INCEPTION OF DATA CREATION PHASE AND PREVENTION OF DATA LEAKAGE IN BIG DATA LIFE CYCLE

Thesis submitted in fulfillment of the requirements for
the Degree of

DOCTOR OF PHILOSOPHY

BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY

LUCKNOW

प्रज्ञा शील करुणा

ESTABLISHED 1996

in

INFORMATION TECHNOLOGY

by

KANIKA

Supervised by

PROF. R. A. KHAN
Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow

Co-Supervised by

Dr. ALKA
Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow

Submitted to

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW

FEBRUARY-2019

DECLARATION

I, Kanika, solemnly declare that this thesis of research on "**Inception of Data Creation Phase and Prevention of Data Leakage in Big Data Life Cycle**" is my original work. The study has been conducted under the guidance of Prof. Raees Ahmad Khan and Dr. Alka, at Department of Information Technology, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow. It is further declared that to the best of my knowledge and belief it has not been submitted earlier for the award of any degree. I also undertake that the thesis is essentially free from all kinds of plagiarism.

Dated: 07/02/2019

Kanika 07/02/2019

(Kanika)
Researcher
Department of Information Technology
Babasaheb Bhimrao Ambedkar University
(A Central University)
Lucknow, Uttar Pradesh, India

CERTIFICATE

This is to certify that the thesis entitled "**Inception of Data Creation Phase and Prevention of Data Leakage in Big Data Life Cycle**" submitted by **Ms. Kanika** is an original research work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other University.

This thesis submitted to Babasaheb Bhimrao Ambedkar University Lucknow satisfies all the requirements as stipulated in the *Doctor of Philosophy (Ph.D.)* regulations-1999 as amended in 2013 and it is fit for submission and evaluation for the award of the degree of Doctor of Philosophy of the University.

07/02/19

Co-Supervisor

07/02/19

Supervisor

07/02/19

Dated: 07/02/2019

Head of the Department

HEAD
Department of Information Technology
School For Information Science & Technology
Babasaheb Bhimrao Ambedkar University
Lucknow

ACKNOWLEDGEMENTS

Undertaking this Ph.D. has been a truly life-changing experience for me and it would not have been possible to do without the support, guidance and blessings that I received from The Supreme Power and many people.

First, I would like to thank God for the gift of life and for blessings all through my life that allowed me to get here.

I am deeply indebted to Professor and supervisor **Prof. R. A. Khan**, in addition to being the best supervisor one could hope for he made the whole experience so exciting and enjoyable throughout. I am extremely grateful for providing me with the valuable insights that allowed me to consider my work in different contexts. He not only showed belief in me and supported me when I started the long journey, but also provided me with constructive criticism that helped me to refine my work. His enthusiasm and involvement in my research have been instrumental in helping me stay motivated and excited about my dissertation all along. I will always be obliged to him for his wholehearted support and kindness extended to me during my entire course.

I would like to extend my sincere gratitude to my co-supervisor **Dr. Alka** for her guidance, invaluable support and consultations during the course of the study. I will remain indebted to her valuable suggestions during the entire thesis work and providing thoughtful feedback to improve its content. I will always be gratified to her for her unconditional support and kind-heartedness extended to me during my entire course.

I would also like to thank to all the faculty members and office staff of the Department for their cooperation and continuous support extended during the thesis work.

Special thanks to my mother **Mrs. Shashi Sharma** and my father **Mr. Devendra Sharma**, for unconditional love, attention, affection, understanding, motivation and teachings always given to me. Cheers to the support my mother-in-law **Mrs. Sarika**

Vats and my father-in-law **Mr. Ajay Mohan Sharma** extended during the course of my journey.

I would like to single out my beloved husband, **Mr.Utsav Vats**, my companion and friend, from whom I borrowed the most precious time of our marital life and did my research, he has constantly been by my side on this long journey, stimulating me to continue and never stop believing in myself.

To my loving niece, **Aadya**, I thank you for love and moments of tenderness and happiness, for making my life lighter and more joyous. Non-academic support also has been apart in succeeding this work. Mentioning which this I can never skip to pen down my humble gratitude toward with my brothers, without them I would not have been able to show enthusiasm and zeal to execute the task. My heartfelt thanks to my sister-in-law **Prof. Niti Sharma** for her wise counsel and never ending support.

Finally, there are my friends **Dr. Shilpi Singh, Ms. Tahmish Fatima, Ms. Richa Verma,** and **Ms. Jasleen Kaur**. We were not only able to support each other by deliberating over our problems and findings, but were also happy to talk about things other than just our papers.

My colleagues have been a source of inspiration to me and I would like to convey my sincere thanks to **Mr. Tarique Ansari, Mr. Neeraj, Mr. Amal, Mr. Ajay, Dr. Prabhishek** and **Mr. Asim** for their feedback, cooperation and of course friendship and support during the entire Doctorate. I thank you all from the core of my heart.

To all, despite not mentioned, who directly or indirectly have contributed to the accomplishment of this work. Thank you for all your effort extended to my personal and professional training.

Kanika

ABSTRACT

Big data has almost become lifeblood for most of the business to gain profit. It is the combination of popular 3V's, which defines its nature with the three basic characteristics Volume, Velocity and Variety. Big data deals with the creation, collection, storage and transformation of different formats from various sources. Google, using online search to mining large customer's data or predicting big data can be used to support medical and health related tasks including other clinical decision support, disease monitoring, and population health. Amazon may know about each and every book, a user viewed or bought by analyzing huge amount of data collected over the years. The National Security Agency (NSA) can know all phone numbers a person dialed. Facebook may and will examine big data and can tell the birthdays of persons that you did not know. With the introduction of various digital methods all this data has become big data and is still growing.

Eventually, big data technologies can improve decision-making and can provide more insights with faster results but with the negative side of data privacy loss. With the development of more advanced analytical tools for big data, increasing availability of large data sets from different sources makes it more difficult to ensure security. From the last few years, big data research has been spread worldwide. Currently, user's data is one of the most essential assets for the organizations. The constant growth in volume of data has raised a crucial and sensitive problem which cannot be managed by the traditional techniques. Big data has created new challenges linked not only to data's volume, variety or velocity, but also to security and privacy of data. However, this big data's proliferation is not without its risks. The gathered data contains personal information about users or corporate secrets which causes great harm if caught by wrong hands.

The attackers create underground markets where someone can purchase and sell stolen sensitive or personal information. This imposes the need for improved security techniques are to secure big data stored at scattered systems from such ruinous attacks. There have been great efforts to employ a wide range of mechanisms to enhance the privacy of data and thus to make environments more secure. The techniques that have been used for securing data include encryption, trusted platform

module, tokens, access control etc. However, building usable privacy-preserving systems to handle sensitive data securely is still an open problem. Existing privacy and data protection legislation demand strong privacy policy, security and transparency of data usage. In addition, prevention from data leakage with a broad range of emerging or existing security solutions to build efficient secure environments is strongly required.

Security is the basic need for the user's sensitive and personal data. The big data's enhancement is approaching to provide the secure environment. Although there are several cryptography techniques (that can secure data) available, yet due to existence issues or problems there is need of more work in this field. First contribution of this thesis is that the researcher has proposed a unique big data lifecycle which introduces various security issues and their possible solutions at each phase. The researcher is sure that the thesis will provide better understanding to the two main big data issues including misuse of personal data and unauthorized access. This study explores the relation between privacy and security of big data. The objective the research has been to collect knowledge on how adoption of big data affects the security and privacy in this connected world as well as their solution. In the context of big data, sensitive information includes data from an extensive range of various domains and areas. Data related to health or even basic information, both can be the example of sensitive information and it is clear that most of the users do not want to disclose their sensitive information. Therefore, in recent days, with the growth of big data, data privacy and security requirements are in the air to secure users against monitoring and data disclosure.

This thesis focuses on the design and development of methodologies for handling sensitive data appropriately in the area of big data. The idea behind the proposed solutions is enforcing the privacy requirements mandated by existing legislation that aims to secure user's privacy in big data environment. The researcher begins with a description of background material in addition to reviewing existing security and privacy solutions that are being used in the area of big data. It then continued to develop an improved lifecycle for big data and phase wise security threats, followed by identifying the problems on very phase of big data life cycle that are essential to be solved.

This lifecycle also addresses the security attacks on the data creation phase as well as their remedies. This is the complete lifecycle. It has five phases: data creation, data collection, data mining, data analytics and decision making phase. Every phase has its own attacks and data is travelling from one phase to another phase. To provide security at every phase is a difficult task. The researcher has tried to secure the two initial phases i.e. data creation and data collection phase. It is observed that if these two phases can be secured less effort will be required to secure further phases. Every phase is minimizing the flow of vulnerable information throughout the lifecycle.

Further, the research aims to tackle the privacy issue. This issue appears when user's personal information is shared or sold to the unauthorized party or to the third party. Therefore, the organizations collecting user's data should provide the customized privacy policies for the user. Hence, the researcher has proposed the privacy policies to address privacy and security of user's data with the aim to minimize security risks and privacy as the second contribution. In the recent years, security of user's data concerns are largely rooted in the rapidly expanding big data ecosystem conceptualized as the privileges of persons whose data is shared with others. Information security and privacy of private data have long been viewed as primary human rights. It is expected that these policies will secure as well as aware individuals about their security. These state that while asking for personal data, organizations must provide a choice to the user where they wish to share their data or not. Policies are clear and concise that clearly explains expectation from users. In the same row confusion matrix is developed to calculate the accuracy of classification. This approach has been implemented in Hadoop 2.5.2 (pre-built i386-Linux native and java version SDK Oracle Java 1.8.0 25 (8u25).

The third main contribution is a novel approach for securing Sensitive Health Information (SHI) in big data heterogeneous environments. Health data is very sensitive to any patient (data owner). The goal has been design an approach to prevent the SHI from unauthorized access. With the help of this scheme, an unauthorized user cannot access the patient's SHI without user's consent. A novel Information Leakage Prevention Scheme (ILPS) has been developed for SHI. The proposed approach aims to secure patient's SHI by improving RSA encryption technique. This approach works on data collection phase. It achieves better security in comparison with existing methods available in the area. The results of the

experiments reflect improvement in the existing RSA. Comparative analysis shows that the proposed approach is taking less time in comparison to the other existing approaches. According to the statistical results the proposed approach is acceptable. This proposed approach has been implemented in Hadoop 2.5.2 (pre-built i386-Linux native and java version SDK Oracle Java 1.8.0 25 (8u25) with the system specifications CPU Intel i7-7700, 4.2 GHz CPU and Microsoft Windows 7 as the operating system.

ABBREVIATION

- ICT- Information and Communication Technology
- HDFS- Hadoop Distributed File System
- PII- Personal Identifiable Information
- SNS- Social Networking Sites
- CS- Commercial Sites
- HER- Health Electronic Records
- RSA- Rivest- Shamir- Adleman
- ABE- Attribute-based Encryption
- OPE- Order-Preserving Encryption
- ECC- Elliptic Curve Cryptography
- AES- Advanced Encryption Standard
- ILPS- Information Leakage Prevention Scheme
- DDES- Dynamic Data Encryption Strategy
- NoSQL- Not Only SQL
- RDBMS- Relational Database Management System
- SEM- Structural Equation Modelling
- TTP- Trusted Third Party
- CMD-Computing on Masked Data
- TSHC-Trusted Scheme for Hadoop Cluster
- HDD-Hard Disk Drive
- SDD- Solid State Drive
- CVA- Cross Validation Accuracy

TABLE OF CONTENTS

LIST OF THE TABLES

LIST OF THE FIGURES

CHAPTER 1: INTRODUCTION

"Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."

- ***Westin***

1.1 Background

Dependency of human life on computers is continuously increasing. Everything is connected through computer and internet. Internet is like oxygen for computer. The interconnection of computers through internet has cut down the limitations of geographical boundaries all over the world. Each and everything is handled through internet. Due to the low cost availability of internet data is growing exponentially and this increasing data is called big data. It is big data era now and is going to set the present and future research boundaries. The amount of data captured by the organizations and the constant growth such as the rise of social media, Internet of things (IoT), and multimedia, has generated a huge flow of data in a structured and unstructured format [1]. Big data is characterized by three perspectives: (a) data is huge, (b) data cannot be classified into traditional databases, and (c) data is created, captured, and processed fast [2].

In addition, big data is changing engineering, healthcare, science, finance, business, and ultimately the society [1]. The intensive applications for processing and managing big data need security and privacy of the information and asset. Everywhere security intensive information is floating; anyone having malicious intent can misuse the information. This may cause harm to an organization or individual. Organizations are a rich source of user's personal information and it gives opportunity to financially

motivated cyber attackers etc. For example, once a user applies for rental properties, then users have to give photocopied documents such as personal identifiable information (PII): legal name, date of birth, and passport etc., economic information – salary slip and other information, contact information such as email addresses, applicant's address etc. [3].

Cyber attackers do not create security holes in the organization's boundaries. They just find the loop holes present in the organization's boundaries. Now days, sharing information poses a peril to a user's organizational secrecy. The dilemma of information use and privacy protection becomes more acute these days. According to a study, large amount of personal data is freely available to be gathered and analysed. In such technical surroundings, after performing a business transaction or reaching out to people in highly connected societies, physical distance is no longer a hindrance. In June 2011, according to the Australian Bureau of Statistics, the number of internet subscribers was 10.9 million [3]. Software industries are facing various privacy issues with the increase in the use and demands of big data. Privacy of an individual has become a major problem and providing safety has become very critical [4].

Due to the enhanced dependence on technology and its pervasive interconnectivity with the infrastructure of Information and Communication Technology (ICT), it unfortunately, changes the mind of the attackers to operate their plans. Flaws in the ICT infrastructure are fruitful lands for attacker's to get unauthorized access, denial-of-service attacks, and information theft etc. Attackers may persistently explore new fields and possibilities of control and exploit weaknesses [3]. For example in 2006 a privacy breach released, 20 million search queries of users. To facilitate research on information retrieval, these queries were posted by users in a three month period. On behalf of this information, two New York Times reporters were able to discover the identity of user number 4417749, based on search history [5]. So there is a demand to propose such approaches that not only support the collection of huge size of data, but also efficiently manage or operate enormous data requests with minimum time and maximum security [5, 6].

Protecting the big data is an important question that needs to be answered for a particular data whether it is in the public or private category [6]. Finding out the weakness has become vital requirement. Even a single flaw in security of user's

information can completely destroy the privacy of any user or organization. Securing user's data from attacks is not just applying patches every time weaknesses are discovered. Weaknesses, if not uncovered and mitigated during initial phase, can incur gigantic loses in terms of money, time and efforts after implementation. Integrating phase's wise security, according to the big data life cycle has proven to be the most effective way to provide secure environment. Developing, modifying and maintaining the big data environment is difficult. Applying the privacy policies with encryption approach can secure the whole system at a higher level. Big data life cycle has five phases: data creation phase, data collection phase, data mining, data analytics and decision maker phase. With the help of this cycle, any privacy and security issues can be fixed at data creation and data collection phase. Also, people can easily understand the procedure and will be aware of their information. So, there is an urgent need to develop privacy and security approach for big data.

1.2 Big Data

Big data is a popular phenomenon that intends to solve the issues present in traditional solutions based on databases and data analysis. It is not about storage or access to data only. Big data refers to data sets that are terabytes to exabytes in size. Graphical representation of big data has been shown in figure 1.1. The enormous sizes of these datasets expand beyond the capacity of average database software tools to store, capture, manage, and analyse them efficiently [7]. Though it is pervasive, however, as a concept big data is newly born and has uncertain origins. Diebold argues that the word "big data" probably in the mid-1990s Silicon Graphics Inc. (SGI), was initiated during lunch-table conversations, in which John Mashey figured prominently. In this context, from the longitudinal studies starting with data collected in 2005 and its expansion by 2020, our analysis continuously shows the expansion, the increasing complex, and ever more interesting digital universe. The sixth annual study of the IDC for the digital universe is full of new findings [8]:

The digital universe will rise by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes from 2005 to 2020. The digital universes will double every two years from now until 2020. The investment in spending on IT software, hardware, telecommunications and employees which can be considered the "infrastructure" of digital universe and broadcast communications will increase by 40% between 2012 and

2020. The quantity of previously data about people like — downloading music, taking pictures, writing documents, etc. — is much less than the quantity of data is being made about them in the digital universe [8]. Big data definitions have developed fast, by various researchers and practitioners on their own way which has raised several confusions [9]. According to Tech America Foundation's Federal Big Data Commission,

Big data is a word that explains huge amount of high velocity, unpredictable and variable data which needs advanced techniques to store, capture, distribute, and analyse the data.

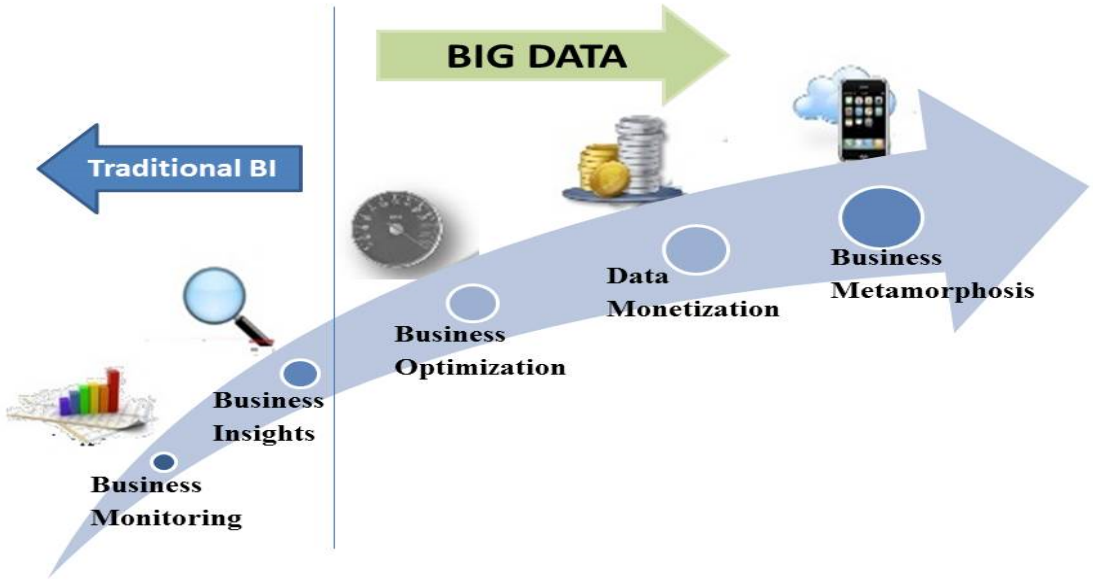

Figure 1.1: Evaluation of Big Data

There is a set of 'big data' definitions that expose diverse aspects of the concept. For example, while some focus more on different aspects of data sources, other writers stress on storage and analysis needs while dealing with 'big data'. IDC categorized three main features of 'big data': the presentation of data itself, data analysis, and the analysis results which allow business value creation in context of new services or products. Eventually, Boyd and Crawford (2012) proposed another definition of 'big data' which includes:

Technology (for example, storage, computation power), analysis (for example, patterns recognition for financial, social, legal claims, and technical), and

mythology (for example, the widespread conviction is that 'big data' provides a high level of valuable insights) [10].

In recent times, various claims and definitions related to "Big Data" have been extended because of the concept [11]. In the last few years, the National Institute of Standards (NIST) has created a big data working group. It is a community of industry, academia and government whose goals are to increase taxonomies, consensus definitions, secure reference architecture, and an innovation guide [12]. This group has described 'big data' as diverse datasets including semi-structured, structured, and unstructured data with variety, volume, and velocity [1]. Figure 1.2 shows the four main parts of the big data architecture: data sources, data storage, data transformations processing, and analysis [1].

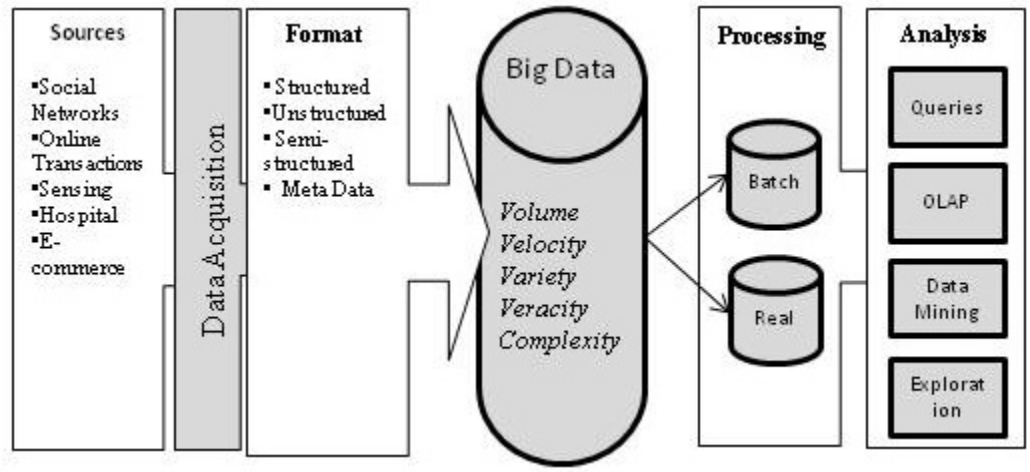

Figure 1.2: Architecture of Big Data

Furthermore, there are supporting subsystems to ensure big data security and big data management. These subsystems give services to other parts of the big data architecture. The data source part of the architecture needs that big data to be served for a specific purpose that can change in various ways. When various sets of big data are first gathered, the datasets with homogeneous source structures are combined. Then metadata is made to facilitate lookup techniques for the collective data [13]. Metadata with dissimilar datasets is also collected into a huge collection for matching goals such as, applying security policies after correlating the data with identifiers. Data mining can be used for analysing the collected data from various perspectives or to extract the particular information from the data [13].

1.3 Big Data Characteristics

Regardless of the wide use, big data has no universally and strictly accepted definition [14]. Adding to Gartner's three Vs.: Velocity, Volume, and Variety, the software company, SAS, has suggested two other big data dimensions: variability and complexity. For example, to create highly customized offerings (e.g., "pregnancy prediction score"), an organization may need to analyse large amount (volume) of unstructured and structured data (variety) from various sources. In other cases, this process can also include the use of high velocity data [15]. Figure 1.3 shows the characteristics of big data. One of the major considerations for potential security issues is related with outsourcing. According to a report, in 2012, 64% of security breaches included outsourcing providers. The information stored in the servers is a rich collection for cyber criminals. But the organization's responsibility for securing data using reputational and regulatory approaches doesn't end here [16]. To add more insight the big data characteristics has been explored in details:

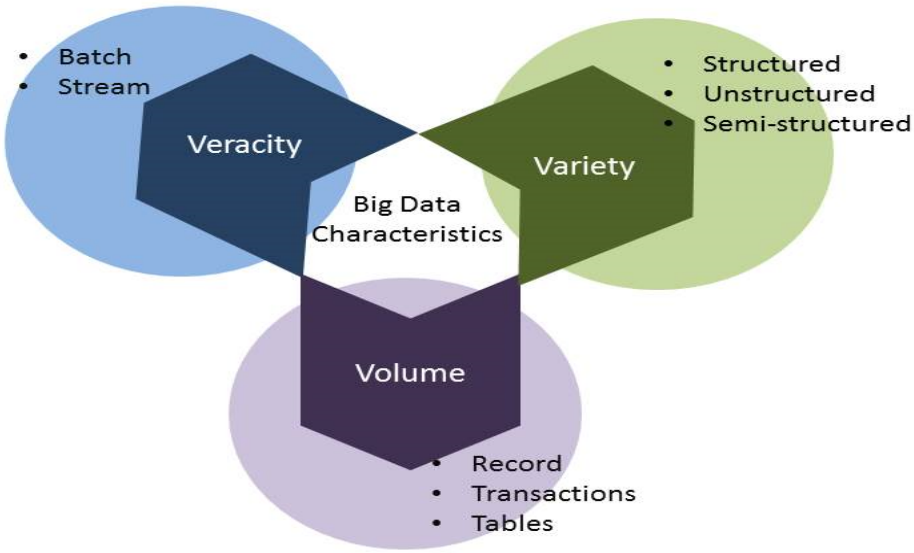

Figure 1.3: 3Vs of Big Data

'Volume' refers to the large amount of data. Sizes of big data are reported in numerous terabytes and petabytes. In 2012, a survey conducted by IBM showed that more than 1144 respondents considered datasets to be bigger than terabyte to be big data. One terabyte is enough to store around 16 million Facebook pictures. In Facebook, one million pictures are processed per second. A petabyte is equal to 1024 terabytes. Earlier estimates proposed that Facebook could store 260 billion pictures using storage space of

more than 20 petabytes [9]. The new age of social big data has stated. The amount of social media data is vast and this amount of data has turned into the main provider to consistently extending digital universe [17]. The advantage of collecting huge amounts of data includes the formation of patterns and hidden information through data analysis. Whang et al. presented a distinctive collection of location based data from smart mobile devices and this collection is available to the researchers. The above scheme is called Nokia's mobile data challenge [18].

'Variety' means various formats of data for example structured, e-mail, numeric data in traditional database and video, unstructured text documents, audio, financial transactions etc. [16]. However users on Internet also generate an enormously different set of unstructured and structured data. Online social networks have brought various complexities to many companies because of new data formats that are semi-structured, unstructured and traditional like text, video, audio, data logs, and images, etc. [1]. Unstructured data is created through mobile phone and sensors for example, text messages, blogs etc. This unstructured data does not follow any particular format. It has become a challenge. From the perspective of data size, unstructured data is becoming dominant in the whole information available within an organization. Examples of unstructured data include: technical documentation of equipment, images taken by infrared thermal imagers etc. [17].

The semi-structured data are primarily text based and conform to particular rules. Within semi-structured data, some other types of tags or markers are used to recognize certain elements. Examples of semi-structured data include emails, log files and XML files with specified formats. On the contrary, semi-structured and unstructured data are moderately difficult to reduce, classify and analyse using traditional techniques [19]. Companies are collecting shapeless data from external sources (such as social media) and internal sources (such as sensor data). Though, the rise of new data management technologies and analytics is an innovative factor which enables associations to take advantage of data for their business processes. For example, facial recognition techniques make brick-and-mortar retailers empowered to obtain intelligence about store traffic, the age or gender of their clients [9].

'Velocity' refers to the data generation rate and the speed at which it should be examined and acted up on. It refers to the speed of data creation, transfer and retrieval

[17]. The creation of digital devices like sensors and smart phones has given rise to the unprecedented rate of data creation and is running a growing requirement for real-time analysis. For instance, Wal-Mart processes more than one million transactions every hour. Generated data from cell phones and through mobile applications can be used to produce real-time, personalized offers for daily customers. [9]. Today, social networking sites generate huge amount of information. Though the amount of data created and shared is relatively small in these sites, but user frequency is high which results in huge data collection [1]. Relativity of big data volumes described before applies to every dimension. Hence, universal benchmarks do not exist for variety, volume, and velocity which describe big data. The defined range depends on the size, area, and place of the firm on the limitations developed over time. It is also important that these dimensions are dependent on each other. However, a 'three-V tipping point' exists for each organization beyond that conventional data management and the analysis techniques become unable to gain intelligence on time. The three-V tipping point is the threshold ahead of which the organization starts dealing with big data [9].

1.4 Security and Privacy Issues in Big Data

Big Data imposes a change from the conventional methods in three unique ways: volume (huge amount of data), velocity (speed of data), and variety (different types of data) [20]. These characteristics are known as the three V's of Big Data. Various researchers have added new characteristics to the initial group, like veracity, variability, or value [21]. The latest progresses in commoditization of mobile technologies, hardware, cloud computing and large-scale networks, have increased the ability to process huge amounts in addition to accelerating the data creation and collection [22]. Increasing mobility and involvement in social media, desire to share more data, new techniques that captures additional data about the data, and the growing business around big data have imposed result — the need for data security [8]. When it comes to information security, Big Data requires a matter of special concern. The lack of standards among e-commerce sites, the openness of users, the phishers' sophistication, and the firmness of the attackers contain considerable personal information on risk. For instance, what a retailer can keep personal about your purchase, like your transaction and user profile data, but other organizations cannot and so data can't be hidden. Still

interconnecting these data sets with different data sets can open up vast security loop holes and can make the private information public [8].

- In September 2018, Facebook confessed that 80 million user's accounts were affected by the hackers. User's private data could be seen by anyone [23].
- Consumer credit score company Equifax had disclosed that hackers accessed up to 143 million customer account details in July, 2017. The details hacked includes names, social security numbers, drivers licences, and credit card numbers of around 200,000 people [24].
- In July 2017, a Verizon company's had told that more than 14 million user personal data has been disclosed. This personal data hacked includes names, contact numbers, PINs etc. [24].
- In November 2017, Uber disclosed that more than 57 million drivers and rider's personal information has been stolen by the hackers. It's one of the mega data breaches that happened in October 2016. The hacked data contained names, phone numbers, email addresses of riders and drivers [24].
- In 7 February 2017, IHG, the company announced that 12 hotels including Holiday Inn, Crowne Plaza, InterContinental etc. properties has been affected by a malware. The Malware has been found on servers which processed on the payments made [25].
- In 19 March 2017, Buzz Feed disclosed the news that more than 10,000 customer information has been available by a link on a website. The available information includes phone numbers; email addresses, product codes etc. [25].
- In 2018 June, 19.5 million records of users has been breached. An attacker has hacked two databases. The hacker demanded a ransom amount in exchange to gain the access on data [23].

1.4.1 Data Privacy Issues

Data privacy is probably the topic about which ordinary people are most concerned, but it should also be one of the greatest concerns for the organisations that use Big Data techniques. A Big Data system usually contains an enormous amount of personal information that organisations use in order to obtain benefit from the data.

Organisations should not have total freedom to use that information without our knowledge, although they also need to gain some benefit from the use of that data [18]. Several techniques and mechanisms with which to protect the privacy of the data, and also allow companies to still make a profit from it, have therefore been developed, and there have been attempts to solve this problem in various different ways [26]. Social media applications help the societies to come closer to each other, hence offering the real-time security and privacy, for voluminous amounts of data generated by social media are key concerns. To protect the social big data, new security and privacy mechanisms are required. Current security technologies are based on static data and not sufficient to protect the data which is changing dynamically. Traditional security and privacy technologies do not completely consider important characteristics of a large amount of data, such as data pattern, and variation of data [27]. Thus, it has become a challenging task to design and implement new privacy and security mechanism in this complex circumstance. In the context of ensuring security and privacy, several research efforts have been carried out. These research efforts can act as guidelines for the new researchers to let them know about the current research for strengthening the security in the cloud as most of the social media service provider use cloud to store the data [28-30].

There are various methods for data privacy, and these are from technical encryption and anonymisation solutions for design, access and rights management solutions. In case of organizational micro-data, many approaches are used to ensure confidentiality such as de-identify data. Synthetic data is created to mimic few features of the original data and Trusted Third Party (TTP) mechanisms is used to reduce the perils of the exposure of a person's identity or data loss [31]. Since user's delicate data gets disclosed to a number of clients including employers, health care providers, social sites etc. So, an attacker can "connect the dots" and can slice up the user's information, leading to even more privacy loss. The more complete the integrated information, the more privacy will be compromised [8] [20].

1.4.2 Privacy Issues in Social Networks

Social networks are all around us. The popularity of social networks is currently huge, and almost everybody with access to the Internet has at least one account with them. People share a lot of personal information on these networks without actually worrying

about what the organisation behind them will do with their data. This data, along with the strong analysis capability of Big Data, is a huge threat to personal privacy. Addressing this problem is not an easy task, and some authors suggest new legislation to increase the protection of data privacy [32]. Another work, meanwhile, proposes a technique that can be used to increase the control that users have over their own data in social networks [33].

1.4.3 Confidentiality Issue

In addition to those attempting to steal sensitive information or damage user data, storage service providers are also assumed to be untrustworthy third parties. Data transmissions among tiers in a storage system provide clues that enable the service provider to correlate user activities and data set. Without being able to break the cipher, certain properties can be revealed [26]. Although privacy is traditionally treated as a part of confidentiality, decide to change the order owing to the tremendous impact that privacy has on the general public's perception of Big Data technology [34]. The authors that approach this problem often propose new techniques such as Computing on Masked Data (CMD), which improves data confidentiality and integrity by allowing direct computations to be made on masked data, or new schemes, such as Trusted Scheme for Hadoop Cluster (TSHC) which creates a new architecture framework for Hadoop in order to improve the confidentiality and security of the data. It protects data from unauthorised alterations during its lifecycle [30].

1.4.4 Integrity Issue

Integrity has traditionally been defined as the maintenance of the consistency, accuracy, and trustworthiness of data [35]. It protects data from unauthorised alterations during its lifecycle. Integrity is considered to be one of the three basic dimensions of security. Ensuring integrity is critical in a big data environment, and authors agree on the difficulty of achieving the integrity of data when attempting to manage this problem. In a heterogeneous environment, information is stored in numerous nodes with replicas for fast retrieval. But if any copy or information is deleted or modified from another node by a hacker, then it will be challenging to recover data [36, 12].

1.4.5 Availability Issue

Researchers have also dealt with the subject of availability in Big Data systems. One of the main characteristics of big data environments, and by extension of a Hadoop implementation, is the availability attained by the use of hundreds of computers in which the data is not only stored, but is also replicated along the cluster. Finding an architecture that will ensure the full availability of the system is, therefore, a priority. Wang et.al, have been proposed a solution with which to achieve high availability by having multiple active Name Nodes at the same time [37]. Other solutions are based on creating a new infrastructure of the storage system so as to improve availability and fault tolerance [38, 39].

1.4.6 Record linkage Attacks

Hacker is able to recognize the item of a target user by linking the item to data from different sources, for e.g. liking the item to an item in a published data table. It is a difficult task because single entity identifiers are not accessible in every database that is connected. Thus, the common attributes available which are sufficiently well correlated with entities, known as quasi-identifiers (QI), have to be used for the linkage [11] [40].

1.4.7 Degree Attacks

Hacker re-identifies the nodes that belong to a target individual from a series of available graphs by comparing the degree of the nodes with the degree advancement of a target. The ability of this attack is that the attacker can actively manipulate the degree of the target individual by interacting with the social network [11]. Structural attack uses the additive degree of n-hop neighbours of a vertex because of the regional feature, and joins it with the imitation annealing-based graph matching methodology to re-identify vertices in anonymous social graphs [41].

