Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Survey on Privacy Preservation Techniques in Big Data Processing: A Review

Nurjahan V A Research Scholar
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Chennai, India
noorji84@gmail.com
Dr. S. Jancy
Assistant Professor
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
Chennai, India
jancy.cse@sathyabama.ac.in

Article Info
Page Number: 59-69
Publication Issue:
Vol 71 No. 3s2 (2022)

Abstract
Big data is a collection of large volume of heterogeneous data. Due to the rapid growth of online social network users, large amount of data is generated every day. Big data processing becomes crucial because of the fast growth of data. Big data includes personal information such as personal identification, salary details, health records etc. As the volume of data increases privacy and security violations may also increase. Privacy refers to the protection of individual's data. Researchers have developed various privacy preservation techniques. One of the most effective methods for big data privacy is anonymization technique. In this paper we are focusing on different privacy preserving methods such as anonymization, randomization and differential privacy. It also reviewed some merits and demerits of different anonymization techniques such as k-anonymization, l-diversity and t-closeness etc.

Article History
Article Received: 28 April 2022
Revised: 15 May 2022
Accepted: 20 June 2022
Publication: 21 July 2022

Keywords: - Big Data, Privacy, Anonymization Techniques, K- anonymity, L-diversity, T- closeness, Randomization, Differential Privacy, Mondrian, MRA

I. INTRODUCTION

Big data refers to the huge amount of heterogeneous data. The data may be structured, unstructured or semi structured formats. Big data processing is a set of techniques for extracting meaningful information from enormous amounts of heterogeneous data in order to make better decisions [1]. Big data analytics is the method used to investigate a large volume of heterogeneous data [3]. The major challenges of big data processing include capturing, analyzing, processing, storing, sharing, visualization [3] etc. Because of its large size and complexity, none of the data management tools can process it effectively. Thus, distributed programming frameworks like Hadoop, Map Reduce, Spark, etc. are used for big data processing.

Bigdata is measured by the following characteristics [2].

- Volume: Represent a large amount of data from various business organizations, various institutions, individuals, etc., which are now frequently larger than terabytes and petabytes. As the volume of data grows, the possibility of information leakage grows, potentially compromising an individual's privacy.
- Variety: The data may be structured, unstructured or semi structured formats.
- Velocity: Represents how fast new data being generated. It determines speed of data generated and meets the demands [4].
- Veracity: Data veracity refers to how accurate or true a data set is. It reflects the quality of the data analysed.
- Validity: denotes the correctness and accuracy of data in order to make right conclusions. [4]
- Visualization: This is the act of displaying data in visual forms that clearly express a concept while also simplifying complex data and making it accessible to a large number of people.
- Value: This refers to the value that bigdata can provide.
- Viscosity: Indicates how challenging the data is to use or integrate. It calculates the amount of resistance to data flow in a given volume.
- Virality: It's a metric for how quickly data is disseminated and shared among individual nodes.

The remaining portions of this paper are organized as follows. Section II describes data privacy and different privacy-preserving techniques. This section mainly concentrates on anonymization, differential privacy, and randomization. Section III concludes this study.

II. DATA PRIVACY IN BIG DATA PROCESSING

Data Privacy concerns the proper handling of sensitive data such as personal information, financial information, medical related data etc. In practical sense, data privacy deals with processing and sharing of data with third parties, how and where data is stored etc. In most cases, data is anonymized before being handled in a distributed framework. As a result, with big data analytics, privacy is a fundamental concern. [17].

A. Privacy Preservation in Bigdata Processing

Privacy represents the protection of individual’s data. Traditional privacy methods cannot handle the data protection properly in big data since it comprised of large and complex data set [5].

Several approaches assure privacy preservation in big data processing which include anatomization, anonymization, and permutation. In Anatomization sensitive attributes are grouped together for eliminating attribute disclosure. Anonymization concentrate on quasi-identifiers for preventing identity disclosure. Permutation is the process of creating different groups based on quasi-identifiers and then shuffling the values of sensitive attributes to each group [18].

i. Anonymization

Because bigdata may contain personally sensitive information, it is crucial to keep it safe from unwanted access [6]. One of the most frequent strategies for concealing personal information is data anonymization. It is the process of removing personal identifiable information from a dataset. In comparison to randomization, perturbation, and other methods for preserving privacy, data anonymization is the most effective.

