A

Major Project-II Report

On

PRIVACY PRESERVATION USING AES ALGORITHM IN HADOOP ENVIRONMENT

Submitted in Partial Fulfillment of the Requirement for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

MAHESH KUMAR

2K13/CSE/09

Under the Esteemed guidance of

Mr. Manoj Sethi

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahabad Daulatpur, Main Bawana Road,

Delhi-110042.

JUNE, 2015

CERTIFICATE

This is to certify that Major Project-II Report entitled **"Privacy Preservation Using AES Algorithm in Hadoop Environment"** submitted by **Mahesh Kumar, Roll No. 2K13/CSE/09** for partial fulfillment of the requirement for the award of degree Master of Technology (Computer Science and Engineering) is a record of the candidate work carried out by him under my supervision.

Mr. Manoj Sethi
Project Guide
Department Of Computer Science & Engineering
Delhi Technological University

DECLARATION

We hereby declare that the major Project-I1 work entitled "**Privacy Preservation Using AES Algorithm In Hadoop Environment**" which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of degree of Master Of Technology(Computer Science and Engineering) is a bonafide report of Major Project-1I carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Mahesh Kumar

2K13/CSE/09

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Mr. Manoj Sethi for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to Dr. O.P.Verma, HOD, Computer Science & Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Mahesh Kumar

University Roll no: 2K13/CSE/09

M.Tech (Computer Science & Engineering)

Department of Computer Engineering

Delhi Technological University

Delhi – 110042

ABSTRACT

Many government agencies and businesses are focusing on security issues because of cybercrime.

If we talk about Facebook which maintains large data for every user account but provides little privacy since privacy is very essential factor nowadays. Security helps in gain of the user's interest and increase the growth of business.

Privacy-Preservation is an important feature for any individual since his personal data should not be accessible to unauthorized people. In recent years, data mining has been viewed as a threat to privacy because wide range of data is spread in digital form by many companies. The data is stored in digital form in the computer. So some privacy should be preserved during data collection and data mining. Nowadays there are so many attackers attacking our data. That's why privacy is so important factor that reduces the chances of attacks and increase the security. In this project, we proposed a method for processing the data file in encrypted form that is unreadable by unauthorized user and reduces the encryption and decryption time as compared to AES. This research also shows the results that have been practically proved by the implementation of existing algorithm and proposed algorithm. In proposed algorithm, we can execute the dataset that can be of size greater than computer RAM which cannot be done in AES. To give an approach for HDFS environment, useful for the improvement of HDFS Hadoop while dealing with big data, Privacy-Preservation should be our main concern. This Means, in HDFS the data is in plain form, which can be accessed and misused by anyone. So data should be in encrypted form to provide the privacy because even if a small piece of information is leaked by attacker, then it becomes difficult to handle data of TBs. So this approach for HDFS environment can be used to preserve the privacy. Here I have also compared the results with different kinds of parameters. The approach used here breaks the file into blocks and then block level encryption is used. The main objective here is to reduce the time of encryption and preserve the privacy to increase the security in HDFS environment.

Keywords: Privacy Preservation, SSH Shell, AES, Hadoop, HDFS, Data Mining, Big Data, Map Reduce, Encryption, Decryption.

List of Figures

List of Abbreviations

AES	Advance Encryption Standard
HDFS	Hadoop Distributed File System
PPDM	Privacy Preservation Data Mining
SSH	Secure Shell

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1. DATA MINING

Data Mining is what, it discovers the knowledge about the large relational databases which is coming from different sources and summarize into a useful information to improve the performance of the business and cut costs, revenue and both.

Data mining software is of the tool to analyzing the data. It provides an interface to allow users to identify the data from different dimensions or angles. Users can also divide the data and can specify the relationship between the data.

Data mining is a group of technologies that find the relationship which have not previously discovered. Data mining is attractive field of industry to keep the information in well manner way. [1]. Due to very large or wide availability of data imminent requirements, data takes place into very useful information and knowledge [1].

This information and knowledge can be used in various applications of real world to improve the business, based on the past scenario. This can be also used for various kinds of applications ranging from market based analysis, loss detection, fraud detection and customer holding. [1].

Due to wide range of data there privacy threats arise. So industry takes care of this type of threats. That's why privacy preservation takes place. [1]. so privacy is associated as an important component with data mining for achieved information and knowledge.

1.2. INFORMATION

Information is a set of the relationships, patterns, and associations between all these data is called information.

Example:-Find the details of transactions of store corresponding which product is sell and when.

1.3. DATA

Data can be anything, facts or texts that will process by the computer machine. Todays, the data of many organizations is growing in different formats and different databases and become a big data because users of internet are increasing day by day and read the data from internet organizations and write the data in much more quantity on the internet organizations. Three types of data-

- Operational or Transactional Data: contains the data of sales, cost, inventory, accounting data, payroll data.
- Meta-Data:-Meta data is a type of the data that describes itself. It is self-describing in nature.
- Non-Operational Data:-This type of the data includes data such as industry sales, economic data, and forecast data.

1.4. KNOWLEDGE

Information is just converted into knowledge when it useful for something. Suppose a bank contains many accounts in bank but want to information which one have ATM or not. Now this information treats as knowledge.

1.5. BIG DATA

The term 'Big Data' first shown on silicon graphics (SGI) slide deck in 1998 by John Massey having title "Big Data and the Next Wave of InfraStress". But Weiss and Indrukya refer a name "Big Data" which is nothing, data mining book [24]. Data mining was much related from starting. Nowadays the data is producing by the very high rate and data is growing with more than 40% every year and is the in size of zeta bytes [24].

Big data following are the characteristics:

1. Velocity
2. Volume
3. Variety
4. Veracity
5. Variability
6. Complexity

The main conferences which contribute in the area of big data and big data mining & Privacy-Preservation are KDD, ICDM, ECML,IEEE and journals like "Data mining and knowledge Discovery" or "Machine Learning".

Actually Big Data is the very wide term for data which is large in size, complex that cannot be processed by classical methods or processing. Information privacy, visualization sharing, storage, data duration and capturing data are the main challenges of the Big Data [24]. Big data means, it is a related to very large or very complex data with very big data because data of some companies are increasing day by day. That's why the data of every company or some company is becoming big data. When data is in very large quantity then it is come under the concept of big data and from this the term big data is introduced [24].

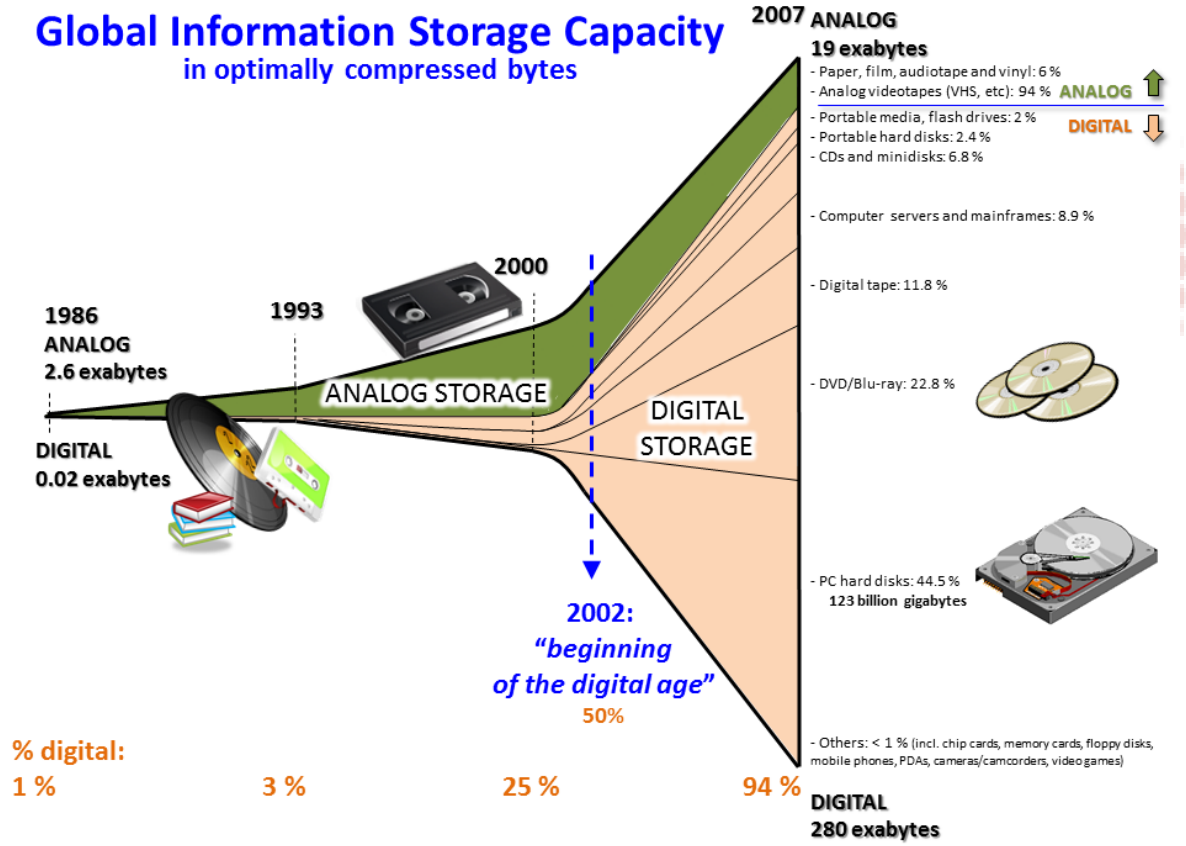


Figure -1 [24].

1.6. PRIVACY-PRESERVATION

1.6.1. PRIVACY

Privacy is what, it means this is a condition that prevent the very sensitive information from illegal use by others.

In nowadays, there are so more data is transferred between many parties then privacy is a very important factor to prevent the information from unauthorized users.

In real world, data is increasing day by day, that's why privacy is the very interesting topic to protect the data.

Privacy is very important when people are participating in competitive situation to show the interest. [1].Privacy is that has an ability to maintain the relationships between different types of social site people. [1].

1.6.2. PRESERVATION

Preservation is what, it means that preservation is the very important functionality that preserves or provides legal access to authorized users to access information.

This plays a very important role in case if we want to provide the privacy about our data. That will preserve some rights to authorized users, they can access the data on the basis of these rights.

Example:-If we have library management system in our college, only 4 books are allowed to a particular student. When a book is issued for a student then a counter for book count for a particular student is increased by one by library clerk because he has only right to access this counter. Now an attacker attacks the system and increases the counter for particular student. Then it is very bad for that student & preservation is important factor to preserve the access rights.

1.6.3. NEED FOR PRIVACY PRESERVATION

Privacy-Preservation is most of important feature to a particular since he should not embarrassed by unauthorized people. [1]. Privacy Preservation is very important aspect of data mining to maintain the privacy through many of various methods. [1]. Privacy Preservation is to protect very sensitive data associated with individual. [1]. in recent years, data mining that has been viewed as a threat to privacy because of wide range of data is spread in digital form by many companies.

Privacy preservation is considered as an important factor for very effective utilization of the sensitive massive volume of data. The data is stored in digital form in the computer. So some privacy should be preserve during data collection and data mining.

Actually in nowadays there are so many attackers to attack on our data. That's why privacy is so important factor to preserves some privacy to the data that will reduce the chances of attacks and increase the security. Privacy, data integrity, Security and are considered as challenging problems in data mining [1].

INTERCEPTION:

- When an Unauthorized user has achieved access to a service or information or data.

INTERRUPTION:

- Services or data become unavailable.
- File is corrupted or lost.
- Denial of service attacks.

MODIFICATION:

- It is what, when Unauthorized changing of data or tampering with a service so that it no longer available to its original features.
- Changing transmitted data tampering with database entries.
- Changing a program.

FABRICATION:

Additional data or activity are generated that would normally not exist,

- When an attacker, add an some wrong entry into a file of password or database
- Replaying previously sent messages

1.7. PRIVACY PRESERVATION TECHNIQUES

Defines actions of entities that are allowed or prohibited.

Security mechanisms are:

1. **Encryption:** Transforms data into something an attacker cannot understand. So data integrity and confidentiality retains.
2. **Authentication:** Identity of identifier.

3. **Authorization:** Permission to access.
4. **Auditing:** (trace which clients accessed what, and which way) – analysis of security breach.

1.8. APPLICATIONS OF PRIVACY-PRESERVATION AND DATA MINING

GOVERNMENT

Many government bodies are working on data mining and after data mining, these bodies try to preserve some privacy on mining data because to protect from illegal access of data mining data.

Example: Suppose if government bodies do no provide privacy on data.

If an election is happening then someone occurs across the very sensitive data of last year's predictions and theft the data. Then can you think what will be happen?

That's why government bodies are using privacy preservation to protect the data mining data (Very sensitive data).

INTERNET

Like Facebook.com, eBay.com, Amazon.com are handling big data in peta bytes, as on 2012 Google was handling about 100 billion of queries per month. So they are using privacy-preservation techniques to hiding the very sensitive information from third party.

If we are talking about the field of retail, Wal-Mart is handles 1 millions of transactions per hour which have size 2.5 PB. Now again about in the area of banking the data is going to double every 1.2 years which is according to [1]. And last is about in the area of Real Estate Windermere real Estate achieving 100 million driver's GPS data in a day. So at last there is so much need of privacy-preservation.

1.9. MINING AND APPROACHES

If we are talking about the data mining process, the mining process requires high computational power and analyzing power. So we require a well-defined platform that can access the data with effectiveness and must have a system with very high computational power. If we have a data which is big the single system cannot handle the mining of the big data. So that to solve this type of problem of mining the big data researchers and privacy-preservation researchers have proposed so many algorithms as well as approaches. These approaches contain mainly distributive computing or parallel or on grid, cluster (which

Contains so many computers according to the requirement of the problem) or Map Reduce framework (which is used to parallel processing).

