Privacy Preserving Data Analytics: A Study of Processing Model

***[1]Shradha Soni, [2]Shraddha Masih**

[1]*Research Scholar, School of Computer Science & Information Technology*
Devi Ahilya Vishwavidyalaya, Indore (M.P.)
[2]*Assosiate Professor, School of Computer Science & Information Technology*
Devi Ahilya Vishwavidyalaya, Indore (M.P.)

***Corresponding Author: shradha.idwork@gmail.com**

Abstract-

Data Analytics is the buzzing field. Every day a vast amount of data is being generated by various ways. Companies or organizations need to analyze such data to make impactful decisions for their growth and customer satisfaction. At the same time, it is equally important to maintain the confidentiality of the data. This paper carries out the study of different privacy preservation algorithms for data analytics used by the researchers. Along with this, it also presents the experimental scenario and uncovers the future research directions.

1. Introduction

Usage of internet and its services is growing rapidly which is resulting in an explosive growth of the data. Social Media, E-Commerce, Banking, Healthcare, Government Agencies, Education and Entertainment industries are among the largest producers of the data. Mostly these organisations analyse raw data to make decisions through data analytics. They excavate hidden patterns, unknown correlations, market trends and customer preferences. The data collected for analysis may contain some private and sensitive information. For example, in health care industry, medical history of patient could be helpful for accurate diagnosis and treatment of diseases. But this data may contain some personal information about patient. Similarly, for the Finance sector customer's data, bank records, tax records, sales revenues, forecasts, accounting records, investment holdings and wages or income information are the example of data that could be misused while performing the analytics. Likewise, Education Industry, Government Sector, Media and Entertainment Industry, Weather Patterns, Transportation Industry, Manufacturing and Natural Resources, Retail and Wholesale trade and Internet of Things (IoT) also have the private data that need to be protected. Such data is an asset for any organisation and individual as well. Processing this type of data without masking encourages privacy disclosure. Maintaining the privacy of the sensitive data during analysis is known as Privacy Preservation. The crux of the privacy preservation is that, the application of Privacy Preservation Techniques should not change the accuracy of the outcome [1]. Efficient privacy preservation algorithms are applied to data but increasing volume and diversity is making the analysis process complex and time consuming [2]. Thus, rapid growth and heterogeneity of data mark challenges for the data analytics to have more advanced Privacy Preserving Data Analytics Model. This paper is intended to address research challenges of privacy preservation in data analytics.

The paper is organized in following sections. Section I is the introduction; Section II is giving the overview of present Data Analytics Techniques. Section III presents the Standard Privacy Preservation Techniques and respective glitches. Section IV explores Privacy Preserving Data Analytics approaches, attainment and recommendations. Section V discusses about the Datasets used by Researchers in Privacy Preservation. Section VI suggests the Processing Platforms for Privacy Experiments. Section VII narrates the Metrics of Privacy Preservation and Section VIII gives the conclusion and directions for future research.

2. Data Analytics Overview

Application of statistical techniques to derive meaningful information from the massive and diverse data is known as Data Analytics. Industries like social media, e-commerce, healthcare, banking, entertainment etc. are continuously digging out the information and patterns using data analytic. Such analysis could be of any type based on the expected output. Major categories of Data Analytics are: **Descriptive Analytics, Diagnostic Analytics, Predictive Analytics and Prescriptive Analytics**.

Descriptive analytics gives information about what happened while Diagnostic analysis reveals why it happened. Based on the findings of Descriptive and Diagnostic analysis, Predictive analytics predicts about future trends. Next, Prescriptive analytics tells that how to deal with a future problem.