A dataset comprised of four kinds of attributes.

- Personal Identification: Attributes that are used for identifying individuals and have unique values.
- Sensitive attributes: Attributes that should be hidden from others during the process of publishing and sharing data eg: - salary
- Quasi- identifiers (QI): The attributes like gender, date of birth, zip code can be joined with external data to reidentify individuals are known as quasi-identifiers.
- Non sensitive attributes: The remaining attributes are called non sensitive attributes [7].

B. Different Anonymization Techniques

Anonymization techniques for privacy preservation in bigdata mainly classified into three

1. k- anonymity
2. l- diversity
3. t- closeness

Consider a sample health dataset (Table 1), here Name is the personal identification attribute, Age, Zip code and Sex are the quasi-identifiers and disease is the sensitive attributes. Anonymization refers to the process of removing personal identification information. In this table Name is the personal identification attribute.

TABLE 1: SAMPLE HEALTH DATSET

Name	**Zip code**	**Age**	**Sex**	**Disease**
Alice	22324	29	M	Heart Disease
Peter	22332	22	M	Infection
Joy	22348	27	M	Cancer
George	22856	43	M	Cancer
Anu	22837	32	F	Heart Disease
John	22884	47	M	Stomach Problem

If we remove Personal Identification attribute such as Name from the dataset, it will not provide complete privacy to data. To provide privacy to the dataset we also have to anonymize the quasi-identifiers [9]. For anonymizing the quasi-identifiers, use k-anonymity and l-diversity anonymization techniques.

1. k-anonymity

If one record in the data set has a value for QID, then at least k-1 other records in the data set have the same value for QID [8]. In other words, at least k entries in the data set must have the same QID value, and the resulting table is called k- anonymous [10].

TABLE 2: AFTER K-ANONYMITY ON TABLE 1

Zip code	**Age**	**Sex**	**Disease**
223**	2*	M	Heart Disease
223**	2*	M	Infection
223**	2*	M	Cancer
228**	3*	F	Heart Disease
228**	4*	M	Cancer
228**	4*	M	Stomach Problem

When we apply k-anonymity to Table 1 (for example, the value of k is 2), we get Table 2. Because at least k-1 records have the identical QID values, so it is difficult to an outsider to uncover sensitive information. The first three records in this table make up one equivalence class, while the last two records make up another.

Two methods are used for implementing k-anonymity are

- Generalization
- Suppression

Generalization

In generalization method the values of the quasi-identifiers are substituted with a general value [13]. The attribute Age, for example, can be represented in a generic way. If the Age attribute has a value of 32, it is indicated as Age <30.

Suppression

Suppression is a technique used to hiding the values of the quasi-identifiers. The suppressed value can be represented using asterisk (*). For example, some values of the Zip code attribute are hidden by using * symbol [15].

Apply generalization and suppression on Table 1 we get Table 3.

TABLE 3: EXAMPLE FOR GENERALIZATION AND SUPRESSION IN-K-ANONYMITY

Zip code	**Age**	**Sex**	**Disease**
223**	<30	M	Heart Disease
223**	<30	M	Infection
223**	<30	M	Cancer
228**	<40	F	Heart Disease
228**	<50	M	Cancer
228**	<50	M	Stomach Problem

If an attacker has some background knowledge about the person or if any equivalence class has the same sensitive information, the attacker can easily obtain the sensitive information. Use another anonymization approach, l- diversity, to solve these issues.

2. l-diversity

In l-diversity each equivalence class has at least l- “well represented” sensitive values [9]. A dataset is said to have l- diversity if it satisfies the following properties

- If each table equivalence class has l- diversity
- If equivalence class contains at least l “well represented” value [16].