1.4.8 Network level Attacks

Network level issues are related to network security and network protocols, such as Internodes communication, distributed nodes, distributed data. Many nodes are present in clusters and on those nodes, processing or computation of data are done. In a cluster, this processing of data can happen anywhere among the nodes. So it is challenging to discover the node, where data processing is happening. Because of this problem, providing security to processing node becomes complicated [35].

1.4.9 Authentication Level Attacks

User authentication level issues deal with encryption/decryption methods, authentication techniques such as administrative rights for applications' nodes authentication and nodes, and logging. Several nodes are present in a cluster. Each node has a different priorities or rights. If malicious nodes get the administrative rights, then it will steal or modify the confidential data. In case of no authentication, affected node can destroy the cluster. In big data, Logging plays an essential job. If logging is not given, no action is recorded. If a new node enters in the cluster then that will not be identified because of logging absence [35-36].

1.4.10 Generic Type Issues

In distributed environment, various technologies are used for data processing, as well as some traditional security tools for providing security features. Traditional security tools have been developed over the years. So these tools may not be compatible with new distributed structure of big data. As big data uses several technologies for data retrieval, storing, and processing, certain complexities may occur because of the wide use of different technologies [35, 36].

1.5 Need for Big Data Security

The quantity of information in the world is large and increasing exponentially. For example, multiple social networking sites like Twitter and Facebook produce terabytes amount of data per day such as pictures, gif, videos, posts, etc., and will generate huge amount of data in the future. Therefore, it is clear that the volume, variety, and velocity of big data not only increases privacy and security challenges, because they are generally addressed in conventional security management, but also generates new ones which should be dealt in a special way [20] [42]. Today the amount of data is unprecedented and cannot be analysed with traditional techniques. Still, being capable to efficiently "make sense" out of big data is becoming more important than ever before in different areas. In the health area, health devices produce large amounts of data which reflect the patients' status by monitoring their heart rate, sleep, and other medical conditions. In finance, the stock market creates huge amount of transaction. Though many organizations use big data for collecting non-personal information, there are others that use it "in ways that implicate individual privacy," that the various type of

data collected can disclose an individual's health issues, purchasing habits, browsing history, social, financial data, religious and political preferences etc. [43].

The data can help companies to maximize profits [44]. In homeland security, the U.S. government collects more terabytes data every day than the quantity of text in the Library of Congress. This data can be analysed to identify the potential risks to the country. It requires addressing a variety of privacy and security issues as well [45]. One of the main issues in data analytics is to gather data from various sources and add them together so that data analysers can extract information use efficiently in an integrated manner. When collaborate the multiple data, a basic issue is identifying that which portion of information define the real-world entity [43]. Another aspect of data integration is the threat on individual's personal information, as it is now more open to the public. For instance, life insurers are exploring ways to predict the life span of their customers by linking personal health information together on the Web [17].

One of the main issues for big data is the security of a user's privacy; for example systems frequently contain a rich amount of PII. Thus, a main question arises; what access policy and privacy-preservation approach should be enforced to assure a suitable security? In case of big data removing, PII is not easy, when the data is unstructured. To do this, one has to first classify the sensitive information contained in big data and then carefully remove sensitive information to ensure compatibility [20]. Another potential violation of security is the loss of control which may happen where users' data and applications are held at the server's locations. The users don't have any control over their data, which can lead to various security challenges because it makes possible for cloud providers to perform data mining task on the users' data. Furthermore, when the cloud providers backup the data on various data centres, the users are not aware that the data is totally removed ubiquitously when they erase their data. There are various ways to misuse the undeleted data [22].

In big data, the security problems cause many privacy concerns, because privacy is a delicate topic which has different interpretations based on communities, contexts, and cultures. Privacy is different from person to person. Moreover, security and privacy both are two separate topics. Security is usually essential to provide privacy [18, 26]. Right to privacy has been recognized by the United Nations as the fundamental human right [46]. In order to provide a better understanding of privacy, various efforts have

been made to conceptualize the privacy by philosophers, jurists, psychologists, sociologists and researchers. For instance, in 1960 Alan Westin's research was considered to be the first important effort on the problem of user's data protection and privacy [47].

Westin to "Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is to be communicated to others [47].

1.6 Hadoop

Hadoop was created in 2005 by Goug Cutting and Mike Cafarella to support a distributed search Engine Project. It is an open source framework that helps to store, access and gain huge resources from large data in a distributed fashion at low cost, high scalability and high degree of fault tolerance [48]. It handles huge amount of data from several sources such as audios, Images, sensor data, videos, Communication data etc. Hadoop consists of two parts i.e. HDFS (Hadoop Distributed File system) and MapReduce [49]. HDFS provides storage of data and MapReduce provides analysis of data in clustered environment. MapReduce model supports data-intensive distributed applications. HDFS is a distributed, scalable, and portable file system written in Java [50].

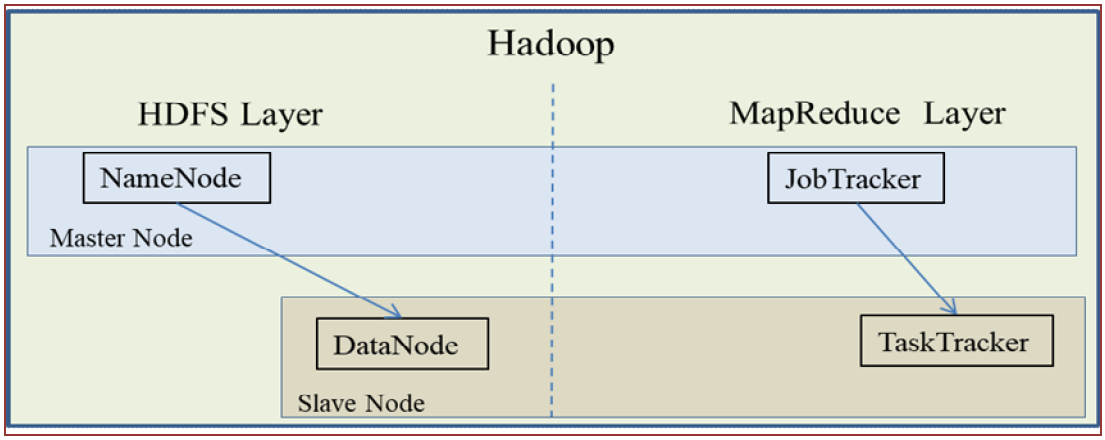

Figure 1.4: Hadoop Architecture

The Architecture of Hadoop is represented in figure 1.4. In the next 5 years, approximately 50 per cent of big data projects are expected to run on Hadoop. Using Hadoop financial organizations started to store their sensitive data on Hadoop clusters. Thus, there is a requirement for a highly strong authentication and authorization

mechanism to secure the sensitive data and also there is a need for highly protected authentication system to restrict the access of sensitive business data stored in an open framework e.g. Hadoop [49].

1.6.1 Security in Hadoop

Primarily, Hadoop had no safety framework and it is believed that the cluster, the client and the environment were fully trusted. Even if it had various authorization controls such as permission to access file, any malicious user may without difficulty impersonate a trusted user as the authentication was depending on the password. Later, Hadoop cluster moved to private networks, where clients have equal rights to access the data stored in the cluster [51, 49]. Due to less security concerns, many government organisations never use Hadoop environment to store valuable data. They are providing security outside of Hadoop Environment such as firewall and Intrusion Detection System [48]. Hadoop doesn't follow any classic interaction model as the file system is partitioned and the data remains in the clusters at diverse points. One of the two situations may happen: the job runs on other node different from the node where the client is authenticated or a different set of jobs may run on the same node. Security breach in Hadoop may be as follows [50]:

a. Unauthorized client may access the HDFS file
b. Unauthorized client may read or write the data block
c. Unauthorized client may submit a job, alter the priority, or remove the job in the queue.
d. A running task may access the data of another task through the operating system interface, few of the possible solutions may be -:
    - Access control at the file system level.
    - Access control checks from start to read and write
    - Protected way of user authentication
e. Replay attack – The attackers copy the stream of communications between the two parties and it reproduces the same to one or more parties.
f. Stolen verifier attack – The stolen verifier attack occur when the attacker break the password verifier from the server and makes himself a genuine client.

1.7 Research Issues

In this case, the questions include the investigation of the main challenges and problems that can be found with respect to the topic of big data security, along with another question whose objective is to discover the main security dimensions on which researchers are focusing their efforts. Finally, we wished to discover what different techniques, methodologies or models have already been developed in order to deal with these problems. Thus, reviewing the literature about big data and associated security, the following research questions arise which needs to be addressed:

- What are the major challenges with respect to Big Data security?
- What is the current status of big data security?
- Is there any security mechanism available to secure big data at the time of data creation?
- Does the information leakage problem in big data have been addressed previously?
- Can privacy policy be imposed on the organization collecting user's data?

1.8 Research Problem

From the foregoing discussion, it is pertinent that security is a big challenge in the current digital era where insecurity is everywhere. The user's personal and sensitive data may be collected, evaluated and reused by the other third party users for several purposes. A major problem with user's personal data is that his sensitive data is shared with unauthorised party and he is not even aware about it. Privacy and security of personal data have long been declared as primary human rights. For adequate secure environment, there must be a secure approach that can prevent from unauthorised and data leakage issues. Keeping this in mind, the researcher has formulated a problem as under in order to improve the privacy and security of big data.

"Inception of Data Creation Phase and Prevention of Data Leakage in Big Data Life Cycle"

1.9 Objectives of the Research

In order to achieve the most general goal to improve the security of big data techniques, the following objectives have been set forth:

- To review and critically examine the big data security and to identify key issues that needs to be addressed in the real-world deployment of this technology.
- To study the alternatives to the present state-of-the-art approaches for big data, and to identify the ones that offer practical advantages,
- To examine the impact of threats and attacks on big data.
- To appreciate the need, importance and significance of data creation phase in big data security life cycle.
- To study and analyse the effect of privacy paradox on user's data.
- To design the privacy policies following which the user's personal information becomes more secure.
- To conduct a detailed study on encryption techniques.
- To design an algorithm to prevent the data leakage problem in big data.
- To evaluate the performance of the proposed system.
- To compare the performance of the proposed information leakage prevention scheme with the other existing approaches in the area.
- To validate the proposed system.

1.10 Methodology Followed

The basic methodology is tantamount to a list of things that we might try in order to reach out the ultimate goal.

- Review of the available literature
- Conceptualization of the approaches
- Expert-Review and Revision of the proposed approaches
- Implementation of the proposed approach
- Experimentation
- Assessment of Effectiveness
- Documentation and Finalization

1.11 Expected Deliverables

The following are the expected outcomes of the research:

- A checklist to identify the security and privacy threats in the big data that can be addressed at data creation and data collection phase.
- Design and development of privacy policies and their implementation in Hadoop.
- Design and development of a novel approach for Data Leakage Prevention at data collection phase.
- An experimental study showing the usefulness of the proposed work.
- A comparative study to prove that the proposed approaches are better one.
- A study carrying out the validation of the proposed approaches.

1.12 Limitations

In order to keep the research precise and within the time boundary, the thesis has few limitations. These are as follows:

- The proposed work has been tested on the small data only.
- Implementation of the proposed work has not been done in real scenario.
- The researcher has implemented this approach on text data only.

1.13 Thesis Outline

A thesis of the research has been prepared to present a detailed study on the research problem and to answer the research questions. The thesis has six chapters apart from annexure, references and other components. The chapter summary shows in figure 1.5. The chapter wise summary of research thesis is presented below.

Chapter- 2: Literature Review

This chapter provides various available definitions and discussion about big data security. It presents the literature review, basic terminology of big data and privacy preserving techniques in details. It also presents the general overview of Hadoop, and consequently introduces the literature review on security o big data.

Chapter- 3: Development of Improved Security Threat Model for Big Data Life Cycle

In this chapter a new lifecycle for big data has been proposed which explains security threats phase wise. This lifecycle has been unique in itself. After that the researcher has

discussed the need of data creation phase and explained the role of users in their privacy disclosure. To strengthen the need and importance of new phase, the researcher has collected 165 key cases of privacy violations from numerous sources.

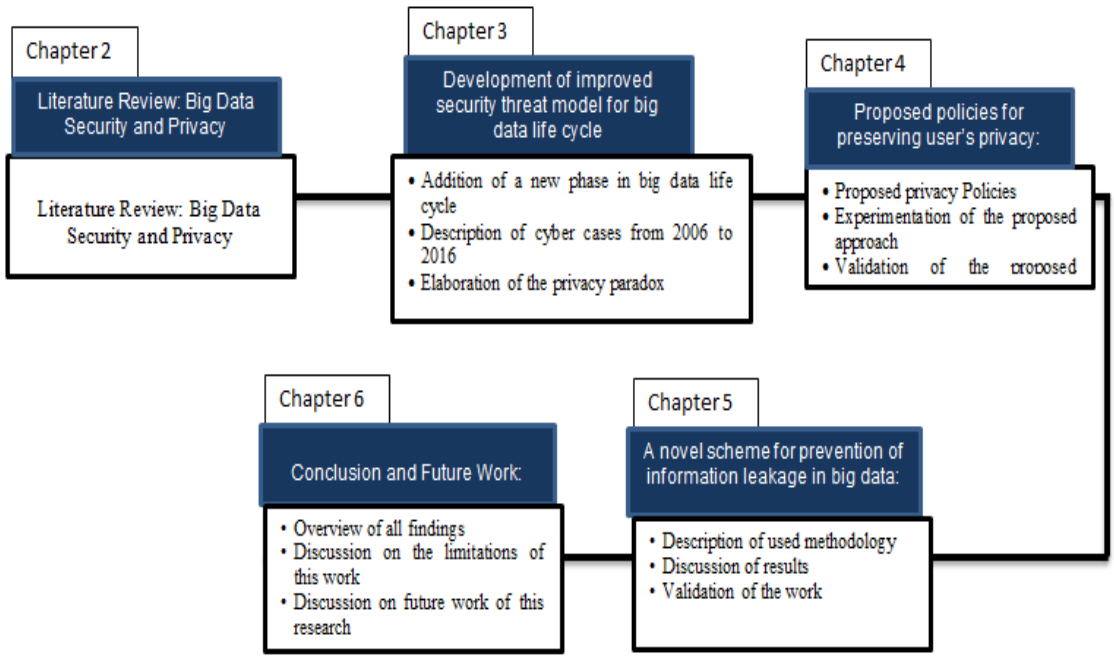

Figure 1.5: Summary of Chapters

Chapter- 4: Proposed Policies for Preserving User's Privacy

In this chapter, the researcher has proposed big data privacy policies. The idea is dependent upon the user's choice on whether or not to disclose their information to third party. These policies have been implemented in Hadoop. This tool is suitable to handle big data. The experimental results are also presented.

Chapter- 5: A Novel Scheme for Prevention of Information Leakage in Big Data

In chapter 3 and chapter 4, the researcher tries to improve security of data creation phase. In this chapter researcher provides security on data collection phase. Proposed approach specially deals with the health data. This introduces a novel scheme for prevention of information leakage in big data. The experimental results are also presented. Experimental results are presented graphically. The approach is compared with the available approaches in the area and found that the proposed one is better in terms of encryption time, decryption time and throughput. In this chapter, concept of

validation has been discussed and methodology for validation has been designed. During validation, hypotheses have been formulated and have been tested on the basis of statistical analysis. Student t-test has been used for testing the hypothesis.

Chapter- 6: Conclusion and Future Work

This chapter describes the findings of this research. It presents an overview of the research along with its major findings. In addition, it demonstrates the significant contribution of this research in reference to security of big data. It also discusses probable limitations of the research and proposes directions for future research.

CHAPTER 2: LITERATURE REVIEW

"Risk of security cannot be removed completely when spreading the data over the Internet"

- ***Mavridis***

2.1 Background

Advancements in computer and telecommunications technologies with comparatively reduced in costs have led to tremendous growth and data availability, both the unstructured and structured formats. The phenomenon called as big data includes several costs, benefits and issues [52]. Owing to the increased utilization of big data, it is understandable that there has been a high degree of interest on this topic. It is argued that 2011 has marked as the year when big data gained widespread interest [53]. Big data is becoming a key source of firms' competitive advantages and national competitiveness. For instance, McKinsey Global Institute estimated that annual overall economic gains from big data would be US$610 billion in annual productivity and cost savings [46]. At the same time, big data's characteristics are tightly linked to privacy, security and effects on consumer welfare, which have attracted the attention of scholars, businesses and policy makers. For instance, a huge amount of data means security breaches and privacy violations are likely to lead to more severe consequences and losses via reputational damage, legal liability, ethical harms and other issues, which is also referred as an amplified technical impact [52].

In 1980 by Warren & Brandeis, a theoretical vision has defined privacy as the "the right of being alone" and "being free from intrusion". They have discussed their concerns over the effect of portable instantaneous camera in their article, which was a technical innovation during that time in the late 19th century. According to the authors, the

instant photo camera allows persons to infiltrate each other's private space, which might be harmful. The authors defined the right to privacy and argued for this right to be known in legal rights [54]. Indeed, the practices of everyday life and the places in which people live are now augmented, monitored and regulated by dense assemblages of data-enabled and data-producing infrastructures and technologies, such as traffic and building management systems, surveillance and policing systems, government databases, customer management and logistic chains, financial and payment systems, locatives and social media [55]. Given the soaring popularity of smartphones, retailers will soon have to deal with hundreds of thousands of streaming data sources that demand real-time analytics. Traditional data management systems are not capable of handling huge data feeds instantaneously. This is where big data technologies come into play. They enable firms to create real-time intelligence from high volumes of 'perishable' data. Enabling security related information to float over the internet will always be associated with risk elements [56].

In the information technology world, big data is known as one of the next big things [57]. Due of its characteristics and architectural design it imposes various security issues such as centralization of security, redundancy, data segmentation and high availability [58]. While adopting, big data has various benefits but there are also some important obstacles associated with its adoption [59]. Big data introduced new theories including external data warehousing, resource sharing and computation outsourcing. It creates new privacy and security challenges. Thus, the popularity of big data is increasing gradually and extra security concerns have been raised. Generally, Security is associated with the three important aspects; confidentiality, integrity and availability. In big data, these three vital aspects of security apply to the three categories of assets that must be protected: information, software and hardware resources. Additionally, efforts by some researchers present surveys on big data security needs including confidentiality, transparency, integrity, availability, and accountability etc. [11][36][19][60] .Various practitioners and researchers have discussed about big data security challenges and issues. Many of them work on vulnerabilities and threats, identifying attacks and strive to provide recommendations, strategies, countermeasures, frameworks and other security solutions [61][62][63][64][65][66][45].

2.2 Relevant Research in Big Data

Jeremy W. Crampton has been described the contradictions and complications that lie at the heart of geo spatial big data. This was a geo-political assemblage that includes nexus of interests including state, military, legislative, knowledge producers, and corporate world. None of these are different from each other; In fact the combination theory enables them to investigate collectively and to find relationships and flows of causes and effects. He has investigated the intelligence community as a kind of case where people can gain perspective on these problems. While more work is needed. It looks at the three particular complexities of big data which are knowledge, identity and power [67]. Archana R.A et.al have discussed that big data is the extension of data mining. They explained MOBAT technique in this paper and proposed a data mining technique which is able to secure original set of data [12]. Vikram Phaneendra et.al, have demonstrated that in the past days, the data was in a smaller amount, and RDBMS simply operated it. But today it is hard to handle vast data i.e. big data through RDBMS tools. They have pointed out, that big data differs from the other data in five dimensions such as volume, velocity, variety, value, and veracity [68].

Ashwin Machanavajjhala et.al, have discussed the confidentiality and data misuse which may cause harm to data Provider. They have described that distribution of such micro data facilitated advances in public and science policy, helped citizens to learn about their societies, and enabled students to develop data analysis skills [60]. Jonathan Stuart Ward et.al, have surveys on big data definition; two ideas mainly connect big data: data storage and data analysis. This, therefore, raises the question that how big data is dissimilar from conventional data processing techniques. The contribution attempted to assemble the diverse definitions which have expanded the number of degrees of traction to secure a clear and brief definition [69]. Amir Gandomi et.al, have explained the broad definition of big data that captures its distinctive characteristic. The industry has been forced to catch academic presses with rapid development and adoption of big data which has led to lecture for popular outlets. They highlighted the requirement to develop suitable and effective analytical approaches to leverage enormous amounts of heterogeneous unstructured data such as text, video formats, and audio [9]. Some tools of big data have been shown in table 2.1.

Table 2.1: Tools of Big Data

Big Data Tools	Description	Use	Merits	Demerits
Dryad [71]	To improve the parallel and distributed programs and scale up the capability of processing from a small to a large number of nodes	Infrastructure and platform	• Good Programmability • Allows multiple inputs and outputs • High performance • Distributed execution engine	• Unsuitable for iterative and nesting program • Conversion of irregular computing into data flow graph is very difficult.
Hadoop [72]	To perform the processing of data intensive applications	Infrastructure and platform	• Reliability • Independent tasks • Easy programming Model • High scalability	• Restrictive programming Model • Joining of multiple data set that makes it tricky and slow • Hard cluster management • Single master node • Unobvious configuration

				of the nodes
Tableau [73]	To process large amount of datasets	Data visualization, Business analytics,	• Faster and ease of use dashboards	• Lack of predictive capabilities • Risky security • Change management issues
Karmasphere [74]	To perform business analysis	Big Data Workspace	• Rapid pattern discovery • Parallel collaboration	High complexity
Pentaho [74]	To generate reports from a large volume of structured and unstructured data	Business analytics platform	• Robustness • scalability • Easy access to data • Detailed visualization • Seamless integration	• Less advanced analytics as compared to Tableau

Richards and King have recently recognized three contradictions which are at the heart of big data. While they have acknowledged that Big Data can make beneficial outcomes, they argued that the evidence of these benefits has not been balanced with a view of the Big Data limits or undesirable results. First of all, Big Data suffer from ‘‘transparency paradox.’’ i.e., where the operation of big data itself is almost shrouded in commercial privacy [70].

2.3 Brief Description of Hadoop

Hadoop is an open source framework which supports distributed processing of enormous data sets across clusters of commodity servers. It is the core platform for structuring big data, and prepares the same for making it useful for analytical purposes.

It is designed to scale up from one server to number of machines, with a high fault tolerance [72]. It has its own file system, HDFS and MapReduce applications. Bernice Purcell et.al explained that big data is the collection of huge amount of data sets that cannot be managed by traditional techniques. It includes different formats of data. The technique of data storage used for big data contains multiple clustered object based storage and network attached storage (NAS). The Hadoop architecture is used to process structured and unstructured using the map reduce to locate all relevant data. The start of big data has posed challenges as well as opportunities to business [72].

Islam N et.al, have been studied the performance of Hadoop in Solid State Derives (SSD) and identified low bandwidth as obstacle [76]. Ahn et.al have been proposed a performance model by queuing network to simulate the completing time of MapReduce and therefore Hadoop, came up with a cost performance model for HDD and SSDs [77]. Hong et.al, have been explored how to improve a Hadoop MapReduce framework with SSDs in terms of cost performance [78]. Krish et.al, have been proposed a VENU an extension of Hadoop which used SSDs as a cache only for those who are expected to take advantage from the use of SSDs. But this work raises an open question that, about the applications that are fruitful from the SSD's performance characteristics [79].

2.3.1 Hadoop Distributed File System (HDFS)

Hadoop contains a fault tolerant storage system known as HDFS or Hadoop Distributed File System. HDFS is capable of storing large amount of information, without losing the data, thereby surviving the failure of the essential parts of the storage infrastructure. Hadoop makes clusters of machines and organizes the work amid them. If one cluster fails (without losing the data or without interrupting the work) Hadoop continues to operate another cluster, by shifting the work on the remaining machines in the cluster. HDFS handles the cluster storage by dividing the incoming files into pieces, known as "blocks," and redundantly storing every block across the server's pool [75]. It stores files in clear text and regulates the file security via a central server. So in Hadoop, HDFS has no security when the communication occurs between data nodes and clients because data nodes are not encrypted [80]. Yang et.al have been proposed a triple encryption approach which depends upon RSA as public key encryption algorithm and DES as symmetric key encryption scheme, but the complete computation overhead is an obstacle for this hybrid technique. This triple encryption scheme incorporated in

Hadoop. To encrypt the files of HDFS using DES and RSA depends upon the IDEA to encrypt RSA's private key of users. This paper enhanced the security in Hadoop and performance of files encryption and decryption using MapReduce [81].

2.3.2 MapReduce Framework

MapReduce is the processing pillar of Hadoop. The MapReduce allows the operation to be applied on a large data, breaks both the problem and data, and runs it parallel. According to an analyst's point of view, it may occur on various dimensions. For instance, an enormous dataset can break into a smaller section where analytics can be applied [80]. In Hadoop, such operations are written in java as MapReduce jobs [75]. Albert Bifet et.al [82] stated that streaming data analysis in actual time is flattering the greatest and the most disciplined way to acquire valuable knowledge; allowing organizations to respond quickly when problem emerge or identify to improve performance. The techniques used for big data are a storm, cascading, Apache Hadoop, Scribe, Apache big, etc. Yingyi Bu et.al, have been used HaLoop, improved version of Hadoop MapReduce Framework, because MapReduce does not support iterative programs. HaLoop solves the problem, without modifications. It allows iterative programs to be collected from existing Hadoop programs and significant improvement of efficiency. The researchers have presented the HaLoop implementation, design, and evaluation. It is built on the top of Hadoop and extended by a new programming model and various essential optimizations such as a task scheduler for loop aware, caching for competent fix point verification and loop-invariant data caching [83]. Katal Avita, et.al have been explained that big data requires new technologies and architectures to extract valuable information. They have discussed some issues, challenges regarding big data, some tools, and techniques such as Hadoop; map reduces, etc. [84].

2.4 Hadoop and Big Data Security

Hadoop environments may include variety data in classifications and security issues. Data Collection from one environment also raises the security risks such as accidental disclosure and data theft. Gathering and storing data from various sources may cause uncontrollable problems with management and access control as well as access of data and ownership in the Hadoop environment [85]. The prominent security issues for Hadoop Clusters are access control and data security due to Internet based storage of data. Cloud computing uses Hadoop clusters as big data. For storage and business

purposes, operations users have to provide their data to the servers and mostly based on Hadoop clusters. Big data service providers are normally commercial enterprises which can't be fully trusted [84].

In today's environment, where insecurity is everywhere, security has been one of the important issues. Initially, Hadoop was developed without any security. It has no data privacy, no security model, and no authentication of services and users, so anyone could submit arbitrary code to be executed [35]. It is deployed in various organizations; all of them don't require extremely protected deployment. However, only Yahoo! has deployed Secure Hadoop clusters. Several organizations are planning to provide secure environment. Thus, Hadoop needed options for being secure configured through strong authentication. Providing security to Hadoop is arduous because the interactions don't follow the standard client server pattern: the file system is split and distributes, which requires authorization checks at several points [86].

Park S et.al have been proposed an approach to secure the Hadoop architecture by applying encryption and decryption operations on the HDFS. For encryption and decryption of data AES has been used. Experiments on Hadoop indicated that the on encrypted HDFS, MapReduce job generated less than 7% computation overhead [87]. O'Malley et al. have been proposed a security enhancement for Hadoop, which provides strong mutual authentication by using Kerberos. The central server performs access control for stored files on storage servers. Since files are stored in clear text, when storage servers are compromised by an attacker, data confidentiality is broken [88].

Hadoop architecture's base layer is HDFS; it holds different classifications of data and is more sensitive to security issues. It has no appropriate role based access for controlling security problems. Also the risk of data access, theft, and unwanted disclosure takes place when a data embedded in single Hadoop environment. The replicated data is also not secure which needs more security for protecting from breaches and vulnerabilities [48]. Some authors represented that the HDFS in Hadoop environment is prevented with security for avoiding the theft, vulnerabilities only by encrypting the block levels and individual file system in Hadoop Environment. Even though other authors encrypted the block and nodes using encryption technique but no perfect algorithm is mentioned to maintain the security in Hadoop Environment [88, 50]. Relevant research in the area has been summarizes in Table-2.2.

Mikko Siponen et.al, have been discussed about the importance of information security in research. They have proposed five guidelines to increase the practical applicability of field survey research using employees 'deliberate Internet Service Provider (ISP) violations [61]. Waldo Rocha Flores et.al, have been presented an empirical investigation on which behavioural information security governance factors drive the establishment of information security knowledge sharing in organizations. The investigation followed research design method. The result showed the major influenced on the establishment of security information sharing in organization [89]. Kalyani Shirudkar et.al, have been described that the big data performing data operations and computation for huge amounts of data. They found some challenges such as scalable data mining and analytics, computation access control, and secure communication etc. They have used diverse security methods like Type Based keyword search [75].

Edith Ramirez has talked about the security of big data. He explained that big data brings big benefits. In this paper, he explained some privacy challenges that had been discussed such as unauthorized access, data provenance, etc. He gave some case studies like how an attacker can steal user's private information. He has described four steps of big data lifecycle i.e. infrastructure security, data privacy, data management, and reactive security [90]. Sung Hawn Kim has explained two important elements for big data security. First attribute significance in big data is a key element for fetching information. Second, they have defined that it is not possible to secure all attributes of big data [4]. According to IDC [8], there are five levels of security: compliance, privacy, custodial, confidential, and lockdown. If the data will continuously grow with this exponential speed, the expected volume of data would be 40 trillion gigabytes by 2020. Richard Baskerville et.al, have defined the security framework that focuses on managing the proper balance between prevention and response paradigms. They have conducted a comparative case study with a European organization. This study analysed and empirically verifies why and how organizations balance between their prevention and response policy [91].

S. Sicari et.al, have discussed that traditional security techniques cannot be directly applied to the heterogeneous environment. They have emphasized the need of flexible infrastructure to deal with the security threats. In this paper, they have presented some research challenges, identifying open issues, and suggesting some hints for future research [62]. Chanchal Yadav et.al, have been described architecture for big data. They

have presented a review of various algorithms from the year 1994 to 2013. While pointing out the various security issues, like data integrity, the drawback of the hash function, they have listed out various tools for analysing the big data [64].

Table 2.2: Some Security and Privacy Approaches of Big Data

Author	Organization	Approach	Problem identified	Year
Clarkson et al. [103]	IBM Almaden Research Centre	Identity anonymisation	Identity attack	2010
Zhou and Pei [104]	School of Computing Science, Simon Fraser University, Canada	Vertex anonymisation	Anonymising sensitive attributes	2011
Zakerzadeh and Osborn [105]	The University of Western Ontario London	Cluster-based *k*-anonymity Approach	Numerical data anonymisation	2011
Abawajy et al. [106]	School of Information Technology, Deakin University, Burwood 3125, Australia	Four-tier classifier based on random forest	Detection of malware	2014
Zhang et al. [107]	Department of Computer Engineering, Pune	Two phase top down Specialisation	Scalability of large datasets	2014
Bhattacharya et al. [108]	School of Computing and Mathematics, Australia	Hierarchical multi population approach	Search space	2014
Fard and Wang [109]	School of Computing Science, Simon Fraser University, Canada	Link anonymisation	Link privacy	2015

Nawsher Khan et.al, have described that increasing amount of collected data generates various serious challenges and issues like security issues. They provided study surveys and classified the attributes of big data [92]. Colin Tankard has explained some recommends providing better control over big data sets, such as archiving, access control and data release. He has described that big data centralized storage is so sensitive. It creates new security challenges [93].

Joseph Turow et.al, have discussed about internet vulnerability. They have investigated a study where 1500 adults, 79% concurred with the statement 'I am worried about sites containing data about me'. They have explained the extra issues about the industrial use of individual data, including the behaviours tracked by the profiles [94]. Joseph L has discussed about the threats of telehealth. They explained that the use of telehealth solutions without proper security and privacy policies provider will lack trust [95]. Shui Yu has explained about the privacy of big data with the advancement of technology. He in his paper explained that research in privacy just started and field of big data is almost untouched. He addressed the privacy threat before user can execute big data applications and enjoy the benefit. He has presented the mathematical frameworks and models related to the privacy [96].

LEI XU et.al have discussed the privacy preserving data mining. They have explained various data mining techniques to reduce the privacy risk and unwanted disclosure of sensitive information. They have identified four different roles of the user involved in data mining applications [63]. Vassilios S. et al. have been thoroughly investigated and presented five algorithms to hide the sensitive association rules. They have concluded that the proposed algorithms did not find the best solution for all metrics, including: first of all, execution time is required and second, proposed algorithms produced some issues [97]. Shyue-Liang Wang instead of hiding sensitive association rules has proposed approaches for sensitive data. The proposed schemes required small database [98]. Ali Amiri has been proposed heuristic scheme to hide sensitive data, with maximizing the data utility at the cost of computational productivity [99].

2.5 Big Data in Healthcare

The concern of privacy and confidentiality on electronic medical data is not a new concept. Numerous researchers and organizations have been working on systems and mechanisms to provide these two essential needs on health information systems. To

give more perspective about the encryption algorithm's performance, this subsection defines and studies related research work has been done in the field of information security in terms of encryption method [100, 101]. Encryption is a procedure to convert simple text into an unreadable form, so that, delicate information can be secured from malicious attacks. In order to protect the user's personal health records in big data environments, various works deal with the data security issues. A person's privacy is becoming an important issue [102]. Yan et.al, have been proposed an approach to encrypt de-duplicate records using re-encryption method that is stored in cloud-based. To protect the outsourced databases, various encryption approaches are available that indicate the way to hide data frequency and data distribution is essential to design OPE approaches. They have proposed a new OPE model. This model worked against the mischievous attacks, to use message space expansion and nonlinear space split to protect data distribution and frequency [110]. Comparative analysis techniques in the area has been summarizes in table 2.3.

Few authors have focused on the particular privacy and security needs for Health Electronics Records (HER) systems, where they analysed the concepts of HER privacy and security maintenance, such as data access authorization, data confidentiality, user's (patient) consent, information ownership, data, information consistency etc. [111-113]. Bernice S. Elger et.al have discussed some more legal standardization and ethical frameworks are the essential need to implement interoperable atmospheres on e-health systems, to enable the transfer of information among the diverse shareholders [114]. Boyang et.al have been proposed the first method of privacy-preserving that permits open auditing on shared data and shared data is stored on cloud. They have exploited ring signatures to calculate the required verification information to test the shared data's integrity [115]. Ruoyu Wu et.al, have been proposed a framework to bridge a significant difference between HIPAA regulations and healthcare systems. Their framework has supported the analysis of compliance-oriented to ensure that the healthcare system is complied with HIPAA rules [116].

