IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

# Scalable Privacy Preservation in Big Data

**Ankita Hatiskar**
Department of Information Technology
Sir Sitaram and Lady Shantabai Patkar College of Arts and Science, Mumbai, Maharashtra, India

**Abstract:** *In the realm of data analytics, big data has welcomed a revolution. Data that was abandoned a few years ago is now seen as a valuable resource. Big data is increasingly widely employed across all aspects of society for knowledge extraction. It is generated by practically all digitization, and it is saved and transmitted over the internet. This dependence on the web approach for massive data raises severe security problems. Due to its vast volume, diversity, and volume, traditional security measures cannot be used to big data. Privacy is also a huge security nightmare since a large dataset appears to contain confidential details. Traditional privacy preservation approaches are utilized to solve privacy difficulties in Big Data, and K-anonymity is the most often used strategy for protecting privacy for data disclosure.*

**Keywords:** Big data, Data privacy, Anonymization, K-Anonymization, Privacy, Scalable.

## I. INTRODUCTION

CLOUD computing and Big Data are two disruptive innovations that are now affecting the industrial and research sector. For data mining, processing, and sharing, a great number of big data services are being installed or transferred to the cloud. Due to the perceived key qualities of cloud computing, such as extreme scalability and pay-as-you-go pricing, Big Data is inextricably associated with public cloud infrastructure. Personal private information data, such as digital medical database systems and financial transaction history, is frequently included in data sets used in Big Data applications. The data sets are frequently shared or published to third-party partners or the general public since the study of these data sets gives significant insights into a variety of crucial sectors of society. As a result, significant data security protection is required. In the non-interactive access control and distributing process, data anonymization plays a critical role in maintaining privacy. Data anonymization [1,4] refers to hiding the identity of sensitive data so the privacy of an individual is preserved even if certain aggregate information can be still exposed to data users for diverse analysis and mining tasks. Several privacy models and data anonymization approaches have been proposed and extensively reviewed [8,12]. However, due to the total 3Vs i.e., Volume, Velocity, and Variety, applying these classic methodologies to large data anonymization offers scalability and performance issues. The study of scalability challenges in large data anonymization has come to light, although it is lacking in certain key areas.

The format of this document is as follows: in Section II we portray the Basic Concept of Anonymization. Traditional Data Privacy Preservation Approaches are discussed in Section III. Section IV explicates Privacy Preservation in Big Data. Finally, in Section V, we bring this paper to a conclusion.

## II. PRIVACY ISSUES IN BIG DATA

Data sharing is a decision made by the individual based on the protection of privacy. Due to the evolution of internet applications in many fields, a large volume of data is produced on the internet. A public database puts the privacy of individuals at risk. Many businesses analyse their customers' buying habits and suggest various products with multiple offers [5]. The recommendations are generated by tracking the consumer's activities, which poses a personal privacy risk. Assume that a data owner provides anonymized data with a third party. In that case, the third party can link the data with external sources' data to identify sensitive information [3]. Data exposure is a serious invasion of privacy. Data exposure might potentially lead to prejudice. When a person's medical data is released, for instance, it might result in personal disgrace and abuse. Although data analytics is important for decision-making, it must be protected from privacy concerns such as homogeneity assaults, background attacks, and data breaches. As a result, privacy-preserving approaches in data analytics are critical.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

## III. Basic Concepts of Anonymization

### 3.1 Big Data

Big data is a vast combination of associated data that organisations utilise to summaries, evaluate records, and modify their business tactics, depending on their business model. Big data, according to Francis Diebold, is the explosion in the quantity and quality of data that is readily available and possibly useful.

### 3.2 Data Privacy

With the increase in the number of active netizens, it is more important than ever before for individuals and organisations to enhance privacy awareness in place to evade bogus assaults. The term "data privacy" refers to the measures that ensure that data should only be used for the purposes for which it was collected. From granting a mobile application inappropriate access to large-scale network assaults, a platform that improves security by accurately recognizing harmful code is required.

### 3.3 Identifiers

Whatever confidential material about an individual is frequently referred to as an identifier. To defend against data theft, this metadata needs encryption. Credit card numbers, ID numbers, bank account numbers, names, and other identifiers are samples of identifiers.

