculations are intended to distinguish proficiently and to foresee precisely designs inside multivariate information. They give investigators computational apparatuses to help prescient demonstrating and the understanding of associations of information. The examination of giant volumes of stand-out variable geospatial info utilizing machine learning estimations thus offers out of the question confirmation to trade and analysis within the geosciences. Geosciences info square measure currently and once more delineated by a restriction within the variety and transports of direct acknowledgments, static amendment in these info associate degreed an uncommon condition of interclass fancy and interclass likeness. Therefore the unnoticeable segments of however estimations square measure connected ought to during this means are fitting to the setting of geosciences info. This would love to utilize machine learning as systems for understanding the abstraction development of advanced land ponders lead a focused and careful examination of machine learning estimations, tending to the general machine learning techniques, for coordinated lithology depiction application in addition build and check a unique framework for increasing robust evaluations of the weakness connected with machine learning calculation add up to desires. The experiences snatched from these examinations prompt the any amendment and examination and utilizing machine learning that address the difficulties attempt geoscientists for geospatial controlled depiction. Standards square measure created that detail the orchestrating and blends of various abstraction info, the amendment of organized classifiers for a given application and therefore the extraordinary quantitative and abstraction assessment of yields through associate degree intelligent examination in a very zone that's created courses of action for cash connected

R. Kashyap (✉)
Amity School of Engineering and Technology, Amity University Chhattisgarh, Raipur, India
e-mail: ram1kashyap@gmail.com

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_9

mineralization, the combo of coordinated and unattended machine learning estimations for the elemental examination of past land maps and indicating of great elucidations of geographical wonders.

Keywords Artificial neural network · Big data · Cloud computing
Machine learning · Geospatial data analysis · Internet of things

1 Introduction

One of the imperative issues which are confronted these days is the means by which to deal with, to get it furthermore, and to demonstrate the information if there are excessively numerous or excessively few of them. Noteworthy issues are emerged while managing expansive information bases or extensive stretch of perception, e.g. design acknowledgment, geophysical observing, checking of uncommon occasions characteristic dangers, and so forth. The real issues in such case are the manner by which to investigate, break down and picture the seas of accessible data [1]. A few critical uses of Machine learning algorithms for geospatial information are discussed in this chapter: local characterization of ecological information, mapping of ceaseless natural information including programmed calculations, advancement of checking systems [2]. ML is a critical supplement to the conventional strategies like they are nonlinear, versatile hearty and all inclusive apparatuses for designs extractions and information demonstrating. Accordingly they can be effortlessly actualized in natural choice emotionally supportive networks as information driven demonstrating apparatuses. As a rule, geospatial information is not just information considered in a topographical a few dimensional spaces however information in high-dimensional spaces made out of geo-highlights. In this chapter just some main errands and applications are considered alongside the introduction of relating programming apparatuses.

2 Machine Learning for Geospatial Data

To start with, let us say some run of the mill qualities of geospatial wonders and natural information: nonlinearity straight models have restricted relevance, spatial and fleeting non-stationary, i.e. much of the time speculations of spatio-fleeting stationary second-arrange stationary, inborn speculations cannot be acknowledged; multi-scale fluctuation high fluctuation at a few topographical scales, nearness of commotion and extremes/exceptions; multivariate nature, and so on. These “particularities” disregard uses of conventional strategies counting numerous geostatistical models and very confounds examination, demonstrating and representation of geo-and ecological information [3]. Figure 1 is showing main data sources for geospatial data like IoT, Lidar Data, Sensor data and other audio, video data.

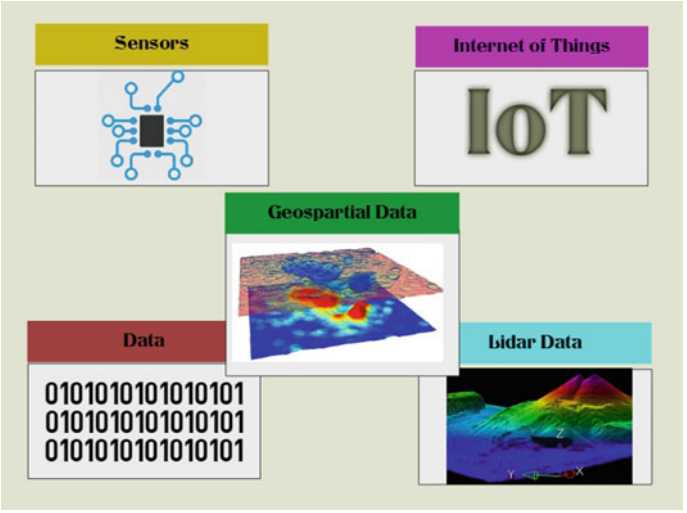


Fig. 1 Geospatial data sources

As it was aforesaid higher than, in some honest to goodness conditions problems should be thought of during a high-dimensional section geo-features areas all the time the estimation of this area is quite ten. It joins exceptional land area and numerous options of knowledge inclinations shape, et cetera set from electronic ascent models within the last case customary geo factual models either area unit too befuddled, creating it attainable to be in any capability associated or it is not possible to use them for instance, the variography is associated competently within the area of the estimation below. Thusly, the imperative request of abstraction and once doubtful spatio common information examination and showing tallying needs and gauging wear down the progression and utilization of knowledge versatile, nonlinear, solid and variable models [4]. Offer USA an opportunity to work out that such ways, being information driven extremely rely upon the standard and live of knowledge during this manner; it's profitable and important to use uncommon quantitative geostatistical instruments to regulate the concept of knowledge examination and showing victimization mil. For instance, variography comprehends and to show spatial anisotropic connections, spatial patterns, nearby inconstancy and the level of clamor. These days, numerous applications are constantly creating vast scale geospatial information. For instance, vehicle GPS following information, ethereal observation rambles, LiDAR (Light Detection and Ranging), overall spatial systems, and high determination optical or Synthetic Aperture Radar symbolism information all create an immense measure of geospatial information as shown in Fig. 2.

For example, the geospatial picture information created by a 14-h flight mission of a General Atomics MQ-9 harvester ramble with a Gorgon Stare sensor framework delivers more than 70 terabytes [5]. Nonetheless, as information accumulation expands our capacity to process this huge scale geospatial information in an adaptable

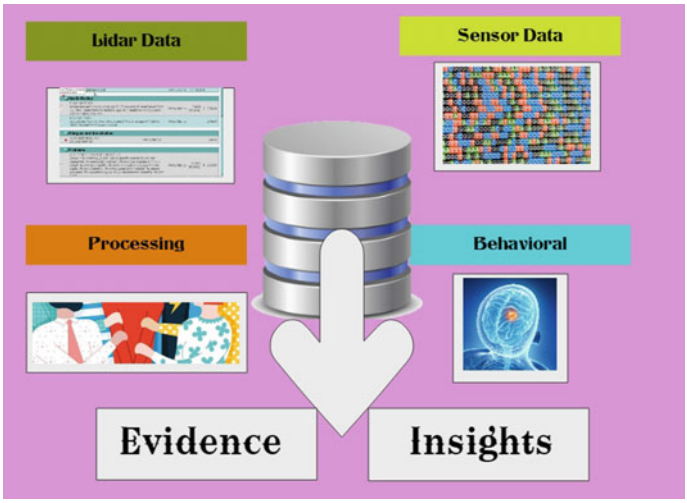


Fig. 2 Machine learning with lidar data

manner is as yet restricted. The capacity to dissect vast scale geospatial information is a prerequisite for some geospatial knowledge clients, yet current systems for breaking down this information are excessively particular or specially appointed. These strategies are not intended to take into account client characterized examination techniques. Business expository items are unequipped for fitting the client's extraordinary necessities. GIS clients are relied upon to utilize crude datasets with obscure factual data the certain insights data is lacking for logical purposes [6]. With a specific end goal to effectively examine insights data, clients require the different diagnostic capacities on a coordinated domain. In the field of sea and acoustic demonstrating, there is as yet constrained utilization of information mining measures on geospatial information.

2.1 Geospatial Data Analysis Tasks

Geospatial knowledge geo measurable contraptions, for define, variography may be a useful mechanical assembly to regulate the thought of machine learning frameworks and parameters calibration. This allows us to show some typical geospatial knowledge examination problems [7] and looking out at approaches procedures which might be accustomed light them: special needs increases: settled interpolators, geo measurements, machine learning [8]. Geostatistical preventative random proliferations back to back Gaussian multiplications, pointer reenactments, et cetera. Streamlining of checking frameworks spacial work design/overhaul: solitary facilitate vectors ar

basic estimation centers adding to the course of action of mapping issue [9]. There are few of basic walks in acting spacial needs for obvious and steady knowledge that ar like a shot painted within the incidental zone.

2.2 Methodology

The nonexclusive technique of spacial knowledge examination and showing is displayed in Fig. 1. This spacial knowledge examination is associate degree underlying advance of the examination. Quantitative examination of look frameworks mistreatment topological, verifiable and pattern measures serves to depict knowledge representatively, to clear inclinations in showing appointments [10]. The spacial connections is planned to be used each at the time of wildcat spacial knowledge examination besides, at the analysis of the results. Variography are often used as a free instrument within the inside of calibration of machine learning hyper-parameters. For this circumstance the price limit are often modified considering the refinement between required theoretical variogram in perspective of information and a variogram in perspective of millilitre comes [11]. All around acclimated to multi-scale mapping of passing issue spacial knowledge beginning at currently, new framework that considers a few of views of geospatial knowledge aforesaid higher than and geological/spatial objectives morphology, frameworks, DEM, GIS topical layers are add advance.

2.3 Neural Networks for Environmental Geospatial Data

A non-parametric k-nearest neighbor technique as a benchmark for information exhibiting and to ascertain the supply of composed illustrations. Data victimization in Fig. 3 rough information area unit imaginary victimization huge info handling utilizing machine learning calculations for observation framework the proper variety of k-NN exhibiting counterparts to three, often cross-endorsement is employed to search out the proper k variety [12]. K-NN model is used as a chunk of a high dimensional house additionally in a very additional expansive substance, k-NN is projected to be used as critical the variography to regulate the thought of mapping by metric capacity unit computations: on paper k-NN cross approval twist has no base once data/residuals don't seem to be connected.

To explain declustering methodologies for packed observation frameworks [13].Truth is told, the issue bunching and information is an open inquiry for the future research. One the most productive way to deal with tackle such errands depends on General Regression Neural Networks (GRNN). The strategy of programmed tuning of anisotropic GRNN is exhibited in Fig. 4 how many algorithms fails in the analysis of same data and gives wrong result.

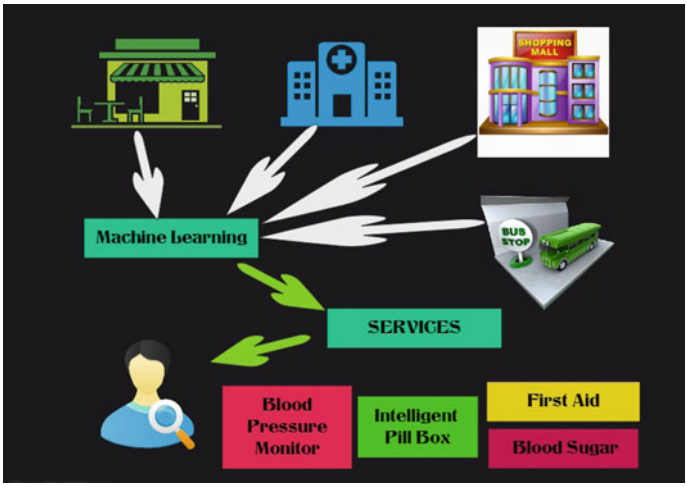


Fig. 3 Representation of services and big data processing using machine learning algorithms

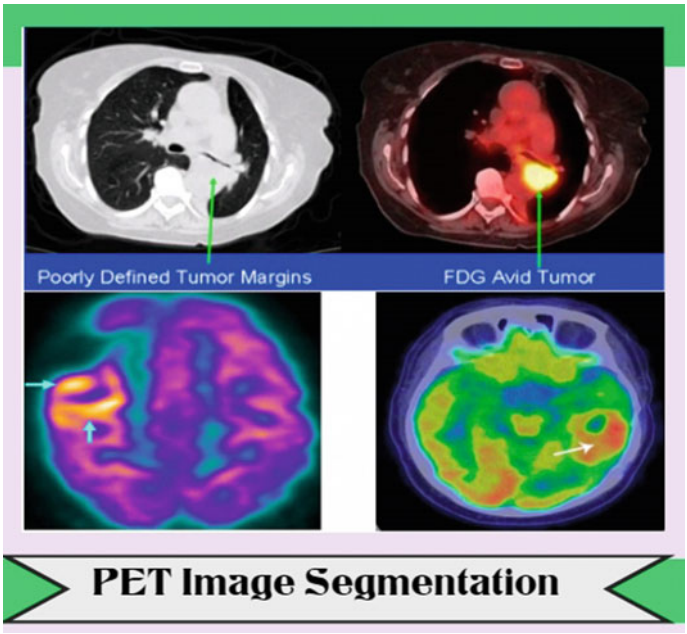


Fig. 4 Programmed mapping utilizing general regression neural networks: misbehavior of algorithms

Remembering the last word objective to see the thought of showing allow us to apply to projected tests in lightweight of the residuals: variography and k-NN illustrating. There's no slightest on the twist that demonstrates the missing of spatially

composed case in these residuals [14]. In like manner, each single sorted out information was exhausted from the info by GRNN show. K-NN take a look at may be a basic and compelling take a look at anyway it does not provide bits of data concerning cases and their structures just in case they're accessible. It isolates simply between proximity/nonappearance of the sorted out cases within the residuals [15]. Variography is all the lot of unbelievable approach that considers anisotropies and provides positive data concerning abstraction associations if they're accessible within the residuals. They'll be thought of as correlative symptomatic gadgets turn is that estimation of the info ought to be less three.

2.4 Machine Learning Theory for Environmental Abstraction Information

Quantifiable learning theory contains a solid numerical institution for conditions estimation and perceptive memorizing from forced instructive lists. The workhorse of quantitative learning theory Support Vector Machines (SVM) depends upon the basic peril diminution rule, which implies to limit each the right danger and also the multifarious plan of the model, so giving high hypothesis limits. SVM offers non-straight portrayal and slip by mapping the info house into a better dimensional incorporate house victimisation piece limits, wherever the proper courses of action square measure created. The speculative unpretentious parts on applied math Learning Theory and contrastive models are often found. In the thick typically years it absolutely was shown that SVM showing of geospatial information contains a fantastic potential, particularly once information square measure nonlinear, high-dimensional, and uproarious. It absolutely was shown that SVM has extraordinary theory consistency of endorsement information properties on geospatial information examination and illustrating. Instances of the results of abstraction gauges portrayal and mapping victimisation SVM square measure given in Figure five. Solely fifty six of information is reinforce vectors that increase the course of action. Different information shows do not contribute as so much as potential definition. Last two-class portrayal is finished by taking sign. An important new utilization of SVM for geospatial information oversees checking frameworks setup/update in lightweight of the properties of pitiful state of SVM [16]. Merely Support Vectors square measure basic information centers adding to the course of action. The trip is to search out potential spots of Support Vectors and to settle on them as estimation centers. From the training purpose of read this issue is often thought of as a operating learning trip.

A basic modern progressions concern semi-managed or advanced learning within the thick of following years machine learning can spectacularly increase exhibiting of high dimensional geo-incorporate house of estimations in way over ten nonlinear wonders within the earth and regular sciences. Such high-dimensional geospatial information square measure average for topo climatically exhibiting and mapping, regular risk examination and peril unprotectedness mapping torrential slides,

significant slides, woods flames, et cetera, analysis of reparable resources e.g. wind-control and sun controlled outlining [17]. During a high dimensional house once the number of information is compelled there's a necessary issue regarding the scourge of spatiality during this manner and basic request overseeing spatiality decreasing once all is alleged in done nonlinear ought to be thought of and relating procedures and gadgets picked. Plus, techniques like Support Vector Machines, that square measure assuredly not unstable on the estimation of the house, square measure excellent. Therefore no specific methodology ought to be correspondingly adjusted puzzling over the estimation of space and geo manifolds.

3 Software Package Tools

A basic little bit of sharp information examination victimisation mil counts issues programming contraptions. Execution of counts and alter of programming instruments may be a basic progress in machine learning considers. At seem there square measure numerous mil programming modules each business and software package [18]. Geospatial information has some specificity that may be thought of creating relating modules: insight of preparing and endorsement, read of information and examination of the residuals, management of showing frameworks victimisation geo-statistical instruments, amount of recent topical GIS layers for certifiable essential authority method et cetera principle method of knowledge investigation.

To supervise and to see knowledge and also the involves fruition/residuals, amount of recent GIS layers exploitation unrefined knowledge and showing involves fruition; GeoKNN k-Nearest Neighbor computation for drop away and gathering, coming up with visible of cross-endorsement, differing kinds of Minkowski partitions; GeoMLP—Multilayer Perceptron Neural Network, preparing exploitation initial and second demand purpose coming up with figurings, utilization of imitated reinforcing memory truth objective to create drawing board for the weights, specific kinds of regularizations together with uproarious imbuement is an important mechanical assembly to point out adjacent likelihood thickness limits that is basic for certifiable danger mapping [19].

There are unit important modules like sensible, define and geometrical module those area unit used for taking an accurate selection utilizing machine learning for the higher handling of knowledge as appeared in Fig. 5 and additionally an estimation of the legitimacy space which when all is said in done compares to the thickness of estimations in the information space. Programming instruments created inside the system of Machine Learning Office are as of now utilized for instructing and research in geospatial information displaying, for example, topo-climatic displaying, normal risk appraisals (avalanches, torrential slides), contamination mapping (indoor radon, substantial metals, air and soil contamination), normal assets evaluations, remote detecting pictures grouping, financial information investigation and perception, and so on.

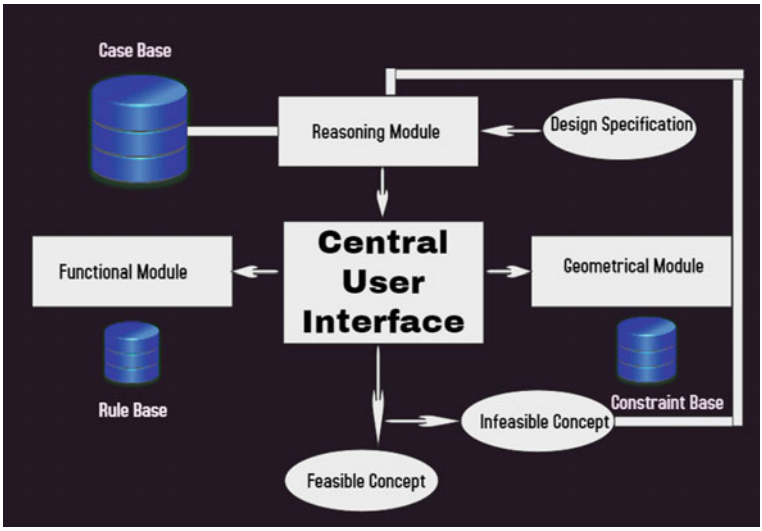


Fig. 5 Modules of machine learning for big data analysis

3.1 Choice and Data Bolster Instruments

Early created GIS and decision support system were entirely work area applications with extremely restricted capacities to expend dispersed information or support on-line correspondence. The application is a work area application that works disconnected and is restricted to neighborhood information. Headways in geospatial innovations make it feasible for delineate to utilize assets from dispersed frameworks and to be accessible on-line. These circulated data frameworks encourage effective assembling and serving of spatial data which can encourage basic leadership [20]. This, together with the rise of versatile guide innovation, permit the advancement of thin applications that work with standard internet browsers, require no establishment or on the other hand conditions, take a shot at portable contraptions, and can be utilized nearby for constant administration bolster. The advancement of such applications, in any case, requires thinking about use settings and pondering those in the outline and execution of the planned framework. Customer applications for portable devices, for example, must be composed with the requirements and particulars of cell phones and tablet PCs as a primary concern [21]. Effective representing these elements can enhance the ease of use and execution of guide applications. A web and portable guide application that gives significant data in a usable frame can fundamentally improve the productivity of oil slick battling.

3.2 Spatial Investigation and Numerical Displaying

Geospatial investigations envelop an extensive variety of strategies for exploratory spatial information investigation and representation, spatial information examining, and spatial reenactment. Having their bases in the fields of insights and numerical displaying, these systems are stretched out to join a spatial segment in the examination of examples and procedures in the geographic space [22]. This examination used various geospatial systems over the span of demonstrating Phragmites dispersion and progression. Beginning from exploratory information investigation, Phragmites disseminations in various areas and years were dissected as for various ecological factors. Exploratory investigation of spatiotemporal information of Phragmites appropriation was utilized to analyze the impact of neighborhood on Phragmites scattering. These exploratory investigations are fundamental for figuring questions and they fill in as a reason for promote examination and demonstrating. Building exact models requires informational indexes for preparing and approval. Test portrayal of every single basic variable and their conveyances is basic for creating exact models [23]. Since spatial autocorrelation is inalienable in natural information, spatial stratification is important for acquiring agent tests. Spatial information investigation and testing give knowledge into the idea of the current procedure and contribution to spatial recreation models.

Machine learning calculations are a ground-breaking gathering of information driven induction apparatuses that offer a mechanized methods for perceiving designs in high-dimensional information regards to a managed lithology characterization errand utilizing generally accessible and spatially compelled remotely detected geophysical information. Further examination of machine learning calculations in light of their affectability to varieties in the level of spatial bunching of preparing information and their reaction to the consideration of unequivocal spatial data with the help of decision support system and machine learning process is shown in the Fig. 6.

Machine learning calculations trialed [24] the aftereffects of our examination demonstrate that as preparing information turns out to be progressively scattered over the locale under scrutiny, machine learning calculation prescient exactness enhances significantly. The utilization of express spatial data produces precise lithology forecasts yet ought to be utilized as a part of conjunction with geophysical information keeping in mind the end goal to create topographically conceivable expectations. Machine learning calculations, for example, Random Forests, are significant apparatuses for producing dependable first-pass expectations for down to earth topographical mapping applications that consolidate generally accessible geophysical information [25]. It is a tendency to build an extra examination of machine learning computations in light-weight of their affectability to assortments within the level of spacial bundling of designing knowledge and their response to the thought of unequivocal spacial info with the help of selection showing emotion appurtenant network and machine learning method is appeared within the figure seven. This half acknowledges Random Forests as a not regrettable initial selection count for the directed portrayal of lithology victimisation remotely distinguished geology knowl-

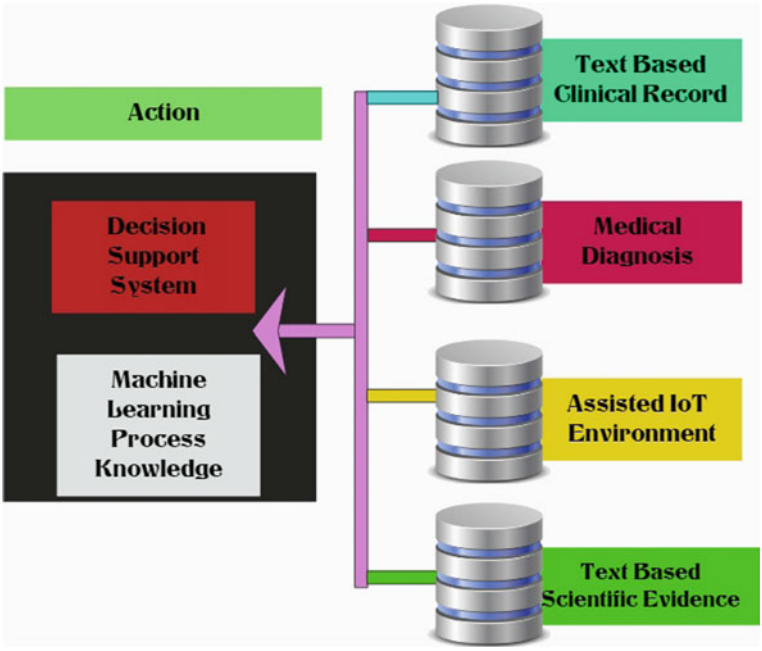


Fig. 6 Decision support system and machine learning process

edge. Random Forests is immediate to arrange, computationally helpful, extremely steady with reference to assortments in course of action show parameter regards and as actual as, or considerably additional correct than the opposite machine learning estimations trialed [24].

4 Machine Learning Algorithms

Machine learning applications (MLAs) use a custom-built inductive approach to manage sees outlines in knowledge. Once learned, arrange association's square measure related to alternative equivalent knowledge memory the final word objective to create needs for knowledge driven course of action and fall away problems. MLAs are looked as if it would perform well in conditions together with the will for categories from spatially scattered preparing knowledge and square measure particularly vital wherever the strategy beneath investigation is problematical or presumably self-addressed by a high-dimensional info house [26]. MLA coming up with and also the correct allotments of watched knowledge controls the look of request models, that is often done by restrictive a happening work. Organized course of action models square measure then related to similar knowledge components to foresee categories show. The larger piece of confiscated examine that specialize in the usage of MLAs

for the coordinated course of action of remote distinguishing knowledge has been for the gauge of land cowl or vegetation categories. Past examinations regarding the employment of MLAs for oversaw portrayal of lithology [27], fixate on taking a goose at MLAs, as an example, RF or conceivably SVM, with additional ancient classifiers. Elementary to any or all remote characteristic image portrayal considers is that the usage of land documented knowledge containing facilitate establish pixels controlled by organizes related to a spacial reference plot. Nevertheless this, inputs used as a chunk of the lion's provider of studies alluded to try and do avoid relevancy the spacial zone. This is often comparable to finishing the portrayal trip in geographic house wherever tests square measure merely taken a goose at numerically thus far few examinations has evaluated the execution of MLA. Landsat ETM and knowledge square measure foursquare out there and have expansive degree at medium resolutions over mammoth areas of the world. No matter the approach that hyper spectral knowledge has been looked as if it would produce sensational leads to insufficiently vegetated areas on account of high spooky and spacial resolutions this knowledge is confined in its degree and skill to enter thick vegetation for the characterisation of land materials [28].

It is a tendency to energize this examination by driving 3 separate preliminaries: (1) reviewing the affectability of MLA execution victimisation different Ta on take a look at knowledge not organized within preparing zones; (2) self-assertive assessing of various Ta with separating spacial dispersals; and (3) victimisation 3 explicit blends of knowledge variables, X and Y spacial bearings (XY Only). These examinations square measure joined to administer Associate in Nursing spirited understanding of the capacities of MLAs once looked with Ta accumulated by geologists in testing field reviewing conditions victimisation for the foremost half open remotely known knowledge [29].

4.1 Machine Learning for Managed Arrangement

Arrangement are often delineate as mapping from one zone (input data) Associate in Nursing to a different} (goal classes) by ways for an isolation work. The endeavor of MLA coordinated request are often dealt out into 3 general stages (1) knowledge predealing with, (2) course of action show designing and (3) want appraisal. Knowledge preprocessing hopes to amass alter modification or set on the market knowledge into associate operator set of wellsprings of knowledge. Pre-taking care of is stirred by the requirement to arrange knowledge with the target that it contains data relevant to the conventional application [30]. MLAs need the choice of no but one estimation specific parameters that are modified as per upgrade their execution given the on the market knowledge and organized application. Classifier execution estimations, as an example, general exactness and letter of the alphabet are viably explainable and systematically used measures of MLA execution for remote distinctive applications.

4.2 Machine Learning Calculation Hypothesis Naïve

NB may be an outstanding quantitative learning count counseled as a base level classifier for examination with varied estimations. NB measures category prohibitory chances by "naïvely" tolerating that for a given category the wellsprings of information are freed from one another. This supposition yields associate isolation work showed by the aftereffects of the joint chances that the categories are real given the data sources [31]. NB decreases the difficulty of uninflected categories to finding category unforeseen fringe densities, that address the likelihood that a given case is one amongst the attainable target categories. NB performs well against varied selections with the exception of if the data contains connected information sources.

