Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of System Engineering and Information

Master's Thesis

Big Data and its Security Challenges

Bc. Mayank Jitendrabhai Modi

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Bc. Mayank Jitendrabhai Modi

Systems Engineering and Informatics
Informatics

Thesis title

Big Data and its Security Challenges

Objectives of thesis

Emerging big data technologies raise many security concerns and challenges. There are huge amounts of data from different sources divided into structured and unstructured data. Big data technologies also have issues with attacks and threats like data theft and data hack. My main objective is to research and find out security issues and their challenges in big data technologies. As an outcome I wish to present and explain the security threats and attacks on big data and its challenges.

Methodology

I will research on previous resources and case studies and find out security issues in big data.

The proposed extent of the thesis

60 pages

Keywords

Big Data, Big Data Security, Big Data Security challenges, Data Hacks, Data Theft, Big Data Analysis

---

Recommended information sources

Benjelloun, Fatima-Zahra & Ait Lahcen, Ayoub. (2015). Big Data Security: Challenges, Recommendations and Solutions.

Bernard Marr. 2018. Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Audible Studios on Brilliance Audio.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work and think. London: John Murray.

Meier A, Kaufmann M (2019) SQL- & NoSQL-databases: models, languages, consistency options and architectures for big data management. Springer, Heidelberg

Ottenheimer, D. (2015). The realities of securing big data.

Expected date of thesis defence

2021/22 WS – FEM

The Diploma Thesis Supervisor

Ing. Himesha Prabhakara Wijekoon

Supervising department

Department of Information Engineering

Electronic approval: 23. 2. 2021

Ing. Martin Pelikán, Ph.D.

Head of department

Electronic approval: 23. 2. 2021

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 29. 11. 2022

Declaration

I declare that I have worked on my diploma thesis titled "**Big Data and its Security Challenges**" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague, 2022 _____

Acknowledgement

I would like to thank my family for always encouraging and supporting me in my life, and also my God for giving me with strength, again for chance to travel here and study in Prague, Czech Republic.

I would also like to thank The **Czech University of Life Sciences Prague** and my supervisor, **Ing. Himesha Prabhakara Wijekoon**, for providing me with this excellent thesis topic and allowing me to conduct research on it. They were constantly encouraging while I was researching on my thesis.

I'd also like to thank all of my family, friends, and incredible worldwide friends and acquaintances who have consistently believed in me and assisted me in achieving my goals.

Big Data and its Security Challenges

Abstract

In this study I have proposed work consisted of four entities including Trusted Center (TC), Data Owner (DO), Data User (DU), and Cloud Server (CS). The data stored in the cloud is always prone to attack mainly due to the anonymous, virtual, shared, replicated, multi-node storage, and execution. Unauthorized access or manipulation leads to data theft or data loss. Big Data outsourcing from data owners to the cloud is in three phases as follows: Authentication, Compression and Encryption. The volume of data in the globe is increasing with each day. The application of the internet, smartphones, and social networking sites has resulted in an increase in data. Big Data consists of data sources that are both vast and complicated. Petabytes and Exabytes are the most common data sizes. Regular protection and security processes are deficient and therefore unfit to adjust to the quick colossal measure of information in such a muddled registering climate because of the inborn information volume and qualities of Big Data, specifically speed, volume, and an assortment related with enormous scope public clouds and the Internet of Things (IoT).

Large-scale information management and analysis tools that go beyond the capabilities of conventional data processing tools are referred to as Big Data analytics. Large Data design is distributive in nature, increasing to huge number of information and handling hubs. Among these a huge number of hubs, the information gets parcelled, imitated and dispersed for strong calculation, and in light of execution reasons, information is likewise portioned into classes. Highlights like auto-tiering, continuous handling and gushing of information have been significant patterns in large information examination. The review gives practical arrangements created from cutting edge innovations like information respectability, cryptography and validation, information examination, and blockchain, featuring their connected certifiable application for Big Data security and the issues they go up against. My analysis in this paper provides an example of how Big Data may be secured. This work would clear the way for future scholarly investigations into this basic area of information security.

Keywords: Big Data, Big Data Security, Big Data Security challenges, Data Hacks, Data theft, Big Data Analysis

Velká data a jejich bezpečnostní výzvy

Abstraktní

V této studii jsem navrhl práci sestávající ze čtyř entit včetně Trusted Center (TC), Data Owner (DO), Data User (DU) a Cloud Server (CS). Data uložená v cloudu jsou vždy náchylná k útoku především kvůli anonymnímu, virtuálnímu, sdílenému, replikovanému, víceuzlovému úložišti a provádění. Neoprávněný přístup nebo manipulace vede ke krádeži nebo ztrátě dat. Outsourcing velkých dat od vlastníků dat do cloudu probíhá ve třech následujících fázích: ověřování, komprese a šifrování. Objem dat na světě se každým dnem zvyšuje. Aplikace internetu, chytrých telefonů a sociálních sítí má za následek nárůst dat. Velká data se skládají z datových zdrojů, které jsou rozsáhlé a komplikované. Nejběžnější velikosti dat jsou petabajty a exabajty. Pravidelné procesy ochrany a zabezpečení jsou nedostatečné, a proto se nehodí přizpůsobit rychlému kolosálnímu množství informací v tak zmateném registrujícím klimatu kvůli vrozenému objemu informací a kvalitám velkých dat, konkrétně rychlosti, objemu a sortimentu spojeného s obrovským rozsahem veřejnosti. cloudy a internet věcí (IoT).

Rozsáhlé nástroje pro správu a analýzu informací, které přesahují možnosti konvenčních nástrojů pro zpracování dat, se označují jako analytika velkých dat. Návrh velkých dat je svou povahou distributivní a zvyšuje se na obrovské množství informačních a manipulačních center. Mezi těmito obrovskými centry jsou informace rozděleny, napodobovány a rozptýleny pro silný výpočet a ve světle důvodů provedení jsou informace rovněž rozděleny do tříd. Nejvýznamnější prvky, jako je automatické vrstvení, nepřetržité zpracování a proudění informací, byly významnými vzory při rozsáhlém zkoumání informací. Recenze poskytuje praktická opatření vytvořená na základě špičkových inovací, jako je důvěryhodnost informací, kryptografie a ověřování, zkoumání informací a blockchain, včetně jejich připojené certifikovatelné aplikace pro zabezpečení velkých dat a problémů, se kterými se setkávají. Moje analýza v tomto článku poskytuje příklad toho, jak lze zabezpečit velká data. Tato práce by uvolnila cestu pro budoucí vědecká zkoumání této základní oblasti informační bezpečnosti.

Klíčová slova: Big Data, Big Data Security, Big Data Security výzvy, Data Hacks, Krádeže dat, Big Data Analysis

Table of Content

1. INTRODUCTION 1

1.1 Overview 1

1.2 Big Data Application 3

1.3 Challenges of Big Data 4

1.3.1 Data Storage & Access Issues 4

1.3.2 Management Issues 4

1.3.3 Processing Issues 5

1.3.4 Security 6

2. OBJECTIVE AND METHODOLOGY 8

3. LITERATURE REVIEW 9

3.1 Introduction 9

3.2 What is Big Data? 9

3.3 Benefits of Big Data 11

3.4 Challenges in Big Data 13

3.4.1 Challenges in Managing Storage 13

3.4.2 Transmission and Sharing Challenges 14

3.4.3 Implementation Challenges of Big Data 14

3.4.4 Analytical Challenges 15

3.4.5 Big Data Security and Privacy 16

3.4.6 Cloud Computing and Security Issues 21

3.4.6.1 Need of Cloud Computing 24

3.4.6.2 Example of Cloud Computing Application 25

3.4.6.3 Cloud Sharing and Infrastructure 27

3.4.6.4 Data Security and Privacy Protection in Cloud 29

3.5 Big Data Security and Privacy Solution in Literature 30

3.6 Discussion of Big Data Privacy Issues and Challenges in Literature 37
4. PROPOSED TECHNIQUES FOR BIG DATA SECURITY 42
4.1 Authentication 45
4.1.1 Hashing Algorithm 47
4.2 Compression 49
4.2.1 Ma for Data Compression 49
4.3 Encryption 50
4.3.1 Asymmetric Encryption 51
4.3.2 SALSA20 Algorithm 53
4.3.3 SALSA20 with Map Reduce 54
4.3.4 Map Reduce Programming Model 55
5. DISCUSSION AND CONCLUSION 57
6. REFERENCES 60

List of Figures

Figure 1 The principal obstacles to Big Data Security (Moreno, Serrano et al. 2016)..... 6
Figure 2 Big Data Architecture with Real time streaming (Bhandari, R., Hans, V. and Ahuja, N.J., 2016) 10
Figure 3 Big Data Security and Privacy Area (Bashari et al., 2016)..... 20
Figure 4 Work Flow of Big Data Enabled Cloud Environment security systems (Mishra et al., 2022) 43
Figure 5 Proposed System Architecture with Big Data Outsourcing, Sharing, and Management (Narayanan et al., 2020)..... 44
Figure 6 Data User Registration with Trusted Center (Narayanan et al., 2020) 46
Figure 7 User Authentication with Trusted Center (Narayanan et al., 2020)..... 46
Figure 8 Time is taken for Key Generation and Authentication 48
Figure 9 The process of Encoding and Decoding..... 50
Figure 10 Symmetric Key Cryptography Process 51
Figure 11 Public Key Cryptography Process 51
Figure 12 SALSA20 Algorithm (Narayanan et al., 2020) 54
Figure 13 Encryption time for different algorithms in Map Reduce 56

List of Tables

Table 1 A summary of Big Data Security and Privacy Challenges presented by Oguntimilehin and Ademola (2014)..... 13
Table 2 Performance Test for Cipher Suites 55
Table 3 Encryption time for different Algorithms in Map Reduce 56

1. INTRODUCTION

1.1 Overview

John Mashey popularised the term Big Data, which has been in usage since the 1990s. Big Data often consists of data sets that are too large for frequently used software tools to capture, curate, manage, and process in a reasonable amount of time. The concept of "Big Data" is already nearly ubiquitous in our daily lives. The phrase "Big Data refers to a wide variety of huge data sets that are nearly hard to manage and analyse using standard data management methods - not just because of their size, but also because of their intricacy. Big Data may be observed in economics and banking, where a huge quantity of stock market, financing, digital and onsite buying information travels via computer networks nearly every day, where it is then collected and kept for managing inventory, consumption patterns, and financial markets. An increasing number of organizations are employing the innovation to collect and manage Exabyte's of data, such as site logs, click stream data, and media platforms material, in order to acquire a deeper understanding of their consumers and their organization. As a result, information categorization is becoming increasingly important. A majority of software companies have already been developing Big Data apps and services with the goal of bringing the potential of data analysis to the consumers (Adnan, N.A.N. and Ariffin, S., 2018).

Numerous research investigations have revealed multiple advantages of Big Data technologies. Related literature evaluations on Big Data security, on the other hand, show that hostile hackers pursuing large data are on the climb. In the Big Data world, the primary challenges and strategies around security threats and personal privacy have yet to be thoroughly investigated. These problems inspire new inventions and research efforts to uncover unresolved concerns that will lead the way for future research and practise. In order to fully comprehend Big Data's fundamental concepts, security concerns, and possible methods, this research first explores its many facets. Also, it deals with the security and privacy issues that Big Data is facing in real world problems and how to mitigate and control the risks related to Big Data through current and upcoming technology (Alhanahnah et al., 2018).

Unstructured data is the major focus of the Big Data philosophy, which also includes semi-structured and structured data. As of 2012, the "scale" of Big Data ranged from a few dozen gigabytes to many zettabytes of data, making it a shifting target. To extract insights from

diverse, complex, and enormously scaled data sets, Big Data necessitates a collection of approaches and technologies with various forms of integration.

Some organisations modify it by adding "variety," "veracity," and several other "Vs"; this revision is contested by some industry leaders. The "3 Vs," "4Vs," and "5Vs" were common names for the Big Data Vs. They highlighted the volume, variety, velocity, veracity, and value of Big Data. Big Data frequently includes variation as an additional quality. (Sahafizadeh and Nematbakhsh. 2015)

As the meaning of enormous information has been laid out in the past segment, it is currently essential to represent its attributes. Huge information is recognized by its attributes from conventional advances. Three key huge information highlights, volume, variety, and velocity are generally eluded as 3Vs. It developed from 3V to 7V and added esteem, inconstancy, intricacy, furthermore, veracity (Ding et al., 2017).

- Information Volume: In large information, the word huge is a direct result of the gigantic size of enormous information that alludes to the immense amount and extent of information that is created consistently, moment, hour and day in our computerized world. Virtual entertainment, monetary foundations, clinical establishments, government organizations, sensors, logs which create terabytes of information consistently. Large information size is in the request for Terabytes (TB), Petabytes (PB), Zettabytes (ZB) and Exabytes, separately.
- Information Velocity: Velocity in large information is an idea that arrangements with the information speed from various sources and the rate at which it is totalled and communicated. It likewise alludes to the speed at which the information should be handled to fulfil the interest. Large Data examination's goal is to handle the information progressively to match their creation rate as it is produced.
- Information Variety: Data blast caused unrest in the sorts of information designs. The assortment of information alludes to the different sorts of information put away, broke down, and utilized. There is no predefined Big Data structure; it is completely not quite the same as customary information. It very well may be organized (for instance, exchange information, calculation sheets, interface data sets), semi-organized (for instance, web server logs and XML) and unstructured (for model, posts via web-based entertainment, sound, video, pictures).

- Information Veracity: Veracity alludes to the unwavering quality and consistency of the information being investigated and portrays whether information is substantial for examination. The separating of experiences cannot be founded on huge volumes of information that isn't definite or legitimate. Since all information created and consumed into the large information stage isn't depend ably ensured to contain clean information.
- Information Value: Value alludes to the nature of put away information and tests the significance of information for navigation. It is the main component of any enormous information-based application since it empowers valuable business data to be produced. The whole information contained in the data set isn't valuable for everybody and consequently it is critical to separate the helpful piece of the data set.
- Information Visualisation: Data the executives in an association, especially from assorted sources, is an extremely confounded task. The information must be connected, associated, and related with different sources to comprehend the data that should be conveyed. Information intricacy delineates the trouble of managing different information sources, i.e., interfacing, cleaning, and changing them prior to handling.
- Information Variability: notwithstanding the rising information speeds and assortments, information streams with occasional pinnacles can be profoundly conflicting. Information changeability alludes to the different information stream rate with various pinnacles and conflicting information speed. (Agrawal, D., Budak, C. and El Abbadi, A., 2011).

1.2 Big Data Application

The utilization of Big Data applications increments over the long haul because of the simple administration of tremendous amounts of information. It uncovers data and information about purchasers, providers, what's more, other business partners, markets and exercises, the underlying drivers of difficulties and costs, and the potential dangers that the organization might confront. This multitude of realities and bits of knowledge would be covered up in any case. It can determine forecasts of future patterns and open doors through newfound way of behaving and designs, which will improve functional and strategic decision-production so as to speed, quality, and importance. Large information permits information to be utilized actually for business benefit and value creation. Enormous information examination is progressively being utilized both by people in general and confidential areas (Fan et al., 2015).

1.3 Challenges of Big Data

The assortment, combination, handling, and examination of Big Data makes a few difficulties in different fields. The large information investigation incorporates a scope of stages, including information handling and assortment, data extraction and cleaning, information mix, collection, and portrayal. Challenges like heterogeneity, scale, practicality, intricacy, and security are presented in every one of these stages. These different challenges and problems associated with bringing this technology and adapting it must be addressed. In this section, I present various challenges of Big Data.

1.3.1 Data Storage & Access Issues

When new storage devices were invented, the amount of data which can be stored expanded. Vast quantities of data are produced, generated by everyone connected to digital devices which creates the need for enormous storage. Also, stored data should be available to access in the correct format at any time so that analysis can be carried out. Cloud storage is the solution to this problem where multiple servers are used to achieve ample storage in the distributed mode and to have high storage capacity remotely.

1.3.2 Management Issues

The data sources vary by size, format, and collection method. Big Data management means maintaining data integrity, transparency, control, access, and documentation of massive data sets. Resolving access, use, update, governance, and reference problems have proved to be significant obstacles. In the event of any failure, large clusters in clouds must be managed effectively (Ghosh, N., Ghosh, S.K. and Das, S.K., 2014).

- Scale: In Big Data, data volume grows at a very fast pace while computing resources speed is static; volume is scaling faster than computational resources. The challenge is to manage large and quickly increasing data volume with existing resources. The management of increasing volumes of data is not possible with traditional software tools. Also, there exists the problem of data retrieval & analysis due to the constraint of scalability and data complexity. To process these data, new innovative methods or techniques are required.
- Fault Tolerance: Since the advent of emerging technology such as cloud computing and Big Data, it is often expected that the damage done will be within appropriate limits if

the failure happens, rather than beginning the whole process from scratch. Fault-tolerant computation, involving complex algorithms, is extremely difficult. It is simply impossible to design machines or software that are fool-proof, 100% reliable, and tolerant to a fault. Therefore, the key goal is to reduce the risk of failure to an "acceptable" level.

1.3.3 Processing Issues

Big Data also faces problems with processing. Therefore, improved parallel processing and modern computational algorithms are required to provide rapid information and efficient processing of big (Goh, E.J., 2003).

- Heterogeneity and Incompleteness: Big Data analysis has difficulties both because of its large scale and mixed data that are available based on different patterns in data collected. Multiple sources produce organized, nearly fully, and unstructured and semi - structured. They have no set structure and are quite dynamic. When experimenting with Big Data technology, working with this varied data essence is a significant problem. The handling of such different data formats is also very time consuming and expensive. Therefore, before the analysis, the data must be carefully organized. This transformation of data into a standardized form for subsequent review is a significant challenge in Big Data mining. Moreover, deficient information makes vulnerabilities during information examination that should be overseen during the review (Ateniese et al., 2006).
- Timeliness: As the size of informational collections to be broke down builds, handling can take more time. In any case, in certain conditions, the results of the examination are required right away; for instance, in the event that a transfer via Mastercard is thought to be deceitful, it is ideally fundamental for the exchange platform to forestall the transaction before it is concluded. It won't be allowed to properly analyse a user's shopping history in real-time. Therefore, it is important to develop preliminary results in advance so that a quick conclusion may be reached with just a tiny amount of additional calculation using fresh information.
- Data quality: In general, Big Data are aimed at storing reliable data, not having incredibly large, meaningless data to obtain better results and conclusions. That poses numerous questions, such as how to ensure that the data is valid, how much data will suffice for decision making, and whether the data stored is reliable or not, etc. Usually,

a good process can make poor decisions if it is based on bad data. But due to the volume, it is not practical to validate each data item. New approaches are needed to qualify and validate data.

