# Secure Data Mining in Cloud using Homomorphic Encryption

*Thesis submitted in partial fulfillment of the requirements for the award of degree of*

**Master of Engineering**
in
**Computer Science and Engineering**


*Submitted By*
**Deepti Mittal**
**(Roll No. 801232007)**

Under the supervision of:

**Mr. Ashish Aggarwal**
Assistant Professor

**Dr. Damandeep Kaur**
Assistant Professor



COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004


**June 2014**

# CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Secure Data Mining in Cloud using Homomorphic Encryption"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Mr. Ashish Aggarwal* and *Dr. Damandeep Kaur* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature:

(Deepti Mittal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Mr. Ashish Aggarwal
Assistant Professor

Dr. Damandeep Kaur
Assistant Professor

**Countersigned by**

**(Dr. Deepak Garg)**
Head
Computer Science and Engineering Department
Thapar University
Patiala

**(Dr. S. K. Mohapatra)**
Dean (Academic Affairs)
Thapar University
Patiala

# ACKNOWLEDGEMENT

No volume of words is enough to express my gratitude towards my guide **Mr. Ashish Aggarwal** and **Dr. Damandeeep Kaur** Computer Science & Engineering Department, Thapar University, Patiala, who have been very concerned and have aided for all the materials essentials for the preparation of this thesis report. They have helped me to explore this vast topic in an organized manner and provided me all the ideas on how to work towards a research-oriented venture.

I am also thankful to **Dr. S. K. Mohapatra**, Dean of Academic Affairs, **Dr. Deepak Garg**, Head of Computer Science & Engineering Department and **Mr. Ashutosh Mishra**, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work.

I would also like to thank the staff members and my colleagues who were always there at the need of hour and provided with all the help and facilities, which I required, for the completion of my thesis work.

Most importantly, I would like to thank my parents and the almighty for showing me the right direction out of the blue, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

(Deepti Mittal)
801232007

# ABSTRACT

With the advancement in technology, industry, e-commerce and research a large amount of complex and pervasive digital data is being generated which is increasing at an exponential rate and often termed as big data. Traditional *Data Storage* systems are not able to handle *Big Data* and also analyzing the *Big Data* becomes a challenge and thus it cannot be handled by traditional analytic tools. *Cloud Computing* can resolve the problem of handling, storage and analyzing the *Big Data* as it distributes the big data within the cloudlets. No doubt, *Cloud Computing* is the best answer available to the problem of *Big Data* storage and its analyses but having said that, there is always a potential risk to the security of *Big Data* storage in *Cloud Computing*, which needs to be addressed. Data Privacy is one of the major issues while storing the *Big Data* in a Cloud environment. Data Mining based attacks, a major threat to the data, allows an adversary or an unauthorized user to infer valuable and sensitive information by analyzing the results generated from computation performed on the raw data. This thesis proposes a secure k-means data mining approach assuming the data to be distributed among different hosts preserving the privacy of the data. The approach is able to maintain the correctness and validity of the existing k-means to generate the final results even in the distributed environment.

# TABLE OF CONTENTS

# LIST OF FIGURES

# Chapter 1

# Introduction

This chapter introduces cloud computing, data mining and other concepts which include various issues, need, threats discussed in detail in later chapters of this Thesis.

The present era of internet is generating a massive amount of digital data and information which is highly complex and pervasive too in nature. This rapid or exponential increase in amount of data can be credited to industry, commerce and research. The field of science, as biological medicine, earth science field, astronomy, all are producing very large data sets daily whether by observation or by simulation. But the traditional database systems are unable to handle such large data sets as they were designed to according to enterprise infrastructure and thus do not meet the requirements of scalability, fault tolerance etc. This problem is solved by cloud computing. Also, to extract useful information and some hidden or meaningful pattern in this large dynamic and distributed data, cloud computing is combined with data mining

## 1.1 Cloud Computing

Cloud computing refers to the web-based computing, providing users or devices with shared pool of resources, information or software on demand and pay per-use basis. It frees a user from the concerns about the expertise in the technological infrastructure of the service. It allows end user and small companies to make use of various computational resources like storage, software and processing capabilities provided by other companies. The cloud services can be divided into three categories: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS)[2]. Amazon, Microsoft, Google are some of the major cloud service providers. Google App Engine (GAE) is a type of PaaS provided by Google which allows web application hosting. Windows Azure, SQL Azure are some of the services offered by Microsoft providing processing and storage capabilities for large datasets [3]. Amazon Web Services (AWS) including Simple Storage Service (S3), SQS, EC2 are cloud services provided by the Amazon [1].

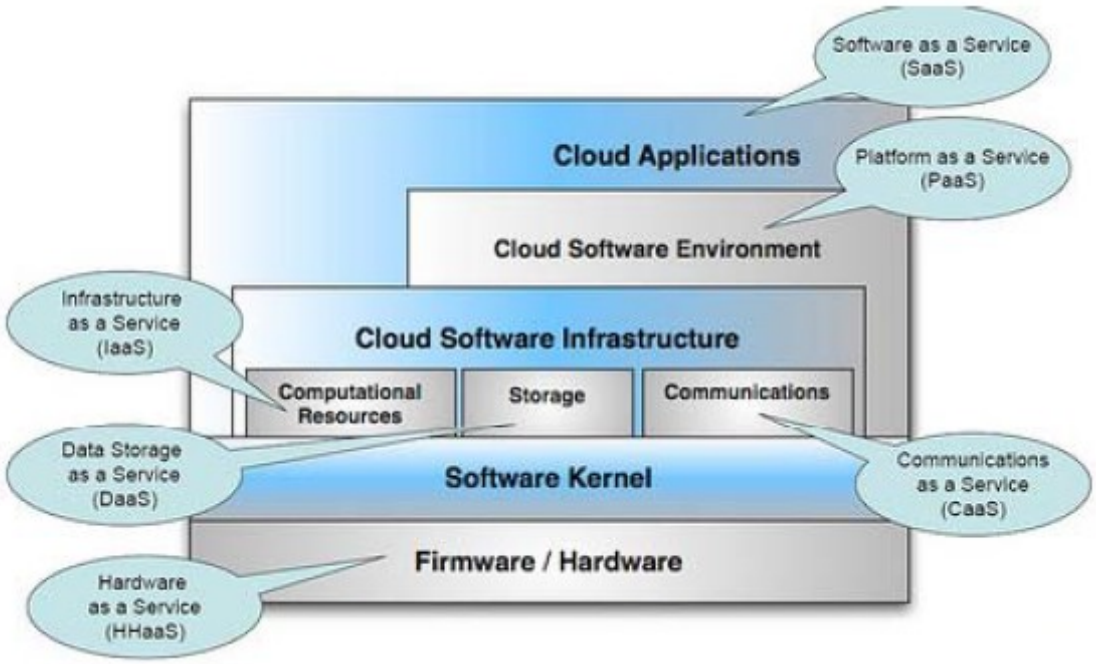


Figure 1.1: Various Cloud Services [48]

According to NIST Cloud Computing is composed of [48]:

- 5 Essential Characteristics
- 3 Service Models
- 4 Deployment Models

### 1.1.1 5-Essential Characteristics

1. *On-Demand Self-Service* – A user can get resources like server time or storage according to his/her requirement and without any other human intervention.
2. *Broad Network Access* – Computing and storage capabilities can be accessed from anywhere and by any means over the broad and easily accessible network.
3. *Resource Pooling* – Large amount of multitenant and location independent physical and virtual resource pool which can be dynamically assigned.
4. *Rapid Elasticity* – The provisioning of the capabilities is so elastic and agile that it gives a sense of infinite capacity. Scaling up and Scaling down of capabilities is very fast.

5. *Measured Services* – Customer has to pay for the services he/she used and the quantity or the time quantum for which the services were used.

### 1.1.2 3-Service Models

1. *Software as a Service (SaaS)* – The customer uses the application service provided by the cloud to perform the computations and is not bothered about underlying hardware or infrastructure like storage, Servers, O.S, processing peeds, bandwidth requirements etc. . The consumer does not have any control over the Application, storage, servers, O.S, networking capabilities etc.
2. *Platform as a Service (PaaS)* – In this model cloud provider offers the customer with the platform on which the user can deploy his/her application and use the underlying infrastructure to run the application. The consumer has control over only the application but not on any underlying hardware, Software, O.S, networking capabilities etc.
3. *Infrastructure as a Service (IaaS)* – This provides consumer with the Storage, Server, O.S, firewalls and other hardware capabilities but not with the underlying cloud infrastructure.

### 1.1.3 4-Deployement Models

1. *Private Cloud* – operated and owned solely by an organization. It can be managed by the organization or by a third party and can reside on the premises of organization or off-premises.
2. *Community Cloud* – this type of cloud belongs to a group of organizations or community with common goal or mission. Managed by community or by a third party.
3. *Public Cloud* – is generally open to the general public or owned by a very large organization and is owned by a cloud service provider company.
4. *Hybrid Cloud* – is the combination of any two types of model (Private, Public, Community) that although unique or independent entities are bounded by standardized technology so as to enable data and application portability.

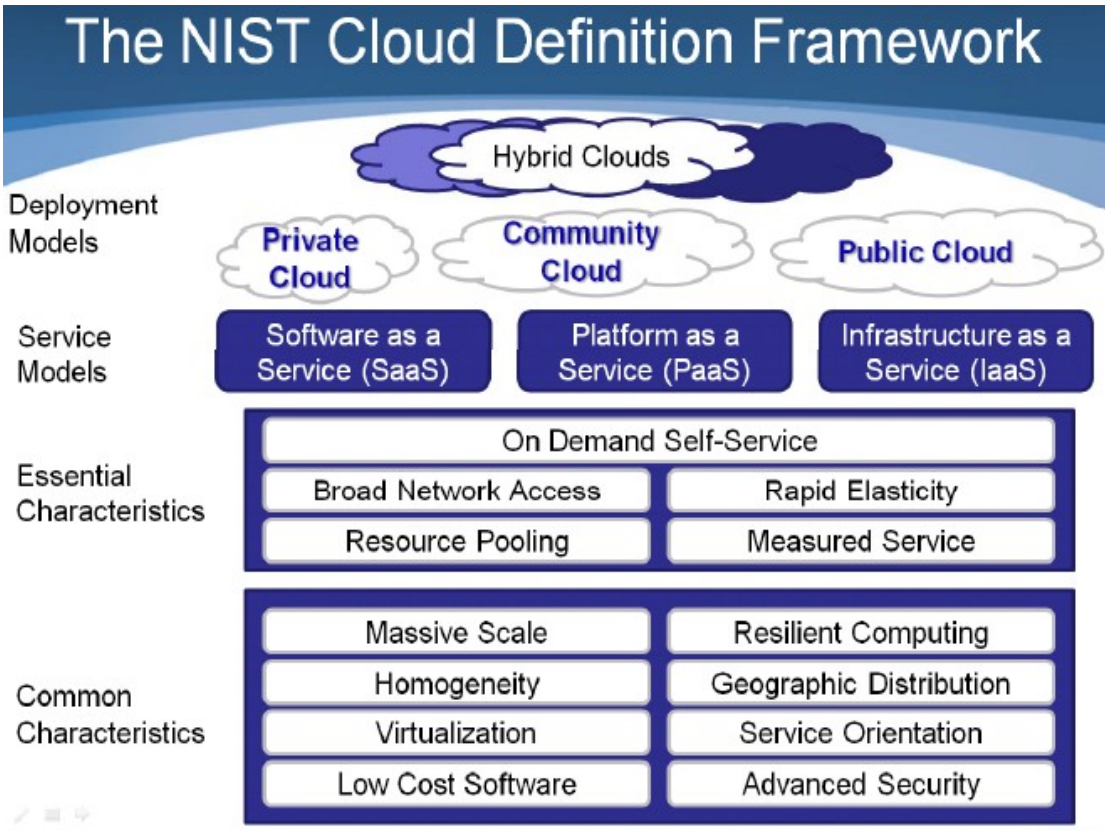


Figure 1.2: Cloud Framework [48]

Thus convenience, on demand measured access, shared easily configurable computational resources; rapid provisioning, location independence and self-service are some of the major characteristics of a cloud environment [2].

