IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

Review: Big Data Privacy and Security Risk and Solutions

Madhavi Tota
Assistant Professor
Rajiv Gandhi College of Engineering, Research and Technology, Chandrapur

Abstract: *Big Data is very dynamic issues in the current year, enables computing resources as a data to be provided as Information Technology services with high efficiency and effectiveness. The high amount of data in world is growing day by day. Data is growing very rapidly because of use of internet, smart phone and social network. Now size of the data is in Petabyte and Exabyte. Traditional database systems are not able to capture, store and analyze this large amount of data. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the limits. However, the current scenario the growth rate of such large data creates number of challenges, such as the fast growth of data, access speed, diverse data, and security. This paper shows the fundamental concepts of Big Data. Privacy threats and security methods used in Big Data. With the development of various research application and recourses of Internet/Mobile Internet, social networks, Internet of Things, big data has become the very important topic of research across the world, at the same time, big data has security risks and privacy protection during different stages such as collecting, storing, analyzing and utilizing. This paper introduces security measures of big data, then proposes the technology to solve the security threats.*

Keywords: Big Data, security risks, information security, information security technology, DP.

I. INTRODUCTION

Big Data is used to explain large amounts of structured and unstructured and semi-structured data that are so large and it is very difficult to process by using traditional databases and software technologies. Due to the development of new technologies like social networking sites, the amount of data produced by mankind is growing rapidly every year. Big data means large data; it is a collection of large datasets that cannot be processed by using traditional techniques. Big data is not just a data; rather it has become a complete subject, which involves various tools, techniques and frameworks.

The rapid growth rate of data is expected to be double every years, from 2500 Exabytes in 2012 to 40,000 Exabytes in 2020. At the KDD BigMine 12 Workshop Usama Fayyad in his invited talk presented following data numbers about internet usage, that is each day Google has more than 1 billion queries, Twitter has more than 250 million tweets per day, Per day Face book has more than 800 million updates, and YouTube has more than 4 billion views per day. Big Data is a heterogeneous mix of both structured data (traditional datasets –in rows and columns like DBMS tables, CSV"s and XLS"s) and unstructured data like PDF documents, e-mail attachments, images, manuals , medical records such as x-rays, ECG and MRI images, rich media like graphics, audios and videos, contacts, forms and documents. Businesses are primarily concerned with managing unstructured data, because about 80 percent of enterprise data is unstructured.

II. CHARACTERISTICS OF BIG DATA

Big data has certain specific characteristics with which it can separate from the normal data. There are six key characteristics that define big data. These characteristics are also known as Six V"s of big data[1].

- **Volume:** It refers to the vast amount of data generated rapidly in every second. Many factors contribute towards increasing volume such as storing transaction data, live streaming data and data collected from

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

sensors, human interaction on systems like social media etc. Latest big data tools use distributed system so that it can store and analyse data across databases that are dotted around anywhere in the world. The quantity of data generated is not in terabytes but in Petabyte or zettabytes.

- **Variety**: Variety Refers to the different type of data that is being stored. Today data comes in different types of formats from different sources. With the invention of sensors, smart devices and social media technologies, data is being generated in countless forms, including text, web data, tweets, sensor data, audio, video, click streams, log files and more. There are three categories on the variety of data, structured data (relational databases), semi-structured data (xml data) and unstructured data (text and multimedia contents).
- **Velocity:** It means how fast the data is being produced and how fast the data needs to be processed to meet the demand. It refers to the speed at which new data is generated and the speed at which data moves around. Like by using social media messages are going viral within seconds. New Technology allows us now to analyze the data while it is being generated known as in-memory analytics, without storing it into databases.
- **Veracity:** It refers to the level of reliability associated with certain types of data. Veracity means data uncertainty and unpredictability. The requirement for high data quality is an important Big Data has big challenge, but even the best data cleansing methods cannot purify the inherent unpredictability of some data, like the weather, the economy, or a customer's buying decisions.
- **Value:** Every data is important and carries Value. Good information may be hidden in unstructured non-traditional data. The challenge is identifying what is valuable and then transforming and extracting that data for analysis.
- **Variablity:** It refers to the messiness, inconsistency or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable but technology now allows us to work with this type of data. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent.

