American-Eurasian Journal of Scientific Research 11 (2): 93-97, 2016
ISSN 1818-6785

DOI: 10.5829/idosi.aejsr.2016.11.2.22734

Analysis of Bigdata Security

D. Aruna Kumari and L. Anusha

Department of electronics and computer engineering,
Kl University, Vaddeswaram, Green fields, Guntur Dst., AP., India

Abstract: Big data has the large data and it is an ongoing technology which is used by every organization. It has flexibility to store the structured, semi-structured and structured data. It is faster to store the data than we use before. The data can be generated by connecting devices from pc's and smart phones to sensors. Big data can collect heterogeneous data; it can be in any format like text, document, image, video etc. Since it has a large amount of data, security is fundamental right it should contain. The security should provide to every individual, organization and society data. Without privacy safety, diversity, innovation… etc will be in risk. New technologies may known to other countries which are implemented by own. Like these many more problems effect to country. So to overcome to these problems we are using Big data security. In this paper we present some of the technologies to provide security to big data.

Key words: Big data · Security · Flexibility · Technologies

INTRODUCTION

In olden days, we store the data in books, later we have gone to maintain small amounts of data in computers by using files. In most enterprise scenarios the volume of data is too big and too faster as it exceeds current processing technologies. So the new technology to store the data is "BIG DATA". Big data can be in petabytes (1024 terabytes) or exabytes (1024 petabytes) of data. It can consist billions to trillions of records of millions of people all from different sources (e.g. web, sales, social media, mobile data etc). The data is typically loosely structured data that in often incomplete and inaccessible. The term big data indicates the large volumes of data of any type being generated at high velocity. Big data is generated from the sources like online transactions, quires, mobile phones, emails, videos etc. It can be stored by partitioning along various servers. By this we can say that big data sources and storage systems are scattered all around internet. Since it web model security is the major challenge in big data. If the security is not ensured, it may lead to great data lose for the people and organizations. Exposing whole big data may yield good results but at same time it has security problems. The data hidden in big data is more valuable to hackers and invaders. So, there is tradeoff between big data availability and big data security and should have balance between them. This paper provides various methods to protect the knowledge hidden in big data during whole analytics process. Normal security called traditional security mechanisms fail to handle big data because it has large volume and velocity. Security is the main issue because big data generally compromises of person specific information. Big data information can be stored, linked on web. If there is unsecure data personal identifiable information, this causes customers to lose faith in organizations.

Description

Types of Big Data: There are three types of big data:

- Streming data 2.social media data 3.publivly available sources

Streaming Data: This data depends on the internet. The data reaches to the system from a web of connected devices. is type of data is helpful to the organizations, we can analyze which data is arriving and which data should keep and which data should not. This data can be used for further analysis.

Corresponding Author: D. Aruna Kumari, Department of Electronics and Computer Engineering, Kl University, Vaddeswaram, Green Fields, Guntur Dst., AP., India.

Social Media Data: The data will have the social interactions which are attractive now a days. Particularly for marketing, sales etc, even the data may be unstructured or semi-structured forms. It poses a unique challenge when consuming and analyzing the information like company is marketing to mobile and social customers.

Publicly Available Sources: This data is open source like European Union Open Data portal.

Purpose of Security: The purpose of security for data is the data should be left alone, the selected information can share, we can intimate personal decisions without government interference, we can protect the personal characteristics, in social network we can have private communication, tracking, stalking by product of locational tracking, false conclusions about individual byproduct of group and sometimes personal profiles from big-data analytics.

Process to analyze the big data:

Fig. 2.0:

First we have to discover the data available and understand the data. By implementing the solution we have to integrate the data. After integrating we have to secure the integrated data by locking. After completing these procedures we have to monitor the data.

Security Techniques: There are many techniques to secure the data.

Application Software Security: To secure the data we have to use secure versions of open-source software. For example using technologies like Apache Accumulo or 20.20.x version of Hadoop can help to secure the data because these are secure versions. The other technologies like Cloudera, Sentry or Data Stax are enhanced security at the application layer.

Maintenance, Monitoring and Analysis of Audit Logs: In this we have to implement audit logging technologies to understand and monitor big data. The security engineers in the organization need to be tasked with examining and monitoring the files. It is important to ensure that auditing, maintaining and analyzing logs are done consistently across the enterprise.

Secure Configurations for Hardware and Software: The servers will be based on secure images for all systems in big data architecture. The patching is up to date on these machines and that administrative privileges are limited to small number of users. Organizations should use frameworks, like puppet, to automate system configuration and ensure that all big data servers in the enterprise are uniform and secure.

