GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

PRIVACY PRESERVATION TECHNIQUES FOR PERSONALIZED DATA IN BIG DATA

M. Santhiya Devi[*1] & Dr.K.Arunesh[2]
[*1] Research Scholar, [2]Associate Professor,
[2] Department of Computer Science,
[1] Madurai Kamaraj university – Madurai
[2] Sri S.Ramasamy Naidu Memorial College – Sattur

ABSTRACT

The recent advancements in this digital world huge amount of information are generated and shared, and the management of such large data is the most difficult and challenging task. Due to its size and variety of data, its name big data was derived. In the management of this data, some information may be disclosed. This type of disclosure can lead to leakage of Personal Identifiable Information (PII), as it contains individual's information. The voluminous data generated from the various sources can be processed and analyzed to support decision making. However, data analytics is prone to privacy violations. Due to this, privacy has become one a major challenge of big data. There are different approaches to privacy preservation is processed like encryption based methodology, anonymization based method and noise based techniques in big data. Here the issues and methods are discussed and successfully analysed with privacy preservation and Big Data. This research paper examined various privacy threats, privacy preservation techniques and models with their limitations, and the analysis shows which one is the best method that guaranteed true privacy.

Keywords: *Anonymization, Big Data, Encryption, Privacy Preservation, Privacy Threats*.

I. INTRODUCTION

In the past years, the amount of data created by the human has increased tremendously. From 2005 to 2020, the total amount of data is increased 300 times, from 130 exabytes to 40,000 Exabyte. Big Data is due to the drastic increase in data [7]. The data generated are from handheld devices and machine communication, and also with the network sites [6]. Due to its volume, velocity and variety, the traditional security approaches fail to handle the challenges of big data. There are many issues and challenges in big data, but privacy is the most important issues. As big data generally consist of person-specific information and such information is used on the web. This type of unsecured data can lead to the leakage of Personal Identifiable Information (PII). This will cause users to lose faith in firms [3]. In today technology world there is an exponential growth in volume and variety of data as due to diverse applications of computers in all domain areas. The growth has been achieved due to the affordable availability of computer technology, storage, and network connectivity. The large-scale data, which also include person specific private and sensitive data like gender, zip code, disease, caste, shopping cart, religion etc. is being stored in the public domain. The data holder can release this data to a third party data analyst to gain deeper insights and identify hidden patterns which are useful in making important decisions that may help in improving businesses, provide value-added services to customers, prediction, forecasting and recommendation [10].

In generally, Cryptography is a method which uses algorithms and other techniques for protecting the data. In the cryptography the original text is converted into cipher text using the encryption schemes [11]. This technique alone can't implement the privacy demanded by big data services. The data which are useful as well as sensitive are encrypted so that the user can use those data. Data anonymization changes data that are used or published in such a way that it prevents the identification of the data. Anonymization generally refers to the hiding of identifier and attributes [5]. With anonymization the data are vulnerable to various attacks such as unsorted matching attack, complementary release attack. Notice and consent is another method for privacy preservation. In this methodwhenever a person accesses the application a notice for privacy is displayed. The user needs to consent the privacy notice before using the application. It requires changing the notice every time it is used for a variety of purposes. The consent andnotice becomes a burden for the user as big data is processed and used by a large amount of users. With differential privacy, the user can extract useful data from the database and also provides privacy. This minimizes the chances of identification of individuals. In opposite with anonymization, data are not changed in differential privacy. Users cannot have direct access with database [8].

The structure of the paper is formatted as Section I provide a general introduction related to the big data and its preserving techniques. Section II deals with the related works of big data preserving techniques. Section III discusses the big data challenges with its privacy. Section IV design the methodology of the preserving techniques for personalized data in big data. Section V deals with the comparative analysis of the techniques. The conclusion is defined as the next section VI.

II. LITERATURE REVIEW

P. Ram Mohan Rao et al., [12] examines various privacy threats, privacy preservation techniques and models with their limitations. And also the author proposes a data lake based modernistic privacy preservation technique to handle privacy preservation in unstructured data. They propose a novel privacy preservation model based on Data Lake concept to hold the variety of data from diverse sources. The authors discussed that the algorithm will be trained with existing data sets with known sensitive attributes and rigorous training of the model will help in predicting the sensitive attributes in a given data set.

Mohammad Tarique Mohammad Salem and A.P.Kankale [9] viewed privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, the authors identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker.

III. BIG DATA CHALLENGES AND PRIVACY

A collection of large and complex data which are unmanageable for database management and processing is known as Big Data. The big data issues and challenges with its privacy are explained in the following section.

3.1 Big Data

With the information growth in the data analytics field, big data has become the most known emerging technology. It has the capacity to provide understanding of the unseen features of the data analysis. Hence this asset has become more profitable and makes the business management smarter [1]. Traditional data analysis and management are based on the relational database management system (RDBMS). But, RDBMSs does not apply to unstructured data and semi-structured but only to structured data. For this, the solution for permanent storage and management of large-scale datasets, distributed file systems, the SQL databases will be suitable. In generally Big data can be characterized by,

- Volume, size of data
- Variety, structured or unstructured
- Velocity, static or stream data
- Veracity, accuracy and reliability of data [4].