1.3.4 Security

Security is one of the main pressing concerns in the data innovation industry. Large information gives tremendous benefits and chances to end-clients, yet it is likewise answerable for the issues of safety. It contains large volumes of private data; maintaining high-security standards is critical. The data will have to be authenticated, encrypted, and authorized. There are many challenges in managing security in a wide range of data. Another significant issue is data privacy and one that is growing in the sense of Big Data. Effectively handling privacy is both a technological and a sociological challenge that must be tackled simultaneously from all sides to fulfil Big Data's promise. To solve this security and privacy issue, precise techniques and suitable algorithms must be established to improve data protection (Herodotou et al., 2011).

According to the Big Data Working Group of the Cloud Security Alliance organisation, infrastructure security, data privacy, data management, and integrity and reactive security are the main four different components of Big Data security. The International Organization for Standardization used this breakdown of Big Data security into four main topics to develop a security standard for Big Data.

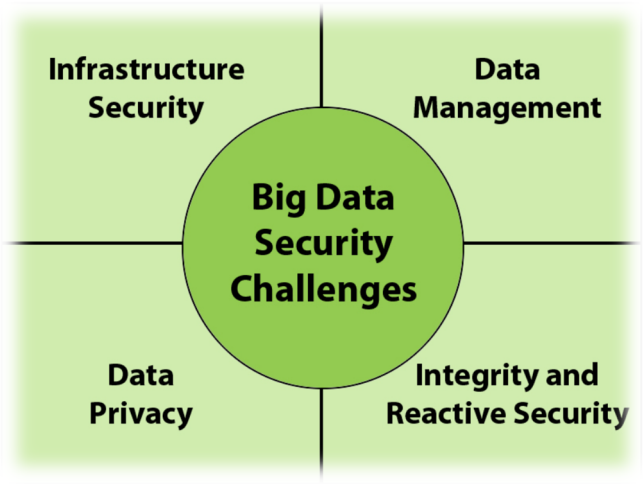

Figure 1 The principal obstacles to Big Data Security (Moreno, Serrano et al. 2016)

Every new disruptive technology introduces fresh problems. In the context of Big Data, these problems involve not just the quantity or diversity of data but also the quality, privacy, and security of the data. The privacy and security of Big Data will be the main topics of this essay.

Big Data not only enlarges the scope of the privacy and security concerns that are addressed in traditional security management, but also generates new ones that require a different methodology of approach. More restrictions are required to address these issues as organisations and governments keep and analyse more data. As a result, ensuring Big Data is secured has emerged as one of the most significant obstacles that could hinder the development of technology, without sufficient security guarantees. Thus, a huge responsibility comes with Big Data to garner the necessary degree of trust.

2. OBJECTIVE AND METHODOLOGY

Big Data is a topic of more and more interest in the world. All companies, big and small, need to be at least aware and ready to adopt Big Data for commercial reasons (if they have not already done it). However, like any new technology, Big Data also has its own issues. The biggest issues are related to security and privacy. People are producing more data and companies are storing more customer data and using them for new Big Data analytics for business reasons. This means there is more data than before which is at the risk of being accessed by unauthorised people, stolen or abused. From the literature review, I believe that security is the biggest problem with Big Data so I have focused on that and the solutions in literature.

The objective of this literature is: To research and find out the different types of security issues with Big Data and some of the solutions.

The methodology is: to study a large amount of Big Data research papers and understand the security and privacy problems and their common solutions. After this, I will also extend one of the solutions in the next section to find an efficient way to encrypt information for protection in Big Data system.

The outcome of this thesis will be to create an overview of Big Data security and privacy issues in the literature and to make one practical recommendation for using a more efficient encryption system. This is discussed in detail in the next system.

3. LITERATURE REVIEW

3.1 Introduction

This section reviews the literature on Big Data, with a focus on the security challenges and suggested solutions. The review will go into some detail on cloud computing-based solutions since these are common in many papers, and it provides the infrastructure to put some other solutions into action. First, there will be an overview of security issues most commonly found in Big Data, then a brief description of different solutions. Then, I will discuss some of the cloud computing solutions which are the most popular ones. Finally, I will write about some unique techniques and software used to overcome Big Data security challenges. This software will be discussed in detail in my solution in this thesis. The final section of this literature review will summarise the relevant key points for discussion in the following sections.

3.2 What is Big Data?

Throughout the last several decades, Big Data has been increasingly popular in virtually every industrial sector, and tremendous work has gone into developing new approaches for Big Data analytics. While many businesses recognise the importance of Big Data analytics, organisations are still in the early stages of realising its benefits because they must redesign their business operations and architectures in order to accommodate the enormous and rapidly increasing volume of data that is being generated (Venkatraman, S. and Venkatraman, R., 2019).

As a result of the proliferation of Big Data, the way data is managed and used is evolving. Healthcare, traffic management, finance, retail, education, and a number of other industries are among the uses. As a result of this tendency, businesses are become more flexible and open. Unavoidably, new data types will result in new problems. This research aims to raise awareness of significant Big Data issues. This article will discuss a wide range of Big Data-related issues. The topics addressed include defining Big Data and going over the many standards used to describe it. Velocity, volume, and variety the three Vs of Big Data are all included in this categorization (Bhandari, R., Hans, V. and Ahuja, N.J., 2016).

Big Data is a group of data sets that are so numerous and intricate that they become challenging to process using typical data processing software or readily available database management tools. Capture, curation, storage, search, sharing, transfer, analysis, and visualisation are among

the difficulties. Big Data is a term used to describe novel database administration and analytical techniques created for the analysis, storing, and manipulation of big or complicated data sets. Big Data investments can be made in people (such as data scientists) as well as in business and technological solutions, such as database management systems (such as Hadoop, IBM/Netezza), analytics and visualisation tools (such as Revolution R), or text-processing and real-time streaming options. Big Data is also used to describe datasets that are too large to be captured, stored, managed, and analysed by conventional database software tools. The size of a dataset required to qualify as Big Data is not specifically defined. In order to manage the Big Data phenomenon, new technology must be implemented. Big Data technologies are described by IDC as a new generation of tools and architectures that enable rapid data capture, discovery, and analysis in order to cheaply extract value from very large amounts of a variety of data. Big Data is data that can be processed more quickly than by traditional database systems. The data does not match the structures of the current database architectures because it is too large, moves too quickly, or both. There must be a different method of processing these facts in order to extract value from them.

Big Data Infrastructure is dispersed in design, with the potential to expand to millions of data and computing units. The data in a Big Data platform is split, duplicated, and dispersed over thousands of servers. Data is split into two types for performance reasons: hot data and cold data, resulting in a useful aspect of Big Data architecture known as auto–tearing (Eldawy, A., Levandoski, J. and Larson, P.Å., 2014).

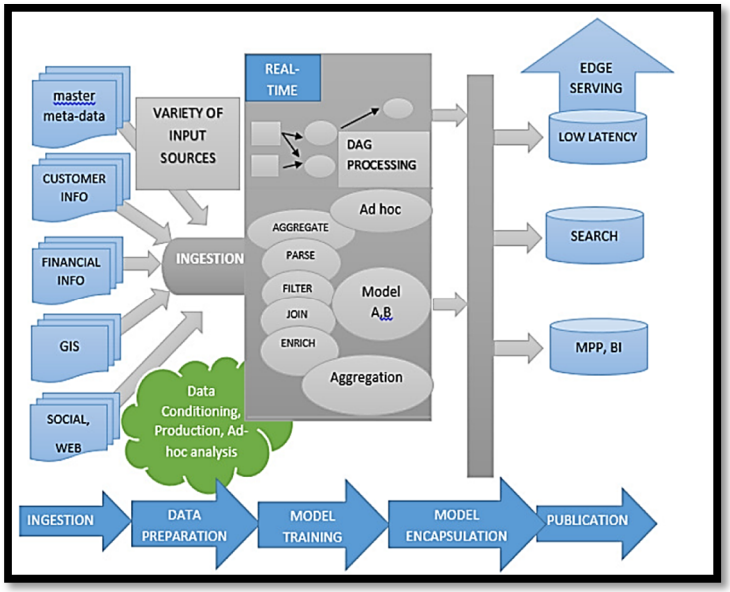

Figure 2 Big Data Architecture with Real time streaming (Bhandari, R., Hans, V. and Ahuja, N.J., 2016)

Present day Big Data analytics' design empowers ongoing calculation (continuous examination), information gathering from different info sources figure 2 and customary information move to huge information arrangements, in addition to other things. Impromptu quests, as well as huge parallelism and an intricate programming style, accommodate more noteworthy versatility and adaptability. There are a variety of Big Data processing frameworks to choose from. Hadoop makes use of the Map Reduce structure, in which a programme is divided into multiple maps, each of which is executed on a large number of data nodes before the results are merged into a single result set. Topology-based computing is yet another important paradigm in the field of massive data processing and computation. Storm, a large data tool for real-time analysis, makes use of this technique. Storm makes use of a network structure that is made up of Spouts and Bolts. Spouts are used as data sources, whereas Bolts are used as data processing nodes (or nodes that handle data). It is possible to use a variety of frameworks and topologies for data processing and administration in the context of Big Data computing.

3.3 Benefits of Big Data

The development of computation technologies and the capture of large volumes of data has resulted in the development of storage technologies that are high in capacity, efficacy, and long-term viability. There are as a result storage choice that are based on load balancing or scale-out technology. This type of extensibility is most suited for situations in which it is hard to predict changes in storage requirements; rather than acquiring anticipated storage in order to meet brief use requirements, more capabilities can be added as needed. Because each node has its own storage and backup capabilities, adding networks that function in parallel enhances the efficiency, capability, and bandwidth available to the system as a whole Big Data has the potential to alter not just research, but also education and other fields.

For example, using knowledge extracted from Big Data, educators may create the most efficient teaching methods, which can range from simple reading, writing, and math to complex college-level courses. Although we are still a long way from having access to such data, there are many signs that this is the way things are moving. There has been a significant change in recent years toward the widespread deployment of educational activities online, which will provide an ever-increasing amount of precise data about students' performance (Oguntimilehin, A. and Ademola, E.O., 2014). It, as a rule, is accepted that the sending of data innovation may

extraordinarily diminish the expense of medical services while at the same time work on the general nature of therapy by making care more preventive and customized.

Another example of advantage of Big Data is for individual firms facing competition and development will find that the use of Big Data will be a vital basis for their success in the future. All firms must take Big Data seriously if they are to realise its full potential, according to the perspectives of competitiveness and prospective capture. Both established competitors and newcomers compete in the vast majority of industries. Both companies will innovate through the use of data-driven strategies. Deep to real-time information is sought after, and value is extracted from it.

Big Data analytics are used by small and medium-sized enterprises to find patterns and gather useful data from the vast volumes of data at their disposal. Big Data analysis not only helps with understanding the data but also with locating the information that will be most crucial to the company's future actions and decisions. The major goal of Big Data Tools is to assist organisations in making better decisions based on their data. Big Data Analytics enables Computer Scientists, Business Analysts, and other machine learning professionals to get access to huge numbers of data records that would otherwise go unnoticed by standard business processes, such as sales and marketing. Big Data Architecture is built on a set of capabilities that allow for the establishment of long-term, trustworthy, and completely automated data centres, among other things. In part due to the fact that Big Data is still a relatively new topic, there is no "standard Big Data framework" that has been in use for an extended period of time (Toshniwal, R., Dastidar, K.G. and Nath, A., 2015).

Overall, Big Data analytics can be used by small and medium-sized enterprises to find patterns and gather useful data from the vast volumes of data at their disposal. Big Data analysis not only helps with understanding the data but also with locating the information that will be most crucial to the company's future actions and decisions. The major goal of Big Data Tools is to assist organisations in making better decisions based on their data. Big Data Analytics enables Computer Scientists, Business Analysts, and other machine learning professionals to get access to huge numbers of data records that would otherwise go unnoticed by standard business processes, such as sales and marketing. Big Data Architecture is built on a set of capabilities that allow for the establishment of long-term, trustworthy, and completely automated data centres, among other things. In part due to the fact that Big Data is still a

relatively new topic, there is no "standard Big Data framework" that has been in use for an extended period of time. A Big Data design must take into account factors such as latency, volume, velocity, diversity, veracity, ad hoc query capabilities, sustainability, robustness, and fault tolerance (Toshniwal, R., Dastidar, K.G. and Nath, A., 2015).

3.4 Challenges in Big Data

The following section briefly introduces some of the general Big Data problems (storage, sharing and analytics). However, the focus is on security and privacy challenges which are discussed in more detail and the first three issues are analysed in the context of security challenges.

Big Data Security & Privacy Challenges		
Data Acquisition	Data Storage	Data analytics
Collecting right data • Relevant data sources • Real- time processing • Checking endpoints	Providing right data • Data integrity • Data access controls • Backup & recovery	Performing right actions • Timely decisions making • Credible judgments • Contextual insights

Table 1 A summary of Big Data Security and Privacy Challenges presented by Oguntimilehin and Ademola (2014)

3.4.1 Challenges in Managing Storage

The huge volume and quick proliferation of Big Data distinguish it from other types of data. The quantity of data has increased from the Petabytes level to the Exabytes level, and it is continuing to expand in response to the needs of prospective implementation and business expansion, and it is soon nearing the Zettabytes level (Zettabytes = one trillion bytes). Dispersed data kept on cloud servers is difficult to secure, and this is especially true for sensitive information. Using data directly on a cloud server is difficult for users; moreover, they have no way of ensuring that the data is not accessible by unauthorised persons. Account passwords are used for data security in management in order to secure the security of cloud-based servers. Administrators of cloud servers with substantial power, on the other hand, may collaborate to share critical Big Data (Bashari et al., 2016).

Examination cycle and social worries. The first test shows up in quite a while of protection. The most sensitive issue is the protection, which has mechanical, legal, and calculated implications. With reference to extensive information, this concern becomes more relevance. The International Telecommunications Union defines protection as the "right of persons to control or effect what data related to them could be revealed" in its broadest definition. From a broader standpoint, security may also be seen as including both corporations looking to protect their seriousness and customers and stages seeking to preserve their power and residents. In both these translations, security is a general worry that has a large number of suggestions for anybody wishing to investigate the utilization of enormous information for improvement with regards to information procurement, capacity, maintenance, use and show (Bao, R., Chen, Z. and Obaid at, M.S., 2018).

3.4.2 Transmission and Sharing Challenges

Users' secret communications are not utilised or exchanged in accordance with standards and monitoring, with the majority of companies depending on their own self-discipline. As a result, customers are unable to determine the purpose of their confidential communications. As previously said, the large volume of Big Data increases transmission time and increases the risk of personal data being compromised. Moreover, to achieve business greatness and seriousness, numerous associations would connect as buyer and business information sharing, which would raise the gamble of client security being presented to outsiders. The sharing of information over a wide reach, as I would like to think, isn't generally a helpful turn of events. It would be a disaster for Big Data if there was no safe environment and no reliable technology to work with it (Kaisler et al., 2013).

3.4.3 Implementation Challenges of Big Data

Due to their inherent features, Big Data faces multiple problems as new security tests are introduced. It is difficult to enforce safety measures for large quantities of data. The data encryption and decryption protection method reduce efficiency; also, data processing slows in case of fast streaming. It is challenging to maintain quick response time while maintaining privacy and protection. Furthermore, highly heterogeneous data types pose a problem for the privacy and security of Big Data; in particular, the integrity of data is undermined by the declaratory attitudes, diversity of collection points, different data formats, and the activities of many fake internet identities (Burdonov, I., Kosachev, A. and Iakovenko, P., 2009).

Therefore, privacy and safety measures for this ambiguous data are challenging to enforce. Big Data clusters contain data that represent fluidity quality and are spread across many servers. For parallel computation, data processing is entirely dependent on resource availability. In such cases, clusters are challenging to control the user and to determine the data access policy. The development of a fool proof security measure for such an extensive scale and volume is extremely complex and complicated. While several new techniques and methods have been developed, they are slowed down to some degree when large quantities of data are involved. Strong security measures are required to secure, track, and audit Big Data processes that include solutions for user authentication and access verification, policy enforcement, and data encryption. Security measures are necessary to bring security controls closer to the sources of data and data themselves so that protection can be assured at the source of data (Jin et al., 2011).

3.4.4 Analytical Challenges

Big Data analysis is influenced by the structure of the data, which can be classified as organised, semi-structured, or unstructured, among other things. Big Data, in whatever form it takes, has characteristics associated with volume. A consequence of this is that scaling Big Data is the most challenging aspect of data processing to deal with. When working with large amounts of data, data encryption is an excellent strategy for maintaining privacy. The sheer volume of data, on the other hand, increases the security load. Aside from that, the dangers of data analysis are heightened as a result of scattered storage of Big Data. For continuous big information applications, for example, person to person communication locales, modern robotization, migration, statistical surveying, etc, the necessity for idealness for information stream handling represents a more prominent test to encryption approaches and security insurance (Tan, J., Meng, X. and Zhang, L., 2013).

Distributed computing execution this is a Business to business (B2B) has more layers of members for that reason there is a more significant level of intricacy. At the point when the information regulator is cloud administrations client, the shortfall of control diminishes their capacity to conform to legitimate commitments connected with their own information insurance. The association between the obligation and the activity of information control in these commitments. In the event that the information can't be confined cloud clients go about as information regulators, it is normal that the specialized parts of the issue are the absence of control from cloud suppliers. Utilization of ill-advised measures should safeguard the

information and the business administrators. In the event that multiple clients of the information gave, cloud probably won't meet the cloud regulator.

3.4.5 Big Data Security and Privacy

There have been new challenges in terms of data security as a result of the rise of large amounts of data (also known as Big Data). Investigational technology that can manage vast volumes of data while also keeping it secure is becoming increasingly important. Even when dealing with massive amounts of data, data security systems are slow to respond. Two of the most important security vulnerabilities in Big Data are the absence of security measures and the absence of the use of encrypted technologies to access data, such as cryptography (Benjelloun, F.Z. and Lahcen, A.A., 2019).