In the healthcare area many studies and researches on the privacy protection of health data have been reported [117]. Various privacy policy models are adopted in health care system for the protection of patient's privacy [118-120]. The combination of information security and cryptography protocols is used to handle privacy and security challenges [121-122]. Narayan S. et.al, have been proposed a techniques that combined

public-key encryption and attribute-based cryptography with keyword search to provide privacy of the electronic health record management system. They have extended their proposed solution to support many authorities to decrease the trust need in the key generators [121]. The previous research studies have dealt with the problem of record-level's security that addressed privacy issues to use and access the patient's medical data. These studies result to focus on the patient's control of their medical data access [118, 119]. Claudio et.al, have proposed an access control solution to provide intension to break the glass exceptions in healthcare system. Their solution has based on the composition algebra, a language, and different policy spaces to access the patient data [120].

Jing Jin et.al have proposed an integrated access control approach that supports patient-focused selective sharing of EHRs with dissimilar levels of granularity, and privacy preserving requirements. They have also addressed the issues and methods on policy irregularities which happen in the structure of discrete access control policies from various data sources [122]. With the increasing amount of different types of data in big data, security and privacy of data has become an important issue. Various researchers have happened to dealt with the issue of data privacy and its security [45]. Large amount of sensitive data of several organizations is easily available online. Main issues of sensitive data' security is sources by the fact that attackers have opportunity to reach the data [123].

Table 2.3: Comparative Study of Different Encryption Techniques on the Basis of Different Parameters

Author	Techniques	Descriptions	Domain	Year	Issues Identified
Hui-Feng Huang et al. [131]	Elliptic Curve Cryptography (ECC)	An efficient key management approach to facilitate inter-operations among the applied cryptographic	Healthcare	2011	Information Security

		mechanisms.			
Jian et al. [132]	Advanced Encryption Standard (AES)	The PHR data is secured by encrypting the zip files using 256 bit symmetric Key encryption.	Healthcare	2011	Information Security
Wuchner and Pretschner [133]	UC4Win, data control approach	A data loss prevention solution for Microsoft Windows operating systems	Corporate Information	2012	Data Loss
Ming Li et.al. [66]	Attribute-based Encryption (ABE)	A patient-centric framework to secure PHRs stored in semi honest servers.	Healthcare	2013	Privacy Exposure
Zheli Liu et.al. [134]	order-preserving encryption (OPE)	Proposed a novel OPE model for securing data that how to hide data distribution and data frequency and discussed the statistical attack for OPE schemes.	Not indicated	2014	Privacy Exposure
Shehzad Ashraf Chaudhry et.al. [135]	ECC	Proposed an authenticated encryption technique and an	Internet Banking	2015	user anonymity

		e-payment method based encryption. They excluded the requirement of digital signatures for authentication.			
Puneet Kumar et.al [136]	AES	Secure the data with AES algorithm Increase the number of rounds (Nr) to 16 for the encryption and decryption	Internet Banking	2015	Information Security
Ximeng Liu et.al. [137]	Paillier Homomorphic Encryption	Secure the patient information without leaking information	Healthcare	2016	Information Security and privacy concerns
Qi Jiang et.al. [138]	ECC	Three factor authentication protocol based on ECC which attempts to fulfil three-factor security and prevent from various attacks	Healthcare	2016	Identity Attack
Keke Gai et.al. [139]	Dynamic Data Encryption	Encrypt the information using privacy	Not indicated	2016	Privacy Concerns

	Strategy (DDES)	classification method with time limitation. This strategy is developed to increase the privacy protection by using a (DDES).			
Sangram Ray et.al. [130]	RSA	Patient centric e-health model based on RSA that allows secure sharing of health data.	Healthcare	2012	Information security
Zhiwei Wang et.al. [140]	ABE	The shared file can be encrypted with the specific policy only once, and it can be decrypted by any receiver whose attributes are satisfied.	Healthcare	2017	Data Leakage
Nai-Wei Lo [141]	ECC	Framework which allow to access historical data EHR of a user and it managed by authentication and authorization protocol.	Healthcare	2017	Data Integrity

Qingchen Zhang has proposed deep computation approach, to secure the personal data using BGV encryption method and to perform the back-propagation algorithm to efficiently encrypt the personal data [124]. Various studies have suggested that PHR (Personal Health Records) should be encrypted before outsourcing the data [125] [126]. Li et.al have been proposed an SA-EDS model has proposed, which divide the file and stores data independently in a heterogeneous environment [123]. Pradhan et.al, have been proposed a BM prime scheme has been developed from Mprime RSA and batch RSA using CRT to faster the decryption process. This scheme is least vulnerable to various attacks on RSA [127]. A hybrid encryption scheme has been developed that combined asymmetric and symmetric key's encryption and decryption process using RSA which overcome the overhead of secret's keys and made efficient encryption of data [128-129]. Sangam ray et.al have been proposed RSA based on public and private keys which permitted to share medical information in an e-healthcare system. To access the PHI of patient, doctors and others staff has to take permission from patients. In this scheme, a random session key has been generated for every appointment to upload and access PHI data from MCS (Medical Center Server). They have been developed a CA based e-health system [130].

2.6 User and their Privacy Paradox

The rise of online communication has changed the lives of individuals, helping them expand their social circle and acquaintances, and connect with others asynchronously or synchronously [142]. Online social media has turned out to be so widespread that social media users perceive a level of emotional support and sympathy while sharing their information with others [143, 145]. Hyun Jung et.al have presented a study to investigate whether supportive communications on Social Networking Sites (SNS) facilitate the effect of SNS use and the number of SNS friends influence the sense of community and life satisfaction [143]. Haein Lee et.al have proposed a scheme that the idea of benefit's encourages the users to tolerate the existence of risk, and the users actively seek to update their information to adjust the level of profits and risks. In their qualitative study, they have observed the types of benefits and risks existing in sharing and their control. For example, Facebook has exploited the unaddressed need of users to maintain and develop social networks in a rapidly busy modern life [144].

Kimberly gender et.al have tested the design principles in four field experimentations including community's members to recommend a movie online. In every experiment,

the participants have given various descriptions about their contributions [146]. Evidence supports that SNS and social mobility are changing at both micro and macro levels, as well as the offline and online effects on life [59]. The situation has so changed that the requirement of social relations is more important for SNS users [147]. Moira Burke et. al. have examined the interaction's role amid pairs—including likes, comments, and wall posts. Consumption of friends' content, such as friends' conversations with other friends, photos, and status updates. They have found that directed communication is related to more emotional bonding social capital and less aloneness [148]. In the last decade, research has shown that most of the population has developed social relations through SNS [148].

Nicole et.al have conducted a study in college that Facebook used by the undergraduate's students. This research has explored two main questions. First, what were the social capital implications and second, how was Facebook integrated in student's daily communication? Specifically, are users articulating existing relationships in Facebook, or are they using the site to discover and interact with strangers [149]. Nicole B. Ellison et.al, have been used the survey data, to explored the relationship between perceived bridging social capital and especially Facebook-enabled communication behaviors. They have explored the role of Facebook behaviors which support the relationships maintenance [150]. Majority of the relation category includes strong connection and weak connection. For example, friendly friends and family reflect strong connection and friends of friends, common interest groups reflect weak connection [151].

Li has been studied review on empirical studies, this research have been summarized the privacy concern's concepts and predictions and concepts of results. An integrative framework has been developed to demonstrate the connections between the factors [152]. Additionally, users enthusiastically post large amount of delicate personal data, though being aware of and worried about the threat to their information privacy [86]. Researchers have also focused on the impact of privacy assumptions on behavioral intention. Raschke et al. have been observed the effect of privacy safety and privacy risk assumptions on the behavioral intention to reveal information through location based application (LBS). Their results have explained that privacy safety assumptions are negatively related with privacy concerns, while privacy risk assumptions positively

affect privacy concerns. User's concern reflects privacy concern on personal information disclosure [153].

Krasnova et.al have been developed Structural Equation Model (SEM) on self-disclosure. They have found that users are mainly inspired to reveal the personal data due to the ability of maintain and develop relations and platform enjoyment [154]. Pew has been examined the privacy opinions and behaviors after the disclosure about U.S. government of Americans. [155]. Monika Taddicken has been studied that the potential impact of age on psychological traits, privacy concerns, attitudes, and self-disclosure. This conclusion indicated that privacy concerns rarely affect the self-disclosure [156]. Spyros Kokolakis has been presented the result of research literature on privacy paradox. They have analyzed the studies that provide a contradictory dialect among the behavior and attitudes [157]. In the environment of social media, users measure potential loss of privacy and of social advantages previously disclosing personal data [158]. Tobias Dienlin et.al, have been proposed that featured a multidimensional operationalization approach of privacy by distinguishing among psychological, social, and informational privacy [159].

Hanna has been adopted privacy calculus to examined culture role in self-disclosure of users [160]. Debatin et.al, have been discovered the concept of information privacy paradox. On the other hand, she did not ready to replicate the outcomes with the sample, since privacy concern didn't significantly load on self-disclosure behavior in their model. Their outcomes uncovered the known phenomenon as 'information privacy paradox' or privacy paradox, where users' privacy concerns didn't directly relate to their online behavior [161]. The privacy conflict has been the focus of many studies. Some, though, have tried to explain it, and regrettably, few have provided contradicting and partial results [162]. Turow et.al have been study described the continuation of the privacy paradox in online self-disclosure by implementing construal level theory for a behavioral model [163].Cory Hallam et.al have been examined the concept of privacy paradox used construal level lens theory. They have been shown how a privacy breach is not far from experienced and psychologically, and compared to social networking activities more solid and psychologically, there was less weight in everyday choices and discussed on the implications of research and practice [164].

Literature provides a rich discussion on the concept of nature, definition and secrecy. Privacy can be characterized as a right, a service, and a state. It can be portioned into

various classifications, i.e. psychological privacy, data security, social privacy, and physical privacy [135, 165, and 166]. Despite the fact that users are exceptionally worried about the potential violation of their privacy, only one part makes adequate use of the privacy options available on SNS to protect their information from unwanted onlookers. The severity of security concerns is confirmed by reports and studies on Internet service companies [167]. Tufekci has been examined the mechanisms used by a survey sample of college students, the majority of Facebook and Myspace users, to communicate boundaries between public and private. There is no connection between disclosure of information on online social concerns and online social network sites. The author has also found important differences in racial and gender. There are other issues about the industrial use of individual's data, including the behaviors tracked by the profiles [168]. An Investigation of approaches for online tracking in digital marketing has shown that people have rejected digital data tracking.

2.7 Relevant Findings

After reviewing available literature and a careful study of the big data security, the following inferences have been drawn:

- One reason for the security loop holes that is predominant in today's digital environment is believed to be a general lack of security focus in the initial phases of the security threat model for big data lifecycle.
- Development of secure environment which can withstand attacks is an emergent need for today's environment.
- Security related weaknesses vulnerability minimization is one way to improve security in big data security.
- Minimization of security vulnerabilities in data creation and data collection phase are feasible. Securing these phases will minimize the effort to secure the further phases.
- Conventional security approaches designed to securing small scale, stable data on firewalls and semi isolated networks are insufficient. Issues needs to be addressed.
- In order to protect the data itself, information dissemination should be preserving privacy and sensitive data should be secured through the cryptography and granular access control.

- A little work has been done to protect data provided by the data creator. For improving security of that data, recent procedures depend greatly on threat models and attack types.
- In general, the security philosophy of heterogeneous databases depends upon the external enforcement approaches. To minimize security attacks, the organization should review the security policies and at the same time; provide the high level security of NoSQL database without compromising on its operational facilities. NoSQL was designed to handle large data sets, with partial emphasis on security which has initiated several security flaws in NoSQL.
- Vulnerable security approaches can be exploited for insider attacks. These attacks may remain ignored due to the poor log analysis methods, with other basic security mechanisms. As sensitive data is stored under a thin security layer, or the data owner it is difficult to maintain and control.

2.8 Conclusion

This chapter reviewed several big data security and privacy issues. Several key concepts such as virtualization and containers have been discussed. The researcher also discussed several security challenges that are raised by existing or forthcoming privacy legislation, such as the HIPAA. The contribution in the area of cloud security and privacy are based on cloud provider activities, such as providing orchestration, resource abstraction, physical resource and cloud service management layers. Security and privacy factors that affect the activities of cloud providers in relation to the legal processing of consumer data were identified and a review of existing research was conducted to summarize the state-of-the-art in the field. The results of our survey reveal that currently there is no methodology to identify privacy requirements according to privacy legislation for processing sensitive data in cloud computing. There are comprehensive security threat models but privacy is not emphasized. Also, there is not a common protection solution or a simple combination of existing mechanisms to build usable cloud systems while ensuring security and privacy of sensitive data. In the remainder of this thesis, we explore this gap through proposing several usable solutions that can be used to ensure security and privacy in cloud computing environments.

CHAPTER 3: DEVELOPMENT OF IMPROVED SECURITY THREAT MODEL FOR BIG DATA LIFE CYCLE

"There were 5 Exabytes of information created between the dawns of civilization through 2003, but that much information is now created every 2 days."

- ***Eric Schmidt***

3.1 Background

Now a day's information is created and processed very speedily and in turn producing huge amount of data with different formats i.e. 'big data'. This heterogeneous data is generated from various sources such as YouTube, social media, online transactions, sensors, etc. Big data has three main characteristics volume, velocity, and variety. It is used in a variety of ways in several areas such as health sector, social networking sites, public sector, government sector, etc. [26]. According to IBM, 80% of the data is unstructured i.e. generated by various sources, and this is in a variety of formats such as video, text, audio, images and combinations of the formats [87]. For the past two years, users are facing several problems due to its large size. This large amount of data cannot be handled through traditional techniques due to its complexity and formats [169]. With the increase in the use and demands of big data, organizations are fighting various privacy and security issues. In terms of privacy and security perspective, researchers should see big data security from every angle and perspective. They should think about keeping the data safe, the process of the big data, and the output of big data process [65]. User's privacy is still a big problem and providing security is very important. In addition, securing big data is a major concern which needs to be addressed [12].

In today's digital atmosphere, physical distance is no longer obstructed when conducting business transactions or reaching individuals in extremely linked societies. Users disclose personal data with the organizations for benefits or services. Most of the

data is created by the users and they are not aware about the security breaches [3]. For providing security and privacy, big data should be examined from every angle. A careful thinking should be there for the protection of data itself. Big data is created by the users and at the creation time several security issues exist. In this chapter a new lifecycle for Big Data has been proposed. The lifecycle is an improvement over previously available life cycle [65]. It explains security threats phase wise. This lifecycle has been unique in itself. The main idea of the proposed lifecycle is to emphasize on the role of users while creating their own data. The privacy loss has happened due to the lack of awareness of user because user isn't aware about selling and purchasing game [43]. After that the researcher has discussed the need of data creation phase and explained the role of users in their privacy disclosure. To strengthen the need and importance of new phase, the researcher has collected 165 key cases of privacy violations from numerous sources. The same data in analysed statistically and the significance of the data creation phase has been proposed.

In an open access environment, huge collection of personal data is available online. Privacy of human is generally challenged by the advancement of technology. In today's scenario, security of huge data is a problem of this information age and to protect it in heterogeneous environment is a challenge. According to some researchers, big data refashions the way of working, living and thinking [170]. In this technological era for a little convenience to live a life with minimal efforts, user hand over the personal data to others. Internet, cell phones and other technologies have provided various facilities to the user. Users on the other hand, accept to their privacy in order to avail the services. Hence, big data improves the services but it raises the privacy issues for users [171]. This is the reason; users are facing the challenges of privacy loss in cyberspace.

However, most of the previous big data security studies do not focus on the threats and attacks. Furthermore, there is lack of focus on the big data lifecycle and how to correlate the threats and attacks based on the lifecycle model. Yazan's et.al [65] data lifecycle threat model has given the idea about threats and attacks on security that come in each phase. There are many techniques and approaches available for securing big data. This is a new kind of big data life cycle. In their threat model, they have described four phases i.e. data collection phase, data storage, data analytics, knowledge creation phase. But this big data lifecycle threat model is incomplete because it does not provide the

source of data clearly. They did not explain the role of data creator. It ignores the user who provides or creates data. With the creation and increased use of big data, several issues are cropping up such as storage issues, processing issues and security issues etc. [84]. Research must be carried out to protect the data provided by the user. Individual's data is very sensitive and it travels from one source to another. In heterogeneous environment users create their own data with a little concern about privacy and security of their information. Usually, organizations provide a platform to access their services at one click. Inappropriately, they seek personal information in exchange [47].

Once user's information is collected, it can be misused. Sensitive data can be sold to third parties. Once the users handover their data, they have no control over it [172]. So the user should be aware about the flow of their information. For the same, the researcher has proposed a new threat model for big data life cycle and added a new phase which explains creation of the data. Most of the data is created by the users and they have no control over it. Personal information about user's online activity, health, electricity usage, and location can be exposed to analysis, raising worries about confidentiality, integrity, accessibility, and loss of control [173]. Commercial sites and social networking sites simply ask basic information which is created by the users. Data is created and after collection the whole data is analysed. The user's data is openly unveiled or this huge amount of sensitive data is combined or processed with other data [174].

3.2 Proposed Big Data Security Life Cycle

From the creation to the decision making, big data has different phases [65]. Data is created, collected, mined, analysed and then meaningful information is extracted. When a user creates his data and submits to data collector then, the main game is started and the privacy of user is abandoned to the hands of others i.e. Data collector. Big data is basically dependent upon the theory of vague data collection, in addition on connecting and reuse the collected data for another aim by the unauthorised parties [175]. Basically privacy varies among the phases and depends upon the consideration of user (creator) that whether the data is sensitive or not. Data creation phase must be the main phase of this threats and security model for big data life cycle. So, it is highly desirable to introduce another phase namely data creator in the existing big data life cycle threat model to make it complete. By adding this phase, the aim of the researcher is to secure

data in the heterogeneous environment. Big data lifecycle threat model in figure 3.1, presented by the researcher, has the following five phases, i.e.

- Data Creation Phase
- Data Collection Phase
- Data Mining Phase
- Data Analytics Phase
- Decision Making Phase

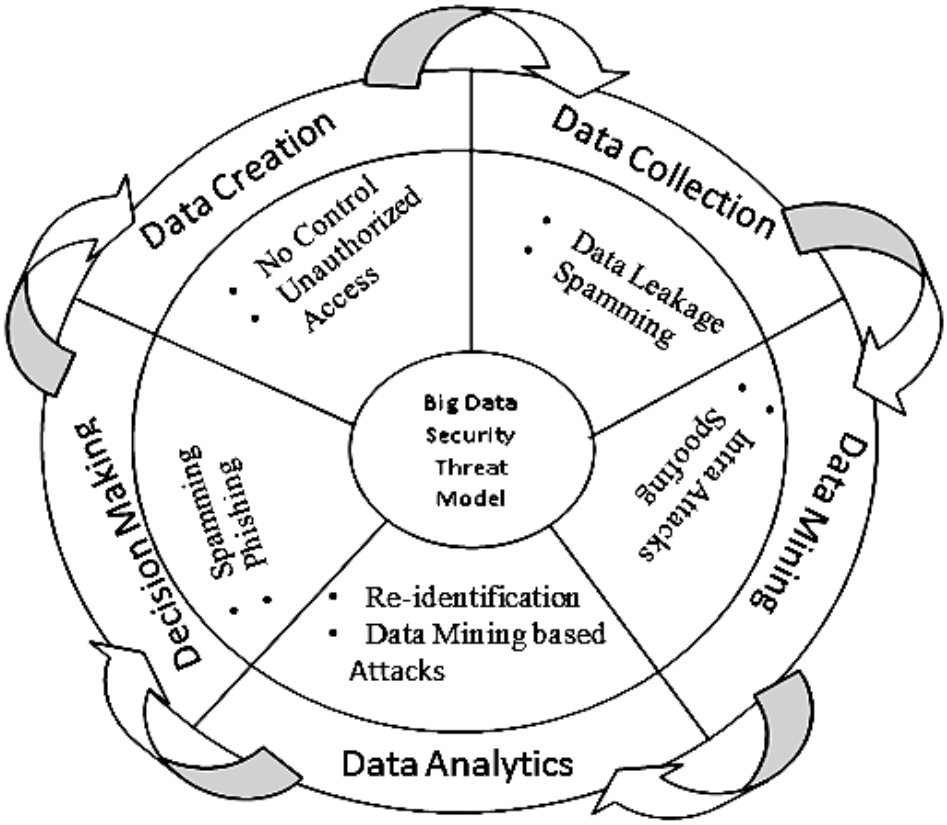

Figure 3.1: Big Data Lifecycle Threat Model

3.2.1 Data Creation Phase

Data creation phase is the most essential phase of improved security threat model for big data life cycle. A data creator is that person who provides data to the data collector (organization). Keeping the security in mind, it is a very important phase. A creator is a person who creates data and submits it to the data collector. He/she is a person who discloses information with his/her own responsibility. When a data creator puts his/her information to other hands then, there is always a possibility of attack. Invaders may misuse his/her information provided to the data collector. A number of cases have been reported, where an attacker has stolen user's sensitive data [23]. So a creator should be aware of these types of attacks on their data. The following suggestions should be kept in mind for a data creator to avoid attacks on his/her data, if a data creator considers its data sensitive, i.e. any theft of data can harm his/her reputation or privacy. While providing the data, the data creator should keep in mind the following suggestions:

- The creator must keep in mind that once any information is submitted from the creator side, the control over it is lost, irrespective of the sensitivity of data.
- The creator should provide only relevant data to the data collector.
- The creator must be sure about the authenticity of the data collector.
- The creator must not provide his sensitive data until required.

Once the data has been put down on the collector's hand, privacy of the same can't be ensured. Now the data privacy depends on the data collection phase as well as later phases. It is required for the data creator to disclose only relevant information to others only as the end user (creator) is prone to security attacks such as authenticity, phishing, etc. In this phase security countermeasures must be implemented.

3.2.2 Data Collection Phase

The next phase of the proposed Big Data lifecycle threat model is data collection phase. At this phase, data collector acts two roles: one is data collector as well as a data provider. This phase plays the role of data collector phase and for the data mining phase, data creator phase plays the role of the data provider. This phase is vulnerable to many attacks such as data leakage attacks and spoofing attacks [169]. To avoid these attacks, data collector must provide only relevant data to data miner. When data collector/provider provides data to the data miner, he/she should always be concerned about the data privacy. There are several approaches to preserve user's privacy at collector's level [65].

3.2.3 Data Mining Phase

Data collector provides data to the data miner. The main concern of the data miner is to secure sensitive data from attackers. Data miner needs to change data, what he/she gets from the data collector. In this phase, there are various attacks. This phase is susceptible to numerous attacks such as intra-attacks, spoofing attacks etc. [63, 65].

3.2.4 Data Analytics Phase

Data analytics phase is the fourth phase of this proposed security threat model. In this phase, data will be analysed which is obtained from data mining phase. For data analysis, the data analyser observes large amounts of data to find hidden patterns to fetch sensitive data. The data obtained from this phase is used to make a decision [63, 65].

3.2.5 Decision Making Phase

The last phase of this big data lifecycle is decision making. This phase uses verified and valuable information received from the previous phase. The right purpose of data mining is to deliver the useful information to the decision maker. Thus, the decision maker may choose a better way to fulfil organization's goals. There is no responsibility for the security of data of a decision maker [63, 65].

3.3 Importance of the Data Creation Phase

Researcher has proposed a new phase in big data lifecycle threat model i.e. ‘Data Creation Phase’. The data creator creates information at this phase and this information is stored in data collection phase. A creator can create information such as name, mobile number, email, etc. Once the information is stored at the collection phase, the data creator has no control over this. Now it is the choice of data collector (organization) to maintain the privacy of the collected information or not. If a creator wants to access any organization’s services, then he has to provide personal information and he has no option to deny the information. Once a creator discloses his basic information to data collector his privacy might be compromised due to the unexpected privacy breaches. If the creator decide not to provide the information, disclosure of which can arise a problem for him later then he can avoid many problems related to privacy issues. It justifies the need of “data creation phase”. In order to strengthen the need and significance of the creation phase, the researcher has collected data of various related

privacy violations from several sources [177-233]. The detailed study has been performed on these cases. The study is summarized in table 3.1. Data is analysed keeping in mind to show the need of data creation phase. The results show that if the data creators are careful in advance about what data should be provided and what not, then frauds could be avoided.

Table 3.1: Some Major Cases of 2007-2016

Year		Description	Mode
2016	I	According to FACC, a cyber-fraud happened, where intruders stole nearly £ 50 million by email.	Email Id
	II	In Balajinagar a 49-year old victim of tele-phishing attack almost lost Rs. 24978. The person was reportedly been duped by the unknown suspect.	Mobile No.
	III	A 61 year old retired government employee was cheated of Rs 3.35 lakh when he booked an air ticket.	Mobile No.
	IV	Cyber-crime officials arrested 'Nigerian fraud' gang. This gang sent a fake e-mail to raw material supplier. They have chetaed 13 lakh form him.	Email Id
	V	A businessman from Raipur, Tushar Mohanty was trapped by false call. According to IT expert and lawyer Mahendra Limaye, in Nagpur June 25, 2014, a same complaint have registered, by Bodhiratna Fulzel. The amount was transferred to the same accused — SK Ibrahim's account, but in different Odisha Axis Bank branch.	Mobile No.
	VI	A banker, Srinivas Reddy received a mail saying that the 'Rio Olympic Committee' has offered 40% discount on air tickets which was going to be happen in August.	Email Id

	VII	A man has cheated by a fake call and lost Rs 36,480. On 11 Oct 2016, victim received a call from an unknown number. The caller called from the victim's bank account.	Mobile No.
	VIII	A fake email was sent to the Carole Gratzmuller, boss of a medium-sized French company. The scammers have duped €500,000.	Email Id / Mobile No.
	IX	A woman has lost 11 lakh rupees. She received a message declaring that she had won 45 lakh and shelled out Rs 11 lakh to claim the amount.	Mobile No.
2015	I	In Eastern Europe a malware, was developed by cyber attackers. After collecting all online banking details, the attackers stole money from all over the world. Britain lost nearly £20 million in this case.	Email Id
	II	The hackers called the victim and presented themselves as a service team and told victim that they were calling from "Action Fraud Litigation Services team" form legal prosecutor's department. They will help to recuperate the amount of victim that he has lost in past fraud. Once the victim finished his transaction, the attackers were not available to make contact.	Mobile No.
	III	A BPO from Kadubeesanahalli has claimed to lose over Rs. 99 lakh when the company received a phishing mail. The BPO received a phishing email for payment. The phishers reflects themself as a representative of foreign firm. They changed an alphabet in the official email id, the recipient couldn't notice.	Email Id

	IV	A Graeme Smith, told the Guardian that he has become a victim by the hackers and after the fraud, £2,815 has been lost and his computer was affected to install a piece of malware and pass the bank details to attackers. In this case the fraudsters used the customers' personal details to gained access of the victim's computers and bank.	Mobile No.
	V	The Oil and Natural Gas Corporation Limited (ONGC) has lost 197 crore after cyber criminals duplicated the official e-mail address of the public sector firm with slight changes and to transfer payment to the customer in Saudi Arabia.	Email Id
	VI	Sayara a victim woman registered a complaint to police; the online transfer of money was done without her knowledge. The police said victim received a call from the man last month. He represented himself as a bank officer and asked for the ATM details. She had accounts in State Bank of India and the Bank of Maharashtra. She has shared the information with him and the attacker withdraws the amount Rs 34,890 from her bank account without her consent.	Mobile No.
2014	I	In this case a victim received a text message from a "potential buyer" who asked additional information about the bed. After some conversation approving to buy the bed, then he asked to pay money by PayPal.	Mobile No. / Email Id
	II	Attackers were cold calling victims and telling them that they need to pay £150 for "release" reimbursement for which they were entitled.	Mobile No.
	III	During April and May, all claimed marketing	Mobile No.

		limited sent approximately 6 million unwanted spam text messages on victim's cell phones.	
	IV	Action Fraud has got reports of an attacker cold calling members of the public that they have to pay the debt or face sent by a bailiff. The attacker was claiming to be from a legitimate company called Jacobs.	Mobile No.
	V	The 'number spoofing' scam, where attackers cloned the organisation's mobile number. Where attackers called people and represented as police officers, bank staff, or other trusted companies to persuade their victim to part with personal and financial details. Once the attackers won the victim's trust they tried to get victim's OTP, PIN and thereafter the money was stolen from the victim's bank accounts.	Mobile No.
	VI	The attackers disclosed details and passwords of almost 5 million Gmail accounts on Bit coin Security.	Email Id
	VII	A man in Surrey has been arrested by the Police on a scandal, which has been prepared to cheated university's students. The attackers have been offering students financial benefits to sign up for personal details. After that students disclosed their sensitive details about accounts, credit or debit cards to the scammers.	Mobile No.
	VIII	One scam, masterminded by two 26 year olds, revolved around the hijacking of mobile phone accounts. The two individuals began by creating a fake company that purchased lists containing customer details.	Mobile No.

	IX	In this case the LinkedIn website was hacked. The phishing emails are not from the LinkedIn. When a user clicked on the fake link it directed to the fake website exactly looked like LinkedIn website.	Email Id
	X	The fraud was started and maintained by online platforms; often used the actual employer's names and recruitment agencies. The website or email has advertised for the job. After submitting the CV, the victim was offered a job and said to pay an advance fee almost £50.	Email Id
	XI	The attackers sent the scam emails represented Neil Trotter from - who publically disclosed winning amount £107.9m of Euro millions lottery. The attackers have been taken advantage of last lottery winners and they sent phishing emails to the victims.	Email Id
	XII	The attackers have been stolen wallets and purses of victims' from various locations such as care homes staff rooms, and health centres. The fraudsters stolen every detail about the victims and called the victim pretending to be bank. In one of the case the fraudsters stole a person's purse and withdrawal thousands of pounds from banks.	Mobile No.
	XIII	The fraudsters have been sent fake emails that claimed to victims that they had to pay a false parking fine which has a suspicious attachment.	Email Id
	XIV	Trading Standards have said that a tenant got a call, from which the caller pretend to be "Scottish & Southern Energy" company. The caller was trying to sell the "Green Energy" to resident who paid him £1000 cheque as cash back".	Mobile No.

	XV	The cold callers were pretending to be from the Office of Fair Trading (OFT) that they could recover payment insurance such as bank or other fees. The fraudsters claimed that the service requires an advance fee paid by money transfer or that victims required to disclosed their bank account details. In some cases the victims have been asked to pay between £200 and £250.	Mobile No.
	XVI	The officials' authorities, supported by the European Cybercrime Centre in Europe, have arrested group of 12 members, with the €15,000 in cash and seized vital digital evidence in the vishing case, according to a statement from the European Union's law enforcement agency.	Mobile No./ Email Id
	XVII	A victim with Alzheimer was called by a firm pitching diamond investment and was handed over more than £90,000 in three months. Every time a transaction was planned, he was asked to keep a secret.	Mobile No.
2013	I	The cumbria police warned the society to ignore the fake calls from fraudsters, who claimed that they are calling from OnlineTec company, to get access to computers. The attackers on call requested to remote access of user's computer. Some people became stupid and have provided remote access to attackers.	Mobile No.

	II	In this case attackers those users who were looking for a job. A scammers group posted a fake online ad for job seekers. People clicked on the fake link and the malware downloaded on their computer that recorded their keystrokes, have gathered financial data which was transferred to the attacker's gang.	Email Id
	III	Essex County Council issued a warning after a parent of students. People got a call from the Education Welfare Service, stated that their children's did not attend the school on a certain day and or that they had to pay fine of £340.	Mobile No.
	IV	When attackers broke into Twitter account of AP, their fake tweet about Barack Obama being wounded in an explosion. AP employees received an email asking them to click on a link that was believed to have been in the Washington Post article.	Email Id
	V	A fake mail sent to the user's Gmail account. The fake emails stated that "You got a new message from Skype voicemail service". The aim of this fake email is to get users download the attached malicious "voicemail" file which contained Zeus Trojan.	Mobile No.
	VI	It is estimated that all over the UK, the average household got nuisance calls in one month; it was found that about 40% of the calls received by older and susceptible residents in Scotland are nuisance calls.	Mobile No.
	VII	FFA UK has seen a total increase of £36m in remote banking (telephone and online) and application fraud in the previous financial year.	Mobile No.

		Early estimated shown that at least £7m worth can be attributed to the scam, known as 'Vishing'.	
	VIII	An email circulated through Action Fraud, threatens legal action due to failure to offer Western Union details to PayPal. The personally email stated a recent transaction with "Mercy Peters" and threatens legal action due to refusal to provide Western Union details to PayPal. The attackers claimed that Mercy Peters has contacted Action Fraud to report the violation of the PayPal agreement.	Email Id
	IX	A resident in Cumbria alerted the police after got a cold call from those who had been fraudulent to Sky. The attackers told the owner that their Sky card was due to expire soon and was more payment on his account which was needed to be refund.	Mobile No.
	X	Facebook disclosed that a bug in its system revealed 6 million users email addresses and phone numbers to be exposed. Facebook stated that this bug meant that the site has collected information about users to create master records such as email addresses and phone numbers that the owners never wanted to share.	Mobile No. / Email Id
	XI	A former director of Indian Institute of Technology (IIT), Kanpur Sanjay Dhande (65), has been cheated of Rs 19 lakh through an online fraud. Rs 19 lakh were transferred from account of Dhande in ICICI Bank without his knowledge.	Mobile No.
2012	I	In Kerala, a non-resident Malayali had an account in bank. He was lost $10,000 when bank officers observed that a phishing email request to transfer money to another bank's account in Ghana.	Email Id

	II	The attackers withdrew the money Rs. 11.14 lakh and fraudsters transferred this amount to various accounts through internet mobile banking of doctor's account. The fraudster cloned the victim's SIM card and blocked his number and duplicate cards were used to commit fraud.	Mobile No.
2011	I	On 23 August 2011, in Kathua (Jammu) an attacker stole the victim's user ID and password of account through phishing email. With the help of this information attackers hacked the net banking information and transferred Rs 80.1 lakh to other PNB account.	Email Id
	II	A school teacher in Nalgonda Andhra Pradesh got an SMS stating that he won four lakh pounds in an anniversary promo of Microsoft lottery. The teacher fooled by the attackers. Five Nigerians have been arrested by police for alleged duped school teacher of 18.42 lakh rupees in the fake British online lottery racket in Andhra Pradesh.	Mobile No.
	III	An Indian woman and five Nigerians have been arrested for cheating on a charge to cheat a student in Hyderabad. A B.tech student has been duped of Rs 18.73 lakh through a phishing email that claimed he had won 7.5 lakh pounds prize money.	Email Id
	IV	The AIIMS administration has registered a complaint in cybercrime cell on the hacking of two senior professor's email accounts in gynaecology department.	Email Id
2010	I	In this case Rs 3.39 lakh were illegally transferred from the account of victim. A Victim received a fake email and trapped by the attackers and	Email Id

		transferred money to attacker's account. He had reacted to phishing mails sent by attacker.	
	II	In New Year, the phone scandal also hit a credit union and a bank in Indiana. In this scandal the hackers asked for the debit card details. Several users trapped by the attackers and send the debit card details.	Mobile No.
2009	I	On June 12, Raj Babbar claimed that he got an email asking about his personal and financial details. he said that he supposed that it was a junk email and didn't respond but next day he could not access his office's mail. Reason: the attacker sent emails to all his contacts and was saying that Babbar was trapped in China and had lost his luggage.	Email Id
	II	"In Delhi, many fraudsters were circulating emails and SMS to the people who won the lottery. Some email and SMS offer attractive offers to provide loans at very low interest rates.	Mobile No./ Email Id
2008	I	The criminal hacked the account of Bank of Baroda linked to Gujarat Informatics Limited (GIL) and stolen Rs 7.39 lakh in one month.	Mobile No./ Email Id
	II	The Surat police had arrested a criminal from Mumbai's western suburbs for stole Rs 12 lakh from at least eight HDFC Bank accounts of customers.	Mobile No.
2007	I	Software professional Atin Kapoor, found that Rs 85,500 transferred in three bank accounts from his ICICI Bank account using net banking without his knowledge. He got an SMS about the transfer and found out that he had been cheated.	Mobile No.