3.2 Big Data Challenges

1. Storage challenge: Storage system should be capable of storing ever-increasing data every day [4].
2. The Processing challenge: The technology should be used such that it should be fast and yield error-free data.
3. Privacy and Security: At the time of data sharing or retrieving the leakage of data should be prevented.

3.3 Big Data Privacy

The common meaning of data privacy is preventing the leakage of sensitive information.
1. Context-Based Privacy: Deciding what type of data is present and providing privacy for them [4].

2. Co-related and Aggregated Datasets: Considering the side knowledge of each data and providing privacy to those data.
3. Threat Modeling: With the threat model the threats and attacks to data should be identified and a secure technique must be provided.
4. Privacy Budgeting: For organizing such a large amount of data deciding the cost to be spent is difficult.

IV. PRIVACY PRESERVATION TECHNIQUES

Privacy-Preserving (PP) is the emerging topic in the today research world. The basic idea of Privacy-preserving is to modify the data in such a way so as to perform cryptographic algorithms effectively without compromising the security of sensitive information contained in the data [14]. Current studies of PP mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. In this research paper discusses with various Privacy preservation techniques which help to preserve the personalized data in big data. The privacy-preserving techniques are discussed as follows,

4.1Encryption based Technique

A framework in big data for privacy-preserving was developed where participants are categorized as in the figure.

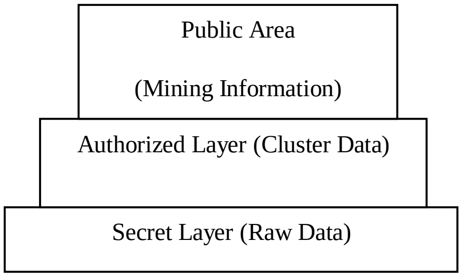

Figure 1. Levels of Access Model

In this encryption based technique mainly done in the secret layer. The raw data is stored on the secret layer. So the customer details are encrypted and then it was saved in the database. For the encryption, an asymmetric based cryptographic algorithm is the acceptable one to encrypt the big data. The decryption is performed in the authorized layer, in the decryption it is the reverse process of encryption. First by the employee with the private key and also identifies the sender. After that, the rule system clustering is done, and then the mining information in the public layer is taken after mining is viewed. This layer is accessible for all the employees within the company. The process details of this encryption based method are discussed as the following steps,
Step 1: Applying for digital ID by the database administrator
Step 2: Customer information is encrypted before saving it to the database.
Step 3: Administrator applies for the digital signature so that no one will alter the information.
Step 4: Decryption is done by the only authorized user and verifies the identity.

4.2Anonymization Based Technique

De-identification is known as anonymization is useful for the protection of data that are stored in databases and cloud. Intel uses internal web portal known as Circuit which contains employee details. For example, a link can be used for a site on the portal which is frequently used. They are also prone to unsorted matching attack and complementary release attack. Anonymization Architecture is shown in the following figure.

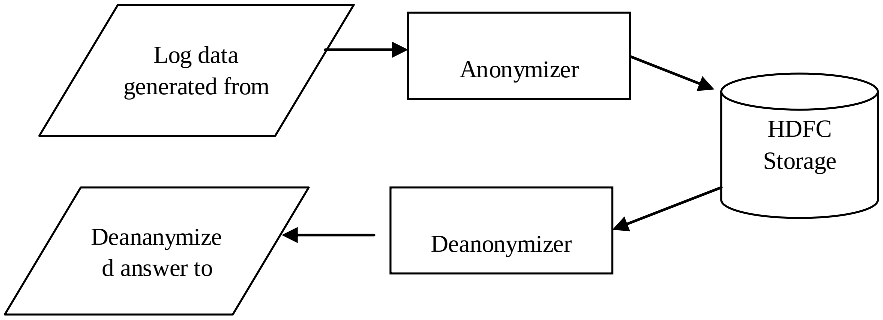

Figure 2. Anonymization Architecture

The Personally Identifiable Information (PII) must be removed from the log data so that data are not disclosed. The above figure shows that data which are sensitive anonymized such as IP address and user IDs in the log. Only fully anonymized data can be moved to Hadoop File System (HDFS) storage. As with anonymization [15] when the data are obscured or removed it doesn't mean that data are not revealed or there is no disclosure of an individual. With some visible log entries or with some other information from the database the attacker can reveal the individual information. So the following changes are made to improve the anonymization,

1. Hide all the references to nodes, sites, languages, and other data of an individual (no local info).
2. Hide all user browser information.
3. The log entries with a certain risk level of disclosure of data are removed.