## 1.2 Cloud Issues [3]:

Despite all the above powerful functionalities and certainty provided by the cloud computing, the migration to cloud computing from the In-house traditional IT-Infrastructure is just the beginning. A lot of perspective customers and users still lack interest for cloud services, reason being the cloud issues which includes:

**Availability:** This is a major concern to all the organizations using cloud services that whether the utility computing service will be able to provide high or adequate availability of the service or resources agreed for. And for cloud provider this is a concern as they

have to prevent “single point of failures” to maintain high availability, for which, the most appropriate step is to distribute the data among multiple cloud vendors or providers.

**Data Lock-In/Interoperability**: Lack of the standardization of data storage and data processing techniques makes it difficult for a user to move from one platform to other thus resulting in a Lock-in which though attractive to a cloud provider is a bottleneck for the customer. Thus, there is a need to standardize the APIs so as a service or storage can be deployed across multiple vendors.

**Data confidentiality/Security:** One of the major issue for all the parties involved in a cloud service whether a provider, customer or the third-party, as a lot of sensitive and confidential data resides in cloud which is a source of immense valuable information. Data in cloud needs to be secured from both the outside as well as inside attackers.

**Data Communication Bottlenecks:** Modern applications are more data intensive as compared to the earlier ones, for which, data needs to be transferred across to boundaries of cloud quickly. So, to transfer such large amounts of data at a high speed cloud vendors have to take in considerations the traffic as well as the communication overhead in terms of cost.

**Performance:** I/O performance rather than memory performance are more serious issues in a cloud computing environment if they want to match the traditional computing practices. Also the appropriate scheduling policies in a virtualized environment concern the performance to a great extent.

**Scalability:** Scalability can be discussed in terms of Storage or in terms of Service. Scalable storage points to the cloud definition which presents cloud as an infinite capacity store and Pay-as-you-go rule. Thus, to provide scalable storage cloud should be able to scale up/down to the storage demands of the customer. Talking in terms of services the cloud should be able to automatically scale up/down according to the load or the requirement of customer whether of processing capability or of resources.

**Debugging the Large Distributed Systems:** Removing the errors in such large distributed system is a very big challenge for a cloud provider.

**Pricing and Cost Models:** Cloud computing is largely publicized as a Pay-Per-Use model i.e. to pay for the services used and the time for which the services are used. Thus, an appropriate or optimized pricing model is required which satisfies the customer as well as the provider so as to provide them with optimal cost of their resource utilization.

**Cross-Border Regulations and Taxation Policies:** This is a serious issue for the cloud providers if cloud is to be a borderless service as the legal and taxation policies differ according to the geographical locations.

**Risk Management:** As the technology changes very fast the cloud providers needs to conform to the new standards as well business approach if they want to remain in the business.

## 1.3 Cloud Security

Of all the issues that bug the modern cloud computing the issue of security or data privacy is one of the biggest concerns of the cloud providers as well as the client. Maintaining the client privacy is not only important to maintain the confidentiality of the sensitive and valuable data of user but also to maintain the reputation among customers for the cloud provider. Responsibility of maintaining cloud security is divided amongst all the parties- the client, the cloud provider as well as any third-party involved. The data stored in the cloud needs to be secured from inside as well as outside the cloud attacker.

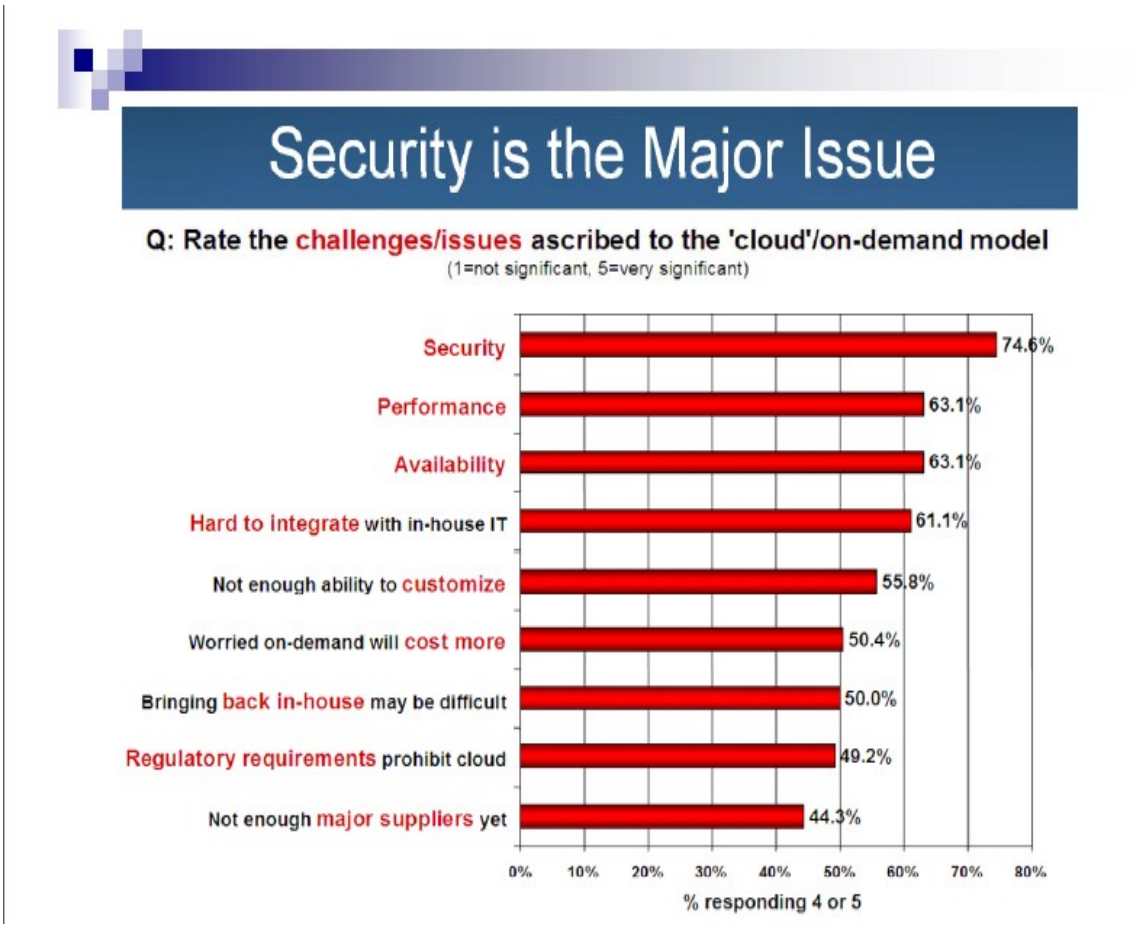


Figure 1.3: Major Issue in cloud [48]

On the inside a client may face attacks from the other user like theft or Denial-of Service attacks. The inside attacks can be prevented by using authentication at the time of access of the data. *Virtualization* is one other technique that can be employed to prevent the users from one another but this still cannot solve the problem completely as all the resources cannot be virtualized and also the virtualized environments are not completely error-free. Any error in the network virtualization environment may lead to incorrect transfers or leakage of sensitive information in the communication process or even the irrecoverable loss of valuable data. On the inside a cloud provider may also pose as a threat to users data if he/she misuses the data as the provider is present at the bottom layer of the security and may know how to circumvent the applied security techniques.

To prevent the user's data from outside attackers is another major challenge faced by the users and providers nowadays. Authenticating the access cannot be the only solution to prevent the attacks as an adversary can gain an unauthorized access to the system. Also eavesdropping by an attacker when the vulnerable data or information is in communication may also allow the outsider to collect the valuable information which can be misused thus violating the privacy and individuality of the client. Various encryption techniques have been proposed in the literature to secure the data even if the authorization fails and the attacker get hold of the data.

## 1.4 Data Mining

Data Mining is the processing of the massive amount of raw data to convert it into useful, valuable information which can be used for decision making and strategy planning by identifying hidden relationships, behavioral patterns or any other statistical data. Data Mining is one step of the Knowledge Discovery (KDD).

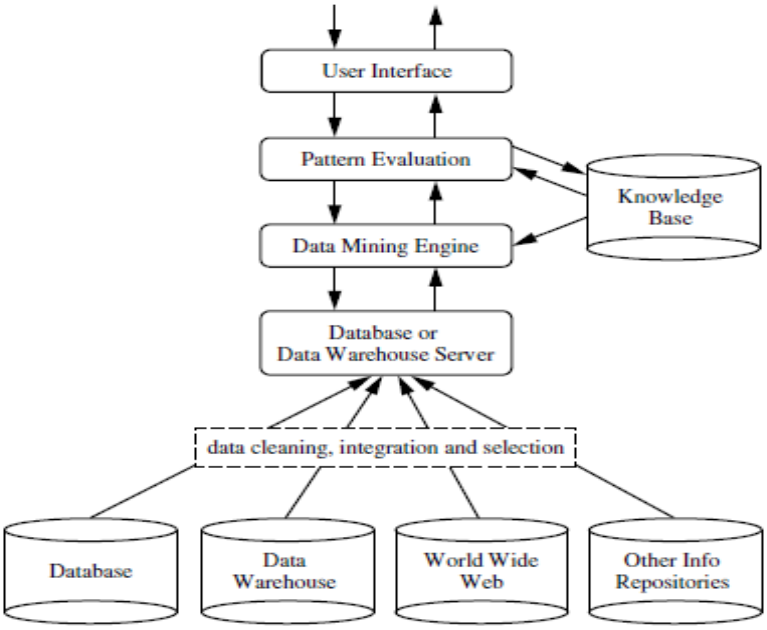


Figure 1.4: Architecture of a Data Mining System [4]

### 1.4.1 Motivation for Data Mining

The era of internet has provided a great boost to each and every field the mankind is involved in whether industry, commerce, education, research, biology, medicine, earth science, astronomy etc. . Each of these fields is generating a massive amount of digital data daily which is stored in large databases called Data Warehouses. This large amount of data in the raw form is incomprehensible and needs intelligent tools and techniques to convert it into valuable information. Here comes the Data Mining which extracts knowledge from this huge raw data.

