A Survey on Big Data and Information Security in It

Sandeep Kumar Yadav
Department of Information Technology
Rajkiya Engineering College, Banda [U.P], India -210201

Abstract— Big data is a term encircling different types of complicated and large datasets that is hard to process with the conservative data processing systems. With the development of information technology big data application prompts the development of storage, network and computer field. It also brings new security problems. These security challenges caused by Big data has attracted the attention of information security. Big data analysis is now used in almost every phase of our society, communication services, banking and research. Big data utilizes huge quantity of data that may be available in the cloud and it may require data processing distributed across many servers. This paper summarizes the characteristics of big data and information securities in it. We also present the future orientation of information security issues and their feasible countermeasures.

Key words: Big Data, Information Security Threats, Data Privacy, Cloud Security

I. Introduction

It is whispered that the term big data started with companies handling web search applications and looked-for queries on large disseminated collection of data. As the data size is huge, it is very difficult to use traditional database and software techniques to develop it. The development of the current big data is still faced with many problems especially security and privacy protection[1]. On the Internet People's behavior are known by Internet merchants[2]. It is not necessary that Big Data only refers to extremely huge data and tools and measures used to process and investigate them, but it also gives directions for new ideas and challenges in research area[3]. Currently many organizations realize the big data security issues and actively take actions on big data information security problems. In 2011, CSA formed a working group on big data to find solutions for data security and privacy issues. In this paper, we present a survey of big data in terms of information security threats and their remediation. We have also focused on information security threats.

A. *Major challenges in Big Data:*

1) Confidential and secure transaction of data from one to another via channel (networking).
2) Conscious or malicious leakage of data.
3) Conscious or malicious leakage of data

II. Features of Big Data

There are many features associated with big data. The excellent aspects are volume, variety, velocity, variability and value.

Volume: Big data uses huge datasets[4] which include data internet searches, online purchases and transactions, social media interactions, mobile information, data from sensors in vehicles and other devices. The amount of big data may be petabytes or exabytes. It is also possible to hold very large datasets, due to the reducing price of storage and the accessibility of cloud-based services. As these datasets are very large, they cannot be analyzed using conventional techniques like spreadsheets or SQL queries. New tools like NoSQL and open source software Hadoop have been developed to analyse big data.

Variety: Big data often necessitate collection of data from heterogeneous sources. Presently it seems that big data analytics primarily employs structured data like tables with defined fields as well as unstructured data. For example, the data is collected from various sources like social media source like twitter, on line products purchases and the comments related to products etc. merging data from diverse sources in this way presents various challenges with respect to IT perspective. Practitioners analyzed and suggested that of the ‘three Vs’, variety is the most significant characteristic of big data. This view propose that, when a company is analyzing its own customer database which is very large, may not essentially publish any innovative ideas in terms of either analytics or data protection. On the other hand, when it joins its own information with the data extracted from various sources, then it will give results that are qualitatively different.

Velocity: In some circumstances like in real time it is essential to analyze data as fast as possible. Big data analysis can be employed to analyze the static data like database of a store as well as the data which is time varying and continuously created or documented like online purchases and credit card payments.

Variability: It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

Value: All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed. Moreover, users can also query the data store to find business trends and accordingly they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

III. Classes Used in Big Data

Categorization of big data falls with major aspects, since this technology involves with multiple diversified fields and ir-related types of information handling. Some of the classes can be framed like storing, sourcing, formatting, staging and processing[5]. Each class instantiates many entities through which the actions carried out.

Sourcing: Data sources identified are Internet Web Pages, Discussion Forums, Chats and message shared in and among social networks, Remote Sensing Networks, All kinds of day to day transactions done through internet based applications.

Formats: Unstructured, partially structured, and structured.

Storing: Image based, Graph based, documents, Key Value Stores (Key Values Store is a way of storing application's data with null schema. It doesn't require a static data model. Unique keys are used to represent values stored in it.)
Organize: Extract, Clean, normalize, transform, Load
Process: Online, Offline
Query: On demand, ad-hoc

IV. THREATS OF BIG DATA SECURITY

Today, big data has penetrated into various industries, and has become a kind of production factor which plays an important role. In the future it would be the highest point of the competition. With the development of rapid processing and analysis technology, the possible information it contained can quickly capture the valuable information in order to provide reference for decision making. However, as big data setting off a wave of productivity and consumer surplus, the challenge of information security is coming either.

A. *Data Acquirement:*

The source of big data is diversity. Therefore, the first step to process big data is to collect data from source and pre process, in order to provide uniform high quality data set to the subsequent process. As a result, due to the inundation of data acquisition, large data become more likely to be "discovered" as a sensitive target, and be more and more attention. On one hand, big data not only means the huge amounts of data, but also means more complex and more sensitive data. These data would attract more potential attackers, and become a more attractive target. On the other hand, with data assembled, the hacker could get more data in one successful attack, and reduce hacker's attack costs.

The discretion of information refers that according to specified requirements, information can not be disclosed to unauthorized individuals, entities or processes, or provided the characteristics of its use. A large amount of data collection includes a large number of enterprises operating data, customer information, personal privacy and all kinds of behavior records. The centralized storage of these data increases the risk of data leakage, and not abused of these data also becomes a part of the personal safety. There is no clear definition to the proprietorship and right to use of sensitive data. And many analysis based on large data did not consider the individual privacy issues involved either.

The integrity of information refers to all the resources which can only be modified by authorized people or with the form of authorization. The purpose is to prevent information from being modified with unauthorized users. Due to the openness of big data, in the process of network transmission, information would be damaged, such as hackers intercepted, interruption, tampering and forgery. Encryption technology has solved the data confidentiality requirements as well as protecting data integrity. But encryption cannot solve all of the safety problems.

B. *Data Storage:*

The formation of network society creates the platform and channel of resource sharing and data exchange for the big data in the field of various industries. Network society based on cloud computation provides an open environment for big data. Network access and data flow provides the basis of rapid elasticity push of the resources and the personalized service. In recent years, from the chain reaction of user account information being stolen on the Internet, it can be seen that big data is more likely to attract hackers, and once being attacked, the volume of stolen data is huge.

Before big data, data storage is divided into relational database and file server. And in current big data, diversity of data type makes us unprepared. For more than 80% of the unstructured data, NoSQL has the advantages of scalability and availability and provides a preliminary solution for big data storage. But NoSQL still exist the following problems: one is that relative to the strict access control and privacy management of SQL technology; secondly, although NoSQL software gain experience from the traditional data storage, NoSQL still exist all kinds of leak.

C. *Data Mining:*

With the development of computer network technology and artificial intelligence, Network equipment and data mining application system is more and more widely used, to provide convenient for big data automatic efficient collecting and intelligent dynamic analysis. On the one hand, big data itself exits leak. Big data itself can be a carrier of sustainable attack. Viruses and malicious software code hidden in large data is hard to find. On the other hand, the technique of attack improves. At the same time of the big data technology such as data mining and data analysis gaining value information, the attacker using these big data technology either, just as the two following aspects.

A large number of facts show that failure to properly handle big data will cause great violations to users' privacy. According to the different contents need to be protected, privacy protection can be further divided into location privacy protection, identifier protection, anonymous connections and so on. The threat People faced with is not only personal privacy leakage, but also prediction and behavior of the people based on big data. In fact, anonymous protection cannot protect privacy very well. Research on social network also shows that user attributes can be found from the group features[6].

Currently collection, storage, management and use of user data is short of specification, and regulation[7]. Users can't determine their privacy information usage. In commercial scenario, user should have the right to decide how their information be used, and realize users' controllable privacy protection.

A general view about big data is: data itself can tell everything, the data itself is a fact. In fact, if not carefully screened, the data can deceive people, just as people can sometimes be deceived by their eyes.