Different types of people introduced different types of approaches to handle the privacy preservation of data or big data computation or big data mining using various types of tools and methods which are based on according to demand of problem nature. But here we are interested in a programming that will proof the privacy preservation using existing algorithms and my algorithm, which comes under normal approach or Hadoop environment containing a Hadoop toolkit which is under comes of Hadoop common, Hadoop distributed file system (HDFS), and Map-Reduce Framework [27].

AES is an attractive approach because it provides a model with simple way through which, leading to significant interest in the educational community.

Frequent item set mining which is base step for Association rules mining. After find out frequent item set mining algorithm like A-Priori [25] or FP-Growth on dataset, you can find out frequent item sets which satisfied the minimum support count criteria. And now after found these frequent item sets, we can easily find out the associations rules which is based on mathematics extension**.**

Association rule Mining concept which was introduced by R. Agrawal in 1993 [25], which goals to find user interesting correlations, frequent patterns, Association rules and a structure among the set of items.

In last 2 and 3 decagons, in the area of association rules mining or frequent item set mining so many types of efficient algorithms have been proposed and developed for the association rules mining. These algorithms are defined based on their categories into 3 types frequent pattern growth based, Apriori based and Vertical database format based.

Apriori [25] algorithm is the first classical algorithm or simple algorithm for mining frequent patterns and generating association rules. It is the very most popular algorithm which is proposed in 1994. In this algorithm there are multiple scans of the database, to generating large number of candidate items which are the main challenging issues of this algorithm. Apriori which was more effective full during the candidate generation process because it used a unique type of method called Apriori_gen () and used new pruning techniques which is different from AIS. AIS is a method which is based on the concept of straight forward approach that require so many passes over the database and still generate too many candidate item sets [25].

Reverse-Apriori [25] which was introduced by Kamrul, Shah and Mohammad in 2008. In this algorithm all approaches are reversed of classical A-Priori Algorithm. Classical Apriori [25] which is based on concept of first, it finds candidate-1, then used these candidate sets to find frequent-1 by pruning candidate-1 item sets. The pruned item set is nothing, those item set which do not satisfy minimum support. Now it generates candidate-2 item sets which are generated from frequent-1 item sets by using a method named as Apriori_gen () method. Now again this process will repeat in same above way to calculate all frequent item sets to find out all association rules. This process will continues till large frequent item set is generated. But you know what, the Reverse-Apriori [25] algorithm is what, can say totally different. Initially, it will generate large frequent item sets by consider all the largest possible number of items in the dataset. And then it will generate large frequent item sets only, you know when, after these satisfies minimum support. Until all the largest frequent item sets, it will decrease the largest possible number of frequent item set and repeat the process in above same way.

1.10. MOTIVATION

If we look at last some years then we can released that there are so many witnesses are existed for dramatic growth in data day by day and look at an ability to manage and handle the data in well-defined manner. There are so many resources to incoming and outgoing the data. So that risk is introduced because there are many security issues are arising day by day. You know why, When data is traveled through a medium channel, then there is no techniques to encrypt the data then you know there are so many attackers are sit alongside your very important and sensitive information to capture the details about your information. If assumes that our data is going on through a secure medium then after getting the data on your system. You cannot say that your data is secure. Attackers can attack on your data in your system through internet or so many techniques. So this time is what, this time is going on to protect your data everywhere and preserve some privacy to access the data which can reduce the chances of illegal access. So that there are so many areas to work in that field. That's why my interest is motivated in that field. As you can see that I am going to discuss some real life problem or can say some real world problems. Which are discussed below.

We have so many sources to getting data like sensors, devices in different and independent formats through various applications and there are sources like structured or unstructured

E.g.-Mail, Yahoo. Although this data has been or has exceeds our ability to handle, mange, process, store and analyze. So many issues arise again. For example, Google indexed itself one millions of pages in 1998, you can see this information on internet and after some time which is increased in a fast way, arrived one billions in 2000 and again Google is indexed one trillions of pages in 2008. This quickly growth of data is also apply to Facebook, Twitter and other websites like Flip kart, Amazon, etc.

For this, so much data available and exponential growth of data day by day, it is not easy to just handle these data and store these information. But another issue is that to manage the data in appropriate form. So, this is the motivation field to work because it attracts the interest to do work. This data is managed a form that will provide some privacy and this data is used to find the knowledge which can be used as suitable for our business by taking support in decisions. We can used this data to and after that find many kinds of attractive or interesting information and patterns, One of the attractive thing is that association rules and after that to provide some privacy to informative data, so that provide some privacy that will preserve some right or reduce the risk of third party.

For example: - If we are taking the example of our library system. Then there is a rule to maximum books limit to a student 6. 3 from book bank and 3 from other department. So there must a counter to count the number of books for a particular student. To access this counter account, someone will be have rights to access this counter to increase/decrease according the student status. If a suppose has been issued two books, means counter should be set on 2 by authorized users. Everything is going on in right way. In between an attacker/third party occur and changes the value of this counter to 6, then it means, this student cannot be able to issue next book because authorized authority is unknown by the event which is happed by attacker. Authority assumes that counter is already set 6. That's why no books are allowed to this student. So that student is only responsible for 6-2=4 books which is not issued by this student. So this student will pay fine of 4 books. That's why privacy is lost. The main goal is what yon know, provide preserve some privacy that will reduce these kinds of attacks.

1.11. RESEARCH OBJECTIVE

Motivation as we discussed already in previous section the objective of this research is solving a problem that will reduce the time complexity than existing algorithm-AES.

Another objective is what, optimization is necessary to solve any problem. Means this research will provide optimization over the existing problem way.

In this research, we will take an authorized dataset of transactions from authorized party and will show how many benefits of my algorithm over existing algorithm. Another objective of this research how much data can processed by my algorithm over existing algorithm.

In below figure 1.2, [2], a classical scenario is described for data collection and publishing is described [2]. In this figure there are two phases, Data Collection phase that is used to a publisher that will collect the data from record owners (Alice and Bob). In the second phase, Data Publishing phase, a data publisher release the data which is collected by publisher to an entity that is a miner or to the another entity like public, which is called by a name data recipient, do you data recipient is what, who is an entity that is responsible for conduct data mining process on collected or published data. For Example: - if we are taking example of a hospital then in this process hospital collects the data from their patients' and do publish these patients' records data to an entity which is an external entity (Medical Center). In this do you who is who? Who is the publisher, hospital is the publisher. Who is the owners, patients records are the records of owners. Who is data recipient, Medical center is what, and it is a data recipient. Now data mining takes place at medical center. In that case, mining can be anything like count the number of men with fever to practiced cluster analysis [2].

Another goal of this research to reduce the time of encryption and decryption using proposed algorithm. The scope is limited to data encryption using combined method Advance Encryption Standard and proposed algorithm. Main objective of thesis is reduce the time of encryption decryption ratio. Algorithm is related the block are break into block having size of 128 bits. Encryption of data using XOR operation and encrypt the block with AES encryption Algorithm.

As I discussed below, what I have done in my research:-

1. To make new encryption approach for data/big data and testing of encryption and decryption and after that evaluate it.
2. To determine the result and compare different the results at various sizes of the dataset with respect of time with a new approach using AES algorithm and Proposed Algorithm.
3. Analysis the various level of the security at encryption level of big data by proposed algorithm and Advance Encryption Standard (AES).

4. Compute the time for encryption decryption and compare the time between AES Algorithm and Proposed Algorithm.

Privacy-Preserving Data Publishing

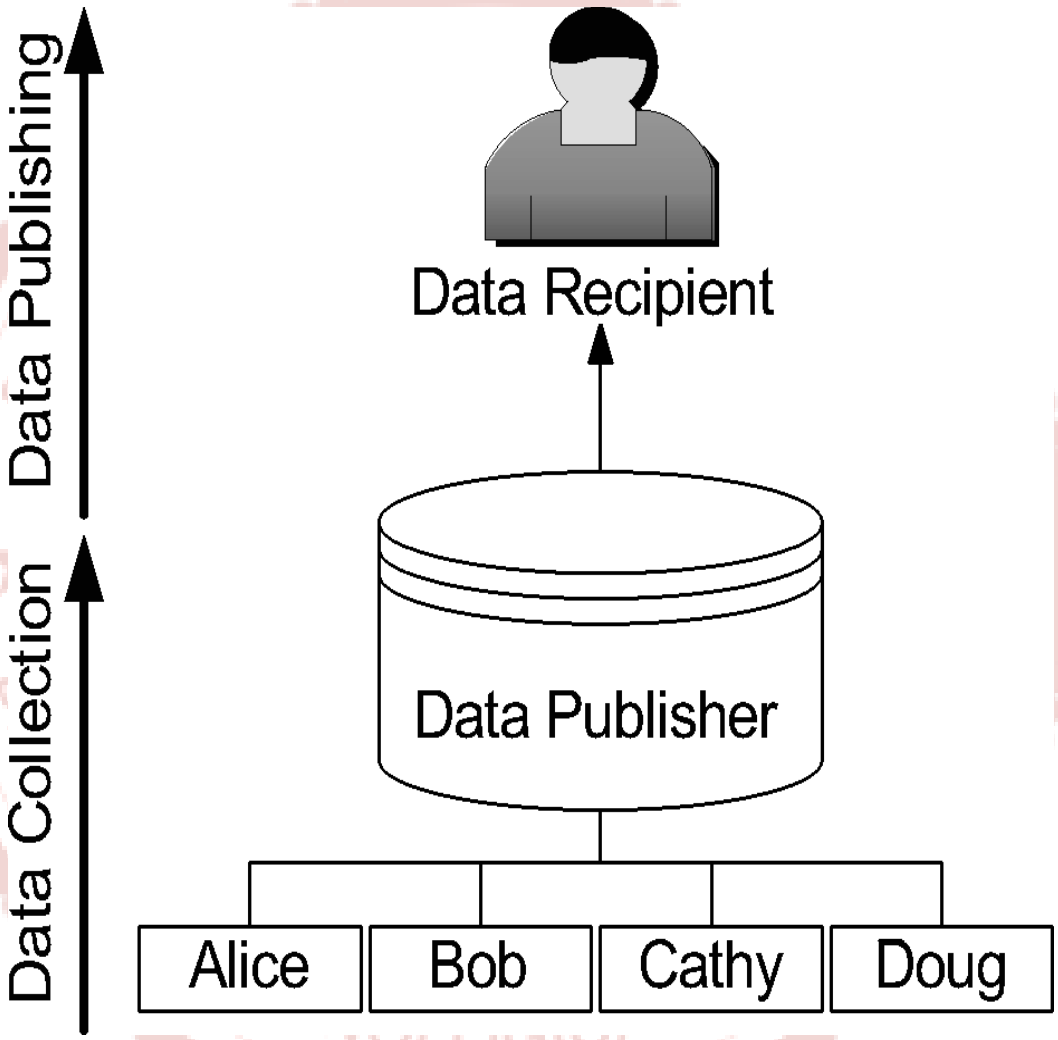


Figure -2 [2]. Data collection and data publishing.

1.12. REPORT ORGANISATION

I started this report with an introduction by a name called introduction in chapter 1. A description with full of details and background which is described in a section called literature review and the description of all related research work and algorithms in chapter 2. In this section all the descriptions are discussed like related work, what is data mining, how data mining is processed, what are algorithms to process it. What is big data and procedure to process the big data like how to manage, how to store, how to handle it. What is frequent item sets, what is support, what is the confidence? What is existing algorithms for privacy-preservation, what is the privacy preservation? What is going in the field of privacy-preservation for data mining? Encryption and decryption of transaction datasets/databases using different kinds of approaches. A details description of Hadoop is described in chapter 3. Privacy-Preservation in Hadoop environment. Privacy-Preservation for data mining. The details description of SSH shell is defined in the chapter 4. Existing work is also defined in this section in chapter 5. There is an proposed work section which is defined in chapter 5.There is a one more section which is called chapter 6 with all the comparisons and results, what is differences between performance of proposed algorithm and existing algorithm. In chapter 7 is described with, what is conclusion and future work.

CHAPTER 2

LITERATURE REVIEW

2.1. PRIVACY-PRESERVATION HISTORY

Privacy-Preservation is becoming a demand of today because data of social site or social networks is growing by an exponential rate. That's why it is in demand due to its popularity. When we are dealing with cloud computing or web 2.0 or big data. In these fields of research it taking a vast place to do work more. We will focused on several techniques of privacy-preservation. It is the main concern of nowadays. We will noticed new challenging issues in the field of privacy preservation compare to existing scenario of the privacy-preservation of social data or private data. And we will identify the three things in this paper names privacy, data utility and background knowledge. In this paper, a survey is based on two categories: Clustering based approaches and graph modification based approaches.

In graph based approach, there are two things which are used in this approach are Edge and Vertices. When we are processing the big data then it is divided into a forms of graph where a node of graph is representing a computer in this network. Do you know, why we are processing the data through a graph because these data is very large in respect of the size? That's why this cannot be processed by a single computer. So we have need to process these data through a graph.

Actually due to instantly popularity of social sites on the web. You know what this means, it means data is becoming big data because users are increasing day by day on the social sites. So that privacy-preservation is a concerned issue. According a survey of by TNS Canadian Facts, a Canadian marketing a firm of social site, teens, and so many young adults are the addicted users of social networking sites. So that privacy preservation is taking the place in nowadays [3]. In background knowledge there are so many attributes which are considered as a major types of privacy attacks is what, it is to identify the a particular entity by joining the table which contains the publish data with some external modeling of the table is basic knowledge of the users. In privacy-preservation with respect of publishing the social networks on the web, there created an issue of complex nature of structure of the graph data. In this paper, there exist an entity which is called adversaries can be modeled in so many ways:-

- **Identifying attributes of vertices: -** In this model, a vertex is connected uniquely to an individual through a set of attributes. There a set of attributes will play a role to a quasi-identifier in a process of re-identification attacks on the data. In this approach vertex attributes are used to design a model in which label will be used to indicate the attribute in a social networks. An adversary can be know the some attributes and corresponding all their values. This kinds of background knowledge can be an abused for privacy attacks [3].
- **Vertex degrees: -** Vertex degree is nothing, it is number of edges which are incident on that vertex is recall by a name as vertex degree [3].

