A Proposed Multilayered Framework for Security and Privacy in Big Data

Danish Bilal Ansari
Department of Computer Science and Information Technology
Virtual University of Pakistan

Muhammad Abdul Khaliq
Department of Computer Science
Lahore Garrison University

ABSTRACT
In this day and age, Big Data is extensively used among numerous organizations. From gathering and storing sensitive data to performing analytics, Big Data plays a fundamental role in any organizational growth in various social and economic areas. With data secrecy, a huge volume of data creates different security and privacy concerns that lead to several threats and vulnerabilities. Conventional methodologies for secured data are no longer applicable when applied in the perspective of Big Data. This paper presents a detailed overview of Big Data characteristics regarding V's and provides an insight into the analytics. It also addresses the security and privacy risks and proposed a multilayered framework to address and mitigate these challenges.

Keywords
Big Data, Security and Privacy Challenges, Multilayered Framework, Big Data V's, Big Data Analytics

1. INTRODUCTION
In former times, the information was gathered and stored on paper documents rather than electronic format. As the volume of the collected information increases, it becomes challenging to store and maintain these documents. With the advancement in technology, various tools become available for sustaining the information and make is beneficial. The gathered information could comprise text, images, videos, etc. that is classified as Data.

In other words, Data is the collection of facts and figures gathered for analysis or study. In recent years, a large amount of data is being produced which becomes a major management issue for any organization. To handle this, a term coined as "Big Data". Big Data is defined by the Ministry of Defense as, "Big Data is a large collection of data that can be stored, scrutinized, evaluated, explored and conveyed – is now a part of every section or work of the worldwide economy. Like other vital aspects such as hard possessions and human resources, the economical movements, modernization, novelty, and development couldn't take place without the essence of data" [1].

Big Data refers to a huge collection of data gathered through numerous sources including the Internet of Things (IoT) where every device is connected to the internet and contributes to generating, collecting and storing an enormous amount of data. Big Data comes into play when the heterogeneous and diverse volume of data is not manageable through conventional techniques. Vast storage with processing and analysis is required to attain usefulinformation from data collected through different sources as shown in Figure 1.

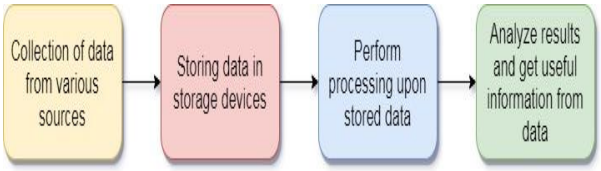

Fig 1. Big Data Structure

The main concern regarding data collection through various sources is the secrecy and privacy of the gathered data. Information is collected at various layers in a Big Data system, and every layer has security concerns with data anonymity, data integrity, and data availability along with the privacy in different phases comprising of generation, storage, and processing phase. To address these, the paper is divided into 6 sections and is arranged as follows. Section 2 presents the Literature Review regarding Big Data. Section 3 focuses on Characteristics of Big Data consisting of 7 V's followed by section 4 which addresses Analytics in Big Data. Section 5 describes the Security and Privacy concerns of Big Data. Section 6 proposed a Multilayered Framework for Security and Privacy in detail, and finally, Section 7 concludes the paper.

2. LITERATURE REVIEW
Over the past time, the majority of the organizations implemented big data for data management and analytics with much emphasis on analysis. The potential outcome of big data analytics is humongous, but organizations have to amend their business procedures and infrastructure to attain the maximum benefit out of the expanding capacity of big data [24]. Thus, researchers tend to focus on another term known as cloud computing, because of its nature to provide more storing capabilities for big data, along with cost-effectiveness, flexibility, resource management and data handling properties [25,26]. However, there are many security and privacy threats associated with cloud computing as well [27,28].

Many surveys and research were conducted in the past few years for data integrity and its anonymity in the IoT and cloud-based systems [29]. But, the concentration of the research is shifting towards the security and privacy concerns in big data systems because of their distributed nature and diverse attributes [30]. With the progression in the e-health systems, research studies in big data have been conducted in the health industry because of its predominant nature [31,32]. However, with excessive usage of social media, and IoT network devices in different organizations and among many individuals, personal data and user behaviors have been gathered progressively over the internet which results in various security and privacy implications that make available solutions insufficient for meeting various consumers requirements [33,34]. Numerous literature reviews, surveys

and research have been conducted recently in the area of big data with the emphasis on security and privacy. The results show the prominence of security and privacy in big data concerning its computational characteristics. The main focus of these literature reviews and surveys is to focus only on well-known research areas usually privacy and security. Among them, only a few have provided with an indication of big data and have proposed a substitute for research in this area [35,36]. However, none of them addresses the security aspect of big data.

Thus, this literature review presents that there is a divergence in the research area regarding the security and privacy concern of big data at different layers and phases of a big data life cycle. Besides, the growth and advancement in new technologies have presented exceptional issues and concerns for the security and privacy of big data. However, no detailed research is available to address the primary issues in big data. It is essential to identify the imminent concerns in big data from the very beginning. First, numerous features and attributes of big data need to uncover that are progressing with the advancement in recent technologies. Additionally, the growing attributes of big data demonstrate specific compelling elements that affect the security and privacy aspect of big data. These implications require a detailed study that could lead to various security and privacy issues and threats in big data, which needs to be classified and tackled appropriately. As far as we know, these issues and concerns have not addressed properly from the very beginning. The lack in the research area of big data regarding security and privacy aspect gives us the prime motivation to address the differences in literature with the first step in this paper.

3. BIG DATA CHARACTERISTICS

To better understand the terminology of Big Data, it can be characterized into 7 V's as follows:

3.1 Volume

The huge amount of data in the form of images, videos, textual, graphical, etc. has been generated in recent years which led to the storage aspect of Big Data also known as "Volume" or the first V of Big Data. Data varies in the form of terabytes, zettabytes to yottabytes [2]. According to this [3], about 1.1 trillion photos were taken in the year 2016 while the number rose to 9 percent in the year 2017. Figure 2 predicts the volume growth over the years [17].

