International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

PROTECTION OF BIG DATA PRIVACY

Chadalavada Lasya Chowdary[1], Sudarsanam Vaishnavi Sudha[2], Tellamkela Pooja Supriya[3], Seelam Hari Haran[4]

[1]- [4]Dept of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Guntur

---

Abstract - In recent years, big data became a hot research topic. The increasing amount of data also increases the prospect of breaching the privacy of people. There are variety of privacy-preserving mechanisms developed for privacy protection at different stages (e.g., data generation, data storage, and data processing) of an enormous data. The goal of this paper is to supply a comprehensive overview of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms. Especially, during this paper, we illustrate the infrastructure of massive data and therefore the state-of-the-art privacy-preserving mechanisms in each stage of the large data life cycle. Furthermore, we discuss the challenges and future research directions associated with privacy preservation in big data.

***Key Words*: Big data, privacy, Privacy Preservation, Data auditing, K-anonymity.**

INTRODUCTION

Due to technological development, the measure of information produced by person to person communication locales, sensor organizations, Web, medical services applications, and numerous different organizations, is definitely expanding step by step. All the tremendous sum of information produced from various sources in different formats with exceptionally rapid speed is called as Big Data. Enormous information has become an exploration region for recent years. The data generation is growing so rapidly that it's becoming extremely difficult to use traditional methods. Meanwhile, big data could be structured, semi-structured, or unstructured, which adds more challenges when performing data storage and processing tasks. Subsequently, to this end, we need better approaches to store and analyze information continuously. Large information, whenever caught and analyzed in an ideal way, can be changed over into meaningful insights.

Users security might be penetrated under the accompanying conditions :

- Personal data when joined with outer datasets may prompt the deduction of new realities about the clients/users. Those realities might be sensitive and shouldn't be uncovered to other people.
- Personal data is now and then gathered and utilized to enhance business. For instance, person's shopping propensities may uncover a great deal of individual data.

Limiting the danger of protection spills and expanding information accessibility are both the objectives of security assurance in information distributing situation. To secure individual protection, a few associations normally erase or encrypt explicit identiers which can plainly distinguish clients from the distributed information tables, for example, name, social protection code, and so on. This methodology can't effectively pre- vent aggressors from connecting other data in the distributed information with information gotten from different sources i.e., linking attacks. In order to solve the leakage problem caused by linking attacks, k-anonymity privacy protection model is proposed.

Anonymization

The anonymization method aims at making the individual record be indistinguishable among a group record by using techniques of generalization and suppression. The information are anonymized by eliminating the identifiers and adjusting the semi identifiers prior to distributing or putting away for additional preparing. How much information should be anonymised predominantly relies upon how much security we need to protect in that information. The security models are fundamentally classified into two classifications the first classification depends on the suspicion that the attacker can recognize the records of a specific client by connecting the records with outer information sources. The second classification depends on the suspicion that the attacker has enough background information i.e., the attacker can make a confident guess about whether the specific client's record exists in the data base or then again not.

There are a few models proposed to manage the abovementioned issues. Some of them incorporate k-anonymity.

Anonymization techniques

The popular k-anonymity techniques typically generalize and suppress the quasi-identifier attributes that will leak data.

Generalization: The basic idea of generalization as shown in Fig-1 is that the quasi-identifier attributes within the same equivalence category ought to get replaced by an equivalent generalized values.

Suppression: This technique can be considered a special sort of generalization, wherever all values are replaced by "*".

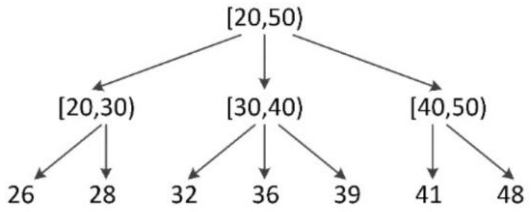

Fig -1: Generalization

Conceptions of k-Anonimity

Turning a dataset into a k-anonymous (and possibly l-diverse or t-close) dataset is a complex problem. To protect the security of a client, the original information are thought to be delicate and private and comprise of various records.

Identifier: An attribute or a set of attributes that can identify a unique individual, such as ID card number, zip code, name, etc. Before publishing, these identifier attributes are usually deleted, encrypted to protect the individual's identity.

Quasi-Identifier: Quasi identifier is a bunch of non-sensitive attributes (e.g., q1, q2,.. .,qm) in the information table that can be SQL-connected with an outer data table so that at any rate one individual can be recognized once more, wherein any single attribute can't distinguish unique person. A bunch of attributes like this is known as a quasi Identifier(QI) or semi-identifier. Connecting a set of quasi identifiers together can conceivably distinguish the person's attributes.is a bunch of semi identifiers including age, sexual orientation, and postal code in clinical protection safeguarding situation, every one of the traits is put away in various tables, and patient's sickness is viewed as a sensitive identifier. Attackers can connect these tables to a medical record table through the semi identifiers above, and afterward they can gather the patient's sickness data without any problem.