TABLE 4 : EXAMPLE FOR L-DIVERSITY

Zip code	**Age**	**Sex**	**Disease**
223**	2*	M	Heart Disease
223**	2*	M	Infection
223**	3*	M	Cancer
228**	2*	F	Flue

It doesn’t consider the semantic meaning of sensitive attributes. If two diseases have distinct name but it is semantically same, so attacker may gain some sensitive information [10].

3. t-closeness

It is an extension of l-diversity. The equivalence class is considered to have t-closeness if the distance between the distribution of sensitive attributes in the equivalence class and the distribution of attributes in the entire table is less than a threshold t. The Earth Mover's Distance (EMD) method is used to compute the distance. With respect to sensitive attributes, T-closeness is calculated for each attribute.

C. Different approaches for privacy preservation based on k-anonymity and l-diversity

K-anonymity, l-diversity, and t-closeness are three basic techniques commonly used for privacy preservation in the big data processing. As the data volume increases these methods are not efficient for handling data privacy. So many researchers developed different approaches based on these three techniques for anonymization. Some of them are described below with a comparison study.

i. Mondrian Multidimensional K-Anonymity

The k-anonymity technique is used in this method. The input data set is first handled as a single equivalence class, and then partitioned into the desired number of equivalence classes based on the k-anonymity condition. The splitting procedure continues until there is no class that meets the k-anonymity criteria [21].

ii. Map Reduce based Anonymization (MRA)

MRA algorithm is also used the concept of Mondrian Multidimensional K-anonymity [21]. Mondrian algorithm cannot run of multiple machines in parallel. But MRA algorithm can be implemented on multiple machines in parallel.

iii. Scalable k anonymization approach using MapReduce (SKA)

The main concern of this approach is scalability issues. Scalability is described as the ability of the system to manage the increasing amount of data without degrading its performance.

Based on all the attributes of data set, SKA divides the input data set into number of equivalence classes. These classes are grouped to make it large enough to satisfy the k-anonymity conditions. This procedure is repeated until all of the classes are processed [8].

iv. Improved l-diversity: Scalable anonymization approach

SKA approaches use the concept of k-anonymity. So, it suffers with record linkage attack. To eliminate this problem, use Improved l-diversity approach. It is an extension of SKA approach. In this approach two techniques are used for improving the running time and decreasing information loss. They are Improved Scalable k- Anonymization (ImSKA) and Improved Scalable l- diversity (ImSLD) [22].

Various researchers have written numerous reviews of security algorithms in a variety of sectors in the literature. [20][21][22][23]]. This study will surely introduce researchers to the idea of employing security measures in a variety of applications. [24][25][26]. In several cases, intrusion detection solutions were also considered. [28][29].

ii. Differential Privacy

It is a privacy preserving technology that provide researchers to access the information from dataset without revealing individual's personal identities. This can be achieved by adding minimum distraction in the data.
In differential privacy mechanism [19] there is no direct communication between database and analyst. Here an intermediary software is introduced between database and analyst for protecting privacy. The software is known as privacy guard.

Steps for implementing differential Privacy Mechanism
- Analyst make a query to the database through privacy guard
- Privacy guard executes the query and earlier queries for privacy risk
- The guard collects answer from database
- Based on privacy risk, add some distortion to it.

If privacy risk is low, small amount of distortion is added. If it is high more distortion is needed.

Differential privacy criticizes the limitations of k-anonymity as it provides poor attribute disclosure [14].

iii. Randomization

During data collection and preprocessing phase randomization techniques can be applied. It is the process of adding noise the data [20]. It can be applied during surveys, sentiment analysis etc. If the dataset is large randomization is not an effective method. According to the increasing in data volume more Mappers and Reducers were used [20].

The following table (Table 4) describes different anonymization techniques and their merits and demerits.