In these cases, there are hazards associated with the removal, insertion, or change of data. In addition to these factors, there are other network security issues to consider, namely in terms of how the connection is being used. These issues are connected to and have prompted the creation of fault management and power management approaches, among other technologies. The difficulty in maintaining a high rate of system operation is one of the results. Another challenge is manageability, as well as data visualisation and analysis. It is necessary to develop massive scaling models that improve the horizontal scalability of handling and preparation across a wide variety of architectures and platforms. On the other side, linkage with legacy systems attempts to lessen the alienating impact of a programme by monetizing older equipment, analytical processing approaches, and enhancing data availability. The use of legacy systems in cloud settings is one of the more recent themes to be discussed.

Due to the rising recognition of the value of Big Data, the issue of confidentiality is becoming increasingly significant, particularly to consumers. Some individuals continue to be worried about where and how their personal information is handled, particularly if it may be negative or hurtful to them. If an organisation loses control over its data, private information, essential intellectual property, and commercially sensitive information are just a few types of data that might swiftly become 'hazardous'. Protecting one's privacy is a huge issue, and it is getting much more problematic in the setting of Big Data. There are stringent restrictions governing what I may and cannot do in our daily lives. The requirements for other types of data, such as those in the are less rigorous. However, there is significant worry about the unlawful use of

private information, particularly when it is linked to other data sources (such as social media). Integration of Big Data into enterprises, regardless of their size, has proven to be one of the most difficult operational difficulties faced by organisations today. The use of this new wealth of knowledge is still a work in progress for many firms, while others are obviously profiting from it, albeit in a limited and constricted way. When building a Big Data management system, it appears that the incorporation, modification, dependability, and regulation of Big Data are all important subjects to consider and handle.

Security alludes to information assurance from unapproved access or a catastrophe of some sort. Protection centres around the utilization and treatment of private information like setting up techniques to guarantee that delicate data about clients is gathered, put away, and utilized in adequate ways. Information security is the Confidentiality, Integrity, and Availability of information (CIA), by the standard definition. It is the procedure to ensure that data stored is protected from unauthorized access and usage, that the information is reliable and accurate, and that it is available for use when necessary (Amudhavel et al., 2015).

Big Data is an immense resource that has generated great business and social opportunities in all fields, helping us to uncover previously obscure trends and gain new perspectives for informing and guiding decisions. At the same time, securing individuals' and organizations' information from cyber threats has become an urgent priority. Because the information previously unusable by the companies is highly valuable and must be secured, subject to privacy laws and enforcement regulations. The scale, assortment, and speed of large information amplify the security and protection issues. Joined with the streaming idea of information assortment, the information sources, arrangements, and information stream, and the huge volume presents explicit security dangers and tremendous difficulties (Jenkins, C., Schulte, M. and Glossner, J., 2010).

Big Data analysis tools were initially primarily tested for their speed and reliability. These are developed to manage large data sets, often to which adequate security or privacy measures are not included. Efforts to secure data are only spent on a second glance. It has become highly vulnerable and more susceptible to malicious attacks due to rapid development and progress in all fields of technical perspectives. Traditional protection measures are not enough to respond to these challenges. These attacks can harm the essential properties of information system security. Since a large proportion of the data generated includes sensitive private data, security

should be considered above all when this vast amount of data is stored and processed. More disturbing is that security risks are often proportional to the quantity of data prone to attack. It can quickly become a significant issue without ensuring adequate security measures; can damage users' privacy if not correctly managed (Tan, J., Meng, S., Meng, X. and Zhang, L., 2013).

The Characteristic Based Encryption (CBE) was initially trait-based encryption for the execution of the entrance control of the utilization of public key cryptography. Its primary capacities incorporate giving adaptability and fine-grained admittance controls to guarantee that the entrance control is scrambled in the distributed computing, the ABE is broadly utilized. The client's confidential key and the code text in the CBE plot connected to the ups also, downs different properties. For instance, it is viewed as a vital standard and straightforwardness has obviously referenced that Article 10 of the EU Data Protection Directive (Bashari et al., 2016). As a component of this strategy, information regulators are expected to illuminate the information subject gives data handling exercises and personality and explanations behind handling. Likewise, structures the premise of the guideline of straightforwardness in different terms. For instance, in article 12 (a) currently said before that the information regulator should affirm the information subject information control. There will be no harm or postpone in conveyance including well defined for the subject of the Individual information handling. Assuming keeping up with the degree of control, cloud is important, specialist organization and the client should be straightforward. There ought to be a cycle moreover gives cloud supplier means and measures (Aditham, S. and Ranganathan, N., 2017).

The pertinent administrative specialists there must likewise be straightforward. The straightforwardness in distributed computing gives expanded hazard of encroachment since it contains different determinations. For instance, information handling including subcontractor's chain. Individual information access can be given to those gatherings to subcontractors and providers of endlessly cloud client. In their exercises, they could deal with the individual information and should consent to the EU Data Protection Directive. It would be excessively costly and intricate, managerial and specialized control of the foundation of the sub-workers for hire and working techniques (Reddy, Y., 2018).

The protection of personal data may unquestionably be compromised by non - linear and non-cloud-based strategies and computing stacks, which seem to be widespread, large, and

technically advanced. A most significant thing is to segregate asset the committee optimal protection from the virtualization layer in order to provide security to protect. The simple technique is to start reversing sustainable virtualized manageability and security inspecting. The design that should trigger a modest call cloud visor log jam for I/O-sensitive applications is now being run. Most cloud platforms with a standard virtualization basis work to minimize costs and enhance efficiency while also portraying the safety of the recruit from harm to the virtual machine.

For every exchange, having enough information and registering security issues are important. There is a considerable helping quality in it. However, since the client doesn't have their data stored locally on the necessary frameworks and tools to gainfully and reliably check required data, the QoS needs and problems generated by online Cloud are much more complicated. (Vorugunti, C.S., 2016).

The cloud that is supposed to fail to influence the customer is technically independent of the computer system collecting and generating knowledge and the increased likelihood of managing customers.

I. Uprightness: Not just did it imply the information honesty of the respectability of the estimations. It noticed that the information ought to tell the truth in legitimate capacity and identify unapproved adjustment, erase, and the corruptness.
II. Confidentiality and Privacy: It focuses on the Confidential or Private assets are not to be unapproved access isn't seen.
III. Availability: Cloud administrations need to guarantee that they work appropriately and clock and open on request. IaaS physical and virtual assets should be accessible to support information.
IV. Sanitization: Data cancellation has been a worry of its observing, checking and following components have been utilized to find data (Zhao, Y., Li, S. and Jiang, L., 2018).

The following section details the current work in literature related to Big Data privacy and security:

Traditional security and privacy solutions are unable to adequately handle the changes that Big Data has brought to the digital world, including the volume of data that is collected, stored, and altered as well as the manner in which that data is altered. Complex encryption techniques,

access control barriers, firewalls, and interruption identification frameworks are just a few examples of organizational security measures that may be broken. Even anonymized data may be re-identified and linked to a single client for malicious purposes (Jiang, R., Lu, R. and Choo, K.K.R., 2018).

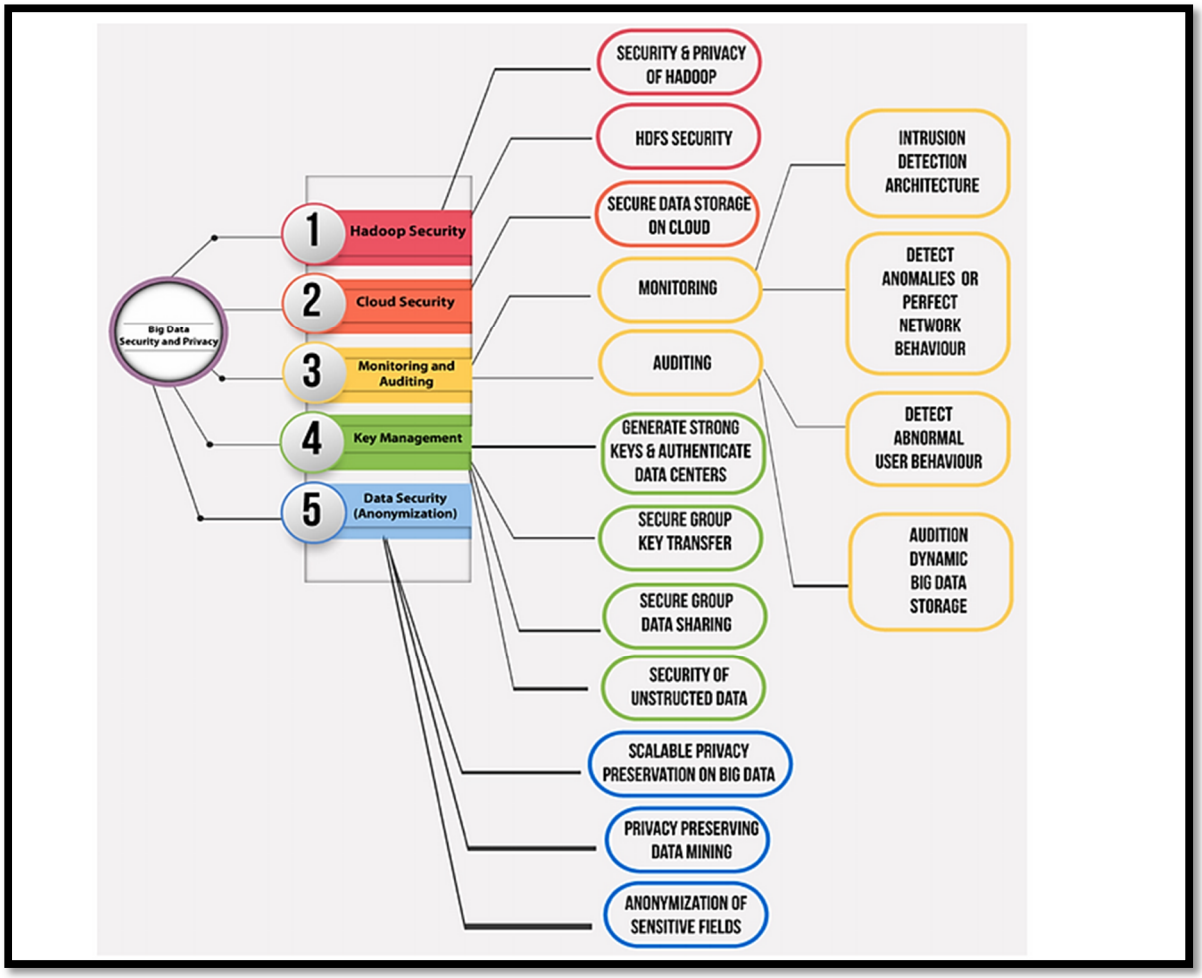

Figure 3 Big Data Security and Privacy Area (Bashari et al., 2016)

Not only is the Big Data phenomenon beset by security worries, but it is also beset by data privacy problems. These days, many organisations are dealing with privacy difficulties and liabilities; nevertheless, unlike security, privacy is perceived as a benefit, making it a selling point for both consumers and other stakeholders. Petabytes of data are now being saved and analysed as a result of the rising usage of Big Data technologies, making information categorization considerably more significant than it was previously (Aditham, S. and Ranganathan, N., 2017).

Author Fan et al. (2018) have introduced a two-step assault identification calculation, which is a secure correspondence convention created to screen the execution interaction of the

framework. The initial step includes the development control of the framework for each interaction. In the subsequent step, directions are coordinating with reproduction hubs. In secure information correspondence, information hubs create arbitrary keys, which spill protection of clients and their information (Fan et al., 2018).

Before diving into the various solution frameworks explained by different researchers, I will take a small detour to explore cloud computing, as most of the security solutions related to Big Data are either designed or implemented using cloud computing.

3.4.6 Cloud Computing and Security Issues

Because of its disseminated computational worldview, Big Data handling is better possible through distributed computing. Its design offers a brilliant answer for enormous scope information capacity and handling that meets two fundamental Big Data prerequisites. In view of their using time productively and cost benefits, distributed computing frameworks have become great places for information examination administrations. The expense proficiency, unwavering quality, and versatility of cloud-based administrations have been utilized for business, non-business, and scholastic administrations like capacity, foundation, stages, and programming. Distributed computing permits clients to send administrations in minutes with practically no specialist organization support (Ward, J.S. and Barker, A., 2013).

Virtualization is a cloud-based structure that permits the reflection and partition of lower-level usefulness and hidden equipment, as well as the convey ability of higher-level capacities and the sharing of actual assets. Distributed computing works on the principal of on-request network admittance to, in view of pay-per-view way. Well known cloud computing association models integrate the Platform as a Service (PaaS), Software as a Service (SaaS), and Infrastructure as a Service (IaaS). PaaS offers clients stages for making, running, and managing applications without having the fitting establishment. The SaaS gives associations cloud-based applications that work on virtual servers. Under IaaS, the client pays per-use for the utilization of offices to help capacity, PCs, servers, and systems administration foundation activities.

There are three kinds of arrangement or cloud modes i.e., Hybrid, Public, and Private. A private cloud is regularly worked by the separate organization, despite the fact that capacities are not unveiled straightforwardly to the client. Undertakings can utilize other cloud highlights on

account of a public cloud and proposition their administrations to clients outside the organization (Liu, Y., Esseghir, M. and Boulahia, L.M., 2016).

Many cloud sellers have given alluring stockpiling administrations that empower enormous scope distributed storage and information handling to be utilized by clients, including Amazon, Drop box, Google Drive, and Microsoft's One Drive. Enormous information device Hardtop structure in spite of shaping a significant job for conventional dispersed frameworks likewise shapes a necessary piece of a cloud climate. Specialist corporations like Amazon and Google are provisioning enormous information examination as administration, which help the clients to run their Hadoop occupations on overseas offices. Aside from the advantages, the shift from nearby to distant capacity and calculation additionally presents new difficulties. While the assortment and examination of large information in cloud can have a profoundly certain effect, it can additionally end up being very inconvenient without legitimate safety efforts, which will be talked about in the impending areas (Macías, M. and Guitart, J., 2016).

As of late, distributed computing has thrived because of its capacity to give shoppers with on-request, versatile, solid, and reasonable administrations. It is utilized across different enterprises to incorporate administrations like capacity, organizations, and applications, and has served as a response to various the Big Data prerequisites The utilization of huge scope cloud frameworks, spread over wide PC organizations, additionally expands the whole framework's target zone. Information is more or less defended less against assault once put away in the cloud, as access can be made on the web from anyplace. It is a basic issue in huge information and cloud applications, as the information proprietor no longer has actual information control. For the developing accessibility of cloud applications, information security is a fundamental concern and is presently of essential significance (Manuel, P., 2015). A distributed storage administration commonly provides various computational spaces, to give, for all intents and purposes limitless processing ability to end-clients. Accordingly, information from various sensible fields might dwell on a similar physical or virtual server, or the information might be portioned and put away across different spaces on numerous servers. The numerous clients utilizing a similar actual framework pool and the utilization of a similar organization involves the chance of making secret information accessible. On the off chance that client and aggressor utilize similar actual gadgets, the assailant will effectively access information on the off chance that satisfactory safety efforts are not upheld.

Most cloud-based frameworks contain important and touchy information; dangers against this information can jeopardize cloud frameworks. Confidential clouds are exceptionally secure on the grounds that they are introduced in firewall got region of the venture, in any case, in public clouds; information isn't protected and is powerless against a wide range of risks in light of the fact that the area is commonly known. Information put away in a circulated framework that is running in a public cloud is available to everybody. Dangers from shared public cloud are undeniably bound to emerge than dangers from the confidential server farm (Noor et al., 2013). The utilization of cloud computing makes overseeing enormous volumes of information more straightforward for organizations in any case, it might cause information security issues, information access issues, information for executives, and progressing business issues. Furthermore, conventional techniques for security are produced for frameworks that are not quite as adaptable as cloud conditions. This infers those procedures that have been worked over the long run cannot be applied straightforwardly to this disseminated distributed computing.

Distributed computing provides enormous private or public organization access to interconnection to give on-request, adaptable foundation for an application, document and digital storage capacity. Distributed computing is a higher-level application that not just permits us to encounter direct expense adequacy, yet additionally may make changes to the server farm of the capital-escalated to set climate factors.

In the context of cloud computing, confidentiality alludes to the fact that client information ought to be kept classified from cloud specialists or other unapproved clients. Protection is as yet the major issue, chiefly because of the deficiency of real control cloud. Like privacy, uprightness, harmony idea is applied to the respectability of the information and the uprightness of the cycle. Information honesty implies that information should be put away in the cloud servers and any abnormalities should be genuinely recognized (for instance, information misfortune, obliteration or harm) (Qu, L., Wang, Y. and Orgun, M.A., 2013).

Privacy is another main point of contention in cloud computing, on the grounds that the client information distributed computing climate is situated on a remote disseminated administration server which may not be supplied by a dependable cloud administrations supplier. There is subsequently a possible gamble of secret or individual data might be unveiled to unapproved elements. Obviously, if secrecy, privacy and trustworthiness could be guaranteed, cloud computing would become more significant and popular.

Going against the norm, on the grounds that a lawful obligation for two-factor authentication can provide a solution to this privacy struggle. This implies that the responsibility framework has the capacity to recognize the gathering of obvious proof. You can recognize an extraordinary unit or even vindictive, truth be told programs utilize the ID is limitless. “Distributed computing" alludes to an Internet or Intranet. Cloud administrations can provide security to the general population and confidential organizations. Distributed computing method for controlling, arranging, and admitting to the remote equipment and programming assets. It additionally gives online information stockpiling, foundation and applications. Cloud Delivery stage autonomy since programming doesn't need to be introduced locally on your PC. It is consequently our distributed computing business applications can help your portable and joint effort (Viji Rajendran, V. and Swaminathan, S., 2016).

The principal thought is that the distributed computing is subject to a vital guideline of PC abilities for re-use. "Distributed computing", contrasted with a customary dispersed figuring matrix processing" "utilitarian registering" or "independent registering ideas", it is an expansion of the level across hierarchical limits. Cloud technology, exceptional adaptation, and a network of extraordinarily complex registering establishment that really can keep up with end technology and stock use are qualities given by Gartner.

