A Research on Bigdata Privacy Preservation Methods

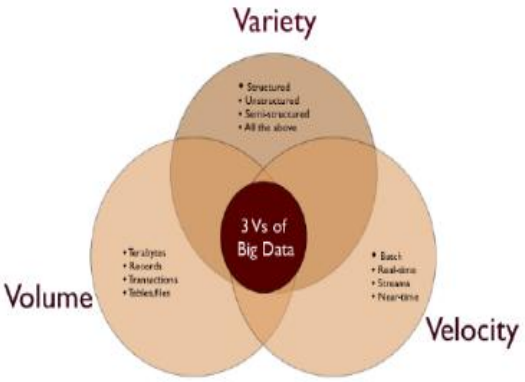
S.Subbalakshmi, K. Madhavi

Abstract:Big data is massive collection of big and complex data sets that cannot be stored and processed using traditional data processing systems. Hence big data requires high computational power and storage and big data uses distributed system. Big data Analytics means analyzing invisible data patterns from the larger data sets. The data sets are gathered from various sources i.e. social media, Business sector, healthcare, data governance, various institutions, etc. So, privacy and security is main concern in big data. This paper mainly focuses on l anonymity techniques preserve the privacy of data. This research aims to highlight three main Anonymization techniques used in a medical field namely, k-anonymity, l-diversity, and t-closeness.

Key words:Big data, Privacy, security, Data Anonymization-anonymity, l-diversity, t-closeness.

I. INTRODUCTION

Big data is related as the massive collection of data.so this information can't be processed by traditional technologies.Big Data is differentiated from the traditional technologies in 3 ways[1]:



Volume represents large amount of data. Data generated from multiple sources i.e. social media, Business sector, Healthcare, data governance, various institutions, etc. **Velocity** describes speed of data

Variety of data is coming from different sources in many formats such as images, audio, video, files, emails, financial transactions, simulations, 3D models, etc. In this paper we report several methods that can assist privacy in big data.

This paper is categorized as follows: Section II big data security and privacy challenges, in section III. Difference between security and privacy, in section IV.privacy preservation in big data, in section V. conclusion discussed.

II. BIGDATA SECURITY AND PRIVACY CHALLENGES

Security focus on protect the enterprise. Data privacy focuses on individual user's information. There are mainly three objectives of security are secrecy, reliability and accessibility. As indicated by late article by Cloud Security Alliance (CSA) [2], there are primarily 10 difficulties in the field of Big Data security and protection as referenced beneath:

1) Secured calculations in dispersed programming systems
2) Security best practices for non-relational information stores
3) End-point input approval and sifting
4) Real-time security observing
5) Privacy-safeguarding information mining and examination
6) Cryptographically upheld information driven security
7) Granular access control
8) Secure information stockpiling and exchanges logs
9) Granular
10) Data provenance

III. DIFFERENCE BETWEEN SECURITY AND PRIVACY

Security mostly center around ensure the undertaking. Information protection centers around individual client's data. There are for the most part three objectives of security are secrecy, reliability and accessibility .Security implies shields the information from unapproved clients. It primarily centers around the information and data as opposed to the individual data of people. Security can be accomplished without protection. Whereas privacy cannot be achieved without security.

Table 1: Distinction between security and privacy [3]

Security	privacy
It protects an enterprise	It protects individual users information
Security can be accomplished without protection	privacy cannot be accomplished without security

Revised Manuscript Received on June 10, 2019.

S.Subbalakshmi,Research Scholar,Department of CSE,JNTUCEA,Ananthapuramu,India (Email: subbalakshmi.btech@gmail.com)

Dr.K.Madhavi, Associate Professor of CSE, Department of CSE,JNTUCEA,Ananthapuramu,India (Email: kasamadhavi@yahoo.com)

It utilizes Various procedures like Encryption, Firewall and so forth.are utilized with the end goal to keep information trade off from innovation or vulnerabilities in the system of an association	The association can't offer its client/client's data to an outsider without earlier assent of client

IV. PRIVACY PROTECTION IN BIGDATA

Privacy is main concern in big data so we need efficient privacy preservation methods. Privacy directly related to customers. Privacy mainly focuses on user's individual data rather than entire collection of data. The privacy preservation techniques can be used protect the persons sensitive information. Privacy is important in 3 stages i.e. data generation, data storage, data processing. In this paper focusing on data Anonymization, l-diversity, t-closeness in big data storage.