### 1.4.2 Types of Data Mining

Data Mining algorithms can be broadly classified into [4]:

1. *Association Rule Mining*- extracts useful information in the form of relationships between data items from massive data, which can further be used for Market Analysis and strategy planning.
2. *Classification Algorithm* – is a type of Supervised learning algorithm which maps the data items to one of the pre specified categories.
3. *Clustering Algorithm* – unlike classification is a UnSupervised learning algorithm which maps the data items to classes without any prior knowledge of categories.
4. *Stream Data Mining Algorithm* – performs mining on the stream of data which is continuous or dynamic in nature as opposed to the traditional static data.

### 1.4.3 Clustering

Clustering is a form of Unsupervised Learning that groups the data on the basis of similarity, with high intra-cluster similarity and low inter-cluster similarity. Hierarchical clustering and partition clustering are the two most popular techniques. Figure below shows the clustering of similar news item based on the query of the user.

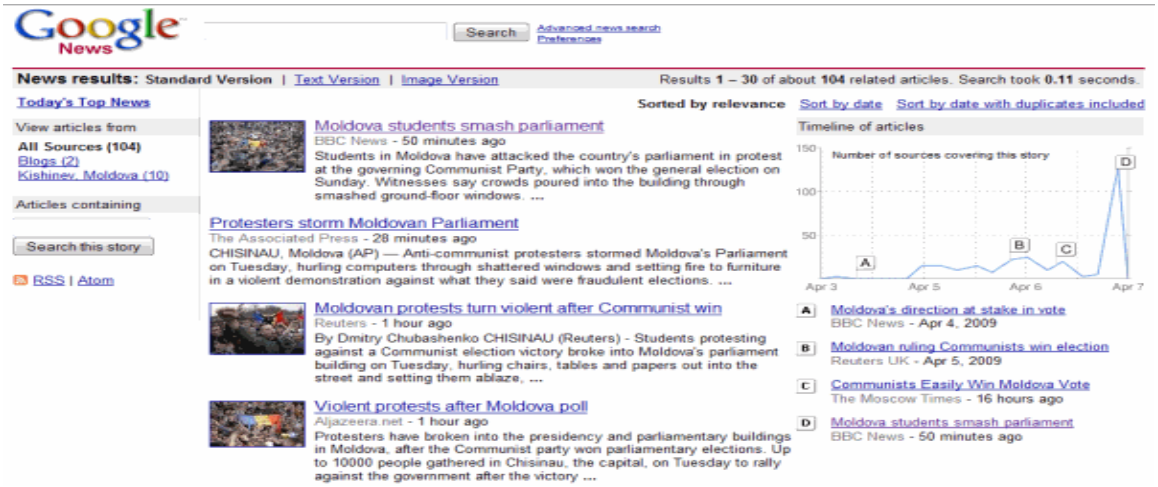


Figure 1.5: News Clustering

**k-means Clustering:**

K-means is one of the most widely used partitioning algorithms to create clusters of data through unsupervised learning. This algorithm tries to find cluster centers so as to minimize the summation of squared distance of each and every data point to its closest cluster center. The basics of k-means follow the following steps [Jain and Dubes (1988)] [6]:

1. Initial partitions are selected with k-clusters, go to step 2 until the stabilization of cluster membership.
2. New partition is created by assigning each sample to nearest cluster centroid.
3. Updation of cluster centers.

Numbers of clusters, distance metric and initial cluster centers are user-specified parameter for k-means algorithm. The numbers of cluster k and the initial cluster positions affect the results a lot.

## 1.5 Data Mining in Cloud

Data mining processes the raw data to extract useful information and cloud computing offers the scalable and flexible infrastructure providing platform, Software and

infrastructure as a service. Thus, combining data mining and cloud computing results in an agile and quick access to the vast amount of valuable information stored in the enormous amount of day-to-day data, which otherwise is incomprehensible. Cloud computing provides a firm and capacitive platform to deal with such huge amount of data, whether in the form of storage capability or in the form of processing capability. Thus, integration of data mining and cloud computing results into discovery of helpful information and generation of new knowledge.

Despite all the above powerful functionalities provided by the relationship between cloud and data mining, it can prove to be a serious issue for both client and cloud provider. Data privacy issue or confidentiality of data is one of the major. As all the data resides with the cloud provider, a serious data privacy issue arises if the provider misuses the data or the information. Also any outside attacker or adversary having an unauthorized access to the storage on cloud can mine the data and retrieve large amount of confidential data. Various data analysis techniques or algorithm are available today which can be used successfully to mine valuable information from the large datasets by analyzing the behavioral and statistical data. Many cloud providers offer these data mining facilities to users which can be used by an adversary. Google also uses some data mining technique to predict search results by analyzing the user behaviors [7]. So, data mining can be a serious threat to the cloud security. If the data mining provider Company is itself a client to cloud, violation of privacy or data intrusion can also lead to loss of reputation and customers. Also, for the organizations dealing with the financial, governmental, education or legal issues of people, leaking of information can sometime result in national catastrophes for e.g. collection of financial, health etc information by TIA (Total Information Awareness) in 2002 [8] and analysis of phone records of people gathered from phone companies by NSA for identifying the possible terrorists in May 2006[8]. Also, according to a survey conducted by Rexer analytics, 7% of the data miners analyze the data using the cloud [9], due to the cheap and elastic computing powers offered by the cloud computing. So, maintenance of client privacy is synonymous with data privacy in cloud and is a major area of concern for the cloud provider as well as cloud user.

## 1.6 Organization of Thesis

The rest of the thesis is organized as follows:

**Chapter 2-** This chapter contains the survey of the work already done in the past or any ongoing research activity related to this Thesis work.

**Chapter 3-** This chapter states the problem that is addressed in this Thesis.

**Chapter 4-** This chapter focused on the detailed solution of problem stated in the Thesis with the details of the hardware and software technology and tools used.

**Chapter 5-** This chapter gives a detailed analysis of the findings and the experimental results.

**Chapter 6-** This chapter states the conclusion drawn from the experimental analysis and the future scope in the concerned problem.

# Chapter 2

# Literature Survey

This chapter gives the literature survey conducted to study the various algorithm to test the performance and need of cloud mining and the techniques being used to secure this cloud mining.

## 2.1 Data Analysis in Cloud

A lot of work has been done in the field of data analytics based on cloud platform and it can be deduced that cloud mining is an efficient technique in nearly each area of study.

[10] discusses various cloud computing based parallel data mining algorithms and gives the problems and future scope of the data mining integrated with cloud computing.

A detailed discussion of the need, issues and necessities of big data analytics in a cloud computing environment is given in [11]. They states that with the increasing pace of data generation or with the increase in amount of big data, some new, smart, efficient, fast and scalable programming tools and applications are required which are satisfied by the cloud computing platform. It states that the high-performance infrastructure of cloud satisfies all the needs of the computation-intensive big data analytics. It presents the three cloud service models in terms of data analytics as below :

**Table 1. Cloud-based data analytics services.**

| Cloud service model | Features | Users |
|---|---|---|
| Data analytics software as a service | A single and complete data mining application or task (including data sources) offered as a service | End users, analytics managers, data analysts |
| Data analytics platform as a service | A data analysis suite or framework for programming or developing high-level applications, hiding the cloud infra-structure and data storage | Data mining application developers, data scientists |
| Data analytics infrastructure as a service | A set of virtualized resources provided to a programmer or data mining researcher for developing, configuring, and running data analysis frameworks or applications | Data mining programmers, data management developers, data mining researchers |

Figure 2.1: Data Analytics as a Service [11]

and further discusses some issues and necessities relating to big data analytics tools based on cloud.

[12] proposes a framework, ‘datacloud’ based on ‘Rabbit’, for massive data analysis on large volumes of data due to the weakness of traditional data analysis techniques to mine data from large storage and the incapability of the traditional tools to scale to cloud.

A case study for price function is presented in [13] by applying frequent pattern mining algorithm. They performed the experiments on top of cloud computing platforms and deduced the increase in efficiency and speedup as a result of high capabilities of cloud.

[14] presents their research which aims at designing the IT Management solutions to leverage the data analytics techniques, which otherwise is not efficient in traditional techniques due to large volume and variety of the data and lack of high-performance scalable frameworks or tools. This paper uses Hadoop, Cassandra and RHadoop as the test bed and shows that the performance increases with the decrease in volume of data divided on several nodes of cloud as compared to a single node data mining.

As more and more survey was performed on data analysis in cloud computing environment, each experiment or analysis supported the theory of increased performance of Data Mining in a cloud environment.

[15] conforms to all the work done earlier on big data analysis on cloud which states the performance enhancement. This paper gives the technical architecture of the Data Mining system based on Hadoop platform and shows with experimental results performed on HSprint algorithm that if the calculation on a single traditional module is distributed on multiple nodes of Hadoop, increased performance and parallel computing is achieved.

Map-Reduce cloud computations is employed in [16] to find out a real-time platform suitable for the dynamic data or stream analytics and for distributed data.

SeRuM [17], BC-PDM[18], CAM [19] and FIU-Miner[20] are some of the other algorithms proposed to perform Data Mining for a particular type of data based on the high-performance cloud computing.

**k-means Clustering** is the most simple and frequently used Data Mining algorithm which can be easily scaled to run on a cloud environment. The popularity of k-means algorithm is due to its simplicity but as the size of dataset increases, the time complexity increases. [39] conducted experiments on Hadoop/Mahout and found the algorithm to scale well. [5] ran the k-means algorithm on Mahout/Hadoop on a noisy, realistic big dataset for document clustering and found that k-means algorithm does not fare well in the situations where there is inherent overlapping in clusters. [41] used the Map-Reduce programming model to implement the k-means algorithm for the cloud platform. It stated the basic requirement of parallel k-means algorithm is to decompose the dataset into smaller datasets so that each subset can be parallel processed. It was concluded that when the dataset is small, the running time of serial k- means is lesser but as the dataset becomes large, parallel k- means becomes faster.

In [42] a hierarchical virtual k-means approach is used which include data processing as a part. As single data node is unable to handle the large data analysis demand, thus approach integrates the concept of virtualization into the simple most basic k-means approach of the unsupervised learning or the clustering.

The k-means approach is combined with hierarchical clustering [39], which can either be bottom up (merging k clusters into k-1 clusters) or top down (dividing k clusters into k+1 cluster)

### 2.2 Secure Cloud Mining

Integration of Cloud Computing with Data Mining is the most famous technique for efficient and fast mining of valuable information from the massive amount of data being generated every day. But as Cloud Mining techniques makes it easy to mine data from cloud storage, at the same time it makes it easy for the adversary also who can misuse the data to create harm to the user or for personal benefits. So, it is the need of the hour to

secure these cloud based Data Mining techniques to prevent any attacker with malicious intent from getting hold of the sensitive or confidential information. A lot of secure Data Mining techniques have been proposed in literature and many of them have been applied in Cloud. A brief discussion of these is presented here.

