A Journey on Privacy protection strategies in big data

Ms.D.Viji
Assistant Professor, Department of computer science and engineering
SRM University, Kattankulathur
E-mail: dviji2k@gmail.com

K. Saravanan
Assistant Professor ,Department of computer science and engineering
K.R.S college of engineering, Vandavasi
sksarwan234@gmail.com

D. Hemavathi
Assistant Professor, Department of information and technology
SRM University, Kattankulathur
E-mail:hemavathi.d@ktr.srmuniv.ac.in

Abstract— In this modern world providing security for the data is the great challenging task. Especially handling of big data is a great issue because of its volume and variety of data structure. There are various strategies for storing the big data in an efficient way. But the consideration of privacy look up is very important. Privacy preservation varies from different stage of big data life cycle. Due to multi tenancy and massive computation issues, it is become a demanding task. While considering the Framework security, data security, integrity constraints management protecting big data privacy is plays an important role. This paper surveys the privacy requirements, obstacles and the techniques to handle privacy protection strategies in big data.

Keywords: Attribute based encryption, homomorphic encryption, De identification, differential privacy

I. INTRODUCTION

Big data is nothing but large collection of meaning full information. It is also referred as pool of vast and complex data sets. So that the traditional database systems cannot handle these types of data set. Since its contains large volume with different kinds of data like text, audio, video etc., are generated by different streams. Due to technology development massive data are generated each and every day. Have to focus on security and privacy of data in big data analytics. Both are considered as a vital challenges of big data. Data privacy is the most important thing since now-a-days most of them using social media in that personal information are shared at the time of data transmission through internet easily can track the identification of person [1]. Security of data is different from the data privacy. Privacy is only concentrating on distinct person of data are gathered, shared and used in right perspective way. Security focusing on caring data from intruders attack and abuse of data for the some other purpose based on money minded [2]. Privacy required in big data analytics due to involved in numerous administrations; a bulky data are not able to consume these normal security and privacy protection schemes for that in this section focus on upgrading big data platforms with the facility of privacy protection abilities.

A. Privacy requirements in big data

The following environments may be break the users' privacy in big data technology[4]:

- During data transmission over the internet personal data or information's are shared with some external resources. Due to this type of incidents third party may conclude realities about the users.
- Sometimes personal data are gathered and exploit for the business purpose. For example todays' online shopping is the hectic modern technology through this easily can predict the users habits and lot of personal information.
- Data trickling occurred at the stage of storing and data processing phases. In the data life cycle three important phase first data gathering or collection, second data storage and data processing. In that data privacy is more important on second and third phases of the life cycle.

Traditional DBMS techniques are insufficient for the big data framework due to desired to manage huge data volumes, heterogeneity of the data, and high velocity must be concentrated on that.

II. RELATED WORK

A. Big Data Life Cycle

Have to design well-organized and effective methods to process massive amount of data gathering at different velocity from various sources. Big data technology has to give privacy of data in all phases of life cycle [3].

Data are spread over the world for that new systems are established to store and process big amount of data. The following technologies cloud computing, Hadoop map reduce are recently used for bid data storage and processing. Privacy of big data how achieved in cloud computing when data are processed in storage and processing of big data life cycle.

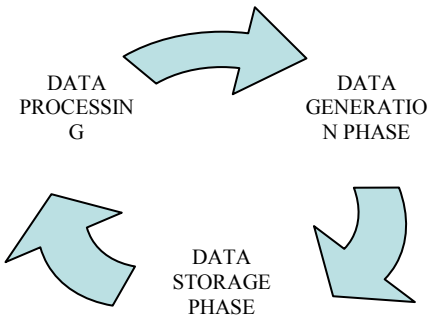

Fig 1 Life cycle of Big data analytics

B. Big data in cloud computing

Big data are needs large amount of computation and storage systems for that cloud computing involved in big data analytics it offers massive processing and vital storage ability. Even though cloud computing also contain issues of privacy of data. Following challenges are involved in processing and storage system on cloud [7].

Outsourcing: Outsourced data means user does not have control on that data this is the main reason for the cloud insecurity it will create major problems on privacy of cloud system users.

Multi-tenancy: Cloud environment share the same storage location to the various cloud users at this stage very easy to hack the data by the third party user really not related to that data. This will makes a very serious issues on data break and computation break. This leads privacy and security risks.