3.4.6.1 Need of Cloud Computing

Associations need to change their thoughts and spending plan and time utilization of the current foundation (e.g., programming, equipment and administrations). Yet, associations neglect to assembling or offering types of assistance in the face of increasing need for cloud computing. Distributed computing gives exceptionally advanced climate effectively as stage, data administrations, programming and equipment prerequisites. Associations can utilize these settings to the idea of administration charge. The organization additionally can effectively develop or contract as per necessities. These attributes are introduced as follows: (Chandramouli, R., 2011).

- On-Demand Self-Service-Cloud clients can get additional resources, for instance, the use of limit breaking point and execution, no manual intercession.
- Expansive Network Access-A broad assortment of admittance to the framework numerous heterogeneous framework contraptions, (for instance, cells, computers, tablet PC, advantageous furthermore, each and every static device that should have the ability

to get to cloud benefits through standard parts. This part in the cloud can get too generally through a standard Internet show support. Regardless, is believed to be unfrosted furthermore, in this way a couple of fundamental framework attacks can be basic in the cloud.

- Shared Pool of Resources considering multi-occupant, cloud resources are shared by different clients. It should be seen that no resources for a specific client. These fundamental workplaces are distribution and re-task to the requesting specialists (overall). Shared resources are basically reinforced by more than one expert association of virtualization development sharing in light of various working system (OS) on a similar actual PC. In any case, certain security necessities for combination described that isolated data and methods. Moreover, staggered condition, the deficiency of actual control also, the shortfall of strong cloud expert centres to the relations between the client system and the necessary security that standard new structure.
- Versatility - despite the self-managerial resources, cloud is depicted by its amazing ability to find and useful appropriation of resources. This property exhibits a great deal of resources. One of the best weaknesses to consider is the information move limit of the security of noxious clients can take advantage of organizations or applications settings to attack. As needs be, arrange flexibility and it is the most crucial to unflinching quality part stays to ensure that the versatility natural in the distributed computing model.
- Metered Services - This property insinuates a business mode use cloud organization for clients to pay for usage and thusly basic expense venture reserves. Thusly approval and obligation necessities should be considered as essential. In addition, assessment server offers assistance for a combination of checking contraptions to ensure business rationality and data mining (Singh, S., Jeong, Y.S. and Park, J.H., 2016).

These properties portray the qualities and traditional IT support illustrates more useful and more versatile cloud organizations. Cloud Based on its area can be designated a private, open or closed structure. As a mix of gear and programming resources for distributed computing.

3.4.6.2 Example of Cloud Computing Application

Dispersed registering is a useful on-request coordinate get to mode to an arrangement of configurable figuring assets (e.g., structures, servers, collecting, applications, and associations) can rapidly make and spread an irrelevant blueprint of clever relationship association or master

focus. Cloud-based data and applications can be used to get to the Internet and empowers clients to get to the Internet. Yahoos, Gmail, Hotmail email are all around common examples of distributed computing. The Email Management programming and server absolutely by Google, Yahoo, etc. for organization and control of everyone in the Cloud (on the Internet).

The latest examples in the IT part have uncommonly unravel limit and sharing data remote. New applications, for instance, online casual associations and on the web, documentation is an incredibly invaluable way to deal with far off internet-based data server can store and offer a grouping of data, including fundamental individual data, electronic reports. Distributed computing, since this is the future IT designing, even ensured to unlimited and versatile limit resources (and other registering resources) is really useful distributed computing organization clients (Somu, N., Kirthivasan, K. and Shankar Sriram, V.S., 2017).

Regardless of the way that distributed computing is at this point a nice time give cautious thought, as well as has the advantage of attracting a growing number of clients in the close by server ranch re-evaluating organization allows a remote cloud server. Moderate gear, organize affiliation, middleware and virtual machine development has incited another, universally conveyed processing stage that distributed computing power and limit organizations appear wherever on the Internet, if there is no colossal interest in new establishment, getting ready or programming grant.

The above part supports three cloud-based organization modes. Dependent upon the gear limits, these organizations show worked in different courses.

I. Infrastructure as a Service (IaaS) - In this organization mode, resource organization, collection, and gives clients the limit furthest reaches of your, (for instance, Amazon), sort out centre or registering limit). After that client would be able act in each functioning structure and programming that best location their issues. They couldn't manage or control the principal cloud structure. The aggregate application open to client (if crucial). The idea is to continue to run on the client contraption event, return helps that are running on the server. No prior interest in servers or programming approving. (e.g., Amazon). An all-out application is proposed to a client as and when required. Believed is to execute an event on client's machine, backend organizations to be executed on servers. No need of direct interests in servers or for approving of programming’s.

II. Platform as a Service (PaaS) - This help model gives a turn of events climate or stage clients run their applications. This implies that the client would be able redo the application to target explicit stage accessible apparatuses. They likewise have full command over the conveyed applications and related design. The advantages of the pawn are more significant. These private companies can fabricate, truth be told what's more, convey your own endeavour applications and programming without purchasing a server and the functioning gathering to oversee them. Google App Engine and Microsoft sky blue stage is an extremely commonplace PaaS administration mode.

III. Software as a Service (SaaS) - - In this mode, cloud expert centre is an item benefit. For example, without the need to purchase and present the item on a substitute system that clients can use for a charge. Admittance to these applications can be used by different contraptions use client connection point or application interface. In any case giving security organizations, SaaS is the obligation to keep up the latest programming refreshes and significant scale cloud organization. The SaaS by and enormous mirrors the growing number of offers, for instance, Google docs and Drop Box sales force (Phaneendra, S.V. and Reddy, E.M., 2013).

3.4.6.3 Cloud Sharing and Infrastructure

The independent computing cloud this principal feature applies to publish time self-management features of computing resources to adapt to unpredictable changes, while hiding the complexity of the to overcome system operators and users manage increasingly complex computing and reduce the complexity of an obstacle. In addition to these unique service mode, cloud services portfolio continually enriching (Chandramouli, R., 2011). This distribution of cloud services and components and architecture of the complexity of new security and privacy challenges (e.g., virtualization). Threat, vulnerability and related risk, technology for a specific environment. This vulnerability of cloud service models and technology and critical environment. Characteristics. There are three main issues that are closely related to properties cloud.

- Outsourcing - Re-appropriating shared with the save information or business capabilities with outsiders. By rethinking, clients to eliminate the necessities for the foundation and upkeep of neighbourhood stockpiling framework. Nonetheless, the revaluating additionally implies that clients will lose a few commands over their information and errand. Many cloud suppliers don't meet the expected level to give

dependable security engineering. Truth be told, information can peruse, alter, or erase when it is revaluated. Then, at that point, the proprietor should comprehend the security, uprightness and secrecy issues. They additionally need to stress over help accessibility, information recuperation and business coherence. Therefore, this issue has become one of the underlying causes of loss of control of the Cloud Security Issues. Consequently, information and interaction security stay a significant obstruction to the turn of events and broad utilization of distributed storage space. Moving to address security issues, cloud specialist co-ops should guarantee the dependability and security of put away information. Furthermore, the rethinking of information should be safeguarded, observing and testing to guarantee that the privacy, respectability, and other security administrations. In the case of external events, outsourcing service provider may be subject to breaches of confidentiality, the service provider must provide (Bocek et al., 2008).

- Multi-Tenancy - Cloud infrastructure rental multi-language is shared by multiple users and use. It is therefore in a virtual environment information belongs to a different user can be on the same physical machine. According to the specific resource allocation policies. Although many are an important select cloud-based provider of rental, this is a cost-effective approach provides a new platform for cloud vulnerability. This means that a malicious user can use this coexistence issues to attack.
- Big Data - A lot of information and uses of information development and introduced new difficulties in the information checking and security highlights, for example, picture handling, information mining in cloud conditions. This means that traditional security mechanisms are not enough, because there is no efficiency computing and communications. For example, cloud storage security settings specify the external data integrity. Therefore, current privacy technology mainly based on static data sets, but the data will still be dynamic change, including the data model, properties, change, and access rights. Thus, new strategies and plans (Gain, U. and Hotti, V., 2013).

A variety of security issues is widespread sharing of technologies and new concepts for the new cloud-based objectives such as virtualization. Many people do not know what cloud computing is, particularly because of the excessive use of the term. In this context, it describes the highly customizable, resources for external services via the Internet, charges for each type of use. A specialized distributed computing architecture that meets the needs for dynamic

configuration is cloud computing. This new human-adaptive mode is different from the traditional network. This is an abstract offers three levels of service. The Economic and attract cloud computing is the user can use only what they need and pay only for what they actually use. Resources can at anytime and anywhere is available via the cloud. No need to worry about continuing to do so.

3.4.6.4 Data Security and Privacy Protection in Cloud

This area covers the most critical perspective is the manner by which to set up and deal with your IT framework must meet the security necessities of the key substantial information applications. One of them is the privilege to protection. This is critical on the grounds that in their clients, gadgets and PCS that is progressively sharing individual data and substance on informal organizations and open cloud. In this manner, the government managed savings organize framework is an exceptionally intriguing issue seek. In this part, one of two sections manage contextual analyses. What's more, customary security instruments, for example, firewalls and Support for substantial information, militarization of the general population can't be utilized for PC framework. The SDN is another administration arrangement would turn into a handy component to ensure expansive data framework, as appeared in the figure toward the finish of the section in the second case. It likewise identifies with the present work and uncovered the open inquiries (Vidyasagar, S.D., 2013).

Anonymous authentication within Cloud services provides a new security solution. The following focusses on effective verification process and the confidentiality of the user. Overview of the system model, security requirements, encryption and encryption protocols. Environment The system model consists of three basic parts:

- Cloud Service Provider (CSP): The Cloud Service Provider is normally an organization as a trusted party. The CSP provides cloud services to authenticated users when they entered into the cloud services. The CSP also give users access to the attribute. However, when the CSP detects malicious user authentication, they are able to log-out relevant users (including managers), to enforce stronger security against unauthorised access.
- Revocation Manager (RM): RM is a partial trust assembly, such as the public authorities to determine the identity of the user on the legality of the revoked or not.

Only the chief superintendent and the RM collaboration can detect user identity. The RM and the Civil Service Bureau to register a user in the user properties will issue.

- User (U): User client users to access shared storage cloud and cloud services. Anonymous users if they comply with the rules of the CSP are correct. To increase security, users can use antivirus software or the security of local storage. The non-linear grouping features' provider provides cloud - based facilities internet - based authentication. When you generate and store user key, user's fail-safe device prevents an attacker incompatible. The authentication phase of the client-side and server-side, this is more efficient than the lack of expensive pairing and less exposure. They therefore use our solutions can also verify that the cloud service provider more clients. We really would like to alter the recollection method in the future. I wish to use another methodology to minimize the effects of the long-standing black list. I would also like to implement adjustments that can really resulting in cause error identities. (Wang et al., 2013).

The above proposed architecture drew attention to the important security questions: to what extent should be outsourced to the customer's organization, system administrator and service provider? Remote access to the company network is already an essential part of the business environment. Staff of customer systems uses a company intranet and the distance between the inside. In the last two cases (RM and U), the application requires a valid query and use of cloud-based architecture in flight. Privacy issues when the confidential data to the cloud. Use the encrypted cloud server that is, it does not allow the system administrator to access the content of the database.

3.5 Big Data Security and Privacy Solution in Literature

After briefly looking at cloud computing processes, I return to the discussion of solutions to Big Data privacy issues in literature:

I. Encryption

The non-secure data has been considered one of the most significant challenges in massive data management and access control on the cloud. The combination of access control and anomaly can be utilized to protect data from malicious client and to maintain the screen and data management, Spike can be used as a solution (Reddy., 2018). Interestingly, the variety of

information sources, such as text, photos, voice, and recordings, has been overlooked. However, another promising approach is PPMUS (Privacy-Preserving Mobile User Authentication framework) which uses a few Big Data features, such as limit cut off, usability, strength regions, and data the board (Vorugunti., 2016). Also, two computations fuzzy hashing and Fully Holomorphic Encryption (FHE) estimation has been used to reduce client insurance in this approach. However, the main drawback of this structure is that the client's secret key creation situation is taken into account when determining if the client is trustworthy or a faker.

Elliptic Curve Cryptography (ECC) has been considered as an option to guard against two security threats, such as Disconnected Password Guessing Attack and Impersonation Attack (Zhao and Jing., 2018). Similarly, the issue with client protection while sending inquiries over scrambled multi-faceted big metering information can be solved using Locality Touchy Hashing (LSH) calculation, Policy Attribute Based Encryption (CP-ABE) and scrambled information sent to heterogeneous big information capacity frameworks (Jiang Lu et al., 2018). The proposed strategy has got great execution concerning information secrecy and protection in a semi-confided in cloud climate. Looking through time is high for multi-layered big metering information over semi-confided in cloud conditions.

The client protection and information secrecy in the big information organizing climate can be boasted by introduction of layers for client's key age where the layer 1, highest keys are utilized to encode lower keys for ensuring clients security (Fan et al., 2018). Long encryption and decoding time for bigger measure information acts as a disadvantage in this approach. However, a different perspective has been developed to resolve scenarios including circulating processing by dividing the input data into several fixed sized block, and then a Generic Algorithm (GA) is used to encrypt the block of parts (Mall and Saroj., 2018). Though Generic Algorithm and hereditary computations increases security in small blocks but the large blocks remain vulnerable.

Half and half cryptology and symmetric and lopsided calculations (AES, ECC, and SHA-1) have proposed to calculate the encryption of data and to achieve data mystery and dependability (Goyal and Kant., 2018). Although 3D-AES model, which a Block-Cipher has been used to estimate with a few restrictions, such as the turn-key limit, replace key brick figure and mixing key limit, as well as three round work patterns (Minimum) (Adnan and Ariffin., 2019). In comparison with AES, 3D-AES has provided better results in terms of complexity, security,

and execution with the testing in discretion; nonetheless, more haphazardness calculations need be examined to show the framework has produced better results.

A security saving disseminated K-Means bunching of evenly apportioned information along with the utilization of the mystery moving system battered to code based zero information has been to reduce the security worries in the vindictive not well-arranged model (Patel et al., 2013). Another proposed method was to use bisecting k-implies bunching to protect saving cooperative sifting plans (Bilge and Polar., 2013). An improved version was proposed where an imaginative tree-based irritation approach was utilized for handling the bothering information mirroring the secret movements (Roy, 2018). In their approach, they utilized a k-d tree trick to recursively partition a dataset into various little subsets so that the information records inside every subset turned out to be additionally orchestrated with each parcel.

The Privacy Preserving Distributed DBSCAN Clustering has been proved effective to handle the issues of the two-party protection (Liu et al., 2012). At the start, two conventions were used for protection safeguarding DBSCAN bunching over evenly and in an upward direction isolated information correspondingly and later extended them to the arbitrarily isolated information.

II. Big Data Privacy and Security Issues in Cloud

Privacy-Preserving Data Obfuscation Scheme is an important tool for data statistic and information mining where unique keys were distributed to clients with dissimilar consents to get to the information. Another different approach has been made based on cloud computing protection where protection arrangements are expected to resolve issues, for example, information revaluating, off-shoring, virtualization, and autonomic innovations (Pearson and Charlesworth, 2009). Lawful planning and administrative approach have been devised for resolving issues in information sharing, its area access, unseemly treatment of information, legitimate issues, programmed direction, and so on. Security chances are decreased by proposing a step in the right direction for security inside distributed computing. Key components, for example, straightforwardness, affirmation, client trust, obligation, and strategy consistence are considered as a forward move toward responsibility.

A centre harmony between security chance and business mission has been proposed to deter information security hazard such as network interchanges (registering climate), controlled

admittance (third party control), information reviewing (the investigation of client's way of behaving to the activity of the information), faculty (staff, organization, and so on) (Weiquan and Houkui, 2013). Information avoidance measures, for example, specialized control, administrative administration, got confirmation, framework security group plan, wellbeing concerns, security reviews, interruption discovery, network security controls, and staff preparing; to diminish the gamble level were planned. A new dependable cloud administration suggestion framework has been proposed to survey the veracity of client criticism and puts together trust calculations with respect to the emotional perspective is the proposed structure (Noor et al., 2013).

Cloud rationale security connected with the framework, information, and organization security runs on the web and stances numerous dangers (Liang, 2014). The creator concentrated on the arrangements on distributed computing security. Infiltration tests for security strength investigation are led and different weaknesses and assaults from the enemy were tried and re-enacted. It has been recommended that the legitimate and standard administration framework and security activity ought to be figured out and executed. SLA-based trust structure for the cloud commercial center has been proposed to evaluate the supplier's standing and reliability (Macias and Guitart, 2016). In spite of the fact that it sticks to the trust proliferation component, the model overlooks the impacts of outlandish client surveys and proposals. By computing start to finish QoS values across the help levels, a help suggestion framework is made. To give a redid determination, this forecast model processes client similitudes utilizing past QoS data.

III. Big Data Handling in Cloud

Various issues and difficulties in the cloud climate have been tried to resolve by cloud administration engineering and the examination of the issues connect with each layer of the cloud administration (Bhatia and Wankhede, 2015). The number of web clients, standard security strategies, regulations, and the executive has been drawn by the cloud administration suppliers. This will help in advancing the administrations by the cloud administrations suppliers and diminish the gamble.

With the quickly developing of information, the need for information handling, and to diminish client cost; distributed computing administrations came into reality. While utilizing these administrations client's information and data travel from his/her framework to a stockpile’s gadget present in a dispersed climate. Hence, the re-appropriated information cross firewall

limits and become powerless. Unwavering quality and accessibility of information become a major worry in such a situation. Information privacy, respectability, and validation become an issue as the client has zero command over such security boundaries in a circulated climate. Various layers of cloud engineering have been considered and the criticality connected with information security is talked about. Specialized issues during security at each layer are examined (Zhe et al., 2017).

A cloud administration determination model has been proposed to coordinate both emotional and objective trust values through fluffy basic added substance weighing framework and destructive information is separated by the deviation from the goal trust, which is assessed by outsiders (Qu et al., 2013). In any case, the outsider's unwavering quality is in uncertainty, and the unique help conveyance climate makes client input sifting ineffectual. Afterward, the creators report their extended exploration on reliable cloud benefits that are setting mindful by doing both goal and emotional trust appraisals. The benchmark level of goal trust assessment is progressively changed relying upon the processed comparability between different assistance conveyance conditions. However, the creators involved objective evaluations as benchmarks to sift through one-sided emotional assessments, the strategy actually comes up short on versatility in doling out a load for trust credits.