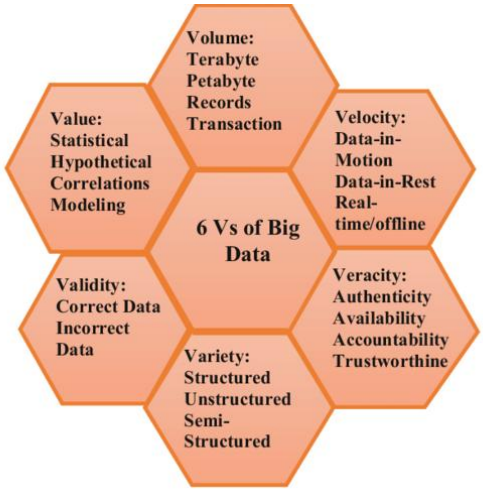

Figure 2.1: Important characteristics of Big Data

III. Big Data Security Threats

Security issues and the value brought by big data become very challenging task for reseachers. Compared with the traditional information security issues, the challenge for big data security is mainly reflected in the following aspects [3].

3.1. Big Data Increase the Risk of Privacy Leakage

The collection of big data definitely increases the risk of leakage of user privacy. Because it contains a large number of user information in the data, the development and utilization of big data is very easy to release the privacy of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

citizens, the technical threshold is greatly reduced. The privacy of citizens will be used and sold poisonously. The cause of information leakage is various, it can be summarized as follows:

1. The abuse of data leads to the privacy leakage Big data relates to several areas and departments, and relates to data collecting, storing, processing, analyzing, reporting and other processes, in this process, if the abuse of data will cause privacy leakage. For example, internal staff of data centre do not abide by the professional ethics, abuse their functions and powers, release the data of the sensitive department, persons and events to the public, resulted in privacy leakage. When big data is shared among multiple relational departments, this should be also a cause of privacy leakage. Data has a life cycle, at the end of its life cycle, if it does not be effectively destroyed, sniffer can get the data by social engineering, results in privacy leakage.
2. Analysis and use of data causes privacy leakage One of the distinctive features of big data is a huge amount of data, the diversity of data source. When analyzing the data, data from different sources are integrated, it may get an unexpected result which can't be got from a single data source. A classic case of privacy leakage in big data environment and mailed for business . All these are obtained from the analysis of the sales records.

3.2. Big Data Becomes the Carrier of Advanced Persistent Threat

APT is a kind of advanced persistent threat, It's a long time of the attack, the attack process is complex and difficult to find. The main features of APT is that the space to attack is very wide, long duration, single point concealment ability is very strong. Traditional protection policies are unable to detect hacker attacks behind big data. Traditional threat detection is based on a real-time feature matching detection on a single point, but APT is an ongoing process, without obvious feature to be detected in real-time, so it can't be detected in real time. Meanwhile, APT code hidden in big data is also difficult to find. In addition, an attacker can also use social network and system vulnerabilities to attack. Hackers can use big data to expand the attack effect, this is mainly reflected in three aspects:

1. Hackers can launch a botnet attack using big data server, they may control millions of puppet machine and attack, a single point attack has not such aggressively.
2. Hackers can enlarge attack effect by controlling the key nodes;
3. Hackers can hide data for attacking in large data, which makes difficulty to analysis of security vendors. Any misguiding hacker set, will lead to deviation from the proper direction of safety monitoring.