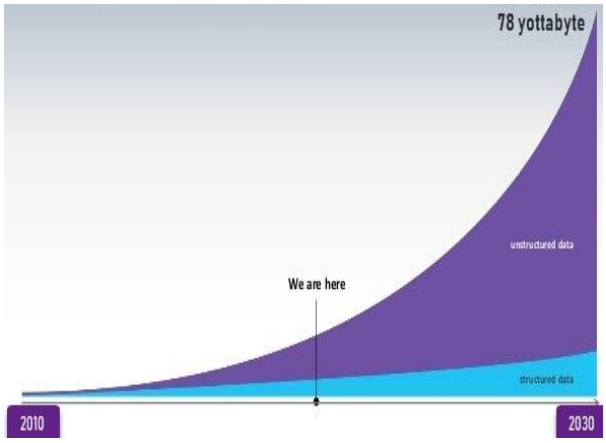

Fig 2. Growth of volume in Big Data [17]

3.2 Velocity

Another V of Big Data is termed as "Velocity" which defines the rate of data being produced, retrieved and processed. With the advancement in internet usage, humongous data is produced just by web site clicks [2]. According to these stats, around 40,000 queries are processed by Google contributing to 3.5 billion searches per day and 1.2 trillion per year [4].

3.3 Variety

Data comprises various forms including structured, semi-structured and unstructured. The structured format is easy to process as it is organized. It is manageable in the form of a repository, typically a database, so analysis and processing become easy. However, unstructured data like videos, images, and audio are difficult to manage as they do not contain any specific set of rules. Apart from above, sensor and machine data, as well as log files, are also an example of different data formats. Such diverse data is more difficult to process and analyze.

3.4 Veracity

The accuracy and trustworthiness of data being utilized refer to the Veracity of Big Data. Data is collected and stored through numerous sources including different websites, social media, etc. which are then processed and analyzed to extract different statistics. Data with inaccurate information often results in misleading facts. Veracity comes into play as it measures the authenticity of data and its usage in the analysis.

3.5 Variability

Another vital aspect of Big Data is Variability which is often considered similar to Veracity. However, data in terms of meaning and context is rapidly changing in variability. For instance, a word in two identical posts on Facebook can have an entirely different meaning. To have a precise and correct analysis, the algorithm for variability should have a better understanding of the data context and should be able to decode the true meaning of a word in that context [5].

3.6 Visualization

Visualization refers to the graphical representation of data in such a way that information and knowledge can be used more intuitively. For example, an e-commerce site like Amazon has millions of registered users that utilize it to buy goods and services and hence, millions of items are sold each month generating a lot of data. To have a better explicability of such data, representation in spreadsheet format is not enough. Here, visualization comes into play. Having a pictorial illustration in the form of graphs or charts would play a significant role in decision making for any corporate sector. Figure 3 shows the amazon revenue for the last five years using a graph [6].

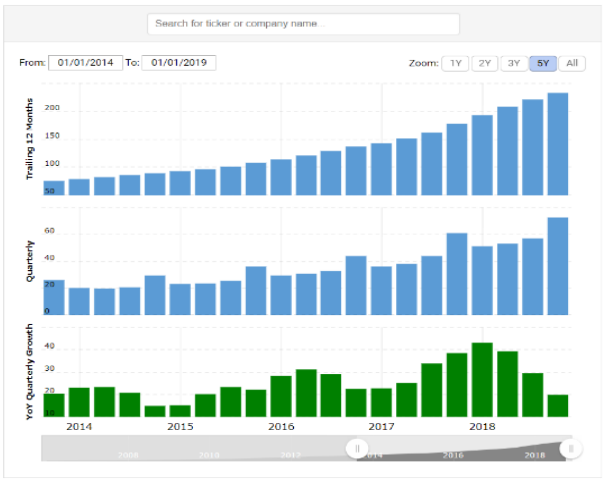

Fig 3. Amazon Revenue [6]

3.7 Value

The last V is of much importance in Big Data known as Value. Data alone is insignificant, but with processing and analysis, it produces value by allowing the appropriate decision making for an organization. Value is considered a vital and essential element in Big Data by researchers because of the valuable information [7] and if assessed properly, big data can provide the business with a significant advantage. Figure 4 shows the 7 V's of Big Data.

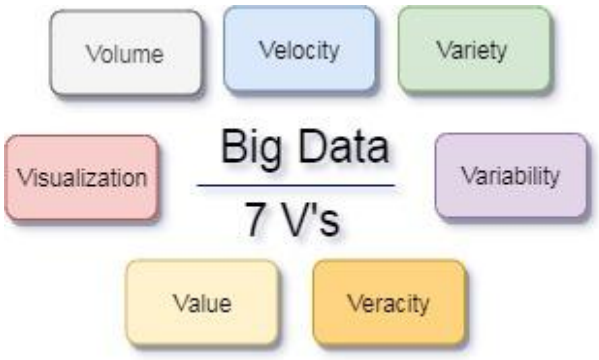

Fig 4. 7 V's of Big Data

4. BIG DATA ANALYTICS

Big data volume is beneficial because of the multiple frameworks which are the basis of big data analytics. These frameworks allow massive analysis to be performed on a small infrastructure. On various equivalent nodes and clusters, these frameworks are installed which consists of a collection of systems. They use "shared nothing" architecture, distributed processing frameworks and parallel relational and non-relational databases [8].

Among the many big data tool, Apache Hadoop is the most known processing tool [9]. Several components are developed by Apache which is open source and the Hadoop ecosystem is formed by them. It is licensed under Apache V2. Apache Hadoop consists of many fundamental components containing Hadoop Kernel, Map-Reduce, Hadoop Distributed File System (HDFS), HBase, Hive and more.

Hadoop is open-source software that is extensively used for Big Data. Unlike other relational database architecture, it is based on a non-relational structure that manages a large collection of distributed data. It also helps in parallel processing on big volumes of data. Two main components of Hadoop are storage and processing. Many clusters of distributed nodes in Hadoop are used for storing the big amount of data which is then processed by MapReduce. These components of Hadoop along with others are shown in Table 1. as follows:

Table 1. Big Data Analytics Techniques

Big Data Analytics Techniques	
Map Reduce	MapReduce was introduced by Google as a processing tool to process large data sets for distributed data. It is divided into two categories, namely Map and Reduce. A key/value pair is processed by Map function which produces an intermediary pair based on key/value, and then all intermediary values are combined with a similar intermediary key using the reduce function.
HDFS	HDFS or Hadoop Distributed File System is the center of Hadoop. It is used in various data warehouses for storing and managing distributed data. The HDFS is established on the basis of GFS (Google File System) for handling large distributed file system. The other file systems are embedded in the kernel of an operating system and executes as a process of the operating system, while the HDFS executes as a user process rather than a kernel process in the operating system processes table. Various nodes are used for storing huge files ranging from terabytes to petabytes in HDFS. The storage is done by dividing data into numerous blocks and assigning these blocks in various locations on a server.
HBase	Google's BigTable was the idea behind the Hadoop HBase. HBase is the distributed database management system for Hadoop and is effective for handling and processing tables with millions of columns. HBase is built over HDFS and provides a fault-tolerant way of keeping a huge amount of data [8]. It is useful where a high throughput of random reading and writing operations are required. It can also use MapReduce for storing and retrieving data. Facebook and Pinterest are an example of using HBase.
Hive	Hive is a data warehousing tool that uses the services provided by Hadoop. Various aggregated functions and queries on data that executes on Hive is provided with a language called Hive QL (HQL). The HQL is identical in structure to SQL. Numerous MapReduce jobs that run of the Hadoop cluster are linked by Hive which splits the HQL queries [10].
Cassandra	Cassandra is known to be a foremost transactional, accessible and exceedingly available distributed database system [15]. Cassandra is highly anticipated among big data platforms because of its scalability, fault-tolerant peer-to-peer architecture [16], multipurpose and adaptable data model, a flexible language like NoSQL known as Cassandra Query Language (CQL), and effective read and write operations that handle millions of transactions per second. It is widely used because of handling nodes and clusters of data centers in case of any malfunction. Cassandra also supports MapReduce processing and is particularly known for its ability to facilitate data access for a large volume of records.
In-Memory	In-memory operations refer to the quick accessibility of data when required. The processing of in-memory is very similar to the procedures performed on RAM rather than on hard drive. However, like RAM, the data is not stored permanently, which can cause data archival issues.
NoSQL	NoSQL or Not Only SQL refers to a non-relational database management system, where tables or views etc. are not required for dealing with a database. For any data operation, SQL based query structure is not used. As established

	ona non-relational database management system, they are suitable where operations on an enormous amount of data are not performed on a relational structure. NoSQL is based on non-relational and distributed database systems used for storing huge amounts of data and performing parallel processing. NoSQL is typically used for ETL based data transformation and performing OLTP. In comparison with relational database management systems, these systems are used for millions of data updates and doing read operations which are the needs of Web 2.0 [11].

5. SECURITY & PRIVACY IN BIG DATA

Because of the potential usage of Big Data, it has been a rapid interest in numerous fields including industry, financial, healthcare, telecom and scientific, etc. However, before any implementation of Big Data in enormous applications, an essential area needed to be addressed: security and privacy.

5.1 Security

The security of big data is quite similar to the security of various information systems that comprise structured data. However, security in big data expects several effective approaches, suitable methods, and progressive skills for data analytics. It also requires a strong security model for managing data collected through internal and external sources. For these points, a few questions arise:

i) How unstructured and heterogeneous data can be securely managed?
ii) How to achieve better functionality while incorporating security techniques into distributed systems?
iii) Without negotiating data security and privacy, how can large data be examined?

Usually, the purpose of security in big data is to discover vulnerabilities, security risks, anomalous activities and provide with role-based access control, strong preservation of secret data and producing the security operation indicators. It also assists prompt decision making in case of any security occurrence. Many big data security issues arise because of significant factors:

5.1.1. Insufficient Conventional Solutions

Many of the encryption techniques regarding data protection, data sharing and storing, and data trustworthiness related to hardware is presented in [12]. Data security is maintained using the following techniques [13]:

- Data is generally stored as plaintext format.
- For data access, authentication is performed on plaintext data.
- Encryption is used when storing the data, while data is decrypted when it is in use.

However, conventional mechanisms like data encryption are slow in terms of performance, and time-consuming regarding big data. Additionally, they are inefficient. For security reasons, only a small fraction of data is processed which results in a catastrophic event as mostly the security intrusion is detected after the occurrence. Big Data platforms are used for managing numerous products and performing parallel processing. Hence, performance is an essential factor for data security and data evaluation in such platforms.

5.1.2 Undeveloped Security Applications

Various threats can be introduced to a big data system by merging numerous technological applications which are not assessed appropriately. Additionally, such security applications are undeveloped. Thus, big data applications may introduce security threats and weaknesses and results in negotiating data integrity, which provides security against any data modification by unauthorized users. Moreover, many security threats emerge from associates, employees and end-users. So, it is essential to have enhanced security tools for protecting big data platforms.

5.1.3 Data Anonymity

Data secrecy or anonymity is considered essential in big data security. Data anonymity should be attained without interrupting real-time data analysis and its value. However, the conventional data anonymization mechanism performs various repetitions and inefficient and slow computations. With large data heterogeneity groups, traditional data anonymization may affect system operations and data consistency with various repetitions. Additionally, it becomes cumbersome to handle and evaluate anonymity in big data.

5.1.4 Data Integrity

Data integrity refers to providing security against any modification of data by an unauthorized user. Many of the data integrity issues occur because of software mistakes, hardware errors and user miscalculations as shown in [13]. The data diddling attacks, trust Relationship attacks, man-in-the-middle attacks, salami attacks, and session hijacking attacks are some of the most known attacks on data integrity [15]. Data integrity can be achieved using the following techniques:

- Data Provenance
- Data Trustworthiness
- Prevention of Data Loss
- Data Deduplication

Data integration security can be achieved by utilizing several techniques as better hardware integration, digital signature, data storage protection and big data query protection [12].

5.1.5 Data Availability

Data availability refers to the availability of data to authorized users. Data availability can be ensured using Highly Available (HA) systems [13]. For designing a highly available system, reproductions, substitute communication links, and backup servers are required. Some of the security vulnerabilities regarding big data availability are Denial of Service (DoS) attacks, Distributed Denial of Service (DDoS) attacks, SYN flood attacks, Internet Control Message Protocol (ICMP) attacks, Server Room attacks, and Electricity Power attacks that require improved solutions [14].

5.2 Privacy

As we known that big data is nothing but the collection of a huge amount of useful information, due to technology advancement massive amount of data are being gathered from different sources i.e. social media, IoT devices, organizations day by day, handling of such massive amount of data is also the biggest challenge, a lot of issues have been addressed and their solutions also have been viewed parallel, but most important of them are its security and privacy. Big data privacy is different from its security. Big data privacy refers to protecting sensitive information relates to individual user

i.e. name, address, social security numbers, login credentials and accounts information. Whereas security protects all types of data and information that an organization collects from any source. Privacy in big data is required as different administration roles have been working at different levels in big data analytics, also privacy required at different stages of the data i.e. its generation, storage, and processing stage.