A trust model has been proposed in view of the specialist organization's verifiable certifications and current capacities where the strategy consolidates SLA boundaries and specialist organization limit for the assessment of trust boundaries (Manuel, 2015). Another review-based examination connected with Information security and protection challenges in versatile remote gadgets has been put forward to explore issues in remote correspondence (Mollah et al., 2017). Another worldview of cloud figuring security has arisen as of late by coordinating cloud and portable registering with one another. It represents a greater test from such an incorporated framework. The protection and security necessity for versatile distributed computing and furthermore examined the accessible plans to deal with different protection and security needs for portable cloud figuring. Significant security and protection challenges in these frameworks are information security, cell phone security, versatile cloud application security, virtualization security, character security, dividing and offloading security, and area protection.

IV. Hadoop Architecture in Big Data

An effective way of managing Peta bytes of Data units in environment welcoming and more affordable way is by using Hadoop which is an efficient, strong open-source Apache license. Hadoop is partaking in an essential work in Big Data. Vidyanagar S.D construed "Hadoop is expected to run on low esteemed item hardware, it consistently handles records replication and centre disillusionment, it achieves the outrageous work - you can focal point of thought on taking care of data, Cost Saving and environment pleasant and solid real factors dealing with that".

A stylish articulation of Big Data applications and about the enormous scope of administered limits that can be applied to huge knowledge sets the preferred programming tools for Big Data applications are Google's Map Reduce design and Apache Hadoop, its open-source execution (Zhao et al., 2014). A decree including these limits is that they make a massive measure of focus data, likewise, this plentiful data is highly sought after the dealing with finish. Pushed by using this insight, a data careful store structure for Big Data applications, which is known as Dacha.

To eliminate the issues related to security, arrangements and restrictions to the Big Data, a sophisticated system has been proposed which incorporate sensitive information in Hadoop by examining the encryption of information (Parmer et al., 2017). Information handling needs scrambled information to be decoded first and afterward the calculations are performed which make it hazardous with the speed of handling, memory limitation, and information sizes influence the presentation of this system. A layer of secure client verification has been added with encryption of information moving and calculation named ChaCha20 is planned and carried out which is a stream code and takes the vital size of 32 bytes also, later it is extended to 64 bytes to scramble 64-byte blocks. It is seen that the most noteworthy speed is come about while utilizing the support size of 500 kb or 1 MB.

Bits of knowledge regarding the privacy and security along with the estimation depicting the survey about the environment, innate science and investigation, Life sciences, etc has been suggested to tackle issues with certificate, methodologies, tests, and troubles of Big Data (Sinanc and Sagiroglu, 2013). Any undertaking in any endeavour having colossal information can take the expansion from its cautious appraisal for the trouble fixing reason. Using the Knowledge Revelation from the gigantic information supportive to get the data of the

multifaceted informational collections. Intricacy, assortment, every now and again changing position and fast improvements of colossal information structures increase extremely extraordinary hardships in gigantic records benchmarking (Gao et al., 2013). The tremendous Data Bench now not only covers gigantic utility circumstances, but rather additionally contains unique and master real factors sets. Differentiated and different benchmarking suites, Big Data Bench has an extremely shallow useful profundity, and the number of records entered doesn't essentially influence the limited scale designing characteristics.

The manufactures try to boost management of the characteristics of Big Data but also possibilities of assessments assessment by sending off the duration of time data management. began processing data, which involves researching for true characteristics that develop from data sources which helps in exploring Big Data and put down plans and resources to ingest, preserve, and recognize authentic components that may be numerous, distinctive, and then maybe evolving rapidly (Gain and Hoti, 2013). The shipments off time span information treatment undermines the accuracy of the characteristic Big Data as well as the possibilities of looking at actual variables.

RDMBS contraption has been considered as an important tool to manage Big Data. Hadoop structure involving recognize centre point, information centre point, side centre point, HDFS to adjust to gigantic bits of knowledge systems which has been classified into 5 perspectives like volume, speed, combination, cost and complexity (Phaneedea and Reddy., 2013). Hadoop structure oversee huge real factors sets, adaptable computation logs association utility of epic estimations can be arranged out in money related, retail industry, clinical consideration, flexibility, assurance.

Security investigation over Cloud of virtualized foundation, which is put away in Hadoop Distributed File System (HDFS) has been studied (Win et al., 2017). A two-step AI calculation was recommended that incorporated strategic relapse and conviction proliferation to process the conviction engendering is utilized to register the faith in presence of an assault. The use of calculated relapse enables the fast count of attack's contingent probabilities. Even more essentially, determined backslide similarly enables the retraining of the person strategic relapse classifiers using the new attack incorporates as they are acquired from attack recognizable proof.

Big Data strategies combined in the cloud scenario, including Big Data organization and security (Hababeh et al., 2018). The best degree of data is gathered and divided into two categories: governmental and non - governmental. Private data is tricky data, but public data is common data (for instance understanding prosperity history). Anyhow, sorted data remains a secret to others. For grouped data that utilizes the Hadoop Distributed File System (HDFS), security has been added. Estimating the risk impact level is used to group the provided data into secret or open categories. Since engineers have designed security calculations for encoding secret information, which aren't suitable in all instances, it isn't taken on for large quantities of information with "n" number of varieties.

3.6 Discussion of Big Data Privacy Issues and Challenges in Literature

It is clear that the security and privacy challenges in Big Data are many and different. Therefore, there are many different solutions in the literature of Big Data. It is seen that a popular method of resolving many security issues is based on Hadoop infrastructure. Further, I have briefly looked at some literature on cloud computing because all Big Data would be expected to be stored and processed in the cloud. Therefore, the security of cloud computing is also related to the security of Big Data and also the solutions (which should be executed using cloud technology). The following is a summary of the overall outcomes or an analysis of the papers I have reviewed in this section.

Managing information security while controlling vast and quick data streams is one difficulty in the Big Data setting. Therefore, security technologies should be adaptable and simple to scale in order to make it easier to integrate future technological advancements and adapt to changing application requirements. Finding a compromise between various security demands, privacy duties, system efficiency, and quick dynamic analysis on various massive data sets is necessary (data in motion or static, private and public, local or shared, etc.).

- **Traditional Solutions Are Insufficient:** Traditional security measures, such as some forms of data encryption, impede performance and take a lot of time in the context of Big Data. Additionally, they are ineffective. In fact, for security reasons, only small data partitions are handled. Therefore, security attacks are typically discovered after the propagation of the harm. Big Data platforms require the management of several

concurrent calculations and applications. As a result, in these situations, performance is crucial for data sharing and real-time analysis.

- **New Security Instruments can have unintended risks:** Combining different technologies may introduce unintended risks that are frequently overlooked or understated. New security tools are also immature. Therefore, Big Data platforms may include unrecognised security risks and vulnerabilities. The value of the data is being concentrated in several clusters and data centres concurrently. These vast data reserves are immensely alluring to business, governments, and industry. They are the object of numerous intrusions and assaults. Furthermore, end-users, partners, and employees pose the majority of security risks. Therefore, it is crucial to implement cutting-edge security measures to safeguard Big Data clusters. Regarding this matter, it is the duty of data owners to establish strict security guidelines that contractors must abide by.
- **Anonymization of Data should not compromise system efficiency:** Data anonymization should be accomplished without compromising system performance (such as real-time analysis) or data quality in order to guarantee data privacy and security. Traditional anonymization methods, however, rely on lengthy computations and multiple iterations. The performance of the system may be slowed down by multiple iterations, especially when handling sizable heterogeneous data sets. Additionally, processing and analysing Big Data that has been anonymized is challenging.
- **Data encryption is popular to ensure privacy in Big Data:** Numerous studies have been done to enhance the effectiveness and dependability of conventional procedures or to develop novel Big Data encryption methods. Homomorphic cryptography, in contrast to some conventional encryption methods, allows computation on encrypted data. As a result, this method guarantees information privacy while enabling the extraction of helpful insight through potential analysis and calculation on the encrypted data. Regarding this issue, suggests a platform that is suitable for handling MapReduce operations in the context of homomorphic cryptography. There is a new key exchange method called CBHKE that ensures the performance of cryptographic solutions in dispersed situations (Cloud Background Hierarchical Key Exchange). It is a secure solution that is quicker than its forerunner methods (IKE and CCBKE). It is based on a two-phased iterative authentication key exchange (AKE) approach (layer by layer). To

better secure huge data volumes in distributed systems, new strategies with improved performance are still required, and operations are essential.

- **Big Data storage issues:** The rapid expansion of data has increased the standards for management and storage. We concentrate on the storage of huge data in this part. Big Data storage is the administration and storage of enormous datasets with the goal of ensuring the availability and dependability of data access. We'll go over significant topics like Big Data storage techniques, distributed storage systems, and enormous storage systems. The storage infrastructure must, on the one hand, deliver information storage services with dependable store space and, on the other, deliver a strong access interface for data analysis and inquiry.

In the earlier, structured RDBMSs have been used to store, manage, search up, and analyse data using data storage devices as auxiliary server equipment. Data storage devices are becoming more and more crucial due to the rapid growth in data, and in order to compete, many Internet organisations seek for large storage capacities. Research into data storage is therefore urgently needed.

The demands of huge data are met by the development of various storage solutions. Direct Attached Storage (DAS) and network storage are two subcategories of the existing enormous storage technologies, whereas Network Attached Storage (NAS) and Storage Area Network are subcategories of network storage (SAN).

In a distributed array system (DAS), different hard drives are directly connected to servers, and data management is server-centric, making storage devices peripheral equipment that each requires a specific number of I/O resources and is controlled by a different application programme. DAS is thus only appropriate for connecting servers of a small size.

When the storage capacity is raised, DAS will, however, demonstrate unfavourable efficiency due to its low scalability, meaning that the upgradeability and expandability are severely constrained. DAS is thus primarily utilised in personal computers and small servers.

Network storage will make advantage of the internet to give users a unified interface for sharing and accessing data. Special storage software, disc arrays, tap libraries, and other types of storage media are all included in network storage devices.

Transmission and sharing challenges: The problems with scalability are closely tied to the problems discussed above. The choice of where to store the huge data is a significant problem as well. Elastic web services that provide petabyte scale data warehouses have been the key new storage models to emerge. Additionally, improved RDBMS with Hadoop Map/Reduce integration schemes and NoSQL DBMS must maintain low input/output latency from huge data repositories in addition to the storage issue.

Because its properties aren't bound to a particular data model but can instead be stored in any suitable structure or format, alternative DBMS schemas (like NoSQL) provide the advantage of enabling updates without expensive reorganisation at the storage layer. As a result, the difficulty in Big Data storage is to manage high-volume data acquisitions and support a range of heterogeneous data structures without experiencing increasing delay. The so-called "Internet plumping problem" is a result of both sharing and the movement of massive data. This is due to the fact that the expansion of wired, optical, and wireless infrastructures has lagged behind the rise of data. For instance, it is predicted that the connection speed in mobile networks would improve seven-fold between 2012 and 2017, but the amount of mobile data traffic worldwide is predicted to increase thirteen-fold during the same time period. Mobile devices (such as smartphones and tablets) will act as "sensors" for data collection and capture during this time period in addition to supplying content to consumers. Furthermore, we anticipate that a sizeable amount of machine-to-machine traffic data would be created, adding to the stress on networking infrastructures. The final challenge of how so much data will "flow" via the network, be shared, and be kept is made up of all the aforementioned problems.

Some of the most commonly used tools in response to Big Data security challenges are:

- **Hadoop:** Big Data processing in a distributed computer environment is supported by Hadoop, a free Java-based programming platform. It is a component of the Apache project, which the Apache Software Foundation sponsors. A Master/Slave structure is used by the Hadoop cluster. Large data sets can be processed using Hadoop over a cluster of servers, and applications can be executed on those servers.

Systems involving thousands of nodes and hundreds of petabytes. Hadoop's distributed file system facilitates quick data transfer rates and enables the system to keep running normally even in the event of some node failures. Even in the event of a substantial number of node failures, this method reduces the likelihood of a system failure as a whole. A computational solution is made possible by Hadoop.

It is fault resilient, versatile, cost-effective, and scalable. Popular corporations like Google, Yahoo, Amazon, and IBM, among others, use the Hadoop Framework to support their applications handling massive volumes of data. Map Reduce and Hadoop Distributed File System are Hadoop's two primary side projects (HDFS).

- **Map Reduce:** The Hadoop Map Reduce framework is used to create applications that efficiently and fault-tolerantly handle massive volumes of data in parallel on clusters of commodity hardware resources. A Map Reduce job first separates the data into discrete chunks, which are then processed concurrently by Map jobs. The reduce jobs then receive the results of the maps sorted by the framework as input. Typically, a file system stores both the job's input and output. The framework handles scheduling, monitoring, and reinitiating unsuccessful tasks.
- **Hadoop Distributed File System (HDFS):** In a Hadoop cluster, a file system called HDFS is used to store data across all nodes. It joins multiple local node file systems to create a single, sizable file system. By replicating data across various sources to overcome node failures, HDFS increases reliability.

4. PROPOSED TECHNIQUES FOR BIG DATA SECURITY

The proposed work consisted of four entities, including Trusted Center (TC), Data Owner (DO), Data User (DU), and Cloud Server (CS). It based on a similar architecture used by Narayan et al. (2020). Their paper provides a robust design for testing security systems in Big Data environments as is the objective here. This design has also been used recently in published papers by Obayya et al. (2022) and Singh and Jha (2022) The usage of the individual components is explained below.

- TC: This is a reliable, solid, generous capacity cap and powerful substance designed for security purposes. It is used to create secure communications between DO and CS, or DU and CS, and to attract information customers and information owners. We are approaching the boot loader of the complete CS security framework. For all information owners and customers, it holds a mysterious key and denies access depending on their past and present activities.
- DO: The DO is registered with the TC and then the approved DO can send that information for storage in the cloud. Basically, records are transferred to CS in three phases: evaluation at the sensitive information level, printing, and encryption.
- DU: DU logs in to TC and gives these important customers access to the records stored in CS. To decrypt the information, the DU asks the TC for a key and decrypts it. Therefore, a customer with a mysterious key reserves the privilege of accessing the information.
- CS: Consists of records sent from DO. It handles several tasks such as information gathering, recovery, execution, and access control. DO has some support. B. Update the text in the diagram, change the text in the code, and remove the record from CS (Win, T.Y., Tianfield, H. and Mair, Q., 2017).
- DO requires the TC for a private key so that the data can be stored in the cloud. The TC registers the DO and issues a private key based on the user's credentials. The data is then encrypted with the private key and compressed before being uploaded to the cloud. The DU can request CS to access the data. They used TC to verify the DU's ID before providing the key for decryption. I have installed HDFS in a cloud environment to process 5V Big Data. This provides ample storage space and access control for both data owners and consumers. The process of the proposed system is shown in Figure 4.

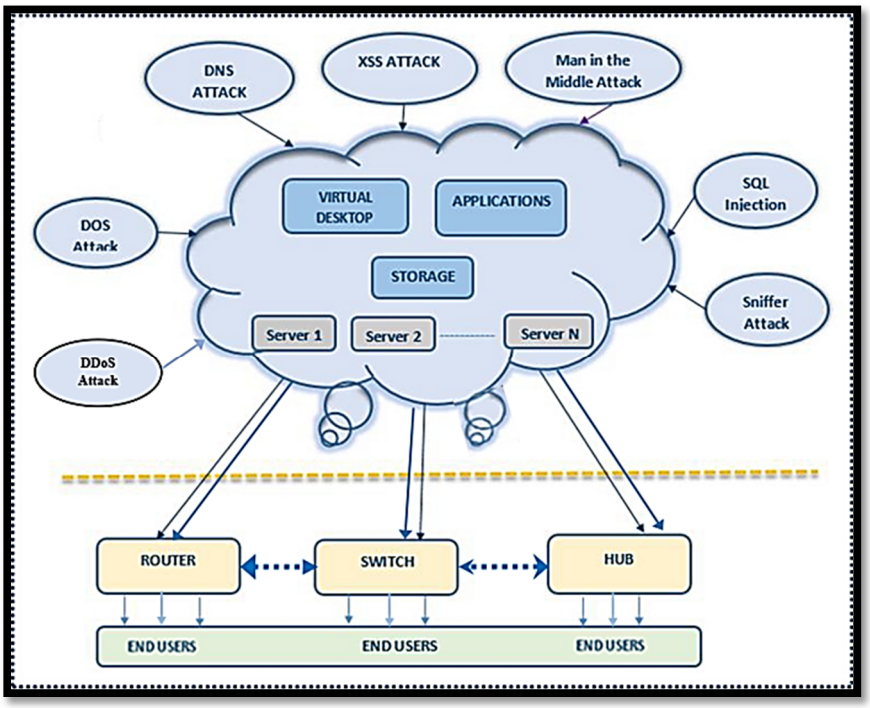

Figure 4 Work Flow of Big Data Enabled Cloud Environment security systems (Mishra et al., 2022)

Indicates the proposed framework design where I represent the interaction obviously. There are three major information methodologies I have introduced in this engineering:

- Big Data Outsourcing
    - Authentication, Compression, Encryption, and Store
- Big Data Sharing
    - Authentication, Decryption, Decompression
- Big Data Management
    - Compression, Clustering, and Indexing

Every system, for example, big information revaluating, big information sharing, and big information the executives depicted exhaustively in the accompanying sections.

