Big Data – Feasible Solutions for Data Privacy Challenges

Miti Jhaveri
Student, Computer Engineering Department
Dwarkadas J. Sanghvi College of Engineering

Resham Gala
Student, Computer Engineering Department
Dwarkadas J. Sanghvi College of Engineering

Khushali Deulkar
Professor, Computer Engineering Department
Dwarkadas J. Sanghvi College of Engineering

ABSTRACT

In this paper, we aim to introduce the challenges that are faced to maintain the authentication and integrity in big data. Big data was developed to provide voluminous data while protecting the integrity of data. We further provide a comparative study on various methods which can help resolve the following issues. The comparative study is based upon a few parameters such as the scope, its prerequisites, advantages and disadvantages of the method.

General Terms

Security, Privacy preservation, comparative study

Keywords

Big Data, Privacy Preservation, Data Privacy, Data Mining, Security Solutions.

1. INTRODUCTION

With the emergence of the voluminous amount of data in the recent past, there has been an increased potential for mining this data to obtain fruitful information, leading to the dawn of the term 'big data' [1]. This data is structured, unstructured and semi-structured depending on its source. Also, it contains a lot of variety, complexity, variability which can ace the process of research and improvisation in current technology.

Data is released by a lot of sources for business analysis, maintaining anonymity. Although, in a few cases it is possible to identify the user to whom the data belongs. This may be possible due to a malicious attacker or an untrusted person who is working in association. In order to preserve the privacy of the user and collect relevant data it is necessary to implement privacy-preserving algorithms.

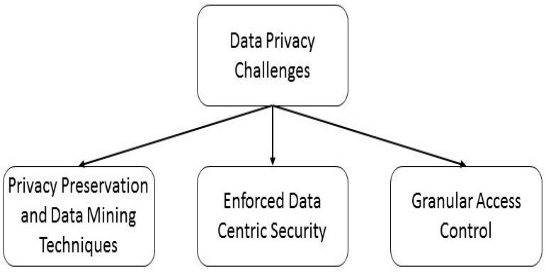
Data Privacy Challenges
Privacy Preservation and Data Mining Techniques
Enforced Data Centric Security
Granular Access Control

Fig 1: Data Privacy Challenges

There are three major issues [2] that we cover in our comparative survey as shown in Fig 1. They are:

- Privacy Preservation and Data Mining Techniques
- Enforced Data Centric Security
- Granular Access Control

2. SOLUTIONS FOR THE MAJOR ISSUES

2.1 Privacy Preservation and Data Mining Technique

In today's world data is being generated at an extremely rapid rate. Various sources like social sites, banks, government records, company records and so on, contribute towards burgeoning growth of data. All this data generated is of useless if it cannot be processed properly. Data mining [3] is a technique through which data can be processed intelligently. The discovered knowledge can be applied to decision making, in process control, managing information, and in query processing. Therefore, data mining helps one in uncovering relations among various clusters of data.

However, if the information goes into wrong hands then the outcome can be disastrous. Cybercrime and forged theft are some of the prime examples of misuse of mined data. Thus privacy of data and personal information is a major concern for all organizations dealing with big data as its source. Privacy-preserving [4] data mining techniques (PPDM) [5,6] deal with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. Some of the commonly used privacy preservation techniques in data mining are: Randomization, K-Anonymity [7] and L-Diversity [8].

2.1.1 Randomization

Randomization is a process of adding high disturbance or noise to the data, in order to mask the actual values of the records. The noise added is high, so that the individual records cannot be recovered. Data mining techniques, therefore, work with the aggregated distributions of these perturbed records. This shows that randomization plays an important role in preservation of sensitive data. However, the distributions are reconstructed independently, due to which data mining algorithms work under an implicit assumption that all records are independent. But most of the data mining algorithms have a lot of relevant information hidden in the inter-attribute relation. This is one of the drawbacks of randomization approach.

2.1.2 K-Anonymity

Publicly available records can lead to privacy getting heavily compromised. Even if the key identifiers such as name, security number are removed from the records, pseudo-identifiers such as age, sex and zip-code can be used to accurately determine the records. This technique of K-Anonymity reduces the granularity of these pseudo identifiers with the help of generalization and suppression. In generalization, the range of the attribute values is reduced to prevent recognition, for example the range of birth date can be reduced to years only. In method of suppression the value of attribute is removed completely. In k-anonymity approach, we use domain generalization hierarchies of quasi identifiers in order to build k-anonymous table. Quasi identifiers are attributes available to the adversary. Thus, the privacy of sensitive data is preserved, but if the background information is available to the attacker then this approach is susceptible to many kinds of threats.

2.1.3 L-Diversity

K-anonymity approach prevents effective identification of the record but it may not always be effective in inferring the sensitive values of the attributes in that record. Therefore, l-diversity approach was proposed, which not only maintains minimum group size of k, the size of the table, but also focuses on maintaining the diversity of sensitive attributes. A block is said to be l-diverse if it has 'l' well represented values for the sensitive attribute S. If every block (a set of tuples such that its non-sensitive value is generalized) is l-diverse then its table is said to be l-diverse. Thus this approach of l-diversity helps in maintaining the diversity of sensitive data along with its privacy.

2.1.4 Comparison between Different Privacy Preservation Techniques

Table 1: Comparison between Randomization, K-anonymity and L-diversity Techniques

Name	Approach	Advantages	Disadvantages	Pre-requisites	Scope of improvement
Randomization	• Preserved by the introduction of high random noise in the data, and mining is applied on the aggregated distributions.	• Method is relatively simple. • Can be implemented at the time of data collection. • Does not require the use of trusted server.	• Outlier records are susceptible to adversarial attack. • Addition of noise reduces data mining utility.	• No prior knowledge of records is required, as the noise added is independent of the behavior of other records.	• Multiplicative perturbation can be used to preserve the privacy of data, and it can also be used for distributed privacy preserving data mining. • Data swapping can also be used to randomize data which works effectively when applied with k-anonymity framework.
K-Anonymity	• Quasi identifiers are used to generalize the attribute values of the records, thereby misleading the attacker.	• Numerous ways are available for anonymization. • Blocks the linking attack.	• K-anonymity is susceptible to Homogeneity and Background knowledge attack.	• Set of pseudo identifiers for every record. • Behavior of locality of each record.	• Micro-aggregation can be used as a transformation technique of k-anonymity approach. • Approximation algorithms can be used to search over a space of possible multi-dimensional solutions.
L-Diversity	• This method focuses on maintaining the diversity of the sensitive attributes.	• Overcomes the homogeneity and background knowledge attack.	• Similarity of sensitive attribute is vulnerable. • Difficult to create feasible l-diverse representation.	• Prior knowledge of multiple sensitive values is useful for creating l-diverse tables.	• T-closeness model is an enhancement of the concept of l-diversity model.

2.2 Enforced Data Centric Security

There are two basic types of approaches to protect the integrity of the data and secure it from any individual or organization. The first approach has basic access denials using the operating system’s access controls. The other approach is to encrypt the stream of data by encapsulating it. Both methods have its pros and cons. In case of the first, it is easier for an attacker to invade into the system by various means such as buffer overflow [9]. Although, in case of the second, it is tougher to extract the private and public keys that are used.

2.2.1 Software-Managed Access Control

At application level, various applications have various types of features to provide security of data. Some applications provide a mere password system which can easily be cracked by various deceitful methods. There are a few applications that block only certain features. The other features are open for all users. Depending upon the confidentiality of data, various levels of security are incorporated. For instance, in case of a digital signature, various levels of security must be ensured to prohibit any type of misuse.

2.2.2 Using Encryption

Encryption is one of the most commonly adopted methods to secure sensitive data. This is done by using public and private keys for encryption and decryption. There are various methods proposed to encrypt data and limit access to digital documents. Cryptolope [10], which are known as cryptographic envelopes, enable a commercial platform providing the content creator and the publisher to give the license of their content to the customers by controlling the decryption keys by their distribution. Cryptolope decouples the distribution of the data and its corresponding decryption keys [11]. There are various such solutions that have different type of model.

2.2.3 Secret Protection Architecture

Secret Protection Architecture [12, 13] was used to protect the secret keys or data depending on the mode. The SP architecture contains the Trusted Software Module. This architecture has been developed to provide integrity and confidentiality. There are various components in the hardware of the architecture such as Storage Root Hash and Device Root Key that facilitate the security.

2.2.4 Comparison Between Different Enforced Data Centric Security Techniques

Table 2: Comparison Between Software Managed Access Control, Encryption, Secret Protection Architecture Techniques

Method	Approach	Advantage	Disadvantage	Prerequisites	Scope of Improvement
Software Managed Access Control	• Application based • Various levels of security is enforced.	• One of the easiest methods. • No additional hardware costs	• Lot of room for malicious attacker to invade into the system.	• No prerequisites for this method	• If the level of dependency on operating system is reduced, it can provide better security.
Encryption	• Data is encrypted cryptographically to avoid provide an additional level of security	• Unlike plain text, one will need to further decrypt data in order to obtain it.	• Security of the data is based on the encryption key. If that is lost, effectively data is lost.	• In order to generate the key, the algorithm or software that will facilitate the generation	• The processing to be made simpler and reduce costs.
Secret Protection Architecture	• Uses hardware in order to perform stringent security	• This is one of the most stringent method and toughest for an attacker to breakthrough	• Additional hardware is required	• The resources and provision for the inculcation of the architecture	• The costs of including the entire architecture is high, which can be minimized.

2.3 Granular Access Control

In order to provide security compliance it is necessary to provide identity management and role management. Granular access control is mainly provided by two such methods. Role based access control gives individuals the ability to perform certain tasks such as creation, modification or viewing a file. Attribute based access control uses attributes to control and permit accesses. This is done by using a language called Extensible Access Control Markup Language [14]. There are various restrictions that are put permitting fine-grained access control by using various combinations of user attributes. They provide support to role based access control. The major advantage of the attribute based access control is the fact that the decisions are made at run-time based on the attributes that are combined in order to form fine-grained decisions. Also these decisions are uniform and are applied at organizational level rather than individual level. A mature identity and access management environment is one of the major prerequisites for this method. The drawback of this method is that it cannot handle network latency. Thus, synchronization to the local database maybe required.

3. CONCLUSIONS

This paper provides a general overview about big data and some of the privacy problems relating it. The above three methods have been evaluated on the basis of various parameters. For instance, if one has a simple database without many repeats then randomization technique can be used for privacy preservation. However, if the database is huge and has many repeatable attributes, then one has to use k-anonymity or l-diversity technique. Thus using the above mentioned techniques, one can be prepared for situations like cybercrime or forged theft. One can choose any of the feasible solutions that can be applied to various systems depending upon the requirement of the system and the gravity of the attacks faced.

4. REFERENCES

[1] M. B. Malik, M. A. Ghazi and R. Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[2] Y.Demchenko, P.Membrey, P.Grosso, C. de Laat, "Addressing Big Data Issues in Scientific Data Infrastructure," in First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California,USA.

[3] Cloud Security Alliance, "Expanded Top Ten Big Data Security and Privacy Challenges", April 2013.

[4] J. Han, M. Kamber and J. Pei "Data Mining: Concepts and Techniques" 2006, Morgan Kaufmann.

[5] Agarwal, R. and Shrikant, R. "Privacy Preserving Data Mining", Proceeding of Special Interest Group on Management of Dat, pp. 439-450, 2000.

[6] Jian Wang, Yong Cheng Lou, Yen Zh, Jiajin Le, "A Survey on Privacy Preserving Data Mining", International Workshop on Database Technology and Application pp. 111-114, 2009.

[7] P. Samarati, L. Sweeney, "Generalizing Data to Provide Anonymity When Disclosing Information" (Abstract), Proceeding of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, IEEE press, Seattle, June 1998, pp. 188.

[8] A. Machanavajjhala, J. Gehrke, D. Kifer, "L-Diversity: Privacy beyond k-anonymity", Proceeding of the ICDE, Atlanta, Apr. 2006, pp. 24-35.

[9] Ulrich Kohl, Jeffrey Lotspiech, and Stefan Nusser, “Security for the Digital Library – Protectiong Documents Rather Than Channels” in Proceedings of the 9th International Workshop on Database and Expert Systems Applications.

[10] Zhiyuan, An; Haiyan, Liu "Information Technology and Applications (IFITA), 2010 International Forum" on, Issue Date: 16-18 July 2010.

[11] Yu-Yuan Chen, “Architecture for Data-Centric Security”, PhD Thesis, Electrical Engineering Department, Princeton, NJ, Princeton University, pp. 130, 2012.

[12] Jeffrey S. Dwoskin and Ruby B. Lee. Hardware-Rooted Trust for Secure Key Management and Transient Trust. In Proceedings of the 14th ACM Conference on Computer and Communications Security.

[13] Ruby B. Lee, Peter C. S. Kwan, John P. McGregor, Jeffrey Dwoskin, and Zhenghong Wang. Architecture for Protecting Critical Secrets in Microprocessors. In Proceedings of the 32nd Annual International Symposium on Computer Architecture.

[14] OASIS "eXtensible Access Control Markup Language (XACML) Version 2.0, February 2005.