TABLE 4: COMPARISON STUDY OF MERITS AND DEMERITS OF DIFFERENT ANONYMIZATION

Anonymization Techniques	**Merits**	**Demerits**
k- anonymity	- Protection against identity disclosure - Prevents from combining sensitive data with external data	- Homogeneity attack - Background Knowledge attack

	- Cost is less compared to other methods [12]	
l-diversity	- Handle homogeneity attack and background knowledge attack - Performance of l-diversity is better compared to k-anonymity	- Doesn't handle semantic attacks - It is difficult to achieve [15] - Insufficient to prevent attribute disclosure
t-closeness	- Identifies semantic closeness of attributes - Protects against attribute disclosure	- Using EMD, hard to identify the closeness between attributes [12]
Mondrian Multidimensional K-Anonymity	- Less Information loss	- This approach cannot run on multiple machines in parallel. - Cannot handle large data set [8]
Map Reduce based Anonymization (MRA)	- Single Map Reduce iteration needed - Less runtime compared to Mondrian Approach	- Information loss high compared to Mondrian and SKA - Enhance Performance in running time compared to Mondrian - Can be execute on multiple machines parallelly [8]
Scalable k anonymization approach using MapReduce (SKA)	- Less information loss compared to MRA - Enhance performance in running time compared to Mondrian and MRA Algorithm [8].	- Only based on k-anonymity so it suffers from record linkage attack.
Improved l-diversity: Scalable anonymization approach	- Information loss is less compared to Mondrian, MRA and SKA approaches - Running time performance increased compared to Mondrian, MRA and SKA approaches [22].	- It doesn't handle semantic attacks

III. CONCLUSION

In this paper made a review on different privacy preservation techniques such as anonymization, differential privacy and randomization techniques. A comparison study is performed on different anonymization techniques including k-anonymity, l-diversity and t-closeness etc. Combination of these three techniques make a balance between ensuring privacy in big data processing. An efficient method is required for privacy preservation in big data processing because of its high data volume and variety of data set.

REFERENCES

[1] N. Gruschka, V. Mavroeidis, K. Vishi, M. Jensen, "Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR," IEEE Proceeding of the International Conference on Big Data, USA, pp. 5027-5033, 2018.

[2] A Classification of non-Cryptographic Anonymization Techniques ensuring Privacy in Big Data, Zakariae el Ouazzani, Hanan El Bakkali, International Journal of Communication Networks and Information Security · April 2020

[3] A brief Introduction on Big data characteristics and Hadoop Technology, Ishwarappa, Anuradha J, Elsevier, Science, Volume, 2015

[4] Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 3, 2016

[5] Privacy-preserving big data analytics a comprehensive survey, Journal of Parallel and Distributed Computing Volume 134, December 2019, Pages 207-218

[6] Data security and privacy: A review Bardi Matturdi; Xianwei Zhou; Shuai Li; Fuhong, Lin. IEEE, January 2014 China Communications 11(14):135-145

[7] Information Security in Big Data: Privacy and Data Mining, ISSN Information: Electronic ISSN: 2169-3536 DOI: 10.1109/ACCESS.2014.2362522 IEEE, Lei Xu; Chunxiao Jiang; Jian Wang; Jian Yuan; Yong Ren

[8] Privacy Preserving Big Data Publishing: A Scalable k-Anonymization Approach using MapReduce July 2017 IET Software 11(5)DOI:10.1049/iet-sen.2016.0264, Brijesh B. Mehta, Udai Pratap Rao

[9] Preserving the privacy of sensitive data using data anonymization, January 2017,International Journal of Applied Engineering Research 12(8):1639-1663 Jyothi Nelahonne, Mohan K L University, M.V.P. Chandra Sekhara Rao

[10] L-Diversity: Privacy Beyond K-Anonymity, Ashwin Machanavajjhala,Daniel Kifer, Johannes Gehrke, Muthuramakrishnan, Venkitasubramaniam, ACM Transactions On Knowledge Discovery From Datavolume 1Issue 1March 2007

[11] T-Closeness: Privacy Beyond K-Anonymity And Diversity Ninghui Li Tiancheng Li Department Of Computer Science, Purdue University {Ninghui, Li83}@Cs.Purdue.Edu Suresh Venkatasubramanian AT&T Labs – Research Suresh@Research.Att.Com

[12] A Study On K-Anonymity, L-Diversity, And T-Close Ess Techniques Focusing Medical Data, December 2017 Keerthana Rajendran,Manoj Jayabalan,Liverpool John Moores University, Muhammad Ehsan Rana