5.2.1 Privacy in Data Acquisition/Generation Phase:

At the time of generation of data, the data can be transfer to the third person in two ways in the first one the data owner are well aware of that the information is shared between two persons and no third person involved between them and in the second way the information of user is collected by unauthorized person without the knowledge of owner. So, while performing any activity on the internet the intruders can intrude on the personal information of the user at any time. In order to avoid such kind of privacy issue, the user should avoid to provide the sensitive information over the internet or on websites that have lower security levels, if needed then user should use any intermediate tools while sharing information over the internet such as anti-virus, anti-malware, anti-spam, and firewalls, for example, while using any card information over any online shopping store the information should be encrypted through any type of tool helps to provide privacy of data at generation time.

5.2.2 Privacy in Data Sharing/Storage Phase

Traditional storage methods are insufficient while dealing with big data technology, as big data required large amount of storage capacity, large size of businesses are stored and process the Big Data with the help of distributed file system such as Google File System (GFS) or Hadoop Distributed File System (HDFS) [23], but for small and medium-sized businesses it is very challenging, the utilization of these systems due to financial problems, for that purpose Cloud Computing can be used.

Storing and processing of big data through the cloud is very helpful for small and medium-sized businesses but it also generates a lot of security and privacy challenges for them, although organization used several security and privacy mechanisms for protecting his information, still, a lot of security and privacy risks exist. Numerous encryption techniques and algorithms have been proposed to ensure these characteristics.

- Identity based encryption:

ID-Base Encryption is a type of Public Key Encryption in which the unique information of a user that represents the unique identity i.e. email, phone, CNIC is used as a public key, [19] so the sender encrypts the message using the unique identity of the receiver as a key and receiver decrypt the message with the key. But the issue with this technique is that its time-consuming technique.

- Attribute based encryption:

It is also a type of public-key encryption technique in which the ciphertext is not encrypted for one specific user as in other encryption technique, instead of all the users their private keys and ciphertexts are related with the set of attributes of a specific user [22]. User can only decrypt the ciphertext if there is a match occurred between his private key and the encrypted text.

- Fully Homomorphic Encryption:

This is also a data encryption technique that provides the privacy of data while transferring or storing it at the cloud. [19] In the Fully Homomorphic Encryption technique, the computation is performed upon the ciphertext and the decrypted data and match the result, the results of that computation remain encrypted or secured. Thus, both the encrypted and decrypted text treated similarly and perform some algebraic functions.

5.2.3 Privacy in Data Processing Phase

The processing of data means to perform some operations on date in-order to extract some useful information. So, while processing the data make sure that the privacy or unwanted leakage of data does not disturb also the information securely stored back into the cloud.

6. MULTILAYERED FRAMEWORK FOR BIG DATA

This section proposed a multilayered approach as a framework for the issues related to security and privacy in big data. Five core layer sections for the framework is as follows:

i) Data Management
ii) IAM (Identity & Access Management)
iii) Data Safety and Privacy
iv) Network and Transport Security
v) Infrastructure Security

These sections are further divided into various sub-sections where each section ensures security and privacy in big data with respect to the vulnerabilities, security risks & threats and, anomalous activities. The proposed multilayered framework for security and privacy is shown in Figure 5.

6.1 Data Management

Organizing, controlling and governing the enormous amount of data is known as data management, both in the form of structured and unstructured. Data management in big data refers to ensuring high data quality and availability for analytical applications. Data management is divided into four components as follows:

6.1.1 Data Classification

Organizations, government agencies, and established firms exercise big data management policies in order to cope with the expanding nature of data containing numerous terabytes or sometimes petabytes of information. In any Big Data environment, effectual data classification is considered one of the most significant aspects that result in employing efficient security regulations. Different firms manage big data by ensuring the security procedures on such a platform by distinctively identifying why data counts, which data needs to be cryptographed, and what data needs to be arranged before others for security. The following are some fundamental points that can assist in data classification matrix for any big data environment.

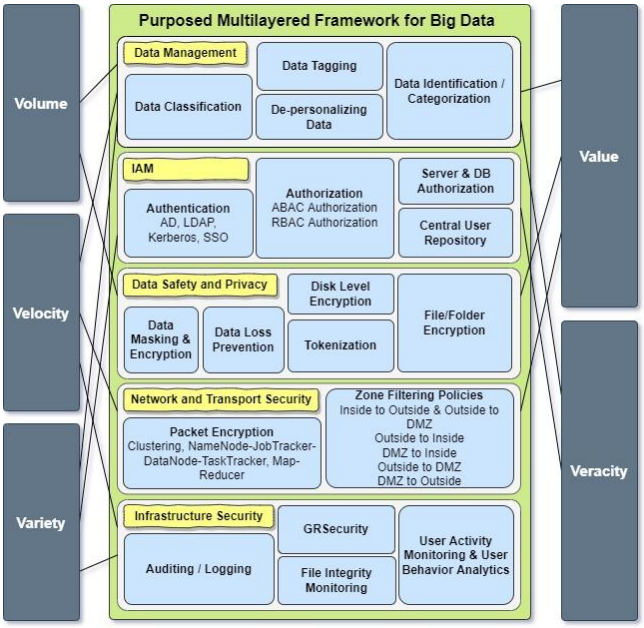

Fig 5. Proposed Multilayered Framework for Security and Privacy in Big Data

- Perform a risk assessment on classified information

Risk assessment on classified information in the context of a Big Data environment can be performed by interviewing business leaders, stakeholders, and compliance members. This can lead to achieving the classified objectives and will help in ensuring an absolute understanding of monitoring and classified requirements.

- Define an official classification policy

Official policy classification should be readily available. Instead of numerous classification policies, only three to five classification categories should be defined which will help eliminating the confusion and ambiguity, and it will also be manageable. Categorize and divide the roles and responsibilities of employees.

- Define types of data categorization

Data can be categorized according to their sensitivity. With the classified information available, challenges are also present. Identifying the sensitivity of data can be a time-consuming activity that should be performed by the business processes and process owners. Following are some of the essential questions which should be considered in the context of data privacy and protection:

  - What type of data about clients is collected by your firm?
  - What proprietary data is collected?
  - What type of transactional information is used?
  - Is all the collected information are classified and sensitive?
- Determine data location

After data types are determined by your firm, it is essential to classify how the data is stored electronically. How the data is stored and what is the procedure of data flow internally and externally? Also, services like OneDrive and Dropbox, etc. are used? Numerous data detection applications help in identifying the data storage, regardless of the data format.

- Maintenance and monitoring of data

After classification, it becomes mandatory for firms to organize systems by making updates essential. Dynamic policies and procedures should be followed for classification.