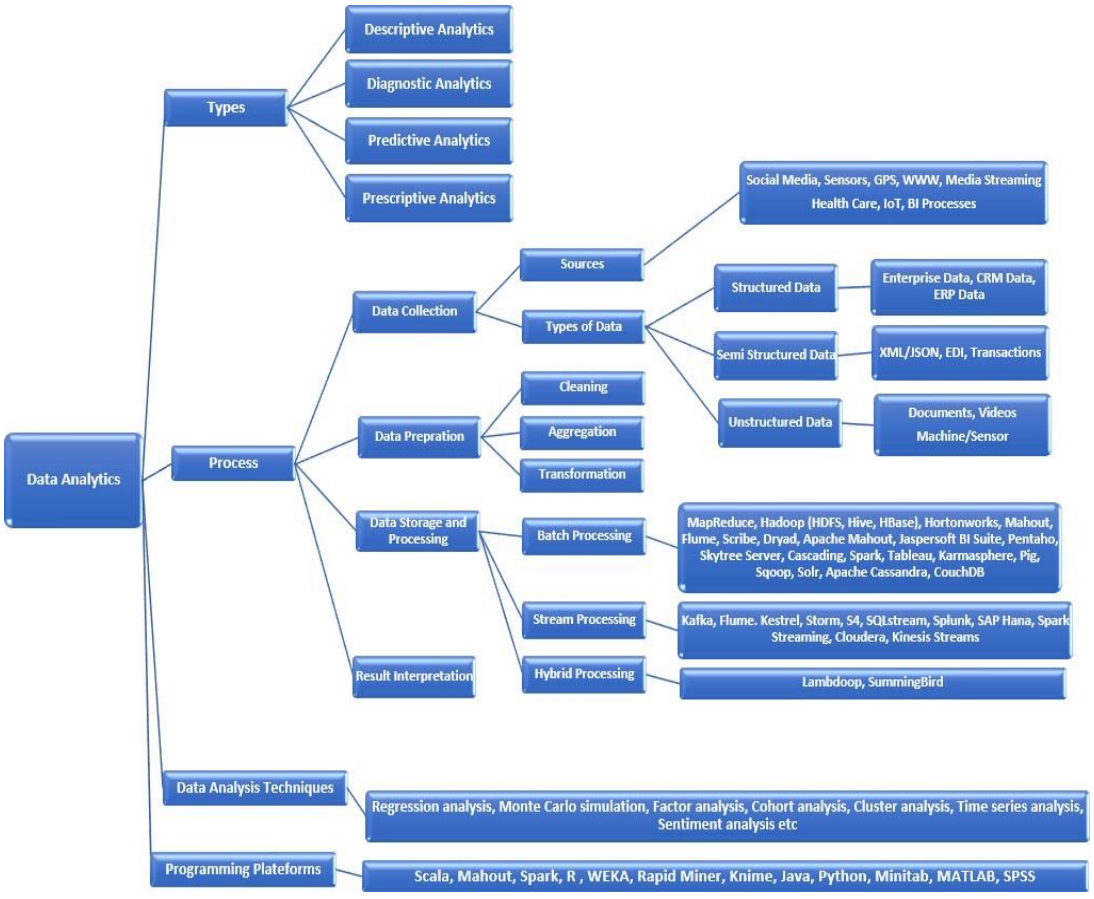

Figure 2.1: Overview of the Data Analytics

Data Analytics is not a one step process, rather it works as a layered approach. At first data gets collected from different sources, in various formats which later get pre-processed. When raw data turns into quality data then Analytics algorithms can be applied. Outcomes of the analytics can be interpreted using numerous visualization techniques. If the data is uniform then the whole process can be simple but due to difference in types and sources of data different types of handling platforms are required to store and process such data. Figure 2.1 gives a brief overview about Data Analytics. While performing any type of Analytics or any step of Analytics process, it is vital to get unaffected outcomes even after having data privacy.

3. Standard Privacy Preservation Techniques

Data privacy is an individual right for the data disclosure. When data is available publicly, Privacy threats may active. Privacy risk domain includes: Surveillance, Disclosure, Discrimination and Personal embracement and abuse [3]. To handle such threats, appropriate Privacy Preservation Technique must be used. To make any Privacy Preservation technique impactful, it should include following features:

a. It should not disclose sensitive information.
b. It should not compromise the access and the use of non-sensitive data.
c. It should not be restricted to any particular domain.
d. It should have low computational cost.

For any technique, attributes must be pre-identified i.e., for which class of data privacy preservation is required. Such data attributes can be classified into four categories:

a. Personal Information Identifier (PII): The attributes which can uniquely identify any individual, such as id number and phone/mobile number, are called PII.
b. Quasi Identifier (QID): Attributes which re-identify individual by linking with external data like gender, age and zip code.
c. Sensitive Attribute (SA): Any individual does not want to reveal such attributes, such as income and medical issues.
d. Non-Sensitive Attribute: The attributes which are not PII, QID, SA are called Non-Sensitive Attributes.