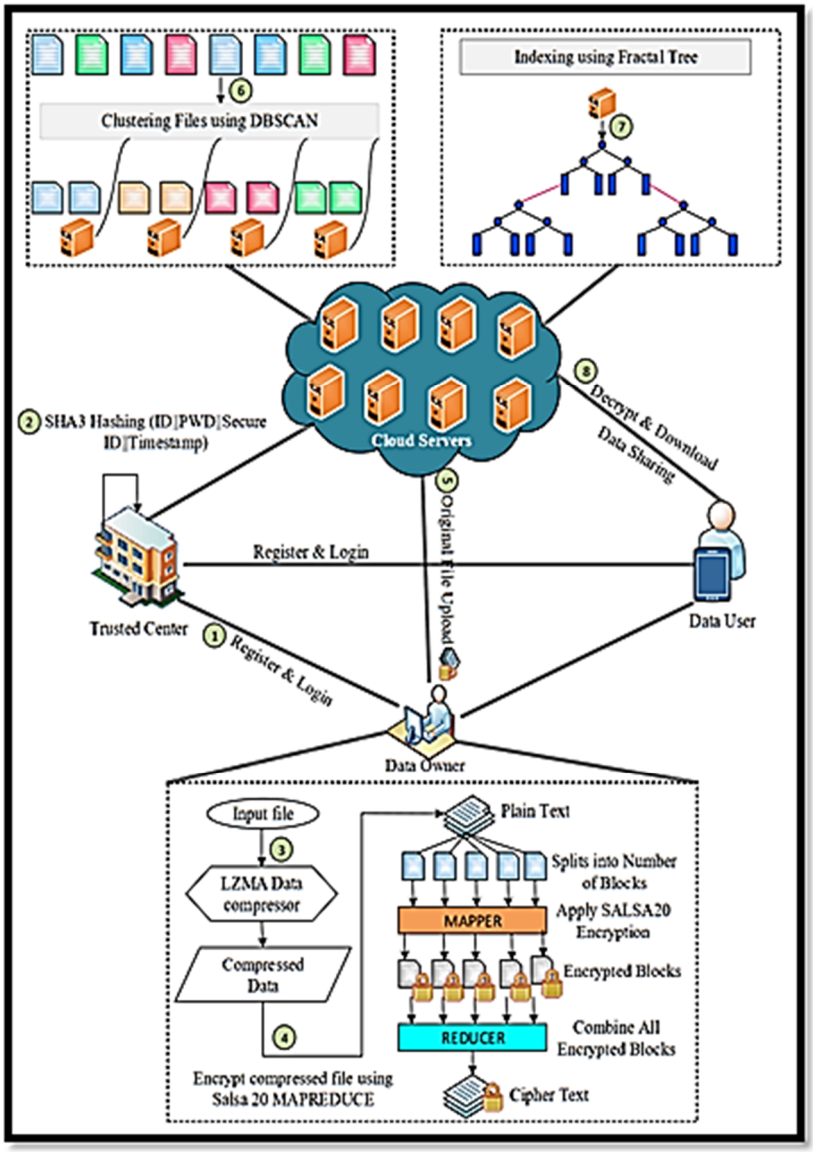

Figure 5 Proposed System Architecture with Big Data Outsourcing, Sharing, and Management (Narayanan et al., 2020)

The information put away in the cloud is consistently inclined to go after predominantly due to the unknown, virtual, shared, repeated, multi-hub capacity, and execution. Unapproved access or control prompts information robbery or information misfortune. Large information re-appropriating from information proprietors to the cloud is in three stages as follows:

- Authentication
- Compression
- Encryption's

4.1 Authentication

The method involved with perceiving a's client's character is known as verification. It is the methodology for partner an approaching solicitation with a bunch of distinguishing qualifications. The capabilities were contrasted with those on a record in a data set or inside a confirmation server containing the endorsed client's information. Identification and true validation are the two stages of authentication. The identification stage provides the security framework with a user identity. This is used to create a user ID. The security framework will search through all of the abstract articles it has access to in order to find the one that the real user is now using.

The user will be identified once this is completed. Authentication is the process of determining a user's identification by checking user-provided proof, and a credential is proof provided by the user during the verification. Conflicting or insufficient authorization, on the other hand, might create security gaps that should be identified and closed as quickly as possible (Mall, S. and Saroj, S.K., 2018). In data outsourcing, Authentication phases consist of three steps as follows:

I. Registration: The Identity Information, Username, Current Timestamp, and Secure ID identities are used by data owners to register with TC. The hashed User ID and Password are then sent to TC for registration. Following registration, TC computes a hash value for DO based on the supplied data using the SHA3-384 hashing method. The figure shows the registration procedure in great detail.

II. Login: The procedure begins when DO enter credentials for CS login. It is crucial for User ID, Passcode, Present Reference number, and Secure Driver's license to be precise. Then, if the authentication is successful, wait for the CS answer.

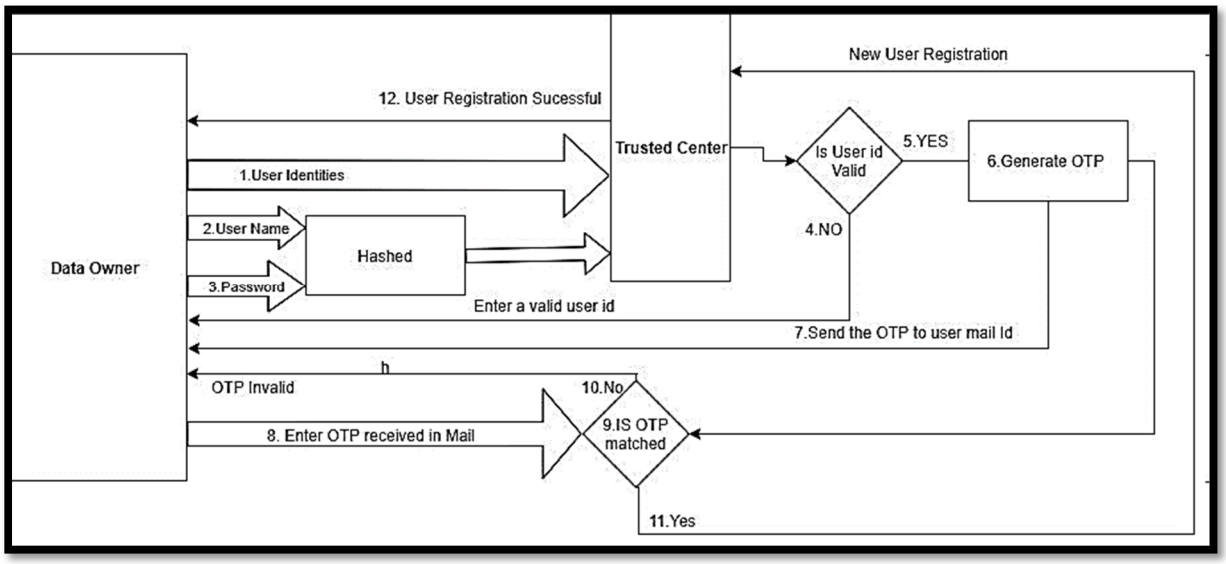

Figure 6 Data User Registration with Trusted Center (Narayanan et al., 2020)

III. **Authentication**: For login considerations, all credentials are scrambled and saved in the memory. To compare the database to the supplied DO information, it is hashed once more. Figure 7 shows the phases in the authentication procedure. In order to use the cloud-based services, the customer must verify their electronic identity. Hence shown as two separate processes.

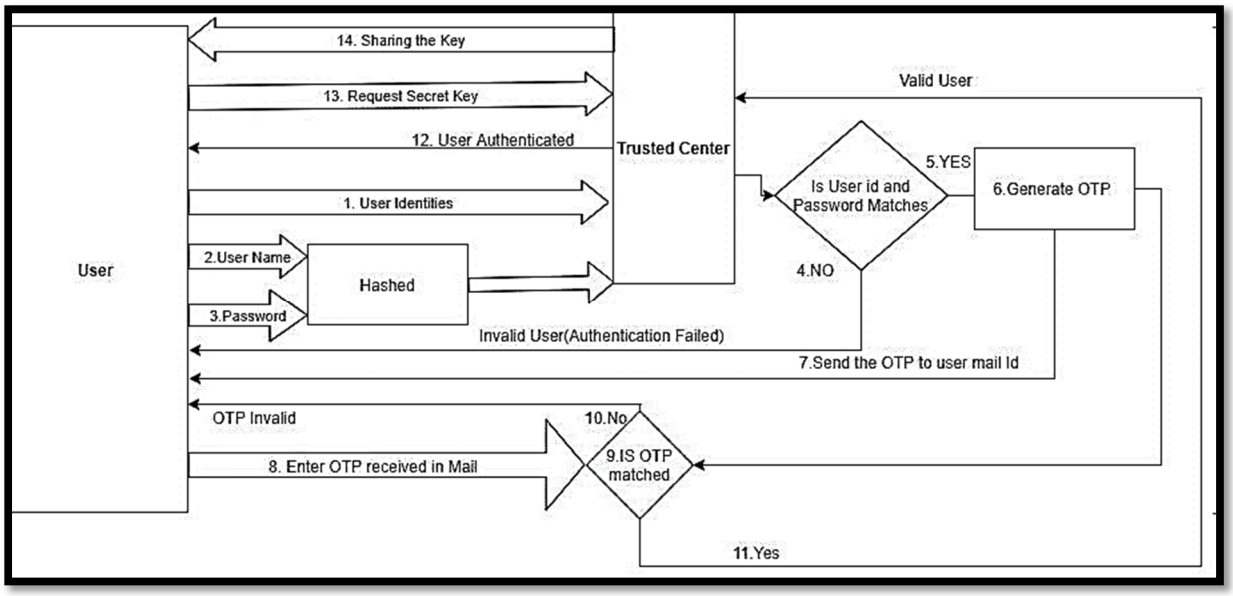

Figure 7 User Authentication with Trusted Center (Narayanan et al., 2020)

4.1.1 Hashing Algorithm

The password is scrambled in this approach, but the user's authenticity may still be checked. MD5 is one of the most straightforward algorithms available. It generates a 128-bit hash that is susceptible to collisions. As a result, I don't utilize it for signatures that are really sensitive. SHA was created in 1973 by the International Bureau of Standards (NIST) as well as the Security Agency (NSA) (Goyal, V. and Kant, C., 2018). It generates a 160-bit hash, although Collisions are possible. Every hash should be self-contained. Because of some of the inconsistencies and vulnerabilities in the original SHA, it was revised in 1995.

However, because that one has been shown to contain collisions, it has been deprecated as well. Both SHA-256 and SHA-512 were developed in 2001 and produce hashes of 256 and 512 bits, respectively. And their block size is going to be 1024, as I specified in the input. Each of these algorithms goes through a particular number of cycles while we're talking about it. As a result, I condense a large document or file into a little algorithm, such as a 160bit, 256bit, or 512bit hash. When I modify a single bit in the information, the output changes completely, indicating that it is extremely sensitive to changes. Authentication is far more difficult to do consistently. I could use hashing to guard against credential sniffing instead of hashing the authentication database. As a result, password hashing is a good security measure.

SHA3 is a secure hashing method that includes the SHA3-224, SHA3-256, SHA3-384, and SHA3-512 hashing functions. SHAKE-128, SHAKE256, and two Extendable Output Functions (XOFs) are also included. In registration, the SHA-3 hashing method is proposed for message authentication. It uses the Keccak algorithm in conjunction with Sponge Construction (Hababeh et al., 2018).

The inputs for hashing include Domain and Bit Rate as well as Padding functions, which are lists of bytes or integers with values ranging from 0 to 255. It considers the website domain Hash, XOF, and KEC, each of which carries out domain separation by recognizing input, correlating to a hash function, and requesting different paddings. The padding method ends by delivering an array of padding message blocks, each of which is made of a list of numbers. The aforementioned method is employed at both TC and CS for hashing in order to ensure user authentication. The method is designated by PROC, and the output length in bits is denoted by N. (224, 256, 384, and 512) (Patel, S., Patel, V. and Jinwala, D., 2013).

The message's type is MT and its message is M. The entire method is carried out in order to hash DO and DU data. The time it took to generate the key and authenticate it is depicted in the graph. In our suggested approach, the time cost of key generation is nearly constant. That is to say, as the number of users grows, the time cost of key generation remains relatively constant. In contrast to key generation, the time cost of authentication increases exponentially as the number of users grows. However, the authentication time cost is still within acceptable limits.

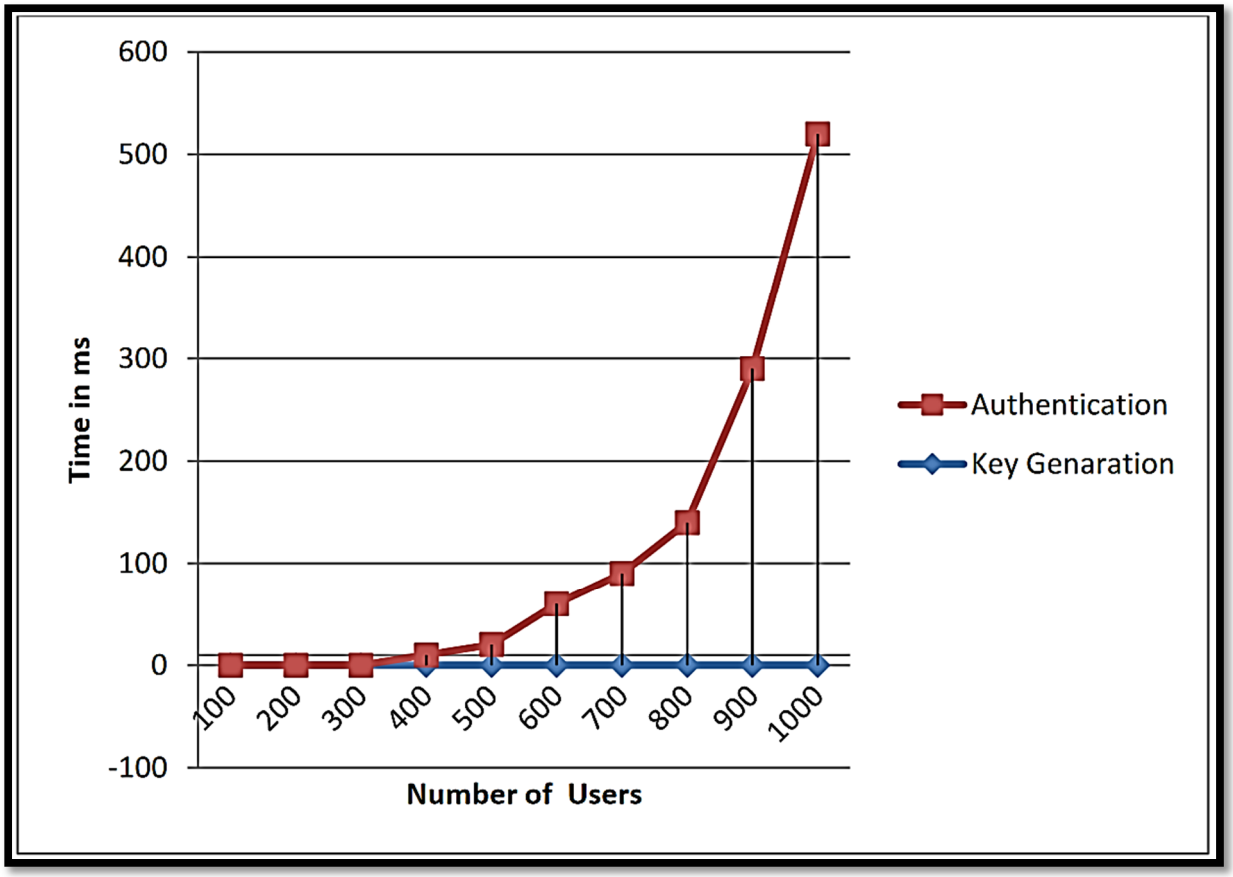

Figure 8 Time is taken for Key Generation and Authentication

Because each user's authentication must be validated, it takes time. As a result, the time curve for authentication has risen significantly. It is simple to produce many keys using a key generation method. As a result, even if there is a rise in key generation time, the time variance is substantially smaller when compared to the time variation for authentication (Roy, B., 2014).

4.2 Compression

Compression is a technique for reducing the amount of data without sacrificing information. Compression allows you to boost transmission speed while lowering storage costs. A measurement that used gauge how often a compression technique shrinks the size of the data representation is the data compression ratio, sometimes referred to as compression power. By multiplying the reduced file size by the original uncompressed file size, the Combustion Engine (CR) of any compression technology may be computed. In general, the compression ratio refers to a compression technique's ability to reduce file size. As a result, if a compression approach has a high CR, it means the compression technique is ineffective at reducing the input file size, or vice versa. The lossless compression methods used in our work is discussed below (Bilge, A. and Polat, H., 2013).

4.2.1 Ma for Data Compression

I first compress the data before I encrypt it. In order to overcome the issues with Huffman compression, I suggest the Lempel Ziv Markov Algorithm (LZMA) in this work. The data is then encrypted using the aforementioned method. The gamma encoder and rolling dictionary generator with LZ77 dynamic dictionary encoding are used in the LZMA encoder, and the output is presented as tuples that have included Offset, Length, and New Symbol. The delta encoding and processor operate in the manner as described elsewhere here:

- **Delta Encoder**: It uses the Sliding Window to produce the input data for compression. It saves and sends data in a consecutive order.
- **Delta Decoder:** The original data streaming is kept through this entity, and succeeding bytes are maintained by appending the recent information syllable to the original data bytes.

The figure 9 depicts the process. Assume the file size was X'ING before compression and was reduced to Z after delta encoding. The delta decoder receives the decreased file and decompresses the image, resulting in a file size of XL.