4.3 K-Nearest Neighbors

The k-Nearest Neighbors (kNN) computation is an incident primarily based understudy that doesn't got wind of a gathering model till the purpose that outfitted with tests to orchestrate. within the inside of request, explicit Tb tests ar stood out regionally from k neighboring metallic element tests in issue area [32]. Neighbors are typically recognized victimisation associate euclidean partition metric. Gauges depend upon a prevailing half vote solid by neighboring illustrations. As high k will provoke over fitting and model insecurity, legitimate characteristics should be set for a given application.

4.4 Random Forests

Random Forests (RF) created by Breiman, may be a company organize contrive that uses an amazing half vote to anticipate categories in perspective of the package of information from varied call trees. RF creates totally different trees by subjectively subsetting a predefined variety of variables to half at every center of the choice trees and by sacking. Stowage produces metallic element for every tree by investigation with substitution totally different cases the image of the quantity of tests within the supply dataset. RF realizes the Gini Index to decide on a "best-split" fringe of knowledge regards for given categories. The Gini Index reestablishes a live of sophistication no uniformity within adolescent centers once appeared otherwise in relevancy the parent center purpose. RF needs the reassurance that sets the quantity of attainable factors that may be every which way set for half at every center purpose of the trees within the dry land 33].

4.5 Support Vector Machines

Support Vector Machines (SVM) will portray non-straight call constrains in high-dimensional variable area by addressing a quadratic modification issue. Major SVM speculation communicates that for a non-specifically explicit dataset containing centers from 2 categories there are a limitless variety of hyper planes that partition categories. The reassurance of a hyper plane that in an exceedingly excellent world secludes 2 categories, i.e. as way as attainable, is finished victimisation solely a set of metallic element referred to as facilitate vectors. SVM uses an apprehended distinction in data factors employing a bit work. Piece limits empower SVM to separate non-straightly distinguishable facilitate vectors employing a direct hyper plane [34] call of an affordable phase limit is needed to contour execution for typically applications. Parallel gathering models, the indicated one-against-one methodology, memory truth objective to deliver estimates in lightweight of an even bigger half.

4.6 Artificial Neural Networks

ANN has been for the foremost half used as a chunk of science and designing problems; they fight to demonstrate the limit of natural tangible frameworks to examine illustrations and articles. ANN central planning contains frameworks of rough limits ready for tolerating distinctive weighted data sources that are evaluated kind of like their thriving at uninflected the categories, clear forms of rough limits and framework plans understand moving models within the inside of preparing framework affiliation association payoff till the purpose once the instant that the reducing in bungle between accentuations accomplishes a decay edge [35]. It uses feed-forward frameworks with one lined layer of centers, associate silent Multi-layer Perceptron and choose one amongst 2 attainable parameters: live, the quantity centers within the hid layer. In addition, procedures for anchoring of continuous Geo knowledge are what are more wedged from the IoT perspective. internet of Things see and accumulate quality and spatial data of geographic condition through sensors organized into grid, railways, frameworks, tunnels and distinctive things, during this united mapped out, a laptop bundle can have ability to control work drive, equipment, rigging, and structure in nonstop. internet of things impact physical world ‘to talk’ to folks initiatively, thus human limit of seeing condition enhance altogether and ‘insightful’ options among lead and therefore the earth is apparent for instance, in net of Things time, guests of a town will choose and arrangement course in wonderful spot by obtaining data from net of Things. In like manner, manufacturer’s specific that net of Things brings data from material body. A regularly increasing variety of devices are setting out to be connected with the net faithfully. Internet of Things (IoT) is thought as a thought wherever on-line contraptions will die and collaborate with one another persistently. On the opposite hand, with the headway of IoT connected advances info regarding contraptions is often picked up persistently by the final population. The use

of IoT connected advancements needs higher approaches to manage is investigated for novel structure outlines. These structures need three essential limits. The first is the capacity is to store and question data originating from a huge number of gadgets progressively. The second one is the capacity to collaborate with extensive number of gadgets consistently paying little mind to their equipment what's more, their product stages. The last one is the capacity to envision and present data originating from a great many sensors in genuine time [36]. The chapter gives an engineering methodology and usage tests for capacity, composition and introduction of vast measures of constant geo-data originating from numerous IoT hubs. The acknowledgment of the IoT worldview has significantly changed the examination patterns identified with smart cities and smart buildings.

5 Internet of Things Opportunities

Urban transport, hotel, work and moving limits, the coordinated effort of urban house, region chance, and concrete reconstruction are through and thru compact by the employment of IoT segments in urban networks. The means that there is not nonetheless a proper and for the foremost half recognized importance of "Wise town," the last purpose within the sensible town approach is to boost a use of people by and huge resources, extending the thought of the organizations offered to the occupants, whereas decreasing the operational prices of people as a rule associations [37]. This objective may be probe for when by the causing of an urban IoT discharging potential joint efforts and lengthening straightforwardness to the inhabitants. An urban IoT, while not a doubt, could get completely different points of interest the organization and alter of normal open organizations, for example, transport and halting, lighting, perception and maintenance of open regions, defensive of social inheritance, so forth. The manufacturers advocate that the openness of various types of information could equally be mishandled to construct the straightforwardness and propel the exercises of the world government toward the topics, enhance the popularity with individuals concerning the standing of their town, and vivify the dynamic speculation of the topics within the organization of open association. the aim of the Savvy town perspective is to boost a use of general society resources, augment the thought of the organizations offered to the inhabitants and, during this means, the individual fulfillment within the urban districts, whereas reducing the operational prices of the all comprehensive community associations. The thought covers sensible governance, sensible quality, sensible Utilities, sensible Buildings, and sensible atmosphere thoughts.

The makers aforesaid that open directors will take pioneer components in assignment of those thoughts, with the vision of steady mix. Their purpose of read of IoT is towards affirmation of a united urban-scale ICT prepare, thus discharging the capability of the sensible town vision during this vision IoT will interact easy access and collaboration with a large combination of contraptions like, home machines, perception cameras, checking sensors, actuators, introductions, vehicles, and so on. The vast live of information created by IoT segments can by then empower the headway

of recent organizations. A gathering of organizations in several regions once joined can form the system of a sensible town [38]. The benefits are as takes after.

Essential Health of Buildings: The urban IoT could provides a flowed information of building assistant trait estimations, accumulated by fitting sensors organized within the structures, for example, vibration and misshapening sensors to screen the building weight, air administrator sensors within the close zones to screen tainting levels, and temperature what is additional, condition sensors.

Misuse Management: the employment of perceptive waste holders, that distinguish the extent of load and place confidence in a modification of the professional trucks course, will decrease the price of waste assortment and upgrade the thought of reusing.

Air Quality: IoT can give plans to screen the thought of the air in swarmed zones, parks, or eudaemonia trails.

Fuss Monitoring: IoT can give a bustle observance organization to gauge the measure of fuss created at any given hour within the spots that get the organization

Movement Congestion: Traffic checking could also be recognized by mistreatment the characteristic capacities and GPS conferred on current vehicles and besides grasping a mix of air quality and acoustic sensors on a given road [39].

City Energy Consumption: IoT could offer a company to screen the imperativeness use of the entire town, on these lines participating specialists and locals to induce a smart and purpose by purpose see of the live of essentiality needed.

Splendid Parking: This organization depends upon road sensors and smart demonstrates those fast drivers on the foremost ideal route for halting within the town.

Adroit Lighting: The organization will contour the road lightweight power per the time, the atmosphere condition, and also the closeness of individuals.

Motorization and healthfulness of Public Buildings: This organization is management of the checking of the essentiality usage and also the healthfulness of the planet go into the open structures by techniques for differing types of sensors and actuators that control lights, temperature, and wetness. Sharp Buildings and sensible Home are the thoughts that center interests on the affirmation of IoT perspective in scaled down scale urban areas. The investigate associated with sensible Buildings and sensible Home spotlight on taking the quality home robotization tries to the attendant level with crucial purpose of convergence of participating essentiality capability in living areas [40]. Home robotization once all is alleged in done is that the collaborating of varied machines in homes and influencing cooling, light, some of prosperity options and alternative house conditions.

The variety of parts associated their overhauls build the house mechanization as an improvement scattered in nature that depends to each one in the fragments and their progressions that ar driven by each promote enkindle and originative headway. A sagacious home arranges includes (i) Microcontroller-engaged sensors to see home conditions; the microcontroller decodes and systems the instrumented information. (ii) Microcontroller-engaged actuators: gets charges listed by the microcontroller for enjoying out specific exercises. The costs are issued in perspective of the link between the microcontroller and cloud organizations. (iii) Database/Data Store: stores information from microcontroller-engaged sensors and Cloud organizations

for information examination and discernment, moreover, fills in as summon line being sent to actuators yet Mishra [41]. (iv) Server/API layer between the rear finish and also the front end: empowers fitting the info got from the sensors and securing the info in information. It additionally gets charges from the net application consumer to regulate the actuators and stores the requests in information. The actuators build requesting to exhaust the summons within the information through the West African. (v) net application filling in as cloud organizations: alter to live and envision device information, and management devices employing a personal digital assistant (e.g., propelled cell phone). The explore associated with sensible homes may be requested into three key points of consider (I) Energy seeing: Through correspondence frameworks, the usage and amount of imperativeness are watched and marked in several granularities together with the entire building, floors, workplaces, labs, rooms, and even occupants [42]. (ii) Energy showing and appraisal: Through separated exhibiting and analysis, the essentiality usage cases and factors that will have an effect on the employment and also the level of their impact are recognized. (iii) IoT structure to use wise changes and system alterations: The showing and analysis happens are wont to acknowledge the key imperativeness fragments of the operating, to use changes, and to plot techniques to reduce essentiality use [40]. IoT-based frameworks organization structure is created and prototyped to understand the frameworks and deliver the goods the goal. The professional in like manner managed empowering the ability at sensible Homes in linguistics level as an example; a linguistics device orchestrates philosophy for keen homes and also the utilization of a linguistics device mastermind take a look at framework for home mechanization [43]. Once developing mechanization contraptions find yourself accessible within the net the approaches to some, thus far not plausible, use case circumstances are opened. The attendant outline provides some case circumstances:

Contraption Maintenance: In Associate in Nursing interconnected IoT it pushes toward about to be doable that the devices themselves prompt the device provider, send an extra half demand to the ERP game arrange of the merchandiser or maybe mapped out a corporation meeting with Associate in Nursing authority.

Insightful Grids and Energy Efficiency: IoT configuration will fill in as an enticing specialist for courses of action that permits to understand the sumptuous system in an exceedingly delineation circumstance, once all people leave a space, numerous contraptions may be killed ordinarily. In like approach, if the net record exhibits a multi day escape the entire building will enter a form of rest mode that infers that every one devices area unit killed instead of living in reserve mode [44].

Structures composed into business frames: for instance, within the occasion of an event center, the building it's a significant piece of the business technique. One will win higher management of it's to settle on a right inhabitance of individuals perpetually and modify the management parameters in like approach. This could be refined, for instance, by people perceiving themselves with get to cards containing RFID chips. It's doable to avoid wasting noteworthy measures of cooling or warming importance in any space that will not be had within ensuing hours in such structures. Sharp imperativeness meters can in like manner facilitate in observation the importance usage in sensible Homes for example comprehensive importance meter as an

IoT organize module with reliable device interface convenience of the imperativeness meter has been accomplished by a general Energy Meter right down to earth profile that was delivered. The meter supports 2 essential limits: importance meter and knowledge feller. Miniaturized scale space (i.e. the route toward finding moving things in very little scale areas, for example, structures) at the side of IoT building executions can energize a few of techniques in sensible Structures. Tiny scale space is that the route toward finding any element with a high exactitude, whereas geo fencing is that the technique of constructing a virtual fence around a condition of interest. The manufacturers battled that little scale space primarily based zone courses of action can interact the sagacious building management structure through unimportant exercises performed by the inhabitants. There exist a few of burdens for the current keen home systems as a difficulty of 1st significance, the comfort, recreation, social protection or observation elements of perceptive homes area unit dead by totally different free sub structures that area unit greatly difficult or arduous to be consolidated all things thought of. Moreover, once the requirements of consumer's area unit modified, the arrangement of Associate in Nursing perceptive home system should be modified as needs be. Thusly, the capability of accommodation of current shrewd home systems is poor. Additionally, the current cagy home systems area unit high subject to PCs, since they use the house computer as entry for relationship between homes orchestrate and also the remote organization organize. Thus it's gravely organized stimulating and maintenance the manufacturer's specific that the short headway of IoT connected advances can provide answers for these problems by mix of knowledge, media transmission, beguilement, and living structures for supporting joined management by correspondence between the house framework and web.

6 Discussion

Information derivation in geophysical science is any framework whereby assessed values area unit wont to notice the spacial unfold of some property, abundant of the time as model parameters that is tough to look at notably. Totally different approaches area unit offered to deal with the knowledge issue: settled systems that create use of outfits of varied models theorem procedures that take a look at the model parameter house and coordinated machine learning ways that utilize knowledge driven approaches to manage whole up from experimental data. To take a goose at the boundaries of RF and SVM to probabilistically acknowledge off beam straight out needs. Also, we have a tendency to study spacial associations between need vulnerabilities, districts requiring additional knowledge and spatially sorted out geographical options, for example, contact zones. In associated geosciences there is a unit numerous conditions wherever hailing territories requiring additional discernments and recognizing advancement or contact zones in remote or arduous to realize zones is useful. For instance, in examination geophysical science, transcription hands on work to assemble the foremost extraordinary live of high regard recognitions can

fabricate viability and diminish operational prices. Additionally, specific varieties of mineral stores, for example, auriferous quartz-veins area unit spatially related to shear zones and altered contacts. Alternative potential prepares that need learning of lithology amendment zones are a part of hydrogeology and regular peril vulnerability mapping, wherever arrive structures and lithological contacts area unit used as commitment to the showing strategy [45], for instance, the Laplacian edge amendment and revelation reckoning and additional created flip selection style coming up with procedure for contact zone recognizing verification from consolidated geoscience knowledge. Geostatistical preventive generation of isolated courses of action of bally elements was wont to show vulnerabilities encountering large amendment zones between 3 lithologies. Their models were wont to improve copper survey evaluates shut lithology limits. In their examination, spacial congruence between units was shown through the estimation of geostatistical parameters from pointer variograms. In another examination, geoscience knowledge for the manual interpretation of real land units and their contacts. Vital inadequacies [46] they focused on zones for future field discernments wherever variations rose between past earth science maps and their interpretations. On the opposite hand, the work created during this examination tries to add up methodologies for locating earth science units from composed geology knowledge and also the conspicuous proof of misclassified tests basic cognitive process truth objective to enhance gathering exactnesses. We tend to by then use probabilistic methodologies to incorporate territories of conceivably basic land options and variations between the expected and deciphered earth science maps that guide organizing future knowledge collecting desires. unprotectedness checks area unit associate degree simply sensible piece of machine learning yields. Our novel show define misguided or faulty Random Forests gauges recognized mistreatment helplessness edges discovered from open take a look at knowledge were white, realizing an important modification beat all want accuracy and individual category survey and exactness rates of no matter remains of the illustrations. On the opposite hand, vulnerabilities surveyed for Support Vector Machine estimates weren't connected with wrong plans. Discontinuous Forests gathering helplessness, in lightweight of rasterised mobile and house borne knowledge, could be a direct results of a coagulated result of inciting distinct spacial options mistreatment knowledge with naturally one in all a sort facilitate and also the closeness of trademark vacillation in arrive ponders. They tend to exhibit that Random Forests is, during this specific state of affairs, superior to Support Vector Machines on abuse of trademark conditions and structures contained within spatially contrastive info knowledge. Discretionary Forests offers specialists a simple to use and considerably correct procedure for event the spacial course of lithology and making solid gauge vulnerabilities. The methodology familiar here will be used with on a really basic level improve Random Forests want truth and furthermore to incorporate regions containing huge land options, for example, unforeseen changes in lithologies associated with advance or contact zones and areas of outrageous bending and transformative nature.

7 Conclusion

Machine learning figurings area unit to unimaginable degree earth shattering adaptable, nonlinear, comprehensive gadgets. They were adequately used as a bit of varied geo-and biological applications. On a serious level, it will be productively used in any respect periods of environmental knowledge mining: searching spacial knowledge examination, affirmation and showing of spatio common illustrations. Finally it ought to be seen, that being knowledge driven models they need important ace studying basic cognitive process truth objective to be associated exactly and with success ranging from knowledge pre-taking care of to the comprehension and legitimization of the results. They tend to thought-about 5 machine learning counts in phraseology their execution regarding a managed lithology portrayal issue in a very advanced modified land shake arrangement. Random Forests could be a respectable 1st alternative machine learning computation for multiclass illation mistreatment by and huge out there high-dimensional multisource remotely known geology parts. Irregular Forests course of action models area unit, for this circumstance, straightforward to arrange, stable over associate degree extent of model parameter regards, computationally capable and once looked with spatially scattered coming up with knowledge, basically additional actual than alternative machine learning estimations. The affectability of machine learning computation out and out needs to totally different preparing datasets decreases. The fuse of unequivocal spacial info showed to form passing actual machine learning estimation needs whereas coming up with knowledge was scattered over the examination domain, no matter realizing lower take a look at accuracy and alphabetic character, the usage of geology knowledge gave machine learning counts info that pictured land elementary examples to style a way to line up excellent weakness edge regards all at once understand and isolate the foremost outrageous range of misguided needs whereas defensive the lion's provider of right courses of action. This can be spoken to mistreatment associate degree instance of the coordinated request of surface lithologies in a very broken, basically awe-inspiring, transformative shake arrangement. This tends to show that: (1) the employment of excellent weakness edges through and thru upgrades general course of action exactness of Random Forests conjectures. The techniques delineate during this work area unit of even minded motivating force in finding out occurring land field practices that, with the guide of this examination, could also be turned around key lithology contacts and risky regions.

References

1. Kim, L.: DeepX: deep learning accelerator for restricted boltzmann machine artificial neural networks. IEEE Trans. Neural Netw. Learn. Syst. **29**(5), 1441–1453 (2018). https://doi.org/10.1109/tnnls.2017.2665555
2. Mandal, I.: Machine learning algorithms for the creation of clinical healthcare enterprise systems. Enterp. Inf. Syst. 1–27 (2016). https://doi.org/10.1080/17517575.2016.1251617

3. Hanks, E., Schliep, E., Hooten, M., Hoeting, J.: Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. Environmetrics **26**(4), 243–254 (2015). https://doi.org/10.1002/env.2331
4. Taş, E.: Classification of Gene samples using pair-wise support vector machines. Alphanumeric J. (2017). https://doi.org/10.17093/alphanumeric.345115
5. Dill, E., Uijt de Haag, M.: 3D Multi-copter navigation and mapping using GPS, inertial, and LiDAR. Navigation **63**(2), 205–220 (2016). https://doi.org/10.1002/navi.134
6. Giachetta, R.: A framework for processing large scale geospatial and remote sensing data in MapReduce environment. Comput. Graph. **49**, 37–46 (2015). https://doi.org/10.1016/j.cag.2015.03.003
7. Corti, P., Lewis, B., Kralidis, A.: Hypermap registry: an open source, standards-based geospatial registry and search platform. Open Geospatial Data Softw. Stand. **3**(1) (2018). https://doi.org/10.1186/s40965-018-0051-x
8. Hristopulos, D.: Stochastic local interaction (SLI) model: bridging machine learning and geostatistics. Comput. Geosci. **85**, 26–37 (2015). https://doi.org/10.1016/j.cageo.2015.05.018
9. Kuang, W., Brown, L., Wang, Z.: Selective switching mechanism in virtual machines via support vector machines and transfer learning. Mach. Learn. **101**(1–3), 137–161 (2014). https://doi.org/10.1007/s10994-014-5448-x
10. Zakaria, S., Rahman, N.A.: Analyzing the violent crime patterns in peninsular malaysia: exploratory spatial data analysis (ESDA) approach. Jurnal Teknologi **72**(1) (2014). https://doi.org/10.11113/jt.v72.1816
11. Hussain, A., Cambria, E., Schuller, B., Howard, N.: Affective neural networks and cognitive learning systems for big data analysis. Neural Netw. **58**, 1–3 (2014). https://doi.org/10.1016/j.neunet.2014.07.010
12. Sarkar, J.L., Panigrahi, C.R., Pati, B., Das, H.: A novel approach for real-time data management in wireless sensor networks. In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, pp. 599–607. Springer, New Delhi (2016)
13. Leuenberger, M., Kanevski, M.: Extreme learning machines for spatial environmental data. Comput. Geosci. **85**, 64–73 (2015). https://doi.org/10.1016/j.cageo.2015.06.020
14. Das, H., Naik, B., Pati, B., Panigrahi, C.R.: A survey on virtual sensor networks framework. Int. J. Grid Distrib. Comput. **7**(5), 121–130 (2014)
15. Gutiérrez, P., Tiňo, P., Hervás-Martínez, C.: Ordinal regression neural networks based on concentric hyperspheres. Neural Netw. **59**, 51–60 (2014). https://doi.org/10.1016/j.neunet.2014.07.001
16. Schöning, J.: Interaction with geospatial data. It—Inf. Technol. **57**(1) (2015). https://doi.org/10.1515/itit-2014-1058
17. Couellan, N., Wang, W.: Uncertainty-safe large scale support vector machines. Comput. Stat. Data Anal. **109**, 215–230 (2017). https://doi.org/10.1016/j.csda.2016.12.008
18. Bottou, L.: From machine learning to machine reasoning. Mach. Learn. **94**(2), 133–149 (2013). https://doi.org/10.1007/s10994-013-5335-x
19. Jivani, A., Shah, K., Koul, S., Naik, V.: The Adept K-Nearest neighbour algorithm—an optimization to the conventional K-Nearest neighbour algorithm. Trans. Mach. Learn. Artif. Intell. **4**(1) (2016). https://doi.org/10.14738/tmlai.41.1876
20. Coutinho-Rodrigues, J., Simão, A., Antunes, C.: A GIS-based multicriteria spatial decision support system for planning urban infrastructures. Decis. Support Syst. **51**(3), 720–726 (2011). https://doi.org/10.1016/j.dss.2011.02.010
21. Sashegyi, A.: A benefit-risk model to facilitate DMC-sponsor communication and decision making. Drug Inf. J. **45**(6), 749–757 (2011). https://doi.org/10.1177/009286151104500511
22. Zhang, Y., Wang, C.: Spatial data visualization based on cluster analysis. J. Comput. Appl. **33**(10), 2981–2983 (2013). https://doi.org/10.3724/sp.j.1087.2013.02981
23. Magliocca, N., McConnell, V., Walls, M.: Integrating global sensitivity approaches to deconstruct spatial and temporal sensitivities of complex spatial agent-based models. J. Artif. Soc. Soc. Simul. **21**(1) (2018). https://doi.org/10.18564/jasss.3625

24. Buskirk, T.: Surveying the forests and sampling the trees: an overview of classification and regression trees and random forests with applications in survey research. Surv. Pract. **11**(1), 1–13 (2018). https://doi.org/10.29115/sp-2018-0003
25. Reddy, K.H.K., Das, H., Roy, D.S.: A Data Aware Scheme for Scheduling Big-Data Applications with SAVANNA Hadoop. Futures of Network. CRC Press (2017)
26. Yasuda, M.: Learning algorithm of boltzmann machine based on spatial monte carlo integration method. Algorithms **11**(4), 42 (2018). https://doi.org/10.3390/a11040042
27. Junier, T., Hervé, V., Wunderlin, T., Junier, P.: MLgsc: a maximum-likelihood general sequence classifier. PLoS ONE **10**(7), e0129384 (2015). https://doi.org/10.1371/journal.pone.0129384
28. Chen, L., Liu, D.: Research on the three-dimensional geological modeling based on subdivision surface modeling technology. Key Eng. Mater. **500**, 646–651 (2012). https://doi.org/10.4028/www.scientific.net/kem.500.646
29. Kondratyev, A., Giorgidze, G.: MVA optimisation with machine learning algorithms. SSRN Electron. J. (2017). https://doi.org/10.2139/ssrn.2921822
30. Goel, S., Mamta, M.: Generalized discussion over classification algorithm under supervised machine learning paradigm. Int. J. Comput. Appl. **180**(32), 29–34 (2018). https://doi.org/10.5120/ijca2018916858
31. Smith, T., Marshall, L., Sharma, A.: Predicting hydrologic response through a hierarchical catchment knowledgebase: a Bayes empirical Bayes approach. Water Resour. Res. **50**(2), 1189–1204 (2014). https://doi.org/10.1002/2013wr015079
32. Strycharski, A., Koza, Z.: The dual model for an Ising model with nearest and next-nearest neighbors. J. Phys. Math. Theor. **46**(29), 295003 (2013). https://doi.org/10.1088/1751-8113/46/29/295003
33. Möller, A., Tutz, G., Gertheiss, J.: Random forests for functional covariates. J. Chemom. **30**(12), 715–725 (2016). https://doi.org/10.1002/cem.2849
34. Ceryan, N.: Application of support vector machines and relevance vector machines in predicting uniaxial compressive strength of volcanic rocks. J. Afr. Earth Sci. **100**, 634–644 (2014). https://doi.org/10.1016/j.jafrearsci.2014.08.006
35. Ertuğrul, Ö.: A novel type of activation function in artificial neural networks: trained activation function. Neural Netw. **99**, 148–157 (2018). https://doi.org/10.1016/j.neunet.2018.01.007
36. Bakillah, M., Liang, S.: Open geospatial data, software and standards. Open Geospatial Data Softw. Stand. 1(1) (2016). https://doi.org/10.1186/s40965-016-0004-1
37. Umar, A.: Spatial pattern of draught occurence in upper benue river basin. Glob. J. Environ. Sci. **5**(1) (2006). https://doi.org/10.4314/gjes.v5i1.2468
38. Jin, J., Gubbi, J., Marusic, S., Palaniswami, M.: An information framework for creating a smart city through internet of things. IEEE Internet Things J. **1**(2), 112–121 (2014). https://doi.org/10.1109/jiot.2013.2296516
39. Li, F., Zheng, B.: Design of the smart city planning system based on the internet of things. Int. J. Smart Home **10**(11), 207–218 (2016). https://doi.org/10.14257/ijsh.2016.10.11.18
40. Singh, P., Rashid, E.: Smart home automation deployment on third party cloud using internet of things. J. Bioinform. Intell. Control **4**(1), 31–34 (2015). https://doi.org/10.1166/jbic.2015.1113
41. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud Computing for Optimization: Foundations, Applications, and Challenges, vol. 39. Springer (2018)
42. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (eds.): In: Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2017, vol. 710. Springer (2018)
43. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)
44. Jammes, F.: Internet of things in energy efficiency. Ubiquity **2016**(February), 1–8 (2016). https://doi.org/10.1145/2822887

45. Nowak, W., de Barros, F., & Rubin, Y. (2010). Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain. Water Resour. Res. **46**(3). https://doi.org/10.1029/2009wr008312
46. Jo, H., Son, T., Jeong, S., Kang, S.: Proximity-based asynchronous messaging platform for location-based internet of things service. ISPRS Int. J. Geo-Inf. **5**(7), 116 (2016). https://doi.org/10.3390/ijgi5070116

Geocloud4GI: Cloud SDI Model for Geographical Indications Information Infrastructure Network

Rabindra Kumar Barik, Meenakshi Kandpal, Harishchandra Dubey, Vinay Kumar and Himansu Das

Abstract In the digital planet, the concept of spatial data, its cloud and Geographical Indications (GI) plays a crucial role for mapping any organization or point and acquired a reputation for producing quality results based on their spatial characteristics, including their visualization. From the twentieth century onwards, the GIS were also developed to capture, store and analyze spatial data, replacing the tedious analogue map making process. The current examine paper put forwards along with develops a Cloud SDI representation named as Geocloud4GI for giving out, investigation and dispensation of geospatial facts particularly for registered GIs in India. The primary purpose of *Geocloud4GI* framework is to assimilate the entire registered GIs' information and related locations such as state wise and year wise registered in India. *Geocloud4GI* framework can assist/help common people to get enough information for their further studies and research on GI as one of the integral part of IPR studies. QGIS is used for GI geospatial database creation and visualization. With the integration of QGIS Cloud Plug-in, the GI geospatial database uploaded in cloud server for analysis cloud infrastructure. Finally, overlay analysis has performed with the help of Google base maps in *Geocloud4GI* environment.