Privacy	Utility	Background knowledge					
		Identifying attributes of vertices	Vertex degrees	Link relationship	Neighborhoods	Embedded subgraphs	Graph metrics
Vertex existence	General graph properties						
	Aggregate network queries						
Vertex properties	General graph properties		[12; 13; 18; 36]		[12; 13; 36]	[3]	[12]
	Aggregate network queries				[39; 38]		
Sensitive vertex labels	General graph properties	[4]		[4; 37]			
	Aggregate network queries				[38]		
Link relationship	General graph properties						
	Aggregate network queries			[5]			
Link weight	General graph properties						[19]
	Aggregate network queries						
Sensitive edge labels	General graph properties			[4; 37]			
	Aggregate network queries						
Graph metrics	General graph properties						
	Aggregate network queries						

Figure-3 [3]. Graph

[1] Data mining is multi dimension approach to gather useful information by extracting the databases with large in the size. If we look at recent years, then we can say that exchange process of data and a process of publishing the data has been very common for their wealth of

opportunities. Security, privacy, integrity is very challenging issues that is arising in recent years. Do you know, privacy is going to main necessary need to protect the user's area of interest in so many conditions? Privacy is a process to provide the ability that is used to create and maintain different kinds of social relationships with the people. Privacy-Preservation is the one of important factor that is for an individual since someone is not embarrassed by some adversary. Do you know that privacy-preservation is the one of important aspect of data mining that will ensure the privacy by different kinds of methods or models or algorithms? Privacy-Preservation is what, it is a process to protect the information which is associated with an individual. In this paper, a survey is discussed to success and an approach that will not distracted the privacy of an individual [1].

2.2. DATA COLLECTION AND DATA MINING

In this section, which is based on an era. This era contains how the data is collected and how all the transactions which contains in that data are recorded somewhere. In the field of research, so many techniques are enhanced till now and have been developed which were totally different with each other to secure the individual's privacy while collecting data and mining data. If we look at present time of people, there is so many demands to exchange and the process of publish the data around various type of parties after the collecting process of data is done. The sources of collecting the data may be or may not be different and data can be or can structured or unstructured. All sources of collecting the data which cannot be shared by directly. The owner of the data that is an entity will collect the data. For collecting the data, there are so many techniques that are using in nowadays to collecting the data from different kinds of sources. Now again, we can say that there are so many techniques that are used to mine the data and after that owner of the data release the data to data recipient.
Do you know, all collected data is prepared in manner which is based on the well-manner way. After the preparation of the data, now data is ready to provide the some privacy and security, confidentiality before publishing the data [1].

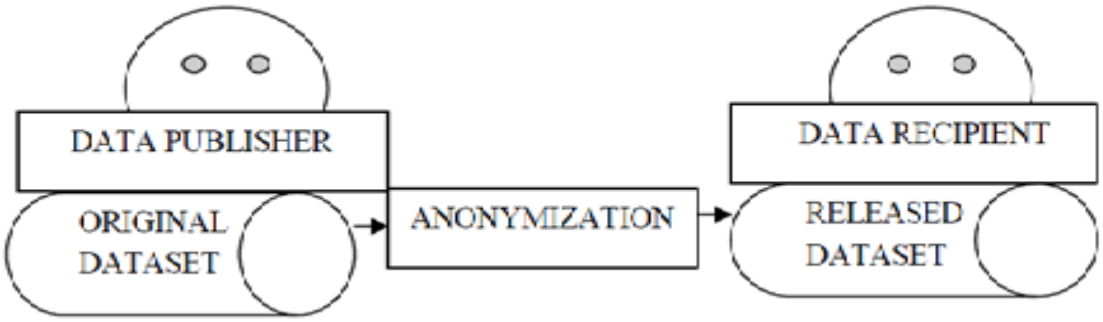


Figure-4 [1]. A Simple Privacy Model

2.3. PRIVACY PRESERVING APPROACHES

According a definition of Sweeny [1], the personal information about someone is stored in a form of table of rows and columns. Several processes are there which are involved in preserve the privacy in data which will make effective process with respect to security or privacy. There are so many methods are exist to solve the problem of privacy, according the nature of application. As I have been seen many approaches which is already discussed in this paper to provide the privacy through these approaches names such as sampling, encoding, randomization, cell suppression, data swapping, and one more name as perturbation that have been already designed for micro data. If data is in a size with large quantity then we can use the cryptography algorithms to provide the privacy that is used to preserve the privacy. Privacy-Preservation can take place any stage of data transmission as well as at the stage of storage. Due to some lackness of existing algorithms, a thought or implementation is the main field for research [1].

2.3.1. RANDOMIZATION

Randomization was a process to preserve the privacy at micro data level. This approach can be used. Randomization has an ability to hide the entire datasets in form of unreadable by the attacker. This can be used to provide to preserve some semantics. If we talk about all the techniques which were already existing, the randomization could be the first and good technique. This approach can be used to provide discovery of knowledge and provides a balance among utility and privacy [1]. In this approach, balance is calculated through by adding the noisy data to the original data. After mixing of randomized data is called balance data.

So randomized data after the balance is achieved, this data is transferred to recipient. The recipient of data has an algorithm which is used by recipient to receive the data. To receive the data through an algorithm is called distribution reconstruction [1].

In this algorithm, some noise data is added to the original data to some fields of the mask of the records. Noise data which is large in the size, so that individual values of the records that cannot be recovered. This is two steps process to implement it. Which are, I am going to discuss below: -

1. An entity that is called data providers which is responsible for randomize their data and transferred this data which is randomized data to the data receiver.
2. Data receiver examines the actual distribution of the original data by using an algorithm which is distribution reconstruction algorithm.

2.3.2. SUPPRESSION

Suppression in which releasing of the actual values is not involved. In this, this algorithm will be used to replace the value of specific attributes which is associated with this attribute to describe it. Actually there is an identifier name Quasi-Identifier that is used as an attribute, with a less description. The name of suppression comes under a process because it hide some details of Quasi-Identifier. In this algorithm a particular value that is replaced though a generic value at the time of suppression. Suppression works as replaced values that are not shared to anyone [1].

2.3.3. ENCRPYTION METHOD

If we are talking about cryptography-based methods, that provide high guarantees for privacy preservation. Encryption algorithm that can be used to resolve the [4] many problems in which people interact with each other and conduct mining tasks, that are under come on the private/secret inputs as they provide. These mining tasks occurred across two competitors or we can say that between untrusted/unknown parties. So privacy-preserving approach requires to secure or provide some privacy based on the nature of parameters of the data. So there are different kinds of PPDM techniques are available for vertically partitioned data and for horizontally partitioned data. Encryption method that ensures the transmission of data should be in secure and exact, but sometimes these methods do not suitable for real life problems.

These kinds of methods are using for data with very large in the size. These methods provide so many methods according the nature of problem. In cryptography there are special kinds of methods that are not easy to break them such as AES, DES and so on [4].

2.4. APPLICATIONS OF PRIVACY PRESERVATIONS

2.4.1. MEDICAL DATABASE : SCRUB SYSTEM

According the perception of author of this paper, if we look at medical database or medical system, then according to the author of this paper, all information is placed in a special form of text. This data includes the information about the patients of that hospital such as address of patients, phone number of the patients, family members of the patients and blood group of the patients. If we discussed as traditional technology that is only responsible for global search and replace all the rules that are used to maintain the privacy. As we can see in this paper according the author of this paper (Sweeny L) found, a Scrub system that is used an algorithm name numerous detection algorithms that is responsible for maintain a order that is used to preserve the privacy [1].

2.4.2. BIO-TERRORISM

This is very essential factor to analysis and summarize the data of medical for privacy-preservation in an area of an application of bio-terrorism. For Example, Agents of Biological are normally, we can found in the environment of the natural like as anthrax [1]. It is most important to find out the attack name anthrax from a prediction of the normal attack [5]. It is Very necessary to trace the all incidences of the diseases with common symterms. The corresponding data that is reported to the agencies of public health. The respiratory diseases that on which reporting factor will not be apply. So, this gives a good solution for most identify able information which is in accordance with the public health law [1].

2.5. COMPARISIONS BETWEEN SOME APPROACHES OF PRIVACY PRESERVATION

In this part of section, we are going to compare some techniques which were already used in privacy-preservation. These are the conceptual description of the techniques. In below figure we are just filtering the techniques according their capability and quality and according the nature of problem. When, Why, What technique is used. Means suitable according the user

perception and user's interest that makes the interest of user will attractive. So you can see below figure which was discussed in a paper [1].

Conceptual Study in following figure

Techniques	Dataset	Parameter used	Advantages	Disadvantages
K-Anonymity	Market Basket Dataset	Number of data points, Dimensionality of data space	High correlation among the tuples	More Number of dimensions would be violated
ℓ-Diversity	Adult Database	Identifiers, Quasi-identifiers, Sensitive attribute	Sensitive attribute would have at most same frequency	Homogeneity and background knowledge attack has lacked
t-closeness	Pension scheme dataset	Identifiers, Quasi-identifiers, Sensitive attribute	Measure the distance between two probabilistic distribution that were indistinguishable from one another	Information gain was unclear
K^m Anonymity	Market Basket Dataset	Distinct items, Maximum transaction size and Average transaction size on distinct items	Similar evaluated approach on k items	Loss of utility
WFC	Iris, Wine, Zoo Datasets	Single, Complete and Average link	Partition the records into equivalence classes	Utility was still not achieved
Distributed K-Anonymity framework (DKA)	Employee Dataset	Public-key, Secret-key, Encryption	Global Anonymization to ensure privacy	Utility and potential were misused
R-U Confidentiality Map	Click Stream data	Maximum transaction size, Average transaction size	Maintain trade-off between privacy and utility	Vulnerable to homogeneity attack
Slicing	Health care Dataset	Identifier, Quasi-Identifier, Sensitive Attribute	Randomization on sensitive attribute	Utility and risk measures not matched
Overlapped Slicing	Health care Dataset	Identifier, Quasi-Identifier, Sensitive Attribute	Duplicate an attribute in more than one columns	Utility was not achieved

Figure-5 [1].

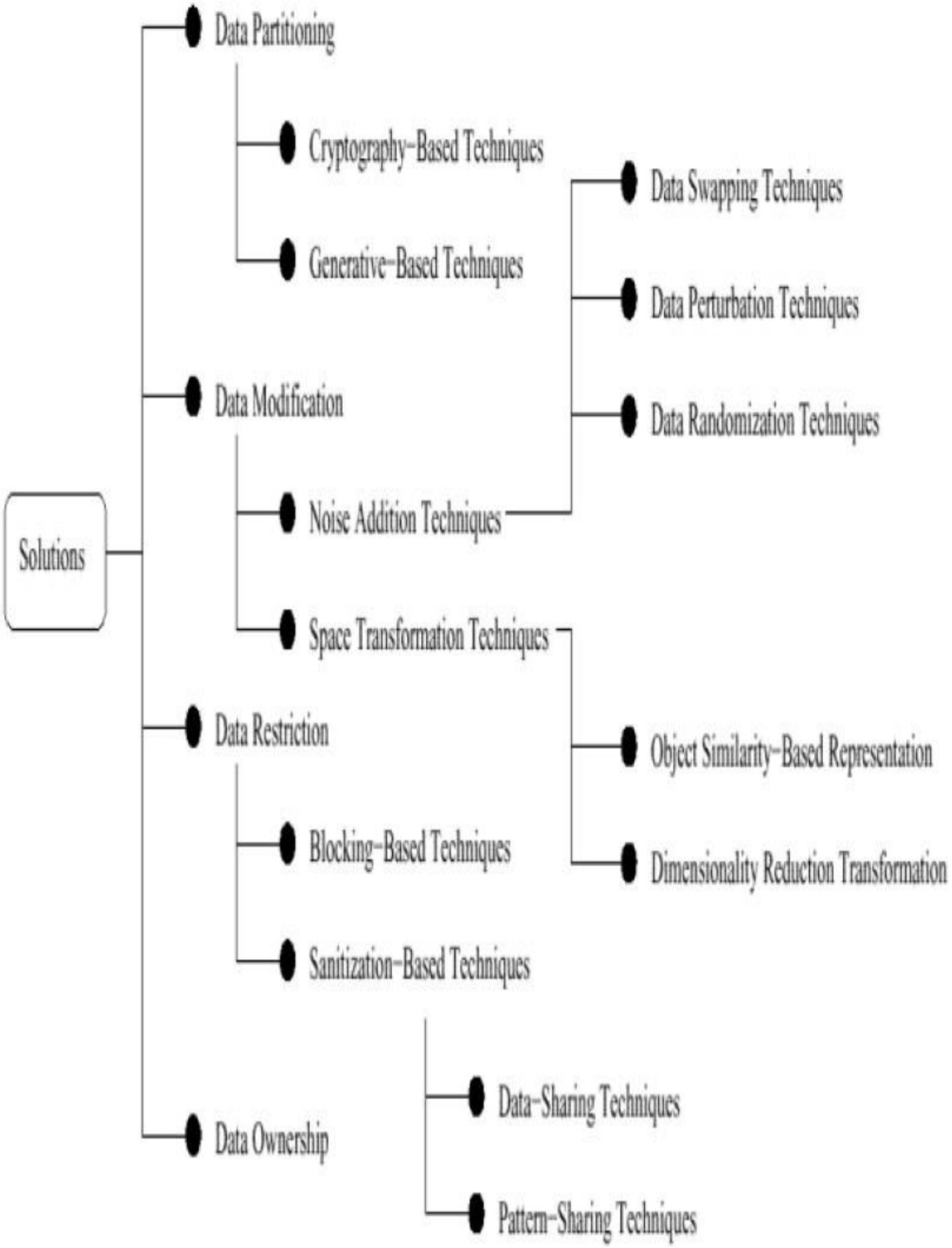


Figure-6 [6]. A Taxonomy of PPDM Techniques

2.6. PARAMETERS THAT IS IDENTIFIED ON WHICH PPDM ALGORITHMS IS EVALUATED

A set of parameters which can be evaluated by a PPDM algorithm. As these are discussed in

1. Privacy level- how much privacy level is achieved.
2. Hiding Failure- An algorithm should be robust, that has power to handle a failure occurred during hiding process.
3. Data Quality-Algorithm should be responsible to give the original quality of data.
4. Complexity-An algorithm should be optimized that takes minimum space and time complexity with a better or optimize solution [6].