[13] K-Anonymity In Practice: How Generalisation And Suppression Affect Machine Learning Classifiers 2021 The Author(S). Published By Elsevier Ltd, Djordje Slijep Cevic, Maximilianhenzl, Lukasdanielklausner, Tobias, Dam, Peterkieseberg, Matthias Zeppelzauer

[14] Differential Privacy: Its Technological Prescriptive Using Big Data, Priyanka Jain, Manasi Gyanchandani & Nilay Khare, Jounal Of Big Data (2018)

[15] Big Data Privacy Methods Jayesh Surana1, Avani Kothari, Meenal Sankhla, Himanshi Solanki, Akshay Khandelwal, Vol-3 Issue-2 2017 Ijariie-Issn(O)-2395-4396

[16] A Review On Anonymization Techniques For Privacy Protection In Data Mining U. Saranya, A. Logeswari, U. Sujatha, International Journal Of Engineering Applied Sciences And Technology, 2020 Vol. 4, Issue 10, ISSN No. 2455-2143, Pages 310-316 Published Online February 2020 In IJEAST

[17] A Survey On Privacy Preservation In Data Publishing. VS Susan, T Christopher. International Journal Of Computer Science And Mobile Computing 3 (3), 188-193

[18] V. S. Susan And T. Christopher, "Anatomisation With Slicing: A New Privacy Preservation Approach For Multiple Sensitive Attributes,"Springerplus, Vol. 5, No. 1, Pp. 1-21, 2016.

[19] Jain P, Pathak N, Tapashetti P, Umesh AS. Privacy Preserving Processing Of Data Decision Tree Based On Sample Selection And Singular Value Decomposition. In: 39th International conference.

[20]RM Gomathi, JML Manickam, A Sivasangari, P Ajitha,"Energy efficient dynamic clustering routing protocol in underwater wireless sensor networks", International Journal of Networking and Virtual Organisations, Vol.22,4 pp. 415-432

[21]Kanyadara Saakshara, Kandula Pranathi, R.M. Gomathi, A. Sivasangari, P. Ajitha, T. Anandhi, "Speaker Recognition System using Gaussian Mixture Model", 2020 ,International Conference on Communication and Signal Processing (ICCSP), pp.1041-1044, July 28 - 30, 2020.

[22] Sivasangari, A., Ajitha, P., Brumancia, E., Sujihelen, L., Rajesh, G.(2021),” Data security and privacy functions in fog computing for healthcare 4.0”,Signals and Communication Technology, 2021, pp. 337–354

[23]A Sivasangari, P Ajitha, RM Gomathi, "Light weight security scheme in wireless body area sensor network using logistic chaotic scheme", International Journal of Networking and Virtual Organisations, 22(4), PP.433-444, 2020

[24] Sivasangari, A., Nivetha, S., Pavithra,, Ajitha, P., Gomathi, R.M. (2020),” Indian Traffic Sign Board Recognition and Driver Alert System Using CNN”, 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315260

[25] Gowri, S. and Divya, G., 2015, February. Automation of garden tools monitored using mobile application. In International Confernce on Innovation Information in Computing Technologies (pp. 1-6). IEEE.

[26] Gowri, S., and J. Jabez. "Novel Methodology of Data Management in Ad Hoc Network Formulated Using Nanosensors for Detection of Industrial Pollutants." In International Conference on Computational Intelligence, Communications, and Business Analytics, pp. 206-216. Springer, Singapore, 2017.

[27]Sivasangari A, Bhowal S, Subhashini R "Secure encryption in wireless body sensor networks",Advances in Intelligent Systems and Computing, 2019, 814, pp. 679–686

[28] Subhashini R, Niveditha P R, "Analyzing and detecting employee's emotion for amelioration of organizations", Procedia Computer Science, 2015, 48(C), pp. 530–536

[29] Ajitha, P.Sivasangari, A.Gomathi, R.M.Indira, K."Prediction of customer plan using churn analysis for telecom industry",Recent Advances in Computer Science and Communications,Volume 13, Issue 5, 2020, Pages 926-929.