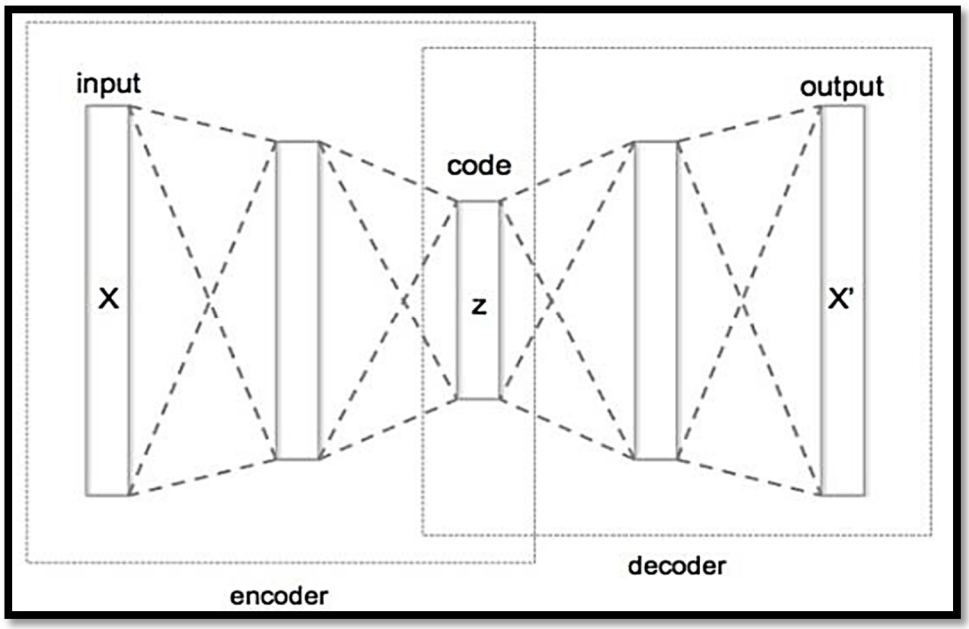

Figure 9 The process of Encoding and Decoding

Advantages of LZMA over Huffman

- LZMA does not in itself required prior knowledge of the raw data stream beforehand.
- The savings deposit can be compacted using LZMA in a single motion.
- LZMA has had the benefit of just being unambiguous, that accelerates operations

4.3 Encryption

I scramble packed information in the third step of huge information moving to guarantee information security in CS. There are two sorts of cryptography, contingent upon the security keys used to scramble and decode the information. Awry and symmetric encryption strategies fall under these two classes. Single key cryptography is otherwise called symmetric encryption. The collector and shipper should settle on a common key for this encryption method to work (Liu, J., Huang, J.Z., Luo, J. and Xiong, L., 2012). Encryption makes tangled data that is generally a similar length as the plaintext, given the plaintext and the key. Unscrambling is something contrary to encryption, and it involves a similar key as encryption. The block graph of the Symmetric key is displayed in Figure 10.

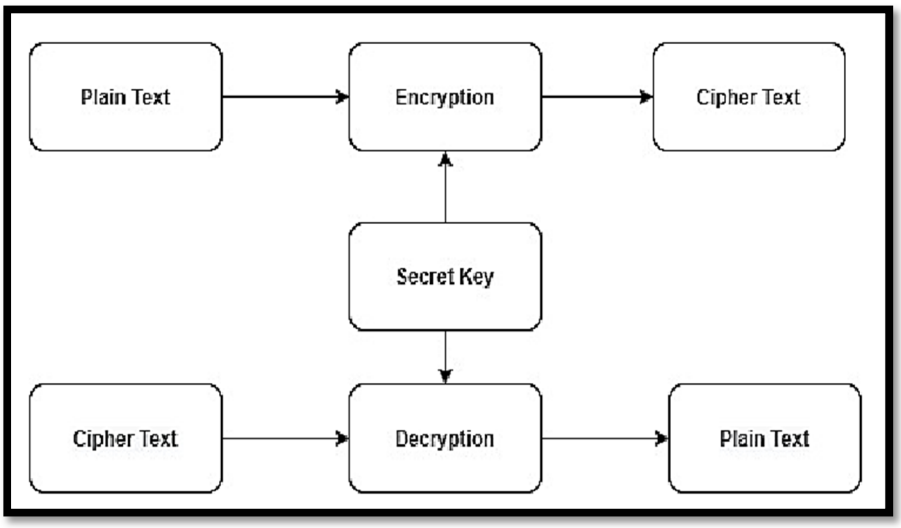

Figure 10 Symmetric Key Cryptography Process

4.3.1 Asymmetric Encryption

It utilizes two keys: a public key that is known to everybody and is utilized for encryption; and a confidential key that is simply known to the client of that key and is utilized for interpreting. The public key and the confidential key are numerically connected to one another. Toward the day's end, data encoded by one public key can be mixed independently by the confidential key related with it. Figure shows the encryption and deciphering framework.

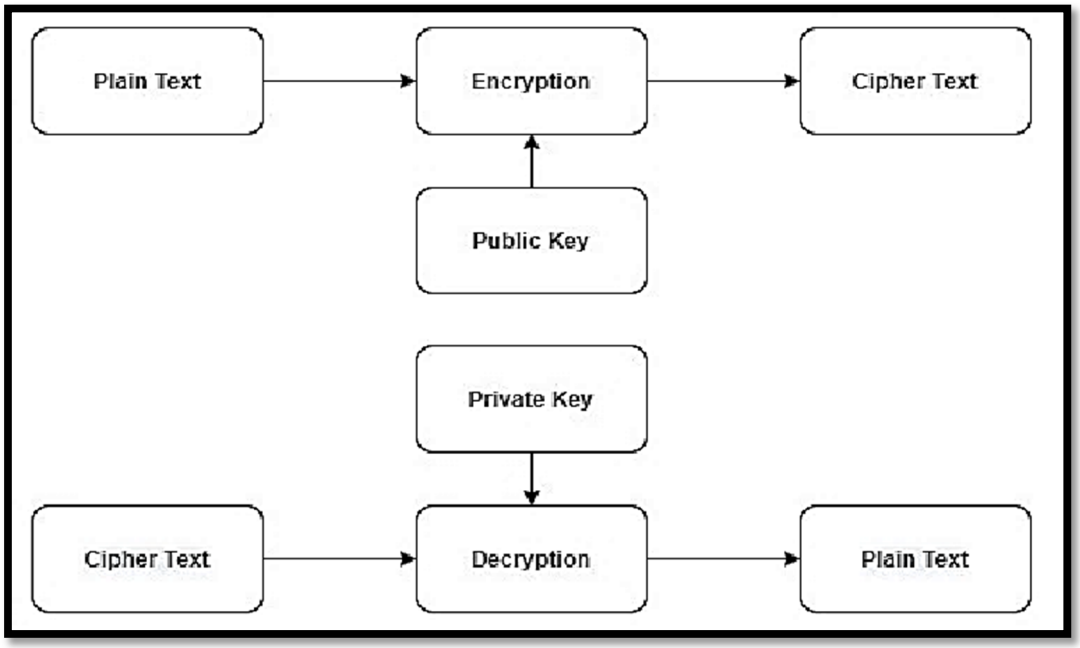

Figure 11 Public Key Cryptography Process

Compared Algorithms (DES, AES, and Blowfish): I selected AES and Blowfish techniques for comparison reasons because they are frequently used in recent works. As a result, I compared

the AES and Blowfish algorithms with our proposed technique on various data volumes ranging from 1MB to 10 GB.

- DES: The National Institute of Standards and Technology (NIST) delivered DES as the essential encryption standard (NIST). In view of IBM's Lucifer figure, it was executed. In 1974, DES turned into a norm (www.tropsoft.com). DES utilizes a 56-bit key and converts 64-cycle contribution to 64-bit yield. The key seems to be a 64-bit key, yet the slightest bit in every one of the eight octets is utilized for odd equality on every octet. There have been different assaults and strategies depicted to date, all of which exploit DES' defects, delivering it a shaky block figure.
- AES: often known as the Rijndael encryption, is a symmetric block cypher that can encrypt a 128-bit data block using symmetric keys of 128, 192, or 256 bits. The only known effective attack against this method is a brute force attack (Liu, J., Huang, J.Z., Luo, J. and Xiong, L., 2012).

Blowfish is a symmetric block figure that can be utilized for data encryption and insurance. It utilizes a variable-length key going from 32 to 448 pieces, making it ideal for information check. In 1993, Bruce Schneider made Blowfish as a rapid, free option in contrast to existing encryption procedures. Blowfish is unpatented, grant free, and open to general society. In spite of the way that it experiences difficulties with feeble keys, no attack has been viewed as compelling against it.

DO demand a private key for data encryption after successful TC authentication. The sensitivity level requested by the DO is used to generate private keys by TC. DO define two levels of sensitivity: (i) Sensitive and delicate (ii). Extremely sensitive. Sensitive information, on the other hand, requires access restriction. As a result, the number of data accesses is tracked in the CS to prevent security breaches. TC generates keys with the SALSA20 Encryption algorithm. SALSA 20 uses buttons with key sizes of 128 bits for non-sensitive data and 256 bits for critical material, respectively. It takes a long time the encrypt and decode enormous volumes of data that must be outsourced to the cloud, which is a time-consuming process (Yang et al., 2013). To deal with these issues, the SALSA20 Encryption-Map Reduce framework is being proposed. The suggested procedures and operating methods are listed below:

4.3.2 SALSA20 Algorithm

The Ultra-Modern Stream Cipher algorithm SALSA20 is utilized for encryption. The major components are Percentage point, Column, Section, as Well as double. High-level security methodology Objectives have been identified is quicker and much more knowledgeable than its early pioneers. The following are the advantages of utilizing SALSA20 encryption:

- It performs encryption and decryption processes at a quicker rate than AES, i.e., it is 3-5 times faster than AES.
- Differential Cryptanalysis, the most common symmetric key, is mitigated.
- While Lookup Indexes are not required, Temporal Attacks are actually suppressed. Easy to implement.
- To speed up encryption, AEs increase the key size first, then perform the encryption operation, which adds extra processing overhead and lengthens the key setup procedure.

It starts with 64 bytes of input and ends with 64 bytes of output. The key stream creation, encryption, and decryption procedures are all part of this technique (Pearson, S. and Charlesworth, A., 2009).

- There's also the quarter round's mathematical function.
- Then there's the row round function, which modifies the matrix's rows.
- Following that, there comes the column round function, which modifies the column in the matrix.
- Finally, employ another double round method, such as SALSA20/20 or Period preceding, which will loop however many repetitions how you choose.
- Just that little ability comes following.

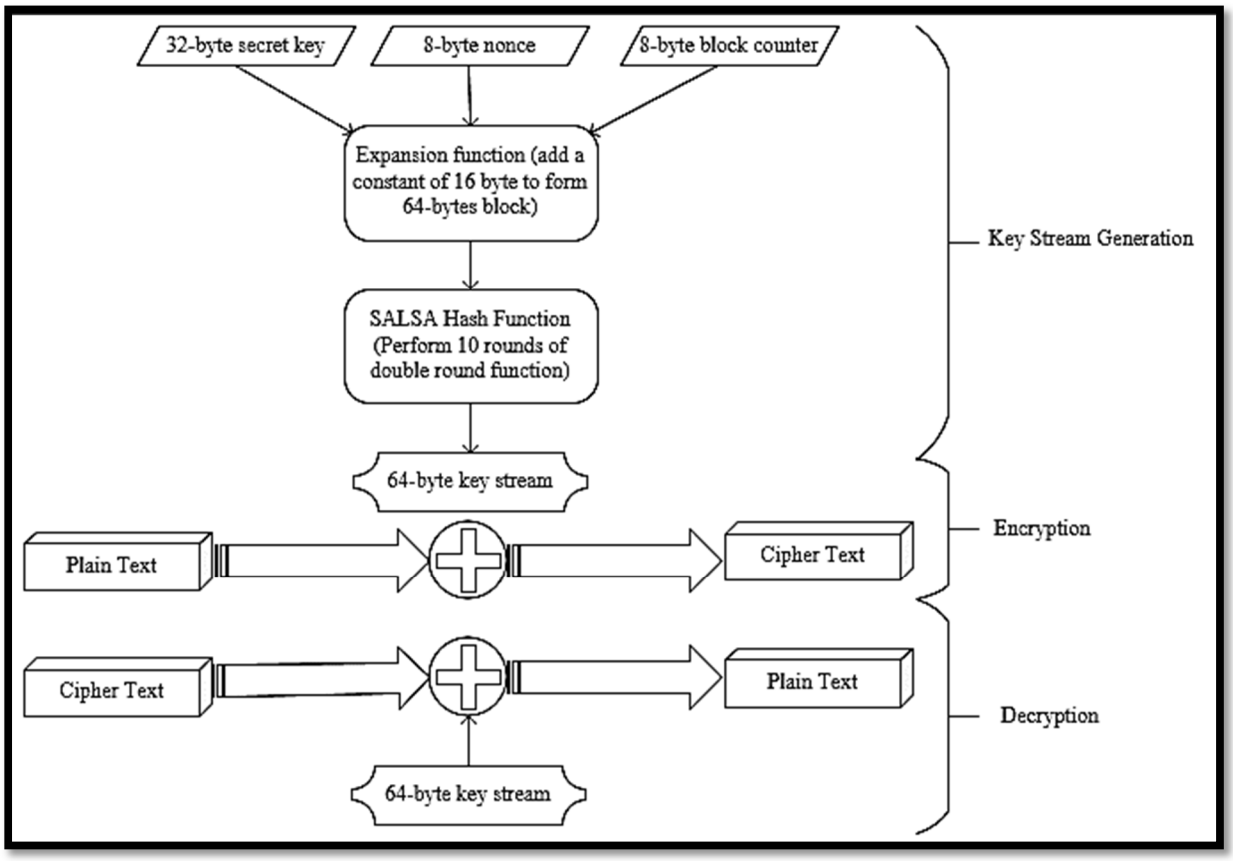

Figure 12 SALSA20 Algorithm (Narayanan et al., 2020)

4.3.3 SALSA20 with Map Reduce

I propose the SALSA20 method with the Map Reduce paradigm for encrypting large amounts of data (SALSA20MR). It's an encryption algorithm that outperforms existing symmetric algorithms in various ways. Hadoop is one of the most scalable platforms available. This is due to the ability to spread and store huge volumes of data across many servers. Servers those are both affordable and capable of running in parallel (Weiquan, X. and Houkui, W., 2013). Due to parallel processing and the addition of servers, more processing power may be added.

Dissimilar to customary social information base administration frameworks, which can't scale to deal with a lot of information, Hadoop Map Reduce programming permits organizations to run applications across huge number of hubs, possibly including great many terabytes of information. Hadoop, which utilizes the disseminated record framework, or HDFS, as a stockpiling strategy, utilizes a planning way to deal with find information in a bunch. The Map Reduce information handling devices, which are regularly put on similar servers, empower for quicker information handling. Regardless of whether you're working with gigantic volumes of

unstructured information, Hadoop Map Reduce just requires minutes to handle terabytes of information, and hours to handle petabytes of information.

4.3.4 Map Reduce Programming Model

Traditional cryptography techniques are insufficient for huge data encryption, necessitating significant improvements in terms of speed and efficiency. Map Reduce is Hardtop's appropriated information handling stage, which was intended to deal with gigantic measures of information. It's adaptable and speedy at handling a lot of information. It is a very useful asset for equal handling. It utilizes Hadoop Distributed File System (HFFS) to run client occupations. The sequence of HDFS write operations is shown in Figure 11, while the order of HDFS read operations is shown in Table 2.

Comparison of SALSA20MR

I implement the following phases in our proposed SALSA20 with Map Reduce paradigm:

- DO data is divided into a set of fixed-size chunks.
- After that, I use the Map Reduce approach to encrypt these blocks.
- The Mapper capability is called to scramble the block, and it encodes the information block in lined up prior to changing all scrambled blocks.
- After that, the Reducer capability is called to solidify all scrambled blocks into a solitary record (Encrypted document) that the Mapper capability has returned.
- After that, the encrypted file is sent to CS.

File Size	**Cipher Suites**		
	AES-128	SALSA20	SALSA20 with MapReduce
10MB	0.85s	0.72s	0.36s
20MB	1.80s	1.55s	0.77s
50MB	4.50s	4.12s	2.06s
100MB	8.78s	8.58s	4.29s
200MB	18.02s	17.44s	8.72s

Table 2 Performance Test for Cipher Suites

Data Set Size	SALSA20	AES	Blowfish
1MB	6408	11851	16933
50MB	9163	22989	151470
100MB	13686	79812	269597
500MB	66749	367194	1474134
1GB	132514	783624	2948050
2GB	271884	1611347	5762432
5GB	687386	3742760	14021810
10GB	1366285	7157654	24211214

Table 3 Encryption time for different Algorithms in Map Reduce

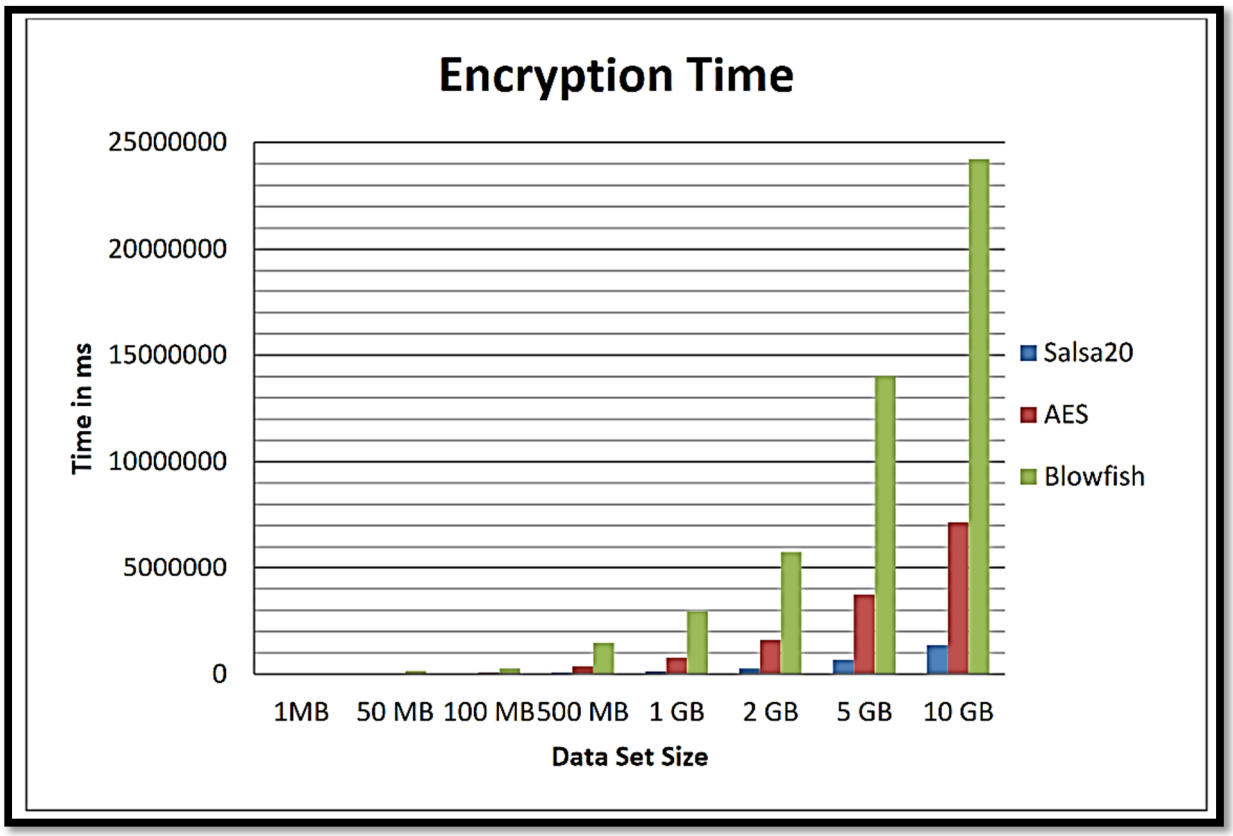

Figure 13 Encryption time for different algorithms in Map Reduce

Our proposed approach requires less time to encrypt even larger data sets, according to the results of the investigation. The same can be seen in Table 2 and its graphical representation in Figure 13. As a result, SALSA20MR appears to be the best fit for our needs. SALSA20MR was created with the goal of being able to encode any arbitrary block of data, making it suitable for use as a block cipher.