Sensitive Identifier: Sensitive identifiers are fields that need to be protected such as disease information in medical data, employee wages, ID numbers, cell phone numbers, etc.

Non-sensitive attribute (NSA): Non-sensitive attributes are which if disclosed will not violate the Protection of Big Data Privacy privacy of the user. All attributes other than identifier, quasi-identifier and sensitive attributes are classified as non-sensitive attributes.

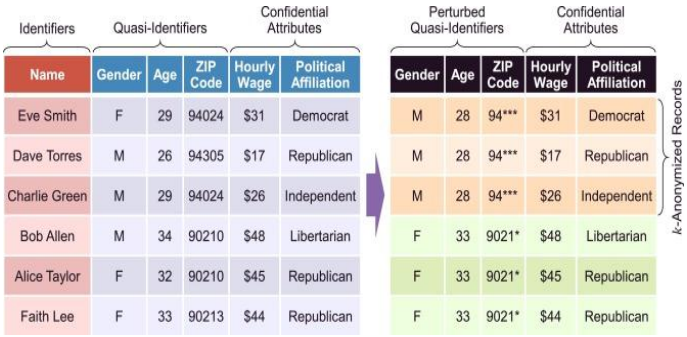

Fig -2: Example of k-Anonymous data

Here the quasi identifiers, gender, age and zipcode are generalized and suppressed. These attributes in Fig-2 are viewed as quasi identifiers because when they are combined with outer data, reveal the identity of an individual. Therefore the attributes like age are generalized as shown in Fig-1 and attributes like zip code are suppressed as shown in Fig-3.

Name	Gender	Zip code	Age	Country
Smith	M	54***	<20	America
Robin	M	98***	25-45	Russia
Simran	F	58***	30-40	India
Tom	M	98***	>50	England
Sam	F	540**	<25	Iran
Jyo	F	943**	45-55	Brazil

Fig -3: k-Anonymized sensitive data

CONCLUSIONS

In this paper, we propose k-anonymity algorithm for privacy preserving in data publishing. Contrasted and the old-style k-anonymity calculations, this calculation can effectively diminish the data misfortune, improve the precision of the semi-identifier bunch in the distributed dataset, and give better information for ensuing information mining. The trial result shows that our calculation could diminish the data misfortune produced by anonymization. It additionally performs well with various estimations of k and n. We likewise propose a few examinations as per our calculation in the situation of information anonymization. In future we can use other techniques as t-closeness.

REFERENCES

[1] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. Zürich, Switzerland: McKinsey Global Inst., Jun. 2011, pp. 1–137.

[2] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652–687, Jul. 2014.

[3] Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream manag ement. VLDB J. 2003;12(2):120–39.

[4] Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1–12.

[5] F. Armknecht, J.-M. Bohli, G. O. Karame, Z. Liu, and C. A. Reuter, "Outsourced proofs of retrievability," in Proc. ACM Conf. Comput. Commun. Secur., Nov. 2014, pp. 831–843.

[6] L. Sweeney, "κ-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowl. Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

[7] 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 on information assurance and security (IAS). 2013.

[8] Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access. 2014;2:1149–76.

[9] S. Singla and J. Singh, "Cloud data security using authentication and encryption technique," Global J. Comput. Sci. Technol., vol. 13, no. 3, pp. 2232–2235, Jul. 2013.

[10] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, 2009.

[11] W. Xiao-Dan, Y. Dian-Min, L. Feng-Li, and C. Chao-Hsien, "Distributed model based sampling technique for privacy preserving clustering," in Proc. Int. Conf. Manage. Sci. Eng., Aug. 2007, pp. 192–197.

[12] Li N, et al. t-Closeness: privacy beyond *k*-anonymity and *L*-diversity. In: Data engineering (ICDE) IEEE 23rd international conference; 2007.

[13] S. Shirkhorshidi, S. R. Aghabozorgi, Y. W. Teh, and T. Herawan, "Big data clustering: A review," in Proc. Int. Conf. Comput. Sci. Appl., 2014, pp. 707–720.

[14] Liu C, Ranjan R, Zhang X, Yang C, Georgakopoulos D, Chen J. Public auditing for big data storage in cloud computing—a survey. In: Proc. of IEEE Int. Conf. on computational science and engineering. 2013. p. 1128–35.

[15] Gionis and T. Tassa, "κ-anonymization with minimal loss of information," IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, pp. 206–219, Feb. 2009.