6.1.2 Data Tagging

It is necessary to comprehend the flow in a Big Data environment with all the inflow and outflow procedures. Following are some of the essential steps in data tagging:

i) The first step is the identification of all the inflow methods which contains automated and manual procedures, and also those methods that are on some meta-layer.
ii) Identifying the data inflow is also crucial as data may be coming using a CLI (Command Line Interface), or it could be entering the system by using any API.
iii) The second step is to identify all of the outflow methods in the Big Data environment.
iv) Identification of the data outflows is also an essential step as it will establish the boundaries and trust zones in the application.
v) These may include reporting jobs using JDBC/ODBC connections, using PIG or distributing it using some Restful API.

6.1.3 De-personalization Data

Depersonalization is the process of altering confidential information with the purpose of securing sensitive data in databases. Data depersonalization should be conducted effectively so that referential integrity should be maintained along with its unique features. In data depersonalization, alteration is generally accomplished using encryption, character mixing or hashing and salting.

Depersonalizing data is becoming a necessity in any Big Data environment, because of the security and privacy concerns. In the case of a huge volume of structured and unstructured data, it is essential to differentiate personal identifiers from datasets or ensure the data anonymity is achieved.

Another technique, known as Pseudonymization can be used for data privacy. It can be achieved by substituting known fields within data by one or more mock identifiers. These identifiers are known as pseudonyms. A single pseudonym also works for single fields or multiple fields that needed to be replaced. Pseudonyms also allow the re-identification of data.

6.1.4 Data Identification/Categorization

Various threats and vulnerabilities to confidential information could be exposed if there is no proper awareness. The crucial aspect is to identify the availability of the confidential data, if present then its location, and then taking essential data security steps. The following are some techniques in securing structured and unstructured data in a Big Data environment.

- Relational databases, CSV or JSON formats are used to store structured data in Hadoop. In the structured format of data, the location, identification, and categorization of confidential data are known. For securing the data, columns and fields can be protected by labeling the engine which allocates view level security to these columns and fields.
- In the case of unstructured data, the location, identification, and categorization of confidential data become challenging. In such a case, identification of data existence and location is the first step towards securing the big data environment.

  Following are the steps for data identification/categorization Big Data environment:

i) The structure and schema for the data should be defined and validated.
ii) Search procedures based on conditions should be available.
iii) Data count should be calculated, as it will help in eliminating data redundancy. For multiple records, only 1 count will be considered.
iv) These results should be analyzed to build a better threat prevention model.

6.2 IAM

IAM (Identity and access management) allows access to entities for the right resources at the right time [18]. IAM allows user access control to sensitive data within the Big Data environment. In any application, IAM comprises of following elements:

- Individuals identification and verification within a system.
- Assigning/de-assigning of roles to users.
- User management including adding, updating and removing a user and their roles.
- Access level mapping for a single user or a group consists of multiple users.
- Security of confidential information and the system itself.

IAM in Big Data environment allows organizations to extend accessibility to their confidential data without compromising security and privacy. IAM can be divided into the following components:

6.2.1 Authentication

Authentication is the process of validating any user against their IDs and password. In a Big Data environment, authentication is performed by validating individuals against their profiles created within the system. When any request to access the system is made, the authentication system validates the user profiles with the information gathered. Any wrong attempt will result in unauthorized access. Authentication in the Big Data environment can be applied by any of the following procedures:

- AD (Active Directory)

Active Directory is a user identity directory service for domain networks. Additional operational overhead and user access to nodes in the Hadoop cluster can be managed using AD.

- LDAP (Lightweight Directory Access Protocol)

LDAP is used to manage distributed directory services including AD. It is based on open standards protocols. Authentication via LDAP also helps in Hadoop clusters.

- Kerberos

Kerberos is an authentication protocol and works by allowing tickets to different nodes (Hadoop nodes) for communication against an insecure network. In this approach, the tickets are issued to the requested users from the Kerberos server, and then these tickets are used for identity verification.

- SSO (Single-Sign-On)

SSO is another authentication protocol that is used for simplifying the complexity of saving and preserving identifications for various applications. One significant benefit in a Big Data environment is that users generally need to sign in once to access the platform. After successful login, users have the privilege to access all platform related technologies without providing their credentials again.

6.2.2 Authorization

Authorization is the process of determining user access rights against a specific resource. It is generally preceded after user authentication. In the Hadoop cluster after authenticating, authorization tells what a user is allowed to do. This is generally managed by file permissions in HDFS. Authorization in the Big Data environment can be achieved by any of the following procedures:

- ABAC (Attribute Based Access Control)

It is also known as Policy-Based Access Control, in which rights access is granted to a user based on defined policies. These policies can be defined against users, objects, and resources, etc. Generally, these policies are based on Boolean logic rules e.g. IF, THEN. ABAC is also used in Hadoop and is applied to the data layer for data retrieval.

- RBAC (Role Based Access Control)

RBAC is also policy-based access control, but unlike ABAC, it is based on roles and privileges assigned to users [20]. In a Big Data system, RBAC act as a fine grain authorization control. Using RBAC, sensitive and confidential information is managed by the roles, instead of users. Rules related to roles are implemented for all data access paths by controlling groups in AD or LDAP.

6.2.3 Server & DB Authorization

Server and DB authorization in Big Data is generally implemented using ACL (Access Control Lists). These are used to associating certain permissions with user identities. Limiting granular access to HDFS files, job execution in servers and restricting access to service APIs are used to an authorization which is supported by HDFS.

From server to HDFS, when distributing queries, integrated security among various modules permit users identity to be passed throughout the Hadoop clusters. Traditional permission used in SQL Server can also be exercised for authorization in the context of Big Data. Big Data clusters are integrated with AD/LDAP using automated deployment. Once configured, user groups and pre-defined identities can be controlled throughout all endpoints.

6.2.4 Central User Repository

Central User Repository is helpful in Big Data environment as it provides storing and delivering information of user identity to various services, and also facilitates credential verification. A logical view is posed by the Central User Repository to different organizations.

Directory services in Big Data use the LDAP standard which has become a prominent figure for Central User Repository. Both virtual-directory and meta-directory are used for this. A virtual directory is used for combining numerous sets of users from various databases containing user identity information and provides a single LDAP view. On the other hand, a meta-directory is used for combining various identity sources into meta-set and provides a logical view of identity information.

6.3 Data Safety and Privacy

Data safety and its privacy are yet another important aspect of Big Data system. Many Hadoop distributers and vendor uses various techniques for the safety and privacy of data. Following are some of the core elements used in Big Data environment for safety and privacy:

6.3.1 Data Masking & Encryption

The most important aspect for any Big Data organization is the safety and privacy of their client data, if breach may result in their reputation spoiled, along with the financial damage, and it can also cost customers trust. Hence, data masking & encryption is becoming an essential requirement for the Big Data environment. Customer sensitive data including health information, personally identifiable information known as PII or any other sensitive data requires precise identification when crucial and critical data is extracted.

