ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

A Privacy Preservation Framework for Big Data (Using Differential Privacy and Overlapped Slicing)

[1] Johny Antony P [2] Dr. Antony Selvadoss Thanamani
[1] Research Scholar [2] Professor and Head of the Department of Computer Science
[1][2] NGM College, Pollachi, Coimbatore

***Abstract:* -- We are in the midst of big data. The rate of data generation is increasing at a very rapid rate. We need to understand and analyze this data as quick as possible. A delay in millisecond to understand the data may cost not only money but also life. There are various processing and analytic mechanisms like Hadoop and MapReduce to process the data. But as big data comprises an enormous amount of personally identifiable information, user privacy and security is a major concern, and it is a massive challenge in big data. It is considered as an absolute prerequisite for exchanging sensitive information in terms of analysis, validation and publishing. The multidimensional anonymization and access control are widely-adopted privacy preservation approaches. Despite much research a method with satisfactory privacy settings are far from being achieved. Owing to the lack of integrating data from multiple sources, manual administration, video surveillance applications, the traditional methods are not feasible to big data. Hence, scalability is concerned as the major problem encountered when the conventional preservation techniques are applied to the big data. Some of the approaches have handled the security and privacy problems at the time of data shared among the different organizations. However, they do not efficiently preserve the data privacy since they fail in handling the attacks. In this paper we present a new frame work for preserving privacy using the basic concepts of differential privacy and overlapped slicing.**

***Keywords:*--Big data, Privacy preservation, Anonymization, Differential Privacy, Overlapped slicing.**

I. INTRODUCTION

We experience a real data deluge today. Big data brings revolution in the world of data. World Wide Web, mobile computing, and wireless technologies have emerged as a leading technology platform for potential applications. These applications generate huge amount of data [1]. The people understand the need for privacy preservation while publishing the data to research centers or other agencies as it contains sensitive information about individuals. Figure 1 shows that today the digital universe is so huge and growing huge exponentially. Big data raises concerns about the tracking and profiling of people and consumers as it expands to all domains. In Table 1, the characteristic of big data is better explained by eight V's.

Figure 1: Growth of Big data

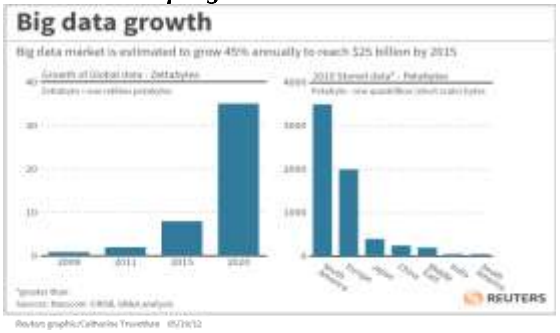
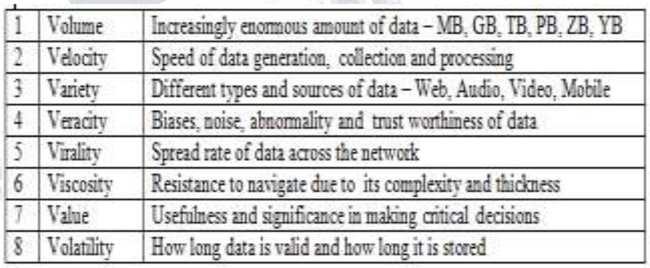
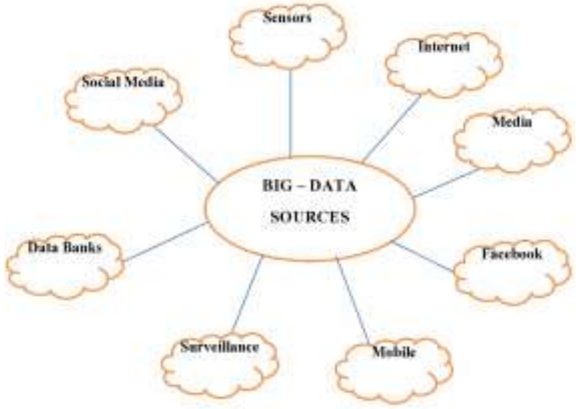


Table 1: Illustration of the 8 V's of Big Data



1	Volume	Increasingly enormous amount of data – MB, GB, TB, PB, ZB, YB
2	Velocity	Speed of data generation, collection and processing
3	Variety	Different types and sources of data – Web, Audio, Video, Mobile
4	Veracity	Biases, noise, abnormality and trust worthiness of data
5	Virality	Spread rate of data across the network
6	Viscosity	Resistance to navigate due to its complexity and thickness
7	Value	Usefulness and significance in making critical decisions
8	Volatility	How long data is valid and how long it is stored