Some of the known techniques for privacy preservation are K anonymity, *l* Diversity, *t*-closeness, Data distribution, Cryptographic techniques, Personalized Privacy and Differential Privacy [3] [4].

Li et. al. discussed about **k-anonymity** which modifies the data before supplying for analytics. Here to maintain privacy, attributes are generalized till each row is identical with at least k-1 other rows. This technique protects the individual identity but for protection of descriptive attributes, found insufficient [5]. In addition, Homogeneity Attack and Background Knowledge Attack put limitations against the technique [6].

Taxonomy Tree is another anonymity-based approach to satisfy the privacy requirement [7] [8] [9]. In this method Top-Down Spinalization (TDS) is used to generalize the data. TDS starts with most general state and then iteratively specialize it. For a categorical attributes, a general value is specialized into a specific value and for continuous attributes, an interval is split into

two sub-intervals. To stop this iteration, violation of the anonymity requirement is the boundary condition. This method effectively preserves utility and privacy of data. With the moderate storage overhead, it scales well for large data sets [7].

***l* Diversity** is mentioned by Machanavajjhala et.al where, values of the sensitive attributes are well represented in each group [6]. The adverse fact is that, it does not consider the distribution of the sensitive attributes [4]. This type of privacy includes Skewness Attack and Similarity Attack [5].
To overcome the problem of skewness attack, new privacy measure ***t*- closeness** has been suggested by Li et.al. In the proposed technique *t*-closeness of an equivalence class depends on the fact that, the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table should not exceed the threshold *t*. Hence for a table to be in *t*-closeness all equivalence classes must have *t*-closeness [5]. In this approach with the privacy increment *t* decreases which results in the loss of information about correlation between quasi-identifier attributes and sensitive attributes [4].

According to **Data Distribution** scheme, data is distributed between multiple sites. The distribution can be done by two ways i.e. vertically and horizontally. Vertical data distribution allows sharing of dataset which have common data records for different attributes. Whereas in Horizontal data distribution sharing of different dataset is performed for the same attributes [10].

Data modification is another approach used to alter the original values before the release. **Perturbation** is the change of value or mixing of noise. It is also treated as **Randomization**. When data is accumulated for mining, adversary can fraudulently use the sensitive information. Data can be made noisy by two ways i.e. Additive Noise and Multiplicative Noise. In **Additive Noise method** noise is added to the data with known statistical distribution [11] whereas in **Multiplicative Noise method** noise is get multiplied [12].

Cryptographic Techniques are used to protect data using encryption methods. The major classification of encryption techniques is Public Key Encryption (PKE) and Symmetric Key Encryption (SKE). Other than this Searchable Encryption, (Hierarchical) Identity-Based Encryption (HIBE), Proxy Re-encryption (PRE), Predicate/Hierarchical Predicate Encryption (HPE) and (Fully) Homomorphic Encryption have also been used rigorously [13].

Person can define the degree of his/her privacy through **Personalized Privacy**. It is based on such table computation where all the attributes exist other than identifier attributes and sensitive attributes should be categorical [14].

Ɛ-Differential Privacy has given a new direction for the data analysis while ensuring the privacy. This scheme guarantees the unaffected outcome of analysis regardless addition or removal of a database item [15]. But its implementation leads to the problem of setting Ɛ and high data perturbation for numerical data [4].

4. Metrics of Privacy Preservation

For any privacy preserving algorithm the success can be measured on few parameters. The possible metrics for privacy preservation techniques could be:

1. **Accuracy/ Data utility -** This is to assess whether the data is meaningful and useful after imposing the privacy mechanism and to what extent the results are accurate.
2. **Privacy Analysis-** From such analysis it can be ascertained to what extent the confidentiality of the sensitive data has been maintained.
3. **Performance Analysis-** Performance of any algorithm could be measured in terms of execution time, memory overhead and efficiency.
4. **Completeness/ Information loss-** It is the measurement that implementation of any privacy technique not causing data/information loss.
5. **Availability-** When privacy is maintained for any data then it is may or may not be available to use when it is required.
6. **Scalability-** For any privacy technique scalability means that even if the data scales, its performance should be unaffected.
7. **Attack resistance-** This is the parameter that specifies that the technique is impenetrable.

5. Data Analytics & Privacy Preservation

5.1 Preliminaries

There are basically four categories of users in Data Analytics [16]. They are:

a. Data Producer: These could be individuals as well as companies directly or indirectly involved in Data Production. Such data may contain some sensitive information which could be misused.
b. Data Curator: These are the units which collect, store and process the data from the data producers. They also have to take care of data disclosure.
c. Data Consumers: These are the users which use the data for Analysis, Mining, Application development and Attackers as well. While analysis inferences also can be made for sensitive information.
d. Decision Maker: These person or organizations makes decision based on the results of the analysis. Here credibility and accuracy of the results are vital.

All of the standard Privacy Preservation Techniques discussed above doesn't ensure privacy maintenance in the operations involved in data analytics. Traditional methods mostly work well with small sized and homogeneous data but not with heterogeneous and scaling data. Hence, development of computationally effective and scalable Privacy Preserving Data Analytics is utmost required to handle variety of data.

Cuzzocrea et.al. have worked on the privacy of Temporal and Spatial data while the analytics operation [17]. While publishing the data, for privacy maintenance they have suggested Temporal Hierarchy Privacy-Preserving Model (THPPM). Although the model performed well in terms of data utility, reduction in number of records, unique record percentage and data compression ratio but still more investigation is required for data analytics operations on other scaling data as well as OLAP.

Zhao et.al. [18] have discussed about the challenges of the data trading in big data environment. Privacy preservation of the data provider is the major challenge for the given scenario. They have suggested a blockchain-based fair data trading protocol, where ring signature, double-authentication-preventing signature and similarity learning integrated. In the given protocol, enhancement in the anonymity of data provider and extension in the double-authentication-preventing signature are further improvement areas.

While using Machine Learning and Deep Learning techniques for analysis, privacy consideration is the need of the hour. In accordance with this sparse denoising autoencoder based DNN approach [19], Objective and Output Perturbation for differential privacy [20] and Independent Component Analysis (ICA) [21] are used. The techniques performed well in context of privacy protection and accuracy but computational effectiveness is yet to be defined. Table 1 shows the some research work according to different phases of Data Life Cycle.

Table1

Data Life Cycle Phase	**Algorithm/Techniques**	**Type of Privacy Technique**	**Privacy Integrated Analytics Operation**	**Evaluation Criteria**
Data Storage	RSA algorithm [22]	Cryptographic Techniques	Cluster Analysis	NIL
Data Analysis/processing	Sparse Denoising Autoencoder [19]		CNN Classification	Efficiency (Mean Squared Error)Accuracy
	Cosine Similarity Protocol [23]		Clustering, Classification	Data Privacy with volume and velocity
	SPPOLAP: State of-the-art algorithm for supporting distributed Privacy Preserving OLAP [24]	Perturbation	Aggregation	Quality Analysis, Effectiveness Analysis, Performance Analysis
	Salted Hashing Technique [25]		Aggregation	Not Mentioned
	Differential Privacy Preservation using Output Perturbation (OPP) and Objective Perturbation (OJP) [20]		Analysis	Prediction accuracy, Data utility and Privacy
	Searchable Encryption [23]	Cryptographic Techniques	Statistical Analysis and Correlation Rule Analysis	Execution Time
	Independent Component Analysis (ICA) [21]		Analysis	Privacy and Data Utility
Data Publishing	Temporal Hierarchy Privacy- Preserving Model (THPPM) [17] [26]	Generalization	Aggregation	Data utility, reduction in number of records, unique record percentage, data compression ratio, count queries, query execution time

	Ring Signatures [18]		Data Trading	Completeness, Confidentiality, Anonymity, Availability, Fairness, Accountability
	Minimum Instance Disclosure (MIDR Ris)-k Angelization [27]		Not Mentioned	Execution time analysis, Disclosurerisk, Accuracy, Scalability
	Scalable K-Anonymization(SKA) [28]	k-anonymity	Not Mentioned	Information loss and Execution Time
	Improved Scalable l-Diversity (ImSLD) [29]		Not Mentioned	Information loss and Execution Time
	Privacy preservation Algorithm for Big Data UsingOptimal geometric Transformations (PABIDOT) [30]	Perturbation	Not Mentioned	Time complexity, Scalability, Memory overhead, Classification accuracy, Biases and post-processing properties, Privacy analysis, Attack resistance
Data collection, Storage, Processing, Publishing	Distributed and Disposable approach (D&D) [31]		All Analytic soperations	Data Privacy