4.3 Noise-based Technique

In the noise based approach, first of all, the noise is added to the data [13]. The flaw that has been found in encryption and anonymization has overcome by differential privacy. With the use of differential privacy technique in big data, the system speed is not affected and it has overcome the k-anonymity problems [2]. In this method the attacker will not be able to extract information of an individual in their presence and absence, if the attacker knows all the information in the database still the information of an individual cannot be extracted from the database. When the outputs of two different databases are computed the result will be almost similar. If the result of two different databases are same.

The curator adds a proper amount of noise that is scaled for privacy as shown in fig 3, this will preserve the privacy of an individual. This noise is the maximum difference between the two neighboring databases. To maintain the privacy the maximum difference should be hidden. Mathematically it can show it as:

$$\Delta f = \max\|f(D_1) - f(D_2)\|$$

V. COMPARATIVE ANALYSIS

In this paper, the privacy preservation techniques of Big data such as encryption based, anonymization based and noised based have been compared. The data security and utility are the most important parameters, so the utility of noised based is efficient than the other two techniques. Similarly, the security and privacy guarantee when compared encryption gives security but it reduces the utility and anonymization does not guarantee full privacy but differential privacy gives both privacy as well as maintains utility. As big data is of large volume data size should also be considered while implementing privacy, in this due encryption keys the data size increases gradually but in anonymization, the size of data is decreased by generalization and suppression this affects the completeness of data. But in the differential privacy techniques, the size of data is maintained as well the privacy of data. Hence, by analyzing the privacy preservation techniques noise based is more efficient than encryption based and anonymization based. The below table shows the comparative analysis between the techniques with security-based parameters.

VI. RESULT & DISCUSSION

Table 1. Comparison of Privacy Preservation Techniques

S.No.	Parameters	Encryption Based Technique	Anonymization based Technique	Noise Based Technique
1	Security	Data is Secure	Data Confidentiality is maintained but the available data is still utilized	Differential privacy promises to be free from the flaws of encryption
2	Data Size	Size of data increased	size of data remains same	Noise is added to the data set in differential privacy
3	Privacy Guarantee	Prone to attacks based on the algorithm selection	Does not guarantee full privacy	Provides Privacy Guarantee
4	Big Data Compatibility	Low	Low-medium	Suitable for data
5	Data Loss	Low-medium	High	Low
6	Privacy VS Utility	Offers higher privacy over utility	Offers better privacy over utility	Provides good balance between privacy and utility

VII. CONCLUSION

Protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. The most critical issues in big data are security and privacy. The analysis and study of various approaches conclude that differential privacy is more efficient and overcome most of the issues and challenges that occur in the privacy preservation of big data. This type of privacy approach is compatible with big data characters and also guarantees the privacy of data as well as individuals. This approach does not compromise with the utility of data and provides privacy inefficient way. So this paper concludes that noise based approach is efficient than encryption based and anonymization based individuals to preserving the personalized data in big data.

REFERENCES

1. *AnjanaGosain and Nikita Chugh, "Privacy Preservation in Big Data", International Journal of Computer Applications, Volume No.17, August 2014.*
2. *Dwork and Cynthia, "Differential privacy: A survey of results" In Theory and Applications of Models of Computation, 2008, pp. 1-19.*
3. *Jeff Sedayao and Rahul Bhardwaj, "Making Big Data, Privacy, and Anonymization work together in the Enterprise:Experiences and Issues", IEEE International Congress on Big Data 2014.*
4. *Katal.A et al., "Big data: issues, challenges, tools and good practices", Sixth International Conference on Contemporary Computing (IC3), IEEE, 2013, pp.404–409.*
5. *N. Li et al., "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity**",** IEEE 23rd International Conference on Data Engineering, 2007, pp. 106 - 115.*
6. *M. B. Malik et al., "Privacy preserving data mining techniques: Current scenario and future prospects", In Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 26-32.*

7. MatturdiBardi et al., "Big Data security and privacy: A review", China Communications Supplement, No.2, 2014.
8. S. Matwin, "Privacy-preserving data mining techniques: Survey and challenges in Discrimination and Privacy in the Information Society", Berlin, Germany: Springer-Verlag, 2013, pp. 209-221.
9. Mohammad Tarique Mohammad Saleem and A.P.Kankale, "Privacy Preserving and Data Mining In Big Data", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume (3) Issue 10, October 2016, pp 197-201.
10. NasrinIrshadHussain and Bharadwaj Choudhury, "A Novel Method for Preserving Privacy in Big-Data Mining", International Journal of Computer Applications, Volume No 16, October 2014.
11. Dr.A.Padmapriya and P.Subhasri, "Cloud Computing: Security Challenges & Encryption Practices", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013, pp. 255-259.
12. P.Ram Mohan Rao et al., "Privacy preservation techniques in big data analytics: a survey", J Big Data, https://doi.org/10.1186/s40537-018-0141-8, Volume 5:33, 2018.
13. Raymond Chi-Wing Wong et al., "(a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing", In Proceedings of the 12th ACM SIGKDD, 2006, pp. 754-759.
14. Salini.S et al., "Survey on Data Privacy in Big Data with K- Anonymity." International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 5, May 2015.
15. L.Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Volume 10 (5), 2002.