When employing data masking & encryption, the following points should be considered:

- Masked and encrypted information should be irretrievable.
- Not all data needed to be masked and encrypt, only sensitive information.
- The final output should demonstrate the same input or source information, and it should be repeatable.
- Referential integrity should be maintained among data.

Numerous masking & encryption practices are used, however, in a Big Data environment, the key is to maintain the format of sensitive information. Therefore, following are some of the masking procedures that can be used:

- **Substitution**: In this technique, the sensitive information is get swapped with a random but significant value.
- **Masking and Spacing**: In this technique, data is get swapped with a piece of non-meaningful information. An example is the CNIC numbers, which are used in transactions and can be interchange with XXXXX-XXXXXXX-X.
- **Date and Numbers**: In this technique, the data containing date or numbers get altered by adding/subtracting any random percentage of the original value. An example is adding/subtracting 20% of a person's salary from its original salary.
- **Encryption**: In this technique, the sensitive information is converted from a plaintext format into a ciphertext or encoded format. The plaintext format is also known as the decoded format. Data once encrypted, can only be processed after it is in decrypted format, which can be done using a decryption key. Various encryption methodologies existed, and implementing the appropriate methodology depends upon the nature of Big Data environment security and privacy policies. Encryption techniques that can be of maximum advantage include:
    - Triple DES 24-Byte Key
    - Advanced Encryption Standard (AES) 128-Bit Key
    - Secure Hash Algorithm 256 Byte (SHA256)
    - RSA – Public/Private Key

6.3.2 Data Loss Prevention

Data Loss Prevention (DLP) comprises of procedures that are useful in preventing data breach or obliteration of confidential information from undesirable resources. Various organization use prevention methodologies to secure their sensitive data. The organization uses DLP because of the following reasons:

- Protect personal identifiable information
- Accomplish information prominence
- Secure sensitive data and information in Big Data systems

Data loss is usually categorized into motion, at rest or in use. In a Big Data system, data loss generally occurs because of some internal threats, extrusion by attackers or by unintended information revelation.

Leakage of sensitive information can be stopped by adopting appropriate monitoring policies and implementing suitable monitoring methodologies. These DLP practices can be performed on either the Network side or at Host in Big Data environment. In addition, data loss blocking procedures can be used and deployed on the network for maintaining, classifying, analyzing network traffic, detecting illicit data storage and applying suitable blocking control.

On the other hand, following are some the recommendations that are useful in DLP at host:

- Securing sensitive information and data that are in motion.
- Securing sensitive information and data that are at rest.
- Securing sensitive information and data that are in use.
- Secure endpoints used for data communication with external agents.
- Use the right data leak detection application like IDS, IPS or SIEM, etc.

6.3.3 Disk Level Encryption

Disk Level Encryption is a technique used in Big Data systems to protect sensitive information placed on disks from unauthorized access by changing it into scrawled code which cannot be decrypted. Some of the advantages of Disk Level Encryption are:

- It provides transparent encryption.
- It helps in processing-based access control.
- It is used to secure metadata, log files and configurations files.
- It uses an external key manager to access data in Disk Level Encryption.

Another terminology, known as Full Disk Encryption (FDE) is also useful in encrypting files and drives stored on hard disk including operating system files. The FDE is generally performed on sectors. Encryption done using FDE requires a key to decrypt the data, even if the drive is replaced with some other will require an authentication key, without it the drives are inaccessible.

6.3.4 File/Folder Encryption

File/Folder encryption is another useful procedure for securing data in a Big Data environment by encrypting directories or folder and individual files. It also refers to as FileBased Encryption (FBE). Various software agents are used in this encryption to intersect read and write operations to disks, and then appropriate policies are used for the identification of encryption or decryption of data. Like FDE, any data stored in folders can also be encrypted. File/Folder encryption is shown in Figure 6. Different types of file/folder encryption are:

- General purpose file system
- Cryptographic file system

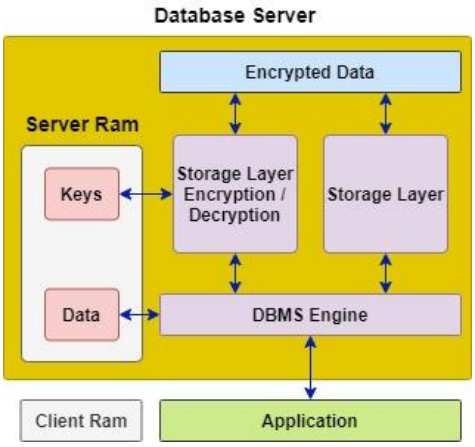

Fig 6. File/Folder Encryption

6.3.5 Tokenization

The process of replacing sensitive and critical information with distinctive classified data to maintain all the necessary information without negotiating security is known as Tokenization.

Encryption and tokenization are referred to similar techniques used for data obfuscation, but they are not quite the same. Unlike encryption, tokenization does not require any mathematical procedure for the conversion of sensitive data into tokens. In tokenization, no algorithm or key is required for obtaining the original data against a token. In this process, a database is used to store the association between the data and token. This is known as the token vault. After this, the encryption is used for securing the original information in the vault. The token value is used as an alternate value for the original information in numerous applications. If the original information is required, then the token is provided to the vault to retrieve the original value.

In Big Data applications, tokenization is used to store information like Personally Identifiable Information (PII) and Protected Health Information (PHI). By using tokenization in PII and PHI, Big Data organizations not only secure and protect sensitive information from various vulnerabilities and threats by creating them obscure, but also it helps their enterprises into conformity with industry directives and governmental principles.

6.4 Network and Transport Security

When data transmission takes place, ensuring its safety refers to network security. Many protocols have been devised for security in Big Data for the network and transport layer. These include SSL/TLS, IPSec, and SSH, etc. It can be divided into the following components:

6.4.1 Packet Encryption

While big data packet transmission in networking, a safe connection is required. There are numerous threats and vulnerabilities available for which it becomes essential to use Transport Layer Security (TLS) over Secure Socket Layer (SSL) to ensure the authentication and privacy of the packet transportation among NameNodes, DataNodes, TaskTracker, and Applications. Following are some the threat possibilities:

- Hadoop consoles can be altered by an attacker and it can have unauthorized access to information.
- This access to information could lead the attacker to retrieve users' credentials.
- Authentication tokens are assigned to users and can be used to imitate them using Kerberos authentication on the NameNode.