3.3. Big Data Is Not Necessarily Credible

Because big data is original, the general view is that the big data is true and reliable. In fact, this is not necessarily true, as people can't always believe their eyes. An important factor affecting the credibility of big data is the correct level of data, If the data comes from the real and reliable production processes, then these data is credible, but if these data is created for a special purpose, so these data is false, because incorrect data will lead to erroneous conclusions . For example, in some review sites, the real reviews information and bogus reviews mixed together, the user is difficult to distinguish between true and false, and sometimes make wrong judgments based on false reviews, to choose inferior products and services. Another factor is the gradual distortion in the data dissemination. Data collection process under manual intervention may introduce errors, data distortion due to errors and deviations, and it ultimately affects the accuracy of the results of data analysis.

3.4. How to Achieve Access Control for Big Data

Access control is an effective and important method to realize data controlled sharing, it is divided into different access control, mandatory access control and role-based access control. While in big data environment, it is difficult to preset the role, to realize the role and to predict the actual authority of each role. Discretionary access control is unable to meet the diversity of the permissions due to the diversity of users, mandatory access control is unable to meet the dynamics of authority, role-based access control is not able to effectively match the role and the corresponding permissions. Therefore, a new security access control mechanisms must be adopted to protect data in big data environment.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

IV. Big Data Security Challenges

Challenges are not limited to only one platforms. They also affect the cloud. The list below reviews the four most common challenges of big data on-premises and in the cloud[2][5].

4.1 Distributed Data

Many big data frameworks divides data processing tasks throughout many systems for faster analysis. Hadoop, for example, is a popular open-source framework for distributed data processing and storage. Hadoop was originally designed without any security in mind. Cybercriminals can force the MapReduce mapper to show incorrect lists of values or key pairs, making the MapReduce process worthless. Distributed processing may reduce the workload on a system, but at the same time more systems mean more security issues.

4.2 Endpoint Vulnerabilities

Cybercriminals can change and manipulate data on endpoint devices and transfer the false data. Security solutions that analyze logs from endpoints need to validate the authenticity of those endpoints. For example, hackers can access manufacturing systems that use sensors to detect malfunctions in the processes. After gaining access, hackers make the sensors show fake results. Challenges like that are usually solved with fraud detection technologies.

4.3 Data Mining Solutions

Data mining is the heart of many big data environments. Data mining tools find patterns in unstructured data. The problem is that data often contains personal and financial information. For that reason, companies need to add extra security layers to protect against external and internal threats.

4.4 Access Controls

Companies need to restrict access to sensitive data like medical records that include personal information. But people that do not have access permission, such as medical researchers, still need to use this data. The solution in many organizations is to grant granular access. This means that individuals can access and see only the information they need to see. The aspects of Infrastructure Security, Data Privacy, Data Management and, Integrity and Reactive Security. Each of these aspects faces the following security challenges are [4] :

4.4.1 Infrastructure Security

1. Secure Distributed Processing of Data
2. Security Best Actions for Non-Relational Data-Bases

4.4.2 Data Privacy

1. Data Analysis through Data Mining Preserving Data Privacy
2. Cryptographic Solutions for Data Security
3. Granular Access Control

4.4.3 Data Management and Integrity

1. Secure Data Storage and Transaction Logs
2. Granular Audits
3. Data Provenance
4. Reactive Security
5. End-to-End Filtering & Validation
6. Supervising the Security Level in Real-Time

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

V. SECURITY TECHNOLOGY IN BIG DATA

For the security issues of big data, need to address the security issues of big data from the following points: data privacy protection technology; data integrity and trusted technology; access control technology [9].

5.1. Data Privacy Protection Technology

To protect the privacy of big data, even if the data with privacy leak, the attacker can't obtain the effective value of data. It can use data encryption and Data anonymity technology.