Any organization is a rich source of user's personal information and shows another opportunity for hackers. For example, when a user applies for rental properties, then it is generally compulsory for the user to give personal identity information such as name, birth date, and passport etc. financial information – proof of income and user's details and other basic contact details such as mobile number, email addresses, applicant's permanent address, and other references [46]. Here user is creating basic information and handovers his personal information onto the other hands.

3.4 Statistical Analysis and Results

For the aim of the research, Social Networking Sites (SNS) (Twitter, Facebook, Wechat, Instagram, Snapchat, and LinkedIn) and Commercial Sites (CS) (lifepartner.com, policybazar.com, shaadi.com, and jeevansathi.com) have been chosen. It is found that these sites regularly ask for basic information such as Mobile No., Email Id, Birth date address, gender, etc. This is the basic and compulsory information which is asked by SNS and commercial sites. Figure 3.2 and figure 3.3 represent the percentage of basic information of SNS and CS correspondingly. It is discovered that Email Id, Mobile No., and date of birth are the most frequently asked information. Although this basic information does not seem very important in relation to security and privacy but it has been noticed that disclosure of the same are causing several incidents related to security. For the research, major cases have been taken for analysis during the year 2006-2016.

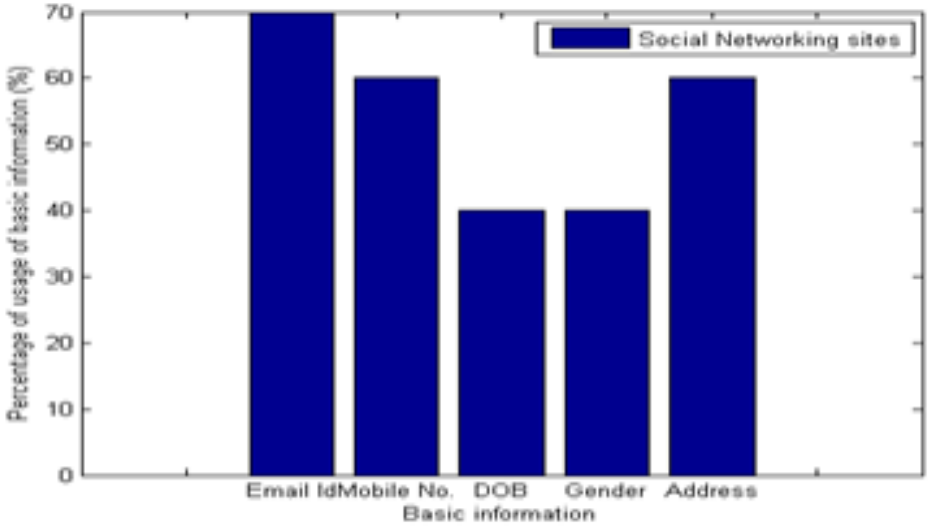

Figure 3.2: Basic Information asked by Social Networking Sites

The analysis (refer figure 3.4) represents that incidents are taking place even with the only disclosure of Mobile No. and Email Id. Increase many incidents related to security with the year enforces the need of the data creation phase. The researcher has also calculated the loss of money in these incidents. Figure 3.5 and 3.6 represents annual loss in terms of money due to expose of Mobile No. and Email Id respectively. This again proves the need of data creation phase in the big data life cycle.

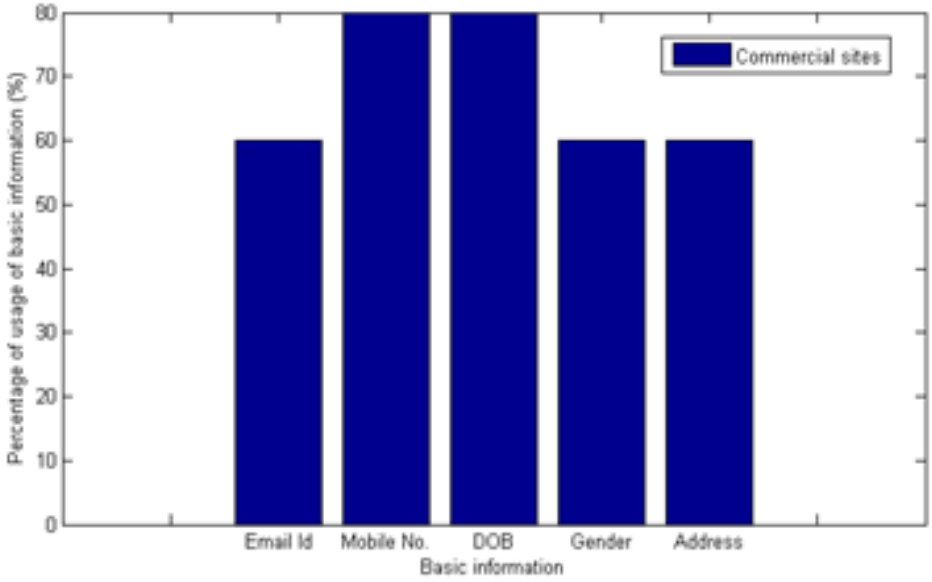

Figure 3.3: Basic Information asked by Commercial Sites

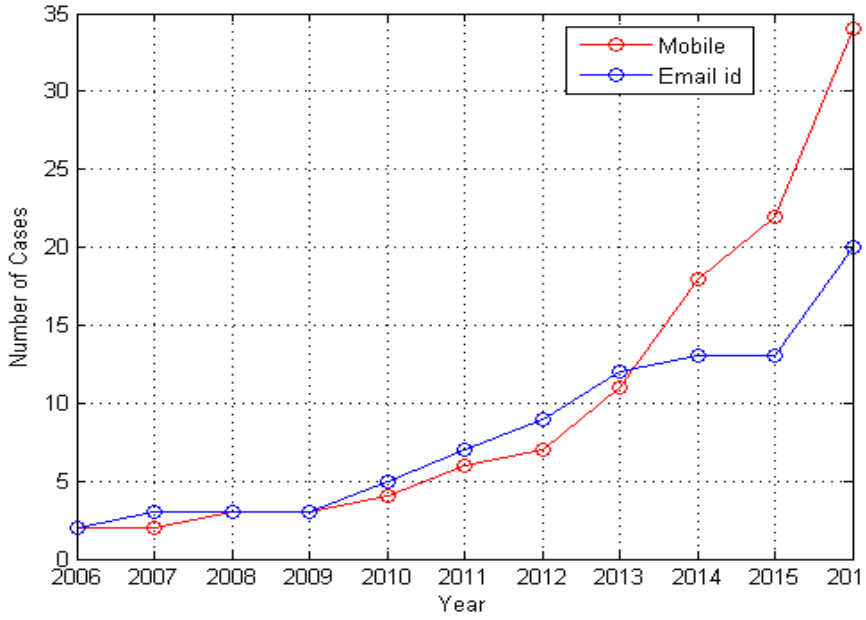

Figure 3.4 Number of Cyber Cases Occurred Due to Reveal of Mobile Number and Email Id

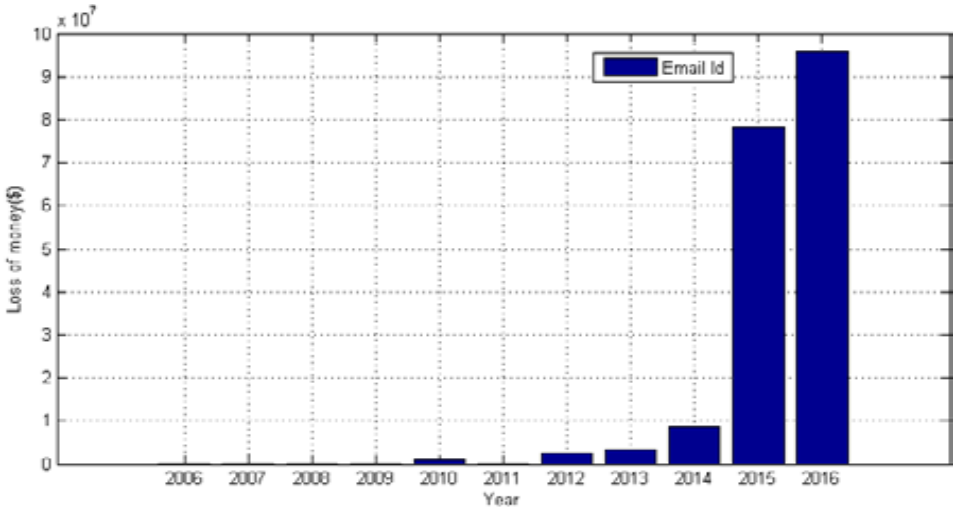

Figure 3.5: Loss of Money by Misuse of Email Id

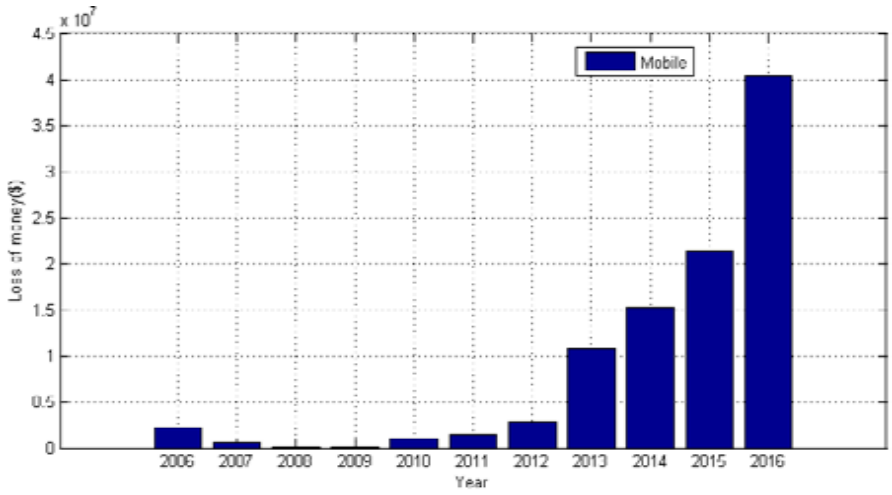

Figure 3.6: Loss of Money by Misuse of Mobile Number

3.5 User's Role in their Privacy Breach: Privacy Paradox

On one hand, life has entered the era of ease and expediency because of social media and mobile phones, but on the other hand it has offered numerous issues related to privacy, security, and data availability. Between all the issues addressed, privacy and security of users is the most critical one because users are disclosing their personal information to access services of organizations. The disclosure of users own personal information is called self-disclosure. On the basis of self-disclosure hidden activities, priorities, and other information of a user can easily be found. More accurately, it can be said that self-disclosure puts user's identity and integrity at risk [176]. The rise of data mining on the internet has been driven by various factors which include: expanding accessibility of information on users and their online activities as more social activities

take place online; the low cost of gathering, storing and processing information; and the development of social media platforms. Mined online information can be matched and then joined with the information from other sources. For example, in a prominent case in United States, accused has admitted about the information mining operations of the National Security Agency and Government Communications Headquarters in the United Kingdom [234].

The use of social media is so widespread that many consider it a regular part of everyday life [235]. Social networking sites regularly offer new apparatus to build and keep up connections and are hence of specific significance for psychosocial improvement [150]. One of the main features of Social Networking Sites (SNS) is clearly designed to facilitate the creation and maintenance of links between individuals through their self-disclosure of information. Hence, there is a connection between use of SNS and social capital which is viewed as a positive result generated from the creations of relationships [149] [163]. However, there is a controversial issue regarding users, its online disclosure. The reason behind the issue is the information created online. It results in the potential loss of privacy of users if sensitive data is made public [236]. Literature shows that various SNS users are afraid that their privacy can be violated online, although few users apply all necessary means to preserve sensitive data [163]. Generally, people ignore the risk in the wake of profits. Hence this may affect human behaviour in various fashions [145].

3.5.1 Research Model and Hypothesis Development

When users disclose their personal information in any situation, they consider risk and profit simultaneously. Collection of an individual’s information can create patterns about the individual’s real life. For example, collected information of location context can disclose the individual's activity area which can be pursued for crime or stalking [236]. In addition, information related to any context can be private and it is closely related to its provider. Hence, researcher expects that context information is personal information and at the time of sharing, people can keep in mind the risks and benefits. As proposed in the previous section, self-disclosure behaviour is the result of a trade-off between two deliberate and effortful thoughts: the evaluation of expected benefits from the use of SNS and threats to personal privacy. Researcher expects that benefits and risks are considered simultaneously in the process of context information sharing, even though they are ambiguous in an internet setting. According to Social Penetration

Theory (SPT), which explains interpersonal relationship development, people tend to predict reward of disclosure, and take into account the result of the comparison. Communication Privacy Management (CPM) theory also points out that the benefit–risk ratio is a crucial reference while making decisions about personal information disclosure [145]. Hence, when people share their private information in social situations, they consider benefit and risk together.

The research model in figure 3.7, describes the self-disclosure behaviour as an independent variable. On the basis of literature, authors proposed two intention variables, both with an opposite result, assume dependent variables. Personalized balance or trade-off in dealing with two-way intentions is the risk of privacy and benefits [237]. The researcher expects a negative collaboration between distant-future intentions and privacy concern. The expected degree of privacy risk and benefits is based on the information sensitivity. Information sensitivity is explained as 'depth' in SPT [238]. Information with high sensitivity is considered to be more risky due to the vulnerability to loss made by the disclosure [239]. Therefore, expected privacy risk increases as the sensitivity of information increases. Due to the reciprocal nature of self-disclosure it gives others sense of obligation to disclose a similar level of information. Besides, users knowingly disclose information to highlight benefits or obligation feelings from others. It is expected that exchanging delicate information can make a strong connection between the users [145].

The present study explores privacy paradox by designing three experiments to examine how people's opinions on online privacy change (or are maintained) when encountering messages arguing for either protection or disclosure of personal information on the Internet. This study modifies the typical counterargument experiment in order to address the privacy paradox problem. As already discussed, ordinary users do not give serious consideration to the online privacy issue, implying that people's opinions on this issue are pliable and fickle. A potential breach of sensitive information in the future, due to online disclosure in the present, is thus considered psychologically distant and consequently related to distal intentions. The researcher also believes that the privacy paradox exists because privacy concern is not predicting the disclosure behaviour. Consequently, researcher proposed H0 and H1, as follows:

H0: Self disclosure positively affects benefits.

H1: Self disclosure negatively affects benefits.

Privacy risk is the uncertainty related to the negative consequences of using any service in which the disclosure of personal information involves potential loss [240]. Xu and Gupta specified that privacy risk associated to sharing personal information with a service provider [27]. Users are afraid to adopt any service in which a provider can share personal information with third parties without the user's consent.

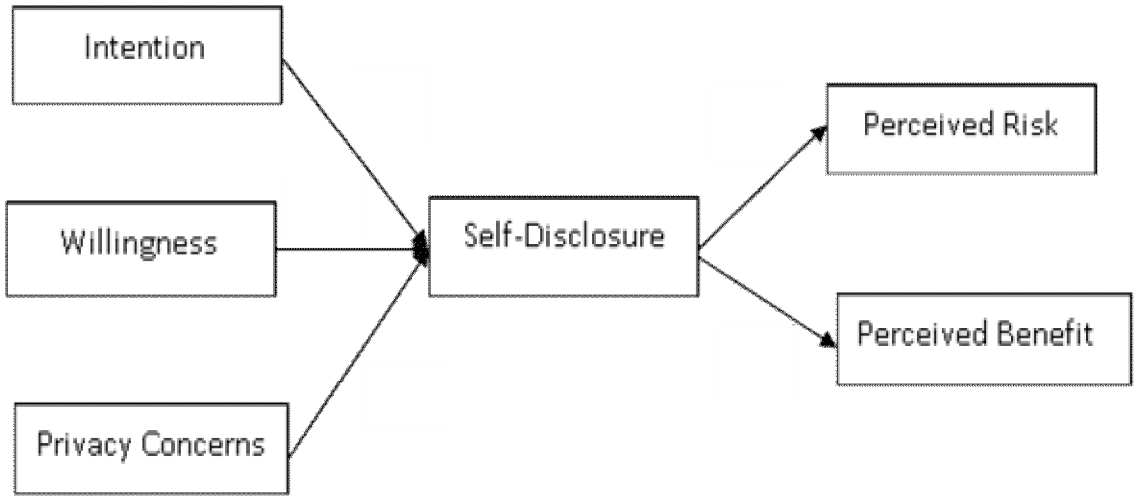

Figure 3.7: Research Conceptual Model

3.5.2 Methodology

For the purpose of proving that the self-disclosure negatively affects benefits i.e. self-disclosure increases privacy risks, the researcher has adopted the following procedure.

a) Participants

Thinking of users about social media data mining has been explored. In context of limited resources, our decisions about how to do this were opportunistic but nonetheless robust. This study has been conducted in India. A pilot study has been used to assess the appropriateness of the questionnaire items, validity and reliability of the research constructs, yielding satisfactory results. In total, 267 samples have been received with 67 partially completed surveys. A descriptive statistical analysis has been conducted to summarise respondent profiles and the characteristics of all the tested variables. Participants in the initial sample were students not specially related to the topic of the research. In age distribution, this process yielded a useful sample of 200 participants.

The respondents' ages were between 18 and 43, with the mean of 28.5. 46% of respondents were male, and about 32 per cent of the sample had a college degree. The average number of social media accounts per person was 3.9, and about 43 per cent of our sample reported using smart phone apps to disclose their daily activity.

The survey uses a 5-points Likert-type scale [241]. The questionnaire items have been developed based on the literature. All of these items have been measured using a 5-point Likert scale, with 1 indicating 'strongly disagree' and 5 indicating 'strongly agree'. [Appendices A].

b) Measurement Accuracy Analysis

The research model included six constructs. Every variable has been measured with several items. The final sample has been tested for the validity and reliability. The Convergent Validity should be the Cronbach's Alpha (>0.7) and average variance extracted (>0.5). Table 3.2 indicates the mean, standard deviation (SD), the Average Variance Extracted (AVE), and the Cronbach's Alpha for the constructs. Table 3.3 shows the CR ranged exceeding the 0.7 threshold from 0.84 to 0.89, and AVE ranged exceeding the 0.5 threshold from 0.59 to 0.78, showing that all constructs got these reliability criteria.

Table 3.2: Correlation Table

Research Constraints	**Correlation**							
	Mean	SD	PC	I	W	SD	PR	B
Privacy Concern (PC)	9.85	3.027	1.00					
Intentions(I)	12.00	3.138	.51	1.00				
Willingness(W)	7.50	2.566	.64	.12	1.00			
Self-Disclosure(SD)	7.20	2.257	.44	.93	.18	1.00		
Privacy Risk(PR)	4.80	1.473	.38	.44	.22	.53	1.00	
Benefits(B)	5.05	1.602	.56	.22	.56	.14	.31	1.00

The calculated values fulfil the convergent validity criteria [242]. All the Pearson correlation coefficients constructs are shown in Table 3.2. The values on diagonal are the AVE square roots for the constructs. The measurement model test that indicates suitable reliability, convergent validity these values are used here to evaluate the proposed research hypotheses. For convergent validity, we have used Standardised values of λ. The researcher has suggested that standardised λ values should be at least .5 and are statistically significant. Hence it is statistically significant at $p < 0.001$, confirming convergent validity.

Table 3.3: Internal Reliability and Convergent Validity Test Results

Constructs and Measurements Items	Λ	Composite Reliability	AVE
Privacy Concerns (PC) **Restricted profiles to others.**	0.73	0.89	0.61
I consider my data sensitive because it contains my personal information.	0.87		
When online companies ask me for personal information, I think twice before providing it.	0.87		
I'm concerned that online companies are collecting too much personal information about me.	0.64		
It is very important to me that I am aware and knowledgeable about how my personal information will be used.	0.80		
Intentions (I) **Sometimes I provided wrong information to protect my personal data.**	0.85	0.85	0.65
To obtain a free gift, I would share my data online.	0.78		
When I buy a new iPhone then I openly share my information.	0.79		

Willingness (W) **From my online profile, it would be easy to understand what type of person I am.**	0.85	0.86	0.75
I want to share my personal thoughts, experiences.	0.89		
Self-Disclosure (SD) **I share personal information like age, home address, and favourite restaurants.**	0.93	0.88	0.78
I share medical history and financial information.	0.84		
Privacy Risk (PR) **When I check in my actual house, who knows? Someone could come and find me.**	0.67	0.85	0.59
Since my boss does Facebook a lot, he is my friend on Facebook he might know everything that I post.	0.82		
Believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.	0.81		
If I forget to logout my id on public system then it is very easy for someone to misuse it.	0.78		
Benefits (B) **I fulfil my social needs in some way by connecting with others.**	0.80	0.84	0.63
I derive satisfaction from disclosing online.	0.79		
Ease of access	0.78		

3.5.3 Result Analysis

The researcher has the hypotheses using regression analysis implemented in IBM SPSS Statistics 2.0 (trial version) to evaluate whether the user considers benefit and risk at the same time. Researcher used regression analysis between perceived privacy risks and perceived benefits on self-disclosure (Figure 3.8). To get support for the hypotheses,

the path coefficient was examined. The result indicates perceived privacy risk and perceived total benefit on self-disclosure is statistically significant ($p<0.05$) and the impact of every path has been 0.53 and -0.043. In contrast to the effect of positively or negatively affects benefit, R^2 increased to 0.53 both privacy risks and benefits are dependent variables. R^2 represent the percentage defining dependent variable based on independent variables. It is particularly interesting from researcher's point of view that perceived privacy risks are affected more than the perceived benefits.

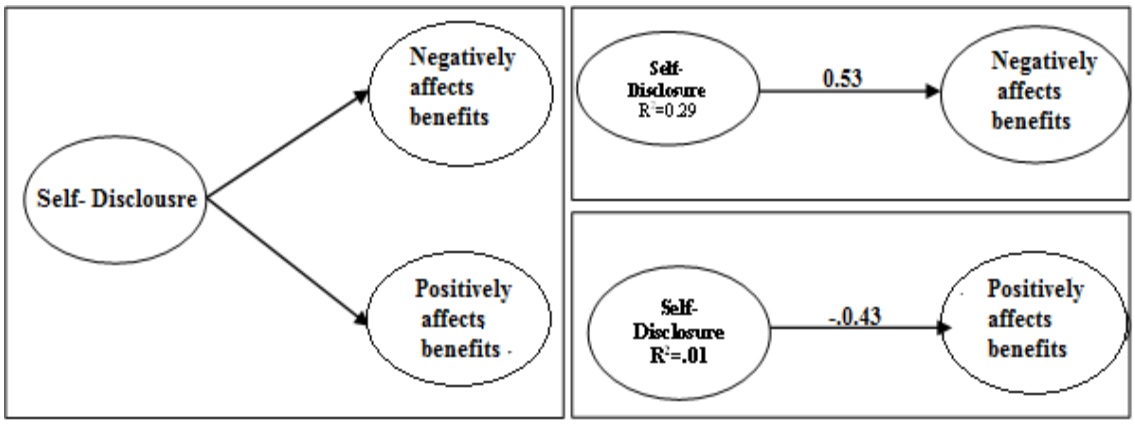

Figure: 3.8: Regression Analysis

a) Chi-Square

The Chi-Square statistic is used to test the relationships between classified variables. Chi-Square's null hypothesis means that there is no relationship between classified variables. These are independent in the population [243]. The researcher has conducted Chi-Square test to analyse the influence of different context and control types on privacy risks and benefits (H0, H1). There was main effect of context type to perceived privacy risk. First, the effect of different types of context information has been examined and found that self-disclosure would induce a distinguishable difference for perceived total benefit and perceived total risk.

- The degree of freedom is 4.
- ($CV>\chi^2$), statistical observation is chi value (18.41) is greater than from chi critical value (9.488) (for $\alpha=0.05$).
- Hence the researcher would reject the null hypothesis (H0). Thus, it is validated that Self disclosure negatively affects benefits (H1).

3.6 Discussion

Our model shows that the privacy concerns, intentions and willingness are significantly related to the self-disclosure behaviour. Users actively share their personal data including reference information regardless of high privacy concerns, since they don't think about the privacy risk. They just consider expected benefit. They don't think what can be the consequences or loss to openly sharing their personal or sensitive data. Moreover, there exists a significant indirect impact of privacy concern on the behaviour of self-disclosure through the future intention. This future intention interferes with the effect of privacy risk on self-disclosure. Future intentions accrue the positive impacts from privacy risk with the negative effect from benefits. Researcher believes this study provides proof of the privacy mechanism. Specifically, expected privacy risk has a greater impact than the expected benefits did on self-disclosure behaviours.

3.7 Conclusion

In this chapter researcher has presented an improved lifecycle for big data and phase-wise security threats. Researcher has presented the vulnerability of basic information asked by the SNS and commercial sites. Researcher has collected 165 major cases for the theoretical and statistical analysis. This study shows that many incidents are taking place only with the disclosure of Mobile No. and Email Id. Increase in the security incidents related to Mobile No. and Email Id shows the need of the data creation phase. Understanding of behaviour of security incidents and their effects of these incidents on society will help to find out the means to avoid the situation. In this chapter, researcher has also examined the patterns of use of disclosing information on internet. These are focusing on the users' concern of expected privacy risk and expected benefits. The results show that the willingness and intention to disclose reference information is influenced by the privacy risk and benefits simultaneously.

In particular, the impact of privacy risk is greater than the benefits. The results showed that privacy concerns, intentions and willingness have a positive impact on perceived self-disclosure, which consequently decreased benefits and increased privacy risk. The researcher has tried to describe the existence of the privacy paradox in online self-disclosure on a behavioural model. This study has been focused on, the self-disclosure behaviour and related privacy awareness to secure users information. This study utilizes the SNSs users, and the data based on user's response. Results indicate that most of the

respondents discloses significant amount of personal information and not conscious of the visibility and information leakage to the third party service providers and unknown users.

CHAPTER 4: PROPOSED POLICIES FOR PRESERVING USER'S PRIVACY

"User confidence is crucial for digital economy. Customer as a product and unsafe privacy are not sustainable business models. Digital is sophisticated enough to combine Security, Convenience and Personal Privacy."

- ***Stephane Nappo***

4.1 Background

Within the several effects of digitalization, the debate has begun on 'big data'. Now a days information sharing may pose a threat to someone's privacy or organizational secrecy. The dilemma of information usage and privacy seems to be fast and acuter. A user's online individuality includes a great choice of traceable characteristics such as email-id, mobile no, DOB, name, passwords, id proofs, etc. These characteristics are enough to uniquely identify any user on the internet. The ease and availability of rich data sets and the use of data analysis undoubtedly affect security concerns to extract significant information on that data. It may also be possible that data collected for one reason may be reused for other reason without the concerned user's consent [172]. Particular information can be sold, mined with other source of information, and try to make useful information which improves the services. Several organizations appoint the persons with mathematical and analysing abilities to available information and make this information more useful [174]. User's data is fragmented everywhere, the risks of security exists every time in this cyber space. Hence, there should be big data privacy policies to avoid misuse of individual's information. These privacy policies should be followed by both the creators and the collectors in order to avoid privacy loss. However, there are various studies going on in the area [175, 171]. There is still a major challenge to develop a strong and protected platform. There have been innumerable efforts to

improve the privacy of digital users [28] [244] but it still needs to be improved. On the contrary, while providing its services the user's security and privacy can easily be maintained with minor modification in organizational policy.

However, along with some advantages, big data opportunities are also raising some security issues. In today's environment user privacy is just one single click away. There is only a single layer between individuals and their online privacy. Anything they update on the internet, it can be easily spread across the globe. The organizations trace the online activities of the Internet users [96]. For example in 2010, a wall street Journal, conducted a survey on 50 popular mobile applications on android and iOS operating systems and examined that both apps gather and share data to third parties without user's consent. Another study has been conducted in 2015, on 110 popular mobile apps which are freely available on Apple and Google play stores and examined that these applications provide private and sensitive data about the users to third party, for example health data, locations, and job data etc. This study shows that mobile applications share user's personal and sensitive data to the unauthorized parties without user's permission [245]. In today's scenario, protection of big data is a major issue and a major challenge in this heterogeneous environment. According to various scientists, big data modifies the way of working, thinking and living. In this technical era for the little ease to living a life with slight effort, customers pass on the personal data on to other hands. Mobile phones, Internet, and other technologies have provided such facilities through various organizations. If they want to achieve something they would accept to be monitored. Big data increases services but it raises privacy concerns for users [171]. In cyberspace, users are facing the issues of privacy loss. Privacy loss is happens due to the lack of awareness of user because users are not aware about selling and purchasing of their information. Normally organizations provide services. User accesses these services and creates data about them or their organizations. At time of accessing the services of any organization, a user provides personal information in exchange of it.

Organizations have their own regulations, rules and policies. Every organization runs the entire business and they have a large repository. Any company managing business-to-customer transactions can easily become a huge repository of user's data [174]. They can sell users data to third party without users consent. To get organization's services,

users have no options; they are bound to give their details to them. Their day to day online activities can also be observed. But this is very risky for the safety point of view because users have left their privacy with their public freedoms [246].The collection, mining, reusing or selling of user's information to others has now become a big business. The user's sensitive information combined with an external datasets can disclose the new facts about users [175]. These in turn may be helped for the mining purpose but may breach privacy of an individual. Recent years, security concerns of user's data are expanding with the expansion of big data environment conceptualized as privileges of those users whose sensitive data is shared with others. The privacy and security of personal data have long been declared as primary human rights [172]. There is a vital requirement to safely gather, store, manage, share and analyse the huge data in order to determine the patterns and trends for improving the confidentiality, integrity and availability of data. The most essential security challenge in heterogeneous environment is that the data creator cannot control on who can access the data and location of the data [247]. A data creator can create both types of data: sensitive or non-sensitive. If data is not sensitive, then there is no concern for its security but the problem begins with the sensitive data. After the data is released from the creator side, the organizations collecting data can sell the same to the third party according to their need. Hence, awareness about preservation of the confidentiality of enormous amount of data is must [172].

In this chapter, the researcher has proposed big data privacy policies. These policies have been implemented in Hadoop. It is expected that these policies will secure and aware individuals for their security. The idea is dependent upon the users choice on whether disclose or not to disclose their information to the third party. It should be the user's choice. It contains complete mechanism to maintain the privacy and security of user's personal information. Protective privacy is a way to share sensitive information to ensure security against identity revelation of a user. With the help of proposed big data privacy policies, a user can avoid misuse of his information. These proposed policies address the privacy and security of user's data to reduce privacy risks. There have been various studies [60] [94] [175] discussing about the user's data misuse harming data creator‘s confidentiality. The researcher has mainly focused on data creation phase. If a user becomes aware of his data then he can protect his data himself. And through these policies a user can be aware about his data. Everyone is bound with

the organization's policy if one wants to access its services then it is important to obligate with these policies and have to provide his data to organizations. In the previous chapter, researcher has proposed lifecycle for Big Data and Phase Wise Security Threats. The researcher has shown the need and importance of data creation phase. This phase and threats on this phase are completely neglected by the researchers. These policies are implemented on this phase because a user can be securing his data at the time of creation. This time a user considers his data sensitive or non-sensitive. Those users who consider their data sensitive can demand for privacy. With the help of these policies, privacy loss can be reduced because currently users have no option to protect their data and they provide the information in exchange to get services.

4.2 Proposed Privacy Policies

In our daily life computer has become an essential part. Internet is used to improve the survival and to upgrade lifestyle. In addition, information security related cases have grown extremely in the last decade [61]. To survive among the growing risks, with the technical solutions the privacy policies should also be proposed. In recent years, security of user's data been strongly rooted in the rapidly increasing big data system conceptualized as the rights of people whose data is shared with others. Security and the privacy of personal data have long been seen as primary human rights [172]. The extended dependence on the pervasive interconnectivity of information and communication technologies (ICT) and big data have completely changed the behaviour of the attacker's on data vulnerabilities and their exploitation. The big data heterogeneous system has provided the ground for attackers for steal the data such as information theft, phishing, etc. [3] [41].

Today's digital user is no longer completely unknown, because of online communication and activities. A creator creates data that can be collected, combined and examined. It is always possible that information collected for one reason can be used for another reason without user's consent [172]. A data creator can create both types of data: sensitive and non-sensitive data. There is no worry with the non-sensitive data but the problem arises with sensitive data. If sensitive data is disclosed in public then, it can harm user's privacy. After disclosure of the sensitive data by the data creator, the organizations collecting the data may sell this data to unauthorised person according to their requirement [172]. Usually organizations provide a platform to get

their services at one click. Unfortunately, they demand for the private information in exchange. Now users are left with only two options; reject or accept the offered terms and conditions. Rejection will debar from the services and acceptance will compromise their privacy. The organizations impose their own policy to the users and the users have to accept the policy to avail the services. The current scenario of information flow is shown in figure 4.1 (a) and (b). A User creates information and pass on to data collector. A collector combines the entire data and reuse or sells it on its own terms and condition. Now the, critical issue arises of user's privacy. Once the information disclosed to others is always in danger [47].