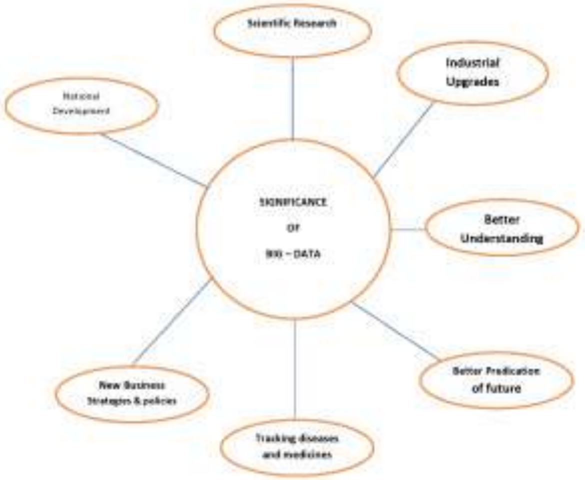


People want to keep privacy of the data before it is given for publishing or some research work. The PPDP is the problem of deciding how to publish useful data while preserving privacy-sensitive information according to the privacy requirements of data holders. According to the concept of the privacy protection, it is defined as such the accessing of published data must not allow the unwanted users to identify anything about the targeted individuals. The conventional privacy preserving mechanisms in data mining limited to small scale and static data [6].

The traditional techniques are inadequate in big data due to its characteristics such as variety, volume, velocity, diversity of data sources and formats, and streaming nature of data acquisition. Hence, it is essential to extend these approaches for preserving the privacy of big data. A strong, efficient, and scalable technique is essential to surmount the shortcomings. The connection between personal data and personal identification should be vanished. Such an anonymization must not only satisfy underlying privacy requirements but also safeguard the utility of the data. One needs to ensure the balance between privacy and utility.

In this paper we give a short view of basic models in privacy preservation and recent approaches and techniques to preserve privacy of big data. Moreover we discuss various challenges in the area of research in privacy preservation of big data. This paper is organized as follows. The basic models in big data and various privacy preservation techniques in the literature. Then we discuss the current approaches to this problem and its limitations. Finally a new framework is proposed for the big data privacy preservation.

II. SIGNIFICANCE OF PRIVACY PRESERVATION


Scientific Research
Industrial Upgrades
National Development
SIGNIFICANCE OF BIG – DATA
Better Understanding
New Business Strategies & policies
Better Predication of future
Tracking diseases and medicines

If the big data is captured, managed and analyzed effectively without any privacy, it has the power to change every industry performance, including cyber security, healthcare, transportation, education and the sciences. Individual activities are stored via social media and search engines which are part of big data, and it is stored on the web. Privacy breach in this information tends to lose the customer believe in a particular organization. Improper big data analytics creates the potential harms to individual privacy. For instance, improper usage of GPS
tracking data about an individual. Privacy preservation that plays a major role in a secure outsourcing of the private data to the cloud. The conventional security techniques used in small-scale static data are not applicable to big data due to its features.

III. BASIC MODEL IN PRIVACY PRESERVATION

3.1. Randomization

The method of randomization can be described as follows. Consider a set of data records denoted by X={x1,….xn}. For record xi ε X, we add a noise component which is drawn from the probability distribution fy(y). These noise components are drawn independently, and are denoted y1....yn.. Thus , new set of distorted records are denoted by x1+y1…..xn+yn.. we denote this new set of records by z1…….zn.

3.2. k-anonymity

K-anonymity is one of the basic privacy preservation model. In the k-anonymity, every published record has to be indistinguishable from at least (k-1) others on its QI attribute. The "quasi-identifiers" are the attributes available to an adversary. It is defined as: A table T satisfies k-anonymity if for every tuple t ∈ T there exist k−1 other tuples ti1 , ti2 , . . . , tik−1 ∈ T such that t[C] = ti1 [C] = ti2 [C] = ... = tik−1 [C] for all C ∈ Q. [11].

3.3 l-diversity

l-diversity is a group based anonymization model that assists to preserve the privacy of data through reducing the granularity of a data representation using generalization and suppression. In l-diversity, an equivalence class is said to have l-diversity if there is at least l "well-represented" value for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. It is defined as: A q block is ℓ-diverse if it contains at least ℓ "well-represented" values for the sensitive attribute S. A table is ℓ-diverse if every q block is at least ℓ-diverse [7].