Massive computation: Traditional systems are not sufficient to protect individual privacy due to the proficiency of cloud system for managing enormous amount of data storage and computations..

1)Big data privacy in data generation phase:

In data generation phase data can be passed to third party in two ways, through the user's knowledge information are passed to the other and second thing without knowledge of user, data are collected by unauthorized person. At the time of user performing actions on online some intruders can hack the user personal information. In this case data owner want to protect their privacy data from the third person. To avoid this kind of data privacy problem first user avoid to share most sensitive data on the internet. In some cases its necessity means user can hide their data with the help of some tools, for example using Socket puppet tool data owner can able to hide their personal information from the third party. While using credit card and debit card details on web for online shopping at that time some security tools mask the data owner's identities.

2) Big data privacy in data storage phase

In big data technology normal storage methods are not sufficient so obviously data are stored and maintained in cloud environment. While data are stored on cloud data privacy has to met following three dimensions, confidentiality, integrity and availability. Confidentiality and integrity are fully related to the privacy of the data. But availability of data make sure that can able to access only by authorized persons. At the storage phase some encryption technologies preserves privacy of the data owner[3].

Identity based encryption: Access control based on the identity of the user but limitation of this method is time consuming, granular access control.

Attribute based encryption: Access control full based on user's attribute more secure and flexible as granular access control. But handling different category users very difficult. Updating cipher text receiver not possible.

Proxy re-encryption: Updating cipher text receiver is possible.

Homomorphic encryption: Computation on the performing data and very secure.

3) Big data privacy preserving in data processing

In data processing phase divided as a bi-part and protect privacy of the data. First part to concentrate on protection of data from unwanted leakage during processing stage since the collected information's are more sensitive data. Second part mainly focusing on the extracting meaningful information from the data without losing privacy.

III. OBSTACLE IN BIG DATA PRIVACY

Many issues are associated with big data privacy that are classified into four categories such as Framework security, Data privacy, Data administration, and Integrity security.

Framework security: Big data technology follows distributed computing infrastructure in this structure multiple users works across parallel so that identification of intruders or malicious users very crucial. Today many institutions transferred from traditional database into NoSQL database for handling unstructured and semi structured data. NoSQL provides architecture flexibility for the multi sourced data but also leaves it unsafe attack .

Data Privacy: Data are collected from the various sources and have to maintain privacy in analytic stage. For that in this paper already discussed access restriction on data and through some encryption technique can achieve preserving privacy to the data.

Data administration: Big data collected from different sources, so that it contain millions of end-users. Origin of the complexity in big data volume grows day by day. In this each data object are attached with metadata that provides information about object's creation. In big data applications provenance metadata will be a complex thing due to large provenance graph generated.

Integrity security: Input validation and filtering process is a vital challenge in big data application. Due to size of the data very difficult to identify whether the input data comes from authorized source or not, if it is not authorized source then we have restrict that source of data for the further process.

Real-time security monitoring is planned to alert the institution at the first stage of attack itself. SIEM systems,

focus to provide feedback of the organization's in real time. Some organizations analyze this feedback to identify real attacks from false results.

IV. TECHNIQUES FOR PROTECTING PRIVACY IN BIG DATA

A. De-identification

De-identification [8,9] is a popular old technique for protecting privacy in data mining and it can be migrated to privacy preserving in big data analytics. Intruders can get more information in the de-identification on the platform of big data. So that de-identification method is not enough to preserving privacy in big data. For that have to enhance this method with the help of some privacy preserving methods such as K-anonymity, L-diversity and T-closeness.

1) K-anonymity: In this method database maintained as a table format like rows and columns. This system constructing and evaluating algorithms for release information and expose about properties of entities that are to be protected. But this method has limitation of Homogeneity-attack, background knowledge. Computational complexity O(k logk) [10].

2) L-diversity: This method is extension of K-anonymity model also resolve some weakness in the k-anonymity model. The l-diversity model preserving privacy in datasets and reducing the granularity of data. In this method each sensitive attributes are represented by well-represented values. Major issue in this method is that it based on the range of sensitive data. Computational complexity O((n2)/k).

3) T-closeness: This method is further advancement of l-diversity that is help to protecting privacy of datasets and also reduce the granularity of a data representation. t-closeness is distance between sensitive attribute in the class and the distribution of attribute in the whole table is less than threshold t. The big advantage of t- closeness is it in prevent about attribute announcements. The issue in this method is if size and variety of data increases, the odds of re-identification process too rose. Computational complexity 2O(n)O(m)[11].

B. Differential privacy

Differential privacy (DP) [12] technology that allows database analysts to get needed information from the database that contains personal information of the users but without enlightening the personal identities of the individuals. This can be achieved through least amount of interruption in the information provided by the database system. Older some methods are did like hiding some information that cannot give guarantee the protection of individual identity. Differential privacy gives solution to this problem. The personal information of the user stored in database. The DP analyst wants to access those information means through the help of some software can able to done. This method restricts direct access of data from the database, for that intermediary software developed and placed between the database and analyst. This software can help to preserving privacy of the data.

Step 1: The data analyst send the request to the database with the help of intermediate software.
Step 2: The software get that request and processed this query.
Step 3: After that the software gets result from the database.
Step 4: Add some alteration in that result depends on evaluated privacy risk and finally handover to the analyst.
The amount of wrap added to the original information is proportional to the privacy risk evaluated at the stage of second. Amount of privacy risk low means wrap added also small. But if privacy risk high then more distortion added.

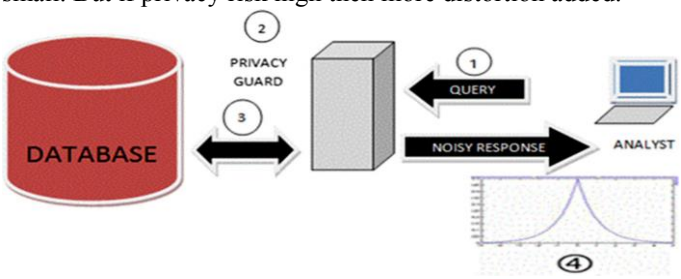

Fig. 2 Differential privacy big data differential privacy (DP) as a solution to privacy-preserving in big data is shown

CONCLUSION

In this paper big data privacy protection strategies of various stages has been analyzed. The techniques of de identification and differential privacy has been analysed for preserving big data privacy. It surveys only about the storage and the various structures of big data. It is a tricky task to preserve the privacy for the streaming data. And as a future work we decided to concentrate on privacy protection policies of continuous frames of data.

References

[1] 1. Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016;3(2):36–45.

[2] 2. Jing Q, et al. Security of the internet of things: perspectives and challenges. Wirel Netw. 2014;20(8):2481–501.

[3] 3. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.

[4] 4. A. Katal, M. Wazid, and R. H. Goudar, ``Big data: Issues, challenges, tools and good practices," in Proc. IEEE Int. Conf. Contemp. Comput., Aug. 2013, pp. 404_409.

[5] 5. B. Matturdi, X. Zhou, S, Li, and F. Lin, "Big data security and privacy:A review," China Communications, vol. 11, no. 14, pp. 135–145, Apr.2014

[6] 6. Z. Xiao and Y. Xiao, ``Security and privacy in cloud computing," IEEE Commun. Surveys Tuts., vol. 15, no. 2, pp. 843_859, May 2013.

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

[8] 8. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. In: Proc. 22nd international conference data engineering (ICDE); 2006. p. 24.

[9] 9. Meyerson A, Williams R. On the complexity of optimal k-anonymity. In: Proc. of the ACM Symp. on principles of database systems. 2004.

[10] 10. Bredereck R, Nichterlein A, Niedermeier R, Philip G. The effect of homogeneity on the complexity of k-anonymity. In:FCT; 2011. p. 53–64.

[11] 11. Microsoft differential privacy for everyone, [online]. 2015. http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf.

[12] 12. 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.

[13] 13. Sokolova M, Matwin S. Personal privacy protection in time of big data. Berlin: Springer; 2015.

[14] 14. Xu K, et al. Privacy-preserving machine learning algorithms for big data systems. In: Distributed computing systems (ICDCS) IEEE 35th international conference; 2015.

[15] 15. Hu J, Vasilakos AV. Energy Big data analytics and security: challenges and opportunities. IEEE Trans Smart Grid. 2016;7(5):2423–36.