2.7. PRIVACY-VIOLATION IN DATA MINING

Do you know before understanding of privacy in data mining, you must have understanding when is privacy property is violating and what is effective meaning of privacy violation? Do you know, most important factor of privacy violation in the data mining is what, it is the misuse of the data? There are many ways in real life world problems, when privacy is violating, due to so many reasons, according the various kinds of privacy with different kinds of intentions. When a particular user have access to an application but due to some privacy reasons, means middle man attacks the system and information is not available to authorized user. Means privacy violation takes place at that time. If there are some algorithm which should be used by this application, then may be, there so many complications occurs to attack on that application. If we are talking about banking sector, means an attacker attacks the banking system and do miss-behave on all account which opened on this bank. Then do you know what will be happen? Privacy violation takes place again. So in above same way we can say what is privacy, when privacy violation will occur, and think about every real world application, where privacy is violated. Then you released that so many applications in this field to do research to find out the applications and apply the some algorithm that provide some benefits and increase the security that will always attract so many users, why because it time, data is growing by exponential rate. That's why user's interest is increasing day by day. When we think about real world, then we examine in our mind, when, where and how, privacy violation is going on. And produce an excitement to work this field. There are so many challenges in this field [6].

2.8. DATA MINING AND BIG DATA

Data mining [24] is to analysis all the steps involved in the process of KDD (Knowledge discovery) or can say to analysis the steps in databases process. Data mining is also involving in the process of machine learning, artificial intelligent, database management systems and statistics. So can say that this is the sub-field of the computer, science that is used to discover some type of pattern in databases of different kinds. The goal of data mining to find out the useful knowledge from the large and very complex database and another step after this has been done, to translate this computer based knowledge into a form that is easily understand by human being. Apart from this, so that the mining process also involved to preserve some privacy that will increase the interest of user's. There are some steps that are involved to do data mining data preprocessing, modeling, a metrics of interestness, a consideration of complexity, visualization of the data, privacy preservation and apart from all this, one more step is there name is post-processing [24].

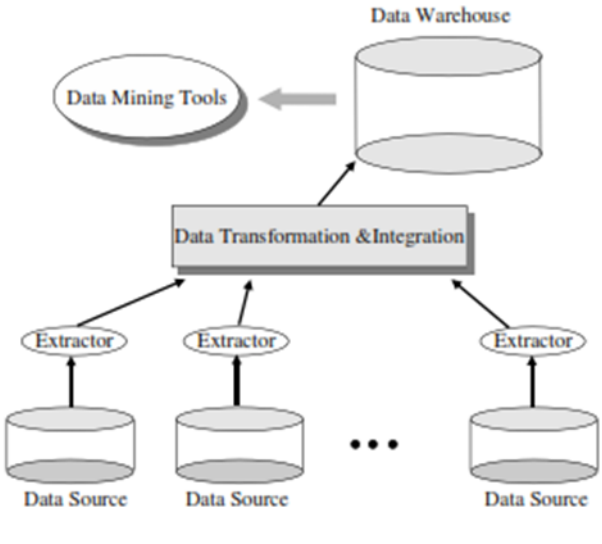


Figure-7 [28] : Data Warehouse

HACE Theorem. Big Data term starts with very large-volume, **H**eterogeneous, **A**utonomous sources with distributed and decentralized control, and try to explore very **C**omplex and **E**stablished relationships among data [24].

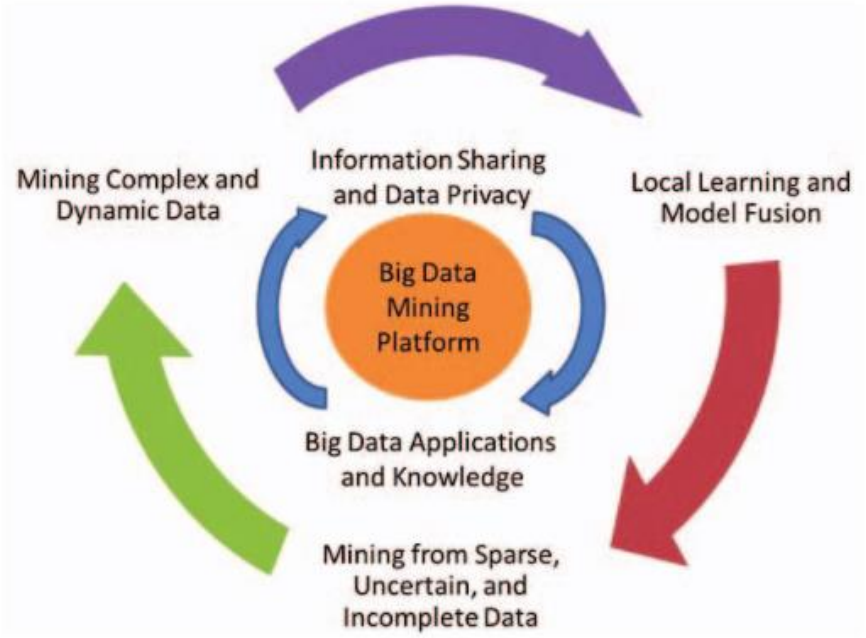


Figure -8 [24] : Big Data

As you can that in above figure shows the process of all the steps involving in

2.9. HADOOP AND HDFS

2.9.1. HADOOP

If we look at the official website of Hadoop [29], then some of features are discussed in below.

"The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing".

The Apache Hadoop is a software library with open source concept that provides a framework to processing of the tasks in distributed manner for the processing of data with very large quantity on a cluster, cluster size will depend upon the size of problem.

A cluster with thousands of computers with single server, with local storage and their computational capabilities. The main reasons behind the design of Hadoop are that:

1.	High Availability	It provides high availability using the replication factor by storing the same data block over more than two sites.
2.	Simple	It allow programmer to easily develop efficient parallel code in map reduce Para diagram
3.	Scalable	Hadoop scales to handle large amount of data by summing up more data nodes to the cluster linearly.
4.	Robust	It is designed with the assumption of having frequent malfunctioning of hardware. And it handles most of the failure very efficiently.

The Hadoop framework is mostly written in the Java, having some native code in C and some command line tools in Shell scrip. For users, even though Map Reduce code in Java is mostly common, but any programming language can also be used along with "Hadoop Streaming" to implement the "map" and "reduce" in the program [29].

The Hadoop Project consist of mainly four modules [29]:

1.	Hadoop Common	The utilities which are common to the all the Hadoop modules and provide their support.
2.	Hadoop Distributed File System	A file system for the distributed environment providing high availability and throughput access.
3.	Hadoop Yarn	Handles management of cluster and its resources for job scheduling by providing an effective framework.
4.	Hadoop Map Reduce	Programming model for parallel processing large scale datasets.

2.10. ENCRYPTION ALGORITHM

2.10.1. AES

AES block cipher developed by the Jon Daemon and Vicent Rijmen. Advance encryption standard (AES) support the any combination of image with key size 128, 192 and 256 bits. In AES algorithm 128 bit data divided into four basic operation block. This block are maintain by 4x4 matrices for the decryption process these 128 bit data passed through different number of round like 10,12,14. This round maintains by following transformation [21]:

2.4.1.2 Sub Byte Transformation:

Sub byte Transformation is S Substation table (S box) having properties of nonlinear substation which is made by multiplicative inverse and affine transformation. The figure show that sub byte transformation.

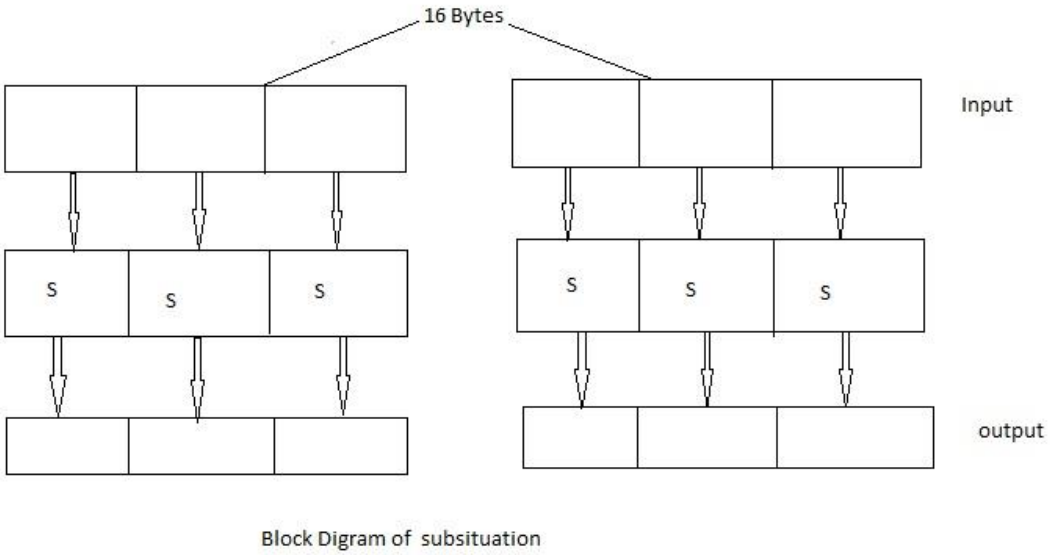


Figure-9 [21].

SIFT ROW TRANSFORMATION:

This is process of byte transposition means last three row are circularly sifted, the offset of left change from one to three bytes.

MIX COLUMNS TRANSFORMATION:

This is process of matrix multiplication of the states. Every Colum multiplied by the constant matrix. It means bytes are treated as polynomial instead of a number.

Add round Transformation:

This is process of XOR operation of round state and round key. This transformation having the property of own inverse.

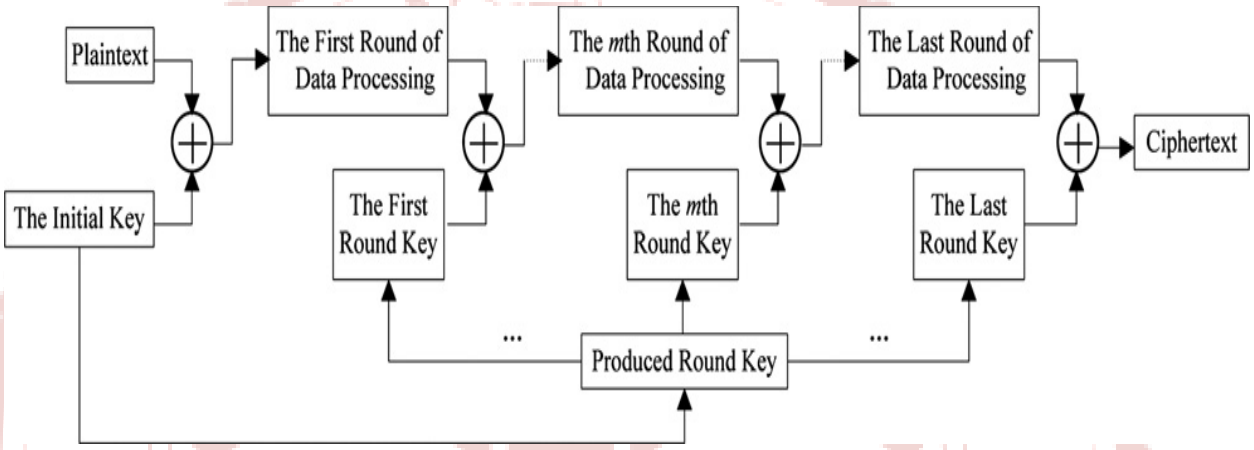


Figure-10 [21].Encryption Process of Advance Encryption Standard (AES)

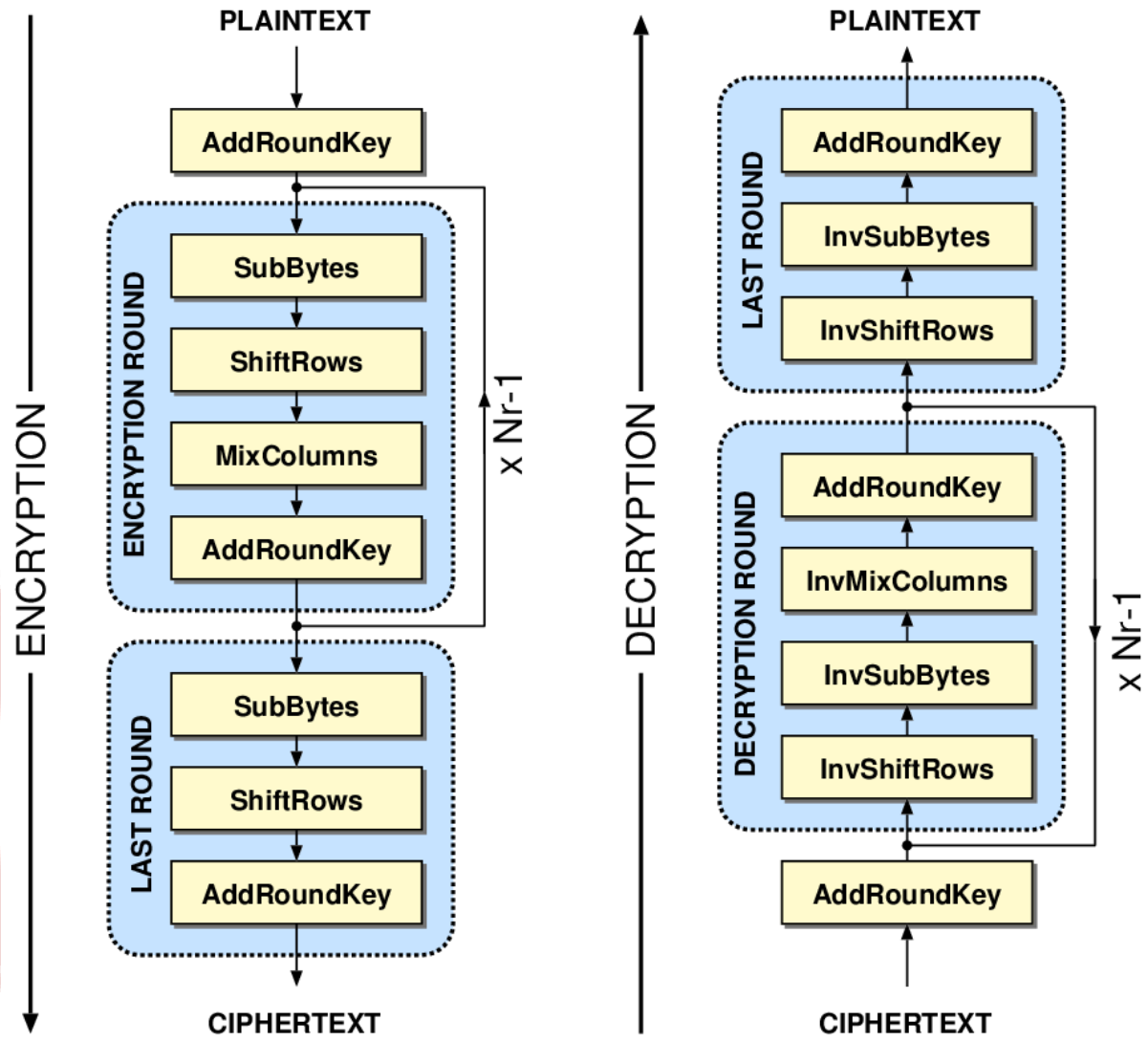


Figure-11 [21]. Encryption/Decryption Model

The encryption and decryption having several step. First add round, a round function work on data block related all sub operation like sub bytes, shift row, mix column and add round key will be perform. This operation will perform many times which is depend on key length. The decryption step same as perform encryption in reverse order. This Advance encryption algorithm having key size 128 bits.