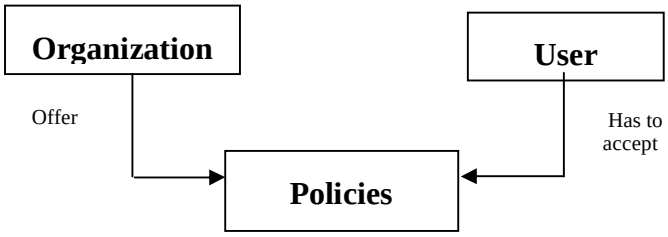

Figure 4.1(a): Existing Organization Policy

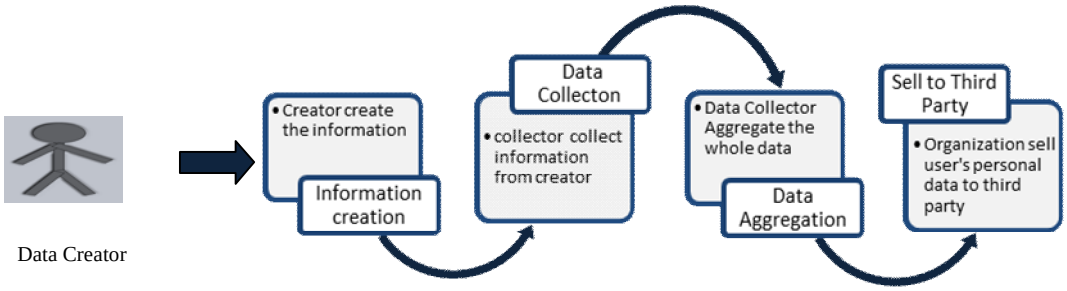

Figure 4.1(b): Existing Flow of Information from Data Creator to Third Party

On the contrary, while providing its services the user's security and privacy can easily be preserved with minor modification in the organizational policy. The idea is to ask the user's options that whether they want to reveal their information to third party or not. In a nutshell, there should be safe information flow from data creator to third party as presented in figure 4.2 (a) and (b). User's privacy can be maintained with a slight vigilance. If a user wants to access services of an organization, then they should provide a choice to optimizing the policies according to their requirement as shown in figure 4.2 (b). Users create information and provide to data collector. When submitting information, if the users consider their information sensitive then they may demand for

privacy. They can choose the option 'do not share information'; if they consider their information to be sensitive i.e. they cease to share their information to others.

On the other side, if users do not consider their information as sensitive then they will not opt privacy, while submitting the information to the organization (data collector). Now, when organization sells data to the third party then it cannot sell the data of those users who have demanded for privacy for their data. In this way, privacy of the users will be maintained. If an organization doesn't adhere to the privacy policies, then it will be violation of privacy policies and users will have the rights to take action against it. Table 4.1 shows an overview of the privacy policies for the users as well as organization.

4.2.1 Privacy Policy (PP)

a) Privacy Policy 1: This policy states that at the time of providing services, organizations must provide services with the option to share or not share their information. Policies should be clear and brief that clearly describes expectation from users. Here, share is a variable which can have two values {0, 1}, share=0 shows not authorised to share and share=1 show authorised to share.

$$share \in \{0,1\}$$

(I) If share= 1 then

Organization is authorized to share creators' data

(II) If share= 0 then

Organization is not authorized to share creators' data

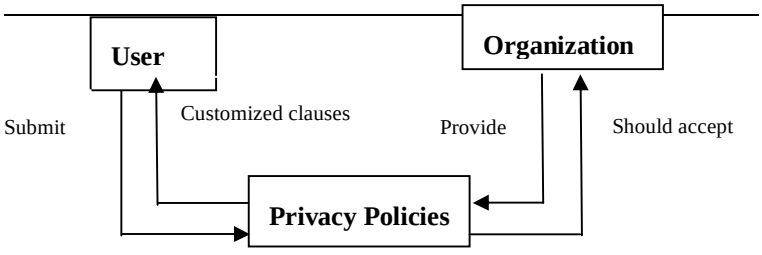

Figure 4.2(a): User/Creator policy

If an organization sells data of a user, then it must be clear in its policies. It should also state those circumstances where it is necessary for the services to share user's

information. It is up to the user to select ‘share’ or ‘don’t share’; it should depend on the user’s choice.

b) **Privacy Policy 2**: Organizations should present users with all their security approaches. Here R stands for rules.

R1: Privacy Policy must clearly define organization’s security strategies.

R2: There should be no hidden policies.

R3: Protection of information should be under the control of the organization.

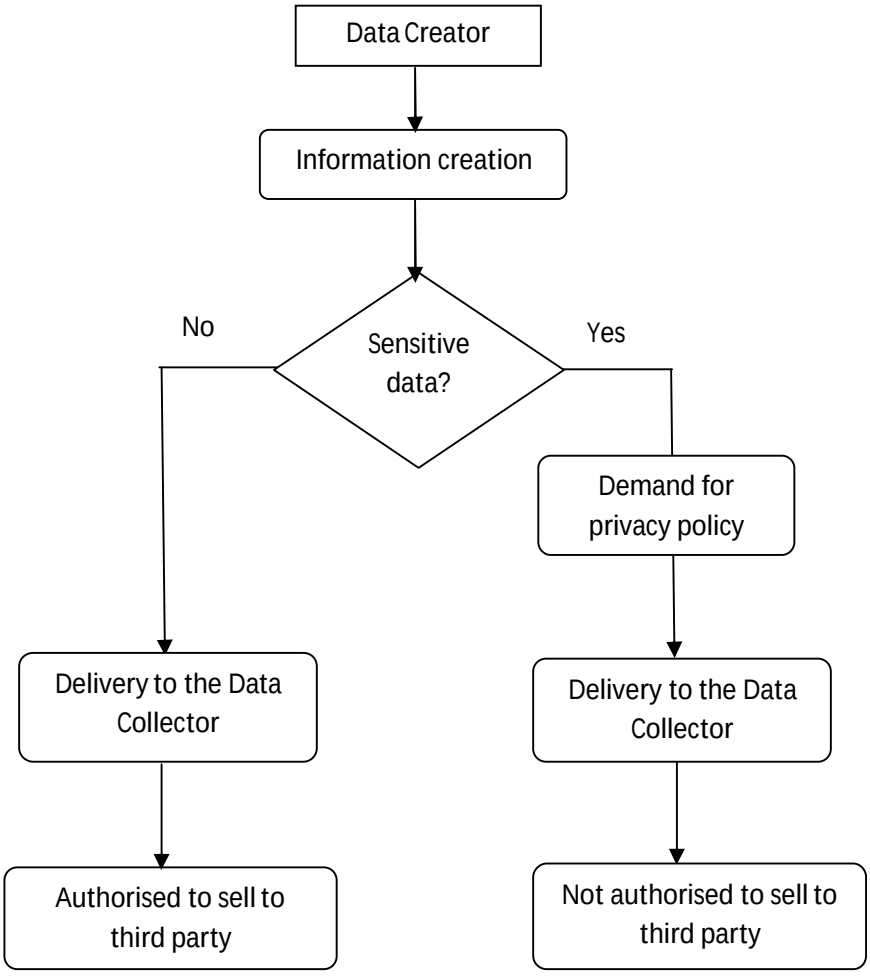

Figure 4.2 (b): Proposed Policy Process of Information Sharing

Table 4.1: Step by Step Process of Privacy Policies

1. Create configuration for Hbase

2. Create connection from cluster

3. Create Scan object to scan the Htable

4. Create ResultScanner

5. while (resultset==empty)

6. Get the Scanner object from the Htable using scan object which contains the content of HBase table.

8. for(each row obtained) do

(a) for (each cell<- rs (Result object)) do.

(b) Get the value of 'share' column from HBase table 'user' and column family 'PersonalData' into 'flag'

(c) if (flag==yes) then

(d) Print the data of the user (referred to as sharing of data)

(f) else if (flag== 0) then

(g) print "Organization is not authorized to share"

end if

end if

9. end for

10. end for

11. end while

Users should be conscious about the use of their information. Thus, organization has to make strong and fair policies. No hidden policies should be there that affect the user's privacy. User's demand must strictly be followed.

4.3 Implementation of Proposed Privacy Policy

Implementation is the process of defining a model, design, standard steps for doing something. It also states normal tasks that must be achieved to happen something truly. The proposed privacy policies for big data security have certain prescriptive steps in data creation phase. These steps are creating data, defining sensitive or non-sensitive data, and demand for privacy and performance evaluation. Implementation of the framework is significant to validate the utility and effectiveness of the proposed privacy policy. Implementation involves developing an organized and well-planned process for creating the data and aware about the use of the data. The proposed privacy policies have been implemented with the system specifications on Hadoop 2.5.2 (pre-built 32-bit i386-Linux native) and java version SDK Oracle Java 1.8.0 25 CPU Intel i3 M370 2.40 GHz, RAM 6 GB 1600 MHz. Hadoop is an open source framework which is used to handle, store, and analyse the huge amount of data [269]. This huge and distributed data cannot be managed by the traditional techniques. Hadoop parallel architecture supports handling hardware failure. It allocates files into huge blocks and distributes between nodes in the cluster. It has two main vital components; HDFS and MapReduce. HDFS is used for storage and MapReduce is used for processing.

Table 4.2: Attributes of Dataset

Attribute	Description
Name	Student's name
Contact No	Student's contact number (numeric)
Email_id	Email id of student
Gender	Student's Gender (binary: Female or male)
Age	Age of student (numeric: from 15 to 22)
School	School name
Address	Student's address (binary : urban or rural)
Fam_size	Family size (binary: ≤ 3 *or* > 3)
Guardians	student's guardian (nominal: mother, father or other)
Share	Share to agree data (binary: yes or no)

A student's dataset of UCI machine learning repository has been used for the experimental assessment of proposed privacy policy. The data has been taken from [248]. It has been prepared according to the need of the experiment. Attributes of dataset are shown in Table 4.2. The data set has 2000 students' records. On the basis of results, accuracy of proposed policy has been checked.

4.3.1 Result Analysis and Discussion

To calculate the usefulness of the proposed privacy policy (PP1), the researcher has uploaded the set of data in HDFS and using hbaseorg.apache.hadoop.hbase.mapreduce.ImportTsv–Dimporttsv package, the data has been imported from hdfs to hbase. It has been saved in hbase in the form of column and column-family. Then hbase has been activated using hbase shell start-hbase.sh command and the connection is established with java code by using the HBaseConfiguration.create function. After execution of the code, the data of the users has only been displayed who were agreed to share their information. The same is shown in Figure 4.5.

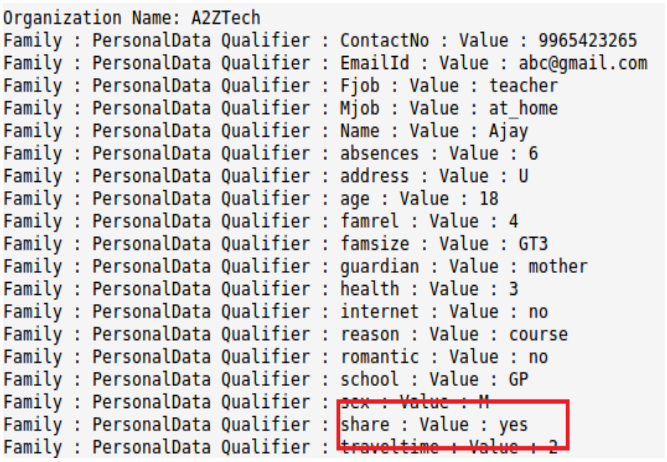

Figure 4.5: Records of Creator Who Agreed to Share

In contrast, in absence of any privacy policy if the same sequence of instructions is executed, record of all data creators are displayed as shown in Figure 4.6.

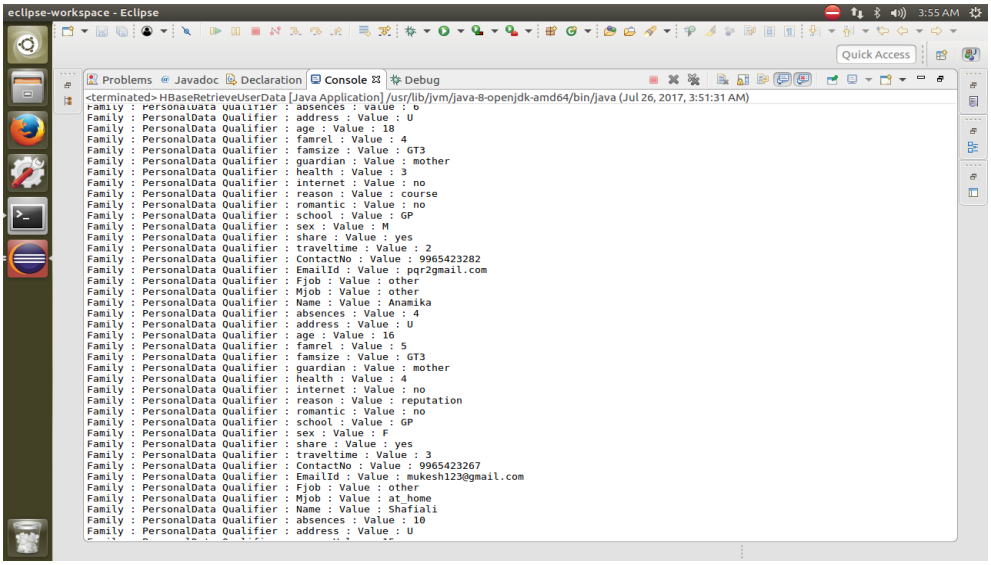

Figure 4.6: Records of All Creators

To calculate the usefulness of proposed privacy policy's, probability theory has been used and privacy loss has been compared in current scenario and in the proposed scheme. In the current scenario of information flow, 1 is the probability of privacy loss's (absolute privacy loss) because organizations don't provide services with the customized policies. But privacy loss can be reduced with the help of proposed privacy policies and therefore privacy can be preserved. A database that contains n users data {u1.....un}, where $n \in \mathbb{R}$(natural numbers) is proposed. For instance, five users u1……u5, with their respective share variable 0, 1,1,0,0 such that$share \in \{0,1\}$, where favourable results denote the number of users who have agreed to share their data. Then

Users	*Share*
u1	0
u2	1
u3	1
u4	0
u5	0

$$Probability\ of\ Privacy\ Loss = \frac{\text{Number of favourable outcomes}}{\text{Total number of outcomes}}$$

$$\text{Probability of Privacy loss } P(L) = \frac{2}{5}$$

$$P(L) = 0.4$$

From the result it is very clear that in the given example privacy loss is 0.4 (by the following the proposed policies) whereas value of privacy loss is 1. Thus, it can be easily concluded that the privacy loss can be decreased with the help of proposed privacy policies. Those users, who follow the proposed policies, have no risk of their data breach because organization has no rights to share their data.

4.3.2 Confidentiality and Privacy Assurance

The proposed policies enable a user to ensure their confidentiality and privacy. Organization are not authorised to sell the user's data to the other unauthorized organization. Hence, unauthorized access cannot be made to the user's data. A user dataset with the basics attributes, for example, U= {Name, Age, Sex, Email-id, Mobile No., DOB, Address, share} are the basic information of an individual. In the database there is another attribute i.e. Share. It is up to the user now that he/she can choose or not choose the share option. In the current scenario user's personal data is sold to others without user's consent. It is supposed that the proposed policies will reduce this behaviour.

4.4 Measures for Performance Evaluation

The performance of the proposed policies has been evaluated as follows:

4.4.1 Confusion Matrix and K-Fold Cross Validation

In this study, the proposed policies were calculated based on the accuracy measures discussed above with confusion matrix (Accuracy, Specificity, Sensitivity and Error) and K-fold cross validation. The results were achieved using 10 fold cross validation for every model, and for each time the test is based on average results obtained from datasets (10^{th} times). Table 4.2 shows the confusion matrix with K-fold cross validation in a tabular format. A confusion matrix is a representation of the results of classification [249]. Confusion matrix is shown in figure 4.7. In a two-class prediction problem, the upper left cell indicates the number of samples is classified as true positives, and lower right cell indicated the number of samples is classified as they were actually false (true false). Lower left and upper right (other two cells) indicates that the number of samples incorrectly. Specially, the lower left cell indicate that the number of samples is classified as false (false negatives), and the upper right cell denoting the number of

samples classified as true while they actually were false (false positives). Once the matrix was constructed, the accuracy, sensitivity and specificity of every section were calculated using the respective formulas presented in the above section [250].

In this study, the researcher has used three performance measures Accuracy, Sensitivity, specificity and error; these are defined as follows:

$$\text{Accuracy} = \frac{\text{true positive+true negative}}{\text{true positive+false postive+false negative+true negative}} \quad (1)$$

$$\text{Senstivity} = \frac{\text{true positive}}{\text{true positive+false negative}} \quad (2)$$

$$\text{Specificity} = \frac{\text{true positive}}{\text{false positive+true negative}} \quad (3)$$

$$\text{Error} = \frac{\text{false positive+false negative}}{\text{true positive+false postive+false negative+true negative}} \quad (4)$$

In stratified k-fold cross-validation, the folds are made in a way that they contain around the similar proportion of predictor names as the original dataset. Observed studies demonstrated that stratified cross-validation have a tendency to produce comparison result with lower bias and lower variance when compared to standard k-fold cross-validation. In this current study, 10-fold cross validation method is used to calculate the performance of classifiers. Observational studies demonstrated that 10 appear to be a best possible number of folds [249]. In 10-fold cross validation the whole dataset is distributed into 10 mutually exclusive subsets with around the similar class dissemination as the original dataset. Classifier performance tested once in every fold that is produced from the combined data of the remaining nine folds, leading to 10 independent performances.

$$CVA = \sum_{i=1}^{f} Accuracy \quad (5)$$

Here CVA stands for cross validation accuracy, f is the number of folds used, and measure the accuracy of every fold.

Since the CVA would rely upon the random task of the individual cases into k different folds, a typical practice is to stratify the folds themselves.

		Yes	No
	Yes	**True Positive**	**False Positive**
	No	**False Negative**	**True Negative**

Figure 4.7: Confusion Matrix

Table 4.3: Confusion Matrix with 10 Fold-Cross Validation

Fold No	Confusion Matrix		Accuracy	Sensitivity	Error
1	13	7	0.725	0.65	0.28
	4	16			
2	16	4	0.825	0.80	0.35
	3	17			
3	19	1	0.85	0.95	0.15
	5	15			
4	15	5	0.82	0.75	0.18
	2	18			
5	16	4	0.731	0.80	0.25
	6	17			
6	12	8	0.775	0.60	0.23
	1	19			

7	18	2	0.875	0.90	0.13
	3	17			
8	17	3	0.775	0.85	0.23
	6	14			
9	13	7	0.63	0.65	0.38
	8	12			
10	13	7	0.74	0.80	0.25
	7	13			
Mean			**0.78**	**0.70**	**0.22**

4.5 Comparison

The proposed policies for preserving confidentiality in big data have been compared with a recent approach for maintaining privacy of big data [175]. Comparison has been shown in table 4.4. During comparison, the critical review of the previous available work has yielded the following observations:

- The previous work has talked about the fundamental rights of privacy and data protection. But it has not mentioned the user's (creator) fundamental rights. Their work is not providing the transparency to user. It doesn't consider user's need of security and privacy. Our proposed policies provided the clear transparency to users, with the help of these policies user can choose or not choose the share option.
- They only mentioned that big data put pressure on important legal principles such as purposes limitation and data minimization but they didn't mentioned how to minimize the data or how to limit the purpose. With the help of these privacy policies a user can puts pressure on core legal framework, because if a user opts not to share his/her data then any organization cannot sell his/her data. Any organisation is bound to agree to user's privacy preferences.
- They have mentioned that less data as possible is collected but how it will be possible without providing the privacy policy on creation phase. They have completely neglected the data creation phase and threats on it and didn't provide any implemented policy which can secure the user's information. But the proposed policies show the importance of data creation phase and these policies have been

implemented at the time of data creation. Through these polices, the collectors will be bound from user's demand of privacy.

- The previous work completely ignore the creation phase and has not stated any mechanism or policy to secure the data on the further phase but through the proposed privacy polices further phases can be secure.
- They have mentioned some baseline on analysis phase that government works legally but government also not providing the customized policy to user.

Table 4.4: Comparison Table

Techniques	Transparency	Awareness	Showed Importance of Data Creation	Implementation
Proposed Policy	YES	YES	YES	YES
Broeders et.al [175]	NO	NO	NO	NO

4.6 Conclusion

The task of ensuring information security and protecting privacy has become harder as information is multiplied and shared more widely around the world. Information about individuals' online activity is exposed to analysis. It raises concerns about security and privacy issues and loss of control. The disclosure or publishing of the data gathered from data creator by the data collector such that user's sensitive information is preserved as well as the published data is useful is highly recommended. The major challenge raised by big data is how to classify sensitive information which is stored in unstructured format. Protection of user's privacy has become one of the main issues in big data. Privacy of human is generally challenged by the advancement of technology. In this interconnected environment, there is no such kind of privacy policy that can protect the individual's privacy or prevent the information from unauthorised access. Organizations sell data to third party without the user consent and form where the data goes, no one is aware about it.

Therefore, it is the need of time to finalise and implement privacy policies to avoid illegal access and misuse of user's information. In this chapter, researcher has

implemented the privacy policies have been proposed in chapter-3 for big data security. The researcher has also claimed that by choosing privacy policy, privacy risk can be reduced. So here researcher has developed such policies that prevent the unauthorised access of user's information. These policies have been implemented in Hadoop, a suitable tool to manage all the big data's issues. For the validation of the proposed work confusion matrix and k-fold cross validation has been used. It is necessary to prove the proposed approach to make it socially acceptable. In this chapter the proposed work has been validated.

CHAPTER 5: A NOVEL SCHEME FOR PREVENTION OF INFORMATION LEAKAGE IN BIG DATA

"When flimsy cyber defense fails, Format Preserving Encryption prevails"

- ***James Scott***

5.1 Background

Big data is produced by users from various sources on daily basis. Internet news portals, e-commerce applications, and web search are providing a different platform where users can interchange the information to obtain the benefits. From these sources, data is collected and processed by several organizations such as medical centres and other companies etc. However, it is challenging to keep the records of personal information safe from unauthorized access. There are several security and privacy issues that disrupt the adoption of big data. Figure 5.1 illustrate the some issues on sensitive data. One of the emerging concerns in the field of big data is to maintain safety and privacy of patient's sensitive health data when they are stored in a server [136]. It is obvious that data is created by the users and collected by various organizations.

In big data life cycle, the first two phases are more susceptible in terms of privacy and data leakage i.e. data creation phase and data collection phase. Literature review suggests that securing these two phases will reduce the burden for the next phases [63]. In a nutshell, encrypting the data at collection phase according to user's preferences will strengthen secure mining. Mining of encrypted data cannot be possible because data is unreadable form [251].

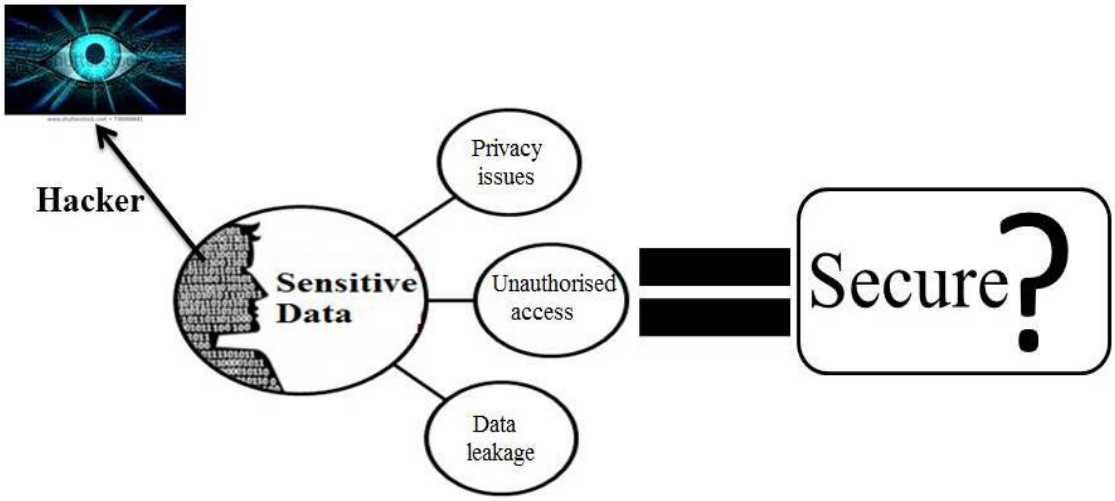

Figure 5.1: Security and Privacy Issues

To mitigate the privacy risks inherent in storing and computing sensitive data, cryptography offers a potential solution in the form of encryption, which metaphorically locks the data in a "box" requiring a key to open. Traditional encryption systems lock data down in a way which makes it impossible to use, or compute on, in encrypted form [252]. In chapter 3 and chapter 4, the researcher has proposed security mechanisms for securing data creation phase. In the current chapter, the researcher has proposed a mechanism to secure data collection phase. In this chapter, the approach proposed by the researcher pinpoints only health data. The approach will help to prevent patient data from unauthorised access as well as misuse of data. Now a day's hospitals, clinics, research institutes, and companies handling medical data are all facing with the mutual problem of security of the health data. Healthcare data is universally considered sensitive (and confidential), so it might seem that the categorisation of less sensitive data is relatively unimportant for medical data research [180]. Sensitive health information (SHI) is a collection of health information related to patients to allow efficient, relevant and worldwide sharing of health information. Due to the sensitivity of health related information, providing safe storage and authorised access of SHI is the core issue in today's systems. Patients lose physical control to their own personal health data as soon as it comes under the control of the servers. As more of our sensitive data comes into various types of social sites, merchants, employer, health care provider etc. so there is a big opportunity that an invader can disclose the data and piece together user's information. The Personal Health Information (PHI) of a patient could be leaked if an insider in the provider's organization misbehaves, due to the high value of the

Personal Health Information (PHI). As a famous incident, a Department of Veterans Affairs database containing sensitive PHI of 26.5 million military veterans, including their social security numbers and health problems was stolen by an employee who took the data home without authorization [136]. So, this chapter introduces a novel scheme for prevention of information leakage in big data which prevents medical data from data leakage.

In this heterogeneous environment, unauthorized users can easily access the information of medical data beyond their right and privileges. Now there is an essential requirement to develop a feasible and promising method which can improve and enhance the security of sensitive health information. The main contributions of the proposed approach are as follows:

- The proposed approach provides the confidentiality of health data in storage. Hence, an unauthorised user will not be able to read patients' files without the patient's consent.
- The proposed approach particularly addresses the issue of accessibility of health record management systems by logically dividing the system into the public and personal domain, which considers both personal and professional SHI users.
- The proposed approach is patient-centric as patients can control their data in big data environment.
- With the help of proposed approach data leakage issues can be prevented successfully.
- The proposed approach minimizes data leak by guaranteeing that the generated private key is efficient and unbreakable.
- To measure the quality and effectiveness of the proposed approach, researcher performed experimental and statistical evaluations with the different algorithms and compare the results of proposed algorithm with the existing works DES, AES, ELGAMAL and RSA.

In the recent years, sensitive health information (SHI) has emerged as a patient-centric model of sensitive information exchange. An SHI service allows a patient to create, manage, and control his/her personal health data in one place through the web. This has made the storage, retrieval, and sharing of the medical information more efficient. Especially, each patient is promised full control of their medical records. They can also

share their health data with a wide range of users, including healthcare providers, family members or friends. Due to the high cost of building and maintaining specialized data centres, many SHI services are outsourced to or provided by third-party service providers. Recently, Microsoft HealthVault.1 has been proposed for storing SHIs in cloud computing [100, 254, and 99]. Personal data of patient's can be collected, analysed and redistributed by the other organizations for various health purposes. Sharing of health data plays an essential role in improving the quality of health services [254]. With the heterogeneous environment of big data, the internet connectivity has provided users the ability to utilize scalable distributing. However, in this environment, data resides in sharing mode. Users have no idea about the actual storage location of their data as well as they aren't aware of the other sources collecting and utilizing the data for their own purposes [66].

In this sharing mode, there are so many privacy and security issues which may impede its broader adoption. When SHI is stored in an untrusted third party server, it may raise the privacy issues for the users. At one side, Health Insurance Portability and Accountability Act (HIPPA) for personal information protection and electronic documents act for medical data are ensuring the sufficient protection for such data, but on the other, various malicious acts are often happening at the third-party storage server which may expose the SHI data [254, 255]. In this environment, there may be unauthorized person who can easily access the information of health data beyond their right and privileges. Now there is an essential requirement to develop a feasible and promising method which can improve and enhance the security of SHI. Health data can be protected in two ways (1) protection by data policy and (2) protection by cryptography [66]. Firstly, the SHI files should be encrypted and the owner should have control on who can access/modify their files. Secondly, the SHI owners have the option to customize the policy of the hospital if they don't want to share their data with the third party. SHI should only be available to those authorized users who have the decryption key. This will maintain confidentiality of SHI.

This chapter considers the issue of privacy protection on health data where the patients can control their own SHI data. It is very important to have access control (mechanisms) for data in semi-honest environments because patient's data can be leaked at any level. The proposed research only concentrates on sharing SHI patient-centric health data in

semi-honest environments and focus on addressing the crucial and sensitive key issues. For the same, RSA for the encryption has been adopted. The researcher has proposed a novel RSA based framework for sharing secure patient Sensitive Health Information (SHI) in big data environment. To improve the security of health records the researcher has generated the key from the fusion of doctor and patient's password. A mechanism has been proposed for key generation so that SHI can specify personalized fine-grained key access policies during encryption. In a personal domain, the patient can directly assign the access privileges for close users and encrypt his SHI file. This will help patients to have full control and privacy of their own SHI records. To accomplish patient centric SHI sharing, patient controlled read/write access is the basic aim for any electronic health data as a system, indicated by Mandl et al. [256] in 2001. The system's security and performance requirements are summarized in brief:

- Data confidentiality: Unauthorized user, who does not have key access privileges, should be banned from decrypting an SHI.
- Data Integrity: unauthorized user must be prohibited from getting modify-access to SHI file, while legitimate contributors should get to the server with accountability.
- The information access policies must be flexible, i.e., dynamic changes may be permitted in predefined policies.
- Scalability, efficiency, and applicability: The SHI system should support users of both the public and personal domains. However the public domain users may be unexpected and large in size, in terms of complexity in key management, storage, and computation, the system should be highly scalable. Furthermore, the data owner's efforts in the user's management and keys should be reduced.

5.2 Overview of the System Model

The proposed system model offers a secure patient focused SHI and access control with an effective key management approach. The key is only used by the authorized person; be it public or private. For secure patient centric SHI, the proposed model involves five main categories of participants such as personal domain, public domain, data owner and big data server as presented in Figure 5.2.

- Patient (Data Owner): The patient has his own SHI file created by him and it is stored in big data server such as name, age, gender, history, disease etc.

- Personal Domain: Each patient has its own private domain, where he/she can rely on his closet people like family members and fast friends. So patient delivers access rights according to his needs. The personal domain is rottenly not a big circle; it is quiet small. So it can manage easily and also decreases the load on the patient.
- Public domain: Public domain is the domain where users have no access privileges on the patient's SHI until the patient allowed them. These users can be other hospitals, insurance services etc.
- Administrator: Every hospital is connected to an administrator who offers the registration ID to patient. First time, when a patient visits to the hospital, he registers himself to take health services. He creates his profile and obtain a private key from the administrator i.e. hospital. This is the one time private key, which is updated every time.
- Big data server: It is a server where a huge size of data can be stored, managed and analysed. Here, the SHI of patient is stored.

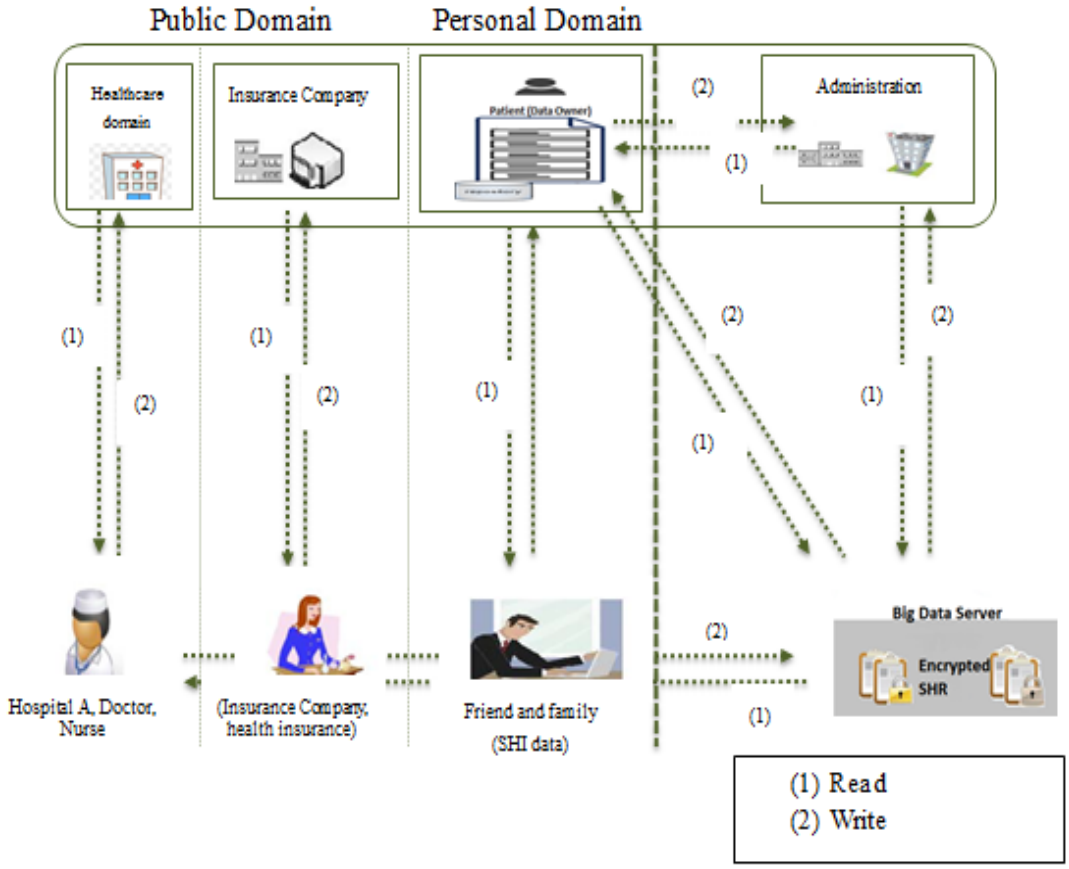

Figure 5.2: System Model

5.3 Proposed Encryption Algorithm

According to the societies of health and medical institutions, a patient's record may contain sensitive health information (SHI). This SHI is created by the patient (data owner) who has full control on his data. SHI is digital information which can be shared with others medical institution. This data serves as the data source to other medical institution like hospitals, labs etc. [100]. Any kind of leakage or disclosure or access by unauthorized party which has no rights of accessing the SHI, can be a big privacy concern of the patient. The proposed encryption approach is an enhanced version of RSA encryption. The proposed algorithm provides security against information leakage; hence with the help of this approach, researcher can prevent the patient's SHI from leakage or disclosure. The proposed algorithm is termed as Information Leakage Prevention Scheme (ILPS). Thus to ensure, that the SHI sharing data should be approved by the patient to prevent data leakage.