Account Monitoring and Control: Organizations should manage the accounts for big data users. Organizations should require strong passwords, deactive inactive accounts and impose a maximum number of failed login attempts to help stop attacks from getting access to a cluster. Monitoring account access can help reduce the probability of a successful compromise from the inside. Cyber criminals are never going to stop being on the offensive and with such a big target to protect, it is prudent for any organization utilizing big data technologies to be proactive as possible in securing its data.

Proposed Solutions: There are some of the existing solutions for data security Cryptography is one of the existing solutions for protecting the data. It consists set of techniques and algorithms. In cryptography plaintext is converted into cipher text using various encryption schemes. There are various methods based on this scheme like public key cryptography, digital signature etc. But big data alone cannot implement the security by common cloud computing and big data services. Because big data differs from traditional data on the basis of velocity, variety, volume, big data architecture is different from traditional information. These changes in the architecture and its complex nature make cryptography and traditional encryption schemes not scalable up to the security needs of big data. So the cryptography is not useful to secure big data. It makes data inaccessible to those who don't have access to decryption key. Data can be stolen or misuse the data. Attribute based encryption can also be used for big data security. This method of

securing big data is based on relationships among attributes present in big data. The attributes that need to be protected are identified based on type of big data and company policies. Encryption or cryptography alone can't stand as big data security preservation method. They can help us to do data anonymization but cannot be used directly for big data privacy.

Existing Solutions: Big data security has raised serious concerns, so there are some of efficient security preservation methods. So we have discussed the methods like data anonymization, notice and consent and differential privacy.
We can implement based on the table.

Data Anonymization: Data anonymization is the process of changing data that will be used that prevents the identification of key information. It is also sometimes referred as data de-identification. Anonymization in this case generally refers to hiding identifier attributes (attributes that uniquely identify individuals) like full name, license number, voter id etc. The main problem with data anonymization is that data may look anonymous but re-identification can be done easily by linking it to other external data it is shown that re-identification of anonymous medical records can be done using external voter list data. The attributes like gender, date of birth, zip code that can be combined with external data to re-identify individuals are called quasi identifier attributes.

Table 1 represents the data set that needs to be analyzed for obtaining income trends without disclosing individual identity.

Table 2 represents data made anonymous by removing identifier attribute Voter ID. This table may look anonymous but can be linked with external data of to re-identify individuals.

K-Anonymity: A dataset is called k-anonymized if for any tuple with given attributes in the dataset there are at least k-1 other records that match those attributes. K-anonymity can be achieved by using suppression and generalization. In suppression, quasi identifiers are replaced or obscured by some constant values like 0,* etc. In generalization, quasi identifiers are replaced by more general values from levels up the hierarchy.

For example, in Table 1, Voter id and name are identifier attributes. Age, DOB, City are quasi identifiers. Income is a sensitive attribute.

Table 1: Base

Age	Sex	City	Income
24	M	Delhi	1,00,000
24	M	Gurgaon	18,000
24	M	Gurgaon	25,500
24	M	Delhi	12,000
26	F	Delhi	20,000
26	F	Delhi	50,000
26	M	Delhi	29,000
26	F	Delhi	48,000
32	M	Delhi	26,000
32	F	Gurgaon	45,000
32	F	Gurgaon	34,000
32	M	Delhi	34,000

Table 2:

Voter ID	Age	Sex	City	Income
	24	M	Delhi	1,00,000
	24	M	Gurgaon	18,000
	24	M	Gurgaon	25,500
	24	M	Delhi	12,000
	26	F	Delhi	20,000
	26	F	Delhi	50,000
	26	M	Delhi	29,000
	26	F	Delhi	48,000
	32	M	Delhi	26,000
	32	F	Gurgaon	45,000
	32	F	Gurgaon	34,000
	32	M	Delhi	34,000

Table 2:

Age	Sex	City	Income
2*	M	Delhi	1,00,000
2*	M	Gurgaon	18,000
2*	M	Gurgaon	25,500
2*	M	Delhi	12,000
2*	F	Delhi	20,000
2*	F	Delhi	50,000
2*	M	Delhi	29,000
2*	F	Delhi	48,000
3*	M	Delhi	26,000
3*	F	Gurgaon	45,000
3*	F	Gurgaon	34,000
3*	M	Delhi	34,000

K-anonymous data can still be vulnerable to attacks like unsorted matching attack, temporal attack and complementary release attack. Therefore we move towards L-diversity method of data anonymization.

L-Diversity: L-diversity technique of data anonymization tries to bring diversity in the sensitive attribute of data. It ensures that each equivalence class of quasi identifiers has at least L different values of sensitive attribute.