1. **Data Encryption Technology:** Data encryption technology is an important means to protect data confidentiality, it safeguards the confidentiality of the data, but it cut down the performance of the system at the same time. The data processing ability of big data system is fast and efficient, which can satisfy the requirements of the hardware and software required for encryption. So the homomorphic encryption has become a research hotspot in data privacy protection. The homomorphic encryption is a model for the calculation of the cipher text, avoiding the encryption and decryption in the unreliable environment, and directly operation on the cipher text. Which is equivalent to the procedure of processing the data after decryption, then encrypting it. Homomorphic encryption is still in the exploratory stage, the algorithm is immature, low efficiency, and there is a certain distance away from practical application.
2. **Data Anonymity Technology:** Data anonymity is another important technology for privacy protection, Even if the attacker gets the data that contains the privacy, he can't get the original exact data, because the value of the key field is hidden. However, in the background of big data, the attacker can obtain data from multiple sources, then associate the data from one source with another source, then will find the original meaning of the hidden data.
3. **Generalization Technology:** The third technology of privacy protection is generalization technology, which is to generalize the original data, so that the data is fuzzy, so as to achieve the purpose of privacy protection. For example: I live in 3-13-2 No. 187 Guangyuanli Lane Xuanwu District Beijing. This address is very detailed, now we change the address to the city of Beijing, so that the value of address has become vague, so as to achieve the purpose of privacy protection.

5.2. Access Control Technology

Big data contains a wealth of information resources, all professions and trades have great demand of the data, so we must manage access rights of big data carefully. Access control is an effective means to achieve controlled sharing of data, but in big data environment, the number of users is huge, the authority is complex, and a new technology must be adopted to realize the controlled sharing of data.

1. Role Mining Role-based access control (RBAC) is an access control model used widely. By assigning roles to users, roles related to permissions set, to achieve user authorization, to simplify rights management, in order to achieve privacy protection. In the early, RBAC rights management applied "top-down" mode: According to the enterprise's position to establish roles, When applied to big data scene, the researchers began to focus on "bottom-up" mode, that is based on the existing "Users - Object" authorization, design algorithms automatically extract and optimization of roles, called role mining. In the big data scene, using role mining techniques, roles can be automatically generated based on the user's access records, efficiently provide personalized data services for mass users. It can also be used to detect potentially dangerous that user's behavior deviates from the daily behavior. But role mining technology are based on the exact, closed data set, when applied to big data scene, we need to solve the special problems: the dynamic changes and the quality of the data set is not higher.

VI. PRIVACY PRESERVING TECHNIQUES IN BIG DATA ENVIRONMENT

Some methods for privacy preserving in big data is described here. These methods being used traditionally provide privacy to a certain amount but their disadvantages led to the advent of newer methods[6].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

6.1 De-identification

De-identification [9, 10] is a traditional technique for privacy-preserving data mining, where in order to protect individual privacy, data should be first sanitized with generalization (replacing quasi- identifiers with less particular but semantically consistent values) and suppression (not releasing some values at all) before the release for data mining.

Mitigate the threats from re-identification; the concepts of k-anonymity [2, 1, 3], l-diversity [10, 11, 13] and t-closeness [9, 3] have been introduced to enhance traditional privacy-preserving data mining. De-identification is a crucial tool in privacy protection, and can be migrated to privacy preserving big data analytics. Nonetheless, as an attacker can possibly get more external information assistance for de-identification in the big data, we have to be aware that big data can also increase the risk of re-identification. As a result, de-identification is not sufficient for protecting big data privacy.

- Privacy-preserving big data analytics is still challenging due to either the issues of flexibility along with effectiveness or the de-identification risks.
- De-identification is more feasible for privacy-preserving big data analytics if develop efficient privacy-preserving algorithms to help mitigate the risk of re-identification. There are three -privacy-preserving methods of De-identification, namely, K-anonymity, L-diversity and T-closeness. There are some common terms used in the privacy field of these methods:
- Identifier attributes include information that uniquely and directly distinguish individuals such as full name, driver license, social security number.
- Quasi-identifier attributes means a set of information, for example, gender, age, date of birth, zip code. That can be combined with other external data in order to re-identify individuals.
- Sensitive attributes are private and personal information. Examples include, sickness, salary, etc.
- Insensitive attributes are the general and the innocuous information.
- Equivalence classes are sets of all records that consists of the same values on the quasi-identifiers.