3.4 t-closeness

t-closeness is another group based privacy model that extends the l-diversity model. It treats the values of an attribute distinctly, and considers the distribution of data values of the attribute to preserve the privacy. It uses the Earth Mover Distance (EMD) function to compute the closeness between two distributions of sensitive values. It is defined as: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is not more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness [8].

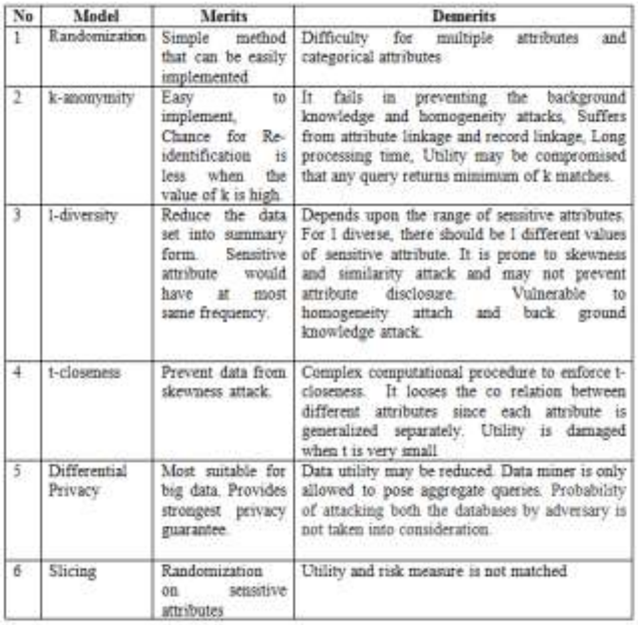
3.5. Differential Privacy

Differential Privacy offers one way forward that to extract insights from a database while guaranteeing that no individual can be identified. It achieves the guarantee of privacy by adding noise to answer to the queries. The amount of noise added must be large enough to conceal the effect of individuals and small enough that does not distort the genuineness of the answer.It is defined as : Let databases (D, D’) differing only in one row, meaning one is a subset of the other and the larger database contains just one additional row. A randomized function K gives ε-differential privacy if for all data sets D and D′ differing on at most one row, and all S ⊆Range(K) [17].

3.6.Slicing

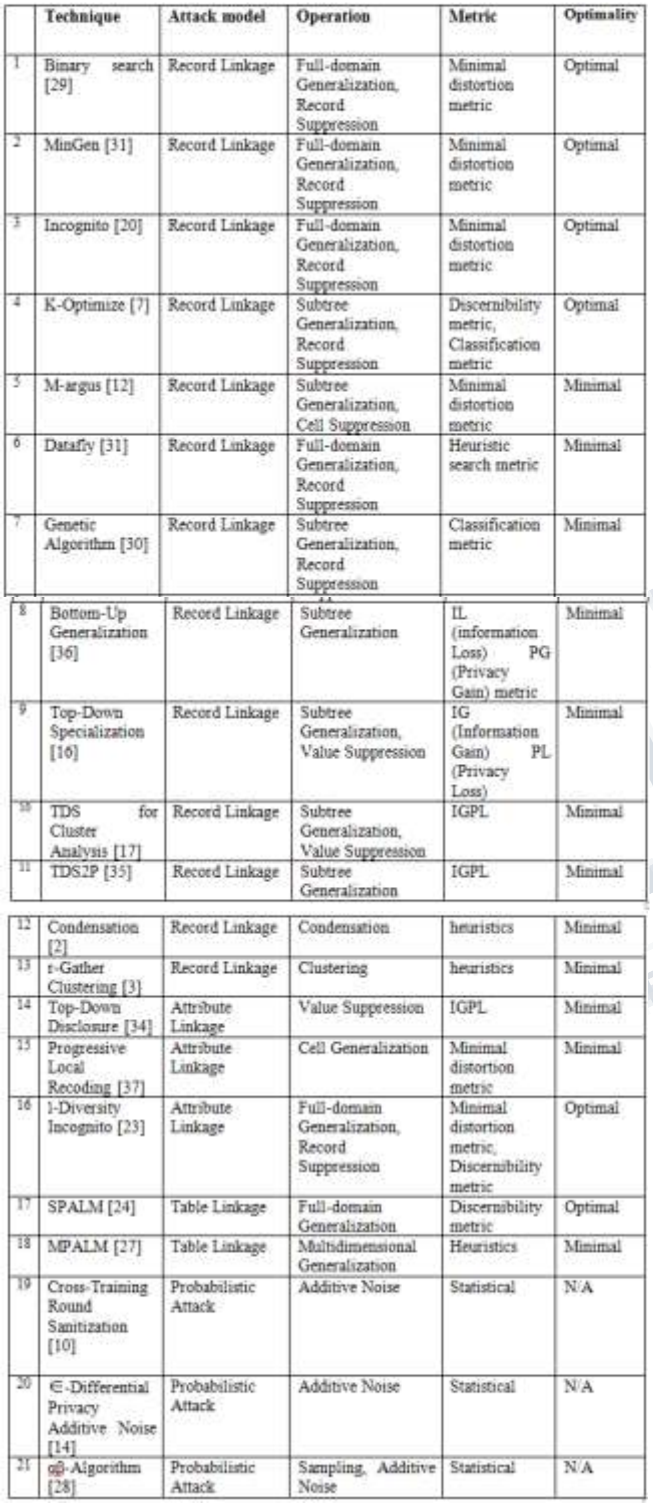
It is a technique that partitions data horizontally and vertically. The basic idea is to break the association cross columns but to preserve the association within each column.

Table 2. Merits and Demerits of basic models



No	Model	Merits	Demerits
1	Randomization	Simple method that can be easily implemented	Difficulty for multiple attributes and categorical attributes
2	k-anonymity	Easy to implement, Chance for Re-identification is less when the value of k is high	It fails in preventing the background knowledge and homogeneity attacks, Suffers from attribute linkage and record linkage, Long processing time, Utility may be compromised that any query returns minimum of k matches.
3	l-diversity	Reduce the data set into summary form. Sensitive attribute would have at most same frequency.	Depends upon the range of sensitive attributes. For l diverse, there should be l different values of sensitive attribute. It is prone to skewness and similarity attack and may not prevent attribute disclosure. Vulnerable to homogeneity attach and back ground knowledge attack.
4	t-closeness	Prevent data from skewness attack.	Complex computational procedure to enforce t-closeness. It looses the co relation between different attributes since each attribute is generalized separately. Utility is damaged when t is very small
5	Differential Privacy	Most suitable for big data. Provides strongest privacy guarantee.	Data utility may be reduced. Data miner is only allowed to pose aggregate queries. Probability of attacking both the databases by adversary is not taken into consideration.
6	Slicing	Randomization on sensitive attributes	Utility and risk measure is not matched

Table 3: Privacy preservation approaches in the Literature



	Technique	Attack model	Operation	Metric	Optimality
1	Binary search [29]	Record Linkage	Full-domain Generalization, Record Suppression	Minimal distortion metric	Optimal
2	MinGen [31]	Record Linkage	Full-domain Generalization, Record Suppression	Minimal distortion metric	Optimal
3	Incognito [20]	Record Linkage	Full-domain Generalization, Record Suppression	Minimal distortion metric	Optimal
4	K-Optimize [7]	Record Linkage	Subtree Generalization, Record Suppression	Discernibility metric, Classification metric	Optimal
5	M-argus [12]	Record Linkage	Subtree Generalization, Cell Suppression	Minimal distortion metric	Minimal
6	Datafly [31]	Record Linkage	Full-domain Generalization, Record Suppression	Heuristic search metric	Minimal
7	Genetic Algorithm [30]	Record Linkage	Subtree Generalization, Record Suppression	Classification metric	Minimal
8	Bottom-Up Generalization [36]	Record Linkage	Subtree Generalization	IL (information Loss) PG (Privacy Gain) metric	Minimal
9	Top-Down Specialization [16]	Record Linkage	Subtree Generalization, Value Suppression	IG (Information Gain) PL (Privacy Loss)	Minimal
10	TDS for Cluster Analysis [17]	Record Linkage	Subtree Generalization, Value Suppression	IGPL	Minimal
11	TDS2P [35]	Record Linkage	Subtree Generalization	IGPL	Minimal
12	Condensation [2]	Record Linkage	Condensation	heuristics	Minimal
13	r-Gather Clustering [3]	Record Linkage	Clustering	heuristics	Minimal
14	Top-Down Disclosure [34]	Attribute Linkage	Value Suppression	IGPL	Minimal
15	Progressive Local Recoding [37]	Attribute Linkage	Cell Generalization	Minimal distortion metric	Minimal
16	l-Diversity Incognito [23]	Attribute Linkage	Full-domain Generalization, Record Suppression	Minimal distortion metric, Discernibility metric	Optimal
17	SPALM [24]	Table Linkage	Full-domain Generalization	Discernibility metric	Optimal
18	MPALM [27]	Table Linkage	Multidimensional Generalization	Heuristics	Minimal
19	Cross-Training Round Sanitization [10]	Probabilistic Attack	Additive Noise	Statistical	N/A
20	∈-Differential Privacy Additive Noise [14]	Probabilistic Attack	Additive Noise	Statistical	N/A
21	αβ-Algorithm [28]	Probabilistic Attack	Sampling, Additive Noise	Statistical	N/A

IV. RECENT APPROACH TO PRIVACY PRESERVATION IN BIG DATA

Abid Mehmood et.al. (2016) presented existing privacy preserving mechanisms in the various life cycles of big data such as data generation (encryption and access restrictions), data storage (hybrid and private clouds) and data processing (anonymization techniques such as generalization, suppression, anatomization, permutation and perturbation) and various challenges of preserving privacy in big data. These methods are described with respect to the factors of scalability, privacy, time, efficiency and utility. Various risks involved in the encryption, anonymization and storage of data in the cloud are also investigated. When these techniques are applied, privacy is protected but the data may loose the meaning in the real world and as a result the utility and significance. An efficient PPDP algorithm must take into account proper trade-off between utility and privacy as the data is prone to any attacks. Therefore the methods/techniques must be modified or extended to handle the big data in an efficient manner [23].

Lei Xu et.al. (2016) created Rampart framework for privacy preservation. It consists of seven procedures namely anonymization, reconstruction, modification, provenance, agreement, trade and restriction to prevent outside intrusion. The framework tries to give high priority to maintain the balance between data utility and privacy. But more ways are to be explored to protect privacy against various threats [39].

Shaden Al-Aqeeli and Ghad Alinfie (2015) investigate privacy preserving problem of big data in the context of hybrid cloud computing and presents frameworks such as Airavat, Sedic, Sac-FRAPP and Hyper-1 based on MapReduce from the perspective of scalability, cost and compatibility. It is recorded that anonymization, encryption, differential privacy are the efficient methods for protecting privacy of data. The final analysis shows that the above said frameworks suffers from limitations such as data distortion and none of them is fully fit for privacy preservation [5].

Liye Fan and Hongxia Jin (2015) present two solutions – SRA and HPA - for privacy preservation based on differential privacy. The empirical studies with real world data show that the solutions enable accurate data analysis reducing the user privacy risk and data storage. The methods take care to achieve a balance between data loss and privacy. But this method fails while encountering with complex data analytical task and to preserve the privacy [4].The work in [35] by Xu Lei and et al reviews the privacy issues related to data mining and different approaches involved in privacy preservation that helps to define new methods to provide privacy to sensitive information. This survey explores various users participating in data mining applications. Moreover, it explains their privacy concerns and suitable privacy methods according to their sensitive information. Moreover, it gives a brief introduction to the basics of related research topics, some ideas for future research. Finally, the survey concludes with the game theoretical approaches that assist to identify the interactions among the discovered users [15].

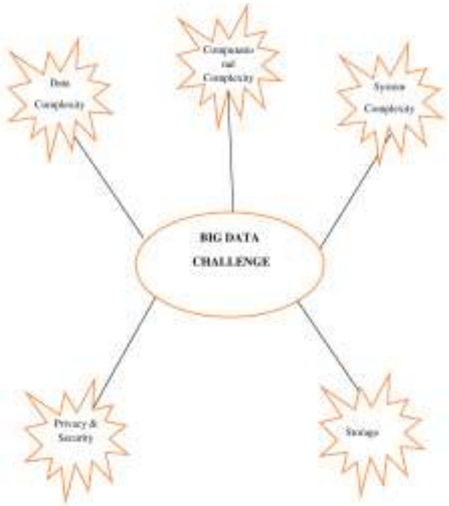
Hui Zhu and et al (2014) mad an approach towards the efficient and privacy-preserving computing in the big data era, and it exploits the new challenges of big data in privacy preservation. Initially, it defines the general architecture of big data analytics and discovers the privacy requirements in big data. Then, it finds out an efficient and privacy-preserving cosine similarity computing protocol [18].

Big data analytics such as purchase histories, medical data, and sensor data are done with the help of cloud. However, the analysis of data using the third-party cloud server has an unauthorized access risk. To overcome this problem, a privacy-preserving analysis technique using searchable encryption was introduced. It performs the text matching of encrypted text for statistical analysis and analysis of correlation rules without decrypting the data [26].

A novel approach in [40] discovers the framework to provide privacy to the social network, according to the user's personal privacy requests. With the intention of satisfying the various users, privacy needs it combines the label generalization and the structure protection techniques based on the gradually increasing attacker's background knowledge. The performance of the system is measured through the experimentation.

Disclosing the private information at the time of comparing the personal profiles of two users is the open problem in Proximity-based Mobile Social Networking (PMSN). The protocol proposed in [21] provides a solution to this problem. This protocol supports a wide range of matching metrics at different privacy levels and gives finer differentiation between a PMSN user.

V. RESEARCH CHALLENGES



Big data comprises great benefits, prospects, and promises while it poses numerous challenges to privacy and data protection. The improper use of data leads to privacy breaches and cause harm to data subjects and the data provider. A large number of privacy preservation techniques are proposed in the literature. However, the support is only limited to the data mining applications that manages small-scale and static data. The characteristics and nature of big data pose an immense challenge to privacy preservation. The scalability is a significant problem that occurs at the time of applying these techniques to big data. Again ensuring the utility of the data and preserving privacy is a great question before the researchers.

Moreover, big data comprises natural data which keeps changing continuously and thus, these techniques become obsolete. The existing privacy preservation approaches for big data lacks in providing efficient security to the data due to the mentioned problems. The storage is the most significant fact in big data processing. It needs massive computation and storage which demands the need of cloud computing. The application of cloud computing environment provides scalable solutions to solve the issues. Nevertheless, storage of data in the cloud increases the security problems. The extension of these techniques with the secure cloud infrastructure enables to preserve the privacy of the big data.

VI. PROPOSED METHOD

The proposed system aims at providing a novel privacy preservation framework to preserve the large medical data privacy using combined anonymization techniques. Fig.1 shows the steps involved in the proposed system. These are,

1. Taxonomy Tree Creation:

The process of the proposed system begins with the taxonomy tree creation which assists closely to classify the input data efficiently. A taxonomy tree is created using the table structure which consists of internal nodes and set of their leaves. For example, the internal node of table is address = {country} and a node country includes two leaves such as City and Taluk. The leave act as a node until it has at least one leave. From multiple table structure, the taxonomy tree can be created.

2. Slicing and Overlapping:

The slicing and overlapping approach is applied to anonymize the data set using the attribute partitioning and overlapping, column generalization and tuple partitioning.

3. Attribute Partitioning and Overlapping:

With the intention of increasing the data utility, the most correlated attributes are overlapped in more than one column through attribute overlapping. The attribute partitioning decides the correlated attributes or fields in the database, by computing the mean square contingency. The related attributes are under same column, for instance {address and phone number} and {Age and Sex}.

4. Column Generalization:

Assuming that each node has a corresponding conceptual leaves. The leaves provide more details than a node. For example, birth date in D/M (e.g. 13/Mar) are leaves and Year of birth (Y) is a node. Generalization replaces values of leaves with node's values. For example, birth D/M is replaced by Y.

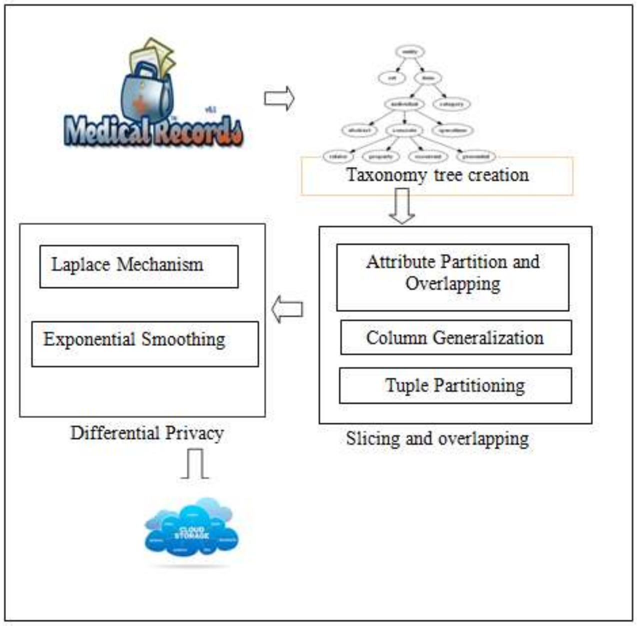
5. Tuple Partitioning:

The tuple values are grouped together to form buckets. The tuple partitioning algorithm is first divides the tuples into buckets and to check whether the each bucket satisfies the l-diversity which ensures "diversity" of sensitive values in each group. For instance, tuples of age is divided into two buckets such as <20 and <40.

6. Differential Privacy:

In order to provide strongest privacy guarantee to the data, the proposed system combines the differential privacy with the overlapped slicing. The differential privacy maximizes the query accuracy from statistical databases and minimizes the individual identification by adding the noise to the data using Laplace mechanism and the Exponential Smoothing takes into account the score of utility function in a differentially private manner. Eventually, the anonymized data set gets stored on the secure cloud infrastructure. Eventually, the anonymized data set gets stored on the secure cloud infrastructure.

Architecture of privacy preservation framework



VII. CONCLUSION

In this paper we presented a survey of big data, privacy preservation in big data and it's relevance and significance. Big data characteristics show that we need different software and techniques for the processing of big data. The paper also presents the basic models of privacy, different privacy preservation techniques, recent approaches and the futures challenges in this field. It provides the overview of privacy preservation techniques and the importance of big data privacy. Finally this work has attempted to provide a new framework for adequate data privacy of big data over the cloud. The system combines the differential privacy and overlapped slicing methods to preserve the privacy of dataset. As the data undergoes various transformations, the privacy is protected and as the truthfulness of data is preserved, an equilibrium between utility and privacy is maintained.

REFERENCES

1. Abid Mehmood, Iynkanran Natgunanathan, Yong Xiang, Gung Hua, "Protection of Big Data", IEEE Access 10.1109/2016.2558446 , May 9, 2016.

2. Aggarwal, C. C. and Yu, p. s., “A framework for condensation-based anonymization of string data”, Data Min. Knowl. Discov. Vol.13, No.3, 251-275, 2008a.

3. Aggarwal, C. C., Pei, j., and Zhang, B., “On privacy preservation against adversarial data mining”, In Proceedings of the 12th ACM SIGKDD. ACM, 2006.

4. Agrawal, D., Das, S., & El Abbadi, A., “Big data and cloud computing: current state and future opportunities”, In Proceedings of the 14th International Conference on Extending Database Technology, pp. 530-533, ACM, 2011.

5. Al-Aqeeli, S., & Alnifie, G. (2015). Preserving Privacy in MapReduce Based Clouds: Insight into Frameworks and Approaches. 2015 International Conference on Cloud Computing (ICCC). doi:10.1109/cloudcomp.2015.7149652.

6. B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, “Privacy-Preserving Data Publishing: A Survey of Recent Developments”, ACM Computer Survey, Vol. 42, No. 4, pp. 1- 53, 2010.

7. Bayardo, R. J. and Agrawal, R., “Data privacy through optimal k-anonymization”, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), pp.217–228, 2005.

8. Bayardo, Roberto J., and Rakesh Agrawal, "Data privacy through optimal k-anonymization", Proceedings on Data Engineering, ICDE, 21st International Conference on. IEEE, 2005.

9. Blum, A., Ligett, K., and Roth, A., “A learning theory approach to non-interactive database privacy”, In Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC). ACM, pp.609–618, 2008.

10. Chawla, s., Dwork, c.,Mcsherry, f., Smith, a., And Wee, h. , “Toward privacy in public databases”, In Proceedings of the Theory of Cryptography Conference (TCC), pp-363-385, 2005.

11. Cox, L. H., “Suppression methodology and statistical disclosure control”, J. Am. Statistical Assoc. Vol.75, No.370, pp.377–385, 1980.

12. Dalenius, T., “Finding a needle in a haystack - or identifying anonymous census record”, J. Official Statistics, Vol.2, No.3, pp.329–336, 1986.

13. De Waal, A. G., A. J. Hundepool, and L. C. R. J. Willenborg, “Argus: Software for statistical disclosure control of microdata”, CBS, 1996.

14. Dwork, C., “Differential privacy”, In Proceedings of the 33rd ICALP, pp.1-12, 2006.

15. Fan, L., & Jin, H. (2015). A Practical Framework for Privacy-Preserving Data Analytics. Proceedings of the 24th International Conference on World Wide Web - WWW '15. doi:10.1145/2736277.2741122

16. Fung, b. c. m., Wang, k., Wang, l., and Debbabi, M., “A framework for privacy-preserving cluster analysis”, In Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 46–51, 2008.