### 3.4 Data Anonymization

In plain terminology, data anonymization implies separating sensitive data from other data kept alongside it. It's a necessary step to ensure that a single identification breach doesn't expose the system server. Data masking, generalisation, data swapping, data perturbation, and synthetic data are all common data anonymization approaches.

When a user agrees to allow cookies on a website, the cookies obtain data about the person's activities. The approaches used to anonymize data eliminate the databases identified, making it difficult to draw inferences and improving the user experience. As a result, we can't utilize anonymized data for advertisements or research.

### 3.5 Sensitive Identifiers

Some identifiers necessitate the most extensive use of flexible resources. Cell phone numbers, medical history, and other personal information are commonly included.

### 3.6 Quasi-Identifiers

Non-sensitive metadata connected to every other access repository to further analyse a certain person's data are referred to as quasi-identifiers. For instance, within the case of account credentials, the IFSC code can perhaps be connected to the bank branch, however, personal data such as debit/credit card information is not.

### 3.7 Generalization

The method for replacing a specific value with a generalised value is known as generalisation. Age and salary are generalised into intervals (for instance, [20-30]), and attribute values are generalised into a collection of unique values. Figure 2 shows how the gender and age attributes are generalised.

**Table 1:** Data Sample

| ID | Gender | Age | ZIP |
|---|---|---|---|
| 1 | Male | 38 | 400101 |
| 2 | Female | 24 | 400109 |
| 3 | Male | 32 | 400163 |
| 4 | Male | 29 | 400104 |
| 5 | Female | 26 | 400105 |
| 6 | Female | 22 | 400102 |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

**Table 2:** K-Anonymity table (k=3)

| ID | Gender | Age | ZIP |
|---|---|---|---|
| 1 | Person | [36,40] | 4001** |
| 2 | Person | [21,25] | 4001** |
| 3 | Person | [31,35] | 4001** |
| 4 | Person | [26,30] | 4001** |
| 5 | Person | [26,30] | 4001** |
| 6 | Person | [21,25] | 4001** |

### 3.8 Distance and Information Loss Metric

When it comes to getting the best clustering results, distance matters a lot. As a result, the distance measure should be carefully chosen. Data leakage is another important consideration when evaluating k-anonymization. Information loss is a type of mistake that happens during the transfer of large amounts of data. To assess information loss, we use an algorithm to measure distance among data as well as between records and clusters. Categorical characteristics do not have the complete classification, hence the difference between quantitative attributes cannot be used.

### 3.9 Equivalent Classes

Each equivalent class must have at least k records. The combination of all the comparable valuable quasi-identifiers is one equivalent class. After k-anonymity, source data yields m equivalent classes, where m = (n/k), where n represents the number of rows in the data table.

### 3.10 Suppression

Quasi-identifiers are darkened by certain fixed binary digits like 0, *, and so on during suppression, making them unlink able to any external sources. The data sample is shown in Figure 1 before it is generalised and suppressed. In Fig. 2, the ZIP attribute from Fig. 1 is disabled.

## IV. Traditional Data Privacy Preservation Approaches

Cryptography is a collection of data-protection methods and algorithms. Various encryption algorithms are used to transform plaintext into ciphertext in cryptography. Various approaches relying on this approach exist, including public-key cryptography, digital signatures, and so on. Cryptography alone can't enforce the privacy demanded by common cloud computing and big data services [6]. This is because big data differs from traditional large data sets based on three V's (velocity, variety, volume) [9,13]. Big data architecture differs from typical information architectures because of these characteristics. Cryptography and typical encryption approaches are not scalable enough to meet the privacy concerns of large amounts of data response to the changes in architecture as well as its sheer unpredictability.

Less confidential material that can be beneficial in big data analytics is encrypted as well, and users are not permitted to view it. It prevents data from being accessed by individuals who may not have the decryption key. Also, if the information is collected prior to encryption or cryptographic keys are exploited, privacy may be jeopardized. Attribute-based encryption can also be used for big data privacy [10, 11]. This approach of protecting big data is dependent on the relationships between the attributes in the data. The properties that must be secured are determined by the nature of big data as well as the policies of the firm. In essence, encryption or cryptography alone is insufficient to protect large data privacy. They can assist us in data anonymization, but they cannot be applied for large data privacy directly.

## V. Privacy-Preserving in Big Data

### 5.1 K-Anonymity

Certain anonymised data have the property of K-anonymity. Provide a lease of the database having scientific assurances that the persons who will be the objects of the data cannot be reidentified while the data remains practically usable, given person particular field structured data. If the information for every person provided in the release cannot be discriminated against at least k-1 persons, it may well have the k-anonymity property.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

**A. K-Anonymization Approach**

Suppression: In suppression, an asterisk mark * is used to substitute particular variables' values. Table 1 shows how all or some values in a column can be replaced by*.

Generalization: Key values of characteristics are replaced by a foreigner category.

**Table 3:** Anonymized data

| Age | Sex | City | Income |
|---|---|---|---|
| 2* | M | Mumbai | 1,00,000 |
| 2* | M | Pune | 18,000 |
| 2* | M | Pune | 25,500 |
| 2* | F | Mumbai | 20,000 |
| 2* | F | Mumbai | 50,000 |
| 2* | F | Mumbai | 29,000 |
| 3* | M | Mumbai | 26,000 |
| 3* | F | Pune | 45,000 |

### 5.2 L-Diversity

L-diversity is a method of anonymization that depends on a group. The term "variety" is used to describe what is utilised in a preserve. The l-diversity model is an evolution of the k-anonymity model that decreases the level of detail of data representation by employing methods such as generalisation and suppression to ensure that each provided record maps onto it at least k other records in the dataset. Additionally, the L-Diversity approach helps handle some of the deficiencies in the K-Anonymity approach in which protected identities at the level of k-individuals do not equate to protecting sensitive values that are generalized or suppressed, especially when the sensitive values are found within a group.

### 5.3 Generalization Approach

By using Map Reduce in the cloud to do Bottom-Up Generalization (BUG) enabling data anonymization and designing a set of new Map-Reduce tasks to concretely achieve the generalization in a highly scalable manner. Moreover, offer a scalable Advanced BUG technique that conducts generalisation on various partitioned data sets and then combines the intermediate anonymizations to discover the final anonymization that is utilised to anonymize the original data set. The results imply that our technique outperforms previous approaches in terms of scalability and efficiency for BUG data anonymization.

### 5.4 Map Reduce: A Large-Scale Data Processing Framework.

A broadly applied parallel data processing framework such as MapReduce was employed to overcome the scalability challenge of the Top-Down Specialization (TDS) technique for big-scale data sets. Over the first half, the original records are lower capacity datasets, which are then anonymized in parallel, providing in-term results. The intermediate findings are combined and anonymized in the second part to provide a consistent k-anonymous dataset.

MapReduce is used to split up massive input data into chunks of more or acceptable size, spinning up a series of processing occurrences for the map phase, allocating data to each mapper, attempting to track each mapper's condition, directing map achievement to their deliver phase, and eventually discontinuing the mappers with reducers when the work is completed. By manually executing the task on a larger cluster, you may easily scale up MapReduce Framework to handle larger workloads or produce results every time. When the MapReduce Framework is not employed, the distribution system breaks down.

### 5.5 T-Closeness

If somehow the gap between the distribution of a protective property in this classification and even the distribution of the attribute in the entire table is less than a threshold t, an equivalence class is said to have t-closeness. A table is said to have t-closeness if all equivalence classes have t-closeness. [7].

T-closeness' key benefit is that it avoids revealing information received. Data anonymization can be used for large data, but the issue is that as the number and diversity of data grow, so does the risk of re-identification. As a result, anonymization's promise in the realm of big data privacy is restricted.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

### 5.6 Top-Down Specialization

TDS is an iterative method that begins with the topmost domain values in the attribute taxonomy trees. Each round of iteration consists of three steps [3]. To reveal the greatest data usefulness, this procedure is repeated until k-anonymity is compromised. A search metric is used to assess the quality of a specialty.

### 5.7 Two-Phase Top-Down Specialization (TPTDS)

In TDS, a TPTDS methodology is a very efficient and scalable method. The two steps of our technique are centered on Map Reduce on the cloud's two layers of parallelization. On the cloud, Map Reduce offers two layers of parallelization. Job-level parallelization deals with several MapReduce tasks that can run in parallel to leverage the most of cloud infrastructure capabilities.