Some of the techniques, when implemented, ensure the protection in data as follows:

- From clients to the Hadoop cluster, using TLS, packet level encryption can be performed.
- Within clustering from NameNode-JobTracker-DataNode-TaskTracker, https can be used.
- Map-Reducer can also be used for packet level encryption.
- To avoid sniffing attacks, LDAP with SSL known as LDAPS should be used when connecting with shared enterprise indexes.

6.4.2 Zone Filtering Policies

Network security comprises imposing a firewall which must ensure the access control policy. It also identifies the access control of traffic which is allowed to pass among NameNode, DataNode, and TaskTracker, etc. In addition to this, end user could only connect to NameNode rather than specific DataNodes. Among these policies, the following are some of the recommended network zone filtering policies:

- Inside-to-Outside/DMZ

Traffic initiating from the inside could either go outside or to DMZ (demilitarized zone). This flow of data should be inspected, and if required should impose traffic restrictions.

- DMZ/Outside-to-Inside

The request should be restricted completely if it is initiating from the outside or DMZ, unless and until it is a request-response from an inside procedure.

- Outside-to-DMZ

Data may flow from the outside to DMZ. In such a case, the data traffic has to be examined using a firewall, by which the permission should be granted or denied. In case of a specific type of data traffic, it should be allowed to pass including the DNS, HTTP or HTTPS, etc.

- DMZ-to-Outside

Data traffic may travel from DMZ to outside. In this case, specific permission should be given based on the firewall and provision guidelines. Also, the firewall will open the port if data traffic is from DMZ to outside.

6.5 Infrastructure Security

Infrastructure Security is also an integral part of Big Data systems, along with other layers concerning V's security and privacy. Infrastructure security can further be decomposed into four components:

6.5.1 Auditing/Logging

Auditing is the process of logging/preserving transactions that takes place within the systems. Regulations help analyze user accesses against any particular data, and logging is also beneficial in numerous scenarios in the identification of any suspicious activities for stored information. For Big Data, every alteration made within the Hadoop clusters and nodes should be audited. Some of the examples are:

- Adding and deleting nodes

- Changes made within nodes including NameNode, JobTracker, DataNode or TaskTracker.

Data movement occurs in majority components of core Hadoop, which results in a great proportion of fragmentations. Thus, this results in a huge trail of audit logs and metadata in all fragments.

MapR is also used for maintaining any accesses made to data, along with the operations performed on various objects and implementations including commands used for MapR clusters alteration. Following are different techniques used for evaluating audit logs that are stored in MapR:

- Security Information and Event Management
- Apache Drill
- Third party tools

6.5.2 GRSecurity

Many secured OS are available for Big Data applications that are based on Linux and UNIX which have optional access rights. This gives users the privileged to act as administrative users. To cater access control rights many secured OS are available including AppArmor, SELinux, and GRSecurity.

GRSecurity comprises of patches used for improving security for Linux kernel. It is used for detecting and preventing threats and vulnerabilities on multi-layer models and is licensed under the General Public License (GPL). Framework for GRSecurity is based on Mandatory Access Control (MAC). Some of the core features of GRSecurity are:

- Without using no or low configuration, it is used to handle RBAC which can be utilized for maintaining minimal privileged policies.
- Hardening is performed of change root also known as chroot. GRSecurity also avoids /tmp racing, which can be used for preventing command injection attacks.
- Extensive auditing mechanism is available.
- Randomization in GRSecurity is based on kernel stack which allows further randomization of stack, library, and heap.
- Any executing or arbitrary code is prevented on kernel base irrespective of the mechanism used i.e. stack, library, and heap, etc.
- In the kernel, protection is available against any bugs related to the null-pointer which can be exploited.
- Limitations are available to the user to view only their tasks.
- Any alert initiated by user are stored in audit logs along with their IP.

6.5.3 File Integrity Monitoring

File Integrity Monitoring or FIM is a process of analyzing operating systems, databases and files to decide if they have tampered or not. The procedure of FIM is to check these files against a trusted baseline by comparing them. If any alteration is made on the files, databases or operating systems, then FIM will initiate an alert for an additional inquiry, and if required, a suitable correction will take place.

In many Big Data organizations, FIM is used to comply, secure and optimize the processes. In the Hadoop filesystem, the trusted or good baselines considered a standard for file monitoring which usually works on the file version, creation date, and modification date, etc. By this, the file legitimacy can be assured. Many tools like Trustwave, Tripwire, LogRythm, and Splunk, etc. [21] are available for file integrity monitoring. The benefit of using FIM in Big Data platforms are:

- Identifying illegal activity
- Locating unintended alterations
- Validate update position and also observe system health
- Sustain compliance mandates including PCI DSS and HIPAA etc.

6.5.4 User Activity Monitoring & User Behavior Analytics

Many anomalies can also occur within the system because of the users performing any abnormal activities. In Big Data organizations, applications employ User Activity Monitoring which monitor and record user performed actions. These actions usually include system commands, text typing or editing, any visited URL, any opened dialogs and every activity perform on screen. Most of the breach happened because of the weak user credentials which results in data exploitation.

User Behavior Analytics (UBA) is another term coined with User Activity Monitoring. UBA is the process of determining anomalies, including harmful threats and vulnerabilities by using detection algorithms and statistical assessment. Hadoop uses it to detect persistent and insider threats. Following are some of the behavior analytics Big Data systems should include:

- Account negotiation, account hijacking and account information sharing
- Data exfiltration
- Personal auditing
- Detection of insider threats
- Stateful session trailing
- Anomalous user behavior
- Risk analysis
- Alert generation in real-time

7. CONCLUSION

Big Data has become a hot topic in modern-day technology because of its data analysis, maintenance, and storage characteristics. This paper provides historical background about Big Data and its utilization in today's organizations. Various companies and corporations tend to use Big Data Analytics to increase customer satisfaction and retention. After providing a brief introduction, we discuss the literature review in which we determine the foundation and boundaries of a big data system to categorized the issues and concerns regarding security and privacy. After that, we proceed with various key characteristics of a big data system and categorized them into 7 V's e.g. Volume, Velocity, Variety, Veracity, Variability, Visualization, and Value. The next section then addresses the Big Data analytics that focuses on various components of Apache Hadoop which include Map-Reduce, HDFS, HBase, Hive, Cassandra, In-Memory, and NoSQL. Then we describe several security and privacy challenges posed by Big Data and how it can put sensitive data to risk. We discuss several security threats because of the insufficient conventional solutions and unavailability of many security applications, and how it is a concern for data anonymity, data integrity, and data availability. This paper also discusses privacy in a big data environment that requires different administrative roles to be working at various levels in big data analytics, and it is also required at different phases of the data i.e. acquisition/generation, sharing/storage, and processing phase. After this, the paper proposed a multilayered framework that consists of five core layer

sections protecting Big Data applications from numerous anomalies and threats. It also laid the groundwork for assessing security and privacy vulnerabilities in Big Data systems. Different V's associated with various layers including data management, identity & access management, safety and privacy, network and transport security, and finally the infrastructure security were also provided. This paper provides an insight to the reader about Big Data applications and it also discussed various challenges and the potential approaches adopted to address the big data challenges concerning security and privacy. Furthermore, this paper will also assist the research society with its intellectual support regarding research and development.