6.2 K-anonymity

A release of data is said to have the k-anonymity [9, 3] property if the information for each person contained in the release cannot be perceived from at least k-1 individuals whose information show up in the release. In the context of k-anonymization problems, a database is a table which consists of n rows and m columns, where each row of the table represents a record relating to a particular individual from a populace and the entries in the different rows need not be unique. The values in the different columns are the values of attributes connected with the members of the population. There are six attributes along with ten records in this data. There are two regular techniques for accomplishing k-anonymity for some value of k.

On the positive side, it will present a greedy O(k log k)-approximation algorithm for optimal k-anonymity via suppression of entries. The complexity of rendering relations of private records k-anonymous, while minimizing the amount of information that is not released and simultaneously ensure the anonymity of individuals up to a group of size k, and withhold a minimum amount of information to achieve this privacy level and this optimization problem is NP-hard. In general, a further restriction of the problem where attributes are suppressed instead of individual entries is also NP hard [5]. Therefore we move towards L-diversity strategy of data anonymization.

6.3 L-diversity

It is a form of group based anonymization that is utilized to safeguard privacy in data sets by reducing the granularity of data representation. This decrease is a trade-off that results outcomes in some loss of viability of data management or mining algorithms for gaining some privacy. The l-diversity model (Distinct, Entropy, Recursive) [9, 1, 4] is an extension of the k-anonymity model which diminishes the granularity of data representation utilizing methods including generalization and suppression in a way that any given record maps onto at least k different records in the data. The l-diversity model handles a few of the weaknesses in the k-anonymity model in which protected identities to the level of k-individuals is not equal to protecting the corresponding sensitive values that were generalized or

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

suppressed, particularly when the sensitive values in a group exhibit homogeneity. The l-diversity model includes the promotion of intra-group diversity for sensitive values in the anonymization mechanism. The problem with this method is that it depends upon the range of sensitive attribute. If want to make data L-diverse though sensitive attribute has not as much as different values, fictitious data to be inserted. This fictitious data will improve the security but may result in problems amid analysis. Also L-diversity method is subject to skewness and similarity attack [4] and thus can't prevent attribute disclosure.

6.4 T-closeness

It is a further improvement of l-diversity group based anonymization that is used to preserve privacy in data sets by decreasing the granularity of a data representation. This reduction is a trade-off that results in some loss of adequacy of data management or mining algorithms in order to gain some privacy. The t-closeness model(Equal/Hierarchical distance) [9, 3] extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute. An equivalence class is said to have t-closeness if the distance between the conveyance of a sensitive attribute in this class and the distribution of the attribute in the whole table is less than a threshold t[14].

Comparative Analysis of De-Identification Privacy Methods:

Advanced data analytics can extricate valuable information from big data but at the same time it poses a big risk to the users' privacy [12]. There have been numerous proposed approaches to preserve privacy before, during, and after analytics process on the big data. This paper discusses three privacy methods such as K-anonymity, L-diversity, and T-closeness. As consumer's data continues to grow rapidly and technologies are unremittingly improving, the trade-off between privacy breaching and preserving will turn out to be more intense.

6.6 HybrEx

Hybrid execution model [7] is a model for confidentiality and privacy in cloud computing. It executes public clouds only for operations which are safe while integrating an organization's private cloud, i.e., it utilizes public clouds only for non-sensitive data and computation of an organization classified as public, whereas for an organization's sensitive, private, data and computation, the model utilizes their private cloud. It considers data sensitivity before a job's execution. It provides integration with safety. The four categories in which HybrEx MapReduce enables new kinds of applications that utilize both public and private clouds are as follows-