5.4 Algorithm of Information Leakage Prevention Scheme (ILPS)

The proposed ILPS algorithm comprises of the following steps which are illustrated below: It contains three algorithms: (1) Key Generation, (2) Encryption, and (3) Decryption.

Input:	**Plain text file**
Output:	**Encrypted file**
Algorithm1:	**Key Generation**
Step 1.1:	X=0, n=P1.length // Where n is the length of password and P1 password of patient
Step 1.2:	for each level i from 1 to n do
Step 1.3:	x = x + r // Select the random number r
Step 1.4:	end for
Step 1.5:	p =getPrimeNumber(x)
Step 1.6:	Repeat step 1 to step 4.
Step 1.7:	q =getPrimeNumber(x)
Step 1.8:	Call function getPrimeNumber():

	(a) if(isPrime(x))
	(b) return x
	(c) Else
	(d) x = x-1
	(e) return getPrimeNumber (x)
	(f) End if
Step 1.9:	Select two different prime numbers p and q and Calculate N=p*q
Step 1.10:	phi=(p-1)*(q-1), where phi is the Euler function
Step 1.11:	Output Public Key PK = {n} and Private Key PR = {p; q; r; s}
Algorithm 2:	**Encryption Algorithm**
Step 2.1:	Selects the file to be encrypted
Step 2.2:	Encrypt a file F
Step 2.3:	Generate cipher-text ci
Algorithm 3:	**Decryption Algorithm**
Step 3.1:	Selects the file to be decrypted
Step 3.2:	Decrypt a file F
Step 3.3:	Check the user is from which domain : Public and Private domain
Step 3.4:	Enter the private key, if it matches, the cipher text ci is decrypted.

5.5 Implementation of the Proposed Algorithm

To implement the prevention of data leakage, a novel algorithm has been imposed. The algorithm is developed by using the modifying Rivest- Shamir- Adleman (RSA). The reason for choosing RSA is because it is strong and more secure in comparison to the other existing algorithms [254]. The performance is calculated by encrypting and decrypting different file sizes. The performance of the proposed approach is also compared with the existing approaches such as DES, AES, ELGAMAL and RSA.

Unfortunately, in the patient-focused atmosphere, it does not have control over data leakage and the patient isn't aware about the exchange of information. The information has been transferred without the patient's approval [118].

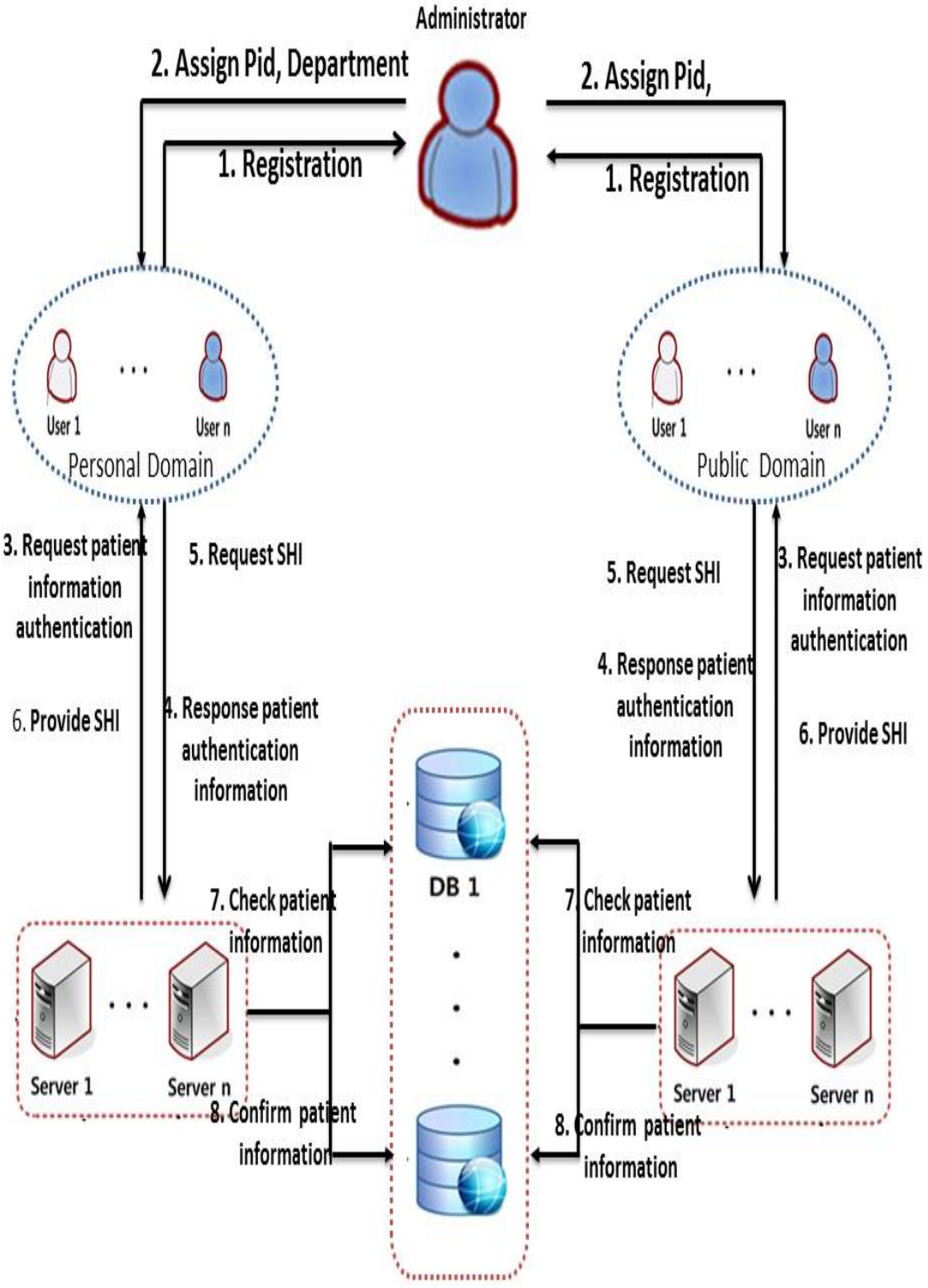

Figure 5.3 (a): Flow Diagram of the Proposed Approach

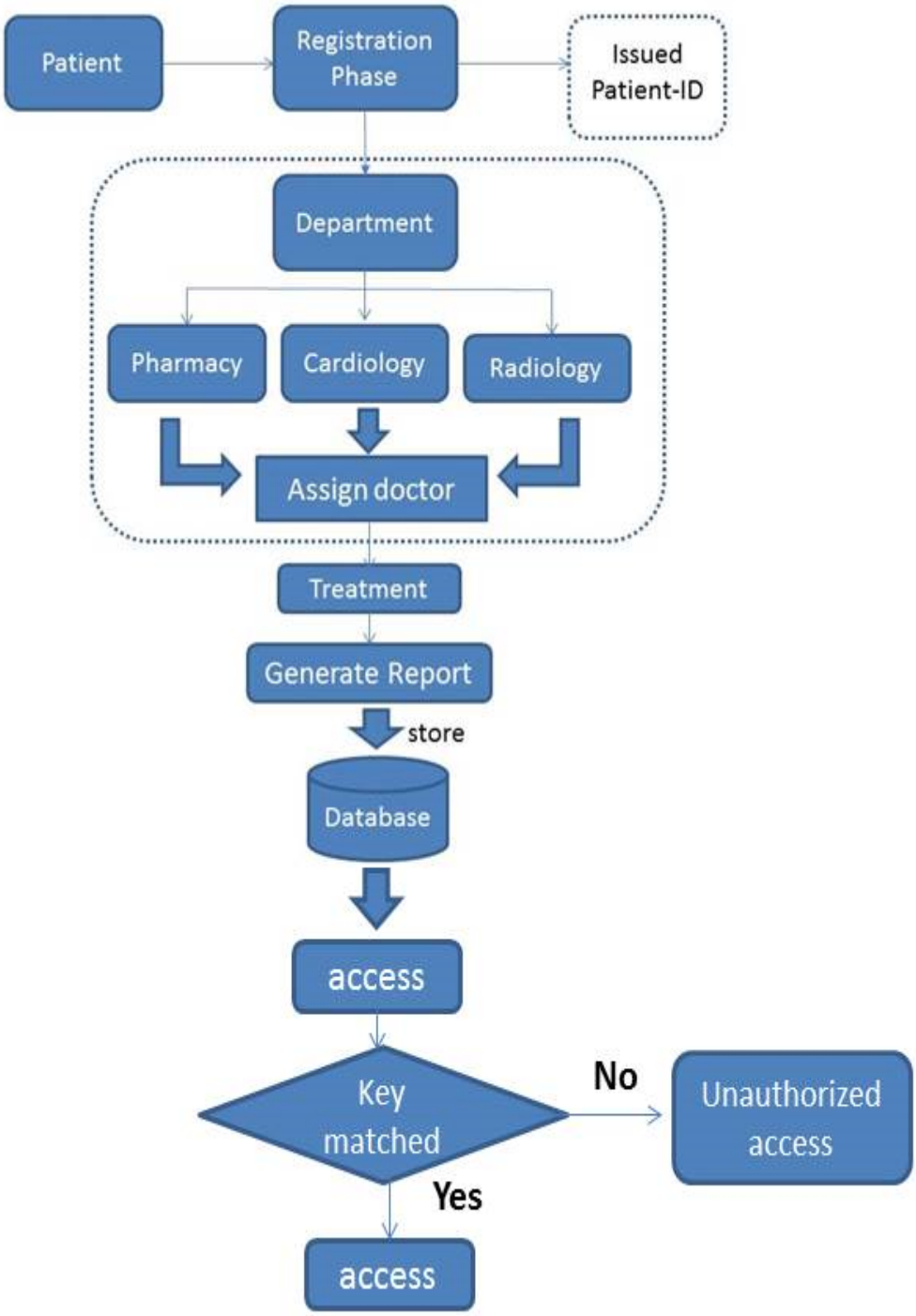

Figure 5.3 (b): Flow Diagram of the Proposed Approach

Figure 5.3 (a) and 5.3 (b) shows the complete process of flow of accessing SHI services in the proposed system model. This system model has two domains: public and private. A patient comes in personal domain and doctor come in public domain. Both the domain users have to register himself in administrator, where they get the patient id (P_id), doctor id (D_id) and department from their requirements. Here the researcher

considers the semi-honest environment. The administrator collects both the password and generate private key from it. A patient enters his personal and medical information such as “patient profile,” “medical history” on big data server. The patient is asked to give the information of the individuals from his personal domain, at the time of registration. Furthermore, he knows about the people who will access his SHI from the public domain. In this model, there are two ways to circulate a secret key. First of all, to access a SHI patient can specify the access privileges from his/her personal domain. Second, a user from the public domain can get the secret key by sending a request to patient. The SHI is stored in big data server and for its authorised use; it is encrypted by a domain based access policy. Patient or authorized user can decrypt the SHI file with private key. Only the authorised users who have access privileges or those which are stated in public and private domain by the patient can decrypt the SHI file. At the time of integrity verification, the big data server confirms who can change the SHI file. Unauthorized users cannot decrypt or alter the SHI file. The server always confirms the patient’s information from the server and from the database also.

5.5.1 The Patient’s SHI Structure

In this system model, the authors believe that a registered patient has an SHI with a SHI _id (unique identification). The SHI’s data attributes may be classified as personal information, Hospital staff, health records, test reports, and insurance information etc. as shown in figure 5.4. SHI is managed in a hierarchical manner for efficient access based decryption, where authorized users are aware of this patient’s SHI. For instance, in this model, here are two domains: public and private. If doctor has a private key for their attributes from the database set S = {doctor, surgeon, hospital A}. Then they can decrypt a cipher text with an access structure A = (doctor ∧ hospital A). The researcher says that set S satisfies the access structure A and shows it as S ∈ A. Someone from another domain who is not authorized to access the data, cannot obtain the private key. The overview of the proposed model has been shown by the help of an example. Suppose, Bob (patient), is a SHI owner of hospital A. During his first visit to the hospital, he is supposed to create his SHI file F1. Apart from Bob's general and health related information, the other authorized person also included in his SHI. The file is stored in the server and is encrypted with ILPS. Accessibility of SHI is possible only by

the key generation policy P. The key policy may be dependent on system's recommended settings, or Bob's (patient) own preference. It seems like

P: (Physician ∧ internal medicine ∧ hospital A) → read & write (1)

P: (Technician ∧ Nurse ∧ hospital A) → read & write (2)

P: (Doctor ∧ Patient ∧ Hospital A) → read & write (3)

P: (family member ∧ Patient ∧ Hospital A) → read (4)

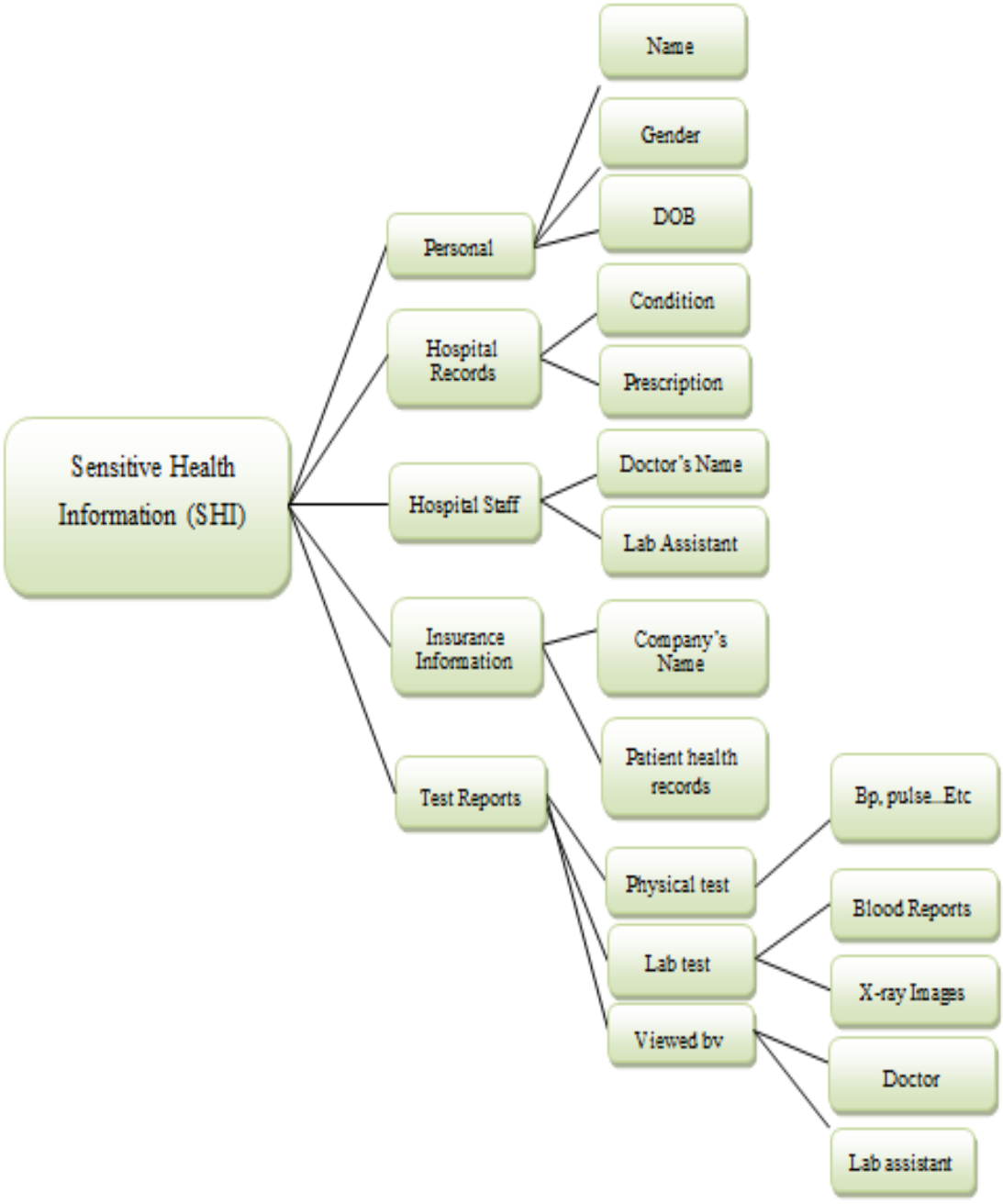

Figure 5.4: Structure of a Patient's SHI

Every patient's SHI is a very sensitive health data. In the proposed model, the patient is aware that as to who can decrypt or retrieve his data. When a patient generates his data, a doctor, a nurse, a technician, family members and close friends are involved. This model can stop the SHI data form leakage and maintain its privacy and confidentiality. In this approach, the key will be created by the patient's and doctors' password and patient's password is compulsory to obtain the decryption key for the access of SHI, after the treatment. As the authors discussed in this chapter, the proposed approach is a multi-users and multi-domains.

5.6 Security Analysis

In this section, the researcher has analysed the security of the proposed approach against internal and external attacks and chosen cipher text attacks.

5.6.1 Definition 1 (Semantic Security): Semantic security captures our intuition that a cipher text is given; the internal or external attackers learn nothing about the encrypted text. Thus, authors can say that it is semantically protected.

5.6.2 Definition 2 (Data confidentiality): If there is a polynomial time attack, an intruder cannot learn anything about the given retrieve encrypted data and the corresponding encrypted secret key. The proposed approach has data confidentiality. In this section, authors examine the security of proposed SHI sharing solutions. The IA based authentication provides cryptographic security against unauthorized access SHI of patient. Thus in proposed approach, it is practical to make sure that the patient's SHI confidentiality is obtained.

5.6.3 Theorem 1: The proposed scheme of ILPS is semantically safe against insider or outsider attacks according to Definition 1.

Proof: The authors may prove that, from the cipher text an attacker cannot get or know anything. Consider the key generation and the encryption process: the patient chooses two large primes p, and q from the passwords of patient's and doctor's, generates private and public key pairs PR = {p, q, d, e} and PK= {n} and then encrypts the SHI file with public key, h = r mod N, li=he mod N and ci = li⊕mi. Analyse that N is integer; an attacker (internal or external) can see only the cipher text ci. Observing that

N factoring is challenging, and the least important bit ln of ln+1 mod N is concurrently secure. Hence, the attacker can guess the pseudo-random bits h, 1≤e≤h. formally, if the N factorization problem is difficult, then the proposed scheme is semantically secure against internal or external attacks.

5.6.4 Theorem 2: According to definitions 2, the proposed scheme satisfies the data privacy. The enhanced ILPS novel approach guarantees the SHI's data confidentiality against the unauthorized users while preserving the collusion resistance against users.

Proof: Here, the researcher proves that privacy of data against cipher text attack. In this attack, the attacker can get temporary access on encrypted files and attempt to decrypt it. This methodology maintains the same cost of encryption and decryption and provides same security against security attacks. Protection of information by proposed approach against cipher text attack can be demonstrated as follows: Let N is the module in RSA i.e. N=pq where p and q are large primes, both are generated from the unique passwords of different size. Let the file of the patient hold a composite number N whose factors p and q are not known. To encrypt the message mi select a random h, then patient calculates ci = li⊕mi and li=he mod N. To decrypt the message mi, the user calculate the mi=hd mod N. The assumed h = r mod N is difficult for an intruder to compute random seed li to access the file. If so, then anyone can make extremely efficient encryption scheme. Therefore, it is safe against the cipher text attack. So, because of security strength of proposed encryption approach against internal or external attacks and cipher text attack from theorems 1 and 2, data privacy is extremely ensured.

5.7 Performance Evaluation

In this section, the researcher has proposed an experimental setup and experimental results of symmetric and asymmetric algorithms has been shown. Performance of any proposed approach is the main requirement. In addition, system consistency is also a key factor. To make more consistent system, longer database is required. Researcher has tried to found the main obstacle which degrades the system performance and found solution to overcome from that. In big data, time is the main factor to evaluate the system performance. The results obtained by the experiments are used for the analysis of system performance. The main objective of this research is to find the approach which is more robust, faster against other existing approaches. In order to find out

whether the proposed approach is robust or not, system performance of available encryption algorithm including DES, AES, ELGAMAL, RSA and proposed approach has been computed [257-258]. The parameters for performance evaluation include encryption time, decryption time and throughput.

5.7.1 Encryption Time: It is the time taken by the algorithm to convert plain text into cipher text. It depends on the key size and the block size of simple text. In this experiment, the researcher has calculated encryption time in seconds. Figure 5.5 demonstrates the encryption time of DES, AES, ELGAMAL, RSA, and ILPS on different file sizes. The figure clearly shows that the encryption time of ELGAMAL algorithm is taking longer than the other approaches such as AES, DES, and RSA. The proposed ILPS approach is faster in comparison to other approaches.

5.7.2 Decryption Time: The time taken to convert cipher text into its original text is called the description time, lower description time indicated better system performance. In the experiment, the researcher has calculated decryption time in seconds. Figure 5.6 represents the decryption time of DES, AES, ELGAMAL, RSA, and ILPS on different file sizes. The proposed ILPS's decryption time is far less than the existing approaches including DES, AES, ELGAMAL, and RSA. Hence, the proposed approach will help to improve system performance.

5.7.3 Throughput: The throughput of an encryption algorithm is obtained by dividing the entire simple text (in bytes) by the complete execution time for an algorithm.

5.8 Comparison between Proposed Scheme and Existing Schemes

In this section, the researcher has compared the proposed scheme ILPS with the available schemes [257] [258] on the basis of throughput for the diverse file sizes. The researcher use dataset of real-world [259] to evaluate the proposed solution. All experiments have been performed on personal computer with Hadoop 2.5.2 (pre-built i386-Linux native and java version SDK Oracle Java 1.8.0 25 (8u25) having the system specifications CPU Intel i7-7700, 4.2 GHz CPU with 8 GB RAM and Microsoft Windows 7 as the operating system. In the experimentation the, calculation costs are analysed independently. It indicates that the proposed scheme is more scalable than the current approaches [257] [258]. The proposed approach uses key mechanism of RSA also where the intruder cannot guess the cipher text. Hence, it can oppose a semantic

attack. Generally, compared with the current work, the proposed scheme attains higher security in key management. To understand the system's behaviour in a real situation, the researcher has considered different files sizes. More specifically, the researcher considered various file sizes of 32 KB, 126 KB, 200 KB, 246 KB, and 280 KB. The system performance is calculated by the computation cost's efficiency of proposed approach. The researcher has conducted the experiments to calculate computing time for the encryption or decryption of various SHI file size.

Table -5.1(a): Encryption Time for Different Size of Text Files

File Size (KB)	DES (msec)	AES (msec)	ELGAMAL (msec)	RSA (msec)	Proposed Algorithm ILPS (msec)
32	270	150	450	130	125
126	830	460	1030	520	492
200	1190	720	1410	740	723
246	1440	950	1750	1110	798
280	1670	1120	1830	1390	825
Total Encryption Time	5400	3400	6470	3890	2963
Throughput(KB/msec)	0.17	0.26	0.14	0.23	0.3

Encryption, decryption time and throughput of proposed ILPS have been shown in Table -5.1(a) and Table -5.2(b) respectively. Performance of ILPS has been compared with existing encryption techniques such as DES, AES, ELGAMAL [257], and RSA [258] with text files of different sizes. These are strong encryption methods but problem with AES is that for both process encryption and decryption, it uses the same key. Hence, it suffers security risks of key interchange which is resolved by RSA [130]. In the proposed scheme the permission to upload and retrieve the patient's SHI is controlled by the administrator and the patient. To decrypt the encrypted SHI, the related private key pk, which is encrypted by using the passwords of patient and doctor should be obtained. Only doctor's password is not enough to get the SHI data, patient's

password is also mandatory. In semi-honest model, doctors cannot disclose the SHI without the patient's consent. Hence the proposed approach is under the control of the patient.

Table -5.1(b): Decryption Time for Different Size of Text Files

File Size (KB)	DES (msec)	AES (msec)	ELGAMAL (msec)	RSA (msec)	Proposed Algorithm ILPS (msec)
32	440	150	430	150	145
126	650	440	850	430	404
200	850	630	1130	660	610
246	1230	830	1300	930	867
280	1450	1110	1640	1230	897
Total Encryption Time	4620	3160	5350	3400	2923
Throughput (KB/msec)	0.20	0.28	0.17	0.26	0.31

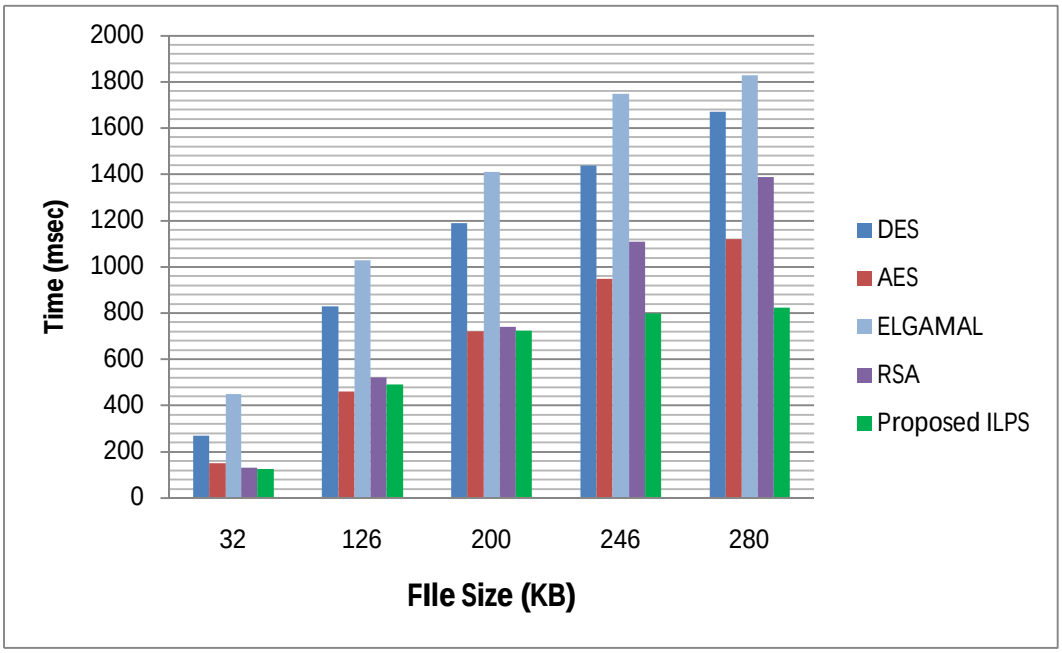

Figure 5.5: Encryption Time of Existing Approaches and Proposed Approach

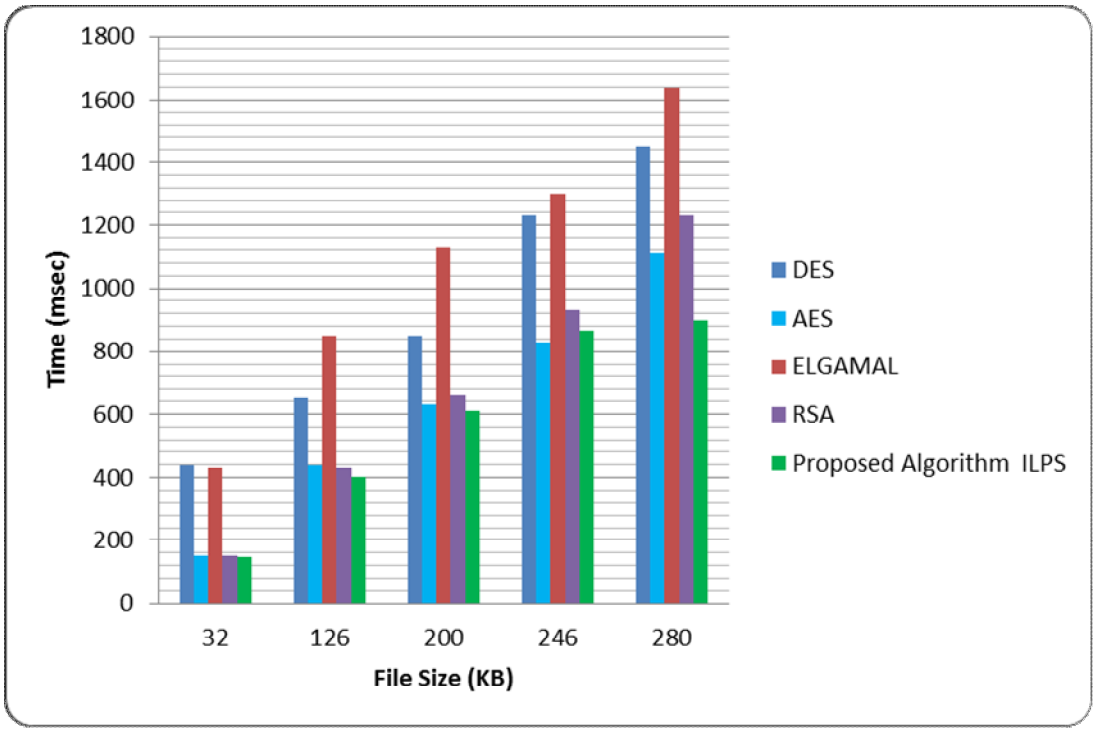

Figure 5.6: Decryption Time of Existing Approaches and Proposed Approach

The proposed approach mainly addresses the accessibility in medical record management systems by logically dividing the system into personal and public domain, which considers both SHI users personal and professional domain. It indicates that the proposed scheme is more scalable than the available approaches. The proposed approach uses key mechanism of RSA method where the intruder cannot guess the cipher text. Hence, it can oppose a semantic attack. Performance evaluation has been compared with the encryption, decryption time, throughput of the proposed approach.

Table 5.2 (a): Encryption Time of RSA and Proposed Approach

File Size	RSA (Min)	ILPS (Min)
64 MB	1.86	1.25
128 MB	2.75	2.06
256 MB	5.56	5.12
512 MB	9.38	8.96
1024 MB	17.03	15.63
Total Encryption Time	36.58	33.02
Throughput of Encryption (MB/Minutes)	54.24	60.09

To analyse the usefulness of the proposed ILPS in Hadoop environment, simulations has been performed with different size's random text files from MB to GB (64MB, 128MB, 512MB, 256MB, and 1GB). Performance of ILPS has been compared with RSA with different sizes of text files. The criteria for comparison have been taken as encryption and decryption time, throughput of encryption and decryption. As shown in table-5.2 (a), RSA the existing RSA took 17.03 minutes but the proposed technique took only 15.63 minutes to encrypt 1 GB file. The throughput of encryption and decryption is depicted in figure 5.7 and figure 5.8. It clearly demonstrates that the proposed algorithm yields better results in comparison to the existing cryptographic approaches. (As RSA, has already been proved to be better than the existing ones [130]).

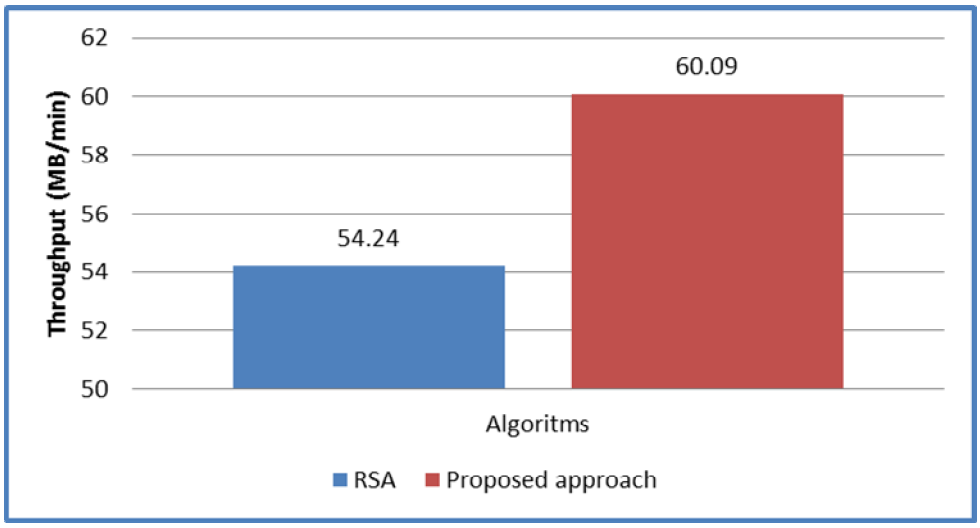

Figure 5.7: Encryption Throughput MB/min of RSA and Proposed Approach

Table 5.2 (b): Decryption Time of RSA and Proposed Approach

File Size	RSA (Min)	ILPS (Min)
64 MB	1.98	1.33
128 MB	2.89	2.08
256 MB	5.52	4.86
512 MB	10.02	9.47
1024 MB	18.21	16.32
Total Decryption Time	38.62	34.06
Throughput of Decryption (MB/Minutes)	51.38	58.26

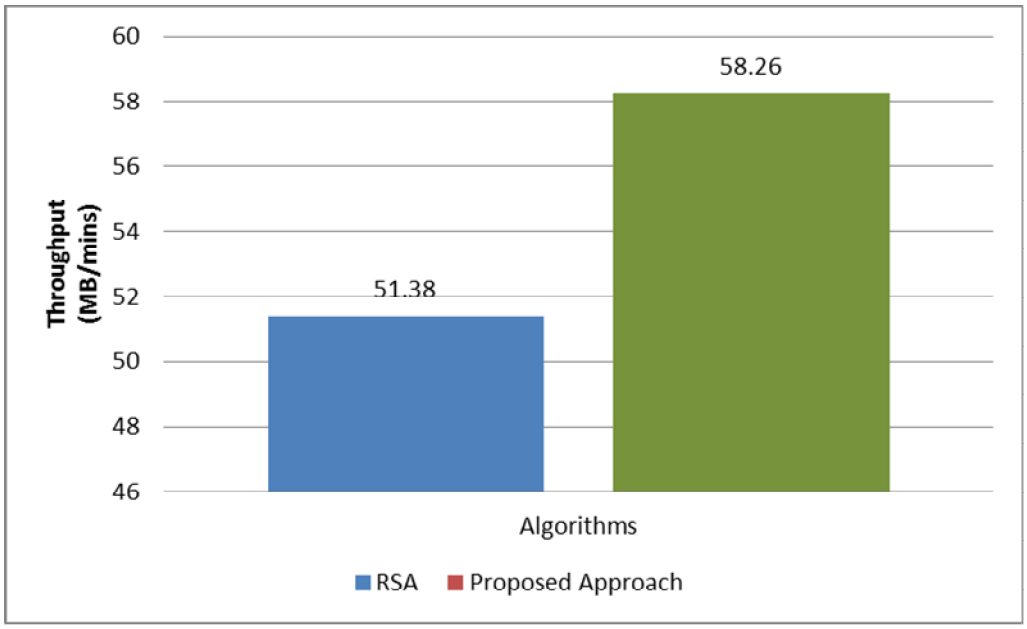

Figure 5.8: Decryption Throughput of RSA and Proposed Approach

Similarity, the proposed approach took 18.32 minutes to decrypt on 1 GB file. On the other hand, the existing RSA is taking 18.21 for decrypting file of 1 GB as shown in table-5.2(b). Pictorial representation of the results is shown in figure -5.8 and figure-5.9. The graphs represent the values for every criterion. Figure 5.8 and figure 5.9 represent the comparative time taken (in minutes) during the encryption and decryption process by existing RSA and proposed ILPS. From both the figures, it is clear that the proposed algorithm is taking less time for encryption and decryption in comparison to other existing algorithm.