Multiple mapper/reducer jobs in a MapReduce project are products are formulated over data divides, which is referred to as task-level parallelization. In the first step, great scalability is achieved by parallelizing numerous processes on data partitions, but the ensuing anonymization levels are not equal. The second phase is required to combine the intermediate findings and then further anonymize full datasets in order to generate finally accurate anonymous data sets.

## VI. Conclusion

As big data is so closely linked to customers, it has now become a major concern. In the age of big data analytics, it is no longer optional for a company to guarantee privacy. Instead, then focusing on information gathering, privacy measures should instead focus on data consumption. They should be changed to account for the magnitude and unforeseen applications of big data. When it comes to huge data, techniques involving anonymization have little utility. The client is additionally burdened by the disclosure and permission procedure for maintaining privacy. Differential privacy might be considered a promising option for protecting large data privacy. One disadvantage of this strategy is that the analyst must first understand the question before applying the differential privacy model. When adapted and used to large data, it has the potential to protect privacy without altering the data itself.

## Acknowledgment

I am grateful to 'Sitaram and Lady Shantabai Patkar College of Arts and Science, Mumbai India' for providing an opportunity to write a research paper in the form of a dissertation on the subject. “Scalable Privacy Preservation in Big Data." I'd also like to thank my research professors for assisting me through the course of writing a research paper. Without their aid and continued role at every phase of the process, this task would not have been accomplished. I'd also like to thank my family and friends who helped me out with ideas and resources that supported me tremendously.

## References

**[1].** AntorweepChakravorty, Tomasz Wlodarczyk, ChunmingRong, ―Privacy Preserving Data Analytics for Smart Homes‖, IEEE Security and Privacy Workshops, pp. 1-5, 2013.

**[2].** Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys 42(4), 14 (2010)

**[3].** Jeff Sedayao, Rahul Bhardwaj and NakulGorade, ―Making Big Data, Privacy, and Anonymization work together in the Enterprise:Experiences and Issues‖, IEEE International Congress on Big Data, pp.1-7, 2014.

**[4].** Linna Li, Michael F. Goodchild and Santa Barbara, ― Is Privacy Still an Issue in the Era of Big Data ―Location disclosure in spatial footprints‖, Proceedings of 21st International conference on Geoinformatics, IEEE, pp.1-4, 2013.

**[5].** Liu Y et al. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans Ind Inf. 2018

**[6].** M. V. Dijk, A. Juels, "On the impossibility of cryptography alone for privacy-preserving cloud computing," Proceedings of the 5th USENIX conference on Hot topics in security, August 10, 2010, pp.1-8.

**[7].** N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, " IEEE 23rd International Conference on Data Engineering, 2007, pp. 106 - 115.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, April 2022
Impact Factor: 6.252

**[8].** Omar Hasan, Benjamin Habegger, Lionel Brunie, Nadia Bennani, Ernesto Damiani, ―A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case‖, IEEE International Congress on Big Data , pp. 1-6, 2013.

**[9].** S. Sagiroglu and D. Sinanc, "Big Data: A Review," Proc. International Conference on Collaboration Technologies and Systems, 2013, pp. 42- 47

**[10].** S.H. Kim, J. H. Eom, T. M. Chung, "Big Data Security Hardening Methodology Using Attributes Relationship," Proc. International Conference on Information Science and Applications (ICISA), 2013, pp. 1-2.

**[11].** S. H. Kim, N. U. Kim, T. M. Chung, "Attribute Relationship Evaluation Methodology for Big Data Security," Proc. International Conference on IT Convergence and Security (ICITCS), 2013, pp. 1-4.

**[12].** Wenyi Liu, A. SelcukUluagac, and RaheemBeyah, ―MACA: A Privacy-Preserving Multi-factor Cloud Authentication System Utilizing Big Data‖, IEEE INFOCOM Workshops, pp. 518- 523, 2014.

**[13].** Y. Demchenko, P. Grzsso, C. De Laat, P. Membrey, "Addressing Big Data Issues in Scientific Data Infrastructure," Proc. International Conference on Collaboration Technologies and Systems, 2013, pp. 48- 55.