A. Big data privacy in data storage phase

Big data stores large amount of data. There are different approaches to preserve privacy in storage. Security consist of mainly three dimensions i.e. confidentiality, integrity, availability [5].Cryptographic encryption mechanisms are Public key encryption, Identity based encryption, Attribute based encryption etc.. In public key encryption enables an information sender to scramble the information under the general population key of recipient, the recipient decrypts the data under private key recipient. So, there may be leakage of information. This cryptographic mechanism does not fulfill every one of the prerequisites of clients in the situation of enormous information stockpiling.

In traditional encryption mechanisms cannot acclaim the anonymity of cipher text receiver /sender. So, anyone can easily obtaining a cipher text (e.g. cloud server), if any one knows the public key of the figure message, that is scrambled under the proprietor of the figure content .So, outsider can without much of a stretch gets the plain text[6].

B. Data Anonymization

Data Anonymization means by removing the personal details to maintain the privacy of users [7,8]. It is also called as de-identification. Whenever organizations releasing the data publically by anonymizing it. Anonymization alludes to concealing the identifier qualities (the traits that particularly recognize the columns) like aadhar number ,bank a/c number ,full name, licence number, voter id etc.This anonymous data links to external data [9]. With the end goal to keep information from re-identification, the ideas of k-anonymity,l-diversity variety and t-closeness have been presented.

Classifications of attributes are

1 Key attributes: Based on the attributes uniquely identifies tuples . Ex: Social security number, pan number, aadhar number, voter id, driving license number etc..

2.Quasi Identifiers: An arrangement of traits that can be conceivably connected with outside data to re-distinguish entities.Ex:ZIP code, date of birth, sex.

3.Sensitive attributes:some of the attributes contains sensitive value with respect to data owner. Ex: salary and disease.

4.Non-sensitive attribute (NSA):disclosing the non-sensitive attributes will not break the secrecy of user.

i.k-anonymity: It can be used prevent record linkage. To preserve the privacy, the following Anonymization techniques are applied to the data [9,10, 11].

Suppression: quasi identifiers are supplanted or darkened by some steady qualities like 0,* and so on. Ex: some values license number ,aadhar number can be invisible using asterisk .

Generalization: Some values are replace by parent values.

The trait age can be written in more broad shape.

Ex: The attribute age can be written in more general form. Ex: age=19.The age attribute can be generalized to age<=20.

The below table1 shows the patient record. Whereas zip code,, Age, Nationality are Non-Sensitive attributes. Disease is sensitive attribute.

Table1: A Non-anonymized data of the patient record list

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Disease
1	11235	27	Indian	HeartDisease
2	11286	25	American	Flu
3	11286	20	Russian	Cancer
4	11235	23	American	Cancer
5	15235	55	Indian	Flu
6	15235	56	Japanese	Cancer
7	15233	45	American	Heart Disease
8	15236	44	Indian	Heart Disease
9	11235	33	American	Viral Infection
10	11235	37	American	Viral infection
11	11286	35	Indian	Flu
12	11286	38	American	Flu

In Table2 Nationality and zip code are suppressed and age attribute is generalized.

Table2 shows the anonymized data of the patient record list

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Disease
1	112**	<=30	Indian	Heart Disease
2	112**	<=30	American	Flu
3	112**	<=30	Russian	Cancer
4	112**	<=30	American	Cancer
5	152**	>=40	Indian	Flu
6	152**	>=40	Japanese	Cancer
7	152**	>=40	American	Heart Disease
8	152**	>=40	Indian	Heart Disease
9	112**	3*	American	Viral nfection
10	112**	3*	American	Viral nfection
11	112**	3*	Indian	Flu
12	112**	3*	American	Flu

K-anonymous data still consist vulnerable to attacks like background information attack,, homogeneity attack[6].Therefore we move for l-diversity Anonymization method.

ii.l-diversity:

In this segment the l-diversity variety idea is an endeavor to avoid homogeneity attacks and back ground data attack[13][14][15]. [16],[17]] let q ? be the summed up estimation of q in the distributed table T ? ; let s be a conceivable estimation of the sensitive attribute; let n(q ?,s') be the quantity of tuples t ? ϵ T ?where t ? [Q] = q ?andt ? [S] = s ' ; and let f(s ' | q ?) be the conditional probabilty of the sensitive attribute conditioned on the way that the nonsensitive attribute Q can be summed up to q ? .Ex:

	Non-Sensitive		Sensitive	
	Zip Code	Age	Nationality	Disease
1	1123*	<=40	*	HeartDisease
4	1123*	<=40	*	Cancer
9	1123*	<=40	*	ViralInfection
10	1123*	<=40	*	ViralInfection
5	1523*	>40	*	Flu
6	1523*	>40	*	Cancer
7	1523*	>40	*	Heart Disease
8	1523*	>40	*	Heart Disease
2	1128*	<=40	*	Flu
3	1128*	<=40	*	Cancer
11	1128*	<=40	*	Flu
12	1128*	<=40	*	Flu

Consider the inpatient records shown in Table 1. We present a 3-diverse version of the table in Table 3.

Comparing it with the 4-anonymous table in table 2 we see that the attacks against the 4-anonymous table are prevented by the 3-diverse table. For example, Alice cannot infer from the 3-diverse table that Bob (a 25 year old American from zip code 11286) has Flu.

iii. t-closeness:

It is extension of l-diversity than average assortment system. It treats estimations of the characteristics are phenomenal. Thusly, the data regards defend insurance. It uses the Earth Mover Distance (EMD) limit can be used to figure the closeness between two scatterings of delicate qualities. An equality class is said to be t-closeness if the separation between the conveyance of the characteristic in the all out table isn't in excess of a t edge. . A table is said to be t-closeness and all proportionality classes have t-closeness [18].

V. CONCLUSION:

In this paper we have shown that different types of privacy preservation methods i.e. k-anonymity, ?-diversity, t-closeness gives strong privacy in bigdata. There are several methods for future work. For future to preserve privacy a novel method is required for when data volumes increases and variety of data.

REFERENCES

1. Big Data Management and Analysis, December 2014
2. A Cloud Securty Alliance Collaborative research, "Expanded Top Ten Big Data Security andPrivacy challenges" , April 2013.
3. A Survey on Big Data & Privacy Preserving Publishing Techniques2017.
4. Xiao Z, Xiao Y. Security and privacy in cloud computing. In: IEEE Trans on communications surveys and tutorials, vol 15, no. 2, 2013. p. 843–59
5. Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage Kaitai Liang, Willy Susilo, Senior Member, IEEE, and Joseph K. Liu 2015.
6. Privacy Preservation in the Age of BigData :A SurveyJohn S. Davis II, Osonde A. Osoba
7. Protection of Big Data Privacy IEEE access January 2016
8. Privacy Preservation in Big DataAugust 2014
9. L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, pp. 557–570, 2002.
10. B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ``Privacy-preservingdata publishing: A survey of recent developments," *ACM Comput. Surv.*,vol. 42, no. 4, Jun. 2010, Art. no. 14.
11. Big Data Privacy Methods , 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939.
12. J. Sedayao, "Enhancing cloud security using data anonymization", White Paper, Intel Coporation.
13. Machanavajjhala, A., Kifer, D., Gehrke, J.andVenkitasubramaniam, M., 2007. *l*-diversity:Privacy beyond *k*-anonymity.*ACM Transactions onKnowledge Discovery from Data (TKDD)*, 1(1), p. 3.
14. Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on.IEEE*, 2007.
15. Truta, T. M., Campan, A. and Meyer, P., 2007.*Generating microdata with* p-*sensitive* k-*anonymity property*(pp. 124–141). Springer Berlin Heidelberg.
16. ℓ-Diversity: Privacy Beyond k-Anonymity AshwinMachanavajjhala Johannes Gehrke Daniel Kifer
17. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity.Availableat http://www.cs.cornell.edu/~mvnak, 2005.
18. Li, N., Li, T., and Venkatasubramanian, S., "t-closeness: Privacy beyond kanonymity and l-diversity", Inoceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 2006.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication