Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering pp. 67–71
DOI: 10.15439/2017R45
ACSIS, Vol. 10 ISSN 2300-5963

Privacy and Security of User's Sensitive Data: A Viable Analysis

J. S. Rauthan
Research Scholar
Department of Computer Science & Engineering
Uttarakhand Technical University, Dehradun, India.

K. S. Vaisla
Associate Professor
Department of Computer Science & Engineering
Uttarakhand Technical University, Dehradun, India.

***Abstract*— Big data is a collection of large amount of data. Big Data is known for any collection of data sets which is enormous and intricate that it becomes challenging to process using on-hand database management tools or traditional data processing applications. Because data stored in these days are too large in enormous in size, Security and privacy of user's sensitive data is a great challenge in this era. There are too many methods and techniques are introduced in past decades for storing and protecting the user's sensitive data such as cryptographically techniques or anonymization method which derive to hiding the sensitive data. While the anonymization techniques cannot fulfill requirements of preserving privacy of sensitive data. So we require a system of techniques, that the sensitive information can be protected from hacker.**

***Index Terms*—Big Data, Security & privacy, User's Sensitive Data, Data Anonymization, Confidentiality.**

I. Introduction

The slang of big data is a watchword used to elaborate a great amount of structured and unstructured data. Traditional database and software's cannot efficiently process big data. Now day's industries, academics and researches generate large amount of data, which access at higher speed and higher than the existing capacity. Big data [1] facilitate to industries, researchers, academics to help their work, operations efficiently, that's why they can find a perfect solutions.

Big data is used by various companies and researchers for online searching applications and they search over those big data for appropriate results. Big data may vary from sizes to petabytes or Exabyte's, which consist of various records consisting of industries, academics, research, and mobile information and so on. Basically these data stored in big data era are in unstructured, unfinished and unapproachable [2].

Big data having characteristics unknown sources, massive amount of data, heterogeneous data, decentralized, unstructured and complex relations among these data. Data sharing is the basic and ultimate object of big data, that's why data privacy in an important challenge of big data. The object of privacy is to preserve the integrity, confidentiality and preventing the leakage of sensitive data.

Big Data may contain massive amount of information and those data may contain more sensitive and confidential information of the users. To preserve the confidentiality of user's sensitive data it is to name as privacy of user's sensitive data in big data platform [3, 4].

An experimental and widely-adopted methods and technique for data confidentiality is to anonymizing the data [5]. Anonymization of data means to hide data confidentiality and the identity of individual user while the information from big data is being analyzed and queried. That's why there are various techniques which focus on methods which may provide privacy of user's sensitive data.

Privacy of users sensitive and confidential data is one of the most important and hot topic in research in big data processing and applications. Due to the limited number of research result, development confidentiality techniques and solutions to provide big data sensitivity. For data privacy of unauthorized access we need to enforced security schemes and policies to secure sensitive data effectively securing and protecting confidential data in big data storage and as well as in transmitting is one of the other challenging task[6].

So protecting the leakage of sensitive data at the time of processing over web and at the time of data rest.

Finally we needed a system that can protect the data confidentiality and privacy of user's sensitive data in a large range of web applications against random server compromises.

II. User Role-Based Methodology

Looking upon the different stages for knowledge searching and discovery from data processing can define four different types of data users [1], named as namely Data Provider, Data Collector, Data Miner, Decision Maker [7] as shown in Figure I. now if we differentiate each four data user with their roles we can easily found out the privacy and security issues in data mining at big data by illustrative ways. All of the users possessive about the privacy and security of confidential and sensitive data, but other user's role views the security issue from its own viewpoint.

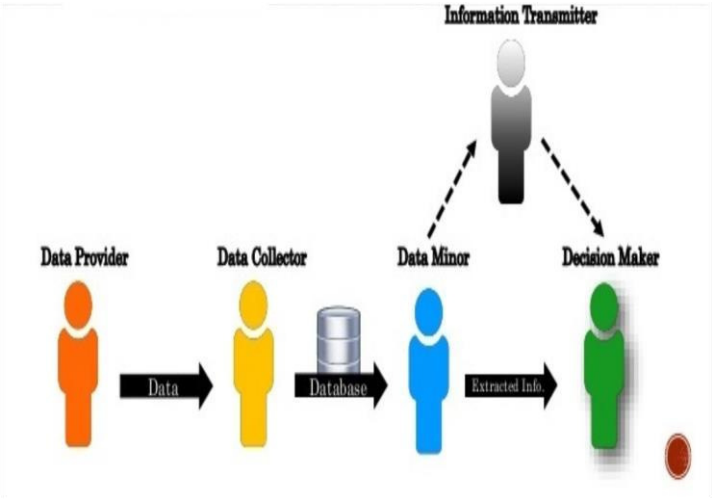

Fig 1. Data Mining in Big Data, Application Functioning

A. Data Provider

Data owner's main concern is to whether to control the confidential and sensitive data, which is provided to others. Where the users owns some data, which is provided for discovery (mining task).

B. Data Collector

The main objective of data collector is to collect data and information from data owners and data provider for publishing purpose for the data miners. Collected data from different data providers may contain some sensitive and confidential data. If that sensitive and confidential data is directly provided to data miners, then the privacy and confidentiality of sensitive data may violates.

C. Data Miner

Data miner, mining to the provided data from data collector from various sources, using different data mining algorithms.

D. Decision Maker

Decision maker can find big data mining result from data miners or from some information transmitter as shown in figure I. information transmitter may harm the data by changing it intentionally or unintentionally, by which it is serious loss to the decision makers.

These different users' roles having their own security and privacy concerns.so we need to focuses upon the data collector phase for privacy concerns. Because if data can be disclosed if the data collector did not take too much security concerns before mining the data to public.

III. RELATED WORK

For privacy of sensitive and confidential data, there have been introduced numbers of techniques implemented and suggested. Mylar can support wide range of applications, whereas others cannot. It can compute encrypted data at the server side and support for processing sensitive data securely for shared data [8].

Each record in storage is identical with each other with at least k-1 records using k-anonymity [9]. In k-anonymity method, confidentiality of sensitive data cannot be achieved, if the confidential and sensitive data having equivalent values in same class.

If there is an equivalence class in a database with the ℓ-diversity [10], if there having at least ℓ well-formed defined values for the user's sensitive and confidential attribute.

Wang [11] proposed a security model, (α, k)-Anonymity model, represents a view of the database table, which is defined as an (α, k) anonymization, where the k-anonymity and α-disassociation properties with respect to the quasi-identifier for modification of the table.

In the t-closeness techniques, if there is distance between distributed classes attributes for sensitive data and distributed of attribute of sensitive data in the global database, where the complete table and database table having no more than a threshold t [12], then that class is defined as t-closeness. The outcome of this method is to preserve the confidentiality of sensitive data with respect to homogeneity and background knowledge attacks.

Web platforms are designed for security purpose, before Mylar is designed. As well as keyword searching is normal and common techniques in web applications. It is often nonpractice to execute because it required too much time to take run client systems due to storing massive amount of data. There are numbers of cryptographically techniques for searching keywords, and they required to encrypt sensitive data using a single unique key [13].

Due to only single key encryption, it is not successful for applying these techniques to the web applications where numbers of users can access the data [14]. There are several numbers of web platforms which support different browsers , on which data can be encrypted before uploading to the server and also can be decrypted before granted to the users [15]. Where the encryption key is stored in the hash fragment of the websites or entered by the users and there the encryption keys and data can be accessed by JavaScript code from the web pages [16]. So the conclusion of this method, the adversary of the java Script code can send to client, where the encryption keys may be able to leak during processing.

A new authentication approach is defined by SUNDR [17] where a specialize protocol is proposed, which helps to

authorized user to identify the updating or alteration of data, which is tried by an unauthorized user in the internet. This technique can define the integrity and consistency of data stored in the unprotected servers.

SPORC [18] and Depot [19] improves the techniques of SUNDR. In this technique, the applications can serve over the encrypted servers. Where the proxy system does not give permission to perform any computation on server side, for example Mylar can define server-side keyword search. In this technique the SPORC, determine the application at run time, when the user visits the server for processing the data.

Now data confidentiality is more interesting to prevent from the threats. Then Crypt DB [20] introduced as new technique, which provide privacy of sensitive data confidentiality from threat over executing SQL queries to the encrypted data in the servers. As well as Crypt DB can perform security from the attacks on the server, even if it does not give guarantees to the users while the user log in to the server is in attack. If there are different keys for the encrypted data, then searching using keywords, Crypt DB does not support computation.

New approach is described in Shadow Crypt [21], in which users can switch transparently to encrypted data for text based web applications. It is to functioning for securing against the potentially malicious and web applications.
This techniques focuses to the privacy of data which is stored into servers is encrypted by key k is only visible to the principals with knowledge of the key k that means the authorized data owner or user. It is having a web browser extension, which replace input data with secure and isolated shadow data and encrypt text with ShadowCrypt. ShadowCrypt does not provide protection against DoS (denial-of-service) attacks by the application.

TABLE I. COMPARING PRIVACY PRESERVATION METHODS

S. No.	***Author***	***Proposed Technique***	***Drawbacks***
1.	Yun Pan et al.[9]	k-anonymity	Not prevent Attribute leakage attack
2.	Ninghui Li et al.[12]	*t*-closeness	Does not preserving the privacy against identity disclosure Attack
3.	Benjamin et al.[10]	ℓ-diversity	The privacy against skewness and similarity attacks can not be preserved
4.	Qiang Wang et al.[11]	(α,k)-Anonymity model	identity disclosure attack does not addressed
5.	A. J. Feldman et al.[18]	SPORC	Does not allow server side computation
6.	Raluca Ada Popa et al.[20]	Crypt DB	if different keys used for data encryption, it does not handle the request
7.	Warren He et al.[21]	Shadow Crypt	denial-of service Attacks by the Application does not resolved

IV. PROPOSED WORK

In Big Data, environment important task is to provide security and privacy of user's sensitive data with confidentiality, even the data is sharing and growing publically over network. For providing, the security against the leakage of the data following proposal is to be considered, which may prevent from the attackers.

A. Problem Statement

Big Data is the term, which refers to the very large amounts of heterogeneous data and information. Big data concept came from the growing and generating a great number of data stored from the different sources, and the internet. Because of these, large amount of data, security and privacy of sensitive data is crucial task.

Today is computing sensitive and confidential data may leak. Now days many of the platform stores the sensitive data on untrusted servers, from where the sensitive data and confidentiality of user may leak.

Upon the conclusion the previous approaches, there we can find two challenges in combating these threats.

In threat 1, an intrusive DBA (database administrator) having the full administrative access to DBMS server may reveal the sensitive and confidential data, for that we proposed a model which may prevent such DBA to gain full access to DBMS server for sensitive and confidential data.

In threat 2, an attacker and intruders can looks after for the full control over the software and hardware and hardware of applications and proxy, DBMS servers, for which we

proposed a model by which we can prevent from such attackers to get access to user's sensitive data.

B. Proposedl Security Model

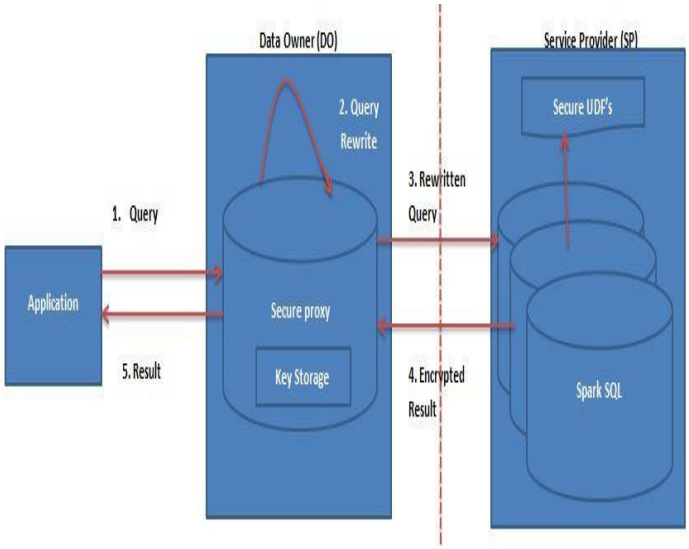

Fig 1. Proposed Security Model Architecture

Our proposed model architecture is implemented as a layer of software, which runs upon the top of Apache Spark SQL architecture. Apache Spark having main features like to manage fast and general-purpose engines for large amount of data processing, where memory primitives can enhance cluster computing. Here we use Apache Hadoop, which provide facility for cluster manager and provide massive amount of information spark's cluster computation. In addition, we use Apache Hive for large amount of data as data warehousing.

For DBMS servers we are using Spark SQL, which supports for structured data and SQL queries as a tool. It also supports for the different users defined functions, by which the user can extend the functionality of cluster database server by adding cryptographically methods, which can be computer in Spark SQL. Using this method the secure server is computed to spark computation engine and performed as user-defined functions.

There is a secure Proxy placed in the client side machine and all the secret keys are stored in the key storage. In the proposed model, there exist five stages of a query processing in this architecture:

a) A query is submitted through client's application to secure proxy.
b) Secure proxy's parses analyses, analyze the inputted query and rewrite the inputted query in terms of user-defined functions.
c) After rewritten query generates, the secure proxy submits the rewritten query for computation to Spark SQL
d) After processing the encrypted query in the spark SQL, the query result is returned to the secure proxy.
e) Secure proxy gets the result in encrypted form, and then it will be decrypted and send it to the user or applications.

V. CONCLUSION

For the above proposed model processing and computing with encrypted data, it will be one of the primary strategies for securing sensitive and confidential data of users in public. It stores confidential, sensitive data in the form of encrypted to the server, and it decrypts that sensitive data only in the users system. Our approach will find the solutions for the given problem statements as given two threats.

In conclusion, we have to define comprehensive description of query rewriting that supports multiple secure operations with data co-operations. In addition, we have to analysis the performance of the system, which is both practically efficient and secure for sensitive data.

References

[1] Agrawal R., Srikant R.,``Privacy Preserving Data Mining.,"In the Proceedings of the ACM SIGMOD Conference,2000.

[2] P.Kamakshi,"Survey On Big Data and Related Privacy Issues", IJRET, 2014.

[3] Hirsch, Dennis D. "The Glass House Effect: Big Data, the New Oil, and the Power of Analogy" , Maine Law Review 66 (2014).

[4] Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: Issues, challenges, tools and Good practices." In Contemporary Computing (IC3), 2013 Sixth International Conference on, pp. 404-409. IEEE, 2013.

[5] Salini . S, Sreetha . V. Kumar, Neevan .R, "Survey on Data Privacy in Big Data with K-Anonymity ", Volume 2,International Journal of Innovative Research in Computer and Communication Engineering, Issue 5, May 2015.

[6] Krishna Mohan Pd Shrivastva1, M A Rizvi, Shailendra Singh, "Big Data Privacy Based On Differential Privacy a Hope for Big Data", 2014, IEEE.

[7] Lei Xu, Chunxiao Jiang, (Member, IEEE), Jian Wang, (Member, IEEE), Jian Yuan, (Member, IEEE), and Yong ren, (Member, IEEE), "Information Security in Big Data: Privacy and Data Mining", Volume 2, IEEE, October 20, 2014.

[8] Raluca Ada Popa, Emily Stark, Jonas Helfer, Steven Valdez, Nickolai Zeldovich, M. Frans Kaashoek, and Hari Balakrishnan MIT CSAIL and Meteor Development Group." Building web applications on top of encrypted data using Mylar .

[9] Yun Pan, Xiao-ling Zhu, Ting-gui Chen," Research on Privacy Preserving on K-anonymity", Jurnal of software, 2012.

[10] Benjamin C.M, Fung, Ke Wang, Ada Wai-Chee Fu and Philip S. Yu, "Introduction to Privacy-Preserving Data Publishing Concepts and techniques", ISBN:978-1-4200- 9148-9,2010.

[11] Qiang Wang, Zhiwei Xu and Shengzhi Qu, "An Enhanced KAnonymity Model against Homogeneity Attack", Journal of software,2011, Vol. 6, No.10, October 2011;1945-1952.

[12] Ninghui Li, Tiancheng Li, Suresh Vengakatasubramaniam,"t- Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity", International Conference on Data Engineering, 2007, pp106- 115.

[13] A. Arasu, S. Blanas, K. Eguro, R. Kaushik, D. Kossmann, R. Ramamurthy, and R. Venkatesa,." Orthogonal security with

Cipherbase", In Proceedings of the6th Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, Jan. 2013.

[14] S. Bajaj and R. Sion."TrustedDB: a trusted hardware based database with privacy and data confidentiality", In Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pages 205–216, Athens, Greece, June 2011

[15] G. Ateniese, K. Fu, M. Green, and S. Hohenberger, "Improved proxy re-encryption schemes with applications to secure distributed storage". In Proceedingsof the 13th Annual Network and Distributed SystemSecurity Symposium, San Diego, CA, Feb. 2006.

[16] D. Akhawe, P. Saxena, and D. Song, "Privilege separation in HTML5 applications". In Proceedings ofthe 21st Usenix Security Symposium, Bellevue, WA, Aug. 2012.

[17] J. Li, M. Krohn, D. Mazieres, and D. Shasha, "Secure untrusted data repository (SUNDR)". In Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI), pages 91–106, San Francisco, CA, Dec. 2004.

[18] A. J. Feldman, W. P. Zeller, M. J. Freedman, and E. W. Felten," SPORC: Group collaboration using untrusted cloud resources". In Proceedings of the 9th Symposium on Operating Systems Design and Implementation (OSDI), Vancouver, Canada, Oct.2010.

[19] P. Mahajan, S. Setty, S. Lee, A. Clement, L. Alvisi, M. Dahlin, and M. Walfish, "Depot: Cloud storage with minimal trust". In Proceedings of the 9th Symposium on Operating Systems Design and Implementation (OSDI), Vancouver, Canada, Oct. 2010

[20] Raluca Ada Popa, Catherine M. S. Redfield, Nickolai Zeldovich, and Hari Balakrishnan, "CryptDB: Protecting Confidentiality with Encrypted Query Processing", ACM, 2011.

[21] Warren He, Devdatta Akhawe, Sumeet Jain,"ShadowCrypt: Encrypted Web Applications for Everyone", ACM, November 2014.