### 2.2.1 Secure Data Mining Algorithms

Preserving the privacy of the data mining algorithm has been a concern of researchers for long and a number of algorithms have been proposed for the same. [8] reviews different methods of privacy preservation and analyzes the underlying techniques for Data Mining. [7] focuses on improving the security of two-party k-means while maintaining the correctness of algorithm. k-anonymity [10], noise transformation and multiplicative transformation [9][17] are some PPDM(Privacy Preserving Data Mining) methods. [14] uses k-NN classifier with an equality test to find the similarity.

### 2.2.2 Security Issues in Cloud Data Mining

[29] is a detailed survey of the key security challenges faced by the developers while designing the cloud application, privacy risks to the Cloud, the various Privacy Requirements and finally gives the design guidelines for developers to tackle the issue of privacy.

A lot of fears bug modern day cloud computing techniques which according to [7] are not sufficient to control the loss of sensitive data from cloud. They propose an extended solution to the current techniques using trusted computing and various new, modified cryptographic techniques for privacy enhanced business intelligence.

The attacks in a Cloud Data Mining system can be listed as *DoS* (Denial of Service) attack, *DDoS* (Distributed Denial of Service), *Sniffing, DNS attack, Man in the Middle attack* etc.[33] gives a detailed survey on the security issues in cloud and a description of the types of attacks possible in a Cloud Data mining environments. It details the above types of attack in the cloud, their impact and possible solution to some of them.

According to [41] data mining attacks in cloud falls in three classes: network-level, application level and virtualization level. [32] discusses about the Network level attacks of the Cloud system. They propose a solution for these type of attacks which is deployed on IBM SCE in the form of “Security-as-Service”. This application prevents the high-level security attacks. Also the effect and challenges of NAT have been discussed.

Application level security is discussed in [33]. This discusses various issues regarding the deployment, moving a service on cloud in detail. It mainly focuses on building transparent cloud application using loosely coupled services.

Virtualization is the key concept of cloud computing these days but it too act as a loophole in the security of the Cloud. [34] discusses the security of the virtual network residing in a virtual environment. They first discuss the security issues in the virtual machines and network and then proposes a solution in the form of a framework to control these security issues.

### 2.2.3 Some Cloud mining Security approaches

[35] discusses the recent advances in the cryptographic security mechanism and try to apply those in the cloud environment. But, [25] states that cryptography alone cannot prevent the attacks on the cloud mining systems and some other form of security must also be imposed.

Fragmentation technique or partitioning of the database into chunks [26] is another method for security which suggest that keeping the data with different cloud service provider or nodes will prevent an adversary from having the access to complete data and thus will not be able to infer correct results. This theory was verified by experimental results sing a cloud distributor to fragment the data.

A different approach is proposed in [27] for secure mining. They employs a privacy preserving repository which with a query plan wrapper limits the task of the data sharing and the access to the shared data with the encrypted results thus, maintaining the confidentiality as well.

[36] discusses the k-anonymity and k-anonymity noise taxonomy in a multi-cloud environment to perform frequent pattern mining. It proves that distributed data or a multi-cloud environment prevents the attacker from getting hold of the complete data thus cannot infer valuable information from the data.

A one-time pass key mechanism [41] can be used to preserve the privacy of the user as well as the service provider. This approach is based on the terminology of the authentication of both user and the provider.

[37] proposes a SCM (Secure Cloud Mining) architecture for the generation of secure forecasting reports for an organization by identifying the interesting patterns and links between variables in a multivariate database system. It employs image based encryption for secure forecasting.

The secure collaborative outsourcing data mining is discussed in [38]. This paper proposes practical scheme as most of the schemes assumes the models to be semi-honest adversary model. But practically every node has a malicious intent. It presents a case study of knn (k-nearest neighbor), SVM (Support Vector Machine) and k-means in the above mentioned outsourced collaborative environment.

A detailed comparative analysis of some of the above studied techniques is given in [44]. They give a brief discussion of the underlying technology in each method and then compare them on basis of certain parameters.

A lot of Privacy Preserving Data Mining (PPDM) techniques exist today. [22] gives a review about all these existing techniques and analyze the representative PPDMs. It finally concludes that most of the existing techniques are an approximation and need to be perfected further if efficiency and accuracy is required as most of the algorithms compromise one for the other and to get a balance between them more robust, dedicated and perfect PPDMs are required.

# Chapter 3

# Problem Statement

Data Security in cloud is one of the major issues that need to be tackled these days. Data mining attacks is one of the most common problems of data privacy. One of the approaches suggested to tackle this is to *Decentralize* the data and to keep it in chunk at different locations so that the intruder doesn't have the access to complete data thus preventing the data mining. But when an authenticated user requires to mine information from the data, a joint computation is to be performed on that data. Thus, we need a secure method to compute the combined result from the decentralized data while maintaining the correctness of the final results.

But Decentralization of data is not the only solution toward data security as it is also susceptible to data/information leakage or addition of malicious data by an adversary, at the time joint computation is performed, if any of the hosts storing data is not honest. So, a secure model or protocol is required to perform the k-means clustering approach for data mining on a multivariate, horizontally partitioned dataset while maintaining the privacy of the content at both the host and also preventing the intermediate values to be leaked to the adversary.

*A semi-honest model* of adversary is assumed, where a host can reveal other host's data, if not secured, while maintaining the privacy of its own. It is desired that the hosts know their inputs, the final outputs and no intermediate values.

## 3.1 Objectives:

- Preventing leakage of Host's data in the process of communication for joint computation.
- Preventing intermediate value leakage.
- Maintaining correctness of algorithm and validity of final results.

# Chapter 4

# Details of the Project

This chapter presents a detailed solution for the problem stated in the previous chapter along with the assumptions and the related concepts required for the proposal.

## 4.1 Introduction

This work proposes a secure multiparty communication method or protocol to solve the security issues for data mining in cloud environment. It aims to perform the k-means clustering over a distributed cloud environment, where data is decentralized, while maintaining the privacy of individual data as well as any intermediate value. It prevents any information from leakage to an adversary. This approach deals with a semi-honest adversary assuming both Host A and Host B to be semi-honest i.e. each host wants to prevent the privacy of their own data but revealing the data of the other. But, in this protocol the only data known to each host is his input and the final outcome. Thus a semi-honest adversary cannot simulate any useful information by using intermediate outputs. Our aim is to preserve the privacy as well as correctness of the algorithm.

## 4.2 Data Distribution

A multivariate relational database depicted as $D= \{d_1, d_2,\ldots. d_n\}$ in which Host A has $D_A=\{d_1^A,\ldots\ldots, d_1^A\}$ and Host B $D_B=\{d_1^B,\ldots\ldots, d_1^B\}$. As the database is multivariate, each data object is denoted by a vector set $d_i=x_{i,1},\ldots\ldots, x_{i,m}$ where m is the number of attributes. Now, let Host A have a set of private clustering centers $H_1^A, H_2^A,\ldots\ldots\ldots, H_K^A$ while Host B has $H_1^B, H_2^B,\ldots\ldots\ldots, H_K^B$ and $(C_1,C_2,\ldots\ldots\ldots,C_k) = \{ H_1^A + H_1^B ,\ldots\ldots, H_k^A + H_k^B \}$ as the joint cluster centers. Here, $k$ is the number of clusters.

## 4.3 Encryption Formulas:

To preserve the privacy of the data of each host and the intermediate results which are communicated to and fro an encryption system is required in which if any specific operation is performed on encrypted data or cipher text, the results generated matches the operation performed on plaintext when decrypted. This system of encryption is known as Homomorphic encryption system. For this purpose the pallier cryptosystem[27] is used which satisfies the need of the approach. Homomorphic formulas $E(a).E(b)=E(a+b)$ and $E(a)^b=E(a*c)$ are used in this approach, where E is the required encryption scheme.

Let M be the message or the plain text which is to be encrypted. The system can be divided into 3 parts (K,E,D):

- A pair of public and private key $(l_k,p_k)$ is generated.
- A ciphertext or encrypted message $c = E_{l_k}(m, r)$ is obtained where $m \in M$ and $r$ is a random value.
- Decryption $D_{p_k}(c) = m$ is used to obtain plain text again.

## 4.4 Proposed Algorithm

**Notations:** $C_i$ represents the combined clustering centers which is the sum of Host A and Host B's share i.e. $H^A$ and $H^B$ respectively where $C_i = H^A + H^B$.

**Input:** 1) Database $D_A$ and $D_B$ belonging to Host A and Host B respectively having n data objects.

2) '*k*' which is the total number of clusters.

**Output:** The *k* cluster which is the combination of $D_A$ and $D_B$ or *D*.

1) Each party performs Data Normalization on local data.
2) Host A and Host B select their respective *k* cluster centers $H_1^A, H_2^A, \ldots\ldots, H_K^A$ and $H_1^B, H_2^B, \ldots\ldots, H_K^B$(locally) randomly.

$$(C_1, C_2, \ldots\ldots, C_k) = \{ H_1^A + H_1^B, \ldots\ldots, H_k^A + H_k^B \}$$

3) Calculate or perform local k-means for Host A and Host B.

4) Save the cluster centers $H_j^{A,i}$, $H_j^{B,i}$ .

5) Perform the secure cluster updation and reassign the data objects to their closest clusters locally

6) Save $H_j^{A,i+1}$, $H_j^{B,i+1}$ . If the difference between the previous cluster center and the current one is less than or equal to threshold value then stop the iteration else repeat step 4 onwards.

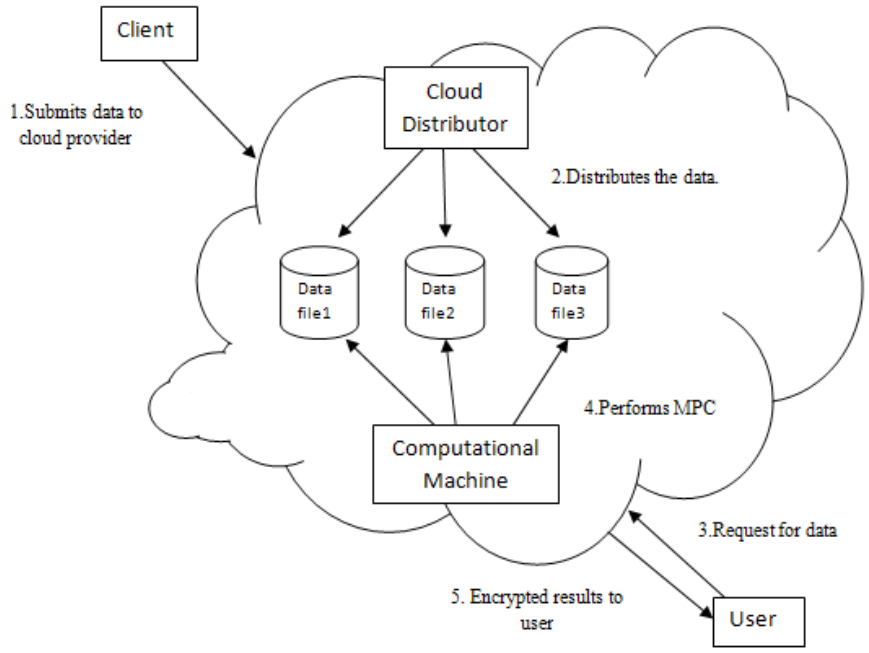


Figure 4.1: Overview of the proposed approach

## 4.5 Detailed Explanation of Methodology

### 4.5.1 Private Data Normalization:

A standard Xml document is used to submit the data so that a data standard is maintained. But as the approach is dealing with multivariate database i.e. a multi-attribute database where the value of variable is obtained as a sum of different attributes, the probability of some variables having large values is high, which can dominate the entire metric. Thus, a

normalization method is used to standardize the multi-attribute data, using private mean computation of the data objects.