CHAPTER 3

HADOOP

[23]If we look at recent days, then so many companies like Google, IBM, VMWare and Amazon etc, these companies have specify so many products and approaches for data mining. One of a product which free and trustable is Hadoop. Hadoop had introduced around 2005 officially as a part of Nutch subproject of Lucene by Apache Software Foundation. Look at past history which is what, it is mostly inspired by MapReduce and Google File System and Google File System – GFS is originally developed by Google Labs. Hadoop was used in starting past days to analysis of the contents on the internet for a purpose of searching keywords. But later, an idea is realized by people that it can be used to solve different kinds of problems depending upon the nature of problem that requires massive scalability.
For an example, if we are working on 10 TB data because 21th century is going on and size of data is increasing day by day and to handle this kind of problem, there are so requirements of parallel processing tool to handle this type of situation in very less time. So that Hadoop became so popular. Hadoop is efficient for these kinds of problems because it provides a parallelism. Hadoop allowed data to enter in system in a parallel manner and process it in very less time and increase the speed of processing. Hadoop is also robust, scalable because it allows operations on data which has a size of PB. In addition, Hadoop placed on servers, which are inexpensive and present to their use for anyone. Hadoop is ideally installed on Linux platform, with a framework which is written in java. Applications that are submitted to the Hadoop, may be developed using other languages like C++ [29].

3.1. HADOOP ARCHITECTURE

Hadoop consists several elements. HDFS is most of the element of the Hadoop that is resides in bottom that stores the files which across storage nodes of a Hadoop cluster.

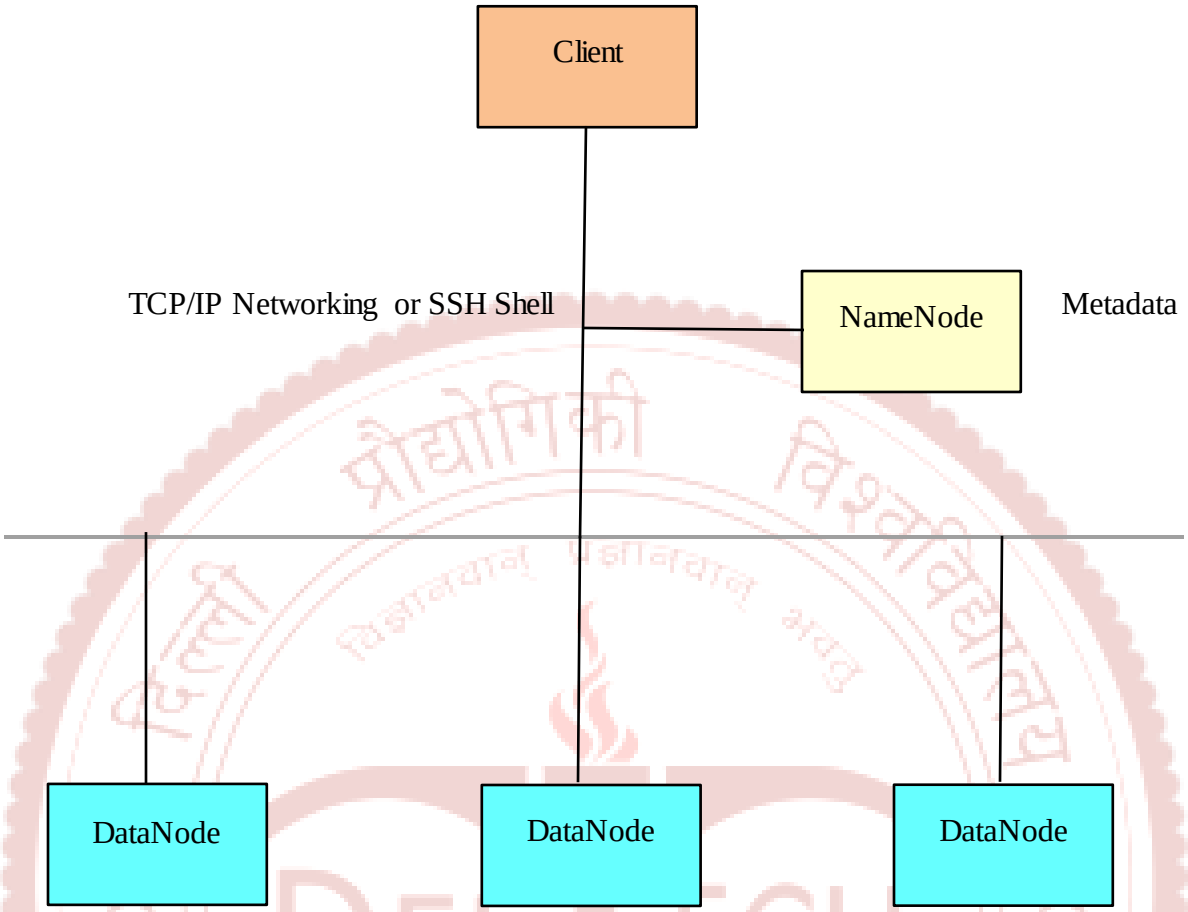


Figure-12:A Simplified View of Hadoop Cluster [29]

3.2. HADOOP DISTRIBUTED FILE SYSTEM-HDFS

HDFS plays a role as traditional hierarchical file system to an external entity called client. Files can be deleted, created, moved and renamed and so on. But HDFS supports some special type of characteristics, architecture of HDFS is built from collection of nodes as you can see that in above figure. There is a NameNode which will be only one in the cluster and there may be several DataNodes in cluster depending upon the requirement of problem. NameNode is only responsible the flow of control or can say that distributed the tasks into different DataNodes. NameNode provides a special type service which under comes metadata service within HDFS. And DataNode that is responsible for serve the storage blocks for HDFS. As we know that there is only one NameNode exist in cluster and do you know that what it is means? Single node failure, now that reason, this created an issue in HDFS.
Due to that reason files are stored in HDFS are divided into blocks and blocks are replicated to multiple computers (DataNodes). And HDFS is totally different from RAID architecture.

The block size is 64 MB but amount of the block size is determined by client, when file is created [23]. All communications take place in HDFS through SSH Shell because provides secure communication channel. Means lossless information communication but if we are talking about HDFS when data is uploaded to the HDFS then data is communicated through secure shell but no encryption or very low level encryption standards are used to preserve privacy in HDFS [7]. As I studied on this topic, there are no technology that are used by HDFS to encrypt the data in unreadable form to attackers. Means Hadoop is the latest technology nowadays and there should be some privacy to preserve because that will also play an important role to attract the people in this field [23].

3.3. FILE OPERATIONS

As studied on the HDFS, it is very clear to say that HDFS is not a file system for the purpose of general. But can say that, it play as a role of mediator between actual file system but some privacy must be preserve to reliability because this time is going on hacking or attacker. Instead, it is designed to support large to provide streaming access that are written once. If a client want to store the file in HDFS, then process is very familiar, process will start with client will cache the file in temporary local storage and when cached limit is out of bound with data and then a request for creation of the file is sent to come which name is NameNode. Now NameNode will response to the client with the identity of DataNode and the address of destination block. The DataNode is also notified that consists the replica of block. Checksum file is also created by client that is saved in HDFS namespace. When last block is sent, the NameNode commits the file creation to its persistent storage [29], [23].

3.4. HADOOP APPLICATIONS

Hadoop is very useful software framework in nowadays and latest technology that is very useful to attract the interest of today's people to do work in this field. Big Data mining is also an application of the Hadoop. Cloud Computing is also very important application of Hadoop [23].

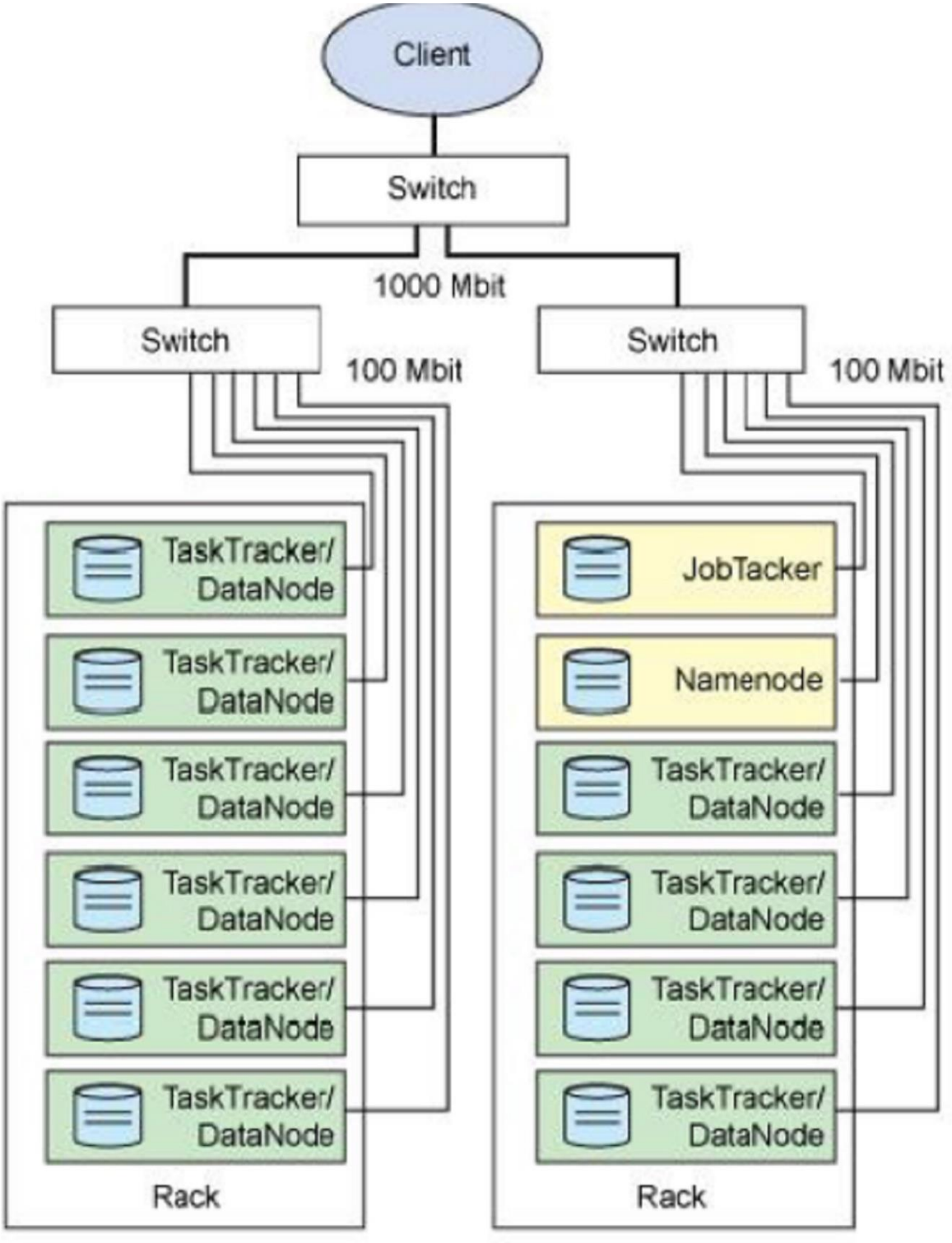


Figure-13: A Typical Hadoop Cluster [23].

CHAPTER 4

SSH SHELL

4.1. SSH SHELL

What is Secure Shell? Secure Shell is nothing, it is used for communication purpose when we are working on a cluster, then we can use the SSH shell, just because of it provides secure communication over a medium. In other words, it provides an interface for the remote communication over the cluster of computers. Do you know that, many industries are focusing on security and privacy, that's why for the present demand of users, industries, SSH Shell is becoming more popular in now days.

It is better way to provide an effective communication when we are dealing with large number of computers for the execution of tasks. Just because of SSH Shell it provides some privacy during the transmission of big data. That means attacker cannot attack on data in during the transmission of the data. It is a protocol that used by many servers to provide an effective communication. There are so many encryption technologies used by SSH Shell.

A mechanism by SSH Shell that provides an interface for establishing a connection with cryptographically secured between two trusted parties, authentication will be there between each other, and passing that is done in the form of commands and output is back and delivered. And now, question is that, how to SSH Shell works and technologies that are used by this shell for secure connection [11].

4.2. HOW DOES SSH SHELL WORK AND ALGORITHM USED BY THE SSH SHELL

[11] You should have basic knowledge of SSH Shell to better understanding of about how does SSH Shell work. SSH Shell protocol applied to a model which will Client- Server model based, that will authenticate two communication parties and uses the existing technologies to encrypt the data, files between them.

The server has a port that is specified to listen the SSH connection from the client to provide a response that is transferred to the client and used by the client at this specified port.

This port is responsible for provide a connection with security, privacy, authentication that is used to verify both parties and produced an environment with correctness and secureness, if the credentials are accepted.

The client is responsible for generating the beginning with TCP handshaking with the server, client will also negotiate the secure connection, and it will the server identity that is already stored in client space and provides credentials to authenticate. The session of SSH Shell is established in two separate stages [11].

1. The first one, agree upon and generate a connection with security with encryption to protect the information for the future use.
2. To authenticate the valid user and checking for weather access of server is granted or not for this particular user is done in second stage.

When a connection is established by the client then, server will response with the version number of the protocols that are supported by it. If requested client is matched with any supported available protocol versions then connection will be continue, otherwise the connection is terminated with an error, requested protocol version is not supported by the server. Now public host key is provided by the server, which will be used by the client to verify weather given host is intended host or not.

At this stage, both parties will negotiate a session key using a version like something suppose Diffie Hellman Algorithm.

Now there is a way to make it possible for each party to combine their own private data with the public data that is belong to other system to make an identical secret session key by above algorithm (Diffie Hellman Algorithm).

To encrypt the entire session, this session key will be used for it. For this part of the procedure, the public and private key pairs used that are completely separate from the SSH keys used to authenticate a client to the server.

Procedure basis for classic Diffie-Hellman is:

1. A large prime number on which both parties agree upon, which will under come in a name as a seed value.
2. An encryption generator on which both parties agree on (typically AES), that will be used a predefined way is used to manipulate the values.

3. Another prime number will take by each party independently which will be secret with each other. A secret key that is this prime number, which is used as a private key for the purpose of interaction with each other. And this private key will be different from the private key used by SSH Shell for authentication.
4. All combination of generated private key, the encryption generator, and the shared prime number is used to generate a key which will be public key that is achieved by a private key, but that can be shared by another party.
5. The generated keys will be exchanged by both parties.
6. At another end, there will be an entity which is called receiving entity will use their own private key and public key of other side and a shared prime number, these combination will be used to achieved a shared secret key. All is done in a manner which is based on independent with each other.
7. The shared secret key will be used to encrypt the all communication in below way.

An encryption that is used for the rest of the connection which has a name binary packet protocol that is based on shared secret encryption. From above process, every party can participate in that and have equally rights to generate the shared secret. From above process of steps, which does not give a permission to one end to control the secret.

The task is completed for generating same shared secret without ever having to send this information by above process over insecure channels.

The generated key that is secret based on a cryptography technique that is called symmetric, means both parties will use the same secret key for the purpose of encryption and decryption. The purpose of SSH just wrap all the data in the form of unreadable (Encryption) by a third party that cannot understand by unauthorized person.

When session is established, after that some stages of authentication begins.

And there are so many techniques used by SSH Shell to secure the data that are discussed below.

4.3. TECHNOLOGIES USED BY SSH SHELL

To make the secure communication, SSH Shell several kinds of technologies according the nature of problem for secure transmission of the data. Which are discussed as follows:-

1. Symmetrical Encryption.
2. Asymmetrical Encryption.
3. Hashing.

1. SYMMETRICAL ENCRYPTION

What is encryption scheme, Symmetrical Encryption or Asymmetrical Encryption, it is determine by the relationship between the components that are used to encrypt the data and decrypt the data. In this technology, where a sender will be there, that will encrypt the data by a key and this same key also used by the receiver of message to decrypt the data. Means there are many chances, when Privacy is lost. Whoever have a key, those can decrypt the data. Symmetric keys are used by the SSH shell to decrypt the connection. In many cases, Asymmetrical public/private key pair that is made to only one purpose, only to provide the authentication. Not encryption the connection. An authentication for password against snooping can be allowed by the Symmetrical Encryption. There are so many techniques that can be used for a purpose, to configured the SSH Shell, these technologies consist AES, Blowfish, 3DES, CAST128, Arc four. The technology that is used by the SSH Shell during the transmission is decided by server and client, what is preferences are there.

The definition of symmetric key algorithm that use the single key which also called shared secret key. In symmetric key algorithm the same key will be used for encryption and decryption process. Here two type of symmetric algorithm, block and stream cipher. A block cipher used encrypt the image into cipher image having same size of image after encryption. For example we take the size to 1024x1024 bmp image or size of bmp image is 2.5 MB as an input for encryption than cross ponding output must be same size. Blowfish, Data Encryption Standards (DES), Triple DES, and IDEA are example of symmetric block cipher. The symmetric key algorithm use a single key for encryption and decryption process.

The blowfish algorithm was designed in 1993 by Bruce Schneider is one symmetric block cipher algorithm. Whereby it can be used changed of Data Encryption Standard (DES) and International data encryption algorithm (IDEA).the blowfish algorithm having two part first one a key expansion part and second one is data encryption part. In this algorithm encrypt the data using block cipher method means image breaks into block having size of block is 64 bit block. It take a range of key from 32 bits to 448 bits length, which show flexibility in its security strength.

If we are talking about a version of Ubuntu operating system is 14.04, in this version AES128, AES192, AES256, Arc Four128 and so on. This means if there are two machines of

Ubuntu 14.04 want to communicate with each other, then by default, a default algorithm is used to encrypt the connection, default algorithm is AES128 [11].

2. ASYMMETRICAL ENCRYPTION

Some times SSH used to encryption technique as Asymmetrical Encryption. In which both parties (sender and receiver) have two keys, one is private and another is public [11].

3. HASHING

Another beauty of SSH shell is sometimes it takes the advantages of hashing for the encryption of entire connection according the demand of application.

There is a function which has a name called cryptography hashed function that is used to achieve succinct "signature" or a set of summary with information. This same hash function is will take a place in way of useful for decrypt the message. This is used in reverse way, when case is of decryption [11]. If one portion of the data is modified then entire hash function will also change.

A user should not able to creation of real message from given hash function vale but able to ask or tell that a given massage will produced the same hash value.

Given these properties, hashes are mainly used for data integrity purposes and to verify the Authenticity of communication. The main use in SSH is with HMAC, or hash-based message Authentication codes. These are used to ensure that the received message text is intact and Unmodified.

As part of the symmetrical encryption negotiation outlined above, a message authentication code (MAC) algorithm is selected. The algorithm is chosen by working through the client's list of acceptable MAC choices. The first one out of this list that the server supports will be used. Each message that is sent after the encryption is negotiated must contain a MAC so that the other party can verify the packet integrity. The MAC is calculated from the symmetrical shared secret, the packet sequence number of the message, and the actual message content. The MAC itself is sent outside of the symmetrically encrypted area as the final part of the packet. Researchers generally recommend this method of encrypting the data first, and then calculating the MAC [11].

CHAPTER 5

PROPOSED WORK

5.1. PROBLEM STATEMENT

If we take the data with large or big size (size in MB or GB) than AES, blowfish algorithm takes more time to encrypted the data so encryption –decryption time ratio will be high.

The objective of encryption -decryption algorithm to make more efficient following constraints:

1. To reduce the encryption and decryption time.
2. Security Analysis for Privacy-Preservation.
3. To show the results that are practically proved that existing algorithm and my proposed algorithm with an implementation.
4. With my proposed algorithm, we can execute the dataset that can be hold the size of greater than computer RAM than AES.
5. To give an approach for HDFS environment that is very useful for the improvement of HDFS Hadoop because we are dealing with big data, then Privacy-Preservation should be main concern. Means the data is in plain form in HDFS, it means when communication is going on all over the world. Then data should be in encrypted form to provide the privacy because if little bit of information is leaked by attacker, then data of TBs, how to handle it. So this approach for HDFS environment can be used.
6. To comparison between the results of different sizes of the data.
7. To compare the results with different parameters.
8. An approach and implementation with some results based on Block Wise encryption using AES that will reduce the time and increase some other factors like a big data processing than original AES.

The main objective of work to reduce the time of encryption and preserve the privacy to increase the security at HDFS environment and at a level of little.

5.2. PROPOSED WORK

Let we take data with size of 100 MB file. Then it will be break or can say stored into a form of the blocks or read the data with specified range of the size of data of blocks.

Suppose file takes place into the memory with 100 blocks of each of size 1 MB.

Number blocks will be there = F/S, Where F is the size of original file and S is the size of each block.

1 MB 1MB 1 MB 1MB

P1	P2	------	------	------	------	-----	----	P99	P100

Figure:14 : Encryption Model

Now applied the AES algorithm:-

Encryption Process:

The encryption process for the data or information with using Advance encryption Standard (AES) having symmetric key 128 bit. Same key will be uses in the encryption and decryption process for enhancement of algorithm.

Here the first step that P1 break into sub block having the size of 1024 bits each block than total number of block will be 100 for P1. Same process do repeat for block P2 and so on.

P1

P1,1 1 KB	P1,2 1 KB	P1,3 1 KB	------	------	------	------	------	P1,1023 1 KB	P1,1024 1 KB

1 1024

Figure:15 : Encryption Model

Taking the, first block P1 (1024 bits) encrypted with Advance encryption Standard (AES)

E= **Encrypt_AES (P1)** …… (1).

Encrypt_AES (P1)

Encrypt_AES P1,1 1 KB	Encrypt_AES P1,2 1 KB	Encrypt_AES P1,3 1 KB	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -	Encrypt_AES P1,1023 1 KB	Encrypt_AES P1,1024 1 KB

1 Figure:16 : Encryption Model 1024

Now taking the sub block **Encrypt_AES (P1,1)** which has a size of 1 KB and expanded it up to size of P1, mean 1024 KB (1 MB) and XOR with next block from P1 which is P2.It means make the duplicate copy of **Encrypt_AES (P1,1)** up to size of block P1. The expandable **Encrypt_AES (P1, 1)** is given in shown in figure.

Exp_Encrypt_AES (P1, 1)

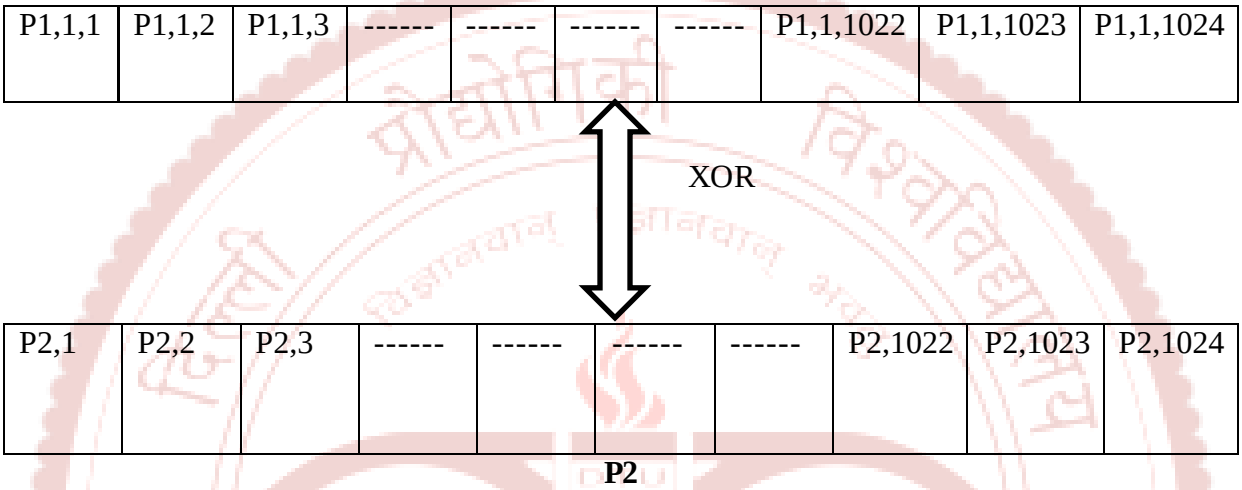


Figure:17 : Encryption Model

Again taking the next sub block of P1 which is Encrypt_AES (P1, 2) and expanded it up to the size of P1 block which 1024 KB.

Take the next block of P2 which is P3 and then XOR it with Encrypt_AES (P1, 2) in same above way.

Exp_Encrypt_AES (P1, 2)

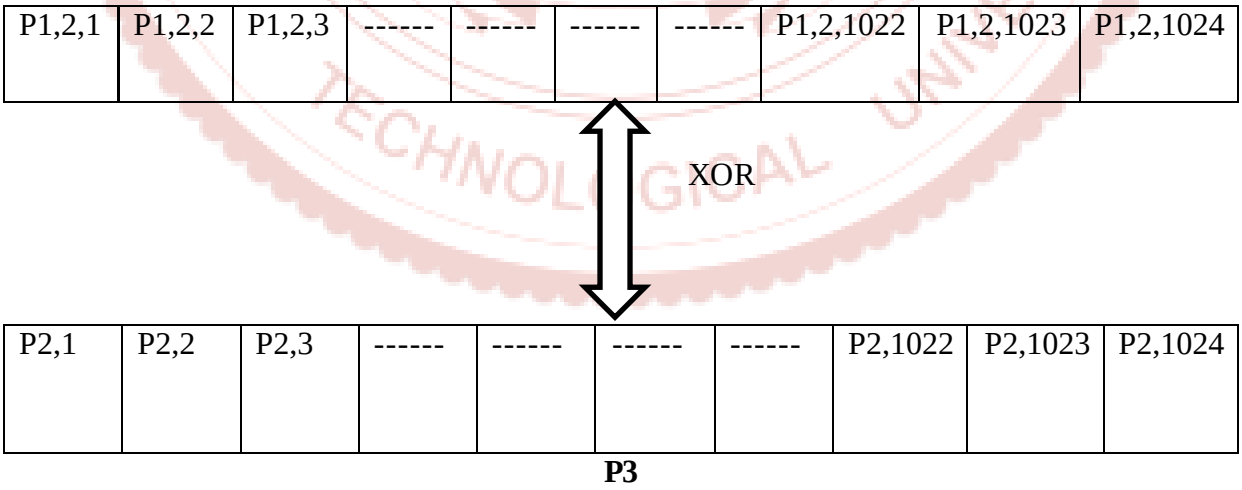


Figure:18 : Encryption Model

Above process will repeat for next block of F which is P4 and next sub block of P1 which is Exp_Encrypt_AES (P1, 3) or until complete file is encrypted and one can only performed 1024 operations. And after that number of block, next block treated as a first block.