R. K. Barik (✉)
School of Computer Applications, KIIT Deemed to be University, Bhubaneswar, India
e-mail: rabindra.mnnit@gmail.com

M. Kandpal · H. Das
School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India
e-mail: meenakshikandpal14@gmail.com

H. Das
e-mail: das.himansu2007@gmail.com

H. Dubey
University of Texas at Dallas, Richardson, USA
e-mail: harish.dubey123@gmail.com

V. Kumar
Visvesvaraya National Institute of Technology, Nagpur, India
e-mail: vinayrel01@gmail.com

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_10

1 Introduction

Spatial Data Infrastructure (SDI) Model has promoted the exchanging as well as sharing of Geospatial Data Proprietor through distinct stakeholders. SDI has also proposed to make surroundings that empower a vast diversity to repossess and distribute geospatial and related attribute data [1–3]. Understanding the importance of the high-tech flourishing SDI concept put in order the facts from corner to corner and organization that has created multitasking, decision-supported atmosphere is crucial. To acquire new geospatial datasets, the SDI concept, saves time, effort and resources [4–7]. The common components for the SDI Model, it has been divided into five vital components. All these five components are shown in Fig. 1. These five components are basically static and dynamic in nature. Accessing networks, policy and standards are coming under the dynamic components where as data and people are said to be static [8–10].

SDI Model is used in several applications including, but not limited to, resource management, healthcare, environmental monitoring and even urban planning. In addition, SDI has been integrated with database operations with overlay operations [2, 9, 11]. SDI has also served a significant role in water resource management, river basin management, mineral resources as well as coastal supervision inside which it has the unprecedented possibility to collect, distribute and evaluate all the hydrological figures, river basin related figures, coastal data and mineral figures (related to geospatial data) in a universal stand. For development of SDI model, it was integrated with cloud computing technology which added numerous services that gave rise to cloud SDI model [11]. Basic concept for the cloud SDI setup, it manages to send the geospatial data to cloud server for analysis and processing [12].

There are several open source software, plug-in and libraries are available meant for the prototype growth of cloud SDI model [2, 3, 8, 10]. For enlargement of

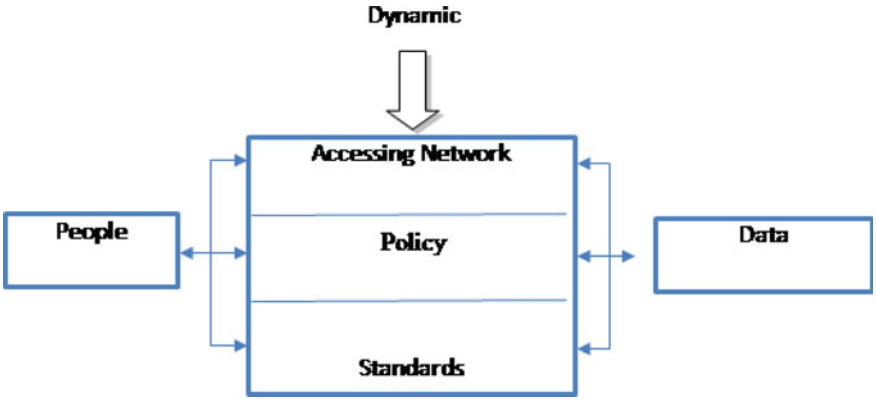


Fig. 1 Five components of SDI Model, as data and people are static components, whereas, access network, policy and standards are dynamic components [11]

cloud SDI model, geospatial database formation is required. After designing of the database, geospatial web services are designed to validate the cloud model [3]. Thus, for designing and implementing of web services and database creations, it uses several open source GIS software [4, 11, 13].

Geographical Indications (GI) which is the part of Intellectual Property Rights (IPR); plays an important and crucial role for mapping and identifying the natural goods such as manufactured with in the territory, such as agricultural or natural goods as originating or manufactured by any organization or point. It has also acquired a reputation for producing quality results based on their spatial characteristics. Thus, the potential of these GIs are required to share each and every aspect for the common people [14, 15].

SDI model with addition of GIs, it can be implement with the help of cloud environments which formulated the Cloud SDI model. This model has provided high processing capabilities and infrastructures that can reduce latency and increase throughput in close proximity to the boundary of the geospatial customers. The model is reduced the storage space required for geospatial information in the cloud. In the present paper, it allows the geospatial data; examined and analysed at the rim by means of cloud computing infrastructure.

The current paper has prepared the subsequent offerings to the Cloud SDI representation for GIs supervision:

- Cloud SDI Model i.e. *Geocloud4GI* is planned and projected to reduce latency and get better result for storing plus analyse which associated with GIs geospatial database
- It performs the case study of all the registered GIs which are registered up to April 2017 at Indian Patent Office
- Geospatial database for GIs and overlay analysis in mobile and thin clients environments are also designed and visualised in *Geocloud4GI* framework

So by adding of cloud computing and SDI concepts, it provides easy to use and perceive GIs information infrastructure network. The main propose of this information network is to awareness the people for proper management of registered GIs in India. This has made the development and implementation of *Geocloud4GI* framework a necessity.

2 Related Works

Cloud computing [16–26] provides an enormous amount of calculable possessions plus storage designed for the execution of different geospatial analysis. The cloud representation provides a changeover starting from desktop to quantifiable with numerous web servers. Cloud computing, along with other web processing [26–31] architectures, have delivered a vast atmosphere on internet to distribute assets [6, 11]. By integration of cloud with SDI, it delivers the Cloud SDI Model [5, 7, 11, 32, 33].

Likewise, Cloud SDI Model deploys multi-tenant design by allowing further clients to share resources devoid of troubling each other. This incorporated hosted provision methods are helping with application advancement and maintenance by installing patches for a better user experience. Several Open Source projects, with the aim of contributing to the society, are currently running on Cloud Computing platforms with different specifications and standards. Skygone Cloud platform and OpenGeo suite are operating in Amazon EC2 whereas QGIS is deployed and employed in the cloud for various Geospatial Web Services [34, 35].

From these several reviewing of the research works, it summarizes that Cloud SDI Model can be implemented on mobile and thin client environment.

3 Objectives of the Present Work

While analyzing various papers of literature review, the main purpose of the current learning is to develop and implement a trial product based on Cloud SDI Model, i.e., *Geocloud4GI* meant for registered GI administration in India. It proposes the organization structural design of *Geocloud4GI,* predominantly for web browsers, mobile and desktop environments. It has also proposed a powerful, sequential loom for the growth of GIs' geospatial database of India with the help of Quantum GIS Ver. 2.14.3. In this swot, it implemented a variety of overlay analysis in different environments, like, broad environment, narrow environment and movable environment.

3.1 Projected Structure Framework of Geocloud4GI

The skeleton of *Geocloud4GI* is separated into two main vital parts. The first and the most significant part is the cloud part in which GIs geospatial database is uploaded among the assistance of QGIS cloud plug-in [35]. Open Street and Google map that are basically data providers are used with the urbanized database for additional processing and investigation. For invocation of Web Feature Services, cloud server managed by QGIS Cloud supplier. QGIS Cloud supplier has been answerable for the organization of cloud layer. The second part i.e. client part is the part which has been categorized into three users for using the developed model. Figure 2 has revealed the projected system architecture of the *Geocloud4GI*.

3.2 Methodology Applied

In the commencing of geospatial catalogue, the chief accent is real-world advancement to broaden plus discover the essential of the cloud SDI model [36, 37]. For the development of *Geocloud4GI*, the multitasking and multiuser is united based on iter-

Fig. 2 System architecture of Geocloud4GI which has client and cloud layer. In client layers, it is categorized into thick, thick and mobile clients. In cloud layer, the geospatial data are stored and used for further analysis

ative model. Figure 3 represents the entirely iterative model process for enlargement of *Geocloud4GI*. In this model, it has gone through 3 iterative phases. In iterative phase 1, it dedicated for the geospatial database creations for registered GIs. In iterative phase 2, it designed and developed the geospatial web services in cloud environment for over lay analysis in thin and thick clients respectively. For complete model development, it has completed in phase 3.

4 Result and Discussions

4.1 Geospatial Database Creation and Visualisation of Registered GIs in India

The established cloud SDI Model for geospatial database formulation is intermittent and continual in nature, and each application upgrades strategically ladder wise during various appraisal and numerous testing of a built components. In built components, QGIS software is use to lay down the geospatial database creation used for registered GIs in India [38]. In the present study, it uses Quantum GIS Ver. 2.14.10 for integrated geospatial data visualisation registered GIs in India. At this point worldwide coordinate scheme WGS-84 with EPSG:4326 coordinate reference system are chosen in the visualisation of geospatial database. In Fig. 4, it has revealed the visualisation of registered GIs among Google maps.

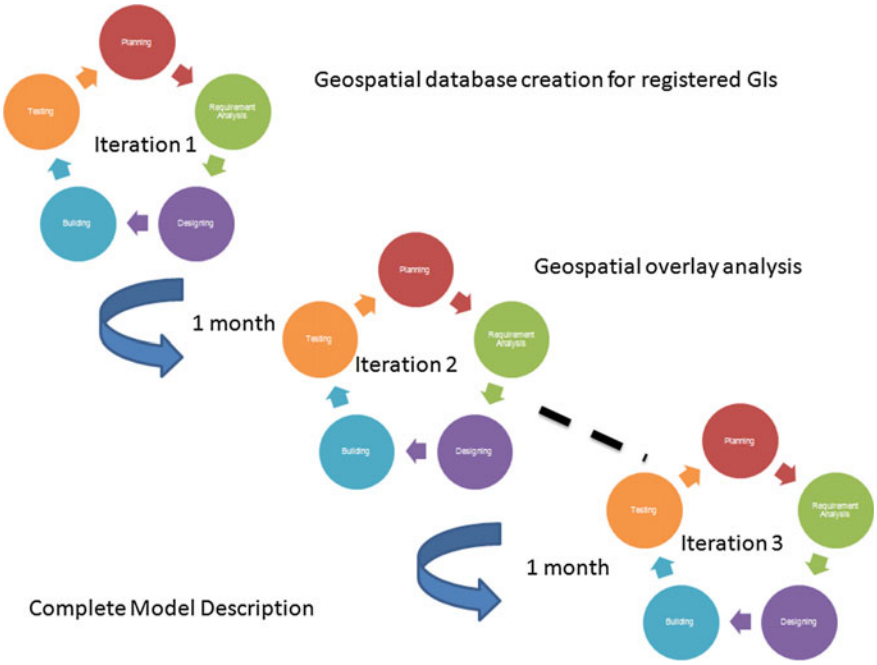


Fig. 3 Iterative model approaches for development of Geocloud4GI

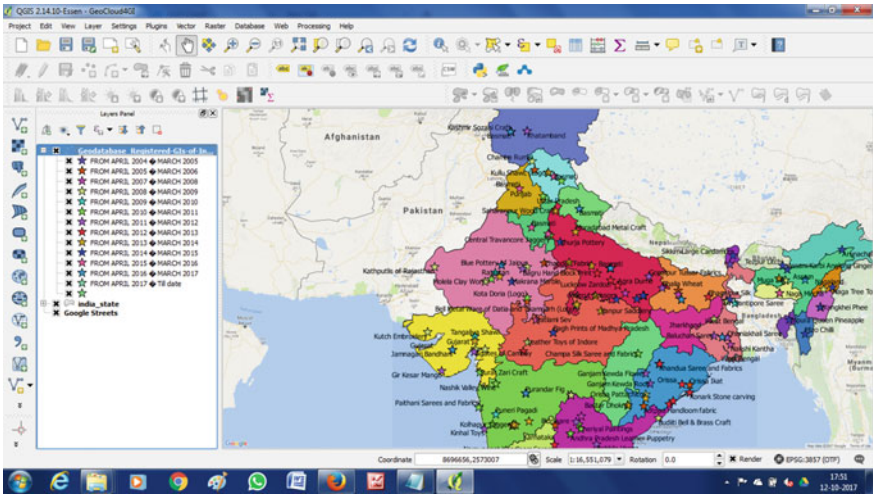


Fig. 4 Integrated geospatial database of GIs according to year wise registered

4.2 Geospatial Overlay Analysis in Geocloud4GI

Within this fragment, geospatial overlay data examination is performed for robust locality type geospatial statistics of registered GIs in India. It superimposes with vector and raster geospatial data which has been collected from IPINDIA [38]. Primarily, the data are downloaded from IPINDIA site [38] in excel format. The excel formatted data made into the trained data as in .csv format. The .csv format data again converted into the ESRI shape file formats with QGIS tools. In the current scenario, it designed the registered GIs geospatial data for processing in *Geocloud4GI*.

After storing the data, overlay analysis has performed with raster and designed vector data. In QGIS, a plug-in named as QGIS Cloud [35] has installed and made updated for storing of created geospatial database. This QGIS plug-in is responsible for the storing created geospatial vector data. These geospatial vector data are stored in the database of cloud layer. After successful uploading of the database, it automatically generated the link for mobile and thin client applications for visualization. The overlay analyses on movable and narrow client are shown in Figs. 5 and 6 respectively. It can examine that the overlay examination is a valuable method for revelation of geospatial data with the help of the other free and open data services like Google, Bing and open street maps.

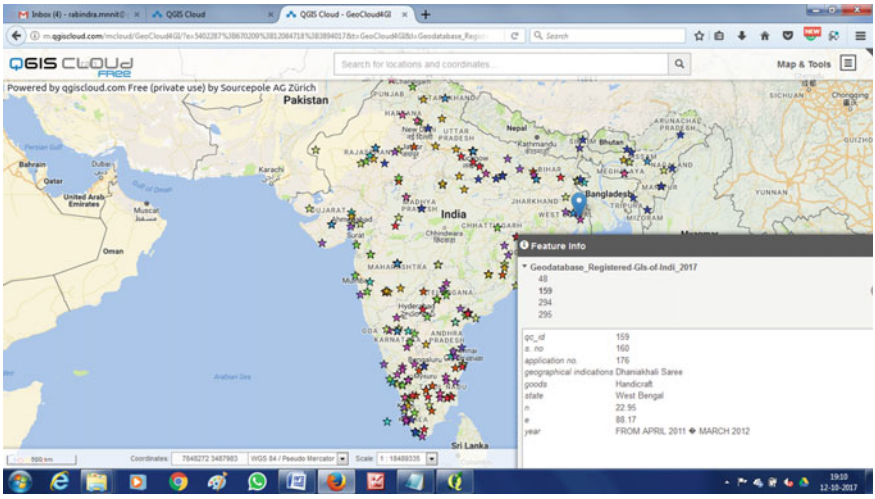


Fig. 5 Overlay analysis with Google street map in thin client environment in Geocloud4GI framework [39]

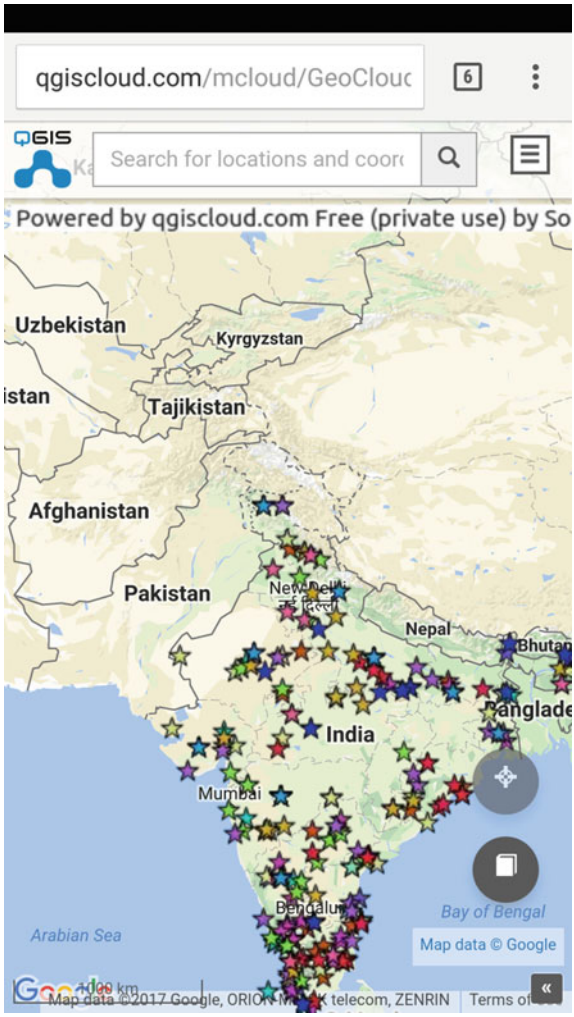


Fig. 6 Geospatial Overlay analysis with Google street map on mobile client environment in Geocloud4GI framework [40]

5 Concluding Remarks

The primary objective of this paper is to developed and validated *Geocloud4GI*, which uses cloud computing gateway in Cloud SDI framework. The development of *Geocloud4GI* augments the overall efficiency of processing of geospatial data, which in turn is beneficial to the decision makers/end users in IPR sector. In order to analyze, the developed *Geocloud4GI* framework, we have considered geospatial database of registered GIs at Indian Patent office as a case study. In future, it has planned to add more intelligence services in the developed framework particularly at cloud layer.

Acknowledgements Authors are thanking to all the experts those are involved for completion of these research work.

References

1. Arundel, J., Winter, S., Gui, G., Keatley, M.: A web-based application for beekeepers to visualise patterns of growth in floral resources using MODIS data. Environ. Model Softw. **83**, 116–125 (2016)
2. Barik, R.K., Samaddar, A.B.: Service oriented architecture based SDI model for education sector in India. In: Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013, pp. 555–562 (2014)
3. Barik, R.K., Samaddar, A.B.: Service Oriented architecture based SDI model for mineral resources management in India. Univers. J. Geosci. **2**, 1–6 (2014)
4. Brovelli, M.A., Minghini, M., Zamboni, G.: Public participation GIS: a FOSS architecture enabling field-data collection. Int. J. Digit. Earth **8**(5), 345–363 (2014)
5. Coleman, D.J., Rajabifard, A., Kolodziej, K.W.: Expanding the SDI environment: comparing current spatial data infrastructure with emerging indoor location-based services. Int. J. Digit. Earth **9**(6), 629–647 (2016)
6. Georis-Creuseveau, J., Claramunt, C., Gourmelon, F.: A modelling framework for the study of Spatial Data Infrastructures applied to coastal management and planning. Int. J. Geogr. Inf. Sci. **31**(1), 122–138 (2016)
7. Giuliani, G., Lacroix, P., Guigoz, Y., Roncella, R., Bigagli, L., Santoro, M., Mazzetti, P., Nativi, S., Ray, N., Lehmann, A.: Bringing GEOSS services into practice: a capacity building resource on spatial data infrastructures (SDI). Trans. GIS (2016)
8. Laura, J.R., Hare, T.M., Gaddis, L.R., Fergason, R.L., Skinner, J.A., Hagerty, J.J., Archinal, B.A.: Towards a planetary spatial data infrastructure. ISPRS Int. J. Geo-Inf. **6**(6), 181 (2017)
9. Leidig, Mathias, Teeuw, Richard: Free software: a review, in the context of disaster management. Int. J. Appl. Earth Obs. Geoinf. **42**, 49–56 (2015)
10. Mwange, C., Mulaku, G.C., Siriba, D.N.: Reviewing the status of national spatial data infrastructures in Africa. Surv. Rev. 1–10 (2016)
11. Barik, R.K.: CloudGanga: cloud computing based SDI model for Ganga River Basin Management in India. Int. J. Agric. Environ. Inf. Syst. (IJAEIS) **8**(4), 54–71 (2017)
12. Patra, S.S., Barik, R.K.: Dynamic dedicated server allocation for service oriented multi-agent data intensive architecture in biomedical and geospatial cloud. In: Cloud Technology: Concepts, Methodologies, Tools, and Applications, pp. 2262–2273. IGI Global (2015)
13. Barik, R.K., Samaddar, A.B., Gupta, R.D.: Investigations into the efficacy of open source GIS software. In: International Conference Map World Forum (2009)
14. Samaddar, S.G., Barik, R.K.: A mobile framework for geographical indication web services. In: Third International Conference on Computational Intelligence and Information Technology, pp. 420–426 (2013)
15. Barik, R.K., Samaddar, A.B., Samaddar, S.G.: Service oriented architecture based SDI model for geographical indication web services. Int. J. Comput. Appl. **25**(4), 42–49 (2011)
16. Das, H., Jena, A.K., Badajena, J.C., Pradhan, C., Barik, R.K.: Resource allocation in cooperative cloud environments. In: Progress in Computing, Analytics and Networking, pp. 825–841. Springer, Singapore (2018)
17. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud Computing for Optimization: Foundations, Applications, and Challenges, vol. 39. Springer (2018)
18. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (eds.): Progress in computing, analytics and networking. In: Proceedings of ICCAN 2017, vol. 710. Springer (2018)
19. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 1–26. Springer, Cham (2018)

20. Reddy, K.H.K., Das, H., Roy, D.S.: A data aware scheme for scheduling big-data applications with SAVANNA Hadoop. In: Futures of Network. CRC Press (2017)
21. Sarkhel, P., Das, H., Vashishtha, L.K.: Task-scheduling algorithms in cloud environment. In: Computational Intelligence in Data Mining, pp. 553–562. Springer, Singapore (2017)
22. Panigrahi, C.R., Tiwary, M., Pati, B., Das, H.: Big data and cyber foraging: future scope and challenges. In: Techniques and Environments for Big Data Analysis, pp. 75–100. Springer, Cham (2016)
23. Kar, I., Das, H.: Energy aware task scheduling using genetic algorithm in cloud datacentres. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 106–111 (2016)
24. Sahoo, A.K., Das, H.: Energy efficient scheduling using DVFS technique in cloud datacenters. Int. J. Comput. Sci. Inf. Technol. Res. **4**(1), 59–66 (2016)
25. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R. K., Pratik, T., Sharma, S., …, Das, H.: Mistgis: optimizing geospatial data analysis using mist computing. In: Progress in Computing, Analytics and Networking, pp. 733–742. Springer, Singapore (2018)
26. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S.A., …, Mankodiya, K.: Fog assisted cloud computing in era of big data and internet-of-things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
27. Das, H., Panda, G.S., Muduli, B., Rath, P.K.: The complex network analysis of power grid: a case study of the West Bengal power network. In: Intelligent Computing, Networking, and Informatics, pp. 17–29. Springer, New Delhi (2014)
28. Das, H., Mishra, S.K., Roy, D.S.: The topological structure of the Odisha power grid: a complex network analysis. IJMCA **1**(1), 012–016 (2013)
29. Kar, I., Parida, R.R., Das, H.: Energy aware scheduling using genetic algorithm in cloud data centers. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3545–3550. IEEE, Mar 2016
30. Das, H., Roy, D.S.: A grid computing service for power system monitoring. Int. J. Comput. Appl. **62**(20) (2013)
31. Das, H., Jena, A.K., Rath, P.K., Muduli, B., Das, S.R.: Grid computing-based performance analysis of power system: a graph theoretic approach. In: Intelligent Computing, Communication and Devices, pp. 259–266. Springer, New Delhi (2015)
32. He, L., Yue, P., Di, L., Zhang, M., Hu, L.: Adding geospatial data provenance into SDI—a service-oriented approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **8**(2), 926–936 (2015)
33. Idrees, M.O.I.: Challenges in coastal spatial data infrastructure implementation: a review. S. Afr. J. Geomatics **4**, 369–383 (2015)
34. Internet-1. http://boundlessgeo.com/products/opengeo-suite/. Accessed 23 Jan 2018
35. Internet-2. http://qgiscloud.com/. Accessed 23 Jan 2018
36. Abrahamsson, P., Salo, O., Ronkainen, J., Warsta, J.: Agile software development methods: review and analysis (2017). arXiv:1709.08439
37. Giardino, C., Paternoster, N., Unterkalmsteiner, M., Gorschek, T., Abrahamsson, P.: Software development in startup companies: the greenfield startup model. IEEE Trans. Softw. Eng. **42**(6), 585–604 (2016)
38. Internet-3. http://www.ipindia.nic.in/registered-gls.htm. Accessed 23 Jan 2017
39. Internet-4. http://qgiscloud.com/mcloud/GeoCloud4GI. Accessed 12 Oct 2017
40. Internet-5. http://m.qgiscloud.com/mcloud/GeoCloud4GI/?e=5402287%3B670209%3B12084718%3B3894017&t=GeoCloud4GI&l=Geodatabase_Registered-GIs-of-Indi_2017&bl=ROADMAP&st. Accessed 12 Oct 2017

The Role of Geospatial Technology with IoT for Precision Agriculture

V. Bhanumathi and K. Kalaivanan

Abstract Precision agriculture is mainly used to make the farming as user-friendly to achieve the desired production of a crop. With the latest Geospatial technologies, the analysis related to any type of application using the Internet of Things (IoT) made each and everyone, to materialize the things whatever is imagined. The geographic information collected from various sources and with this, IoT establishes a communication to the entire world through an Internet. The information will be helpful in the maintenance of the farmland by applying the required amount of fertilizer at the right time in the right place. It is expected that in the future, this type of smart agriculture with the application of information and communication technologies including IoT will definitely bring a revolution in the global agricultural scenario to make it more resource-efficient and productive. The main goal in combining the Geospatial technology with IoT for precision is to monitor and predict the critical parameters such as water quality, soil condition, ambient temperature and moisture, irrigation, and fertilizer for improving the crop production. It can be expected that with the help of Geospatial and IoT in smart farming, the prediction of the amount of fertilizer, weeds, and irrigation will be accurate and it helps the farmers in making decisions related to all the requirements in terms of control and supply.

Keywords Wireless sensor networks · Precision agriculture · Internet of Things Smart farming · Crop production

V. Bhanumathi (✉) · K. Kalaivanan
Department of Electronics and Communication Engineering,
Anna University Regional Campus, Coimbatore 641046, Tamil Nadu, India
e-mail: vbhanu_02@yahoo.com

K. Kalaivanan
e-mail: kalaivaanankk@yahoo.com

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_11

1 Introduction

Agriculture sector plays a premeditated role in building a backbone of economic development in India. Due to the rapid growth of the population, the demand for the food also raises. The rapid development in the space-research technology and satellite communication also will help to provide regular updates and inputs about the weather forecasting, crop production statistics to the farmers for attaining the sustainable agriculture. And also it reduces the risk in the agriculture, increases the crop productivity and economic level of the farmers and fulfills the demands of the foods. Precision agriculture is defined as the technology-based crop system, designed for long-term, site-specific applications, and aimed to increase the quality and profitability of crop production and farmer's economic growth by adjusting the agriculture inputs based on the local requirements of the croplands and also reduces the environmental impacts and risks. The crop diseases and the reduction in crop yield invariably depend on many factors such as bacteria, virus, fungus, rats, and also environmental factors like wind speed, direction, radiation, humidity, temperature, soil and water acidity. The precision farming plays a vital role in the reduction of negative impacts on the crop productivity by analyzing and monitoring spatial variability on the water, soil, and environmental conditions. In precision agriculture, the detailed information about the spatial variability of cropland is required to find the current status of the crop field and take the suitable decision on crop field management. The Geospatial technologies such as remote sensing, Global Positioning System (GPS), satellite imagery, and Geo-fencing, etc., are used to obtain valuable geographic information from various sources and with this, IoT establishes a communication to the entire world through the Internet. Thus, the users can extract the valuable information from anywhere and anytime. The Internet of Things (IoT) is one of the key components that integrates hardware, computing devices, physical objects, software in order to establish the communication, collect and exchange data among each other [1]. A general schematic of IoT is shown in Fig. 1. These technologies can be suitable for almost all the applications for analyzing the issues and meeting out the challenges. This chapter mainly focuses on the technologies in precision agriculture. WSNs are one of the simplest technologies to collect and monitor the actual spatial variability of the crop field and these have the advantages such as cost-effective, real-time suitability, and ease of deployment. The decision-support system in a farmhouse or mainframe computer integrate the prior knowledge and on-field sensed data with GIS to make optimized control decisions in water irrigation, pesticides, and the duration of fertilization [2]. This leads to many open challenging unexplored research areas in precision agriculture and act as a guiding tool for the farmers in diagnosing the crop or plant diseases accurately in a timely manner in order to protect and enhance the yield and in estimating and utilizing the available resources for e.g., water and manure at the right time and in the right quantity.