17. Fung, b. c. m., Wang, k., Wang, l., and Hung, p. c. k, “Privacy-preserving data publishing for cluster analysis”, Data Knowl. Engin. Vol.68, No. 6, pp.552–575, 2009.

18. Hui Zhu, Ximeng Liu, Liu, J.K., Jun Shao, “Toward efficient and privacy-preserving computing in big data era”, Network, IEEE, Vol.28, No.4, 2014.

19. Iyengar, V. S., “Transforming data to satisfy privacy constraints”, In Proceedings of the 8th ACM SIGKDD. ACM, pp. 279–288, 2002.

20. Lefevre, K., Dewitt, D. J., and Ramakrishnan, R., “Incognito: Efficient full-domain k-anonymity”, In Proceedings of ACM SIGMOD. ACM, pp. 49–60, 2005.

21. Li, Ming, et al. "Privacy-preserving distributed profile matching in proximity-based mobile social networks." IEEE Transactions on Wireless Communications, Vol.12, No. 5, pp.2024-2033, 2013.

22. Li, N., Li, T., and Venkatasubramanian, S., “t-closeness: Privacy beyond k-anonymity and l-diversity”, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 2006.

23. Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M., “l-diversity: Privacy beyond k-

anonymity", ACM Trans. Knowl. Discov. Data, Vol.1, No.1, 2007.

24. Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of Big Data Privacy. IEEE Access, 4, 1821-1834. doi:10.1109/access.2016.2558446

25. Mohammed, N., Chen, R., Fung, B., & Yu, P. S., "Differentially private data release for data mining", In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 493-501, ACM, 2011.

26. Naganuma, Ken, et al, "Privacy-preserving Analysis Technique for Secure, Cloud-based Big Data Analytics", Hitachi Review, Vol.63, No.9, pp.577-583, 2014.

27. Nergiz, m. e., Atzori, m., And Clifton, c. w., "Hiding the presence of individuals from shared databases", In Proceedings of ACM SIGMOD Conference. ACM, pp.665–676, 2007.

28. Rastogi, v., Suciu, d., and Hong, s., "The boundary between privacy and utility in data publishing", In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), pp. 531- 542, 2007.

29. Samarati, P, "Protecting respondents' identities in microdata release", IEEE Trans. Knowl. Data Engin. Vol.13, No.6, pp.1010–1027, 2001.

30. Samarati, P. and Sweeney, L., "Generalizing data to provide anonymity when disclosing information", In Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART (PODS), ACM, Vol.98, pp.188, 1998a.

31. Sweeney, L, "Datafly: A system for providing anonymity in medical data", In Proceedings of the IFIP TC11 WG11.3 11th International Conference on Database Securty XI: Status and Prospects. pp.356–381, 1998.

32. Sweeney, L., "k-Anonymity: A model for protecting privacy", Int. J. Uncertainty, Fuzziness, Knowl. Based Syst., Vol.10, pp.557-570, 2002b.

33. Wang, K. and Fung, B. C. M., "Anonymizing sequential releases", In Proceedings of the 12th ACM SIGKDD Conference. ACM, 2006.

34. Wang, k., Fung, b. c. m., and Dong, G., "Integrating private databases for data analysis", In Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI). pp.171-182, 2005.

35. Wang, k., Fung, b. c. m., and Yu, p. s., "Template-based privacy preservation in classification problems", In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM). pp. 466–473, 2005.

36. Wang, K., Yu, P. S., and Chakraborty, S., "Bottom-up generalization: A data mining solution to privacy protection", In Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), 2004.

37. Wong, R. C. W., Li., J., Fu, A. W. C., ANDWANG, K., "(a,k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing", In Proceedings of the 12th ACM SIGKDD, ACM, pp.754–759, 2006.

38. Xu, Lei, et al. "Information Security in Big Data: Privacy and Data Mining", Vol.2, pp.1149-1176, 2014.

39. Xu, L., Jiang, C., Chen, Y., Wang, J., & Ren, Y. (2016). A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining. Computer, 49(2), 54-62. doi:10.1109/mc.2016.43

40. Yuan, Mingxuan, Lei Chen, and Philip S. Yu. "Personalized privacy protection in social networks." Proceedings of the VLDB Endowment, Vol. 4, No. 2, pp. 141-150, 2010.