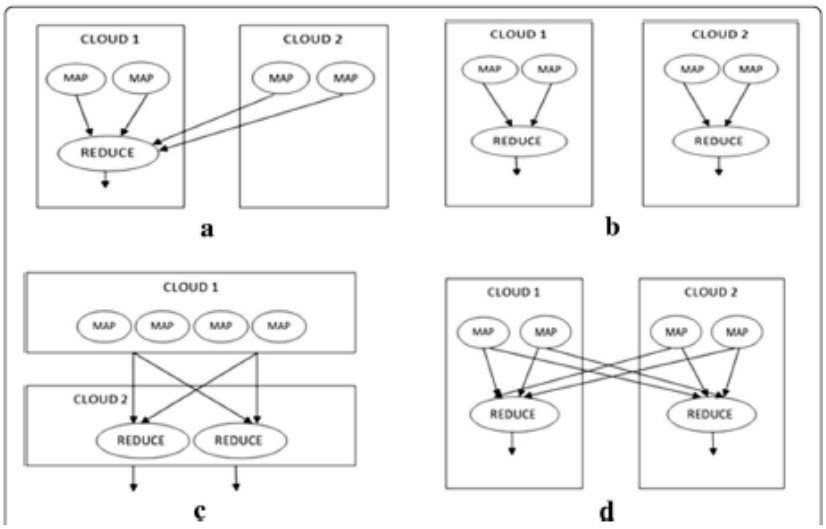

Figure 6.2.1: HyberEx methods.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

1. Map hybrid The map phase is executed in both the public and the private clouds while the reduce phase is executed in only one of the clouds as shown in Fig. a.
2. Vertical partitioning It is shown in Fig. b. Map and reduce tasks are executed in the public cloud using public data as the input, shuffle intermediate data amongst them, and store the result in the public cloud. The same work is done in the private cloud with private data. The jobs are processed in isolation.
3. Horizontal partitioning The Map phase is executed at public clouds only while the reduce phase is executed at a private cloud as can be seen in Fig. c.
4. Hybrid As in the figure shown in Fig. d, the map phase and the reduce phase are executed on both public and private clouds. Data transmission among the clouds is also possible.

Integrity check models of full integrity and quick integrity checking are suggested as well. The problem with HybridEx is that it does not deal with the key that is generated at public and private clouds in the map phase and that it deals with only cloud as an adversary.

6.7 Privacy Preserving Aggregation

Privacy-preserving aggregation [8] is built on homomorphic encryption used as a popular data collecting technique for event statistics. Given a homomorphic public key encryption algorithm, different sources can use the same public key to encrypt their individual data into cipher texts [9]. These cipher texts can be aggregated, and the aggregated result can be recovered with the corresponding private key. But, aggregation is purpose-specific. So, privacy- preserving aggregation can protect individual privacy in the phases of big data collecting and storing. Because of its inflexibility, it cannot run complex data mining to exploit new knowledge. As such, privacy-preserving aggregation is insufficient for big data analytics.

6.8 Differential Privacy

Differential Privacy [10] is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals. This is done by introducing a minimum distraction in the information provided by the database system. The distraction introduced is large enough so that they protect the privacy and at the same time small enough so that the information provided to analyst is still useful. Earlier some techniques have been used to protect the privacy, but proved to be unsuccessful.

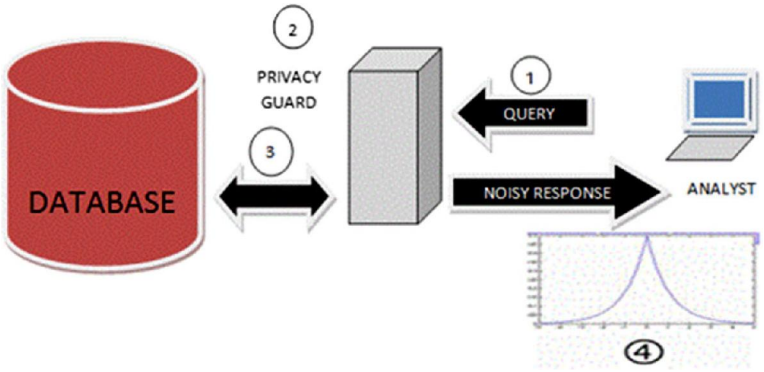

Figure 6.8: Differential privacy as solution to privacy preserving in big data

Differential Privacy (DP) deals to provide the solution to this problem as shown Fig. 6.8. In DP analyst are not provided the direct access to the database containing personal information. An intermediary piece of software is introduced between the database and the analyst to protect the privacy. This intermediary software is also called as the privacy guard.