Let Host A has $d_A = \sum_{i=1}^{n} d_i^A$ with *n* data entries

And Host B has $d_B = \sum_{i=1}^{m} d_i^B$ with *m* data entries

Then mean $M = \frac{d^A + d^B}{n+m}$

This mean is generated using Pallier Homomorphic cryptosystems. Now, the data is standardized locally using the above mean value as

$x_i = x - M$ for all data objects $x_i$

### 4.5.2 Distance Measuring and Updation of Clusters

After the standardization of the data a local k-means is performed by all host on their respective datasets and initializes the cluster center for each attribute and assign data objects to the nearest luster center using Euclidian or Manhattan distance which can be chosen according to the application or database, i.e. $h_1^A$,……,$h_k^A$ for Host A and $h_1^B$, ……, $h_k^B$ for Host B. As these cluster centers are calculated locally there is no need of any security protocol but in the next step of updating the cluster centers, joint centers are to be found which needs to be calculated privately.

**Cluster Updation:** for every data object's values in the j[th] attribute in i[th] cluster, calculate sum as

$S_j^A = C_{i,j} * n_j$ where $n_j$ is number of data objects for j[th] cluster

$S_j^B = C_{i,j} * m_j$ where $m_j$ is number of data objects for j[th] cluster

Now, new *i*[th] cluster center for *j*[th] attribute is

$$C_{i,j} = {S_j^A + S_j^B}/{n_{j+} m_j}$$

Pallier homomorphic cryptosystem [27] is used to do the above computations privately as: Host A, B and the Third Party (Computational Machine) randomly generates a pair of public/private keys ($l_k, p_k$). Host A and B encrypts the sum value with Third party's public key and send it to third party along with their Public keys.

**Iteration Stopping Criteria:** As known that k-means is iterative in nature, so there must be a criteria which when met stops the iterations. This iteration stopping criteria is reached when output requirement are satisfied. For k-means this criterion is that the eucledian distance between two consecutive cluster calculations is less then ∈ (threshold value). .i.e. Dist ($C_j, C_{j+1}$) = Dist ($H_j^{A,i+1} + H_j^{B,i+1}, H_j^{A,i} + H_j^{B,i}$) < ∈ or ($H_j^{A,i} + H_j^{B,i}$)-( $H_j^{A,i+1} + H_j^{B,i+1}$) < ∈. To check this Host A computes Enc ($H_j^{A,i} - H_j^{A,i+1}$) and host B Enc ($H_j^{B,i} - H_j^{B,i+1}$) locally with third party's public key. Then third party do multiplication of intermediate encrypted values and HOST A and B decrypt with their private key as follows:

$$T = Dec\ [Enc\ (H_j^{A,i} - H_j^{A,i+1}).\ Enc\ (H_j^{B,i} - H_j^{B,i+1})]$$

If $T < \epsilon$, then the desired output is reached and the iterations can be stopped.

## 4.6 Assumptions

- A semi-honest model of adversary is assumed by the proposed approach in which a host can reveal other host's data, if not secured, while maintaining the privacy of its own.

- This approach assumes that the data input by client is stored as chunks [26] at different locations instead of storing whole of the data centrally, as, the centrally stored data is more vulnerable to the attacker. Thus the client's data is stored in a decentralized manner by partitioning the database horizontally. Horizontal

partitioning is referred to the partitioning scheme where each site has different records which contain same or equal set of attributes.

## 4.7 Implementation Steps:

### 4.7.1 Preliminaries

1. **k-means** - k-means is one of the most widely used partitioning algorithms to create clusters of data through unsupervised learning. This algorithm tries to find cluster centers so as to minimize the summation of squared distance of each and every data point to its closest cluster center.

$$D = [min\ k = (1...k)\ d\ (x_i, c_k)]^2$$

Where $d$ is distortion, $k$ is the no of clusters, $c_k$ is the cluster centroid and $x_i$ is data point. k-means can be summarized as follows:

k initial cluster centers $C_1$,……..…..,$C_k$ are chosen randomly for 'm' attributes of datapoints $x_i$.

**Repeat**

**For** each datapoint $x_i$

Assign data point to the closest cluster center .i.e. having minimum squared distance.

**End for**

Calculate new cluster centers as mean of all datapoints belonging to that cluster $C_1^{'}, \ldots \ldots \ldots C_k^{'}$.

**Until** the change in mean for consecutive iteration is less than the threshold value.

2. **Tools:**

- **Hadoop** [46] - Hadoop is a Java framework that runs applications on large clusters of commodity hardware and comprise of features like Google File System (GFS) and the Map Reduce computing prototype. Hadoop's HDFS is distributed file system that is extremely fault-tolerant and, is designed to be mounted on low-cost hardware. It is most suitable for applications with large datasets and has a high throughput access to the data being used by application.

- **Mahout** [45] - Mahout is a machine learning library provided by Mahout and is open source. Currently it primarily implements *recommender System*, *clustering*, and *classification algorithms*. It's also provides scalability across machines. It can be the machine learning tool for the processing of collection of large data, which may be too large for a single machine. Mahout should be run on top of Hadoop when a large amount of data is to be processed.

3. **Parameters Used:**
   - k –no of the clusters. Default is 6 clusters.
   - x – No. of iterations which is taken to be 10.
   - dm – distance measure used is Cosine Distance.
   - cd – convergence delta or threshold value which is taken as 0.5.

### 4.7.2 Testbed

**Dataset used**: Synthetic_control data, which is control charts exhibiting the time series comprising of 6 different classes, from UCI Machine learning repository is used [49].

- 600 records are there with 60 attributes per record.

**Technology used:**

- Linux 12.04, 64-bit – 40GB hardisk,1.5 GB RAM
- Jdk 1.7.0_60

- Hadoop-1.2.1
- Apache maven-3.2.1
- Mahout-0.9

### 4.7.3 Steps:

1. Installation of Hadoop [46] in a single-node cluster format and mahout [47] on top of that is the preliminary step for running the mahout k-means algorithm. After installation, hadoop services are started, as hadoop namenode is the one doing all the processing task or Mapreduce job on the data submitted to HDFS.

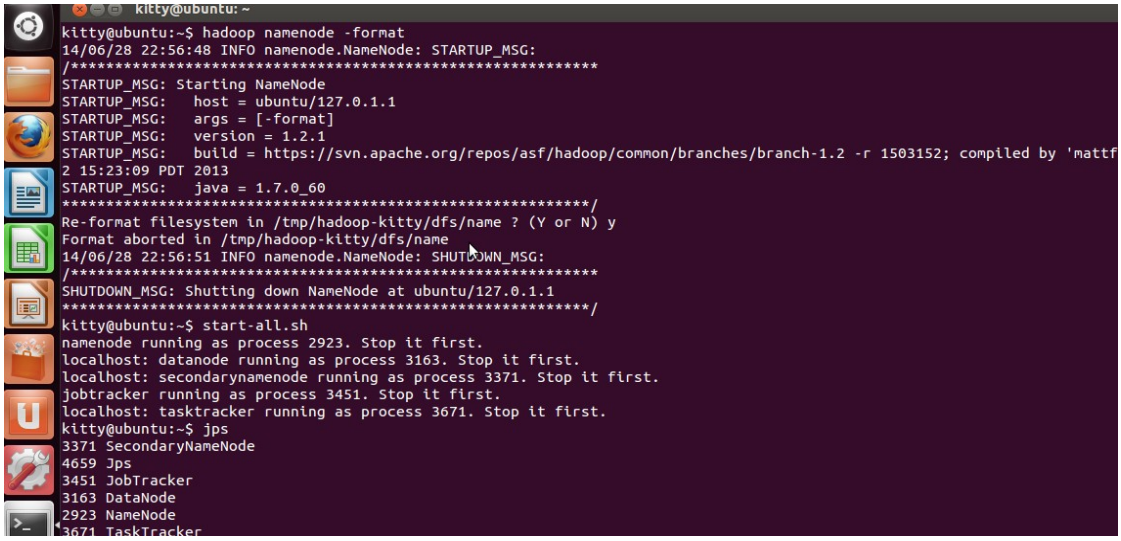


Figure 4.2 : Hadoop services startup

2. As per the assumption that the data resides at two different locations, the dataset is horizontally partitioned and stored in hadoop file system as ‘testdata1’ and ‘testdata2’ containing the respective sequence files converted to vector format using mahout’s ‘seqdirectory’ and seq2sparse.