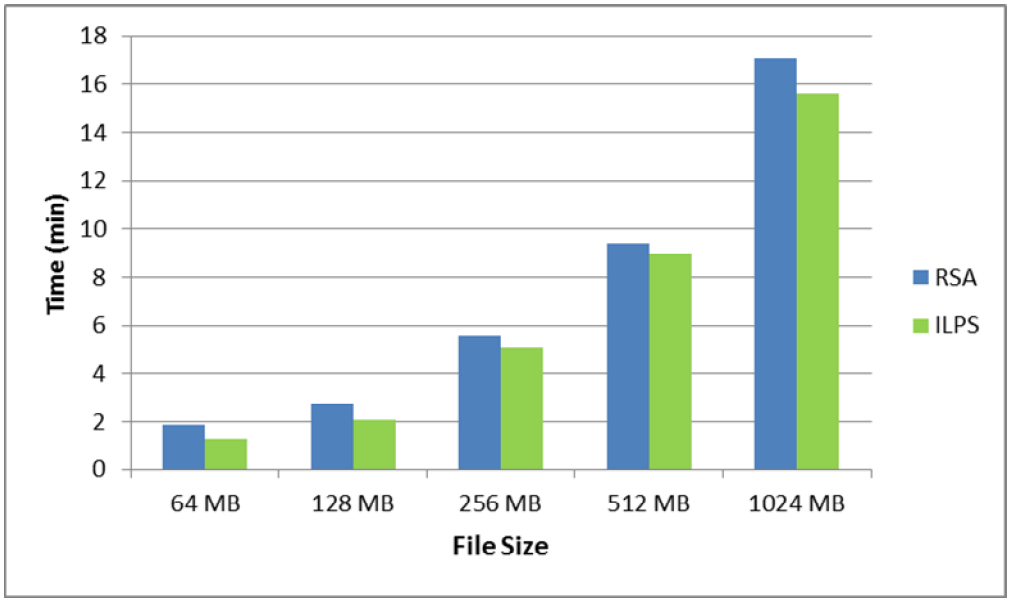

Figure 5.9: Encryption Time of RSA and Proposed ILPS

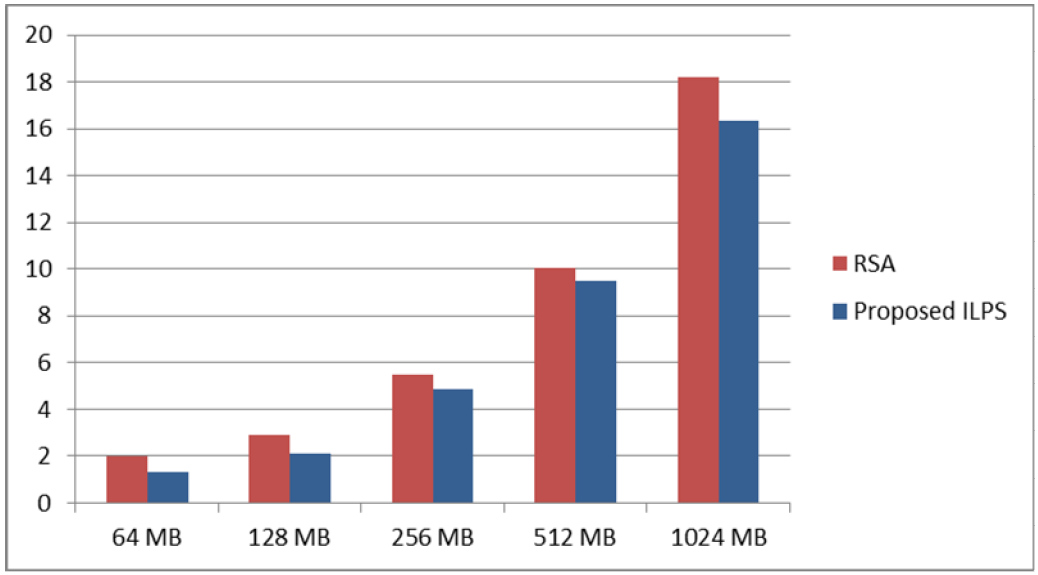

Figure 5.10: Decryption Time of RSA and Proposed ILPS

5.9 Statistical Validation

Validation of the research work is required to provide certain definite reassurance for consistency, accuracy and robustness of an automated system [260]. It is the process of finding or testing the truthfulness of an approach. The acceptance of any approach/model/methodology is based on its validation. This is the validation which makes one to believe in the result of any approach. It is a process that can be defined, designed and deploy data validation using different available methods [261]. Validation is the procedure to test the trustworthiness of any model/methodology/approach. It assumes that a product, system and service are fulfilling the necessary condition [261-263].There are some features which mostly include in the process of validation such as:

- Accuracy
- Specificity
- System suitability
- Repeatability
- Reproducibility
- Quantification limit

System suitability testing is an essential part of many analytical procedures. Validation tests are based on the concept of analytical operations, analysis samples, equipment and

electronics. The system suitability process tests parameters to establish for a specific method based on the type of method to be validated. Model validation is not a decision even rather it is a requirement in a dynamic modeling development [242] [264].

5.10 Methodology for Validation

Validation is a process to estimate the comparison among computational results and the actual (hypothesis) data from the system [242]. The primary aim of validation is the identification and quantity of error, uncertainty in conceptual models, and calculation of numerical error in the computational results, evaluation of simulation uncertainty, and finally, evaluation between computational results and real data. Therefore, accuracy is calculated in terms of actual/hypothetical data. But this method doesn't believe that the actual and hypothetical data is more accurate in comparison to computational results [265-266]. Various validations' characteristics are discussed below:

- A model should be evaluated for its effectiveness rather than its complete validity.
- A model cannot have complete validity however it should be fulfil the requirement for which it is constructed
- For Validation to pass test does not assure that model is valid and failing a test help to reject a negative hypothesis.

Validation is the most accurate part of developing models. Without validation test, no model can be accepted. It is a process to determining the trust of the model and usually it is framework based and dynamic [260]. As author's said in [264][265] that validation is applicable to all simulation models even if any existing system is present or it will be developed in the future. Though other authors have explained [260] that there is a complete valid model, the reliability of a model can be justified for the proposed use of the model and the recommended conditions under which the model has been tested [264].

5.10.1 Hypothesis Testing

Statistical validation is done to study the effectiveness of the proposed methodology. A hypothesis about the outcome of the proposed technique or measurement is a wise assessment. The student's t- Test sample is one of the most universally used techniques to test a hypothesis on the basis of difference among sample means [267]. For small observations, the student t-test is applied to check the level of significance and rejection of the null hypothesis. Since the rejection or acceptance of a null hypothesis is based on

(0.05) alpha (α) or (0.01) alpha (α) level of significance for one tailed or two tailed test, (0.01) alpha (α) level of significance for a two tailed test is taken to reject the null hypothesis [267-268].

Statistical analysis process involves few sequential steps in the process. Firstly, null hypothesis and alternate hypothesis has been formulated and after that statistical analysis has been performed. As a result of statistical analysis, it may be concluded that whether there is significant difference among the previous approach (RSA) and the proposed approach (ILPS). The received t value will determine whether to reject null hypothesis and accept the alternative hypothesis.

A null hypothesis shows that there is no essential relationship among two or more parameters, though the alternate hypothesis proves the relationships. Rejecting a null hypothesis provides a strong basis for accepting the relationship or accepting the alternate hypothesis [268]. As among available approaches RSA is a well-established and more secure approach, it has taken as basis for validation as
The formulated hypotheses are mentioned below:

- **Null Hypothesis (H_{01}):** The proposed approach ILPS takes more time in encryption compared to RSA.
- **Alternative Hypothesis (H_{11}):** The proposed approach ILPS takes less time in decryption compared to RSA.
- **Null Hypothesis (H_{02}):** The proposed approach ILPS takes more time in encryption compared to RSA.
- **Alternative Hypothesis (H_{12}):** The proposed approach ILPS takes less time in decryption compared to RSA

5.10.2 Encryption Data Set

For the purpose of validation, the researcher has taken text file having different sizes (32 KB , 126KB, 200KB, 246KB, 280KB, 64 MB, 128 MB, 256 MB, 512 MB, and 1024 MB). Total time taken for encrypting there files by the proposed ILPS as well as RSA has been tabulated in table-5.3.

a) Level of Significance of the Proposed Approach

To find out the significant difference between run time of RSA and proposed approach ILPS shown in table -5.4. Pearson's correlation coefficient (r) test is used to measure the strength of a linear relationship between two quantitative variables. Pearson correlation is (0. 999038776) which represent that the two variables are rather closely correlated. The degree of freedom is 9 (10-1). Statistical observation t -Statistic is (2.464787964) that is larger in absolute t critical value 2.262157163.Since the p (0.035875734) value is less than alpha, 0.05, researcher strongly reject null hypothesis (H0) and alternate hypothesis (H1) is accepted. Finally it can be concluded that with ILPS, the system will be more secure as well as robust.

Table-5.3: Encryption Time Taken by Proposed ILPS as Well as RSA

File Size	RSA (msec)	ILPS (msec)
32 KB	130	125
126KB	520	492
200KB	740	723
246KB	1110	798
280KB	1390	825
64 MB	111600	75000
128 MB	165000	123600
256 MB	333800	307200
512 MB	563000	537600
1024 MB	1021800	937800

Table -5.4: T -Test Paired Two Sample for Means

Statistical Observation	RSA	ILPS
Mean	198416.3	219909
Variance	98896468532	1.14283E+11
Observations	10	10
Pearson Correlation		0.999038954
Hypothesized Mean Difference		0
Df		9
t Stat		2.464787964
P(T<=t) one-tail		0.017937867
t Critical one-tail		1.833112933
P(T<=t) two-tail		0.035875734
t Critical two-tail		2.262157163

5.10.3 Decryption Data Set

For the purpose of validation, the researcher has taken text file having different sizes (32 KB, 126KB, 200KB, 246KB, 280KB, 64 MB, 128 MB, 256 MB, 512 MB, and 1024 MB). Total time taken for decrypting there files by the proposed ILPS as well as RSA has been tabulated in table-5.5.

Table -5.5: Decryption Time Taken by Proposed ILPS as well as RSA

File Size	RSA (msec)	ILPS (msec)
32 KB	150	145
126KB	430	404
200KB	660	610
246KB	930	867
280KB	1230	897
64 MB	118800	79800
128 MB	173400	124800
256 MB	331200	291600
512 MB	601200	568200
1024 MB	1092600	979200

a) Level of Significance of the Proposed Approach

To find out the significant difference between run time of RSA and proposed approach ILPS; table -5.6 shows the mean of RSA and proposed approach ILPS. Pearson's correlation coefficient (r) test is used to measure the strength of a linear relationship between two quantitative variables. Pearson correlation is (0.998848513) which represent that the two variables are rather closely correlated. The degree of freedom is 9 (10-1). Statistical observation t -Statistic is (2.464787964) that is larger in absolute t critical value 2.262157163. Since the p (0.040909047) value is less than alpha, 0.05, researcher strongly reject null hypothesis (H02) and alternate hypothesis (H12) is accepted. Finally it can be concluded that with ILPS, the system will be more secure as well as robust.

Table -5.6: T –Test Paired Two Sample for Means

Statistical Observation	RSA	ILPS
Mean	204652.3	232060
Variance	1.07759E+11	1.30064E+11
Observations	10	10
Pearson Correlation		0.998848513
Hypothesized Mean Difference		0
Df		9
t Stat		2.384733501
P(T<=t) one-tail		0.020454523
t Critical one-tail		1.833112933
P(T<=t) two-tail		0.040909047
t Critical two-tail		2.262157163

5.11 Conclusion

In the big data heterogeneous and semi-honest environment is the collection of user's personal and sensitive information. In this chapter, researcher has proposed a novel approach for secure sharing of sensitive health information in big data. Considering semi-honest big data servers, researcher argue that to completely understand the patient-centric concept, data owners (patient) shall have full control of their own security by encrypting SHI files to permit fine-grained access. The approach addresses the security issues of various SHI owners. The proposed approach shall greatly decrease the complexity of key management while increasing privacy of the SHI owner. The researcher has improved RSA to encrypt the SHI data, so patients will also be involved to permit access by personal users, as well as other users from public domains with different professional roles.

Through implementation and simulation, the authors have demonstrated that the proposed solution is scalable and efficient. The most essential phase of the research work is to validate the proposed approach. It is necessary to prove that the proposed approach it socially acceptable. In this chapter the proposed work has been validated. Student t-test has been used to test the hypothesis. The obtained values after t-test

disclose that the proposed ILPS is taking less time in comparison to the existing RSA. The results support strong rejection of null hypothesis and acceptance of alternate hypothesis.

CHAPTER-6: CONCLUSION AND FUTURE WORK

"A conclusion is the place where you get tired of thinking."

- Arthur Bloch

6.1 Background

This chapter encapsulates the main contributions of the research and also presents the future work in the area of security of big data. The main issues that have been taken into consideration and those remaining for the area have been discussed in addition to limitations of the research work. Big Data is used for the development of future technologies which will change the world as researcher see it, such as Internet of Things (IoT), or on demand services, and this is the reason that big data has big future. It is broadly used in research and industries. The big data concept begins with the data creation. But the next step is to secure that collected data and it is more difficult. Further, It is not only about collection or retrieval. Data collected from various sources and integrated with other sources to analyses the whole data for generating useful information. The big issue in big data is how to classify and secure sensitive information stored in different formats. Security of user privacy has become another big challenge for big data. An attacker can access user's private data with the help of some basic information of that particular user available in multiple sources. It is evident from the literature that complete lifecycle for big data and phase wise security doesn't exists which can provide the various issues arising in big data such as creation, sharing etc. Yazan's et.al [65] big data lifecycle threat model has given the idea about security threats and attacks that come in each phase. But yazan et.al did not explain the data creator's role. Was it possible to ignore the users' roles that create data? Can a life cycle be completed without data creator phase?

Hence, an improved lifecycle for big data and phase wise security threats has been proposed by the researcher. This lifecycle also addresses security attacks on the data

creation phase as well as their remedies. This is the complete lifecycle. It has five phases: data creation, data collection, data mining, data analytics and decision making phase. Every phase has its own attacks while data is travelling from one phase to another. To in build security at every phase is a difficult task. Though, by providing security at the initial phases further phases can also be secured. Hence, to secure the big data, researcher has worked on the two phases: data creation and data collection phases. The researcher has also addressed the need and importance of data creation phase. To strengthen the need and importance of the data creation phase, the researcher has collected data of 165 major cases of privacy breach from various sources [177-233]. The data is analyzed keeping in view the importance of data creation phase.

Every phase restricts privacy of user's information throughout the lifecycle. Further, the researcher has provided strategies controlling unauthorized access on personal sensitive information. The main objective is to develop lifecycle for big data and phase wise security threats, development of privacy policies and their implementation and an approach for data leakage prevention. The approaches have been implemented. Validation of the approaches has been also performed. Statistical interpretation represents that the proposed approaches are acceptable as well as useful.

6.2 Major Research Contributions

In depth study of the existing literature on security of big data and a rigorous research process done by the researcher has made the following major contributions:

- **Literature survey on big data security:** A significant review has been done on big data security during the starting phases of research work. It has been observed that from the past few years, research in big data has been going on worldwide. Data is currently one of the most important assets for companies in every field. Continuous growth in the importance and volume of data has created a new problem: it cannot be handled by traditional analysis techniques. However, big data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. So, security is the basic need of the people's personal and sensitive data. The development of big data and its security is the demand of time. It is the first step towards providing the secure environment. The research work highlighting the same has been published in [270].

- **An Improved Lifecycle for Big Data and Phase Wise Security Threats:** From the available literature, researcher ignores the creation of data. Users create their own data and they hand over this data to data collector. Hence the researcher has proposed data creation phase. Data creation phase is the most important phase of improved security threat model for big data life cycle. A creator is a person who handover or provides his data to the data collector. Keeping in mind the security of information, it is a very important phase. Once the data has been transferred in collector's hands, the privacy of the same cannot be confirmed. Now, the privacy of creator's data will depend on the data collection phase as well as further phases. It is important for the data creator to reveals only relevant information to others. An end user (creator) is most prone to the security attacks such as no control, unauthorized access. Hence, in this phase, security remedies have to be applied. To strengthen this phase the researcher has also introduced the concept of privacy paradox. This study has been focused on, the self-disclosure behavior and related privacy awareness to secure users information. Using survey data, researcher has validated and tested the research model. This study has utilized the online users, and has gathered data based on user's response. Results indicate that most of the respondents disclose significant amount of personal information without being conscious about the visibility and information leakage to the third party service providers and unknown users. The results reflect that if the creators would have taken care in advance about what data should be provided and what not, then these frauds could have been avoided. The detailed description is shown in chapter 3 and chapter 4. The study has been published in [101].

- **Development of Privacy Policies and their Implementation:** The researcher has proposed privacy policies to address privacy and security of user's data with architecture to minimize security risks and privacy. In the recent years, user's data security concerns are largely rooted in the rapidly expanding big data ecosystem conceptualized as the privileges of persons whose data is shared with others. Information security and the privacy of data have long been viewed as primary human rights. Today's digital user is no longer completely virtually identified. It is possible that user's personal information assembled for one reason can be retrieved for another without user consent. Organizations sell data to unauthorised party and form the data travels one place to another. A user is never aware about it. So, a big

question that arises here, in this interrelated milieu is there any privacy policy exists that can protect user's information? By analysing the feasibility of solution of the problem, the proposed research has been carried out to cater the need. Privacy policies have been developed and its implementation has been performed. The privacy policies should be adhered by information creators and information collector both in order to avoid loss of privacy. These policies when implemented may ensure privacy to individual as well as organization. So, it is the need of time to impose and finalize privacy policies for big data to evade from unauthorized access and misuse of user's information. It is expected that these policies will aware as well as secure individuals. The policies states that while providing services, organizations must present users with choice that whether they want to share their information or not. Proposed policies are clear and concise which clearly explains expectation from users. A confusion matrix is obtained to calculate the accuracy of classification. The detailed description is shown in chapter 4. The same has been published in [271].

- **A Novel Approach for Data Leakage Prevention:** A novel information leakage prevention scheme (ILPS) has been developed to secure big data system. The developed approach aims to secure patient's Sensitive Health Information (SHI) by modifying RSA encryption technique. This approach works on data collection phase. The proposed approach achieves better security in comparison with existing methods. Health data is very sensitive to any patient (data owner). The goal is to design an approach to prevent the SHI from unauthorised access. With the help of this scheme, an unauthorised user cannot access patient's sensitive health information (SHI) without his/her consent. Contrary from the previous work, the researcher has focused on the data owner (patient) and divides users into several domains such as public and personal domain. It is a novel patient-centric system model to control SHIs which is stored in semi honest servers. The proposed approach greatly decreases the key management complexity for data owners and users. This approach is implemented Hadoop 2.5.2 (pre-built i386-Linux native and java version SDK Oracle Java 1.8.0 25 (8u25) with the system specifications CPU Intel i7-7700, 4.2 GHz CPU and Microsoft Windows 7 as the operating system. The researcher has presented a performance analysis of the proposed approach, in which computation costs are analysed separately. Furthermore, system performance

is evaluated on the basis of encryption time, decryption time and throughput. For validation of the proposed approach, an experimental tryout has been carried out. The detailed description is presented in chapter 5. The study has been submitted for publication [272].

6.3 Significance of the Work

In today's scenario, data is growing exponentially. The world-wide free exchange of information, and the ease availability of internet, users willing to find an easy way to access the information even as the cost of their personal details. In this scenario, security of personal data has become a difficult problem, because there are no universally acceptable privacy policies or mechanisms to provide secure environment. There is no clear definition 'how to sensitive data in this technical era? It is very difficult to find an appropriate way to measure big data security susceptibilities and most of the aspects related to it. In addition, study of security measurement has become an important need for the industry developers and the users. The study carried on improving security by overcoming privacy issues in big data. With its significant addition in the area of knowledge; are significant directly or indirectly in context of the followings:

- The study will help to minimize the privacy issues.
- It may help to provide better security on both the data creation and collection phase of the threat model of big data lifecycle.
- With the help of proposed approach, data leakage issues can be prevented successfully.
- Proposed algorithm is easy to implement but it is very difficult for attacker to revert it back at the same time.
- It will generate the private key which is efficient and unbreakable.
- The proposed work may encourage and enable the users to come out with more efficient approaches for improving big data security.
- In proposed approach, data cannot be leaked or accessed at any point.

6.4 Future Directions

Research is an on-going activity. Reaching one milestone encourages the way to the next. As a future research plan, there may be the following tasks to be performed:

- In future, researcher will try to use different data mining techniques to identify and classify the sensitive data.
- Analyzing a variety of other case studies and experiments using other configurations is also a part of our future work.
- The researcher will try to develop a technique that can send authentication message to those users who demand privacy.
- Researcher will plan to conduct more experiments on industry data to draw more concrete conclusions.
- In future, researcher will implement this approach on different data formats.
- The researcher is planning to enhance proposed ILPS for other areas also.

6.5 Research Findings

During the course's study, the researcher's objective has been to find out the answers to research questions posed in chapter-1. In this line, the researcher has also tried to solve the security challenges acknowledged during survey of literature. The present section answers the research questions raised in chapter-1.

Research Question: What are the major challenges with respect to Big Data security?

Research Finding: Big data is new concept in heterogeneous environment. When it comes to big data, there are many privacy and security issues including unauthorised access, information leakage, loss of control etc.

Research Question: What is the current status of big data security?

Research Finding: Big data is different from the traditional database and it is very new for everyone. Very few techniques and methods are available to manage big data especially in the area of big data security.

Research Question: Is there any security mechanism available to secure big data at the time of data creation?

Research Finding: No, there is no security mechanism available to provide security at the time of data creation. It is up to the data collector to secure it. Information once submitted, the owner has no control over it.

Research Question: Does the information leakage problem in big data have been addressed previously?

Research Finding: Though this issue has been addressed by some researchers but the work still needs improvement.

Research Question: Can privacy policy be imposed on the organization collecting user's data?

Research Finding: According to THE PERSONAL DATA PROTECTION BILL 2018, privacy policy can be imposed on the organization collecting user's data. Now it is up to the user whether to disclose his/her sensitive data or not.

6.6 Conclusion

Every second, there is a large amount of data generated either by human or by machines around us. The era of big data has arrived and there is an essential need to address the issues that come with it. Thus, the researcher has identified a lot of critical security issues on big data security and privacy including data leakage and data privacy. In this thesis, the researcher investigated techniques to improve security of big data. Two closely-related problems have been studied: data privacy and data security. The proposed research has been carried out to address these problems. In the thesis, the researcher has proposed an improved llifecycle for big data and phase wise security threats that give a solution to the security challenges of big data. If the service providers follow the proposed privacy policy, it will definitely reduce the security risks.

Hence, the most essential insight from this research is the data creation phase of big data life cycle that minimizes the risk of the user's data, classifies the policy rules and includes them in clear cyber security guidelines. The researcher has also introduced the ILPS approach that accepts to protect e-Health infrastructures. The privacy paradox concept that refers to the real world problem has also been introduced. These approaches will surely bring significant improvement in system performance. Successful validation of the proposed approaches has also reflected its acceptability.

Appendices A: Questionnaire

QUESTIONS	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Privacy Concerns **PC1-**Restricted profiles to others.					
PC2-I consider my data sensitive because it contains my personal information.					
PC3-When online companies ask me for personal information, I think twice before providing it.					
PC4-I'm concerned that online companies are collecting too much personal information about me.					
PC5-It is very important to me that I am aware and knowledgeable about how my personal information will be used.					
Intentions **I1-**SometimesI provided wrong information to protect my personal data.					
I2-To obtain a free gift, I would share my data online.					
I3-When I buy a new iPhone then I openly share my information.					
Willingness **W1-**From my online profile, it would be easy to understand what type of person I am.					

W2-I want to share my personal thoughts, experiences.					
Self-Disclosure **SD1-**I sharepersonal information like age, home address, and favourite restaurants.					
SD3-I share medical history and financial information.					
Privacy Risk **PR1-**When I check in my actual house, who knows? Someone could come and find me.					
PR2-Since my boss does Facebook a lot, he is my friend on Facebook he might know everything that I post.					
PR3-Believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.					
PR4-If I forget to logout my id on public system then it is very easy for someone to misuse it.					
Benefits **B1-**I fulfil my social needs in some way by connecting with others.					
B2-I derive satisfaction from disclosing online.					
B3-Ease of access					

Appendix B: Hadoop Installation Overview

1. Ip address: - 192.168.X.X
2. Hostname: - hadoopkanika
3. Username: - hadoop password: - XXXXXXXXXXX
4. Hadoop Software Directory: - /home/hadoop/soft/hadoop2.5.2
5. Java: - /usr/java/jdkl 1.8.0_25/bin/java
6. The Java JDK software is the perquisite to be installed before Hadoop:
7. Install Java JDK

https://www.oracle.com//technetwork/java/javase/downloads/index.html

8. To check Java version that is installed in the system:

$java -version

Appendix C: Hadoop Pre-Requisites

I. Commands for using Hadoop

1. $ hadoop fs –ls /
2. $ hadoop fs –ls /user/hduser
3. $ hadoop fs –ls /user/hduser/samplefile
4. $ hadoop fs –put Sample2.txt /user/cloudera/
5. $ hadoop fs –df
6. $ hadoop fs –mkdir /user/cloudera
7. $ hadoop fs –rm /user/cloudera/dezyre3
8. $ hadoop fs –cp /user/cloudera /user/dezyre1/

II. Start HDFS

1. $ hadoop namenode –format
2. $ hadoop home/bin/hadoop fs –cat/user/output
3. $ hadoop home/bin/hadoop fs –ls
4. Stop-dfs.sh

III. Start Hbase

1. Cd/user/localhost
2. Start-hbase.sh
3. $hadoop/user/local/hbase/bin hbase.shell

References

[1] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 98-115.

[2] Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. Journal of big data, 1-20.

[3] Mani, D., Raymond Choo, K. K., & Mubarak, S. (2014). Information security in the South Australian real estate industry: A study of 40 real estate organisations. Information Management & Computer Security, 24-41.

[4] Kim, S. H., Kim, N. U., & Chung, T. M. (2013, December). Attribute relationship evaluation methodology for big data security. In IT Convergence and Security (ICITCS), 2013 International Conference on (pp. 1-4). IEEE.

[5] Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in big data. XRDS: Crossroads, The ACM Magazine for Students, 20-23.

[6] Tole Alexandru Adrian (2013), "Big data challenges." Database Syst, 31-40.

[7] Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. Information Fusion, 45-59.

[8] Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the fareast. IDC iView: IDC Analyze the future, 1-16.

[9] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 137-144.

[10] Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ‘big data can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 234-246.

[11] Gholami, A., & Laure, E. (2016). Big Data Security and privacy issues in the cloud. International Journal of Network Security & Its Applications (IJNSA), 59-79.

[12] Achana, R. A., Hegadi, R. S., & Manjunath, T. N. (2015, December). A novel data security framework using E-MOD for big data. In Electrical and Computer

Engineering (WIECON-ECE), 2015 IEEE International WIE Conference on (pp. 546-551). IEEE.

[13] Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 314-347.

[14] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. IEEE access, 1821-1834.

[15] sas.com. (2013). Big Data: What it is and why it matters. Available at: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

[16] Kshetri, N. (2014). Big data′ s impact on privacy, security and consumer welfare. Telecommunications Policy, 1134-1145.

[17] Wall Street Journal. Insurers test data profiles to identify risky clients, 2011.

[18] Whang, S. E., & Garcia-Molina, H. (2012, August). A model for quantifying information leakage. In Workshop on Secure Data Management (pp. 25-44). Springer, Berlin, Heidelberg.

[19] Warren, P.W. & Davies, N.J., 2007. Managing the risks from information — through semantic information management. BT Technology Journal, 178–191.

[20] Min Chen, Shiwen Mao, Yunhao Liu (2014). Big Data: a survey, Mobile Networks Appl. 171–209.

[21] Moreno, J., Serrano, M. A., & Fernández-Medina, E. (2016). Main issues in big data security. Future Internet, 44.

[22] Wang, H., Jiang, X., & Kambourakis, G. (2015). Special issue on Security, Privacy and Trust in network-based Big Data. Information Sciences: an International Journal, 48-50.

[23] Available at, https://blog.barkly.com/biggest-data-breaches-2018-so-far.

[24] Available at, https://www.checkmarx.com/2017/12/31/recap-biggest-data-breaches-2017/

[25] Available at, https://incyber1.com/2017/09/13/the-largest-data-breaches-of-2017/

[26] Expanded Top Ten Big Data Security and Privacy Challenges, Cloud Security Alliance, April 2013, 1-39.

[27] Xu, H., & Gupta, S. (2009). The effects of privacy concerns and personal innovativeness on potential and experienced customers' adoption of location-based services. Electronic Markets, 137-149.

[28] Yaqoob, I., Chang, V., Gani, A., Mokhtar, S., Hashem, I. A. T., Ahmed, E., & Khan, S. U. (2016). TEMPORARY REMOVAL: Information fusion in social big data: Foundations, state-of-the-art, applications, challenges, and future research directions. International Journal of Information Management, 1231–1247.

[29] Chang, V., & Ramachandran, M. (2016). Towards achieving data security with the cloud computing adoption framework. IEEE Trans. Services Computing, 138-151.

[30] Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., & Qu, W. (2012). Big data processing: Big challenges and opportunities. Journal of Interconnection Networks, 1-19.

[31] Thakuriah, P. V., Tilahun, N. Y., & Zellner, M. (2017). Big data and urban informatics: innovations and challenges to urban planning and knowledge discovery. In seeing cities through big data (pp. 11-45). Springer, Cham.

[32] Mantelero, A., & Vaciago, G. (2014). Social media and big data. In Cybercrime and cyber terrorism investigator's handbook (pp. 175-195).

[33] Estivill-Castro, V., Hough, P., & Islam, M. Z. (2014, October). Empowering users of social networks to assess their privacy risks. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 644-649). IEEE.

[34] Kepner, J., Gadepally, V., Michaleas, P., Schear, N., Varia, M., Yerukhimovich, A., & Cunningham, R. K. (2014, September). Computing on masked data: a high performance method for improving big data veracity. In High Performance Extreme Computing Conference (HPEC), 2014 IEEE (pp. 1-6). IEEE.

[35] Sharma, P. P., & Navdeti, C. P. (2014). Securing big data hadoop: a review of security issues, threats and solution. Int. J. Comput. Sci. Inf. Technol, 2126-2131.

[36] Savant, V. G. (2015). Approaches to solve big data security issues and comparative study of cryptographic algorithms for data encryption. International Journal of Engineering Research and General Science, 425-428.

[37] Wang, Z., & Wang, D. (2013, November). NCluster: Using Multiple Active Name Nodes to Achieve High Availability for HDFS. In High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on (pp. 2291-2297). IEEE.

[38] Meye, P., Raipin, P., Tronel, F., & Anceaume, E. (2014, July). Mistore: A distributed storage system leveraging the DSL infrastructure of an ISP. In High Performance Computing & Simulation (HPCS), 2014 International Conference on (pp. 260-267). IEEE.

[39] Azeem, M. A., Sharfuddin, M., & Ragunathan, T. (2014, October). Support-based replication algorithm for cloud storage systems. In Proceedings of the 7th ACM India Computing Conference (p. 11). ACM.

[40] Vatsalan, D., Christen, P., O'Keefe, C. M., & Verykios, V. S. (2014). An evaluation framework for privacy-preserving record linkage. Journal of Privacy and Confidentiality, 35-75.

[41] K., Agrawal, A., & Khan, R. A. (2017). AN OVERVIEW OF PRIVACY PRESERVATION IN BIG DATA. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 61-71.

[42] Moura, J., & Serrão, C. (2015). Security and privacy issues of big data. In Handbook of research on trends and future directions in big data and web intelligence (pp. 20-52). IGI Global.

[43] Securing the Big Data Life Cycle, oracle, MIT.

[44] Dijcks, J. P. (2012). Oracle: Big data for the enterprise. Oracle white paper, 1-14.

[45] Jeong, Y. S., & Shin, S. S. (2016). An efficient authentication scheme to protect user privacy in seamless big data services. Wireless Personal Communications, 7-19.

[46] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

[47] Alan Pacocha (2002),"Using Security To Protect The Privacy of Customer Information", SANS Institute, 1-11.