- Step 1 The analyst can make a query to the database through this intermediary privacy guard.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

- Step 2 The privacy guard takes the query from the analyst and evaluates this query and other earlier queries for the privacy risk. After evaluation of privacy risk.
- Step 3 The privacy guard then gets the answer from the database.
- Step 4 Add some distortion to it according to the evaluated privacy risk and finally provide it to the analyst.

The amount of distortion added to the pure data is proportional to the evaluated privacy risk. If the privacy risk is low, distortion added is small enough so that it do not affect the quality of answer, but large enough that they protect the individual privacy of database. But if the privacy risk is high then more distortion is added.

6.9 Identity Based Anonymization

These techniques encountered issues when successfully combined anonymization, privacy protection, and big data techniques [11] to analyse usage data while protecting the identities of users. Intel Human Factors Engineering team wanted to use web page access logs and big data tools to enhance convenience of Intel's heavily used internal web portal. To protect Intel employees' privacy, they were required to remove personally identifying information (PII) from the portal's usage log repository but in a way that did not influence the utilization of big data tools to do analysis or the ability to re-identify a log entry in order to investigate unusual behaviour. Cloud computing is a type of large-scale distributed computing paradigms which has become a driving force for Information and Communications Technology over the past several years, due to its innovative and promising vision. It provides the possibility of improving IT systems management and is changing the way in which hardware and software are designed, purchased, and utilized.

Cloud storage service brings significant benefits to data owners, say, (1) reducing cloud users' burden of storage management and equipment maintenance, (2) avoiding investing a large amount of hardware and software, (3) enabling the data access independent of geographical position, (4) accessing data at any time and from anywhere [12]. To meet these objectives, Intel created an open architecture for anonymization [11] that allowed a variety of tools to be utilized for both de-identifying and re-identifying web log records. In the process of implementing architecture, found that enterprise data has properties different from the standard examples in anonymization literature [13]. This concept showed that big data techniques could yield benefits in the enterprise environment even when working on anonymized data. Intel also found that despite masking obvious Personal Identification Information like usernames and IP addresses, the anonymized data was defenceless against correlation attacks. They explored the trade-offs of correcting these vulnerabilities and found that User Agent (Browser/OS) information strongly correlates to individual users. This is a case study of anonymization implementation in an enterprise, describing requirements, implementation, and experiences encountered when utilizing anonymization to protect privacy in enterprise data analysed using big data techniques. This investigation of the quality of anonymization used k-anonymity based metrics. Intel used Hadoop to analyse the anonymized data and acquire valuable results for the Human Factors analysts [14, 15]. At the same time, learned that anonymization needs to be more than simply masking or generalizing certain fields—anonymized datasets need to be carefully analysed to determine whether they are vulnerable to attack.

VII. CONCLUSION

Big data [2, 8] is analysed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses. Yet only a small percentage of data is actually analysed. In this paper, a brief explanation about the privacy challenges in big data by first identifying big data privacy requirements and then discussing whether existing privacy preserving techniques are sufficient for big data processing or not. Privacy challenges in each phase of big data life cycle [7] are presented along with the advantages and disadvantages of existing privacy-preserving technologies in the context of big data applications. This paper also presents traditional as well as recent techniques of privacy preserving methods in big data. Concepts of identity based anonymization [11] and differential privacy [10] and comparative study between various recent techniques of big data privacy are also discussed. It presents scalable anonymization methods [9] within the MapReduce framework based on cloud. It can be easily scaled up by increasing the number of mappers and reducers. By extending privacy preserving methods using Differential privacy, which has a power to prevent some of the security risk but at last it will has some demerits and need to improve by adding different learning techniques.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 4, Issue 1, April 2021
Impact Factor: 4.819