5. DISCUSSION AND CONCLUSION

In 4 entities that up our increasingly considering were the Trustworthy Centre (TC), Data Owner (DO), Data User (DU), and Cloud Server (CS). When interacting with big volumes of information, security / privacy is two of the most crucial factors to take into account. I have found that the best way to implement Big Security and privacy would be through regulation, not technologies. However, regulation cannot keep up with technical improvements and differs from nation to country. As a result, new security technologies and other techniques are continually being developed and refined.

In the preceding section, I looked at a variety of alternative tactics and methodologies for assuring security and privacy in Big Data. Also stressed was the importance of large-scale data correlation in order to fully use Big Data, as well as the source of Big Data security and privacy problems, as previously stated. Our study on "no correlation" has led us to believe that it is possible to accomplish Big Data security and privacy using this method. Though still in its early stages, block chain technology is beginning to have an impact on the way Big Data is processed and analysed in the enterprise. Blockchain can possibly essentially change how Big Data is overseen and handled, with further developed security and information quality being only a couple of the advantages that this innovation conveys to organizations, among different advantages. The force of huge information examination to affect organization activities might turn out to be much seriously engaging from now on. However long advances in this field proceed, I might hope to see extra headways in the association between Big Data examination and block chain innovation.

New and more functional use cases for Big Data the board and information examination will be created and investigated as the innovation propels and new leap forwards are found. Observing how the block chain continues to revolutionise other industries while also enhancing data privacy will be fascinating as more real-time data is acquired. By leveraging the usage of specialised current technologies, I suggested four feasible options for addressing the privacy and security problems associated with Big Data at every step of the data's life cycle. For the purpose of addressing security problems that arise during the Big Data life cycle, existing technologies were examined and relevant adjustments implemented, including data provenance, data encryption and access control, data mining, and distributed ledger technology (Blockchain). Over the course of this report, a range of hurdles, unsolved problems, and

prospective technology adoption approaches in the subject of Big Data security were discovered and discussed. The ramifications of this decision might lead to more research into this critical problem in the future.

The application of artificial intelligence (AI) in the workplace accelerates the process of resolving cyber risks. Malware detection, which was previously impossible to accomplish using traditional methods, has now become straightforward owing to artificial intelligence. Given that artificial intelligence keeps track of all previous data breaches, any suspicious activity may be recognised rather soon. Another distinction between this method and the previous one is that AI apps lower the amount of time that IT employees must spend. In many situations, artificial intelligence is employed in order to discover, identify, simplify, and resolve issues. Artificial intelligence (AI) provides the system with the capacity to monitor risks automatically and address them with automated solutions, among other things. Artificial intelligence also assists in the categorization of cyber-attacks based on their ability to cause harm to the organisation. Despite the fact that some experts believe that 50 percent of all cyber security systems need to be upgraded in order to reduce the number of threats, the fact that these cyber-threats are growing more ubiquitous, impactful, and sophisticated makes them all the more frightening. When a company's infrastructure and resources are completely depleted by several compromised attacks, such as Dados attacks, it is said to be "completely wiped out".

Despite this, the advancement of artificial intelligence technology has the potential to considerably alleviate these issues and aid us in our search for solutions. In 2017, cyber security accounted for $3.92 billion of the global market, and it is expected to expand to $34.81 billion by 2025, according to forecasts. There are a number of products such as Dark Trace and Sophos that are making a name for themselves.

In network anomaly analysis, observation, detection, and investigation, Big Data is a treasure trove of information that may be mined for useful information. In order to successfully handle the problem, the information obtained from diverse sources assists in the finding of security-related information and the identification of potential threats. Using data, problems may be identified and resolved in a fraction of the time it would take without it. It also assists the cyber analyst in predicting the possibility of an attack or an entry into the network. To defend against cyber-attacks in today's environment, the vast majority of enterprises rely on Big Data to build their defences. In order to resist hackers, around 84% of firms are using data, which has resulted

in a drop in the number of cyber security breaches. The data may actually be used as an advanced analytics tool to identify cyber security threats such as hostile insider programmes, malware/ransom ware attacks, and compromised and insecure equipment, among other things.

6. REFERENCES

1. Aditham, S. and Ranganathan, N., 2017. A system architecture for the detection of insider attacks in Big Data systems. *IEEE Transactions on Dependable and Secure Computing*, *15*(6), pp.974-987.
2. Adnan, N.A.N. and Ariffin, S., 2018, August. Big Data security in the web-based cloud storage system using 3D-AES block cipher cryptography algorithm. In *International Conference on Soft Computing in Data Science* (pp. 309-321). Springer, Singapore.
3. Agrawal, D., Budak, C. and El Abbadi, A., 2011, October. Information diffusion in social networks: observing and affecting what society cares about. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2609-2610).
4. Alhanahnah, M., Bertok, P., Tari, Z. and Alouneh, S., 2018. Context-aware multifaceted trust framework for evaluating trustworthiness of cloud providers. *Future Generation Computer Systems*, *79*, pp.488-499.
5. Amudhavel, J., Padmapriya, V., Gowri, V., Lakshmipriya, K., Kumar, K.P. and Thiyagarajan, B., 2015, March. Perspectives, motivations and implications of Big Data analytics. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)* (pp. 1-5).
6. Ateniese, G., Fu, K., Green, M. and Hohenberger, S., 2006. Improved proxy re-encryption schemes with applications to secure distributed storage. ACM Transactions on Information and System Security (TISSEC), 9(1), pp.1-30.
7. Bao, R., Chen, Z. and Obaidat, M.S., 2018. Challenges and techniques in Big Data security and privacy: A review. *Security and Privacy*, *1*(4), p.e13.
8. Bashari Rad, B., Akbarzadeh, N., Ataei, P. and Khakbiz, Y., 2016. Security and privacy challenges in Big Data era. International Journal of Control Theory and Applications, 9(43), pp.437-448.
9. Bhandari, R., Hans, V. and Ahuja, N.J., 2016. Big Data security–challenges and recommendations. *International Journal of Computer Sciences and Engineering*, *4*(1), pp.93-98.

10. Bhatia, D. and Wankhede, S., 2015. A Study Of Security Issues In Cloud Computing Architecture. *International Journal Of Advanced Research In Datamining And Cloud Computing Issn*, 3(5), pp.391-396.
11. Bilge, A. and Polat, H., 2013. A scalable privacy-preserving recommendation scheme via bisecting k-means clustering. *Information Processing & Management*, *49*(4), pp.912-927.
12. Bocek, T., Hunt, E., Hausheer, D. and Stiller, B., 2008, April. Fast similarity search in peer-to-peer networks. In *NOMS 2008-2008 IEEE Network Operations and Management Symposium* (pp. 240-247). IEEE.
13. Burdonov, I., Kosachev, A. and Iakovenko, P., 2009, March. Virtualization-based separation of privilege: working with sensitive data in untrusted environment. In *Proceedings of the 1st EuroSys Workshop on Virtualization Technology for Dependable Systems* (pp. 1-6).
14. Chandramouli, R., 2011, November. Service Model Driven Variations in Security Measures for Cloud Environments. In *IADIS International Conference Applied Computing 2011; November 6-8, 2011; Rio de Janeiro, Brazil* (pp. 527-530). International Association for Development of the Information Society (IADIS).
15. Chang, Y.S., 2018, March. The moderator effect of working memory and emotion on the relationship between information overload and online health information quality judgment. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 345-347).
16. Ding, S., Wang, Z., Wu, D. and Olson, D.L., 2017. Utilizing customer satisfaction in ranking prediction for personalized cloud service selection. *Decision Support Systems*, *93*, pp.1-10.
17. Fan, K., Lou, S., Su, R., Li, H. and Yang, Y., 2018. Secure and private key management scheme in Big Data networking. *Peer-to-Peer Networking and Applications*, *11*(5), pp.992-999.
18. Fan, W.J., Yang, S.L., Perros, H. and Pei, J., 2015. A multi-dimensional trust-aware cloud service selection mechanism based on evidential reasoning approach. *International Journal of Automation and Computing*, *12*(2), pp.208-219.
19. Gain, U. and Hotti, V., 2013. Big Data analytics for professionals, data-milling for laypeople. *World*, *1*(2), pp.51-57.

20. Gao, W., Zhu, Y., Jia, Z., Luo, C., Wang, L., Li, Z., Zhan, J., Qi, Y., He, Y., Gong, S. and Li, X., 2013. Bigdatabench: a Big Data benchmark suite from web search engines. *arXiv preprint arXiv:1307.0320*.
21. Ghosh, N., Ghosh, S.K. and Das, S.K., 2014. SelCSP: A framework to facilitate selection of cloud service providers. *IEEE transactions on cloud computing*, *3*(1), pp.66-79.
22. Goh, E.J., 2003. Secure indexes. *Cryptology ePrint Archive*. [Online] Available at: https://eprint.iacr.org/2003/216.pdf [Accessed 03 August 2022].
23. Goyal, V. and Kant, C., 2018. An effective hybrid encryption algorithm for ensuring cloud data security. In *Big Data analytics* (pp. 195-210). Springer, Singapore.
24. Hababeh, I., Gharaibeh, A., Nofal, S. and Khalil, I., 2018. An integrated methodology for Big Data classification and security for improving cloud systems data mobility. *IEEE Access*, *7*, pp.9153-9163.
25. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B. and Babu, S., 2011, January. Starfish: A Self-tuning System for Big Data Analytics. In *Cidr* (Vol. 11, No. 2011, pp. 261-272).
26. Jenkins, C., Schulte, M. and Glossner, J., 2010, November. Instruction set extensions for Triple DES processing on a multi-threaded software-defined radio platform. In *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers* (pp. 1387-1391). IEEE.
27. Jiang, R., Lu, R. and Choo, K.K.R., 2018. Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. *Future Generation Computer Systems*, *78*, pp.392-401.
28. Jin, B., Wang, Y., Liu, Z. and Xue, J., 2011. A trust model based on cloud model and bayesian networks. *Procedia Environmental Sciences*, *11*, pp.452-459.
29. Kaisler, S., Armour, F., Espinosa, J.A. and Money, W., 2013, January. Big Data: Issues and challenges moving forward. In *2013 46th Hawaii international conference on system sciences* (pp. 995-1004). IEEE.
30. Liang, G., 2014, May. Security of huge data in the digital campus on the cloud computing mode. In *2014 IEEE Workshop on Electronics, Computer and Applications* (pp. 736-739). IEEE.
31. Liu, J., Huang, J.Z., Luo, J. and Xiong, L., 2012, March. Privacy preserving distributed DBSCAN clustering. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops* (pp. 177-185).

32. Liu, Y., Esseghir, M. and Boulahia, L.M., 2016. Evaluation of parameters importance in cloud service selection using rough sets. *Applied Mathematics*, *7*(06), p.527.
33. Macías, M. and Guitart, J., 2016. Analysis of a trust model for SLA negotiation and enforcement in cloud markets. *Future generation computer systems*, *55*, pp.460-472.
34. Mall, S. and Saroj, S.K., 2018. A new security framework for cloud data. *Procedia computer science*, *143*, pp.765-775.
35. Manuel, P., 2015. A trust model of cloud computing based on Quality of Service. *Annals of Operations Research*, *233*(1), pp.281-292.
36. Mishra, K., Bhattacharjee, V., Saket, S. and Mishra, S.P., 2022. Cloud and Big Data Security System's Review Principles: A Decisive Investigation. *Wireless Personal Communications*, pp.1-38.
37. Mollah, M.B., Azad, M.A.K. and Vasilakos, A., 2017. Security and privacy challenges in mobile cloud computing: Survey and way ahead. *Journal of Network and Computer Applications*, *84*, pp.38-54.
38. Moreno, J., Serrano, M.A. and Fernández-Medina, E., 2016. Main issues in Big Data security. *Future Internet*, *8*(3), p.44.
39. Narayanan, U., Paul, V. and Joseph, S., 2020. A novel system architecture for secure authentication and data sharing in cloud enabled Big Data Environment. *Journal of King Saud University-Computer and Information Sciences*.
40. Noor, T.H., Sheng, Q.Z., Ngu, A.H., Alfazi, A. and Law, J., 2013, October. Cloud armor: a platform for credibility-based trust management of cloud services. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 2509-2512).
41. Oguntimilehin, A. and Ademola, E.O., 2014. A review of Big Data management, benefits and challenges. *A Review of Big Data Management, Benefits and Challenges*, *5*(6), pp.1-7.
42. Parmar, R.R., Roy, S., Bhattacharyya, D., Bandyopadhyay, S.K. and Kim, T.H., 2017. Large-scale encryption in the Hadoop environment: Challenges and solutions. *IEEE Access*, *5*, pp.7156-7163.
43. Patel, S., Patel, V. and Jinwala, D., 2013, February. Privacy preserving distributed k-means clustering in malicious model using zero knowledge proof.

In *International Conference on Distributed Computing and Internet Technology* (pp. 420-431). Springer, Berlin, Heidelberg.

44. Pearson, S. and Charlesworth, A., 2009, December. Accountability as a way forward for privacy protection in the cloud. In *IEEE international conference on cloud computing* (pp. 131-144). Springer, Berlin, Heidelberg.
45. Phaneendra, S.V. and Reddy, E.M., 2013, April. Big Data-solutions for RDBMS problems-A survey. In *12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010)(Osaka, Japan, Apr 19 {23 2013)*.
46. Qu, L., Wang, Y. and Orgun, M.A., 2013, June. Cloud service selection based on the aggregation of user feedback and quantitative performance assessment. In *2013 IEEE international conference on services computing* (pp. 152-159). IEEE.
47. Reddy, Y., 2018, May. Big Data Processing and Access Controls in Cloud Environment. In *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)* (pp. 25-33). IEEE.
48. Roy, B., 2014. Performance analysis of clustering in privacy preserving data mining. *International Journal of Computer Applications Information Technology*, *5*(4), pp.35-39.
49. Sinanc, D. and Sagiroglu, S., 2013, November. A review on cloud security. In *Proceedings of the 6th International Conference on Security of Information and Networks* (pp. 321-325).
50. Singh, S., Jeong, Y.S. and Park, J.H., 2016. A survey on cloud computing security: Issues, threats, and solutions. *Journal of Network and Computer Applications*, *75*, pp.200-222.
51. Somu, N., Kirthivasan, K. and Shankar Sriram, V.S., 2017. A rough set-based hypergraph trust measure parameter selection technique for cloud service selection. *The Journal of Supercomputing*, *73*(10), pp.4535-4559.
52. Tan, J., Meng, S., Meng, X. and Zhang, L., 2013, April. Improving reducetask data locality for sequential mapreduce jobs. In *2013 Proceedings IEEE INFOCOM* (pp. 1627-1635). IEEE.
53. Tan, J., Meng, X. and Zhang, L., 2013, April. Coupling task progress for mapreduce resource-aware scheduling. In *2013 Proceedings IEEE INFOCOM* (pp. 1618-1626). IEEE.

54. Toshniwal, R., Dastidar, K.G. and Nath, A., 2015. Big Data security issues and challenges. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(2).
55. Venkatraman, S. and Venkatraman, R., 2019. Big Data security challenges and strategies. *AIMS Mathematics*, *4*(3), pp.860-879.
56. Vidyasagar, S.D., 2013. A study on "Role of Hadoop in information technology era". *GRA-GLOBAL RESEARCH ANALYSIS*, *2*(2).
57. Viji Rajendran, V. and Swamynathan, S., 2016. Hybrid model for dynamic evaluation of trust in cloud services. *Wireless Networks*, *22*(6), pp.1807-1818.
58. Vorugunti, C.S., 2016, November. PPMUAS: A privacy preserving mobile user authentication system for cloud environment utilizing Big Data features. In *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)* (pp. 1-6). IEEE.
59. Wang, S., Wei, J., Sun, L., Sun, Q. and Yang, F., 2013, December. Reputation measurement of cloud services based on unstable feedback ratings. In *2013 International Conference on Parallel and Distributed Systems* (pp. 474-479). IEEE.
60. Ward, J.S. and Barker, A., 2013. Undefined by data: a survey of Big Data definitions. *arXiv preprint arXiv:1309.5821*.
61. Weiquan, X. and Houkui, W., 2013, December. The Design Research of Data Security Model Based on Public Cloud. In *2013 Ninth International Conference on Computational Intelligence and Security* (pp. 607-609). IEEE.
62. Win, T.Y., Tianfield, H. and Mair, Q., 2017. Big Data based security analytics for protecting virtualized infrastructures in cloud computing. *IEEE Transactions on Big Data*, *4*(1), pp.11-25.
63. Yang, P., Gui, X., Tian, F., Yao, J. and Lin, J., 2013, November. A privacy-preserving data obfuscation scheme used in data statistics and data mining. In *2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing* (pp. 881-887). IEEE.
64. Zhao, Y., Li, S. and Jiang, L., 2018. Secure and efficient user authentication scheme based on password and smart card for multiserver environment. *Security and Communication Networks*, *2018*.

65. Zhao, Y., Wu, J. and Liu, C., 2014. Dache: A data aware caching for big-data applications using the MapReduce framework. *Tsinghua science and technology*, *19*(1), pp.39-50.
66. Zhe, D., Qinghong, W., Naizheng, S. and Yuhan, Z., 2017, May. Study on data security policy based on cloud storage. In *2017 ieee 3rd international conference on Big Data security on cloud (bigdatasecurity), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids)* (pp. 145-149). IEEE.