8. REFERENCES

[1] Rusi, https://www.rusi.org/downloads/assets/rusi bigdata report 2013.pdf.

[2] Elisa Bertino, "Big Data – Security and Privacy", 2015 IEEE International Congress on Big Data

[3] http://mylio.com/true-stories/next/how-many-digital-photos-will-be-taken-2017-repost

[4] http://www.internetlivestats.com/google-search-statistics/

[5] Xu, J. S., Zhang, E., Huang, C. -H., Chen, L. H. L., &Celik, N. (2014). Efficient multi-fidelity simulation optimization. Proceedings of 2014 winter simulation conference. GA: Savanna

[6] https://www.macrotrends.net/stocks/charts/AMZN/amazon/revenue

[7] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, VishanthWeerakkody, "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research 70 (2017) 263–286

[8] T.K. Das, P. Mohan Kumar, "Big Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology.

[9] Apache Hadoop, "What Is Apache Hadoop?", https://hadoop.apache.org/, Accessed on 1-Nov-2019.

[10] Abdul Ghaffar Shoro, Tariq Rahim Soomro, "Big Data Analysis: Ap Spark Perspective", Global Journal of Computer Science and Technology: C Software & Data Engineering.

[11] R. Cattell, (2010) "Scalable SQL and NoSQL Data Stores," ACM SIGMODRecord, vol. 39.

[12] L. Xu and W. Shi, "Security Theories and Practices for Big Data", Big Data Concepts, Theories, and Applications, 2016, pp. 157-192.

[13] S. Sudarsan, R. Jetley and S. Ramaswamy, "Security and Privacy of Big Data", Studies in Big Data, 2015, pp. 121-136.

[14] "Types of Network Attacks against Confidentiality, Integrity and Availability", Omnisecu.com, 2019. [Online]. Available: http://www.omnisecu.com/ccna-security/types-of-network-attacks.php. Accessed on 19-Dec-2019.

[15] Apache Cassandra Project, http://cassandra.apache.org/.

[16] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *Operating Sys. Review*, vol. 44, no. 2, pp. 35–40, 2010.

[17] https://www.guru99.com/what-is-big-data.html

[18] https://www.gartner.com/en/information-technology/glossary/identity-and-access-management-iam

[19] V. Reena Catherine, A. ShajinNargunam, Encryption Techniques to Ensure Data Confidentiality in Cloud, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 8, no. 11, 2019.

[20] https://en.wikipedia.org/wiki/Role-based_access_control

[21] https://en.wikipedia.org/wiki/File_integrity_monitoring

[22] Lewko, A., Okamoto, T., Sahai, A., Takashima, K., Waters, B.: Fully Secure Functional Encryption: Attribute-Based Encryption and (Hierarchical) Inner Product Encryption. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 62–91. Springer, Heidelberg (2010)

[23] Siddiqa, A., Karim, A. &Gani, A. Big data storage technologies: a survey, Frontiers of Information Technology & Electronic Engineering, vol 18, pp. 1040–1070(2017)

[24] W. Xindong, Z. Xingquan, W. Gong-Qing, et al. *Data Mining with Big Data*, IEEE T. Knowl. Data En., **26** (2014), 97–107.

[25] Stergiou C.L., Plageras A.P., Psannis K.E., Gupta B.B. (2020) Secure Machine Learning Scenario from Big Data in Cloud Computing via Internet of Things Network. In: Gupta B., Perez G., Agrawal D., Gupta D. (eds) Handbook of Computer Networks and Cyber Security. Springer, Cham

[26] Rajabion, Lila &Shaltooki, Abdusalam&Taghikhah, Masoud &Ghasemi, Amirhossein &Badfar, Arshad. (2019). Healthcare big data processing mechanisms: The role of cloud computing. International Journal of Information Management. 49. 271-289.

[27] S. Bahulikar, "Security measures for the big data, virtualization and the cloud infrastructure," *2016 1st India International Conference on Information Processing (IICIP)*, Delhi, 2016, pp. 1-4.

[28] N. Chitransh, C. Mehrotra and A. S. Singh, "Risk for big data in the cloud," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, 2017, pp. 277-282

[29] C. Liu, C. Yang, X. Zhang, et al. External integrity verification for outsourced big data in cloud and IoT: a big picture, Future Gener. Comp. Sy., 49 (2015), 58–67.

[30] S. Arora, M. Kumar, P. Johri and S. Das, "Big heterogeneous data and its security: A survey," 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, 2016, pp. 37-40.

[31] I. de la Torre-Díez, B. Garcia-Zapirain, M. Lopez-Coronado, et al. Proposing telecardiology services on cloud for different medical institutions: a model of reference, Telemedicine and e-Health, 23 (2017), 654–661.

[32] Mishra K.N., Chakraborty C. (2020) A Novel Approach Towards Using Big Data and IoT for Improving the Efficiency of m-Health Systems. In: Gupta D., Hassanien

A., Khanna A. (eds) Advanced Computational Intelligence Techniques for Virtual Reality in Healthcare. Studies in Computational Intelligence, vol 875. Springer, Cham

[33] S. Agarwal, M. Gupta and A. Sharma, "Big Data Privacy Issues & Solutions," *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2019, pp. 225-228.

[34] K. R. Sollins, "IoT Big Data Security and Privacy Versus Innovation," in *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1628-1635, April 2019.

[35] J. Moreno, E. B. Fernandez, M. A. Serrano and E. Fernández-Medina, "Secure Development of Big Data Ecosystems," in *IEEE Access*, vol. 7, pp. 96604-96619, 2019.

[36] J. Andrew, J. Karthikeyan and J. Jebastin, "Privacy Preserving Big Data Publication On Cloud Using Mondrian Anonymization Techniques and Deep Neural Networks," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 722-727