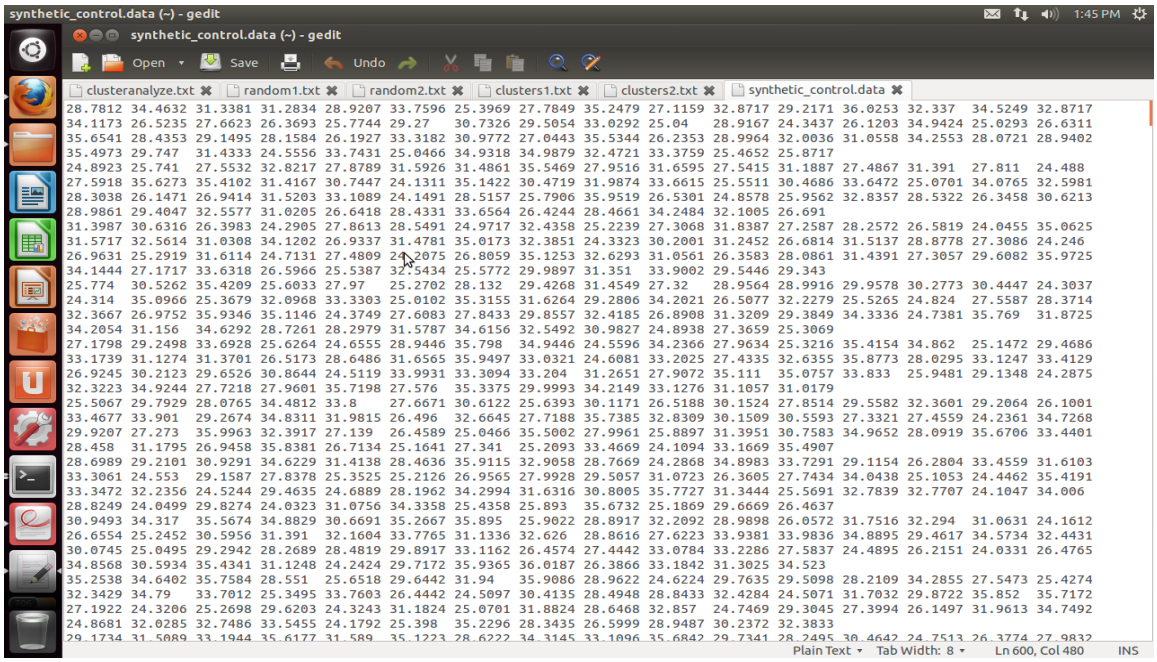


Figure 4.3: Dataset

```
kitty@ubuntu:~$ $HADOOP_HOME/bin/hadoop fs -mkdir testdata1
mkdir: cannot create directory testdata1: File exists
kitty@ubuntu:~$ $HADOOP_HOME/bin/hadoop fs -mkdir testdata2
kitty@ubuntu:~$ $HADOOP_HOME/bin/hadoop fs -ls
Found 2 items
drwxr-xr-x   - kitty supergroup          0 2014-06-28 22:56 /user/kitty/testdata1
drwxr-xr-x   - kitty supergroup          0 2014-06-28 22:59 /user/kitty/testdata2
kitty@ubuntu:~$
```

Figure 4.4: Directories in hadoop file system

3. To run the mahout k-means algorithm with user defined parameters either initial cluster centers are to provide or ‘k’, the number of clusters. Initial clusters represented in vector format are stored in their respective directories in HDFS. After transferring all the required files the HDFS looks like:

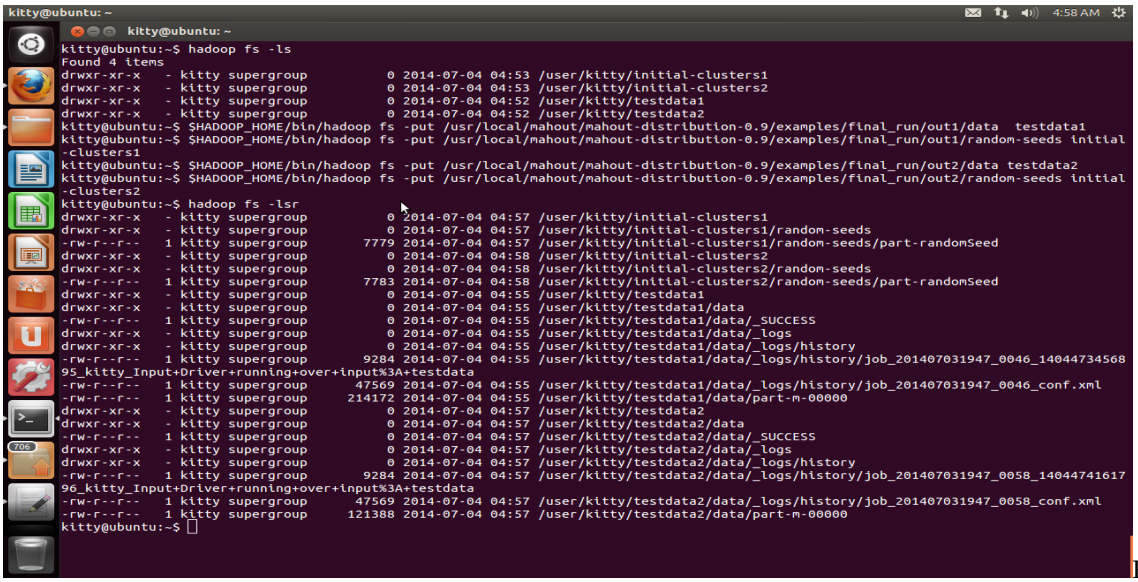


Figure 4.5: Hadoop File System

4. After setting up the environment with required files and tools mahout k-means is run with the required parameters on both the datasets separately and two sets of cluster centers is obtained.

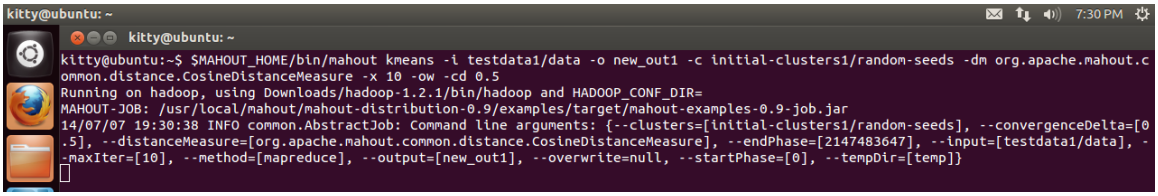


Figure 4.6: k-means with arguments

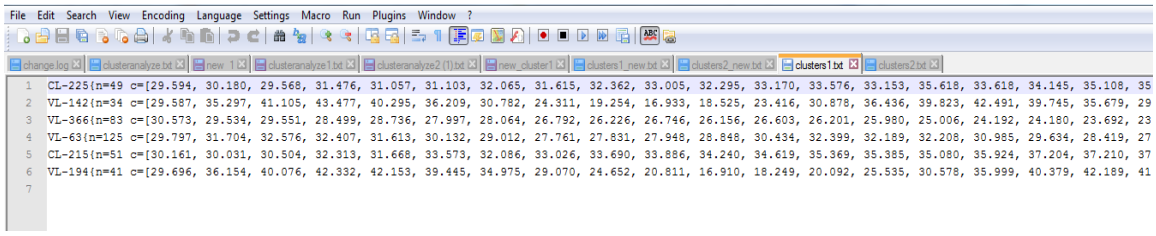


Figure 4.7: Cluster1 (Host A)

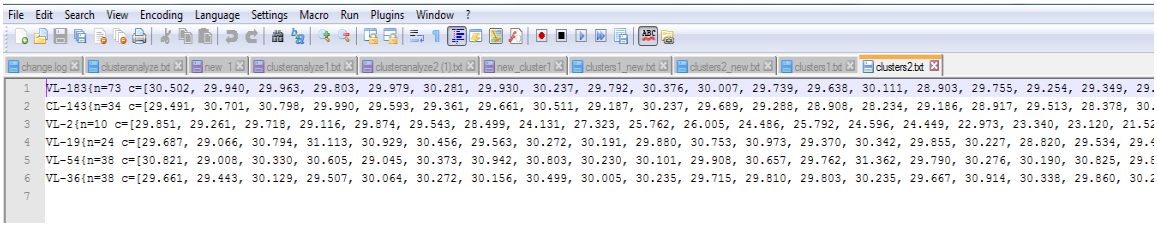


Figure 4.8: Cluster2 (Host B)

5. Now to update the value of the above cluster centers we need to calculate the combined mean of the cluster centers of Host A and Host B privately. So, first we calculate the sum of all the points clustered to a particular cluster for a given attribute as $S_{(i,j)} = C_{(i,j)} * n_i$ where $C_{(i,j)}$ is the cluster center for $j^{th}$ attribute of $i^{th}$ cluster and $n_i$ is number of points in $i^{th}$ cluster.

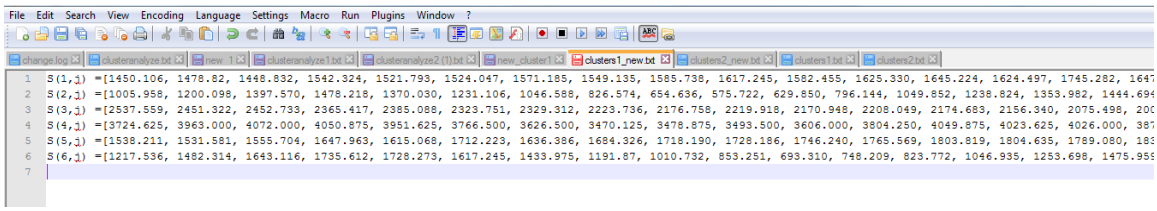


Figure 4.9: Sum of values of Cluster1 (Host A)

After the computation of sum locally Host A, B and a third party generates a pair of (public, private) keys using Paillier cryptosystem. Third party sends its public key to both A and B, which encrypts their data with this public key and sends back to the third-party, along with their public key. Third party performs the homomorphic addition of the data and sends the final result back to the A and B encrypted with their respective public keys.

6. A and B decrypts the result obtained with their private keys and gets the final cluster values.

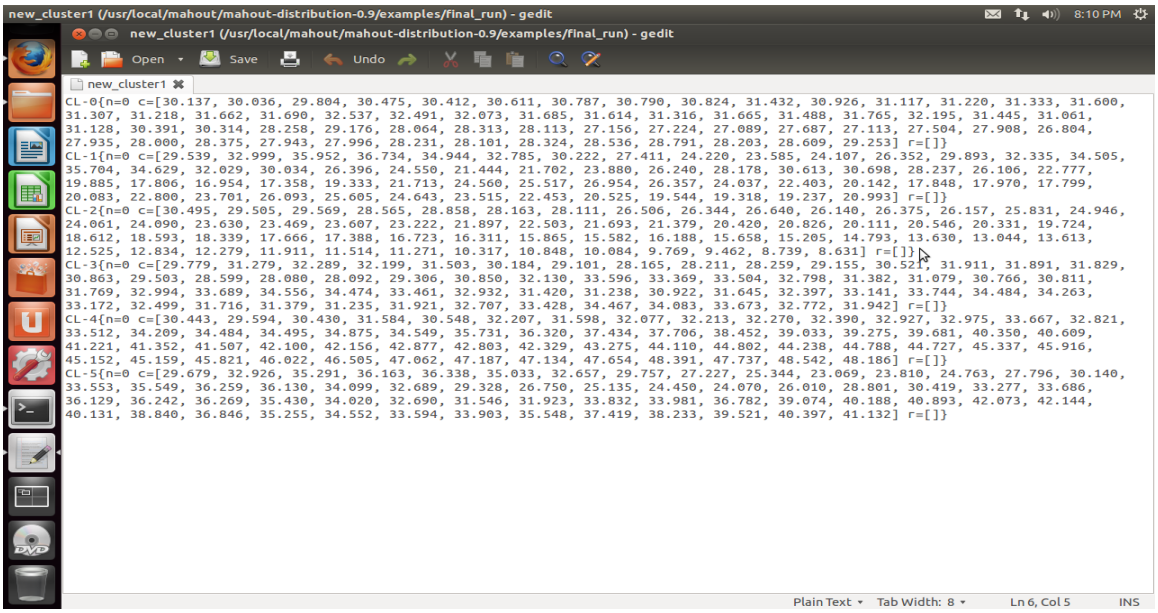

Figure 4.10: Updated Cluster centers

Now A and B again performs k-means on their respective datasets with these new cluster values and follow the above procedure until the threshold value calculated as the distance between current and previous cluster center start lying in the range of a previously set '∈' value.