And now last step applied the AES algorithm on P1 block

E= **Encrypt_AES (Encrypt_AES (P1))** ……………… (2).

Because P1 is encrypted two times that will increase the security level and preserve the privacy.

Decryption Process:

The process of decryption of encrypted of the data has following steps. A same key at the time will be used for decrypt the data in decryption process. Because of its symmetric encryption decryption algorithm.

Suppose we take the decrypted file in form of blocks.

1 MB	1MB							1 MB	1MB
C1	C2	------	------	------	------	-----	----	C99	C100

Figure:19 : Decryption Model

Let take the Decryption process in which algorithm takes the encrypted data and decrypt the first block (**Encrypt_AES (Encrypt_AES (P1))=C1**) with Advance encryption Standard (AES) and we get the block (**Encrypt_AES (P1))** which has a size of 1024 KB.

D1 = Decrypt- AES ((**Encrypt_AES (Encrypt_AES (P1**) ------------- (1)

Second step is that take the decrypted block and breaks it into number of sub blocks in reverse way of encryption process.

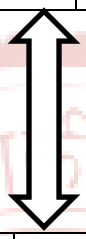
Decrypt_AES (Decrypt_AES (D1))

Decrypt_AES (Decrypt_AES (D1,1)) 1 KB	Decrypt_AES (Decrypt_AES (D1,2)) 1 KB	Encrypt_AES D1,3 1 KB	- - - - - -	- - - - - -	- - - - - -		- - - - - -	- - - - - -	Decrypt_AES (Decrypt_AES (D1,1023)) 1 KB	Decrypt_AES (Decrypt_AES (D1,1024)) 1 KB

1 Figure: 20 : Decryption Model 1024

Next step is that take the first sub block (D1,1) and expanded it up to the size of D1 block and XOR with the block D2.

D1,1,1	D1,1,2	D1,1,3	------	------	------	------	D1,1,1022	D1,1,1023	D1,1,1024



XOR

D2,1	D2,2	D2,3	------	------	------	------	D2,1022	D2,1023	D2,1024

Figure:21 : Decryption Model

From above we get the original Encrypt_AES P1, 1

Next step is that take second sub block and expanded it in same above way and then XOR it with next block of encrypted file.

Next step is that when all the XOR operations are performed then take the first block of encrypted and applied again AES to decrypt the file.

5.3. FLOW CHARTS

I have discussed flow charts of my approach that will show the flow control of the algorithm proposed in this research work. Flow chart that is better way to the purpose of understand the problem statement easily. So I am going to draw the flow charts of my problem that are below.

5.3.1. FLOW CHART OF ENCRYPTION ALGORITHM

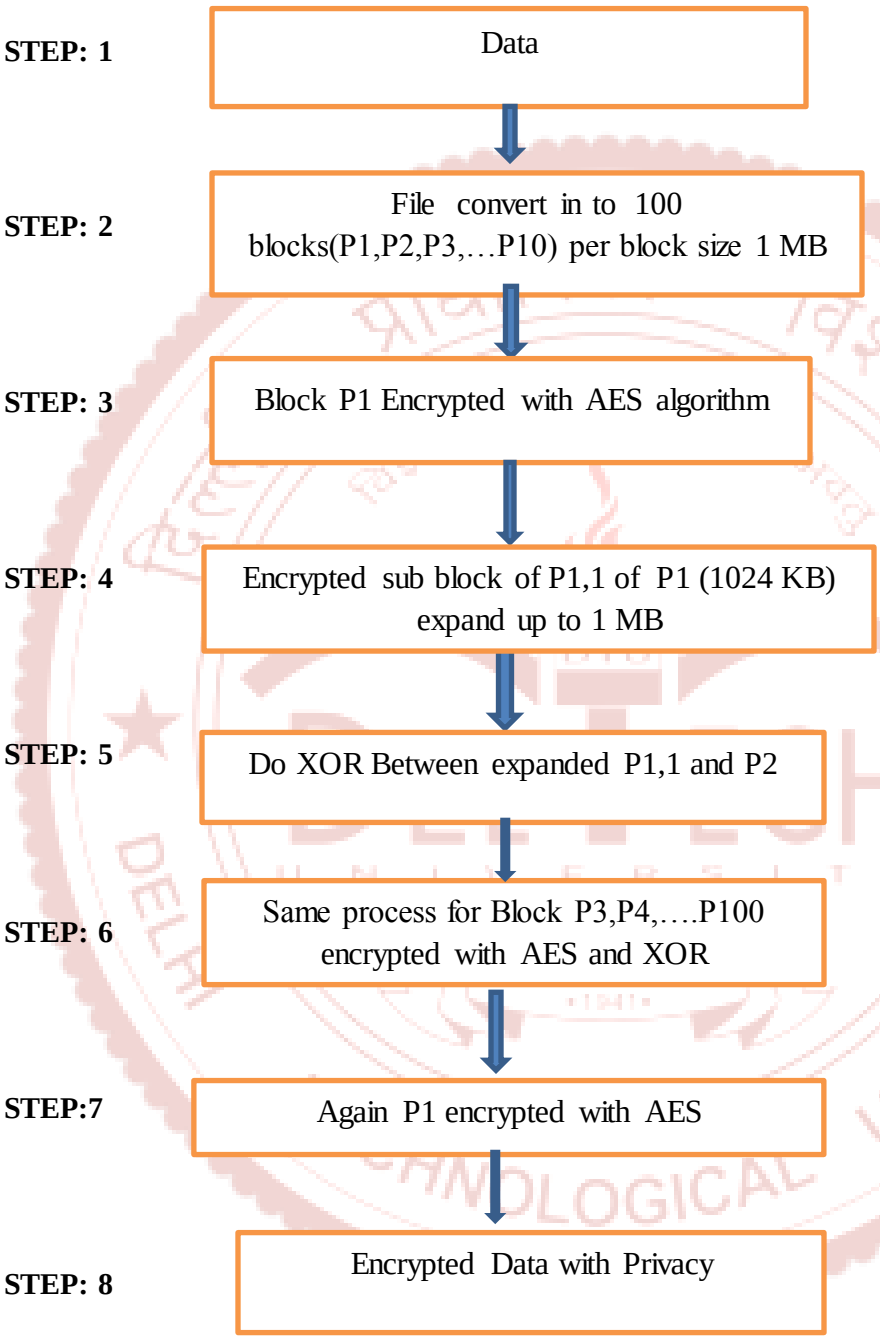


Figure:22 : Encryption Process

5.3.2. FLOW CHART OF DECRYPTION ALGORITHM

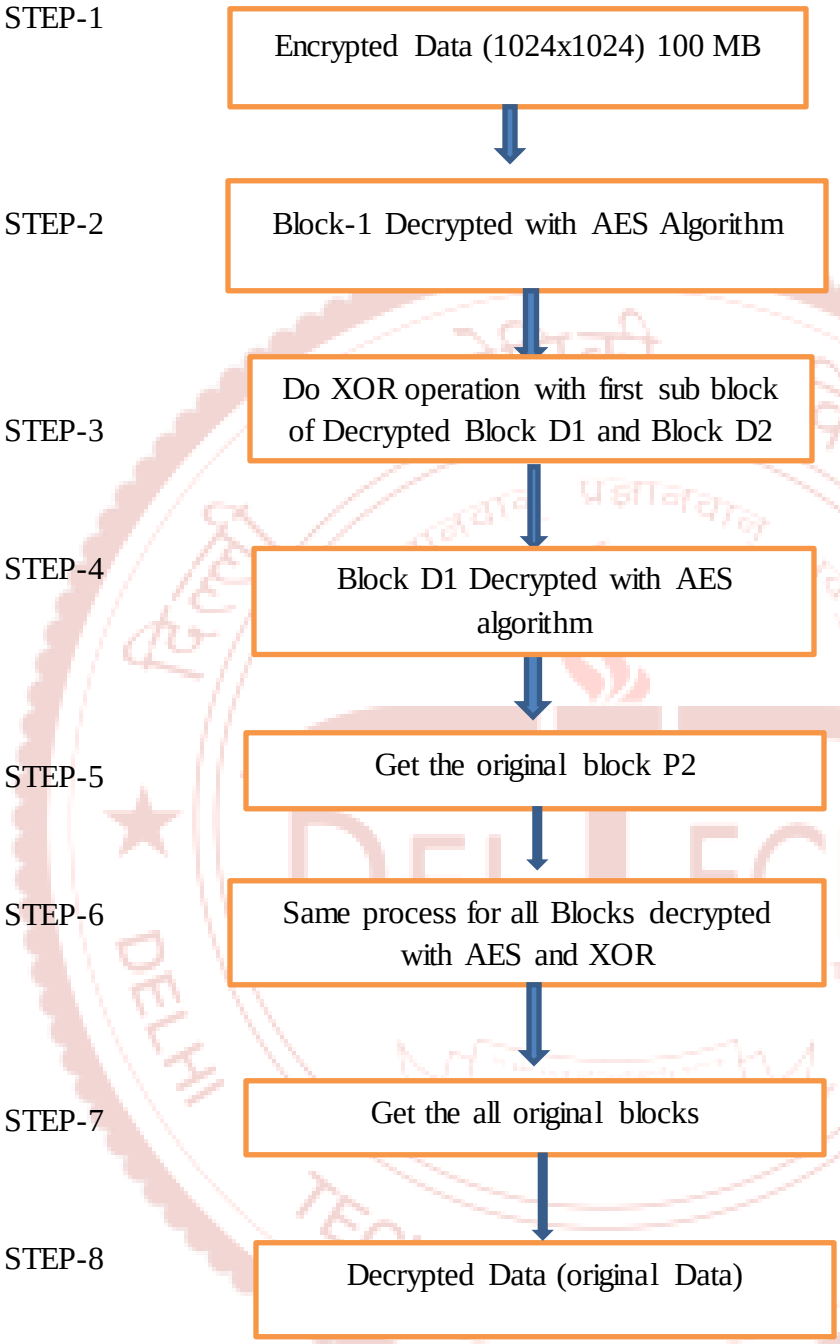


Figure:23 : Decryption Process

5.4. ARCHITECTURE OF PROPOSED ALGORITHM

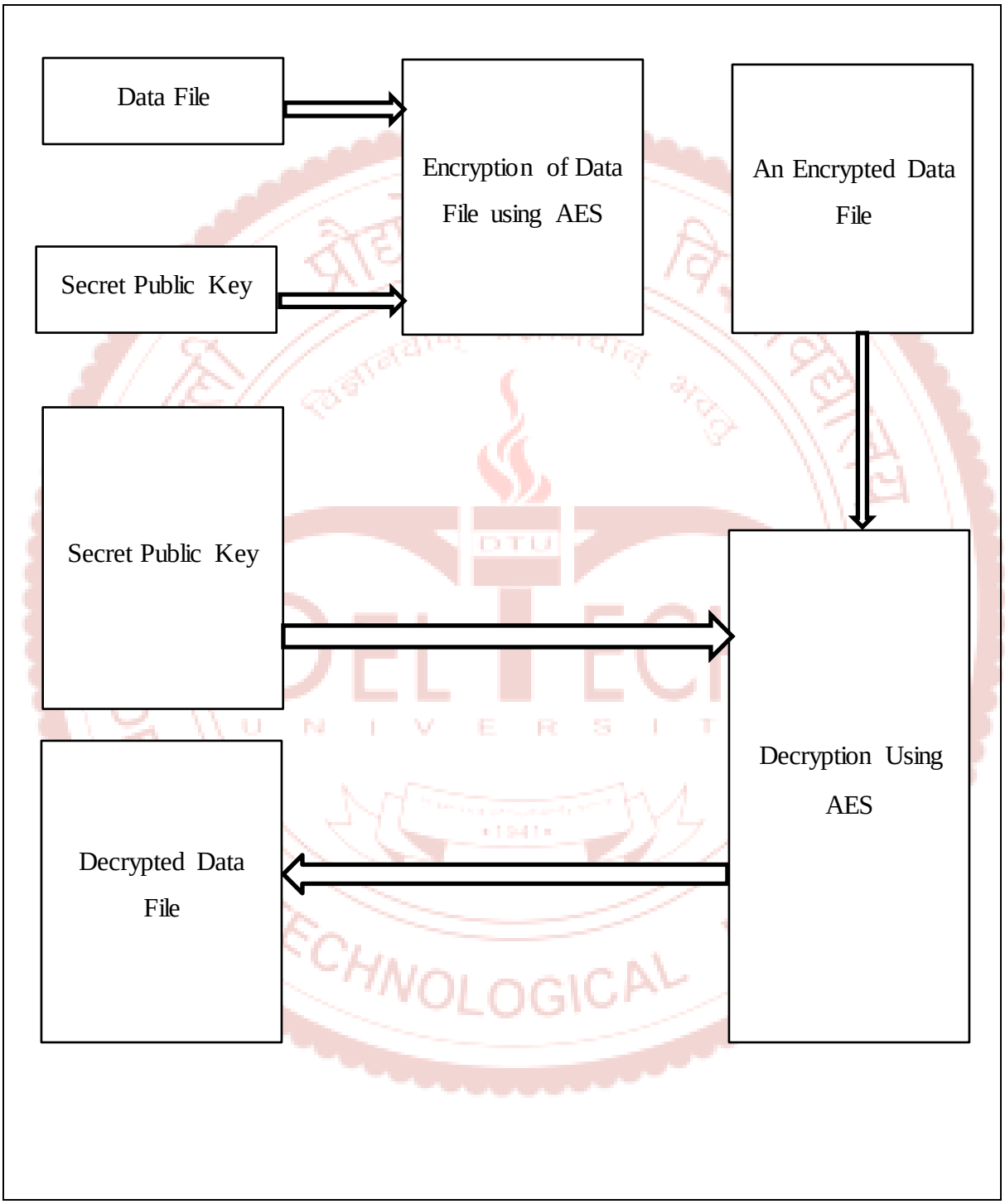


Figure:24, ARCHITECTURE MODEL OF PROPOSED ALGORITHM

CHAPTER 6

IMPLEMENTATION, TESTING AND RESULTS ANALYSIS

We take the different different sizes files File or data set is encrypted and decrypted and then execution time is compared with existing algorithm. The encryption algorithm coding done in Java.This approach is implemented in java.