VIII. Future Scope

With the future direction, are needed to achieve effective and efficient solutions to the scalability problem [10] of privacy and security in the area of big data and especially to the problem of security and privacy models. In terms of healthcare services [9, 4–7] and many more applications of big data need more efficient privacy techniques to be developed. Differential privacy is one such sphere which has got much of hidden potential to be utilized further. Also with the rapid development of IoT, there are lots of challenges when IoT and big data come; the quantity of data is big but the quality is low and the data are various from different data sources inherently possessing a great many different types and representation forms, and the data is heterogeneous, as-structured, semi structured, and even entirely unstructured [11]. This poses new privacy challenges and open research issues. So, different methods of privacy preserving mining may be studied and implemented in future. As such, there exists a huge scope for further research in privacy preserving methods in big data.

References

[1] Pooja Bisht, Kulvinder Singh Computer Science, Uttarakhand Technical University, India " Big Data Security: A Review of Big Data, Security Issues and Solutions" International Journal of Computer Science and Mobile Computing (IJCSMC) Vol. 5, Issue. 7, July 2016, pg.142 – 147.

[2] https://www.dataversity.net/big-data-security-challenges-and-solutions/#:~:text=Attacks%20on%20big%20data%20systems,and%20can%20crash%20a%20system.

[3] Gang Zeng " Big Data and Information Security" , International Journal of Computational Engineering Research (IJCER) ISSN (e): 2250 – 3005 || Volume, 05 || Issue, 06 || June – 2015

[4] José Moura1, Carlos Serrão1 ISCTE-IUL, Instituto Universitário de Lisboa, Portugal" Security and Privacy Issues of Big Data".

[5] Renu Bhandari, Vaibhav Hans and Neelu Jyothi Ahuja "Big Data Security – Challenges and Recommendations", International Journal of Computer Sciences and Engineering Volume-4 , Issue-1 E-ISSN: 2347-2693

[6] Priyank Jain, Manasi Gyanchandani and Nilay Khare "Big data privacy: a technological perspective and review" , Jain et al. J Big Data (2016) 3:25DOI 10.1186/s40537-016-0059-y

[7] P. Ram Mohan Rao , S. Murali Krishna and A. P. Siva Kumar "Privacy preservation techniques in big data analytics: a survey" , Journal of Big Data (Open Access)

[8] Priyanka Gawali , Dhananjay Gawali "Big data privacy preservation using K Anonymization and l-Diversity", IJSART - Volume 2 Issue 11 –NOVEMBER 2016 ISSN [ONLINE]: 2395-1052

[9] Ninghui Li Tiancheng Li Department of Computer Science, Purdue University "t-Closeness: Privacy Beyond k-Anonymity and _-Diversity".

[10] Nandhini.P "A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce", Journal of Engineering Research and Application www.ijera.com ISSN : 2248-9622, Vol. 8, Issue4 (Part -III) April 2018, pp65-78.

[11] Raghav Toshniwal ," Big Data Security Issues and Challenges" , International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2.

[12] Murat Kantarcioglu , "Research Challenges at the Intersection of Big Data, Security and Privacy", Specialty Grand Challenge Article Front. Big Data, 14 February 2019 | https://doi.org/10.3389/fdata.2019.00001

[13] Neha Srivastava; Umesh Chandra Jaiswal, "Big Data Analytics Technique in Cyber Security: A Review", 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)

[14] Bhawna Gupta and Kiran Jyoti, "Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data", (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 5, no. 3, pp. 3867-3870

[15] Alvaro A. Cárdenas, Pratyusa K. Manadhata and Sree Rajan, "Big data analytics for security intelligence", University of Texas at Dallas@ Cloud Security Alliance, pp. 1-2