6. Datasets used by Researchers in Privacy Preservation

Some of the major datasets used during the experiments for the Privacy Preservation while Data Analytics are shown in table 2:

Table 2

Type of Data	Name of the Dataset	Size	Algorithm used
Temporal and Spatial Data	Non-identifiable information relating to parking tickets for the city of Toronto.	1.47 GB	Temporal Hierarchy Privacy-Preserving Model (THPPM)[17] [26]
	Parking tickets for the city of Buffalo in the New York State	413 MB	
Six-dimensional data cubes	APB, TPC-H and UsCensus	1 GB	SPPOLAP [24]
Image Data	MNIST, SVHN, CIFAR-10 andSTL-10	30 MB,2.28 GB, 175 MB, 2 KB	Differential Privacy [20]
Numerical and Categorical Data	CPS power system dataset		Independent Component Analysis (ICA) [21]
	Poker, Synthetic (Poker),adult, synthetic (adult)	2.4 GB, 11.2 MB, 4 MB	Scalable k-Anonymization [28]
	Poker, Synthetic (Poker)	2.4 GB, 11.2 MB	Improved Scalable l-Diversity (ImSLD) [29]
Text Data	Informatics for Integrating Biology & the Bedside (i2b2)	1.1 GB	Privacy Preserving Unstructured Data Publishing (PPUDP) [22]

There are numerous types of data formats are available in today's world such as text, audio, video, image, graphic, streaming, spatial and many more. From the above table, it is

understood that still various data format is still untouched in terms of experimentation for privacy preservation.

7. Processing Platforms for Privacy Experiments during Data Analytics

Hardware Setup-

In majority of experiments, computer with 2.6 GHz Intel(R) coreTM i7 64-bit operating system [17] [30] [26] with 4 GB or 8 GB or 16 GB memory are used. For Wireless bigdata and Edge Computing NVIDIA Tesla K80 GPU accelerator is also added [20]. For scalability Linux (SUSE Enterprise Server 11 SP4) SGI UV30 0 0 supercomputer, with 64 Intel Haswell 10-core processors, 25MB cache and 8TB of global shared memory connected by SGI's NUMAlink interconnect also used [30]. For scalable Bigdata, Hadoop cluster of three nodes [29] and fivenodes are used [28].

Software Setup-

To test the algorithms developed for privacy preservation, Python on Spyder [17] [26], Tensorflow [20], R programming [21], Python script [23], Java [32], Apache Pig and python
[28] [29] are some of the coding platforms.

8. Conclusion and Future Directions

The above study explains the data analytics scenario including types, process, techniques and programming platforms. This analysis also reveals the standard privacy preservation techniques with their implementation and expansion when used while data analytics. This study also suggests the types and names of the datasets where the researches have carried out. In addition, with this it also briefs about the standard hardware and software platforms for experimentation. Some techniques performed well on certain parameters but still they have to completely tested on the other metrics of privacy preservation. The unexperimented data types, data analytics techniques, latest hardware and software platforms to maintain the privacy of data, could be the new research avenues.

References

[1] J. Zhao, Y. Chen and W. Zhang, "Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions," *Artificial Intelligence for Physical-Layer WirelessCommunications*, vol. 7, pp. 48901 - 48911, 2019.

[2] M. Hao, H. Li, G. Xu, S. Liu and H. Yang, "Towards Efficient and Privacy-preserving Federated Deep Learning," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, China, 2019.

[3] P. R. M. Rao, S. M. Krishna and A. P. S. Kumar, "Privacy preservation techniques in big data analytics: a survey," *Journal of Big Data*, vol. 5, no. 1, pp. 1-12, 2018.

[4] R. Mendes and J. P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications," *IEEE Access*, vol. 5, pp. 10562-10582, 2017.

[5] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey,2007.