EXECUTION TIME RESULTS:

File Name: accidents.dat [30]
Size:33.8 MB
Number of Runs: 10

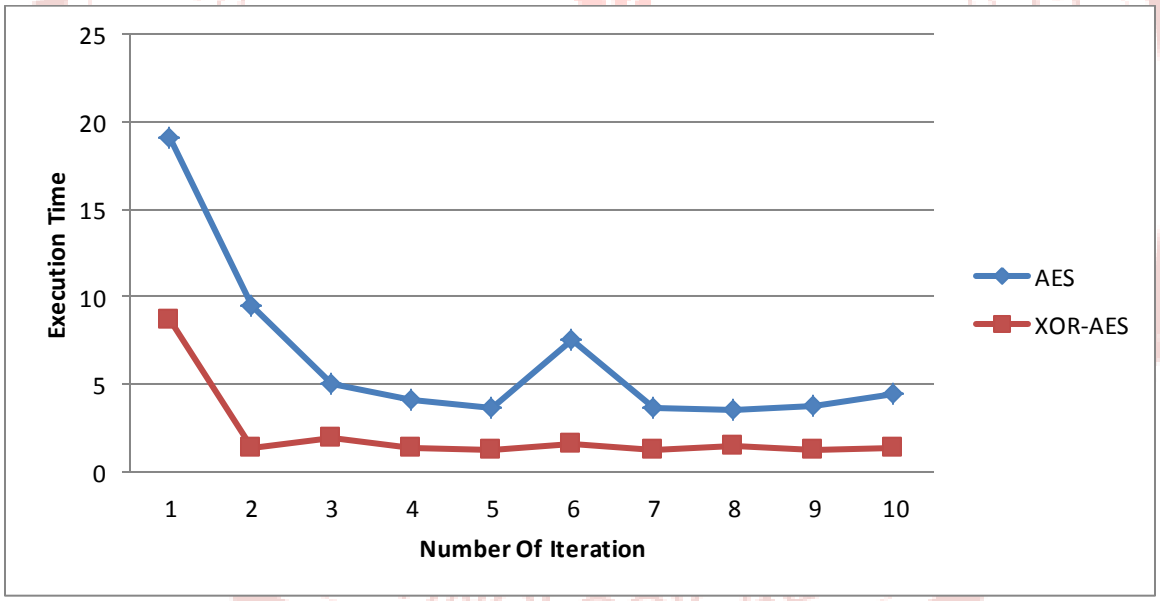


Figure:25 : Execution Time

File Name: T40I10D100K.dat [30]
Size:14.7 MB
Number of Runs: 10

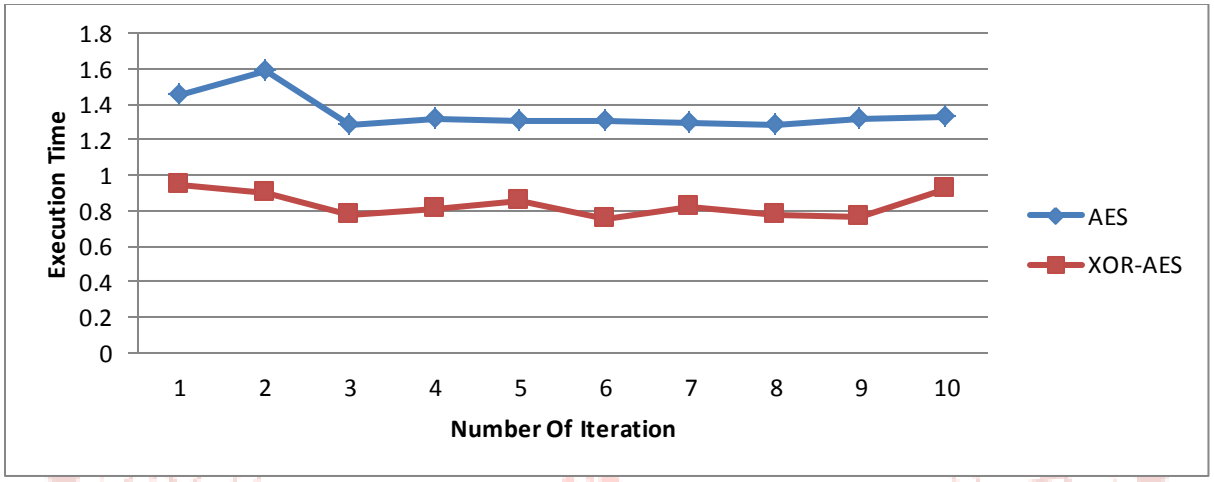


Figure:26 : Execution Time

File Name: retail.dat [30]
Size:3.97 MB
Number of Runs: 10

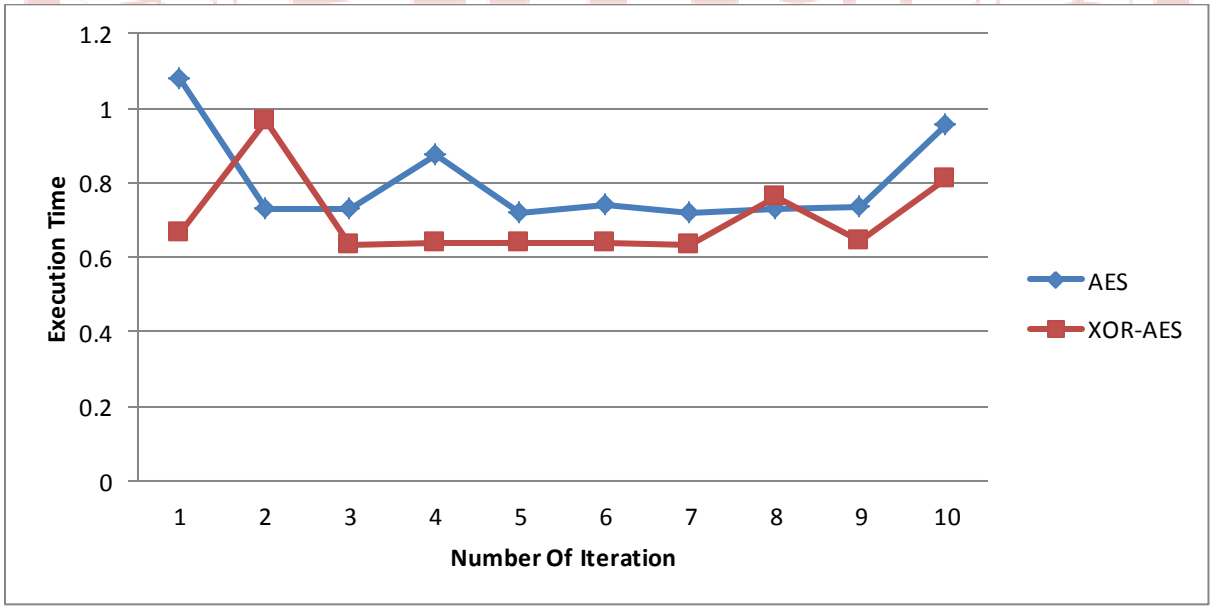


Figure:27 : Execution Time

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this research we proposed a new encryption approach that will provides the highly privacy for data mining sets which is based on AES algorithm. The main objective of encryption algorithm that provide good privacy level from unauthorized person like Facebook cannot see the options that are hide by authorized user. Some confidential data need the security to transmit over the network. We have studies all method of data encryption and got all technique to encrypt the file or datasets.Each method may be possible for different application, some algorithm giving fast encryption and some give the high security level.

In this research we proposed encryption algorithm for privacy preservation in Hadoop environment or data mining that provide the faster encryption method as compare than Advance Encryption Standard. Propose algorithm also provide a good security level because it's double encryption by AES. Proposed scheme suitable for large size of datasets. Here the encryption decryption ratio will less compare than AES. Proposed method mainly useful for when the size of datasets or files in MB or GBs. Its means when the size of datasets will be large then it is very suitable and it is very reliable with respect to the attacking issues. It will give good throughput. The future work of research when we can apply this approach in Hadoop that will reduce the time and reduce the chances of attacks because it is very complicated task to attack on this algorithm because it is double encrypted using AES. If we are thinking about to break this algorithm then just think about AES that is not beaked yet from anyone. With the parallel processing we can encrypt all datasets with large sizes in less time. So this algorithm is very useful for parallel encryption algorithm.

REFERENCES

[1] S.Gokila1, Dr.P.Venkateswari2,(2014), "A SURVEY ON PRIVACY PRESERVING DATA PUBLISHING",DOI: 10.5121/ijci.2014.3101.

[2] BENJAMIN C. M. FUNG,(2010),Concordia University, Montreal,KE WANG,Simon Fraser University, Burnaby,RUI CHEN,Concordia University, Montreal and PHILIP S. YU University of Illinois at Chicago,"Privacy-Preserving Data Publishing: A Survey of Recent Developments",DOI 10.1145/1749603.1749605 http://doi.acm.org/10.1145/1749603.1749605.

[3] Bin Zhou,School of Computing Science,Simon Fraser University,Canada,bzhou@cs.sfu.ca,Jian Pei School of Computing Science,Simon Fraser University,Canada,jpei@cs.sfu.ca,Wo-Shun Luk,School of Computing Science Simon Fraser University,Canada,woshun@cs.sfu.ca,(2007), "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data".

[4] Mohnish Patel, Prashant Richariya, Anurag Shrivastava, (2013),"Privacy Preserving Using Randomization and Encryption Methods",et al., Sch. J. Eng. Tech., 2013; 1(3):117-121.

[5] Charu C. Aggarwal,Philip S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms".

[6] R.Natarajan1, Dr.R.Sugumar, M.Mahendran, K.Anbazhagan,ISSN 2278 - 1021, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 1, MARCH 2012, "A survey on Privacy Preserving Data Mining".

[7] Thanh Cuong Nguyen, Wenfeng Shen, Jiwei Jiang, Weimin Xu, School of Computer Engineering and Science, Shanghai University Shanghai 200072, China, (2013), "A Novel Data Encryption in HDFS".

[8] "IEEE Standard for Encrypted Storage".

[9] Maurizio Dusi, Francesco Gringoli, Luca Salgarelli, (2008)"A Model for the Study of Privacy Issues in Secure Shell Connections", The Fourth International Conference on Information Assurance and Security, DOI 10.1109/IAS.2008.46.

[10] http://csce.uark.edu/~kal/info/private/ssh/ch03_09.htm.

[11] https://www.digitalocean.com/community/tutorials/understanding-the-ssh-encryption-and-connection-process.

[12] http://www.openssh.com/features.html.

[13] http://www.cisco.com/c/en/us/support/docs/security-vpn/secure-shell-ssh/4145-ssh.html

[14] http://www.openssh.com/.

[15] "Implementing Secure Shell", Cisco IOS XR System Security Configuration Guide for the Cisco CRS Router, Release 4.2.x.

[16] Farag Azzedin, Information and Computer Science Department, "Towards A Scalable HDFS Architecture", DOI:978-1-4673-6404-1/13/$31.00 ©2013 IEEE.

[17] Sahil Madaan, Rakesh Kumar Agrawal,Department of Information Technology,Netaji Subhas Institute of Technology, "Implementation of Identity Based Distributed Cloud Storage Encryption Scheme using PHP and C for Hadoop File System".

[18] "A solution dor privacy preservation in mapreduce ".

[19] Kavé Salamatian Universite de Savoie Eiko Yoneki University of Cambridge,"Privacy Preservation in the Context of Big Data Processing".

[20] Mr. Mahesh T.Dhande1, Mrs. N.A.Nemade2, Mr. Yogesh V. Kolhe3, "Privacy Preserving in K- Anonymization Databases Using AES Technique", www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013).

[21] Miss Snehal K. Dekate*, Prof. Jayant Adhikari**, Prof. Sulbha Parate, "Enhancing data mining techniques for secured data sharing and privacy preserving on web mining", International Journal of Scientific and Research Publications, Volume 4, Issue 12, December 2014, ISSN 2250-3153.

[22] Thomas B. Pedersen, Y¨ucel Saygın, Erkay Sava, "Secret Sharing vs. Encryption-based Techniques For Privacy Preserving Data Mining", This work was partially funded by the Information Society Technologies Programme of the European Commission, Future and Emerging Technologies under IST-014915 GeoPKDD project.

[23] Fei Shao, Zinan Chang, Yi Zhang, Department of Information Technology,Jinling Institute of Technology, "AES Encryption Algorithm Based on the High Performance Computing of GPU", 2010 Second International Conference on Communication Software and Networks.

[24] Wei Fan,Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", ACM SIGKDD Explorations Newsletter, 2013.

[25] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", Proc. 20th Int'l. Conf. Very Large Databases, pp.487-499, 1994.

[26] Kamrul Shah, Mohammad Khandakar, Hasnain Abu. "Reverse Apriori Algorithm for Frequent Pattern Mining", Asian Journal of Information Technology 7 (12),:524-530,

ISSN: 1682-3915, 2008

[27] Zhao, Shulei, and Rongxin Du. "Distributed Apriori in Hadoop MapReduce Framework." (2013).

[28] http://en.wikipedia.org/wiki/Data_mining.

[29] https://hadoop.apache.org/.

[30] http://fimi.ua.ac.be/data/