7. After the specified number of iterations the final results of both A and B along with their clustered points are merged to get the combined result.

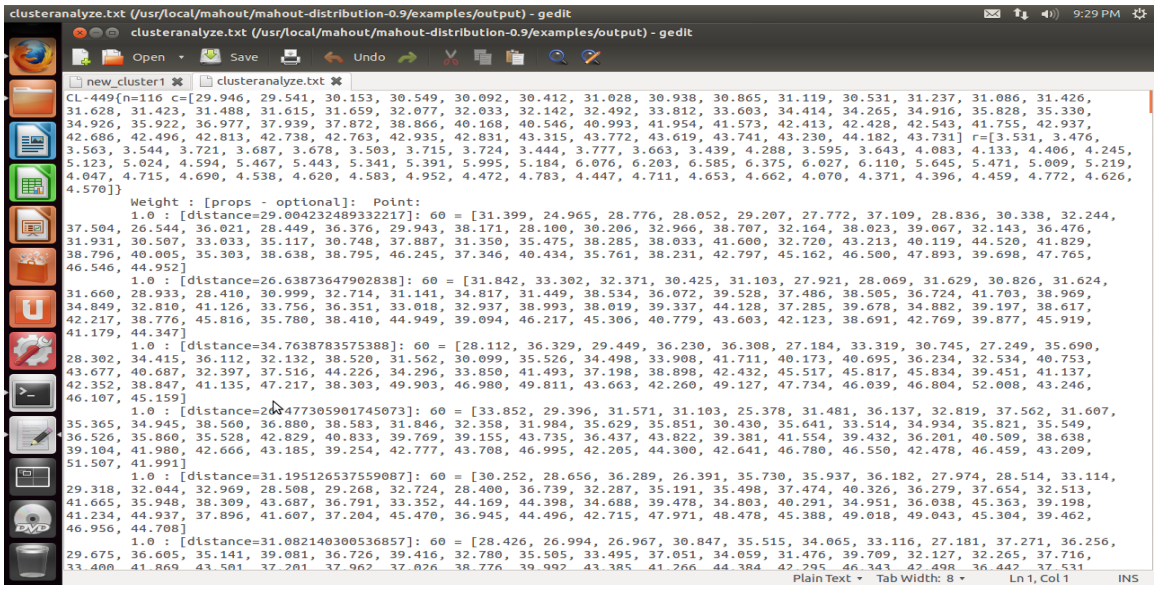


Figure 4.11: Final clusters with clustered points

# Chapter 5

# Experimental Results & Analysis

This chapter presents the discussion and analysis of the results obtained, in the experiments performed in this Thesis, on the basis of certain parameters which are compulsory for the validity and acceptability of the above approach.

## 5.1 Evaluation Parameters

### 5.1.1 Correctness

Correctness refers to the validity of the final results obtained or the outcome of the experiments performed using the proposed approach, on the same hardware and software platform as compared to the original or base approach. The correctness is checked by comparing the deviation of the results from the anticipated results.

### 5.1.2 Security

This parameter evaluates the proposed algorithm in terms of security i.e. the capability of the algorithm to prevent the attackers, with malicious intent, to gain access to the confidential user data and valuable information inferred from the raw data.

## 5.2 Results

- The proposed approach performs k-means clustering on a dataset which is horizontally partitioned and stored on two different locations. The approach first run locally then performs a joint computation on encrypted intermediate results so as to obtain complete result. It was observed that running secure k-means on the partitioned data with same parameters and computation environment as the original single party k-means, produced the same end results and same inference, thus, validating the correctness of the proposed approach.

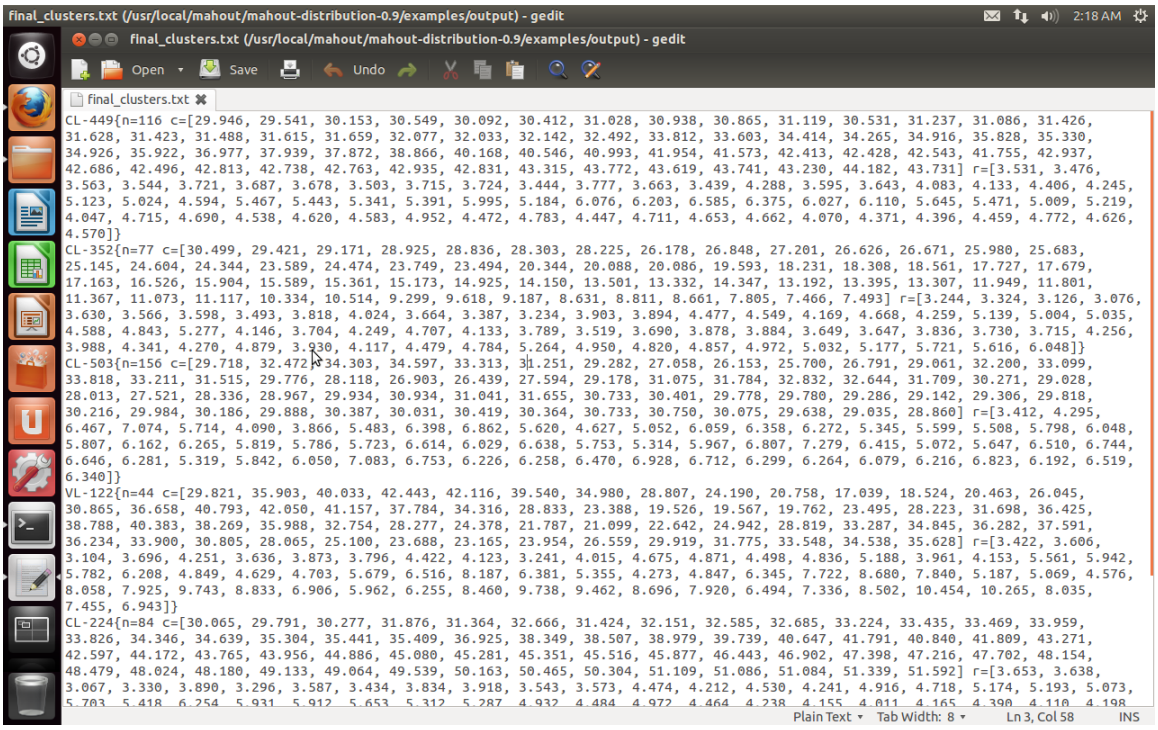


Figure 5.1: Final Clusters on Decentralized data

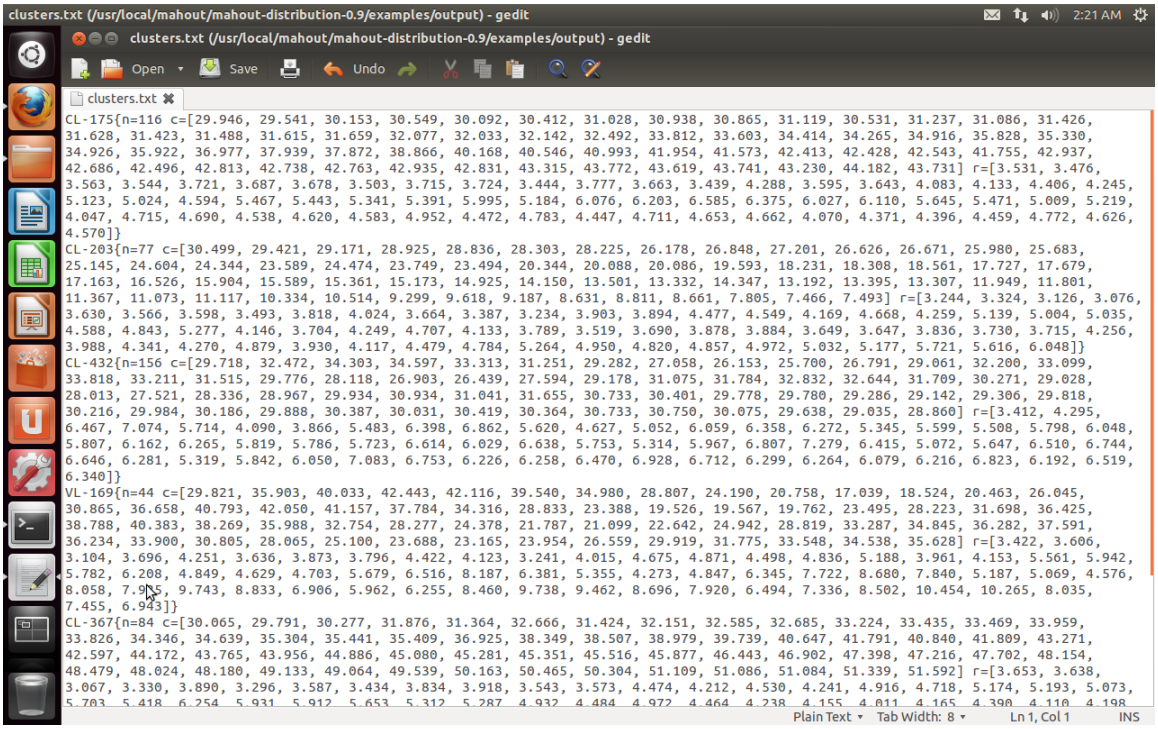


Figure 5.2: Final Clusters on Centralized data

The above figures show the correctness of the proposed algorithm. It can be seen that the final cluster centers obtained by the merging of the clusters and the clustered points obtained in the final iteration of the two-party k-mean computation is similar to the cluster center obtained by the running of k-means algorithm single time. The correctness can be further seen as both the algorithms were run on same environment and platforms with same hardware and software configuration.

Thus, it is proved that the algorithm maintains the correctness and validity of the final result and thus can be applied to all situations where a single party k-means can be used.

- Coming to the security issue we know that the model uses a partitioned approach to store the large dataset .i.e. the dataset is fragmented horizontally with a certain number of records with n attributes stored on Host A and the other set of record on Host B. Thus, *fragmentation* is the first step towards the security against data mining based attacks as the intruder which otherwise could, after getting an unauthorized access or entry to the data storage point, easily use the cheap and simple data mining techniques to extract valuable information from the data. But, as the data is fragmented and kept in chunks at different locations getting the correct information from the incomplete data becomes impossible thus fending off the attack by the adversary.

*Secondly*, the assumed model is that of a *semi-honest adversary* .i.e. participants try to leak the data of one another while maintaining their privacy. This approach deals with this threat as the intermediate results of both the party goes to a third party, and that too in an encrypted form, and it performs the computation on the encrypted data and returns the encrypted results to each party. Thus, each party only knows their intermediate values and the final value but not the data of the other party.

*Lastly*, as the data goes to the third party encrypted with a key, if an intruder is able to pick the data in the transition he/she will not be able to decipher the encrypted data to get the original values and to simulate the approach with those values. This prevents Sniffing attack on the data-in transit.

# Chapter 6

# Conclusion and Future Scope

This chapter presents the conclusion drawn from the results obtained after implementing the proposed approach and the future work which can be done to further improve the efficiency and security against the high-level attacks.

## 6.1 Conclusion

Data Mining attacks in cloud are major concern for the cloud provider as well as client these days. These attacks pose a great threat to the client's confidential and sensitive data. A lot of techniques exist which tries to tackle the privacy issue of cloud mining to some extent but none can solve it completely. Addition of misleading data or perturbation with noise may lead to mining failure or incorrect final results. Fragmenting the data and keeping it decentralized though prevents an adversary from having access to complete data, with an additional advantage of parallel processing but in case an authorized user wants to mine the data from distributed database an algorithm is needed which maintains privacy of intermediate results as well as the correctness and validity of the above results.

The proposed approach tries to fill the above gap by combining the Decentralization of the data along with an encryption scheme which conforms to the correctness and validity of the result while maintain the client privacy at the same time.

## 6.2 Future Scope

The above approach prevents the data leakage to an adversary if he/she intercepts the data in the middle of communication. But, as a Third Party is used to make communication possible between the two Hosts, an adversary can pose as an imposter posing as a Third

Party and can get the data from both the Hosts. To prevent this, a digital signature or hashing technique can be incorporated in the algorithm to prevent an adversary to pose as the third party.

# References


[1] M. Brantner, D. Florescu, D. Graf, D. Kossmann, and T. Kraska, "Building a database on S3." In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, pp. 251-264, 2008.

[2] J. Carolan , S. Gaede, J. Baty, G. Brunette, A. Licht, J. Remmell, L. Tucker, and J. Weise, "Introduction to cloud computing architecture." White Paper, 1st edn. Sun Micro Systems Inc (2009).

[3] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "Above the clouds: A Berkeley view of cloud computing." Dept. Electrical Eng. and Computer. Sciences, University of California, Berkeley, Rep. UCB/EECS 28 (2009): 13

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques (2nd Edition), Morgan Kaufmann, 2006.

[5] R. M. Esteves and C. Rong, “Using mahout for clustering wikipedia’s latest articles: a comparison between k-means and fuzzy c-means in the cloud.” 2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom), IEEE, pp. 565–569, 2011.

[6] A.K. Jain and RC. Dubes. “Algorithms for clustering data.” Prentice-Hall, Inc., 1988.

[7] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, and J. Molina, "Controlling data in the cloud: outsourcing computation without outsourcing control." In Proceedings of the 2009 ACM workshop on Cloud computing security, ACM, pp. 85-90., 2009.

[8] D. J. Solove, "I've got nothing to hide and other misunderstandings of privacy," San Diego L. Rev. 44 (2007): 745.

[9] P. K. Rexer, “data miner survey highlights the views of 735 dataminers” ,2010.

[10] T. Hu, H. Chen, L. Huang, and X. Zhu. "A survey of mass data mining based on cloud-computing.”, International Conference on Anti-Counterfeiting, Security and Identification (ASID),IEEE , pp. 1-4., 2012.

[11] D.Talia. "Toward Cloud-Based Big Data Analytics." 2013.

[12] G. Zhang, C. Li, Y. Zhang, and C. Xing. "DataCloud: An Efficient Massive Data Mining and Analysis Framework on Large Clusters." Ninth In Web Information Systems and Applications Conference (WISA), IEEE, pp. 198-203., 2012.

[13] M. Chen, IJ. Chiang, and CW. Lai. "Frequent pattern mining for price fluctuation based on cloud computing." 2012 IEEE International Conference In Granular Computing (GrC), IEEE , pp. 50-54, 2012.

[14] Y.Song, G. Alatorre, N. Mandagere, and A. Singh. "Storage Mining: Where IT Management Meets Big Data Analytics." 2013 IEEE International Congress on Big Data (BigData Congress), IEEE, pp. 421-422., 2013.

[15] H.Yu, Y. Lan, X. Zhang, Z. Liu, C. Yin, and C. Long. "Research of Data Mining in Cloud Environment." 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), IEEE, vol. 2, pp. 97-100., 2013.

[16] A. Osman, M. El-Refaey, and A. Elnaggar. "Towards Real-Time Analytics in the Cloud." 2013 IEEE Ninth World Congress on Services (SERVICES), IEEE, pp. 428-435, 2013.

[17] D. Apiletti, E. Baralis, T. Cerquitelli, S. Chiusano, and L. Grimaudo. "Searum: a cloud-based service for association rule mining." 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE pp. 1283-1290., 2013.

[18] L. Yu , J. Zheng, WC. Shen, B. Wu, B. Wang, L. Qian, and BR. Zhang. "Bc-pdm: Data mining, social network analysis and text mining system based on cloud computing." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1496-1499., 2012.

[19] Z. Shi, G. Jiang, B. Zhang, J. Yue, and X. Zhao. "Cross-media cloud computing." 2012 IEEE 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS), vol. 1, IEEE, pp. 365-370,2012.

[20] C. Zeng, Y. Jiang, L. Zheng, J. Li, L. Li, H. Li, C. Shen et al. "FIU-Miner: a fast, integrated, and user-friendly system for data mining in distributed environment." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1506-1509, 2013.

[21] C. Su, F. Bao, J. Zhou, T. Takagi, and K. Sakurai, "Privacy-preserving two-party k-means clustering via secure approximation." 21st International Conference on, Advanced Information Networking and Applications Workshops, 2007, AINAW'07. vol. 1, IEEE , pp. 385-391, 2007.

[22] Md. Riyazuddin , Dr.V.V.S.S.S.Balaram , Md.Afroze , Md.JaffarSadiq , M.D.Zuber. "An Empirical Study on Privacy Preserving Data Mining". International Journal of Engineering Trends and Technology (IJETT). V3(6):687-693 Nov-Dec 2012. ISSN:2231-5381

[23] K. Che, and L. Liu, "A random rotation perturbation approach to privacy preserving data classification." (2005).

[24] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification." IEEE 25th International Conference on Data Engineering, 2009. ICDE'09,IEEE, pp. 429-440, 2009.

[25] M. V. Dijk, and A. Juels, "On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing." HotSec 10 pp. 1-8, 2010.

[26] H. Dev, T. Sen, M. Basak, and M. E. Ali, "An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks." 2012 SC Companion In High Performance Computing, Networking, Storage and Analysis (SCC), IEEE, pp. 1106-1115, 2012.

[27] R.Mishra, S. K. Dash, D. P. Mishra, and A. Tripathy, "A privacy preserving repository for securing data across the cloud." 2011 3rd International Conference on Electronics Computer Technology (ICECT),IEEE , vol. 5, pp. 6-10, 2011.

[28] M. D. Singh, P. R. Krishna, and A. Saxena, "A cryptography based privacy preserving solution to mine cloud data." In Proceedings of the Third Annual ACM Bangalore Conference, ACM, pp. 14, 2010.

[29] S. Pearson, "Taking account of privacy when designing cloud computing services." In Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, IEEE Computer Society, pp. 44-52, 2009.

[30] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes." In Advances in cryptology—EUROCRYPT'99, Springer Berlin Heidelberg, pp. 223-238, 1999.

[31] K. P. Lin, and M. S. Chen, "Privacy-preserving outsourcing support vector machines with random transformation." In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 363-372, 2010.

[32] K. Beaty, A. Kundu, V. Naik, and A. Acharya. "Network-level Access Control Management for the Cloud." 2013 IEEE International Conference on Cloud Engineering (IC2E), IEEE, pp. 98-107, 2013.

[33] R. Bhadauria, and S. Sanyal. "Survey on security issues in cloud computing and associated mitigation techniques." arXiv preprint arXiv:1204.0764, 2012.

[34] H. Wu, Y. Ding, C. Winer, and L. Yao. "Network security for virtual machine in cloud computing." 2010 5th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), IEEE, pp. 18-21, 2010.

[35] I. Agudo, D. Nuñez, G. Giammatteo, P. Rizomiliotis, and C. Lambrinoudakis. "Cryptography goes to the cloud." Data Management, and Applications in Secure and Trust Computing, Springer Berlin Heidelberg, pp. 190-197, 2011.

[36] C. Tai, J. Huang, and M. Chung. "Privacy Preserving Frequent Pattern Mining on Multi-cloud Environment." 2013 International Symposium on Biometrics and Security Technologies (ISBAST), IEEE, pp. 235-240, 2013.

[37] ASA. Ansari, and KK. Devadkar. "Secure cloud mining." 2012 IEEE International Conference on Computational Intelligence & Computing Research (ICCIC),IEEE, pp. 1-4, 2012.

[38] Q. Lu, Y. Xiong, X. Gong, and W. Huang. "Secure collaborative outsourced data mining with multi-owner in cloud computing." 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) , IEEE, pp. 100-108, 2012.

[39] R. M. Esteves, R. Pais, and C. Rong. “K-means clustering in the cloud–a mahout test.”, 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications (WAINA), IEEE, pages 514–519, 2011.

[40] Chen, Tung-Shou, Tzu-Hsin Tsai, Yi-Tzu Chen, Chin-Chiang Lin, Rong-Chang Chen, Shuan-Yow Li, and Hsin-Yi Chen. "A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray." 2005 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2005, IEEE, pp. 405-408, 2005.

[41] R. Bhadauria, R. Borgohain, A. Biswas and S. Sanyal. "Secure Authentication of Cloud Data Mining API " arXiv preprint arXiv:1204.0764, 2012.

[42] S. Liu and Y. Cheng. “Research on k- means algorithm based on cloud computing.” 2012 International Conference on Computer Science & Service System(CSSS), IEEE, pp. 1762–1765, 2012.

[43] T. R. G. Nair and K. L. Madhuri. “Data mining using hierarchical virtual k-means approach integrating data fragments in cloud computing environment.” 2011 IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, pp. 230–234, 2011.

[44] M. Modak, M. Aslam, and M. Vijayalakshmi. "Privacy Preserving Data Mining Techniques In The Cloud: A Comparative Analysis."

[45] S. Owen, A. Robin, T. Dunning, and E. Friedman. Mahout in Action. Manning Publications, 2012.

[46] http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster

[47] http://harish11g.blogspot.in/2012/02/configuring-mahout-clustering-hadoop.html

[48] http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-computing-v26.ppt

[49] http://archive.ics.uci.edu/ml/databases

## List of Publications

1. Deepti Mittal, Damandeep Kaur and Ashish Aggarwal, “Secure Data Mining in Cloud using Homomorphic Encryption”, Third IEEE International Conference on Cloud Computing for Emerging Markets (CCEM), IEEE, 2014.[Communicated]
2. Deepti Mittal, Damandeep Kaur and Ashish Aggarwal, “k-means and its Variants in Cloud Data Mining: A Comparative Analysis”, 5th International Conference-Confluence 2014 on Cloud Security and Big Data”, IEEE, 2014.[Communicated]