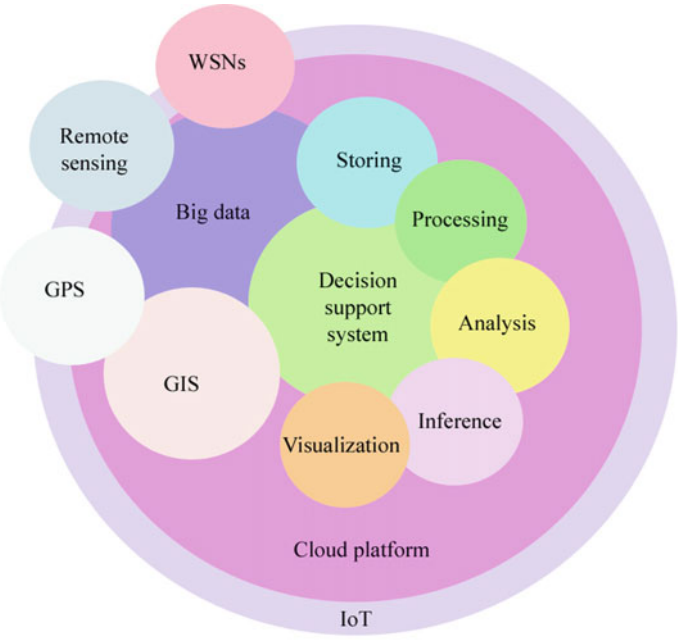


Fig. 1 Concept of IoT

1.1 Applications of IoT in Precision Agriculture

The smart agriculture paradigm evolved to formulate the utilization of new technologies in precision farming. In smart agriculture scenario, numerous services are developed to enhance the standard of the farmer's lives, and economics by implementing the farming applications [3].

1.1.1 Irrigation Monitoring System

Recent advancements in the agriculture systems need an effective irrigation management to optimize the usage of the water in farming. To utilize the available water resources, a cost-effective micro irrigation system is implemented [4]. By using the satellite imagery and remote sensing, the efficiency of the micro irrigation farming can be further extended.

1.1.2 Pest and Disease Control

Nowadays, the WSNs and Unmanned Aerial Vehicle UAV technologies are used to monitor the crop leaves quality and to predict the occurrence and possibilities of

the pest in crops [5]. It can be predicted by using environmental conditions such as temperature, wind speed, and humidity, etc., thereby increasing the crop quality and minimizes the expenditure of the farming cost.

1.1.3 Controlled Usage of Fertilizer

The yield and quality of the crop productivity are directly related to the use of fertilizer at the right time and right quantity. The prediction of the optimal amount of fertilizer is an important task in smart agriculture [6]. The estimation on the amount of fertilizer can be found by using the soil nutrients sensors which can monitor the variation in soil nutrition such as pH, Nitrogen (N), Potassium (K), Calcium (Ca), magnesium (Mg), and Phosphorous (P), etc.

1.1.4 Water Quality Monitoring

The sensors can be used to measure the water temperature, pH, electrical conductivity, turbidity, nitrates and dissolved oxygen.

1.1.5 Green House Gas Monitoring

The changes in the CO_2 and CH_4 are directly affecting the global temperature and have a straight impact on the agriculture. The emission of CO_2 and CH_4 from the various farming lands are monitored by using the WSNs and satellite imagery [7].

1.1.6 Surveillance of Cattle Movement

The latest technologies such as the camera based WSNs, UAV, and Radio Frequency Identification (RFID), etc., are employed to monitor whether any animals are moving or grazing near the croplands or not [8–10].

1.1.7 Assert Tracking and Farming System Monitoring

Presently, various advanced devices along with IoTs are deployed to fetch the information about the crops. These are shared with the cloud, which can ease to control remotely and triggers the automation in agriculture and also makes it easy in smart agriculture with remote tracking and monitoring [11].

To summarize the remainder of the chapters, Sect. 2 gives the details of Information and Communication Technology in Agriculture. Section 3 presents Research challenges and issues in IoT. A framework of Geospatial data with IoT for Precision agriculture is detailed in Sect. 4 and the conclusion is presented in Sect. 5.

2 Information and Communication Technology in Agriculture

The IoT-based smart agriculture monitoring is employed to allow the professional and farmers for monitoring the condition of the environment, farmland, soil, and crop growth. The entire smart agriculture system consists of the various types of sensor, UAV, satellite, GPS, actuators, gateway, cloud server, Internet, and Android mobile phone. The actuator has a provision for driving systems in the smart agriculture which respond to the given command by the central co-coordinator. For example, the central coordinator estimates the soil moisture and schedules the irrigation (turn-on and turn-off) of the actuators based on the readings from the farm field sensors. Each sensor node sends the sensed data to the cloud server through the gateway and Internet. Particularly, the gateway is used to collect the data from the sensor nodes and also to distribute the control message to the sensors, mobile phone, and actuators. Figure 2 shows an example of Information and Communication Technology in Agriculture. The software and the hardware presented in it aids in giving a success in the crop production with the help of these technologies.

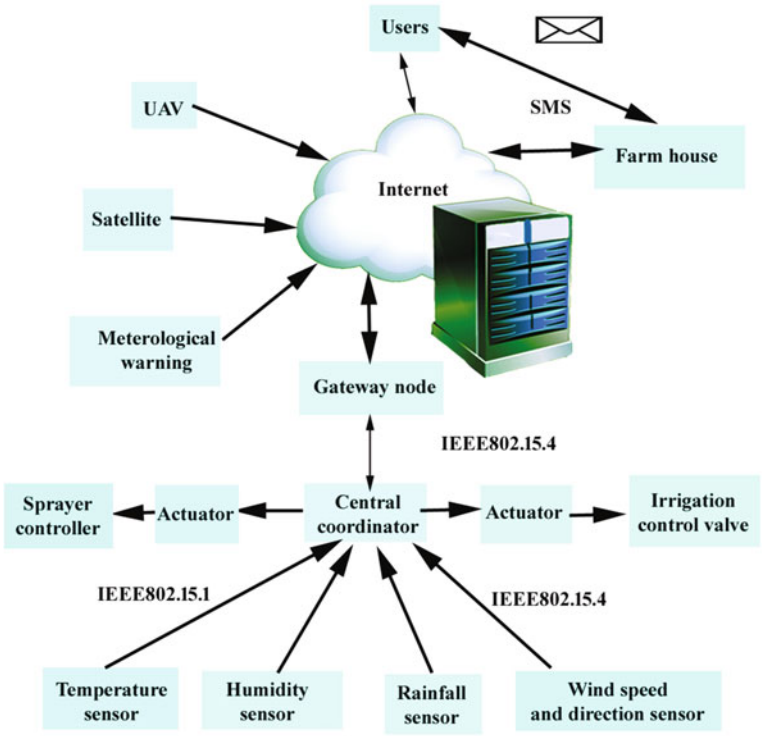


Fig. 2 Example showing information and communication technology in agriculture

2.1 Role of IoT

The Internet of Things is the most promising technologies in precision for providing numerous real-time solutions towards the modernization. It is a network of numerous things connected with various embedded electronics, service provider, automated machinery, persons, software, and hardware through Internet enabling these to gather and share the information for taking suitable solution in the crisis situation or long-term monitoring at anywhere and anytime, resulting in better economic gain, and quality of life [12]. Example of IoT hardware platforms is depicted in Table 1.

Some of the benefits of the IoT in smart agriculture are as follows:

1. Increases the profitability of agriculture.
2. Improves the quality of crop yields.
3. Minimizes the negative ecological effect.
4. Prevents the occurrence of plant diseases and pests.
5. Alerts the farmers during natural calamities, hazards, storm, etc., at the right time.
6. Allows remote monitoring, access, and control.

2.1.1 Role of Big Data in IoT

The evolution of the big data and Internet of Things (IoT) becomes an ever-growing technology in smart agriculture development which provides the enhancement of socio-economics, logistics, and living standards. The IoT continues to integrate the huge amount of data collected from various sources (data on social media and commercial networking, banking transactions, government sector and industrial records, forecasting of the weather, sensor networks, satellites, study materials, GPS, and space other research data) throughout the world [13]. For different kinds of smart applications and services, the big data has undergone the different level of analysis to produce the needed semantic information. But the big data results in the following practical challenges [14].

Data representation: The big data have uniqueness in the dataset including type, pattern, syntax, semantics, size, and accessibility. For computer interpretation and analysis, an improper representation of data will give unreliability in the original data and may also affect the efficient interpretation of various dataset. Thus, proficient data representation is needed to manage the heterogeneous dataset at various levels and also enables the different field to access the data [13].

Storage of big data: Today, various kinds of data are increasing at an exponential rate. In 2020, the amount of data generated and replicated from 4.4 to 44 ZB, said according to the International Data Corporation (IDC). Thus, an intelligent data redundancy reduction and compression techniques are needed to reduce the storage cost-effectively [14].

Responsiveness: Traditional data analysis systems lack maintaining the heterogeneous dataset structure which causes data incompatible issues in various engineering

Table 1 Comparison of various IoT hardware platforms available in the markets

Product	Processor	Memory	Communication	OS support	Operating voltage (V)
Telos	MSP430	10 k RAM, 48 k Flash	USB, IEEE 802.15.4	TinyOS	2.1–3.6
Raspberry Pi 3 Model B	Broadcom BCM2387 chipset	1 GB LPDDR2	802.11 b/g/n Wireless LAN and Bluetooth 4.1	Linux OS or Windows 10 IoT	5
Arduino Yun	Atheros AR9331	Flash Memory 16 MB RAM 64 MB DDR2,	IEEE 802.3 10/100 Mbit/s, IEEE 802.11b/g/n, USB Type-A	Linux based Linino OS	5
Beagle Bone Black System	Sitara XAM3359AZCZ100	512MB DDR3L, 16-bit (LPDDR-400, DDR2-532, DDR3-606)	802.11b/g/n	Linux, Android, Windows Embedded CE, QNX, ThreadX	5
Intel Edison Compute Module	22 nm Intel SoC that includes a dual-core, dual-threaded Intel Atom CPU at 500 MHz and a 32-bitIntel Quark microcontroller at 100 MHz	1 GB LPDDR3 POP memory, 4 GB eMMC	BT 4.0 + 2.1 EDR, IEEE 802.11a/b/g/n, USB 2.0	Linux, windows 10	3.15–4.5
ESP8266EX	Tensilica L106 32-bit RISC processor at 160 MHz	SRAM, ROM	802.11 b/g/n	RTOS, Android/iOS App	2.5–3.6
MICAz	MPR2400	Flash Memory 128 KB, EEPROM 4 KB	IEEE 802.15.4	TinyOS	2.7–3.3
Waspmote	ATmega1281	SRAM 8 KB, EEPROM 4 KB, FLASH 128 KB	802.15.4, ZigBee, Bluetooth, RFID, RFID/NFC, WiFi, GSM/GPRS, 3G/ GPRS	Windows, Linux and Mac-OS	3.3–4.2

and research application. And also it fails to find structures and relationships within big data in a timely manner, thereby causing latency in data retrieval and processing. Hence, an effective data mining and analytical algorithms are highly concentrated, so as to overcome the issue of scalability and manageability in analyzing the huge volume of data [14].

Energy and heat dissipation management: The transmission, storage, and processing of big data by the server will consume more electric power. And also, it dissipates huge volume of heat which reduces the performance and lifespan of the server. Thus, effective energy consumption and heat control and management mechanisms have great attention in preserving the economy and environment [14].

2.1.2 Role of Cloud and Fog Computing in IoT

The cloud based service refers to a variety of resources related to the applications provide to the users on-demand over the Internet. The cloud computing is an effective computation paradigm for analyzing and managing the big data in IoT. The big data related processes such as data acquisitions, processing, filtering, and transmission are rationalized for providing identical information related to the needs of the user through IoT. Cloud computing offers adequate resources and solutions to maintain, store, diagnostic, and analyze to provide a solution for the large amount of data produced by the IoT [15]. These resources are remotely located and managed by the service provider and are used for the integration of host data, diagnostic, visualization, and development of the smart application. The remote user can easily access the cloud platform with their services and resources by using a mobile phone. Numerous sensor nodes are deployed in the targeted area to sense the interest in croplands and environments. These field sensors sense the event, pre-process, and transmit the collected data to the cloud-server directly or through gateway nodes [16]. The performance of the cloud computing is limited due to scalability, congestion and latency [17]. This is because of a huge volume of the gathered data transfer to the cloud and then sent back to the application devices which have a serious impact on the performance. To overcome these drawbacks, IoT architectures are developed with distributed computing model based on the fog paradigms, in which the fog computing node is deployed at the edge of the network to perform the automated decision and learning process, storage, and interface with cloud services [1, 15]. This fog computing paradigm allows the farmers to easily analyze the changes and make the decision in a short period.

2.2 Role of Sensors

The advancement in the MEMS and other communication technologies are used to invent a compact and low-cost sensor node which is utilized in a variety of real-

time application for collecting the information about the environmental and physical parameters associated to the crops [10, 18]. It can support the agriculture professional to make wise decisions and will be informed to the farmers for further action. Advantages of WSNs in smart agriculture are as follows:

1. Ease of node deployment.
2. Low establishment cost.
3. Long-term croplands monitoring.
4. Real-time support and intervention.
5. Increased crop productivity.

Various sensor nodes that are suitable for agriculture are categorized into the plant-related sensor, soil and water-related sensor and environment-related sensor and are discussed below.

2.2.1 Soil and Water Related Sensors

Table 2 depicts that the various types of the sensor are currently utilized for measuring the soil and water related parameters such as moisture, water flow, temperature, nutrients, water content, water level, and conductivity.

2.2.2 Environment Related Sensors

The environment-related information is an essential part of the precision agriculture. The potential uses of these sensor nodes include monitoring the greenhouse gases, pollution, ambient temperature, relative humidity, atmospheric pressure, solar radiation, and lightning. The environment-related sensors are mentioned in Table 3.

2.2.3 Plant Related sensors

These types of sensor are attached to the plants for monitoring the leaf wetness, temperature, conductivity, ice formation, water level, plant stem-cell growth, and yield quality. Table 4 shows the various types of sensors with plant related measurement parameters.

2.3 Role of Remote Sensing

Remote sensing is described as a process to gather information through non-physical contact measurements of radiation reflected or emitted from the particular material [19]. The remote sensing is mainly depending on the reflectance or emission property of the object. The signal emission or reflectance of the object is based on the

Table 2 Comparison of various soil related sensors available in the markets

Sensing devices	Moisture	Temperature	Conductivity	Water flow	Soil nutrient	Water content	Water level
BL-5311, BL-5315B (www.baselinesystems.com)	✓	✓					
SMR110 (www.fecegypt.com)	✓						
FC-28 (www.uruktech.com)	✓						
ECH20EC5 (www.metergroup.com)	✓						
Davis 6440 (www.davisnet.com)	✓						
Davis 6470 (www.davisnet.com)		✓					
Davis 6345CS (www.davisnet.com)	✓	✓					
HydraGO S (www.stevenswater.com)	✓	✓	✓				
Stevens GroPoint Profile (www.stevenswater.com)	✓	✓		✓			✓
HydraProbe (www.stevenswater.com)	✓	✓	✓				
SM150T (www.dynamax.com)	✓	✓				✓	
WET Sensor Kit (www.dynamax.com)		✓	✓			✓	
HFP01 Self-Calibrating Heat Flux Sensor (www.dynamax.com)	✓	✓					
12.545—Soil Nutrient Sensors (www.pbltechnology.com)					✓		
Gropoint lite (www.gropoint.com)	✓	✓					
Gropoint pro (www.gropoint.com)	✓	✓	✓				
SMEC 300 (www.specmeters.com)	✓	✓	✓			✓	
EC 250 (www.rshydro.co.uk)		✓	✓				
SEN0114 (www.mouser.com)	✓						

Table 3 Comparison of various environmental sensors available in the markets

Sensing devices	Humidity	Temperature	Pressure	Wind speed and direction	Lightning	CO_2	Solar radiation	Rain fall
EM60G (www.metergroup.com)		✓	✓					
Met One 083E (www.stevenswater.com)	✓	✓						
Lufft WS-800 (www.stevenswater.com)	✓	✓	✓	✓	✓		✓	
VG-HUMID (www.vegetronix.com)	✓							
Sensirion SCD 30 Sensor Module (www.sensirion.com)	✓	✓				✓		
HDC2010 (www.ti.com)	✓	✓						
IDT HS300x (www.mouser.in)	✓	✓						
ams ENS210 (www.mouser.in)	✓	✓						
Gill wind speed indicators (www.dynamax.com)				✓				
BF5 sensor (www.dynamax.com)							✓	
VG-HUMID (www.vegetronix.com)	✓							
LM 35 (www.ti.com)		✓						
SHT7x (www.sensirion.com)	✓	✓						
DS 1822 (www.digikey.com)		✓						
DHT 11 (www.adafruit.com)	✓	✓						
BME280 (www.digikey.com)	✓	✓	✓					
HMP35C (www.campbellsci.com)	✓	✓						
COR LI200S (www.licor.com)							✓	
PTA427 (www.campbellsci.com)			✓					
TE525-L (www.campbellsci.com)								✓

Table 4 Comparison of various plant related sensors available in the markets

Sensing devices	Wetness	Ice formation	Plant growth	Water level
PHYTOS 31 (www.metergroup.com)	✓	✓		
LWS-L (www.campbellsci.com)	✓	✓		✓
237-L (www.campbellsci.com)	✓			
Davis 6420 (www.davisnet.com)	✓			
EXO-Skin Sap Flow Sensor (www.dynamax.com)			✓	
DEX20, DEX70, DEX100 and DEX200 (www.dynamax.com)			✓	

chemical and physical characteristics of the particular material, and their geographic environment such as temperature, leaf wetness, and chlorophylls, etc. The chemical compounds of the crop, called chlorophylls emit the radiation which is inversely related to the absorption of the electromagnetic radiation. The measurement of the radiated signal is performed through the different levels of spectral bands including Green, Blue, Red, Near Infra-Red (NIR), and Short Wave Infra-Red (SWIR). In agriculture, the Green, Red, and Infra-Red are widely used to obtain the vegetation indices. The common method for measuring the vegetation index is the Normalized Difference Vegetation Index (NDVI). These indices are used to analyze specific characteristics such as crop biomass, Leaf Area Index (LAI), crop count, water content, plant stress, humidity. In agriculture, the variation in chlorophyll concentration is highly sensitive to the variation of green and red wavelength (i.e., the absorption and reflection measurement of electromagnetic radiation through the green and red wavelengths are inversely varied with respect to the concentration of chlorophylls) [19].

For example, highly concentrated chlorophyll plants attract more radiation in red and blue wavelength and less reflection in infrared or green wavelength. On the contrary, less chlorophyll reflects with the high red wavelength. Thus, investigating the visible to the infrared spectral band can give necessary information about the plant's cellular arrangement for measuring the plant stress, and productivity. The analysis of the red edge spectral region provides a better responsiveness under stress induced by chlorophylls content changes and also accurately estimates the LAI than the green and red spectral range. The remote sensing can be performed by satellite-based platform (satellite, UAV, etc.) and ground-based platform (tractor, hand-held sensors, etc.).

Multispectral Scanner System sensors, LiDAR [20], RaDAR, thermal and infrared cameras are frequently used in the satellite and UAV [8], so as to estimate the spatial factors of agriculture objects such as biomass, N stress, yield mapping, pest and disease detection which is based on the observation of the reflectance of electromagnetic radiation from the soil and crop organic contents. The recent improvement in real-time satellite imagery (e.g., Eo-1 Hyperion, Hyperspectral Infrared Imager

Table 5 Comparison of satellites used in the precision agriculture (https://directory.eoportal.org/)

Name	Components used	Applications
RESOURCESAT-2A	(Linear Imaging Self Scanning Sensor) LISS-4 sensor with 5.8 m spatial resolution, LISS-3 sensor with 23.5 m spatial resolution, AWiFS sensor with 56 m spatial resolution	Agriculture, environmental monitoring, disaster management
INSAT-3D	VHRR/2 (Very High Resolution Radiometer), DRT (Data Relay Transponder), Sounder, SAS and R (Satellite Aided Search & Rescue)	Weather forecasting, disaster warning
SPOT-6	NAOMI	Forestry, defense, agriculture, mining, oil and gas exploration, surveillance, environmental monitoring, engineering
RISAT-1	C-band SAR (Synthetic Aperture Radar)	Weather imaging, crop monitoring, disaster management
KALPANA-1	VHRR/2, DRT	Weather forecasting
IKONOS-2	Kodak Model 1000 Camera System	Mapping and monitoring application, forestry, agriculture, mining, construction, engineering, natural disaster
QuickBird Satellite Sensor (0.65 m)	BGIS 2000 (Ball Global Imaging System 2000)	Forestry, agriculture, oil and gas exploration, environmental monitoring, construction
Pleiades-1A Satellite Sensor (0.5 m)	HiRI (High-Resolution Imager)	Crisis monitoring
WorldView-2 Satellite Sensor (0.46 m)	WV110 camera	Defense, land use planning, exploration, environmental monitoring
LANDSAT 7	MSS (LS-1-5), TM (LS-4/5), ETM (LS-6), ETM+ (LS-7).	Forestry, defense, agriculture, mining, environmental monitoring, land coverage and change detection
RapidEye Satellite Sensors (5 m)	REIS (RapidEye Earth Imaging System)	Forestry, defense, agriculture, mining, oil and gas exploration, construction, engineering, governments, cartography, surveillance, environmental monitoring

(HyspIRI)) is high spatial and spectral resolution that could be provided the better classification and decision making in precision agriculture [19]. Examples of satellites that are used in the Precision agriculture are given in Table 5.

Ground-based remote sensing otherwise called as proximal remote sensing, in which the sensor nodes are mounted on to the sprayer, tractor, etc., had important capabilities for determining spatial patterns of soil organic matter, calcium, phosphorous, potassium, carbon, soil temperature, moisture, salinity, soil pH, N stress, water stress, and crop yield. These types of sensing provide a real-time agricultural application specific management such as fertilizer, pesticides, and irrigation. The ground-based remote sensing is more accurate as compared to the sky-based platform since it is less affected by the cloud cover. Asher et al. [21] discussed a ground-based remote sensing, in which linear move irrigation system designed with six IR sensors and moving weather station was used to find the canopy temperatures and dry area of the cropland. This sensed data enabled the grower to open the irrigation control valve.

Applications of the remote sensing in agriculture are listed as follows: (i) cropping system analysis, (ii) agriculture drought assessment and monitoring, (iii) soil mapping and monitoring, (iv) water resources monitoring, (v) crop area estimation and monitoring, (vi) incidence forecasting, (vii) diseases and pests detection.

The limitations of the remote sensing in the agriculture are as follows:

1. Simultaneous sensing of the reflectance radiation from the crop by the sensor causes the confusion in spectral band.
2. The scaling issues due to the inappropriateness between the actual considered and remote sensing data and also it provide the inadequate information for evaluating the historical data.
3. Low accuracy in reflectance measurement of the soil and water due to the organic matter, surface roughness, and moisture, thus characterizing the soil and water properties are very complex.
4. Hard to classify the crop species.
5. The usage of the passive sensor in the remote sensing is severely affected by the weather condition.

2.4 Role of GPS

The GPS and Global Navigation Satellite System (GNSS) give a Geospatial position on the earth surface or it provides the position of the UAV on the sky with respect to the earth surface [22]. The crop field monitoring connected with GPS provides the necessary data to map the agriculture applications and a site-specific cropland management. The tangible benefits of using GNSS and GPS in agriculture are: (i) providing guidance for auto steering the tractors, sprayers, and spreaders on the cropland, (ii) allowing the worker to perform the agriculture activity in low visibility condition.

2.5 Role of GIS

The geographic information system provides a computer-based framework for collecting, storing, processing, analyzing, mapping, and visualization of Geospatial data. With this geographic science with tools, GIS depicts closer discernment into data for understanding the pattern, collaboration, and circumstances which help the people to make smarter decision attaining their goal in real time practices [23]. The GIS technology is mainly used to store the layers of information and integrate the agriculture field with remotely sensed data and produces the number of possibilities in analyzing the geographical data. Numerous modern GIS software packages such as GRASS GIS, ArcGIS, QGIS, OpenStreetMap, and GeoMedia, etc., are easily shared their resources (imagery, features, base-maps, spreadsheets, and tables) and embedded with apps for solving a complex problem and it is freely accessible by the public user [24].

The GPS provides the position of the user or receiver through showing the longitude, latitude, and altitude which is not useful in finding the location. But GIS provides the information about where you are on the map by using computerized mapping system (contains all geographic Information and maps) which also displays or visualizes the ground terrain surface in 2D or 3D scenes. For example, GIS model is incorporated into GPS providing necessary information on the identification of sensor node placement with the earth and the mapping of satellite image to the corresponding farmer's registered cropland.

2.6 Role of Wireless Communication Technology

As depicted in Table 6, the Bluetooth 802.15.1 offers low powered and high speed over short-range wireless communication which connects mobile phones and portable devices and it can connect the devices up to 10 m within its range. Zigbee 802.15.4 is low cost and low powered communication technology which is specially designed for controlling and monitoring of sensor network application and can provide coverage up to 100 m with the transfer rate is up to 250 Kbps. The Wireless Fidelity (WiFi) standards 802.11a and 802.11b are introduced in 1999 and provide data transfer rate up to 11 Mbps, followed by 802.11g is speed up to 54 Mbps. It is commonly used in the educational institution, company and business premises, and railway station which allow every individual to access the internet through WiFi. Worldwide Interoperability for Microwave Access (WiMAX) 802.16 is a new wireless technology in broadband network access. It can provide the services up to 50 Km with data transfer rate up to 70 Mbps. However, WiMAX requires the license for using the spectral bandwidth. The 2G evolved in the 90s comprises the IS-95 and GSM digital voice standards. The added feature of GSM is that it can support circuit switched data (wired communication) at speed up to 14.4 Kbps. The advent of 3G becomes a new era in cellular mobile communication development in the view

of high-speed data communication, security, bandwidth, and support various multimedia applications, etc. it can support the data rate up to 2 Mbps and 384 Kbps for indoor and outdoor application respectively. The 3G licensed network services are operated and owned by the service providers and these services sell to the end users based on the usage of the data or use the network resources per seconds. The 4G is Internet Protocol (IP) based integrated system which provides the data rate between 100 Mbps and 1 Gbps in both indoor and outdoor application. The salient features of 4G are high quality, speed, security, and bandwidth, and also allow the mobile users to access the Internet at anywhere and anytime.

3 Research Challenges and Issues in IoT

3.1 Security Related Issues

The productivity of the agriculture is directly related to the environmental parameters such as humidity, wind speed, temperature, acidity, CO_2, and soil moisture, etc. if any modification, insertion, and deletion of these parameters have a negative outcome on the crop growth and productivity. Hence, the security and privacy are crucial issues in cropland. For example, any modification on the fertilizer quantity leads to reduce the productivity of the crop or the excessive use of fertilizer causing environmental pollution and reduce the groundwater quality. Therefore, great effort is needed to protect the environmental data from the network hackers or eavesdroppers [3, 25–27].