Table 3:

Age	Sex	City	Income
24	Person	ncr	1,00,000
24	Person	ncr	18,000
24	Person	ncr	25,500
24	Person	ncr	12,000
26	Person	ncr	20,000
26	Person	ncr	50,000
26	Person	ncr	29,000
26	Person	ncr	48,000
32	Person	ncr	26,000
32	Person	ncr	45,000
32	Person	ncr	34,000
32	Person	ncr	34,000

In Table 1 income is a sensitive attribute. For data to be L-diverse there should be at least L different values of income associated with each equivalence class. Table 4 shows 3-diverse version of table 1 since each equivalence class has atleast 3 different values for sensitive attribute income.

The problem with this method is that it depends upon the range of sensitive attribute. If we want to make data L diverse whereas sensitive attribute has less than L different values, fictitious data is to be inserted. This fictitious data will enhance the security but may result in problems during analysis. Also L-diversity method is prone to skewness and similarity attack and thus can't prevent attribute disclosure [1-3].

T – Closeness: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. The main advantage of t-closeness is that it prevents attribute disclosure.

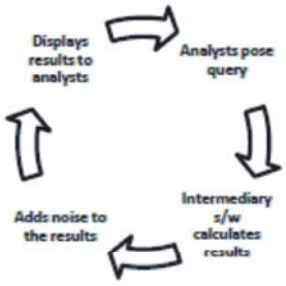
Data anonymization can be applied to big data but the problem lies in the fact that as size and variety of data increases, the chances of re-identification also increase. Thus, anonymization has a limited potential in the field of big data privacy.

Notice and Consent: The most common privacy preservation method for web services is notice and consent. Every time an individual accesses a new application or service, a notice stating privacy concerns is displayed. The consumer needs to consent the notice before3 using the service. This method empowers an individual to ensure his privacy rights. It puts the burden of privacy preservation on the individual.

When applied to big data, this method poses numerous challenges. In most of the cases uses of big data are unexpected or unknown at the time when notice and consent is given. This requires the notice to change every time big data is used for a different purpose. Also big data is collected and processed so rapidly that it creates burden on consumers to consent the notice. A method by which notice and consent can be modified for big data is the use of third parties offering a choice of different privacy profiles [4, 5].

Differential Privacy: Differential Privacy is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protection. It aims to minimize the chances of individual identification while querying the data. The method of differential privacy.

As opposed to anonymization, data is not modified in differential privacy. Users don't have direct access to the database. There is an interface that calculates the results and adds desired inaccuracies. It acts as a firewall. These inaccuracies are large enough that they protect privacy, but small enough that the answers provided to analysts and researchers are still useful.



The advantages of differential privacy over anonymization are:

- The original data set is not modified at all. There is no need for suppression or generalization.
- Distortion is added to the results by mathematical calculations based on the type of data, type of questions etc.
- The distortion is added in such a way that value hidden is useful to analysts.

CONCLUSIONS

Big data privacy has become an important issue since it is directly related to customers. It is now essential for an organization to promise privacy in big data analytics. Privacy measures should now focus on the uses of data rather than collection of data. They should be modified with respect to the size and unexpected uses of big data.

Techniques like anonymization have limited potential when applied to big data. Notice and consent method also burdens the customer for ensuring privacy. Differential privacy may be seen as a viable solution for big data privacy. One problem with this method is that analyst should know the query before using the differential privacy model. When modified and applied to big data, it may ensure privacy without actually modifying the data.

REFERENCES

1. Belloni, Alexandre, D. Chen, Victor Chernozhukov and Christian Hansen, 2012. Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. Econometrica, 80(6): 2369-429.
2. https://www.mwrinfosecurity.com/articles/big-data-security---challenges-solutions/
3. https://crypto.stanford.edu/craig/craig-thesis.pdf
4. http:// www.vormetric.com/ data-security-solutions/ use-cases/ big-data-security
5. https://epic.org/privacy/big-data/
6. https:// www.whitehouse.gov/ sites/ default/ files/ microsites/ ostp/ PCAST/ pcast_big_data_and_privacy_-_may_2014.pdf
7. https://www.drugabuse.gov/sites/default/files/ispab_jun2014_big-data-privacy_blumenthal.pdf
8. http://plato.stanford.edu/entries/it-privacy/
9. Hindman, B., A. Konwinski, M. Zaharia, A. Ghodsi, A.D. Joseph, R. Katz, S. Shenker and I. Stoica, 2011. Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center, NSDI 2011, March 2011.
10. Bliss, N., R. Bond, H. Kim, A. Reuther and J. Kepner, 2006. Interactive grid computing at Lincoln Laboratory, Lincoln Laboratory Journal, 16(1).
11. Kepner, J., *et al*., 2012. Dynamic distributed dimensional data model (D4M) database and computation system, 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, March 2012.
12. FUSE http://fuse.sourceforge.net/