[48] Saraladevi, B., Pazhaniraja, N., Paul, P. V., Basha, M. S., & Dhavachelvan, P.

(2015). Big Data and Hadoop-A study in security perspective. Procedia computer science,596-601.

[49] Somu, N., Gangaa, A., & Sriram, V. S. (2014). Authentication service in hadoop using one time pad. Indian Journal of Science and Technology, 56-62.

[50] Jam, M. R., Khanli, L. M., Javan, M. S., & Akbari, M. K. (2014, October). A survey on security of Hadoop. In Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on (pp. 716-721). IEEE.

[51] Bhosale, H. S., & Gadekar, D. P. (2014). A review paper on Big Data and Hadoop. International Journal of Scientific and Research Publications, 1-7.

[52] Nir Kshetri (2014). Big data's impact on privacy, security and consumer welfare. Telecommunications Policy. 1134–1145.

[53] Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. Big data & society, 1-6.

[54] Warren, D., & Brandeis, D. (1890). The Right to Privacy, Harvard Law Review, 192–220.

[55] Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. GeoJournal, 463-475.

[56] Sasson, R., Jaatun, M. G., & Jensen, J. (2010, February). The road to Hell is paved with good intentions: A story of (in) secure software development. In Availability, Reliability, and Security, 2010. ARES'10 International Conference on (pp. 501-506). IEEE.

[57] Lekkas, D. (2003). Establishing and managing trust within the Public Key Infrastructure. Computer Communications, 1815-1825.

[58] Reese, G. (2009). Cloud application architectures: building applications and infrastructure in the cloud. " O'Reilly Media, Inc.".

[59] Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation computer systems, 599-616.

[60] Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in big data. XRDS: Crossroads, the ACM Magazine for Students, 20-23.

[61] Siponen, M., & Vance, A. (2014). Guidelines for improving the contextual relevance of field surveys: the case of information security policy violations. European Journal of Information Systems, 289-305.

[62] Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. Computer networks, 146-164.

[63] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. IEEE Access, 1149-1176.

[64] Yadav, C., Wang, S., & Kumar, M. (2013). Algorithm and approaches to handle large Data-A Survey. International Journal of Computer Science and Network, 1-5.

[65] Alshboul, Y., Nepali, R., & Wang, Y. (2015). Big data lifecycle: Threats and security model. Americas Conference on Information Systems (pp. 1-7).

[66] Li, M., Yu, S., Zheng, Y., Ren, K., & Lou, W. (2013). Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. IEEE transactions on parallel and distributed systems, 131-143.

[67] Crampton, J. W. (2015). Collect it all: National security, big data and governance. GeoJournal, 519-531.

[68] Phaneendra, S. V., & 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).

[69] Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821.

[70] Richards, N. M., & King, J. H. (2013). Three paradoxes of big data. Stan. L. Rev. Online, 41-46.

[71] Menon, S. P., & Hegde, N. P. (2015, January). A survey of tools and applications in big data. In Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on (pp. 1-7). IEEE.

[72] Purcell, B. (2013). The emergence of" big data" technology and analytics. Journal of technology research, 4, 1.

[73] Rajeswari, C., Basu, D., & Maurya, N. (2017, November). Comparative Study of Big data Analytics Tools: R and Tableau. In IOP Conference Series:

Materials Science and Engineering (Vol. 263, No. 4, p. 042052). IOP Publishing.

[74] Vidhya, S. S. N. S. S., Sarumathi, S., & Shanthi, N. (2014). Comparative analysis of diverse collection of big data analytics tools. International Journal of Computer, Electrical, Automation, Control and Information Engineering, 1646-1652.

[75] Shirudkar Kalyani, et.al (2015), "Big-Data Security", International Journal of Advanced Research in Computer Science and Software Engineering, 1100-1009.

[76] Islam, N. S., Rahman, M. W., Jose, J., Rajachandrasekar, R., Wang, H., Subramoni, H., & Panda, D. K. (2012, November). High performance RDMA-based design of HDFS over InfiniBand. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (p. 35). IEEE Computer Society Press.

[77] Ahn, S., & Park, S. (2015). An analytical approach to evaluation of ssd effects under mapreduce workloads. JOURNAL OF SEMICONDUCTOR TECHNOLOGY AND SCIENCE, 511-518.

[78] Hong, J., Li, L., Han, C., Jin, B., Yang, Q., & Yang, Z. (2016, June). Optimizing Hadoop framework for solid state drives. In Big Data (BigData Congress), 2016 IEEE International Congress on (pp. 9-17). IEEE.

[79] Krish, K. R., Iqbal, M. S., & Butt, A. R. (2014, October). Venu: Orchestrating ssds in hadoop storage. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 207-212). IEEE.

[80] Ateniese, G., Fu, K., Green, M., & Hohenberger, S. (2006). Improved proxy re-encryption schemes with applications to secure distributed storage. ACM Transactions on Information and System Security (TISSEC), 1-30.

[81] Yang, C., Lin, W., & Liu, M. (2013, September). A novel triple encryption scheme for hadoop-based cloud data security. In emerging intelligent data and web technologies (EIDWT), 2013 fourth international conference on (pp. 437-442). IEEE.

[82] Bifet, A. (2013). Mining big data in real time. Informatica, 15-20.

[83] Bu, Y., Howe, B., Balazinska, M., & Ernst, M. D. (2012). The HaLoop approach to large-scale iterative data analysis. The VLDB Journal—The International Journal on Very Large Data Bases,169-190.

[84] Katal, A., Wazid, M., & Goudar, R. H. (2013, August). Big data: issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409). IEEE.

[85] Gaikwad, R. L., Dakhane, D. M., & Pardhi, R. L. (2013). Network Security Enhancement in Hadoop Clusters. Int. J. Appl. Innov. Eng. Manage.(IJAIEM), 151-157.

[86] Das, D., O'Malley, O., Radia, S., & Zhang, K. (2011). Adding security to apache hadoop. Hortonworks, IBM, 26-36.

[87] Park, S., & Lee, Y. (2013, May). Secure hadoop with encrypted HDFS. In International Conference on Grid and Pervasive Computing (pp. 134-141). Springer, Berlin, Heidelberg.

[88] O'Malley, O., Zhang, K., Radia, S., Marti, R., & Harrell, C. (2009). Hadoop security design. Yahoo, Inc., Tech. Rep 1-19.

[89] Flores, W. R., Antonsen, E., & Ekstedt, M. (2014). Information security knowledge sharing in organizations: Investigating the effect of behavioral information security governance and national culture. Computers & Security, 90-110.

[90] Edith Ramirez (2015), Securing the Big Data Life Cycle, MIT TECHNOLOGY REVIEW CUSTOM + ORACLE, 1-8.

[91] Baskerville, R., Spagnoletti, P., & Kim, J. (2014). Incident-centered information security: Managing a strategic balance between prevention and response. Information & management, 138-151.

[92] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, M., Kamaleldin, W., & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. The Scientific World Journal, 1-19.

[93] Colin Tankard (2012), Big data security, Network Security, 5-8.

[94] Turow, J., Feldman, L., & Meltzer, K. (2005). Open to exploitation: America's shoppers online and offline. Departmental Papers (ASC), 1-35.

[95] Hall, J. L., & McGraw, D. (2014). For telehealth to succeed, privacy and security risks must be identified and addressed. Health Affairs, 216-221.

[96] Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. IEEE access, 2751-2763.

[97] Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y., & Dasseni, E.

(2004). Association rule hiding. IEEE Transactions on knowledge and data engineering, 434-447.

[98] Wang, S. L., Huang, K. W., Wang, T. C., & Hong, T. P. (2005, October). Maintenance of discovered informative rule sets: incremental deletion. In Systems, Man and Cybernetics, 2005 IEEE International Conference on (Vol. 1, pp. 170-175). IEEE.

[99] Amiri, A. (2007). Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems, 181-191.

[100] Yang, J. J., Li, J. Q., & Niu, Y. (2015). A hybrid solution for privacy preserving medical data sharing in the cloud environment. Future Generation Computer Systems, 74-86.

[101] **Kanika, Agrawal, A., & Khan, R. A. (2018). An Improved Security Threat Model for Big Data Life Cycle. Asian Journal of Computer Science and Technology, 33-39**.

[102] Souza, S. M., & Puttini, R. S. (2016). Client-side encryption for privacy-sensitive applications on the cloud. Procedia Computer Science, 126-130.

[103] Clarkson, K.L., Liu, K. and Terzi, E. (2010) 'Toward identity anonymization in social networks', in Link Mining: Models, Algorithms, and Applications, pp.359–385, Springer, New York.

[104] Zhou, B. and Pei, J. (2011) 'The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks', Knowledge and Information Systems, 47–77.

[105] Zakerzadeh, H. and Osborn, S.L. (2011) 'Faanst: fast anonymizing algorithm for numerical streaming data', in Data Privacy Management and Autonomous Spontaneous Security, pp.36–50, Springer, Berlin, Heidelberg.

[106] Abawajy, J.H., Kelarev, A. and Chowdhury, M. (2014) 'Large iterative multitier ensemble classifiers for security of big data', IEEE Transactions on Emerging Topics in Computing, 352–363.

[107] Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2014). A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. IEEE Transactions on Parallel and Distributed Systems, 363-373.

[108] Bhattacharya, M., Islam, R., & Abawajy, J. (2016). Evolutionary optimization: a big data perspective. Journal of network and computer applications, 416-426.

[109] Fard, A. M., & Wang, K. (2015). Neighbourhood randomization for link privacy in social network analysis. World Wide Web, 9-32.

[110] Yan, Z., Ding, W., Yu, X., Zhu, H., & Deng, R. H. (2016). Deduplication on encrypted big data in cloud. IEEE transactions on big data, 138-150.

[111] Benson, T. (2012). Principles of health interoperability HL7 and SNOMED. Springer Science & Business Media.

[112] AbuKhousa, E., Mohamed, N., & Al-Jaroodi, J. (2012). e-Health cloud: Opportunities and challenges. Future Internet, 621–645.

[113] Sultan, N. (2014). Making use of cloud computing for healthcare provision: Opportunities and challenges. International Journal of Information Management, 177–184.

[114] Elger, B. S., Iavindrasana, J., Iacono, L. L., Müller, H., Roduit, N., Summers, P., & Wright, J. (2010). Strategies for health data exchange for secondary, cross-institutional clinical research. Computer methods and programs in biomedicine, 230-251.

[115] Wang, B., Li, B., & Li, H. (2012). Oruta: Privacy-preserving public auditing for shared data in the cloud. In Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on (pp. 295–302).

[116] Wu, R., Ahn, G.-J., & Hu, H. (2012). Towards HIPAA-compliant healthcare systems. In Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI ’12.

[117] Fernández-Alemán, J. L., Señor, I. C., Lozoya, P. Á. O., & Toval, A. (2013). Security and privacy in electronic health records: A systematic literature review. Journal of biomedical informatics, 541-562.

[118] Agrawal, R., & Johnson, C. (2007). Securing electronic health records without impeding the flow of information. International journal of medical informatics, 471-479.

[119] Rumbold, J. M., & Pierscionek, B. K. (2018). What are data? a categorization of the data sensitivity spectrum. Big data research, 49-59.

[120] Ardagna, C. A., Di Vimercati, S. D. C., Foresti, S., Grandison, T. W., Jajodia, S., & Samarati, P. (2010). Access control for smarter healthcare using policy spaces. Computers & Security, 848-858.

[121] Narayan, S., Gagné, M., & Safavi-Naini, R. (2010, October). Privacy preserving EHR system using attribute-based infrastructure. In Proceedings of the 2010 ACM workshop on Cloud computing security workshop (pp. 47-52). ACM.

[122] Jin, J., Ahn, G. J., Hu, H., Covington, M. J., & Zhang, X. (2011). Patient-centric authorization framework for electronic healthcare services. computers & security, 116-127.

[123] Li, Y., Gai, K., Qiu, L., Qiu, M., & Zhao, H. (2017). Intelligent cryptography approach for secure distributed big data storage in cloud computing. Information Sciences, 103-115.

[124] Zhang, Q., Yang, L. T., & Chen, Z. (2016). Privacy preserving deep computation model on cloud for big data feature learning. IEEE Transactions on Computers, 1351-1362.

[125] Liu, X., Liu, Q., Peng, T., & Wu, J. (2017). Dynamic access policy in cloud-based personal health record (PHR) systems. Information Sciences, 379, 62-81.

[126] Huang, Q., Yang, Y., & Shen, M. (2017). Secure and efficient data collaboration with hierarchical attribute-based encryption in cloud computing. Future Generation Computer Systems, 239-249.

[127] Pradhan, S., & Sharma, B. K. (2013). An efficient RSA cryptosystem with BM-PRIME method. International Journal of Information and Network Security, 103.

[128] Shehzad, D., Khan, Z., Dag, H., & Bozkus, Z. (2016). A novel hybrid encryption scheme to ensure Hadoop based cloud data security. International Journal of Computer Science and Information Security, 480-484.

[129] Vincent, P. D. R., & Sathiyamoorthy, E. (2016). A novel and efficient public key encryption algorithm. International Journal of Information and communication technology, 199-211.

[130] Ray, S., & Biswas, G. P. (2012). Design of RSA-CA based e-health system for supporting HIPAA privacy-security regulations. Procedia Technology, 954-961.

[131] Huang, H. F., & Liu, K. C. (2011). Efficient key management for preserving HIPAA regulations. Journal of Systems and Software, 113-119.

[132] Jian WS, Wen HC, Scholl J, Shabbir SA, Lee P, Hsu CY, et al. The Taiwanese method for providing patients data from multiple hospital EHR systems. J Biomed Inform, 326–32.

[133] Wuchner, T., & Pretschner, A. (2012). Data loss prevention based on data-driven usage control. In 23rd IEEE International Symposium on Software Reliability Engineering, ISSRE 2012, Dallas, TX, USA, November 27-30, 2012.

[134] Liu, Z., Chen, X., Yang, J., Jia, C., & You, I. (2016). New order preserving encryption model for outsourced databases in cloud environments. Journal of Network and Computer Applications, 198-207.

[135] Chaudhry, S. A., Farash, M. S., Naqvi, H., & Sher, M. (2016). A secure and efficient authenticated encryption for electronic payment systems using elliptic curve cryptography. Electronic Commerce Research, 113-139.

[136] Kumar, P., & Rana, S. B. (2016). Development of modified AES algorithm for data security. Optik-International Journal for Light and Electron Optics, 2341-2345.

[137] Liu, X., Lu, R., Ma, J., Chen, L., & Qin, B. (2016). Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification. IEEE journal of biomedical and health informatics, 655-668.

[138] Jiang, Q., Khan, M. K., Lu, X., Ma, J., & He, D. (2016). A privacy preserving three-factor authentication protocol for e-Health clouds. The Journal of Supercomputing, 3826-3849.

[139] Gai, K., Qiu, M., Zhao, H., & Xiong, J. (2016, June). Privacy-Aware Adaptive Data Encryption Strategy of Big Data in Cloud Computing. In CSCloud (pp. 273-278).

[140] Wang, Z., Cao, C., Yang, N., & Chang, V. (2017). ABE with improved auxiliary input for big data security. Journal of Computer and System Sciences, 41-50.

[141] Lo, N. W., Wu, C. Y., & Chuang, Y. H. (2017). An authentication and authorization mechanism for long-term electronic health records management. Procedia computer science, 145-153.

[142] Jones, S., & Fox, S. (2009). Generations online in 2009. Washington DC: Pew Internet & American Life Project.

[143] Oh, H. J., Ozkaya, E., & LaRose, R. (2014). How does online social networking enhance life satisfaction? The relationships among online supportive interaction, affect, perceived social support, sense of community, and life satisfaction. Computers in Human Behavior, 30, 69-78.

[144] Lee, H., Park, H., & Kim, J. (2013). Why do people share their context information on Social Network Services? A qualitative study and an experimental study on users' behavior of balancing perceived benefit and risk. International Journal of Human-Computer Studies, 862-877.

[145] Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. Academy of Management Review, 17-40.

[146] Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., & Resnick, P. (2005). Using social psychology to motivate contributions to online communities. Journal of Computer-Mediated Communication, 1-30.

[147] Subrahmanyam, K., Reich, S. M., Waechter, N., & Espinoza, G. (2008). Online and offline social networks: Use of social networking sites by emerging adults. Journal of applied developmental psychology, 420-433.

[148] Burke, M., Marlow, C., & Lento, T. (2010). Social network activity and social wellbeing. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1909-1912). ACM.

[149] Ellison, N. B., Lampe, C., & Steinfield, C. (2010). With a little help from my friends: How social network sites affect social capital processes. In A networked self (pp. 132-153). Rutledge.

[150] Ellison, N. B., Vitak, J., Gray, R., & Lampe, C. (2014). Cultivating social resources on social network sites: Facebook relationship maintenance behaviours and their role in social capital processes. Journal of Computer Mediated Communication, 855-870.

[151] Burgoon, J. K. (1982). Privacy and communication. Annals of the International Communication Association, 206-249.

[152] Li, Y. (2011). Empirical studies on online information privacy concerns: Literature review and an integrative framework. CAIS, 453-496.

[153] Raschke, R. L., A. S. Krishen, and P. Kachroo. 2014. “Understanding the Components of Information Privacy Threats for Location-based Services.” Journal of Information Systems, 227–242.

[154] Krasnova, H., Spiekermann, S., Koroleva, K., & Hildebrand, T. (2010). Online social networks: Why we disclose. Journal of Information Technology, 109-125

[155] PEW. (2014). Public perceptions of privacy and security in the Post-Snowden Era. Pew Research Center.

[156] Taddicken, M. (2014). The ‘Privacy Paradox’in the Social Web: The Impact of Privacy Concerns, Individual Characteristics, and the Perceived Social Relevance on Different Forms of Self-Disclosure. Journal of Computer-Mediated Communication, 248-273.

[157] Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers and Security, 1-29.

[158] Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: measuring individuals' concerns about organizational practices. MIS quarterly, 167-196.

[159] Dienlin, T., & Trepte, S. (2015). Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviours. European Journal of Social Psychology, 285-297.

[160] Krasnova, H., Veltri, N. F., & Günther, O. (2012). Self-disclosure and privacy calculus on social networking sites: The role of culture. Business and Information Systems Engineering, 127-135.

[161] Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. Journal of Computer-Mediated Communication, 83-108.

[162] Lutz, C., & Strathoff, P. (2014). Privacy concerns and online behavior–Not so paradoxical after all? Viewing the privacy paradox through different theoretical lenses.

[163] Turow, J., Feldman, L., & Meltzer, K. (2005). Open to exploitation: America's shoppers online and offline. Departmental Papers (ASC), 35.

[164] Hallam, C., & Zanella, G. (2017). Online self-disclosure: The privacy paradox explained as a temporally discounted balance between concerns and rewards. Computers in Human Behavior, 217-227.

[165] Clarke, R. (1999). Internet privacy concerns confirm the case for intervention. Communications of the ACM, 60-67.

[166] Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information systems research, 336-355.

[167] Soghoian, C. (2008). Exclusive: The next Facebook privacy scandal. CNet News. Com.

[168] Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. Bulletin of Science, Technology and Society, 20-36.

[169] Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Dokas, P., Srivastava, J., & Kumar, V. (2003). Detection and summarization of novel network attacks using data mining. Minnesota INtrusion Detection System (MINDS) Technical Report.

[170] Patil, P., Narayankar, P., Narayan, D. G., & Meena, S. M. (2016). A comprehensive evaluation of cryptographic algorithms: DES, 3DES, AES, RSA and Blowfish. Procedia Computer Science, 617-624.

[171] Mayer-Schönberger V, Cukier K (2013). Big Data. A revolution that will transform how we live, work and think. London: John Murray Publishers, 1-7.

[172] Eastin, M. S., Brinson, N. H., Doorey, A., & Wilcox, G (2016). "Living in a big data world: Predicting mobile commerce activity through privacy concerns." Computers in Human Behaviour, 214-220.

[173] Tene, O., & Polonetsky, J. (2011). Privacy in the age of big data: a time for big decisions. Stan. L. Rev. Online, 63-69.

[174] Degli Esposti, S. (2014). When big data meets dataveillance: The hidden side of analytics. Surveillance & Society, 209-225.

[175] Broeders, D., Schrijvers, E., van der Sloot, B., van Brakel, R., de Hoog, J., & Ballin, E. H. (2017). Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data. Computer Law & Security Review, 309-323.

[176] Jagwani, P., & Kaushik, S. (2017, March). Privacy in Location Based Services: Protection Strategies, Attack Models and Open Challenges. In International Conference on Information Science and Applications (pp. 12-21). Springer, Singapore.

[177] Available at: http://timesofindia.indiatimes.com/city/aurangabad/46-year-old-man-duped-by-tele-phishers-in-aurangabad/articleshow/56972419.cms

[178] Available at: http://timesofindia.indiatimes.com/city/aurangabad/61-year-old-retired-government-employee-becomes-victim-of-tele-phishing/articleshow/55303909.cms

[179] Available at: http://timesofindia.indiatimes.com/city/hyderabad/Nigerian-gang-dupes-city-firm-of-rs-13L-3-held/articleshow/55021965.cms

[180] Available at: http://timesofindia.indiatimes.com/city/nagpur/Chhattisgarh-IT-department-asks-PNB-to-payback-Rs-22-36L-to-phishing-victim/articleshow/52423336.cms

[181] Available at: http://timesofindia.indiatimes.com/city/hyderabad/Watch-out-for-Rio-phishing-scam/articleshow/52333371.cms

[182] Available at: http://timesofindia.indiatimes.com/city/aurangabad/zp-clerk-duped-of-rs-36480-via-tele-phishing/articleshow/56408706.cms

[183] Available at: http://www.bbc.com/news/business-35250678

[184] Available at: http://timesofindia.indiatimes.com/city/bengaluru/techies-wife-loses-rs-11-lakh-to-phishing-ends-her-life/articleshow/52966455.cms

[185] Available at: http://www.actionfraud.police.uk/news/nca-are-warning-people-to-protect-themselves-from-dridex-malware-oct15

[186] Available at: http://www.actionfraud.police.uk/news/alert-watch-out-for-fraudsters-purporting-to-be-from-action-fraud-litigation-services-oct15

[187] Available at: http://www.thehindu.com/news/cities/bangalore/Not-easy-to-crack-cyber-crime/article14006807.ece

[188] Available at: http://www.actionfraud.police.uk/millions-of-talktalk-customers-vulnerable-to-fraud-after-data-breach-feb15

[189] Available at: https://indianexpress.com/article/business/companies/identity-theft-ongc-falls-prey-to-cyber-fraud-loses-rs-197-crore/

[190] Available at: http://indianexpress.com/article/cities/pune/online-bank-fraud-again-woman-duped-of-rs-35000/

[191] Available at: http://www.actionfraud.police.uk/news/alert-online-scam-involving-gumtree-paypal-and-western-union-oct14

[192] Available at: http://www.actionfraud.police.uk/news/financial-ombudsman-service-warning-over-150-compensation-scam-calls-oct14

[193] Raided call centre ordered to stop sending millions of spam texts, Available at: http://www.actionfraud.police.uk/news/raided-call-centre-ordered-to-stop-sending-millions-of-spam-texts-sept14

[194] Available at: http://www.actionfraud.police.uk/news/scammer-derek-jones-calling-victims-claiming-they-have-to-pay-a-debt-nov14

[195] Available at: http://www.actionfraud.police.uk/news/alert-watch-out-for-new-number-spoofing-scam-oct14

[196] Available at: http://www.actionfraud.police.uk/news/five-million-gmail-username-and-passwords-posted-online-by-hackers-sept14

[197] Available at: http://www.actionfraud.police.uk/news/police-arrest-fraudster-over-complex-mobile-phone-scam-targeting-students-aug14

[198] Available at: http://www.actionfraud.police.uk/increase-in-the-number-of-younger-fraudsters-say-kpmg-jul14

[199] Available at: http://www.actionfraud.police.uk/alert-watch-out-for-confirm-your-email-address-linkedin-phishing-scam-jul14

[200] Available at: http://www.actionfraud.police.uk/alert-watch-out-for-job-application-fraud-may14

[201] Available at: http://www.actionfraud.police.uk/scam-email-purporting-to-be-euromillions-winner-apr14

[202] Available at: http://www.actionfraud.police.uk/fraudsters-trick-victims-into-revealing-pin-numbers-after-theft-apr14

[203] Available at: http://www.actionfraud.police.uk/alert-beware-of-justice.gov.uk-scam-parking-fine-emails-mar14

[204] Available at: http://www.actionfraud.police.uk/watch-out-for-green-energy-scam-calls-feb14

[205] Available at: http://www.actionfraud.police.uk/ukash-scam-warning-jan14

[206] Available at: http://www.bankinfosecurity.eu/12-arrested-in-vishing-case-a-6839

[207] Available at: http://www.actionfraud.police.uk/cold-callers-promote-diamond-fraud-jan14

[208] Available at: http://www.actionfraud.police.uk/bogus-telephone-calls-from-online-tec-jul13

[209] Available at: http://www.actionfraud.police.uk/fraudsters-convicted-for-fake-online-adverts-nov13

[210] Available at: http://www.actionfraud.police.uk/school-truancy-cold-calls-feb13

[211] Available at: http://akademie.dw.de/digitalsafety/falling-for-phishing-hook-line-and-sinker/

[212] Available at: http://www.actionfraud.police.uk/alert-fake-voicemail-emails-from-skype-contain-virus-nov13

[213] Available at: http://www.actionfraud.police.uk/nuisance-calls-make-up-40per-cent-of-calls-to-elderly-and-vulnerable

[214] Available at: http://www.actionfraud.police.uk/quarter-of-people-in-uk-at-risk-of-vishing-aug13

[215] Available at: http://www.actionfraud.police.uk/threatening-scam-email-from-action-fraud-team-aug13

[216] Available at: http://www.actionfraud.police.uk/sky-card-scam-jul13

[217] Available at: http://www.actionfraud.police.uk/facebook-bug-exposes-6-million-users-jun13

[218] Available at: https://indianexpress.com/article/cities/pune/online-fraud-police-arrest-beneficiary-account-holder-search-on-for-masterminds/

[219] Available at:

[220] http://www.thehindu.com/todays-paper/tp-national/tp-kerala/cyber-crimes-on-the-rise-in-state/article3068112.ece

[221] Available at:

[222] http://www.ndtv.com/cities/doctor-loses-rs-11-14-lakh-in-internet-mobile-banking-fraud-475268

[223] Available at: http://indianexpress.com/article/cities/pune/5-held-in-pnb-cyber-crime-involving-rs-80-lakh/

[224] Available at: http://indianexpress.com/article/india/india-others/sms-scam-how-an-indian-man-lost-rs-18-lakh/

[225] Available at: https://indianexpress.com/article/india/crime/btech-student-loses-rs-19-lakh-in-email-fraud/

[226] Available at: http://indianexpress.com/article/cities/delhi/email-hacker-pits-aiims-professors-against-each-other/

[227] Available at: http://www.dnaindia.com/pune/report-hc-rejects-icici-bank-s-plea-in-phishing-case-1879918

[228] Available at: http://www.bankinfosecurity.in/phishing-scheme-spread-to-3-more-states-a-2058

[229] Available at: https://indianexpress.com/article/cities/delhi/mail-id-stolen-exec-says-hacker-wants-money/

[230] Available at: http://indianexpress.com/article/cities/pune/online-lottery-frauds-on-the-rise-say-cops/

[231] Available at: http://indianexpress.com/article/cities/ahmedabad/hackers-steal-rs-7-39-lakh-from-bank-account/

[232] Available at, http://www.hindustantimes.com/india/rs-12-lakh-in-bank-phished-one-arrested/story-SCUjSjq2UsRzMua9UaAN7I.html

[233] Available at: https://indianexpress.com/article/cities/chandigrah/3-yrs-after-netbanking-account-hacked-cops-fail-to-trace-accused-to-close-case/

[234] Kennedy, H., Elgesem, D., & Miguel, C. (2017). On fairness: User perspectives on social media data mining. Convergence, 270-288.

[235] Lampinen, A., Stutzman, F., & Bylund, M. (2011, May). Privacy for a Networked World: bridging theory and design. In CHI'11 Extended Abstracts on Human Factors in Computing Systems (pp. 2441-2444). ACM.

[236] Duckham, M., & Kulik, L. (2005, May). A formal model of obfuscation and negotiation for location privacy. In International conference on pervasive computing (pp. 152-170). Springer, Berlin, Heidelberg.

[237] Trope, Y., Liberman, N., & Wakslak, C. (2007). Construal levels and psychological distance: Effects on representation, prediction, evaluation, and behavior. Journal of consumer psychology, 83-95.

[238] Mothersbaugh, D. L., Foxx, W. K., Beatty, S. E., & Wang, S. (2012). Disclosure antecedents in an online service context: The role of sensitivity of information. Journal of service research, 76-98.

[239] Lwin, M., Wirtz, J., & Williams, J. D. (2007). Consumer online privacy concerns and responses: a power–responsibility equilibrium perspective. Journal of the Academy of Marketing Science, 572-585.

[240] Wang, E. S. T., & Lin, R. L. (2017). Perceived quality factors of location-based apps on trust, perceived privacy risk, and continuous usage intention. Behaviour & Information Technology, 2-10.

[241] Likert, R. (1932). A Technique for the Measurement of Attitudes. Archives of Psychology, 1–55.

[242] Straub, D., Boudreau, M. C., & Gefen, D. (2004). Validation guidelines for IS positivist research. Communications of the Association for Information systems, 380-427.

[243] Available at, https://www.statisticssolutions.com/using-chi-square-statistic-in-research/

[244] Liu, C., Yang, C., Zhang, X., & Chen, J. (2015). External integrity verification for outsourced big data in cloud and IoT: A big picture. Future generation computer systems, 58-67.

[245] Tsohou, A., & Kosta, E. (2017). Enabling valid informed consent for location tracking through privacy awareness of users: A process theory. Computer Law & Security Review, 434-457.

[246] Schneier, B. (2015). Data and Goliath: The hidden battles to collect your data and control your world. WW Norton & Company.

[247] Reddy, C. K. K., Anisha, P. R., Reddy, K. S., & Reddy, S. S. (2012). Third party data protection applied to cloud and XACML implementation in the hadoop environment with sparql. IOSR Journal of Computer Engineering (IOSRJCE) 39-46.

[248] Student dataset, Available at: http://archive.ics.uci.edu/ml/datasets/student+performance

[249] Rodriguez, J.D., Perez, A. and Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence, .569-575.

[250] Delen, D., Walker, G. and Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 113-127.

[251] Celesti, A., Fazio, M., Villari, M., & Puliafito, A. (2016). Adding long-term availability, obfuscation, and encryption to multi-cloud storage systems. Journal of Network and Computer Applications, 208-218.

[252] Li, M., Yu, S., Ren, K., & Lou, W. (2010, September). Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings. In International conference on security and privacy in communication systems (pp. 89-106). Springer, Berlin, Heidelberg.

[253] Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2017). Manual for using homomorphic encryption for bioinformatics. Proceedings of the IEEE, 552-567.

[254] Vengadapurvaja, A. M., Nisha, G., Aarthy, R., & Sasikaladevi, N. (2017). An Efficient Homomorphic Medical Image Encryption Algorithm For Cloud Storage Security. Procedia Computer Science, 643-650.

[255] Pussewalage, H. S. G., & Oleshchuk, V. A. (2017). Attribute based access control scheme with controlled access delegation for collaborative E-health environments. Journal of Information Security and Applications, 50-64.

[256] Mandl, K. D., Markwell, D., MacDonald, R., Szolovits, P., & Kohane, I. S. (2001). Public standards and patients' control: how to keep electronic medical records accessible but private Medical information: access and privacy Doctrines for developing electronic medical records Desirable characteristics of electronic medical records Challenges and limitations for electronic medical records Conclusions Commentary: Open approaches to electronic patient records Commentary: A patient's viewpoint. Bmj, 283-287.

[257] Maqsood, F., Ali, M. M., Ahmed, M., & Shah, M. A. (2017). Cryptography: A comparative analysis for modern techniques. International Journal of Advanced Computer Science and Applications, 442-448.

[258] Patil, P., Narayankar, P., Narayan, D. G., & Meena, S. M. (2016). A comprehensive evaluation of cryptographic algorithms: DES, 3DES, AES, RSA and Blowfish. Procedia Computer Science, 617-624.

[259] Diabetes 130-US hospitals for years 1999-2008 Data Set, available at: https://archive.ics.uci.edu/ml/datasets/diabetes+130us+hospitals+for+years+1999-2008.

[260] DMSO (Defense Modeling and Simulation Office). “The Principles of Verification, Validation, and Accreditation”, [online], Verification, Validation, and Accreditation Recommended Practices Guide, www.dmso.mil/public, U.S. Department of Defense, Office of the Director of Defense Research and Engineering, Nov. 1996.

[261] Validation and verification, Available at: DOI:www/validation/Verificationandvalidation/Wikipedia/free/encyclopedia.html

[262] Martis, M. S. (2006). Validation of simulation based models: a theoretical outlook. The electronic journal of business research methods, 39-46.

[263] American Institute of Aeronautics and Astronautics. (1998). AIAA guide for the verification and validation of computational fluid dynamics simulations. American Institute of aeronautics and astronautics.

[264] J. D. Sterman (1984), "Appropriate Summary Statistics for Evaluating the Historic Fit of System Dynamics Models", Dynamica, 51-66.

[265] Kleijnen, J. P. (1999, December). Validation of models: statistical techniques and data availability. In Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future-Volume 1 (pp. 647-654). A CM.

[266] Sterman, J. D. (2000). Business dynamics: systems thinking and modeling for a complex world (No. HD30. 2 S7835 2000).

[267] Students dataset, Available at: DOI: Student's%20t%20Test%20(For%20Independent%20Samples).html

[268] Available at: DOI: www.excel-easy.com/ /t-Test.

[269] Chen, T., & Taura, K. (2012, November). A Comparative Study of Data Processing Approaches for Text Processing Workflows. In 2012 SC Companion: High-Performance Computing, Networking, Storage and Analysis (SCC) (pp. 1260-1267). IEEE.

[270] K., Agrawal, A., & Khan, R. A. (2017). AN OVERVIEW OF PRIVACY PRESERVATION IN BIG DATA. INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 61-71.

[271] Kanika, Agrawal A, Khan R .A (2018). Privacy Policy: A Novel Approach to Preserve Confidentiality in Big Data, Jour of Adv Research in Dynamical & Control Systems.

[272] Kanika, Alka Agrawal and R. A. Khan, ''A Scheme for Prevention of Information Leakage in Big Data", communicated in World Wide Web Journal.