3.2 Data Management

In smart agriculture, various kinds of data are collected from the several sources such as sensors in the cropland, satellite images, GPS, meteorological, etc., which are intended to be connected to the cloud. The management of these big data is a challenging task to the designer in agriculture because the determination of the data collection, diagnosis, analysis, decision making, and evaluation are the complex process in the cloud [28]. Thus, the integration of IoT and SDN require a new methodology in deployment, hardware and software services, communication, and resources [29].

Table 6 Comparison of various wireless communication technologies [17]

Communication technology	Frequency band	Speed	Coverage distance	Energy consumption	Cost	Application
Bluetooth	2.4 GHz (v1.x, v4), 5 GHz (v3)	1 Mbps, 24 Mbps	10 m	Medium	Low	Mobile phone, digital camera, etc.
Wi-Fi	2.4 GHz 5 GHz	11 Mbps, 54 Mbps	150 m	High	High	Monitoring based application, smart home, VANET, etc.
Zigbee	868 MHz 915 MHz 2.4 GHz	20 Kbps, 40 Kbps, 250 Kbps	100–300 m	Low	Low	Smart agriculture, city, and health care, etc.
Wimax	266 GHz	0.41 Gbps (stationary), 50,100 Mbps (mobile)	<50 km	Medium	High	Interface, MP3 players, CCTV cameras, PDAs
2G/3G/4G	865 MHz, 2.4 GHz	50,100 Kbps/200 Kbps/0.11 Gbps	Entire GSM coverage area	Medium	Medium	Cellular network, Internet, mobile phone

3.3 Fault Tolerance

It is an important attribute of the WSNs for achieving an efficient smart agriculture. There are different kinds of faults occurred in WSNs due to (i) uncorrected values showed in the sensor's measurement, (ii) deployment of the faulty sensor, (iii) poor network design, (iv) failure in the communication link, (v) exhausting of the battery energy very quickly [30].

3.4 Memory

The large size of memory is required for storing the big data which are collected from the various deployed sensor nodes and other resources. These collected data are used to analyze and derive the conclusion about the current status of the cropland, crop growth, fertilizer and pest level, and also send the valid command to the farmers through SMS or mobile agriculture app. [17].

4 Framework of Geospatial Data with IoT for Precision Agriculture

4.1 Existing Infrastructures

Akkas et al. [31] developed an IoT based greenhouse monitoring in which the Micaz motes were utilized to collect the environmental parameters such as humidity, temperature, light, and pressure. The collected information transmitted through 802.15.4 to the MIB 250 base station, so as to perform the analysis of the collected data and make the inference and operation management. The end user can communicate with BS by using the mobile phone or the internet to get the management information about the green-house. Nagarajan et al. [32] demonstrated a sprinkler irrigation automated system with the help of the pH, temperature, and soil moisture sensor. These sensors were used to gather the field information and forwarded to the BS through Zigbee protocol. The decision support system in the BS made the inference about the cycle of irrigation and sends the command to the irrigation actuator through the PIC microcontroller. Martinez et al. [33] discussed a measurement of greenhouse climatic parameters such as humidity, air-speed, ultraviolet radiation, and globe and air temperature and transmitted to the remote operator through 802.15.4. Foughali et al. [34] developed a blight forecast model based on the weather condition and field sensor (Waspmote 868 SMA 4.5 DBI). The sensed information was transmitted to the IoT cloud platform (Ubidots) through the gateway by using 802.15.4. The ubidots were analyzed and monitored for the incoming sensed and offered the notification service to the farmer's mobile phone.

4.2 Case Studies

Some real-time of agriculture related application by using IOT are discussed in below and summarized in Table 7. These applications related documents are gathered from www.campbellsci.com.

4.2.1 Pest Control and Evapotranspiration System

Radu Carcoana, North Dakota State University described the weather monitoring and pesticide application, in which various kinds of sensors (wind speed and direction, temperature, solar radiation, relative humidity, barometric pressure, and water level) are deployed in the region of North Dakota, USA in order to collect the field information and enhance the crop productivity and quality. These recorded sensor data are accessed through the phone which can be used as input of an automated decision system to monitor the specific crop diseases and plant-growth. This will help the formers to take a decision about when and how much uses pesticide and irrigate.

4.2.2 Flood and Irrigation Monitoring

Rajat Saha, MBK Engineers Narendra S. Raghuwanshi, Indian Institute of Technology Shrinivasa K. Upadhyaya, UC Davis Wesley W. Wallender, UC Davis David C. Slaughter, UC Davis tested the irrigation control in the area of 720 ft long and 50 ft wide in University of California Davis campus which comprises of field sensor and communication system for monitoring the flood irrigation water at the lower end of the cropland. The sensor data are recorded in the micro-logger CR3000, which in turn transmits the collected data to the irrigator through the cellular modem (RAVEN110).

4.2.3 CO_2 Monitoring

Agricultural Research Service (ARS) demonstrated the CO_2 flux cycle measurement in central and western US by monitoring the exchange of carbon dioxide between the rangeland and atmosphere.

4.2.4 Water-Resource Management and Weather Monitoring

South Jersey Resource Conservation and Development Council, Inc., Campbell Scientific, Inc., developed a Resource Information Serving Everyone (RISE) network which aimed for watershed protection and management. In this, the water quality and weather station were established in the New Jersey by deploying various

communication system, environment and the soil related sensor for collecting the valuable field data. These collected data are retrieved from the water quality and weather station through the telephone modem, thus, the user can build up own irrigation management system for their particular cropland.

4.2.5 Korea: Flux Monitoring over a Rice Paddy

Campbell Scientific and Department of Atmospheric Sciences at Yonsei University, Seoul, Korea, jointly calculated the greenhouse gas fluxes such as methane, carbon dioxide, and water vapor over paddy by using a $CH_4/CO_2/H_2O$ vapor profile system and eddy-covariance system.

4.2.6 ISS System for Studying Plants in Space, Feeding Space Travelers

The Utah State University, Space Dynamics Laboratory (SDL) developed a plant growing chamber in space with the help of Campbell Scientific, and International Space Station (ISS). The data logger is used to monitor the temperature and water content, and also control the plant growing environment. This project is used for the micro-gravity food production, as well as it can also be used to perform the research in improving the space traveler life quality.

4.2.7 Costa Rica: Banana Production

Campbell Scientific and Costa Rica for the National Banana Corporation (CORBANA) established the network by using agro-meteorological stations. This network performs the data transmission between the filed sensor and the end users, mail services, Internet, VoIP at the end office through a Mikrotic Wi-Fi. Thus, CORBANA specialist can perform the relevant precautionary measures about the irrigation, fertilization, controlled usage of pesticides, diseases, banana production based on the collected field parameters (wind speed, wind direction, rainfall, soil moisture, relative humidity, temperature, leaf wetness, and solar radiation).

4.2.8 Colombia: Fighting Fungus on Roses

The Agro-Industrial Research Center at Jorge Tadeo Lozano University studied the various environmental conditions which encourage the growth of fungus on the rose. In order to avoid the downy mildew on roses, the leaf wetness sensor (Decagon LWS) and Apogee SI-111 Infrared Radiometers were deployed in the farming land. These sensors were connected to data-loggers (CR1000) through AM16/32 multiplexers. These sensor nodes were used to monitor the environmental changes and made a

Table 7 Summary of the case studies related to the precision agriculture (www.campbellsci.com)

Application	Location	Products used	Contributors	Measured Parameters
A state-wide weather station network that provides data for agricultural applications such as pest control and evapotranspiration	North Dakota, USA	Vaisala HMP35C, Met One 014A and 024A, LI-COR LI200S and LI200X, Vaisala PTA427, TE525-L, thermocouples or thermistors, DC112 phone modems, VS1 voice modems, CR10 datalogger, 21Xs, radio telemetry	Radu Carcoana, North Dakota State University	Wind speed/direction; air, soil, and crop temperature; precipitation; solar radiation; relative humidity; barometric pressure; ground-water level
Flood irrigation monitoring	University of California Davis campus	CR3000 RAVEN110 CFM100	Rajat Saha, MBK Engineers Narendra S. Raghuwanshi, Indian Institute of Technology, Shrinivasa K. Upadhyaya, UC Davis Wesley W. Wallender, UC Davis David C. Slaughter, UC Davis	Arrival of flood irrigation water at lower end of field
Determining the impact of grasslands in CO_2 exchange	Central and Western US	CO_2 flux monitoring network	Agricultural Research Service (ARS of USDA)	CO_2 flux, H_2O flux, thermal flux, wind speed, wind direction, net radiation, solar radiation, precipitation
Weather and water quality stations monitor parameters for irrigation planning and water conservation	New Jersey, USA	CR10X LI200X-L HMP45C-L TE525-L Li-Cor LI200S or LI200X	Steve Quesenberry, South Jersey RC & D	(Weather Stations) Air temperature, relative humidity, rainfall, solar radiation, wind speed and direction (Water Quality Stations) water velocity, pH, turbidity, dissolved oxygen, water level, air temperature, relative humidity and rainfall

Table 7 (continued)

Eddy covariance fluxes and vertical profiles of methane, carbon dioxide, and water vapor	Seoul, Korea	TGA100 Trace-Gas Analyzers, CSAT3 Sonic Anemometers, CR9000 and CR23X hygrometers, and LI-COR 6262 H_2O/CO_2 analyzers	Campbell Scientific, Department of Atmospheric Sciences at Yonsei University, Seoul, Korea	greenhouse gas fluxes
Environmental measurement and control in plant experiment	International Space Station (ISS), Earth orbit	CR10X AM25T	Gail Bingham and Shane Topham, Space Dynamics Laboratory, Utah State University	Temperature of air, leaf surface, and substrate; light; substrate moisture; $COCO_2$; O_2; electrical power
Banana production	Costa Rica	CR10X, environmental, soil, plant related sensors	Otton C. Brenes, Representaciones Corelsa, S.A. Brad Maxfield, Campbell Scientific	Rain, Wind, Solar Radiation, Air Temperature and Humidity, Ground Temperature, Leaf Wetness
Study of growing conditions of roses, gathering data for battling fungus growth	Bogota Savannah, Colombia	LWS-L SI-111 AM16/32 CR1000	Francisco Gonzalez, Durespo	Leaf wetness, leaf surface temperature

better condition for preventing the fungus growth accordingly, resulted in the avoidance of the use of pesticides and environmental pollution.

4.3 Proposed Framework

Precision Agriculture is defined as a method of farm management that retrieves the existing real-time data, process and analyzes it and provides a solution to the farmers in the decision-making process to decrease inputs and increase crop production. The technological developments embarked on a way to use these in all fields of applications. The agriculture industry and automation suppliers are trying to explore and utilize the opportunities for improving their production, profitability, and farming practices with the help of IoT and digital solutions. The proposed architecture for the integration of Geospatial data with IoT to realize precision agriculture is shown in detail in Fig. 3.

It is observed from the figure, those sensors in the farm and farm equipment provide real-time data availability and alarms through the Internet for further process.

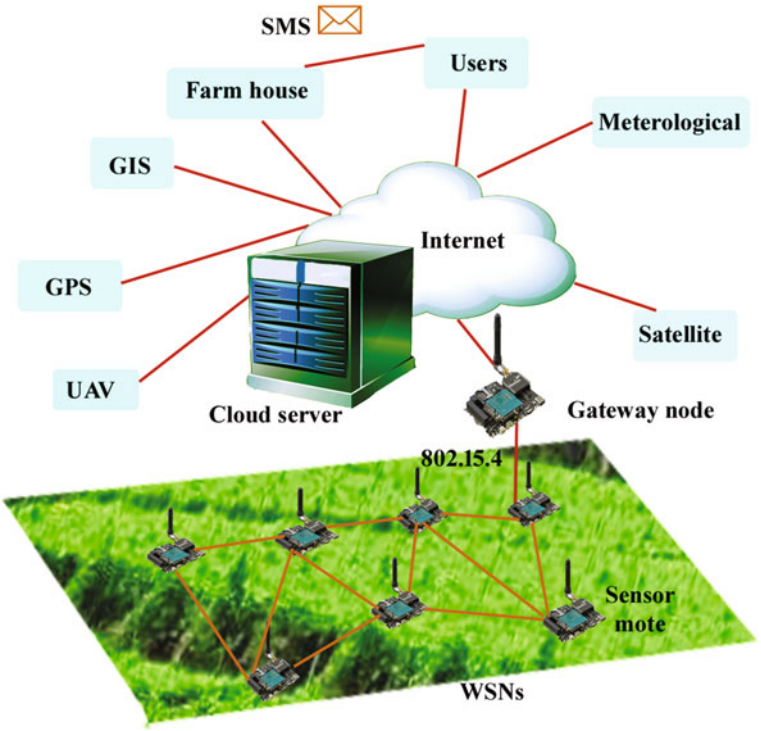


Fig. 3 Concept of IoT

The collected real-time data are analyzed using Big Data analytics and stored in the cloud server for making decisions. The decisions are taken after collecting the Geospatial data. In sensor-based precision agriculture system, data from the sensors are generally shared among the stakeholders either through a local server or the cloud. This information exchange purely depends on the reliability of the communication network and the Internet. Then, the processed data can be retrieved via smart phones and user-friendly apps that are available to represent it in a simple and clear format. It is found from the literature, the Geospatial information has an impact over the crop yield, irrigation and amount of fertilizers etc. Hence for taking a corrective and preventive action, IoT based smart farming is felt to be streamlined. It definitely helps in the prediction of the available and required water resources, fertilizers, and control of weeds. In recent years, Precision agriculture gained popularity among farmers across the globe because of the technological developments. Farmers are in need of these for realizing the optimum production with the available resources. Due to the global warming and sudden fluctuations in the climatic conditions and weather patterns, no one can expect an efficient production in agriculture without the help of these smart farming.

5 Conclusion

IoT is a vital technology in precision agriculture for increasing the crop productivity, monitoring the environmental factor, measuring the quality of water and soil. The IoT technology allows the farmers to monitor the crop remotely and make any decision related to the agriculture tasks like water irrigation, greenhouse maintenance, and pest monitoring, etc., at any time and anywhere through the Internet. In this chapter, the components and technologies involved in the IoT and their strengths, challenges, and issues all are elaborated. The need for the Geospatial data with IoT is discussed. This chapter reveals key strategies to handle the environmental crises such as flood, drought, cyclone, pollution, global warming, and blights, etc. and provides the awareness about the utilization of various technologies which are involved in the precision agriculture. The outcome of the proposal is to develop awareness about the precision farming and available control software application toolkits in the market and to make use of it. With this, the farmers can easily access the zone management office, update the crop details automatically and to retrieve the awareness and control information such as soil irrigation, soil nutrition, crop diseases etc. The proposed framework if it is developed as a model and implemented in real time, definitely it will bring a revolution in the agricultural sector.

References

1. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S. A., Mankodiya, K.: Fog assisted cloud computing in era of Big Data and Internet-of-Things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges, pp. 367–394. Springer, Cham (2018)
2. Thorp, K.R., Hunsaker, D.J., French, A.N., Bautista, E., Bronson, K.F.: Integrating geospatial data and cropping system simulation within a geographic information system to analyze spatial seed cotton yield, water use, and irrigation requirements. Precis. Agric. **16**(5), 532–557 (2015)
3. Tzounis, A., Katsoulas, N., Bartzanas, T., Kittas, C.: Internet of Things in agriculture, recent advances and future challenges. Biosyst. Eng. **164**, 31–48 (2017)
4. Coates, R.W., Delwiche, M.J., Broad, A., Holler, M.: Wireless sensor network with irrigation valve control. Comput. Electron. Agric. **96**, 13–22 (2013)
5. Faial, B.S., Costa, F.G., Pessin, G., Ueyama, J., Freitas, H., Colombo, A., Fini, P.H., Villas, L., Osorio, F.S., Vargas, P.A., Braun, T.: The use of unmanned aerial vehicles and wireless sensor networks for spraying pesticides. J. Syst. Archit. **60**, 393–404 (2014)
6. Alahi, M.E.E., Nag, A., Mukhopadhyay, S.C., Burkitt, L.: A temperature-compensated graphene sensor for nitrate monitoring in real-time application. Sens. Actuators A Phys. **269**, 79–90 (2018)
7. Martnez, J.L., Claraco, J.L.B., Alonso, J.P., Ferre, A.J.C.: Distributed network for measuring climatic parameters in heterogeneous environments: application in a greenhouse. Comput. Electron. Agric. **145**, 105–121 (2018)
8. Pajares, G.: Overview and current status of remote sensing applications based on unmanned aerial vehicles (UAVs). Photogramm. Eng. Remote Sens. **81**, 281–329 (2015)
9. Polo, J., Hornero, G., Duijneveld, C., Garcia, A., Casas, O.: Design of a low-cost wireless sensor network with UAV mobile node for agricultural applications. Comput. Electron. Agric. **119**, 19–32 (2015)
10. Rawat, P., Singh, K.D., Chaouchi, H., Bonnin, J.M.: Wireless sensor networks: a survey on recent developments and potential synergies. J. Supercomput. **68**, 1–48 (2014)
11. Sanchez, A.J.G., Sanchez, F.G., Haro, J.G.: Wireless sensor network deployment for integrating video-surveillance and data-monitoring in precision agriculture over distributed crops. Comput. Electron. Agric. **75**, 288–303 (2011)
12. Afzal, B., Umair, M., Shah, G.A., Ahmed, E.: Enabling IoT platforms for social IoT applications: vision, feature mapping, and challenges. Future Gener. Comput. Syst. Available online 13 Dec 2017
13. Chen, M., Mao, S., Liu, Y.: Big Data: a survey. Mob. Netw. Appl. **19**, 171–209 (2014)
14. DeRen, L., JianJun, C., Yuan, Y.: Big data in smart cities. Sci. China Inf. Sci. **58** (2015)
15. Aazam, M., Zeadally, S., Harras, K.A.: Offloading in fog computing for IoT: review, enabling technologies, and research opportunities. Future Gener. Comput. Syst. **87**, 278–289 (2018)
16. Panigrahi, C.R., Sarkar, J.L., Pati, B., Das, H.: S2S: a novel approach for source to sink node communication in wireless sensor networks. In: International Conference on Mining Intelligence and Knowledge Exploration, pp. 406–414. Springer, Cham (2015)
17. Bhanumathi, V., Kalaivanan, K.: Application specific sensor-cloud: architectural model. In: Mishra, B., Dehuri, S., Panigrahi, B., Nayak, A., Mishra, B., Das, H. (eds.) Computational Intelligence in Sensor Networks. Studies in Computational Intelligence, vol. 776, pp. 277–305. Springer, Berlin, Heidelberg (2019)
18. Barkunan, S.R., Bhanumathi, V.: An efficient deployment of sensor nodes in wireless sensor networks for agricultural field. J. Inf. Sci. Eng. **34**(4), 903–918 (2018)
19. Mulla, D.J.: Twenty five years of remote sensing in precision agriculture: key advances and remaining knowledge gaps. Biosyst. Eng. **114**, 358–371 (2013)
20. Bhardwaj, A., Sam, L., Bhardwaj, A., Torres, F.J.M.: LiDAR remote sensing of the cryosphere: present applications and future prospects. Remote Sens. Environ. **177**, 125–143 (2016)
21. Asher, J.B., Yosef, B.B., Volinsky, R.: Ground-based remote sensing system for irrigation scheduling. Biosyst. Eng. **114**, 444–453 (2013)

22. Kumar, S., Moore, K.B.: The evolution of global positioning system (GPS) technology. J. Sci. Educ. Technol. **11**(1) (2002)
23. Barik, R.K., Lenka, R.K., Dubey, H., Mankodiya, K.: TCloud: cloud SDI model for tourism information infrastructure management. In: Chaudhuri, S., Ray, N. (eds.) GIS Applications in the Tourism and Hospitality Industry, pp. 116–144. IGI Global, Hershey PA, USA (2018)
24. Boyd, D.S., Foody, G.M.: An overview of recent remote sensing and GIS based research in ecological informatics. Ecolog. Inform. **6**, 25–36 (2011)
25. Ammar, M., Russello, G., Crispo, B.: Internet of Things: a survey on the security of IoT frameworks. J. Inf. Secur. Appl. **38**, 8–27 (2018)
26. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of intrusion detection using data mining techniques. In: Progress in Computing, Analytics and Networking, pp. 753–764. Springer, Singapore (2018)
27. Pradhan, C., Das, H., Naik, B., Dey, N.: Handbook of Research on Information Security in Biomedical Signal Processing, pp. 1–414. IGI Global, Hershey, PA (2018)
28. Sarkar, J.L., Panigrahi, C.R., Pati, B., Das, H.: A novel approach for real-time data management in wireless sensor networks. In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, pp. 599–607. Springer, New Delhi (2016)
29. Hammoudi, S., Aliouat, Z., Harous, S.: Challenges and research directions for Internet of Things. Telecommun. Syst. **67**(2), 367–385 (2018)
30. Kalaivanan, K., Bhanumathi, V.: Reliable location aware and cluster-tap root based data collection protocol for large scale wireless sensor networks. J. Netw. Comput. Appl. **118**, 83–101 (2018)
31. Akkas, M.A., Sokullu, R.: An IoT-based greenhouse monitoring system with Micaz motes. Procedia Comput. Sci. **113**, 603–608 (2017)
32. Nagarajan, G., Minu, R.I.: Wireless soil monitoring sensor for sprinkler irrigation automation system. Wirel. Pers. Commun. **98**(2), 1835–1851 (2018)
33. Martinez, J.L., Claraco, J.L.B., Alonso, J.P., Ferre, A.J.C.: Distributed network for measuring climatic parameters in heterogeneous environments: application in a greenhouse. Comput. Electron. Agric. **145**, 105–121 (2018)
34. Foughali, K., Fathallah, K., Frihida, A.: Using cloud IOT for disease prevention in precision agriculture. Procedia Comput. Sci. **130**, 575–582 (2018)

Design Thinking on Geo Spatial Climate for Thermal Conditioning: Application of Big Data Through Intelligent Technology

Divyajit Das, Ashoke Kumar Rath, Dillip Kumar Bera and Bhubaneswari Bisoyi

Abstract This research paper has explored in understanding the design thinking aspect of thermal insulation capacity of rooftops through application of big data and intelligent technologies. The challenges encountered by the technocrats for analyzing the aspects of thermal insulation under varied geothermal conditions are addressed in the research paper. Large volume of information on huge infrastructure can be evaluated through intelligent techniques automatically; which improve the quality of lives. This actionable knowledge unlocking the value; exploring from the raw data shall increase design efficiency and reduce design cost through a proper management system. The smart data developed through intelligent big data analytics with statistical and machine learning shall provide solution to problems. They can provide solution for various geothermal locations across the globe with multi-objective problem solving designs. This research study shall investigate into testing of construction materials through different kind of material mix with various permutation and combinations.

Keywords Big data · Intelligent technologies · Thermal conditioning Geothermal design

D. Das (✉) · A. K. Rath · D. K. Bera
KIIT University, School of Civil Engineering, Bhubaneswar, India
e-mail: divyajitdas10@gmail.com

A. K. Rath
e-mail: akrathfce@kiit.ac.in

D. K. Bera
e-mail: dberafce@kiit.ac.in

B. Bisoyi (✉)
Sri Sri University, Cuttack, India
e-mail: bhubaneswari.b@srisriuniversity.edu.in

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_12

1 Introduction

Big Data has resulted in the introduction of new opening for various study, progress, advancement, and dealing. Basically, four Vs are followed: volume, velocity, veracity, and variety. Nowadays, Big Data and Cloud Computing have an interlinked structure which is taken with utmost importance in the sectors of information technology and geospatial communities. Tera-bytes and pera-bytes of data are recorded each day for earth observation and model simulation. This Geospatial big data has resulted in a large number of spatial datasets which has exceeded present computing capacity of the systems. Every day the size of geospatial data is increasing 20% of its data volume [1]. With the increase in demand for shelter and workplace, buildings are being constructed extensively. With the increase in building construction, the use of energy resources has increased for thermal conditioning. The spatial data play an important role in recording and observing the temperature data and the energy usage data in the form of units of electricity used. This can help to find out a viable media for conserving resource by carefully investing the natural resources by predicting its usage with the help of geospatial big data and resulting in saving the environment.

1.1 Big Data Processing

In the present scenario, big data has surfaced through innovative prospect in the field of research, development, and business. It is characterized by the four Vs: volume, velocity, veracity and variety and it might fetch justifiable assessment all the way through Big Data processing. Whereas, cloud computing has appeared as an innovative platform to impart computing as a facility for mitigating various processing needs with (a) on-demand services, (b) elasticity, (c) broadband access, (d) pooled resources and (e) measured services. In the Big Data domain mainly four geospatial scientific examples are established which includes climate studies, geospatial knowledge mining, land cover simulation, and dust storm modeling [2, 3]. The life cycle of Big Data processing has four examples of framework method supports which comprise of management, access, mining analytics, simulation, and forecasting. Similarly, Big Data is nothing without cloud computing for the geospatial communities distantly located which is explained in the following sequence [4–6].

2 Cloud Computing in Big Data

The two major philosophies in the area of information technology and geospatial communities that have come up in the present time are big data and cloud computing. Various organizations such as AAG, ESIP, AGU and the international GIS Science Conference have originated this special issue to encapsulate the most recent

encroachments on using cloud computing to intercept big geospatial data challenges. From past four years, Big Data and Cloud Computing are gaining tangible escalation and have aroused as a successful domain for research [7–11]. The areas that are covered within cloud computing and big data includes transportation, climate, remote sensing, end-user profiling, data access, and projection. An efficient data processing framework has been proposed by Zhou for drawing out huge trajectories of moving objects using GPS data [1, 12, 13]. Similarly, the framework presented by Li-Yang facilitates the setup, operation, filing, and image of climate modeling in a Model as a Service (MaaS) fashion [14, 15].

2.1 Geospatial Data in Cloud Computing

Daily there is an inflow of data in tetra-bytes or even in pera-bytes for the earth observation and model simulation. Usually, data acquired in non-traditional geospatial data acquisition methods considered to be quicker and quantitative in nature. Adding to the large volume, geospatial data exists in a number of structures and designs for various purposes, their exactness and ambivalence spread transversely through a broad scope and data are required to be formed in a fast velocity by application of sensors [16] (Fig. 1).

Cloud Computing has emerged as a new prototype to proffer computing as an efficacy service having five advantageous characteristics: (a) rapid and elastic requisite computing power; (b) pooled computing power for better utilization and sharing; (c) fast access broadband for communication; (d) on-demand access for computing and (e) pay-as-you-go for the parts used in traditional computing [3, 4, 5, 16]. Cloud Computing has adopted the service-oriented architecture model and enables "every-

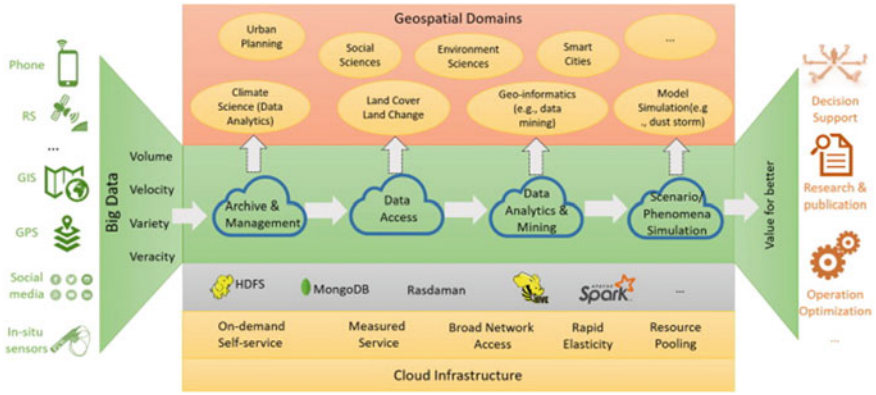


Fig. 1 Types of geospatial domains and various models in cloud infrastructure

thing as a service", which includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

Utilization of Cloud Computing for addressing Big Data issues is still in its early stages, and it is a formidable task. Various studies are being carried out in the area of climate studies, knowledge studies, land-use and land cover change analysis, and stimulation of dust storm through utilization of cloud computing [6, 12, 17].

Figure 2 illustrates the architecture of the cloud-based service-oriented workflow system for a climate model. This model includes: (a) compilation and running models on VMs; (b) the VM status information is provided on the cloud platform which is displayed on the VM service monitor after which the resource scheduling is done; (c) the output of the model is provided by the data analysis service as the input for analytics. The result of the analysis is accessible to the user through the internet. All these services are controlled by GUI. This system helps the application specified by workflow to automatically transition the service. The cloud-based methods consume 10 times less over the traditional method [18–20]. Large geospatial data has grand challenges during the lifecycle which revolves around data storage, access, manage, analysis, mining, and modeling. Some emerging challenges are yet to be addressed which are as follows:

I. Big geospatial data storage and management have become a matter of utmost importance; it includes optimization of the traditional and emerging database management techniques.
II. Space-time Big Data mining has some main components which are real-time data processing, extraction of information and automation in the extraction of information and knowledge.

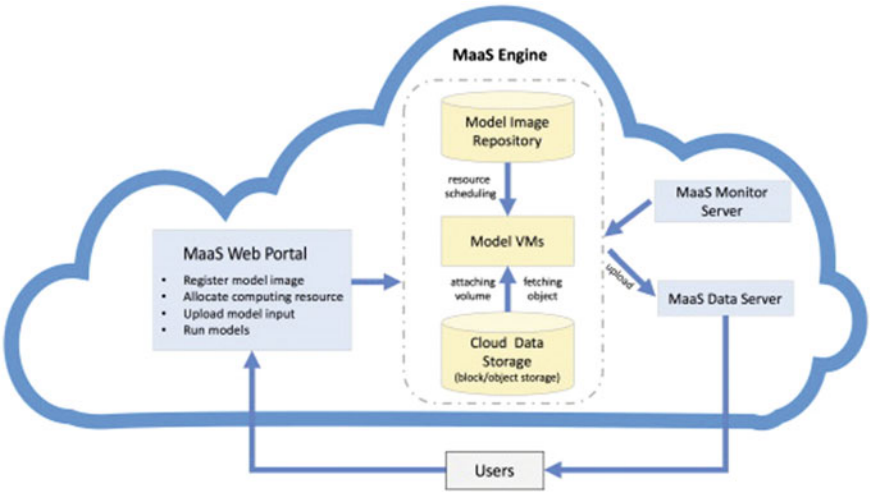


Fig. 2 Climate model study by cloud-based service-oriented workflow system

III. Where there are a sensitive data and corresponding users' privacy there is a challenge in assuring security for its protection.
IV. The energy efficiency and sustainability of the Cloud Computing resources are important as it directly depends upon its usage behavior on the cloud.
V. Space-time thinking methodologies are crucial which should be developed and formalized for optimizing Cloud Computing for big geospatial data processing.

3 Geospatial Big Data

Geospatial big data refers to spatial data sets exceeding the capacity of present computing systems. Generally, a considerable portion of big data is normally geospatial data, which is growing in its size day by day by at least 20% every day. Geospatial data has always been Big data. Big data analytics for geospatial data has a considerable attention to allow users to observe and use large amounts of geospatial data. It is observed that about 25 PB of data is being generated per day in Google, of which large portion of data is spatial-temporal data [4, 18, 21, 22]. With the increase in geospatial big data every day, the capability of high-performance computing needs to be improved majorly for modeling and simulation of those data. With recent improvements, we have a lot of opportunities for using the advanced analytics for geospatial big data.

4 Thermal Comfort in Buildings of Tropical Countries

The rate of urbanization in the tropical countries has raised concerns over depletion of available resources. Wise and efficient utilization of energy is a must to conserve them for a longer period of time as they are fast-depleting. By 2020 the developing countries are expected to use more resources than that of the advanced nations. Besides the transportation and industrial sector, the major consumer of resources is the building sector. This usage of energy is mainly due to the availability of electricity at a very cheap rate by the local governments, and it has become a macroeconomic problem [1, 7]. Even the commercial buildings have more energy requirement than that of the normal buildings as they have to maintain an ambient environment maintaining the humidity and temperature for better indoor comfort. It comprises of about 30–60% of the total energy consumption. It is also identified as the source of largest energy saving possibility. One of the most important and useful methods of identifying thermal perceptions is the thermal comfort analysis of a particular building space and the possibility of energy saving [1, 15]. This analysis can be saved as a spatial data in the form of big data and cloud computing. This data would further facilitate the prediction and usage of energy resources in the different environment. Besides the use of air conditioning machines, the adaptation and acclimatization of occupants

in the buildings in the tropical environment would be able to contribute largely to energy savings [9, 23, 24].

Occupants craving for thermal comfort results in using energy for thermal conditioning which directly relates to the building regions having no natural ventilation and greenery. The temperature changes can be tabulated and observed which can later be saved for further use in the form of geospatial big data [3, 17]. Lack if ventilation results in a rapid increase of building operating cost which requires improvement in energy efficient building designs. Nowadays, air conditioning requirement among the younger generations has increased considerably as they are exposed to thermal conditioning at a younger age which makes then non-resistant to change in temperature. Taking the cost factor into consideration an algorithm can be created for stating the adverse effect of modulating use of air-conditioning which affects the cost by using Geospatial Big Data [25–29]. In many cases, almost all of the non-commercial high-rise buildings have full air-conditioning where occupants have minimal control over the thermostat and air flow speed by mechanical means. Installing fixed air conditioning thermostat set-point can cater to the easiest way to control energy consumption thereby reducing irregular consumption. It would also result in no extra service cost and maintenance charges.

4.1 Engineering Thermal Control Measures

In the civil environment the elimination of risk and hazards caused due to the erratic thermal condition has to be checked and prevented for a better working environment though it incurs investment, it will ensure higher productivity and sustained development in the scales of economy. The plants should avoid installation of mechanical devices like air-conditioners, ventilators, and heaters which are energy intensive and add to the operational cost and maintenance due to consumption of energy. Rather the buildings should have a design thinking approach with predictive temperature adaption techniques. Buildings constructed for the purpose should choose eco-friendly materials and appropriately use them in the design stage. It has to strike a sustainable green option for achieving energy efficacy. The building has to accomplish with an eco-friendly design calibrated to climate change [8, 11, 30, 31].

4.2 Efficient Green Building

The idea of green buildings employs eco-friendly materials to reduce energy demand by optimally using available resources. Zero energy houses have been popular in the western world. In the tropical environment of the developing countries, eco-friendly traditional know-house/knowledge, and thermal resistant materials for the building have been popularly used in the villages. The eco-friendly materials include bamboo, timber, straw, grass, paper flakes, compressed earth blocks, bagged earth,

rammed earth, clay, sisal, sea-grass, coconut, wood fiber have been used as natural materials for buildings. To assess their thermal performance and properties, specific heat absorption capability, transmissibility, heat transfer coefficients, and thermal conductivity are required. As a matter of fact, this critical research review shall focus on the variety of building materials used in the making of green building. It shall also analyze the materials that have the properties to absorb heat which may be applied in the workplace for thermal comforts resulting in productivity [11, 16, 20, 32].

4.3 Building Thermal Comfort: Influencing Attributes of Geospatial Data

Besides energy saving and designing of a house with an eco-friendly approach, thermal comfort of the building are as well influenced by multi-dimensional factors. They are building design, orientation, ventilation, space utilization and eco-friendly materials. It nonetheless has to blend and integrate the modern and traditional energy-saving practices. Usage of geospatial data of thermal reading would play an important role in the prediction of temperature variation over a period of time. The data created by various temperature sensor types are recorded over a period of 2 years and saved on accessible cloud storage. Any building is not a unit away from external environment and is separate. Climatic conditions affect and influence the thermal atmosphere of the building directly and indirectly. Therefore, internal thermal condition affecting aspects have to harmonize with the influence of external humidity and temperature. In India, Middle-East hot temperature and in the western world cold temperature can enter into the building easily through translucent and transparent materials [19, 20, 24]. Thus the thermo-physical properties of the material are significant. Lower thermal conductivity and thermal diffusivity reduce temperature swing inside the building. Similarly, higher thermal conductivity based materials prove ineffective. Nylon, polystyrene foam has optimal thermal comfort for flooring in hot and humid environment. Ventilation is a significant factor for thermal comforts of a building or a structure. Modification of the conventional materials to the modern technology has contributed to the thermal comforts. Moreover natural ventilation improves the comforts in the buildings in the hot and humid atmosphere. In order to ensure natural ventilation the building can use wind tower, wind scoops, ventilation chamber, double façade, chimney, embedded duck and atrium. Persian building architectural design for natural ventilation of the interior space in a building is a dome with the opening at the peak (Bernoulli's Equation) [24, 32, 33].

4.4 Thermal Comforts of Building Appliances

To control the temperature in the buildings the standard three methods are adopted, they are conduction, convection, and radiation. Through these three methods generally, heat gets transferred into the building. They enter through the walls, windows, and roof from all around the surrounding atmosphere. Gravel concrete and marble is an excellent conductor of heat which should be evaded for external construction [34–38]. Hence to minimize transfer of heat from outside to indoors, materials like wood, glass and another alloy should be chosen for the windows, walls, and ceilings. Windows have been scientifically made to check the transfer of heat. The mutual radiation between the walls and the ceiling also has an adverse effect. Materials that absorb radiation shall lower the temperature within the building. Kunzel et al., quotes, "Building materials should be non-hygroscopic and capillary-inactive (hydrophobic)". Water is also silent rouge which can disturb the thermal insulation [39–42]. Currently, modern buildings with load-bearing capacity and durability have threatened the utilization of natural building materials. Natural and conventional materials have their thermal comforts and technological innovation on the materials have a great potential of usage in a tropical climate. Advanced materials like polymer skins, vacuum insulation panels, and gas-filled panels have immense potential for thermal insulation. Currently, a happy combination of synthetic and natural construction materials are being used as hybrid materials [7, 8, 43, 21].

4.5 Indigenous Materials for Buildings

Europe as a cold country has been using cork for insulation for ages. Granules of the cork are compressed at high temperature for low thermal conductivity. They were used in construction application such as flooring, exterior, interior walls and ceilings. They also provide acoustic insulation. Similarly timber and wood are known for their Omni-application in floors, roofs, walls, windows and doors and so far. Wood is a hygroscopic material. And its thermal properties are functions of moisture content. Fiberboard or hardboard panels made of wood pulp have less thermal conductivity values than solid wood because it has air spaces in the fiber. The Indoor climate is moderated by wood-based products for its diurnal changes in the humidity. Wood has a higher heat capacity (1.6–3 kJ/kgK) and comparatively less density in relation to concrete, brick, glass and plastic. Application of straw in Australia, France and Mexico as building material since ages is a popular technology. Straw bell buildings have stupendous thermal performance due to isolative value of the bells and solid plaster shell of the interiors [11, 31, 44, 45]. The thermal resistance value of straw ranges from 6.51 to 7.82 W/m^2 K for 55 cm thick straw bale. But it has inherent disadvantages as they are also highly inflammable. Similarly building construction made of real minerals and rocks for rock wool insulation it has capability to obstruct sound and heat. This method is a superior conductor but stops heat moment and

are sustainable due to eco-friendly properties. Besides brick is the most significant construction material. The thermal conductivity of clay brick is high as compared to red brick which require less energy for sustaining thermal comfort. The internal temperature control of red brick building is stable during extreme fluctuation in temperature in the hot and humid climate. Mud bricks maintain the indoor temperature cooler during summer. Fly-ash bricks (FAB) are becoming popular for green buildings. FAB recycles the by-products and is utilized for thermal conductivity. FAB has thermal conductivity of (0.90–1.05 W/mK) as compared to the traditional bricks having conductivity of (1.25–1.35 W/mK). The versatile insulation material named tuff-stones is amply used in masonry and is a heat insulator due to its porosity. Walls in tuff- stones are better and are bio-degradable. They are affordable by common man [8, 10, 46].

4.6 Synthetic Substance Made Building Materials

The thermal conductivity of the regular concrete used lies within 1.3–1.5 W/mK and the content of moisture is as high as 8%. The character of the most regularly used building material is to store heat and hold-up its transmission in order to provide thermal comfort. The thermal conductivity and heat capacity of cement paste is low. Whereas, Portland Cement Concrete (PCC) made of rubber mix also provides benefit for insulation. A 20.34 cm thick ACC wall with no insulation delivers an R-Value of 13.28 with the benefits of retaining coolness from air conditioning for a longer time. It makes it preferable to be used in tropical hot climates. Its use started in the 1990s. Moreover, vermiculite concrete is hydro-silicate mineral that is classified as a Phylo-silicate. It is lightweight and environmentally beneficial insulator. Rigid foams (polystyrene, extruded polystyrene, polyurethane) and flexible foams (polyethylene etc.,) and spray foams contribute to thermal comforts. Phase Change Materials (PCM) stores the heat by trapping the thermal energy for the betterment of human comfort. Polymer skins create a skin held in between structures, based on the thermal conductivity the skin inflates or deflates [4, 22, 23, 30]. The product like polymer skins consist of a pneumatic cushion divide with a skin sandwiched between the structures. This sandwiched skin in between inflate and deflate in circumstances depending on the thermal conductivity. They are applied in mega-construction projects worldwide in the advanced nations. In smaller buildings the applicability has not been feasible yet; similarly aerogels consisting of porous synthetic materials are one of the lightest of the materials used in the buildings. These arrow gels have been successfully used for thermal insulation. Low tensile strength is the disadvantage. Nonetheless Vacuum Insulation Panel (VIP) similarly has quite low thermal conductivity mostly used in high end insulation. This has a performance of insulation in comparison to the materials used for conventional insulation; VIP is 4–8 times better. A standard VIP product has internal core covering, barrier envelope and a getter along with a heat seal [2, 3, 5, 47]. Thermal conductivity is 0.002–0.004 W/mK which is based on the use of core material. In this material air is better evacuated when core material of the VIP is

good. Barrier envelope protect from the damage of environment. Getters/desiccants absorb the vapours that penetrate through the barriers. Mostly core materials are non-biodegradable. VIP has been applied in zero energy buildings/passive houses. To make a small house high thermal efficient VIPs are used, which insulates ceilings, walls, floors and roofs. Integrating pre-fabricated construction materials with VIP is a reasonable idea. Safe Memory Polymers (SMP) act as sensor for ventilators in the vents which control indoor temperature. SMP can be used in air-conditioners can automatically switch off if the windows are open. It conserves energy. SMPs are largely used in walls improvised for thermal comfort. It creates scope for thin walls, extra space for interiors and low expense for energy. Embodied degree of energy for these materials is higher which has a price for environmental degradation which is recovered again by saving energy by utilization of these materials [12, 20, 24, 46].

5 Components of the Buildings—Sundry Applications

Thermal comfort depends on the optical properties and the transfer of heat coefficient is proportional to the quality of glass used in windows. Optical properties are related to transmittance and absorbance. Materials used in windows relate to the thermal comfort indoors. Triple-glazing windows, coatings on transparent glasses and improvised frames reduce heat exchange [2, 4, 6, 48]. Shutters over the windows reduce 51% of the heat flows as suggested by Paul Becker. Window glassing with low/E-coating improve the properties of insulation compared to ordinary glass. In the building color use of light colour in the outer surface reduce the thermal temperature and provide comfort. High Albedo paint reduces indoor temperature. White color coatings have better performance than aluminum coatings. Thermal control coatings on the rooftop reduce the degree of temperature by 33 ˚C [5, 6, 49]. Coating with white elasto-material together with higher reflectivity is found to be comparatively cooler. Gray wool paint coating with titanium dioxide has reflective qualities and is efficient. Multi mix mineral, ecological paint made out of a mix of milk and vinegar also shown higher solar reflectance and was used in wood, concrete, metal and roofs. Eventually, higher solar reflectivity roofs have higher emissivity. This cool the building roofs can reduce by thermal insulation to 35% less. Red-brown roof tiles have less brightness value (10–20%). Lime-silica brick (0.45 absorptivity and 55% brightness) has better thermal performance than spruce wood (0.4 absorptivity and 50% brightness). High SRI and high emittance are better choice for cool roofs in hot temperatures (43–46 ˚C). Hollow Clay Tiles (HCT) used in Athens in roofs have exceptional energy saving. Application of elasto-metric coating on the roof cools down the indoor temperature. Most cheap and effective thermal comfort in the building can be a roof garden [1, 12, 19]. Green roofs act as a cooling device and insulator for the roofs. Green roof is also environment friendly. Rooftop natural/synthetic man-made shades like pavilion on the rooftops like palm leaves, roof sheets, coconut leaves, straws and vegetation improve thermal comfort.

6 Conclusion

The aforesaid review regarding concern for rise in the thermal temperature due to change in the climate and consequent health problems in the tropical climate has been studied. Eventually thermal comfort has implication to health psychology, happiness and productive work. The literature reviewed on building materials for understanding into the aspect of mechanism desire for thermal comfort inside the building. It is required in the hot, sultry and humid climates. Through minimizing use of energy thermal comfort can be achieved in the living space. Tropical countries with shortage of energy can use this passive technique effectively to control thermal temperature. Array of materials with different properties and characteristics for passive cooling has been studied. In order to adopt a sustainable solution to the rise in the temperature a host of adoption is required to tackle the thermal disturbance. Building living space comfort is pivotal. The review revealed about the method of obtaining the usage of selected materials. Mostly it is related to enveloping the building having natural inherent properties for rendering thermal insulation. To protect from external environment reflective paints, green roof and advanced materials have been discussed [40], [50]. To control the impact of solar radiation and ensure thermal insulation properties of materials were explored. Those include conventional materials, environmental friendly substances and tech-savvy recent objects. The covering of the building needs to be built with materials that have low thermal conductivity, diffusivity and absorptive. The applications of PCMs, VIPs, ACC and Polymer Skin etc. have been assessed to have potential for building envelope. To minimize temperature and heat load inside the building external surfaces can be painted with light colours or reflective colours in the tropical zones. To ensure environmental sustenance optimally utilization of indigenous raw materials, low cost recyclable natural materials are preferable and encouraged. New ideas and concepts about cost-effective materials utilized for reducing temperature, thermal temperature inside the house is essential. Co-benefit of energy efficacy and thermal comfort shall invite innovative ideas in the construction materials for shielding the posterity from the hazard of thermal stumping resulting from unanticipated spike in the degree in temperature due to change in climate. It is important to ensure that the living space and the working space are significant.

References

1. Das, H., Mishra, S.K., Roy, D.S.: The topological structure of the Odisha power grid: a complex network analysis. IJMCA **1**(1), 012–016 (2013)
2. Barik, R.K., Dubey, H., Misra, C., Borthakur, D., Constant, N., Sasane, S. A., Mankodiya, K., et al.: Fog assisted cloud computing in era of big data and internet-of-things: systems, architectures, and applications. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges. Springer, Cham, pp. 367–394 (2018)
3. Barik, R.K., Tripathi, A., Dubey, H., Lenka, R. K., Pratik, T., Sharma, S., Das, H., et al.: Mistgis: optimizing geospatial data analysis using mist computing. In: Progress in Computing,

Analytics and Networking. Springer, Singapore, pp. 733–742 (2018)
4. Bisoyi, B., Das, B.: Necessitate green environment for sustainable computing. Adv. Intell. Syst. Comput. **380**, 514–524 (2015)
5. Bisoyi, B., Das, B.: Green technology for attaining environmental safety and sustainable development. Int. J. Mech. Eng. Technol. (IJMET) **9**(3), 1087–1094 (2018)
6. Bisoyi, B., Das, B.: Development in the field of technology for cooperative problem solving utilizing nonconventional energy resources in India & future trend. Int. J. Sci. Res. Manage. **3**(1), 2321–3418 (2015)
7. Goss, W.P., Miller, R.G.: Thermal properties of wood and wood products. Report (2013)
8. Hayashi, C., Tokura, H.: Effects of head cooling on sweat rate in exercising subjects wearing protective clothing and mask for pesticide. J. Appl. Hum. Sci. (1996)
9. Mishra, B.S.P., Das, H., Dehuri, S., Jagadev, A.K.: Cloud Computing for Optimization: Foundations, Applications, and Challenges, vol. 39. Springer (2018)
10. Panagiota, A., Agis, M.P.: Occupants' thermal comfort: state of the art and the prospects of personalized assessment in office buildings. Energy Build. (2017)
11. Panigrahi, C.R., Tiwary, M., Pati, B., Das, H.: Big data and cyber foraging: future scope and challenges. In: Techniques and Environments for Big Data Analysis. Springer, Cham, pp. 75–100 (2016)
12. Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. **35** (2015)
13. Mohamed, Z., Collier, P.A., Steve, G.S., Davern, M.J., Leech, S.A.: Enhancing the business value of business intelligence: the role of shared knowledge and assimilation. J. Inf. Syst. Fall **27** (2013)
14. Hao, S., Song, M.: Technology-driven strategy and firm performance: are strategic capabilities missing links? J. Bus. Res. **69** (2016)
15. Kalpič, B., Bernus, P.: 2016 business process modelling through the knowledge management perspective. J. Knowl. Manage. (2006)
16. Epstein, Y., Moran, D.S.: Thermal comfort and the heat stress indices. J. Ind. Health (2006)
17. Bisoyi, B., Das, B.: Organic farming: a sustainable environmental ingenuity for biotechnological intervention towards a green world. Int. J. Innov. Res. Sci. Eng. Technol. **6**(9), 179 (2017)
18. Das, H., Jena, A.K., Badajena, J.C., Pradhan, C., Barik, R.K.: Resource allocation in cooperative cloud environments. In: Progress in Computing, Analytics and Networking. Springer, Singapore, pp. 825–841 (2018)
19. Das, H., Panda, G.S., Muduli, B., Rath, P.K.: The complex network analysis of power grid: a case study of the West Bengal power network. In: Intelligent Computing, Networking, and Informatics. Springer, New Delhi, pp. 17–29 (2014)
20. Goldsworthy, M.J.: Building thermal design for solar photovoltaic air-conditioning in Australian climates. Energy Build. (2017)
21. Wilson, A., Piepkorn, M.: "Green Building Products", the Green Spec Guide to Residential Building Materials, Canada (2009)
22. Bisoyi, B., Das, B.: Adapting green technology for optimal deployment of renewable energy resources and green power for future sustainability. Indian J. Sci. Technol. **8**(28), 1–6 (2015)
23. Hyde, R.: Climate Responsive Design: A Study of Buildings in Moderate and Hot Humid Climates. E. & F.N. Spon (2000)
24. Liew, A.: Understanding data, information, knowledge and their inter-relationships. J. Knowl. Manage. Prac. **8**, 2007 (2007)
25. Das, H., Naik, B., Behera, H.S.: Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In: Progress in Computing, Analytics and Networking. Springer, Singapore, pp. 539–549 (2018)
26. George, G., Haas, M.R., Pentland, A.: Big data and management: from the editors. Acad. Manage. J. (2014)
27. Kumar, A., Singh, O.P.: Advances in the building materials for thermal comfort and energy saving. J. Recent Pat. Eng. (2013)

28. Meral, O.: Thermal performance and optimum insulation thickness of building walls with different structure materials. Appl. Therm. Eng. (2011)
29. Pattnaik, P.K., Rautaray, S.S., Das, H., Nayak, J. (eds.). Progress in computing, analytics and networking. In: Proceedings of ICCAN 2017, vol. 710. Springer (2018)
30. Hirunlabh, W., Khedari, J., Kongduan, P.: Study of natural ventilation of houses by a metallic solar wall under tropical climate. Renew. Energy (1999)
31. Nag, P.K., Nag, A., Sekhar, P., Shah, P.: Perceived heat stress and strain of workers. Asian-Pacific News Lett. Occup. Health. Saf. (2011)
32. Lindblom, A., Tikkanen, H.: Knowledge creation and business format franchising. Manage. Decis. **48** (2010)
33. Jani, D.B., Mishra, M., Sahoo, P.K.: Solid desiccant air conditioning—a state of the art review. Renew. Sustain. Energy Rev. (2016)
34. Pramanik, M.I., Lau, R.Y.K., Demirkan, H., Azad, M.A.K.: Smart health: big data enabled health paradigm within smart cities. Expert Syst. Appl. **87** (2017)
35. Salem, A., Saleel, C.A., Abdul Mujeebu, M.: Air-conditioning condensate recovery and applications—current developments and challenges ahead. Sustain. Cities Soc. (2018)
36. Sarkhel, P., Das, H., Vashishtha, L.K. Task-scheduling algorithms in cloud environment. In: Computational Intelligence in Data Mining. Springer, Singapore pp. 553–562 (2017)
37. Shastri, V., Mani, M., Tenorio, R., Impacts of modern transitions on thermal comfort in vernacular dwellings in warm humid climate of Sugganahalli. J. Indoor Built Environ. (2012)
38. Simpson, A., Stuckes, A.D.: Thermal conductivity of vermiculite concrete: effect of inclusion of shape. Build. Serv. Eng. Res. Technol. (1987)
39. Spiegel, R., Meadows, D.: Green Building Materials: A Guide to Product Selection and Specification. Wiley (2010)
40. Ungkoon, Y., Hirunlabh, J.: A Preliminary Study of Hygro-thermal Performance of Autoclaved Aerated Concrete Block Sunder Hot Humid Climate of Thailand. Building Scientific Research Center, King Mongkut's University of Technology, Thonburi (2012)
41. Wu, P.-Y., Cheng, C.-W., Kaddi, C.D., Venugopalan, J., Hoffman, R., Wang, M.D.: Omic and electronic health record big data analytics for precision medicine. IEEE Trans. Biomed. Eng. (2017)
42. Yaodong, T., Ruzhu, W.: Theoretical investigation of a novel unitary solid desiccant air conditioner. Sci. Technol. Built Environ. (2016)
43. Nayak, J., Naik, B., Jena, A.K., Barik, R.K., Das, H.: Nature inspired optimizations in cloud computing: applications and challenges. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges. Springer, Cham, pp. 1–26 (2018)
44. Parsons, K.: Human Thermal Environment: The Effects of Hot, Cold and Moderate Environment on Human Health, Comfort and Performance, 2nd edn. Taylor and Francis, London (2003)
45. Raja, I.A., Nicol, J.F., McCartney, K.J., Humpherey, M.A.: Thermal comfort: use of controls in naturally ventilated buildings. J. Energy Build. (2001)
46. Haghighi, A.P., Maerefat, M.: Solar ventilation and heating of buildings in sunny winter days using solar chimney. Sustain. Cities Soc. (2014)
47. Reddy, K.H.K., Das, H., Roy, D.S.: A data aware scheme for scheduling big-data applications with SAVANNA hadoop. In: Futures of Network. CRC Press (2017)
48. Bansal, N.K., Garg, S.N., Kothari, S.: Effect of Exterior Surface Color on the Thermal Performance of Buildings, Build. Environment (1992)
49. Sahani, R., Rout, C., Badajena, J.C., Jena, A.K., Das, H.: Classification of intrusion detection using data mining techniques. In: Progress in Computing, Analytics and Networking. Springer, Singapore, pp. 753–764 (2018)
50. Zhang, J., Haghighat, F.: Simulation of earth-to-air heat exchangers in hybrid ventilation systems. In: Ninth International IBPSA Conference, Building Simulation, Montreal (2005)

Hyperspectral Remote Sensing Images and Supervised Feature Extraction

Aloke Datta, Susmita Ghosh and Ashish Ghosh

Abstract In the last three decade, one of the significant breakthrough in remote sensing is to introduce of hyperspectral sensors, which acquire a set of images from hundreds of narrow and contiguous wavelengths of the electromagnetic spectrum from visible to infrared regions. Images, which are captured by these sensors, have detailed information in the spectral domain to identify and distinguish spectrally unique materials. To recognize the objects present in hyperspectral images, classification/clustering task need to be performed. But due to the presence of huge number of attributes, classification technique becomes more complex. So, before performing the classification task, reduce the number of attributes (denoted by dimensionality of the data) is an important step where the aim is to discard the redundant attributes and make it less time consuming for classification. In this chapter, few supervised feature extraction techniques for hyperspectral images i.e., prototype space feature extraction (PSFE), modified Fisher's linear discriminant analysis (MFLDA), maximum margin criteria (MMC) based and partitioned MMC based methods are explained. Experiments are conducted over different hyperspectral data set with different quantitative measures to analyze the performance of these feature extraction methods.

A. Datta (✉)
Department of CSE, NIT Meghalaya, Shillong, India
e-mail: alokedatta@nitm.ac.in

S. Ghosh
Department of CSE, Jadavpur University, Kolkata, India

A. Ghosh
Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India

H. Das et al. (eds.), *Cloud Computing for Geospatial Big Data Analytics*,
Studies in Big Data 49, https://doi.org/10.1007/978-3-030-03359-0_13

1 Remote Sensing Images

Remote sensing has been defined as the field of study of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device without coming into physical contact with the object, area, or phenomenon under investigation [1]. In a very common way, it is said that remote sensing measures the object properties on Earth's surface to capture specific information to make decisions. For example, a weather satellite is used to measure the global atmospheric parameters that ultimately help to take decision for weather forecasting. The main three components of remote sensing are: the reflectance/radiated signals from an object or phenomena, the sensor which is on a platform apart from the object, and accumulating knowledge or information through analysis and displaying in a spatial grid, i.e., a two-dimensional images. Depending on the platform from which the remote sensing sensor captures images, the images have different categorizations. If an image is captured from sensor on the platform of aircraft then it is called air-borne image. On the other way, if it is taken from satellite then it is called space-borne image [1, 2]. The main objective of remote sensing systems is to provide a repetitive and consistent view of the Earth facilitating the ability to monitor the Earth system and the effects of human activities on earth. A few application areas of remote sensing images are weather prediction, agricultural forecasting, resource exploration, land cover mapping, environmental monitoring etc. [3–5].

Fundamental basis for space-based remote sensing is that information is potentially available from the electromagnetic energy field arising from the Earths surface and from the spatial, spectral, and temporal variations in that field. Remotely sensed data may be collected in various ways, e.g., multistage sensing, where data from a site are collected from multiple altitudes. It may entail multitemporal sensing, where data from a site are collected on more than one occasions; or, it may involve multispectral and hyperspectral sensing, whereby data are acquired simultaneously in several spectral bands [1, 3, 5].

Multispectral images are simultaneously collected by sensors in several selected bandwidths of electromagnetic radiation from the platform of airplane (called air-borne) or from satellite (called space-borne). Generally, the sensors capture the reflected energy from the visible to near infra-red wavelength range, but they also measure the radiated energy of thermal infra-red wavelength regions with some sensors. Within this range, a few images (mostly 5–12) are taken from some selected bands. A few well-known multispectral sensors are Advanced Very High Resolution Radiometer (AVHRR), the Landsat Multi-Spectral Scanner (Landsat MSS), the Landsat Thematic Mapper (Landsat TM), the Landsat Enhanced Thematic Mapper Plus (Landsat ETM+) [1, 2, 6].

One of the recent advancements in remote sensing and geographic information is the development of hyperspectral sensors. They are able to capture images within a very narrow and contiguous wavelength range. Hyperspectral remote sensing combines imaging and spectroscopy in a single system [5]. An imaging system captures a picture of a remote scene related to the spatial distribution of the power of reflected

(or emitted) electromagnetic radiation integrated over some spectral band. On the other hand, spectroscopy is the study of light that is emitted by or reflected from the materials and its variation in energy with wavelength. Spectroscopy can be used to detect absorption features due to specific chemical bonds in a solid, liquid, or gas. As applied to the field of optical remote sensing, spectroscopy deals with the spectrum of sunlight that is reflected (scattered) by materials at the Earth's surface. Hyperspectral sensors are designed to focus and measure the light reflected from many adjacent areas on the Earth's surface [7]. A few well-known hyperspectral sensors are Hyperion Earth Observation-1 (EO-1), Hyperspectral Digital Imagery Collection Experiment (HYDICE), Airborne Visual Infrared Imaging Spectrometer (AVIRIS) etc. [4, 7].

2 Hyperspectral Images and Dimensionality Reduction

A set of hundreds images with narrow and contiguous wavelengths of the electromagnetic spectrum from visible to infrared regions are captured by hyperspectral sensors. Detection of different targets, identification of material, mapping of existence of mineral in Earth surface, identification of different species in the domain of vegetation, mapping details of surface properties etc. are few of the application area of hyperspectral images. In the above mentioned area, the basic task needs to be performed is to grouping (recognition/classification) of homogeneous pixels with defined [6, 8].

Recognition of patterns can be categorized in two ways: classification (or supervised classification) and clustering (also known as unsupervised classification) [9]. Classification task is a very challenging task in the field of hyperspectral images because of a large number (hundreds) of attributes for each pixel. The efficiency of a classifier depends on the number of patterns, number of attributes and complexity of classifier. The minimum number of training patterns required for proper training may be an exponential function of the number of features present in a data set [10]. Increase the number of features may not increase the efficiency of a classifier, always, due to small sample sizes relative to the features. This paradox behavior is known as "curse of dimensionality" [9, 11]. On the other way, the neighboring bands are generally strongly correlated for hyperspectral images. So, the possibility is that increasing the spectral resolution may incorporate very less relevant information. So, in the domain of hyperspectral images, dimensionality reduction is an important task to perform before classification [12–14].

In dimensionality reduction, the basic two approaches are feature selection and feature extraction [11, 15]. In brief, feature selection [9, 16–22] is the process of selecting a subset of features from the original set of features, whereas, feature extraction [23–27] is a method of transforming the original set of features into a lower dimensional space. The main two advantages of performing feature selection and feature extraction are to improve classification accuracy by avoiding curse of dimensionality and to reduce the computational cost for classification or clustering

of data. Depending on the availability of labeled patterns, feature selection/extraction is categorized into supervised and unsupervised ones. Supervised methods use class label information of patterns and, when no labeled patterns are available, unsupervised method is used for dimensionality reduction.

Due to the presence of large dimensions in hyperspectral images, search strategies employed for feature selection mostly yield suboptimal solutions (subset selection). To increase the efficiency of classification/recognition of land cover types in hyperspectral images, feature extraction is a good choice for dimensionality reduction if it is not necessary to retain the original features. In feature extraction, a newly generated feature inherits the properties of the original features. If class label information of some of the pixels is available, a supervised feature extraction technique is considered [24, 26]. The method relies on class label information of at least few patterns. This chapter concentrates on supervised feature extraction methods in hyperspectral images.

In hyperspectral image classification (i.e., the classification of pixels of hyperspectral image), the features (properties) of each pixel are nothing but the response/reflectance of different bands/wavelengths which are measured by hyperspectral sensors. So, feature and band are synonymously used in the literature of dimensionality reduction of hyperspectral images, as well as, in this chapter.

3 Supervised Feature Extraction in Hyperspectral Images

A brief description of supervised feature extraction methods in the field of hyperspectral images, namely, modified Fisher's linear discriminant analysis (MFLDA) [28], prototype space feature extraction (PSFE) [20], maximum margin criterion based feature extraction (MMC) [29] and partitioned MMC based feature extraction is given in this section.

3.1 Modified Fisher's Linear Discriminant Analysis (MFLDA) Based Feature Extraction Method

Fisher's linear discriminant analysis (FLDA) is a traditional method of supervised feature extraction, which tries to maximize the Rayleigh quotient [30]. Rayleigh quotient is a quantitative measure of class separation which actually calculates the between-class scatter matrix and average within-class scatter matrix ratio.

Let $\mathbf{x}_i \in \Re^D$, $(i = 1, 2, ..., N)$ be D-dimensional pattern and ω_j, $(j = 1, 2, ..., C)$ be the associated class labels, where N and C denote the total number of samples and classes, respectively. Our main aim is to keep information intact in transformed space also, i.e., when $\mathbf{x}_i$ is transformed into $\Re^d$ from $\Re^D$, where $d \ll D$.

The between-class and within-class scatter matrices, denoted as S_b and S_w, are defined as

$$S_b = \sum_{i=1}^{C} p_i(\mu_i - \mu)(\mu_i - \mu)^T, \tag{1}$$

and

$$S_w = \sum_{i=1}^{C} p_i S_i; \tag{2}$$

where the number of classes is C; mean vector is μ_i and priori probability of class ω_i is p_i, respectively. Here μ, overall mean is calculated by following equation:

$$\mu = \sum_{i=1}^{C} p_i \mu_i. \tag{3}$$

S_i is the within-class scatter matrix of class ω_i and defined as

$$S_i = \sum_{j=1}^{n_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T; \tag{4}$$

where n_i is the number of patterns of class ω_i.

The goal of Fisher's linear discriminant analysis (FLDA) is to find a transform vector W such that the Raleigh quotient is maximized, which is defined as

$$q = \frac{W^T S_B W}{W^T S_W W}; \tag{5}$$

W can be determined by solving a generalized eigen problem specified by

$$S_B W = \lambda S_W W, \tag{6}$$

where λ is a generalized eigenvalue.

But due to unavailability of enough training patterns, as well as, complete knowledge of all the classes, the original FLDA is modified (which is called modified FLDA (MFLDA)) to avoid the above mentioned difficulties. In spite of calculating within-class scatter matrix, MFLDA considers the total scatter matrix, which is calculated from the unknown information of all class labels (i.e., patterns without class label); and between class scatter matrix is estimated from the available class signature, i.e., only one training pattern from each class is sufficient to estimate it. The main target of MFLDA is to maximize the ratio of between-class scatter matrix to total class scatter matrix [28].

Let the total scatter matrix S_T be defined as

$$S_T = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \mu)(\mathbf{x}_j - \mu)^T; \tag{7}$$

and it can be related with S_W and S_B by

$$S_T = S_B + S_W. \tag{8}$$

So the maximization of Raleigh is equivalent to maximizing

$$q = \frac{W^T S_B W}{W^T S_T W}; \tag{9}$$

which follows the same idea of FLDA, i.e., the solution will be the eigenvectors of the generalized eigen problem:

$$S_B W = \lambda S_T W. \tag{10}$$

Here, the only available information which is needed, is the class signatures, i.e., at least one pattern from each class which is treated as class means to calculate the S_B.

3.2 Prototype Space Feature Extraction Method (PSFE)

In this method [20], features are represented in prototype space (PS), where feature vectors describe the channel behavior in terms of their reflectance. Then, fuzzy C-means (FCM) clustering operation is performed over features in prototype space to distinguish the highly correlated features. Transformation matrix is formed by a linear combination of reflectance of features weighted by their class membership values. A small number of isolated features, which are not in any cluster may also be included to form the linear transformation matrix, depending on their information content [20]. This method is executed in both supervised and unsupervised manner. In supervised PSFE, the classes's spectra, i.e., class representatives, are computed from the class means of the training data.

3.3 Maximum Margin Criterion (MMC) based Feature Extraction Method

A maximum margin criterion based linear transformation is performed for the hyperspectral image to overcome the drawbacks of Fisher's linear discriminant analysis

(FLDA) based feature extraction methods. Limitations of FLDA based feature extraction are the singularity of within-class scatter matrix which occurs in high dimensional data with small sample size (SSS) problems, and the maximum number of extracted features is limited, which depends on the number of classes present in the data set. To avoid the singularity problem of FLDA, instead of Rayleigh coefficient, the difference of both between-class scatter and within-class scatter, called maximum margin criterion (MMC) [29] is used as a discriminant criterion. Since the inverse matrix does not need to be constructed, the SSS problem in traditional FLDA is alleviated. Geometrically, MMC maximizes the (average) margin between classes and has the advantages of effectiveness and simplicity [29, 31]. This MMC based feature extraction (for hyperspectral images) which uses the difference of between-class and average within-class scatter matrices to calculate the maximum margin criterion. The method is a supervised one as class label information is needed to calculate the between-class and within-class scatter matrices. In this subsection, the MMC based feature extraction method is described in the field of hyperspectral images.

Let $\mathbf{x}_i \in \Re^D$, $(i = 1, 2, ..., N)$ be D-dimensional pattern and ω_j, $(j = 1, 2, ..., C)$ be the associated class labels, where N and C denote the total number of samples and classes, respectively. Our main aim is to keep similarity (or dissimilarity) information intact as much as possible after transforming $\mathbf{x}_i$ from $\Re^D$ to $\Re^d$, where $d \ll D$.

The characteristic of a good feature extractor is to maximize the between-class distances after the transformation. So the feature extraction criterion should be defined as

$$J = \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1} p_i p_j d(\omega_i, \omega_j); \tag{11}$$

where, $d(\omega_i, \omega_j)$ is the distance between two classes ω_i and ω_j, and p_i, p_j are a priori probabilities of classes ω_i and ω_j, respectively. J is called the maximum margin criterion (MMC). It is actually the summation of all the pairs of inter-class margins. If the distance between two classes is measured depending on the distance between the mean vectors, then it is not easy to separate the two classes that have large spread and overlap with each other. So the inter-class distance (or margin) should also consider the scatter of the classes. Thus, $d(\omega_i, \omega_j)$ is defined as follows:

$$d(\omega_i, \omega_j) = d(\mu_i, \mu_j) - s(\omega_i) - s(\omega_j); \tag{12}$$

where μ_i, and μ_j are the mean vectors of classes ω_i and ω_j, respectively, and $s(\omega_i)$ is a measure of the scatter of class ω_i. Using the overall variance $tr(S_i)$ to measure the scatter of data, S_i the covariance matrix of class ω_i, MMC (J) becomes:

$$J = \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} p_i p_j d(\mu_i, \mu_j) - \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} p_i p_j \{tr(S_i) + tr(S_j)\}. \tag{13}$$

By simplifying the above equation, the first part of the equation becomes $tr(S_b)$, and second part becomes $tr(S_w)$ [29]. So, the MMC is defined as

$$J = tr(S_b - S_w). \tag{14}$$

The between-class and within-class scatter matrices, denoted as S_b and S_w, are defined as

$$S_b = \sum_{i=1}^{C} p_i(\mu_i - \mu)(\mu_i - \mu)^T, \tag{15}$$

and

$$S_w = \sum_{i=1}^{C} p_i S_i; \tag{16}$$

where C is the number of classes; μ_i and p_i are the mean vector and a priori probability of class ω_i, respectively. The overall mean, μ, is defined as:

$$\mu = \sum_{i=1}^{C} p_i \mu_i. \tag{17}$$

S_i is the within-class scatter matrix of class ω_i and defined as

$$S_i = \sum_{j=1}^{n_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T; \tag{18}$$

where n_i is the number of patterns of class ω_i.

The feature extraction method transforms sample $\mathbf{x}$ into $\mathbf{y}$ where $\mathbf{x} \in \Re^D$ and $\mathbf{y} \in \Re^d$, respectively, and $d < D$ with a transformation matrix W, i.e.,

$$\mathbf{y} = W \cdot \mathbf{x}. \tag{19}$$

The main aim of transformation matrix, $W \in \Re^{D \times d}$, is to maximize

$$J(W) = tr(W^T(S_b - S_w)W). \tag{20}$$

$J(W)$ is maximized when W is composed the first d largest eigenvectors of $(S_b - S_w)$. In fact, the optimal projection axes $w_1, w_2, ..., w_d$ can be selected as the orthonormal eigenvectors corresponding to the first d largest eigenvalues $\lambda_1, \lambda_2, ..., \lambda_d$, i.e.,$(S_b - S_w)w_j = \lambda_j w_j$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

3.4 Partitioned Maximum Margin Criterion Based Supervised Feature Extraction Method

The two desired properties of any feature extraction technique are ordering constraint and discriminating transform [11]. The characteristics that the features of a data set are correlated should be exploited by the feature extraction technique. Any transformation should involve adjacent group of features to utilize ordering and locality properties of hyperspectral data. Further, the transformation should try to maximize discrimination among classes and thus uses class label information. Use of Fisher's discriminant or maximum margin criterion (MMC) is, therefore, more desirable for feature extraction [32]. Considering the above two properties and to avoid the singularity problem of Fisher's linear discriminant analysis (LDA), this method first partitions all the features of hyperspectral images into subgroups and then MMC [29] based transformation is applied over each subgroup of features. MMC uses difference of both between-class scatter and within-class scatter as a discriminant criterion. Since the inverse matrix does not need to be constructed, the small sample size (SSS) problem in traditional LDA is alleviated. Geometrically, MMC maximizes the (average) margin between classes. MMC has the advantages of effectiveness and simplicity [29, 31].

This method uses the ordering and locality properties of hyperspectral data by partitioning all the hyperspectral features into a number of groups of contiguous features. A transformation is then used on each group to maximize discrimination among classes by using maximum margin criterion. This feature extraction method [24, 26] is basically a two step process: partitioning of hyperspectral features and MMC based transformation.

Partitioning of hyperspectral features At the onset, the D number of features of a hyperspectral image is partitioned into a number of contiguous intervals with constant intensities (i.e., K subgroups). Highly correlated features should lie in a subgroup. Let $I_1, I_2, ..., I_k$, be the number of features in 1st, 2nd, ..., Kth group, respectively. The purpose is to obtain a set of K break points $\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_K\}$, which defines the contiguous intervals $I_k = [\zeta_k, \zeta_{k+1})$. The partition should follow the principle that no feature should be left outside. The total number of features, D, should follow

$$D = \sum_{k=1}^{K} I_k. \tag{21}$$

Let Γ be a correlation matrix of size $D \times D$. Each feature of a pixel is nothing but a reflectance of that pixel of a particular wavelength image (named as band image). So correlation among features, here, is measured as correlation among band images. Each element of Γ is γ_{ij}, where γ_{ij} represents the correlation between band images B_i and B_j.

Let the size of all the band images be $M \times N$. The correlation coefficient between B_i and B_j is computed as follows:

$$\gamma_{i,j} = \frac{\Sigma_{y=1}^{N_R} \Sigma_{z=1}^{N_C} [B_i(y,z) - \mu_i][B_j(y,z) - \mu_j]}{\sqrt{(\Sigma_{y=1}^{N_R} \Sigma_{z=1}^{N_C} [B_i(y,z) - \mu_i]^2)(\Sigma_{y=1}^{N_R} \Sigma_{z=1}^{N_C} [B_j(y,z) - \mu_j]^2)}} \tag{22}$$

where μ_i and μ_j are mean of band images B_i and B_j, respectively, and are defined as

$$\mu_i = \frac{1}{N_R \times N_C} \Sigma_{y=1}^{N_R} \Sigma_{z=1}^{N_C} B_i(y,z); \tag{23}$$

$$\mu_j = \frac{1}{N_R \times N_C} \Sigma_{y=1}^{N_R} \Sigma_{z=1}^{N_C} B_j(y,z). \tag{24}$$

$B_i(y,z)$ and $B_j(y,z)$ are the values of the pixel at position (x, y) of band images i and j, respectively.

$[B_i(y,z) - \mu_i]$ measures the difference between reflectance value of pixel (y, z) from the mean of that image.

It is observed that the correlations among neighboring band images are generally higher than that for band images further apart. Partitioning is performed based on the results obtained by, firstly, considering only correlations whose absolute value exceeds a given threshold, and simultaneously searching for edges in the "image" of the correlation matrix. Each value of the correlation matrix is compared with a threshold. If the magnitude is greater than the threshold value (i.e., denoted by Θ), then it is replaced by 1; otherwise by 0. The value of Θ is determined depending on the value of average correlation (μ_{corr}) and standard deviation (σ_{corr}) of correlation matrix Γ:

$$\Theta = \mu_{corr} + \sigma_{corr}. \tag{25}$$

where,

$$\mu_{corr} = \frac{1}{D^2} \Sigma_{i=1}^{D} \Sigma_{j=1}^{D} \gamma_{i,j}. \tag{26}$$

and

$$\sigma_{corr} = sqrt(\frac{1}{D^2} \Sigma_{i=1}^{D} \Sigma_{j=1}^{D} (\gamma_{i,j} - \mu_{corr})). \tag{27}$$

Result of the thresholded correlation matrix will be a binary image with the square blocks of white color in diagonal direction. These square blocks of white color are treated as a subgroup or partition of features.

Thereafter, MMC based transformation is conducted on each subgroup of features. The method is discussed below in detail.

Linear Band Extraction using Maximum Margin Criterion (MMC) [29] The MMC based feature extraction method transforms sample $\mathbf{x}$ into $\mathbf{y}$ where $\mathbf{x} \in \Re^D$ and $\mathbf{y} \in \Re^d$, respectively, and $d < D$ with a transformation matrix W, i.e.,

$$\mathbf{y} = W \cdot \mathbf{x}. \tag{28}$$

The main aim of transformation matrix, $W \in \Re^{D \times d}$, is to maximize

$$J(W) = tr(W^T(S_b - S_w)W). \tag{29}$$

$J(W)$ is maximized when W is composed the first d largest eigenvectors of $(S_b - S_w)$. In fact, the optimal projection axes $w_1, w_2, ..., w_d$ can be selected as the orthonormal eigenvectors corresponding to the first d largest eigenvalues $\lambda_1, \lambda_2, ..., \lambda_d$, i.e.,$(S_b - S_w)w_j = \lambda_j w_j$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

There are K blocks, and also has I_k number of features in each block, where $k = 1, 2, ..., K$. The MMC based transformation is applied over each block. Let the linear mapping matrix for block k be $W_k \in \Re^{I_k \times d_k}$, which transforms a data set with dimension I_k to that with dimension d_k. The desired number of features, d follows

$$d = \sum_{k=1}^{K} d_k. \tag{30}$$

The target is to form W_k in a way so that MMC becomes optimum, i.e.,

$$J(W_k) = tr(W_k^T(S_b^k - S_w^k)W_k). \tag{31}$$

Here S_b^k and S_w^k are between-class and within-class scatter matrices of block k, and are calculated by following Eqs. 15 and 16. $J(W_k)$ is maximized where W_k composes the first d_k largest eigenvectors of $(S_b^k - S_w^k)$. In fact, the optimal projection axes $w_1^k, w_2^k, ..., w_{d_k}^k$ can be selected as the orthonormal eigenvectors corresponding to the first d_k largest eigenvalues $\lambda_1^k, \lambda_2^k, ..., \lambda_{d_k}^k$, i.e., $(S_b^k - S_w^k)w_j^k = \lambda_j^k w_j^k$, where $\lambda_1^k \geq \lambda_2^k \geq \cdots \geq \lambda_{d_k}^k$.

To determine the value of d_k (i.e., how many eigenvectors will be selected from each block), the eigenvectors with their corresponding eigenvalues from each group are considered at first. Then the ratio of eigenvalues with overall eigenvalues of that block is calculated for each eigenvector, i.e., for each eigenvector w_i^k, corresponding ratio of eigenvalue, Λ_i^k is calculated by

$$\Lambda_i^k = \frac{\lambda_i^k}{\sum_{j=1}^{I_k} \lambda_j^k}. \tag{32}$$

The number of eigen vectors chosen from each subgroup depends on the largest value of Λ_i^k.

An outline of the partitioned MMC based supervised feature extraction method is given in Algorithm 1.

Algorithm 1 Partitioned MMC based supervised feature extraction algorithm

1. Partition all features of hyperspectral images into groups of contiguous features
   - Calculate the correlation matrix of all pairs of features.
   - Compare the correlation value with a threshold.
   - Depict the binary image of the threshold correlation matrix.
   - Square block of white color in diagonal direction in the binary image represents the group of features.
2. Transformation of each group of features using MMC
   - Perform MMC based transformation over each group of features separately.
   - Select first d_k eigenvectors $w_1^k, w_2^k, ..., w_{d_k}^k$ which have the largest ratio of eigenvalues i.e., $\Lambda_1^k \geq \Lambda_2^k \geq ... \geq \Lambda_{d_k}^k$ from the k^{th} group.

4 Experimental Evaluation

4.1 Description of Data Sets

To evaluate the effectiveness of these feature extraction methods, experiments were carried out on three hyperspectral remotely sensed images, namely, Indian Pine [33], KSC [34], and Botswana [34] images corresponding to the geographical areas of Indian Pine test site of Northwest Indiana, Kennedy Space Center of Florida and Okavango Delta of Botswana.

Indian Pine image [33] data was captured by AVIRIS within the spectral range from 400 to 2500 nm with spectral resolution of about 10 nm and has 220 spectral bands. The size of the image is 145 × 145 pixels and spatial resolution is 20 m. AVIRIS acquires KSC images of size 512 × 614 in 224 bands of 10 nm width with wavelengths ranging from 400 to 2500 nm. The Hyperion sensor on Earth Observing-1 (EO-1) acquired Botswana images of size 1476 × 256 at 30 m pixel resolution in 242 bands from the 400 nm-2500 nm portion of the spectrum in 10 nm windows. The details of the data sets are given in [35]. Class name and the number of labeled samples for each class are given in Tables 1, 2 and 3, respectively, for Indian, KSC and Botswana data. Corresponding band 11 images of these three data sets are shown in Figs. 1, 2 and 3.

4.2 Performance Measures

It is better to use more than one performance measures to show the effectiveness of any subset of features. As used in Chap. 2, overall classification accuracy (OA), kappa coefficient (κ), class separability (S) and entropy (E) are calculated for the selected set of features to assess the effectiveness of the proposed method. After performing classification operation, two performance measures, OA and κ, are computed. To measure the statistical significance of the selected subset of features, two statistical

Table 1 Indian Pine data: class names and the number of samples

Class no	Class name	No. of samples
C1	Corn	191
C2	Corn-min	688
C3	Corn-notill	1083
C4	Soybean-clean	541
C5	Soybean-min	2234
C6	Soybean-notill	860
C7	Wheat	211
C8	Alfalfa	51
C9	Oats	20
C10	Grass/Trees	605
C11	Grass/Pasture	351
C12	Grass/Pasture-mowed	17
C13	Woods	1293
C14	Hay-windrowed	477
C15	Bldg-Grass-Tree-Drives	380
C16	Stone-steel-towers	86

Table 2 KSC data: class names and the number of samples

Class no	Class name	No. of samples
C1	Scrub	761
C2	Willow swamp	243
C3	Cabbage palm hammock	256
C4	Cabbage palm/oak hammock	252
C5	Slash pine	161
C6	Oak/broadleaf hammock	229
C7	Hardwood swamp	105
C8	Graminoid marsh	431
C9	Spartina marsh	520
C10	Cattail marsh	404
C11	Salt marsh	419
C12	Mud flats	503
C13	Water	927

measures, S and E are calculated, where the first one is performed over labeled information of the data set and second one is assessed over unlabeled data. Measure of class separability [11] demonstrates the effectiveness of the selected features for classification of data. Our aim is to look for a feature space where the inter-class distance is large and at the same time the intra-class variance is as small as possible.

Table 3 Botswana data: class names and the number of samples

Class no	Class name	No. of samples
C1	Water	270
C2	Hippo Grass	101
C3	FloodPlain Grasses 1	251
C4	FloodPlain Grasses 2	215
C5	Reeds	269
C6	Riparian	269
C7	Firescar	259
C8	Island Interior	203
C9	Acacia Woodlands	314
C10	Acacia Shrublands	248
C11	Acacia Grasslands	305
C12	Short Mopane	181
C13	Mixed Mopane	268
C14	Exposed Soils	95

Fig. 1 Band 11 image of Indian Pine data

A lower value of the separability measure S ensures that the classes are well separated. Orderly or chaotic configuration of data can be determined by the measure of entropy of the data [36]. Entropy is low for stable configuration of patterns (data has well formed clusters), and is high for disordered configuration, i.e., data is uniformly distributed in the feature space. A detailed description of the above mentioned four performance measures are given in [35].

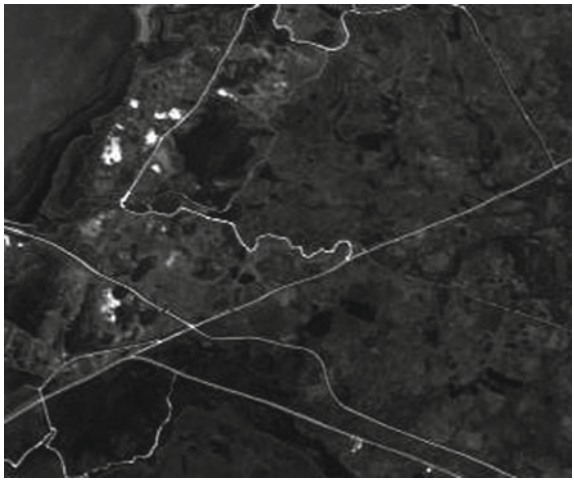


Fig. 2 Band 11 image of KSC data

Fig. 3 Band 11 image of Botswana data

4.3 Parameter Details

Experiments are conducted on three hyperspectral data sets, namely, Indian Pine, KSC and Botswana. As already mentioned, the partitioned MMC based approach follows two consecutive steps: first, the hyperspectral features are partitioned into groups and then MMC based transformation is used on each group separately.

For Indian Pine data, the correlation matrix in image form is shown in Fig. 4a. The threshold value (Θ) is varied depending on data set. For Indian Pine data, average correlation (μ_{corr}) and standard deviation (σ_{corr}) of correlation matrix Γ, are 0.71 and 0.19, respectively. So, Θ is set to 0.90 for Indian Pine data. The corresponding binary image of the threshold correlation matrix (of Indian Pine data) is shown in

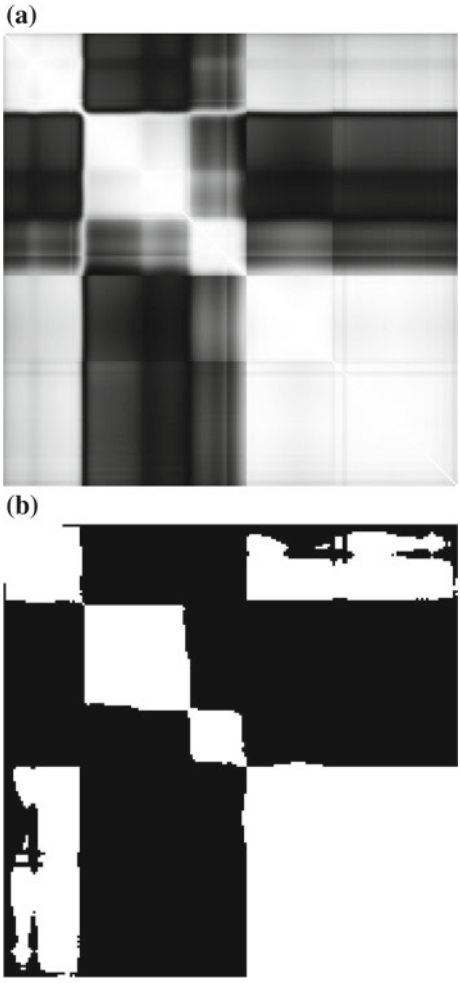


Fig. 4 Image of the correlation matrix of Indian Pine data: **a** gray scale image and **b** binary image; white color indicates higher correlation

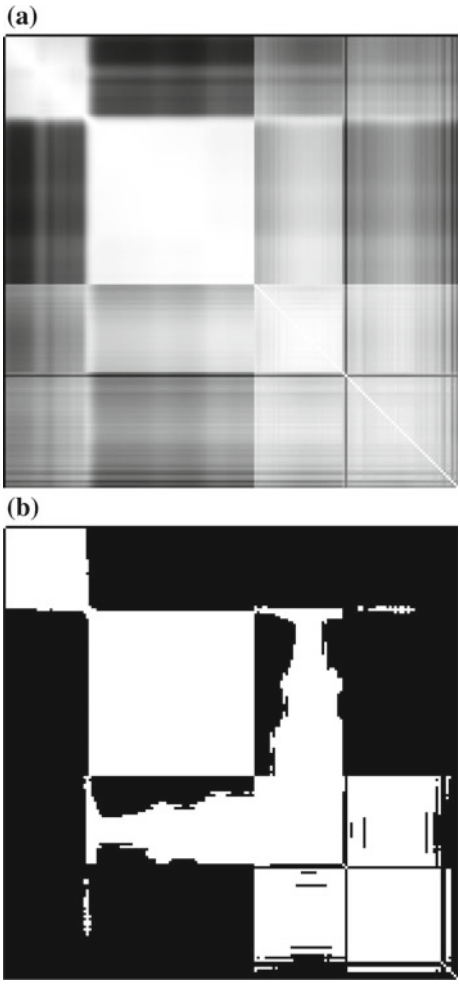


Fig. 5 Image of the correlation matrix of KSC data: **a** gray scale image and **b** binary image; white color indicates higher correlation

Fig. 4b. From the binary image, four blocks of contiguous features of white color are extracted. These four blocks of features for Indian Pine data are 1–33, 34–77, 78–100, and 101–185, respectively. For KSC and Botswana data sets, the correlation matrix in image form and the binary image of the threshold correlation matrix are shown in Figs. 5 and 6, respectively. The values of Θ are 0.85 ($\mu_{corr} = 0.73, \sigma_{corr} = 0.12$) and 0.94 ($\mu_{corr} = 0.88, \sigma_{corr} = 0.06$), respectively, for KSC and Botswana data sets. The four partitions of KSC data are 1–34, 35–98, 99–132, 133–176, whereas, the five partitions of Botswana data are 1–27, 28–47, 48–81, 82–112, 113–145.

Although the main focus of this chapter is on feature extraction of hyperspectral images in supervised manner, classification operation is performed over transformed features to assess the superiority of the proposed method. After completing the feature extraction, fuzzy k-NN based classification (in theory, any good classification

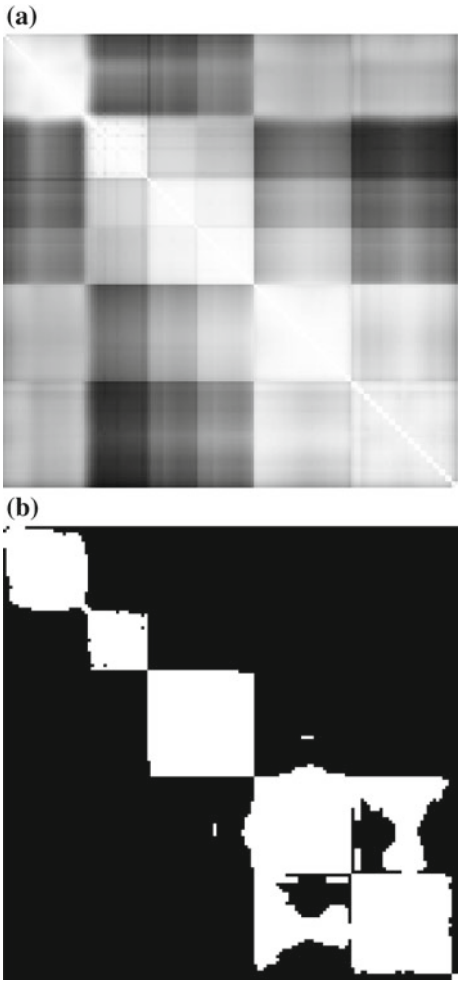


Fig. 6 Image of the correlation matrix of Botswana data: **a** gray scale image and **b** binary image; white color indicates higher correlation

algorithm can be used) operation is performed on the transformed features using 10-fold cross validation. There may be overlapping of information between neighboring pixels of the hyperspectral images. Fuzzy k-NN, rather than other classification techniques, is used to take care of the fuzziness present in the hyperspectral images.

As already mentioned, the performance of the supervised feature extraction techniques, namely, prototype space feature extraction (PSFE) [20], modified Fisher's linear discriminant analysis (MFLDA) [28] and maximum margin criterion based feature extraction (MMC) [29], partitioned MMC based method have been compared. The desired number of transformed features is not known a priori because it varies with data set. In the present investigation, experiments are carried out for different number of features ranging from 4 to 30 with a step size of 2. Overall classification accuracy (OA), kappa coefficient (κ), class separability (S) and entropy

(E) are calculated for the transformed set of features to assess the effectiveness of the feature extraction methods.

4.4 Analysis of Results

The obtained OA and κ for Indian Pine data after applying fuzzy k-NN classifier over the transformed set of features by PSFE, MFLDA, MMC, partitioned MMC based algorithms are given in Table 4. The value of OA and κ are tabulated for PSFE, MMC and the partitioned MMC based method for upto 30 transformed features, whereas for MFLDA, it is only upto 14. As for Indian Pine data, the number of classes present in the data set is 16 and the number of transformed features should be $(16 - 1) = 15$ for MFLDA. It is noticed from Table 4 that the other three methods except PSFE reach the highest value much quickly and then the values of OA and κ become more or less stabilized. From Table 4, it is noticed that the partitioned MMC based method achieves better results in terms of overall classification accuracy and kappa coefficient for different number of extracted features.

It is seen that approximate peak performance is obtained when the number of transformed features is 14 in case of PSFE, whereas for other methods (MFLDA, MMC) it is in between 10 to 12, and 18 for the partitioned MMC based method. The reason behind this finding is that the PSFE method performs clustering over features

Table 4 Overall accuracy and kappa coefficient of PSFE, MFLDA, MMC and partitioned MMC based methods for different number of extracted features for Indian Pine data

No. of bands	PSFE		MFLDA		MMC		Partitioned MMC	
	OA (%)	κ	OA (%)	κ	OA (%)	κ	OA (%)	κ
4	77.10	0.7398	79.81	0.7697	**80.75**	**0.7795**	75.62	0.7199
6	79.74	0.7689	83.20	0.8071	**86.49**	**0.8452**	83.28	0.8079
8	81.25	0.7856	85.89	0.8366	**87.07**	**0.8517**	86.54	0.8452
10	83.10	0.8060	86.27	0.8411	87.08	0.8529	**87.09**	**0.8531**
12	84.92	0.8260	86.11	0.8394	87.38	0.8554	**88.27**	**0.8651**
14	85.12	0.8282	85.91	0.8369	87.18	0.8529	**88.06**	**0.8627**
16	84.74	0.8239			86.96	0.8504	**88.10**	**0.8612**
18	84.64	0.8228			87.09	0.8520	**88.40**	**0.8646**
20	84.98	0.8267			87.20	0.8533	**88.33**	**0.8639**
22	85.03	0.8273			87.31	0.8545	**88.27**	**0.8631**
24	84.61	0.8225			87.14	0.8526	**88.03**	**0.8604**
26	84.29	0.8189			87.24	0.8538	**88.27**	**0.8632**
28	84.99	0.8269			87.08	0.8518	**88.18**	**0.8621**
30	84.76	0.8241			87.31	0.8546	**88.28**	**0.8633**

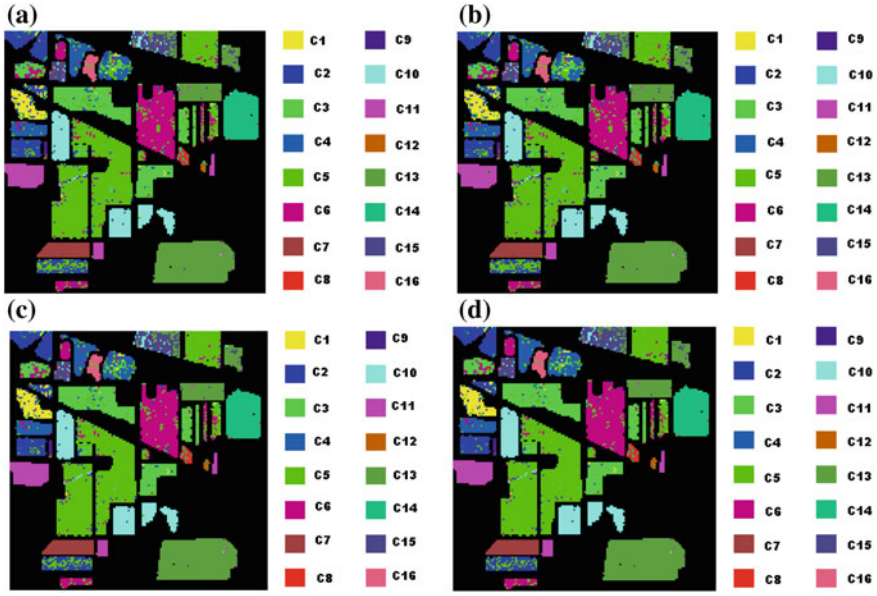


Fig. 7 Classified images of Indian Pine data with extracted feature set using **a** PSFE, **b** MFLDA, **c** MMC, and **d** the partitioned MMC based methods

and selects representative features as well as independent features for transformed set of features. On the contrary, other three methods transform the original set of features into a new set of features where the main aim is to maximize discrimination among classes by using class label information. From the graph, it is observed that when the number of extracted features is small (i.e., less than 10), MMC based method gives slightly better performance than the partitioned MMc based one. Since MMC based method transforms the complete data set at a time, whereas, the partitioned MMC based method partitions the complete data set into blocks and then performs transformation block wise. So the first few transformed features using MMC based method takes the discriminating effect of all the features, whereas, the partitioned MMC based method may consider the effect of some block of features. But, when the number of transformed features are not small, the transformed features using the partitioned MMC based method also include the effect of all the features of the data set and hence better performance is obtained than all the methods compared with.

Figure 7a–d, respectively, show the pictorial representation of the classified images with the best subset of features extracted using PSFE, MFLDA, MMC and the partitioned MMC based techniques. It is clearly observed that the classified Indian Pine image with transformed feature set using the partitioned MMC based method has very less misclassified pixels compared to other methods. Table 5 contains the measurements of class separability and entropy of Indian Pine data with transformed features using four methods. From this table, it is noticed that the partitioned MMC

Table 5 Class separability and entropy of PSFE, MFLDA, MMC based methods and partitioned MMC based methods for different number of extracted features for Indian Pine data

No. of bands	PSFE		MFLDA		MMC		Partitioned MMC	
	S	*E*	*S*	*E*	*S*	*E*	*S*	*E*
4	0.3153	0.7383	0.2969	0.6896	**0.2905**	**0.6726**	0.3254	0.7632
6	0.2974	0.6908	0.2739	0.6285	**0.2515**	**0.5691**	0.2733	0.6585
8	0.2871	0.6636	0.2556	0.5809	**0.2478**	**0.5589**	0.2511	0.5611
10	0.2745	0.6303	0.2530	0.5733	0.2475	0.5587	**0.2474**	**0.5510**
12	0.2621	0.5976	0.2541	0.5762	0.2454	0.5533	**0.2394**	**0.5291**
14	0.2608	0.5940	0.2554	0.5798	0.2468	0.5569	**0.2408**	**0.5330**
16	0.2634	0.6008			0.2483	0.5609	**0.2405**	**0.5323**
18	0.2641	0.6026			0.2474	0.5585	**0.2385**	**0.5267**
20	0.2638	0.5965			0.2467	0.5565	**0.2390**	**0.5280**
22	0.2615	0.5956			0.2459	0.5546	**0.2394**	**0.5291**
24	0.2643	0.6032			0.2471	0.5576	**0.2410**	**0.5336**
26	0.2664	0.6089			0.2463	0.5558	**0.2394**	**0.5291**
28	0.2617	0.5963			0.2475	0.5587	**0.2400**	**0.5308**
30	0.2633	0.6005			0.2459	0.5546	**0.2393**	**0.5290**

based method gives less value of class separability (S) and entropy (E) with respect to others. It shows that the partitioned MMC based method is able to transform better subset of features which, in turn, gives well separated classes as well as stable configuration of patterns compared to other methods.

Overall accuracy (OA) and kappa coefficient (κ) values for KSC and Botswana data sets are put in Tables 6 and 7, respectively. From the table, it is observed that the partitioned MMC based method is producing better results than the other three methods for both the data sets. It is also observed that discriminant analysis based transformation methods (MFLDA, MMC and the partitioned MMC based methods) are found to be better than clustering based methods (PSFE). The partitioned MMC based method gives better results than others, because the ordering and locality property of hyperspectral images are considered with discriminant analysis.

Class separability and entropy values are also calculated for both KSC and Botswana data sets. Results of these data sets provide similar findings with the results obtained using Indian Pine data. Table 8 incorporates the optimum values (for all the three data sets) in terms of OA, κ, S and E. The best results are marked in bold. This table also confirms the fact that the partitioned MMC based supervised feature extraction algorithm gives better transformed set of features for classification than the other methods used in our experiment.

Table 6 Overall accuracy and kappa coefficient of PSFE, MFLDA, MMC and the partitioned MMC based methods for different number of extracted features for KSC data

No. of bands	PSFE		MFLDA		MMC		Partitioned MMC	
	OA (%)	κ	OA (%)	κ	OA (%)	κ	OA (%)	κ
4	83.38	0.8132	82.92	0.8080	**89.51**	**0.8815**	85.04	0.8310
6	84.91	0.8301	86.22	0.8443	**90.64**	**0.8941**	90.03	0.8866
8	86.67	0.8494	87.44	0.8577	**90.72**	**0.8950**	90.30	0.8897
10	87.11	0.8542	88.71	0.8716	90.93	0.8973	**91.56**	**0.9038**
12	87.92	0.8631	88.18	0.8658	90.82	0.8960	**91.81**	**0.9066**
14	88.27	0.8669			90.87	0.8966	**91.89**	**0.9074**
16	88.21	0.8662			90.89	0.8968	**92.20**	**0.9108**
18	88.09	0.8649			90.83	0.8962	**92.27**	**0.9117**
20	87.91	0.8520			90.80	0.8958	**92.01**	**0.9087**
22	87.43	0.8577			90.83	0.8963	**92.12**	**0.9100**
24	87.65	0.8601			90.74	0.8951	**92.06**	**0.9093**
26	87.30	0.8562			90.83	0.8963	**92.10**	**0.9098**
28	87.12	0.8543			90.92	0.9005	**92.27**	**0.9107**
30	87.17	0.8549			90.78	0.8956	**92.04**	**0.9092**

Table 7 Overall accuracy and kappa coefficient of PSFE, MFLDA, MMC based methods and the partitioned MMC based method for different number of extracted features for Botswana data

No. of bands	PSFE		MFLDA		MMC		Partitioned MMC	
	OA (%)	κ	OA (%)	κ	OA (%)	κ	OA (%)	κ
4	82.11	0.8049	83.24	0.8168	**91.59**	**0.9081**	88.02	0.8689
6	85.27	0.8387	84.37	0.8296	**92.27**	**0.9154**	92.18	0.9141
8	86.10	0.8477	87.02	0.8373	**92.16**	**0.9151**	92.14	0.9144
10	86.27	0.8495	88.86	0.8772	92.89	0.9221	**92.99**	**0.9228**
12	87.76	0.8654	88.39	0.8721	92.89	0.9221	**93.29**	**0.9261**
14	88.13	0.8695	88.24	0.8694	93.19	0.9255	**93.28**	**0.9325**
16	88.61	0.8746			92.73	0.9204	**93.39**	**0.9271**
18	87.94	0.8675			92.82	0.9215	**93.94**	**0.9332**
20	87.78	0.8656			92.70	0.9201	**93.73**	**0.9308**
22	88.11	0.8693			92.82	0.9214	**93.54**	**0.9288**
24	87.83	0.8662			92.70	0.9202	**93.76**	**0.9312**
26	87.20	0.8594			92.82	0.9214	**93.85**	**0.9321**
28	87.91	0.8671			92.98	0.9232	**93.66**	**0.9302**
30	87.48	0.8624			92.55	0.9185	**93.88**	**0.9325**

Table 8 Comparison of performances of feature extraction methods for hyperspectral data sets

Data set used	Method	Extracted feature no.	Evaluation Criterion			
			E	S	OA	κ
Indian Pine D = 185	PSFE	14	0.5940	0.2608	85.12	0.8282
	MFLDA	10	0.5733	0.2530	86.27	0.8411
	MMC	12	0.5533	0.2454	87.38	0.8554
	Partitioned MMC	18	**0.5267**	**0.2385**	**88.40**	**0.8646**
KSC D = 176	PSFE	14	0.5680	0.1351	88.27	0.8669
	MFLDA	10	0.5646	0.1332	88.71	0.8716
	MMC	10	**0.5475**	0.1237	90.93	0.8973
	Partitioned MMC	18	**0.5372**	**0.1179**	**92.27**	**0.9117**
Botswana D = 145	PSFE	16	0.4800	0.1024	88.61	0.8746
	MFLDA	10	0.4770	0.1011	88.86	0.8772
	MMC	10	0.4241	0.0777	93.19	0.9255
	Partitioned MMC	18	**0.4154**	**0.0737**	**93.94**	**0.9332**

5 Conclusions

Few supervised techniques for feature extraction of hyperspectral images, namely, modified Fisher's linear discriminant analysis (MFLDA), prototype space feature extraction (PSFE), maximum margin criterion based feature extraction (MMC), and partitioned MMC based feature extraction methods, have been presented in this chapter. To measure the effectiveness of the proposed algorithm, four evaluation measures (namely, overall accuracy, kappa coefficient, class separability and entropy value) have been used. Results of the partitioned MMC based technique have a significant improvement and a more consistent and steady behavior for the same hyperspectral image data sets (Indian Pine, KSC and Botswana data) with respect to the other methods (PSFE, MFLDA and MMC based methods). This improvement may be due to the fact that the strategy considers correlation among neighboring features, as well as, increases discriminating capability among classes by transformation of original set of features into a new space.

Acquisition of labeled data for classification problem often requires human interactions or physical experiments. It is always a hard, time consuming and very expensive process. Sometimes, availability of fully labeled training set becomes infeasible. It is found that the unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. In this sit-

uation, semi-supervised technique may be exploited, by considering a small amount of available labeled data with a large amount of unlabeled data. Developing feature extraction methods under semi-supervised framework will be a future direction of research.

References

1. Lillesand, T.M., Kiefer, R.W., Chipman, J.W.: Remote Sensing and Image Interpretation, 6th edn. Wiley, New Delhi, India (2014)
2. Campbell, J.B., Wynne, R.H.: Introduction to Remote Sensing, 5th edn. Guilford Press, New York, USA (2011)
3. Richards, J.A., Jia, X.: Remote Sensing Digital Image Analysis: An Introduction, 1st edn. Springer, New York, USA (1999)
4. Varshney, P.K., Arora, M.K.: Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data, 2nd edn. Springer, Berlin, Germany (2004)
5. Chang, C.-I.: Hyperspectral Imaging: Techniques for Spectral Detection and Classification, 1st edn. Kluwer Academic/Plenum Publisher, New York, USA (2003)
6. Landgrebe, D.: Hyperspectral image data analysis. IEEE Signal Process. Mag. 17–28 (2002)
7. Eismann, M.T.: Hyperspectral Remote Sensing, 1st edn. SPIE Press, Washigton, USA (2012)
8. Manolakis, D., Marden, D., Shaw, G.A.: Hyperspectral image processing for automatic target detection applications. Linc. Lab. J. **14**(1), 79–116 (2003)
9. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 4–37 (2000)
10. Bishop, C.M.: Neural Networks for Pattern Recognition, 1st edn. Oxford University Press, New Delhi, India (1995)
11. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach, 1st edn. Prentice-Hall International, New Delhi, India (1982)
12. Ghosh, A., Datta, A., Ghosh, S.: Self-adaptive differential evolution for feature selection in hyperspectral image data. Appl. Soft Comput. **13**(4), 1969–1977 (2013)
13. Jia, X., Kuo, B.-C., Crawford, M.M.: Feature mining for hyperspectral image classification. Proc. IEEE **101**(3), 676–697 (2013)
14. Datta, A., Ghosh, S., Ghosh, A.: Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination. Int. J. Remote Sens. **35**(2), 554–577 (2014)
15. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Acacdemic Press, San Diego, CA, USA (1990)
16. Datta, A., Ghosh, S., Ghosh, A.: Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **8**(6), 2814–2823 (2015)
17. Jia, S., Ji, Z., Shen, L.: Unsupervised band selection for hyperspectral imagery classification without manual band removal. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **5**(2), 531–543 (2012)
18. Datta, A., Ghosh, S., Ghosh, A.: Wrapper based feature selection in hyperspectral image data using self-adaptive differential evolution. In: Proceedings of the International Conference on Image Information Processing (ICIIP), pp. 1–6 (2011)
19. Datta, A., Ghosh, S., Ghosh, A.: Clustering based band selection for hyperspectral images. In: Proceedings of the International Conference on Communications, Devices and Intelligent Systems (CoDIS), pp. 101–104 (2012)
20. Mojaradi, B., Abrishami-Moghaddam, H., Zoej, M.J.V., Duin, R.P.W.: Dimensionality reduction of hyperspectral data via spectral feature extraction. IEEE Trans. Geosci. Remote Sens. **47**(7), 2091–2105 (2009)

21. Datta, A., Ghosh, S., Ghosh, A.: Band elimination of hyperspectral imagery using correlation of partitioned band image. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 412–417 (2013)
22. Jain, A., Zongker, D.: Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. **19**, 153–189 (1997)
23. Jia, X., Richards, J.A.: Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. IEEE Trans. Geosci. Remote Sens. **37**, 538–542 (1999)
24. Datta, A., Ghosh, S., Ghosh, A.: Supervised band extraction of hyperspectral images using partitioned maximum margin criterion. IEEE Geosci. Remote Sens. Lett. **14**(1), 82–86 (2017)
25. Fauvel, M., Chanussot, J., Benediktsson, J.A.: Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. J. Adv. Signal Process. **2009**, 1–14 (2009)
26. Datta, A., Ghosh, S., Ghosh, A.: Maximum margin criterion based band extraction of hyperspectral imagery. In: Proceedings of the Fourth International Conference on Emerging Applications of Information Technology (EAIT), pp. 300–304 (2014)
27. Kuo, B.-C., Landgrebe, D.A.: Nonparametric weighted feature extraction for classification. IEEE Trans. Geosci. Remote Sens. **42**, 1096–1105 (2004)
28. Du, Q.: Modified Fisher's linear discriminant analysis for hyperspectral imagery. IEEE Geosci. Remote Sens. Lett. **4**(4), 503–507 (2007)
29. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. IEEE Trans. Neural Netw. **17**(1), 157–165 (2006)
30. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 3rd edn. Academic Press, New York, USA (2006)
31. Yang, W., Wang, J., Ren, M., Yang, J., Liu, L.Z.G.: Feature extraction based on Laplacian bidirectional maximum margin criterion. Pattern Recogn. **42**(11), 2327–2334 (2009)
32. Kumar, S., Ghosh, J., Crawford, M.M.: Best-bases feature extraction algorithms for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **39**(7), 1368–1379 (2001)
33. Jimenez, L.O., Landgrebe, D.A.: Hyperspectral data analysis and supervised feature reduction via projection pursuit. IEEE Trans. Geosci. Remote Sens. **37**, 2653–2667 (1999)
34. Ham, J., Chen, Y., Crawford, M.M., Ghosh, J.: Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. **43**(3), 492–501 (2005)
35. Datta, A., Ghosh, S., Ghosh, A.: PCA, Kernel PCA and dimensionality reduction in hyperspectral images. In: Advances in Principal Component Analysis: Research and Developement, pp. 19–46. Springer Nature, Singapore (2018)
36. Yao, J., Dash, M., Tan, S.T., Liu, H.: Entropy-based fuzzy clustering and fuzzy modeling. Fuzzy Sets Syst. **113**, 381–388 (2000)