One of the threats of big data credibility is forged or intentionally manufacturing data, and the wrong data often lead to wrong conclusions. If data application scenarios is clearly, someone could deliberately manufacturing data, and create a "false scent", to induced analysts come to the conclusion that was on their side. Because of false information often hidden in a lot of information, it make impossible to identify authenticity of information, so as to make wrong judgment. Due to the production and proliferation of false information in network community is becoming more and easier, its effects should not be underestimated and simply using information security technology to identify the authenticity of all sources is impossible.

V. Challenges in Big Data

A. Failure handling:

Devising 100% reliable systems on the go is not an easy task. Systems can be devised in such a way that the probability of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes and the whole computation process becomes cumbersome. Retaining check points and fixing the threshold level for process restart in case of failure, are greater concerns.

B. Data heterogeneity:

Big data deals with unstructured, semi-structured and structured data. Linking unstructured data with structured data, converting data from one form into another required form needs a lot of research.

C. Data quality:

Huge amount of data pertaining to a problem is undoubtedly a big asset for both Business as well as IT leaders[8]. For predictive analysis or for better decision making amount of relevant data helps a lot. But the quality of such data is based on the source through which they are derived. Though big data stores large relevant data, the accuracy of data is completely dependent on the source domains. Hence, there is a question of how far the data can be trusted and it definitely requires appropriate trust agent filters.

VI. Applications of Big Data

Though the term ‘Big Data’ simply looks like a great buzzword today, In the long run, every phase of our lives will be influenced by big data. The applications of big data can be categorized as below:

1) Customer analysis: Big data helps companies in analyzing the customer purchase patterns and predict the future requirement to companies by means of various models.
2) Optimize business processes
3) Improvement in personal performance optimization
4) Improvement in public healthcare system
5) Growth in Science and Research
6) Enhancement in laws of protection and security
7) Improving and Optimizing Cities and Countries
8) Economic improvement
9) Fraud analysis
10) Analysis of social media access

VII. Conclusion

In today’s world many companies and organizations like educational system, credit card companies, insurance companies etc. are using big data for examination purpose. Vast amount of digital information collected from various sources like credit card companies, government institutes, banking, health care systems are used regularly for analysis purpose. Inspite of several applications[9] and advantages, big data has raised new challenges in the area of privacy and security. The main reason behind the privacy problem is that today huge amount of personal information is freely available directly or indirectly[10] in the form of digital information. Many organizations are utilizing Big data for their personal benefits, profit and to achieve their goals by using the personal information of customers. Misusing of such information is causing loss of trust and faith of customers in organizations. In this paper the strength and applications of big data as well as various privacy issues are discussed.

This paper presents comprehensive surveys of the issues and information security in big data.

References

[1] Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work and Think. Boston: Houghton Mifflin Harcourt, 2013.

[2] Meng Xiao-Feng, Ci Xiang. Big Data Management: Concepts, Techniques and Challenges. Journal of Computer Research and Development, 2013, 50(1): 146-169 (in Chinese).

[3] Big Data Analytics for Security Intelligence, Cloud Security Alliance, September 2013.

[4] IDC, Digital data to double every 18 months, worldwide marketplace model and forecast, Framingham, MA. available at www.idc.com May 2009

[5] R. Agrawal, R. Srikant, “Privacy-preserving data mining”, In: Proceedings of the 2000ACM-SIGMOD on management of data, Dallas, TX, USA, May 15-18, 2000.

[6] Narayanan A, Shmatikov V. How to break anonymity of the Netflix prize dataset. ArXiv Computer Science e-prints, 2006, arXiv:cs/0610105: 1-10

[7] Goel S., Hofman J.M., Lahaie S., Pennock D.M. and Watts D.J.. Predicting consumer behavior with Web search. National Academy of Sciences, 2010, 7 (41): 17486–17490.

[8] A. Katal, M. Wazid, and R. H. Goudar, Big data:Issues, challenges, tools and Good practices, Proceedings of Sixth International Conference on Contemporary Computing (IC3), (Page: 404-409 Year of Publication: 2013 ISBN: 978-1-4799-0190-6).

[9] Ira S. Rubinstein, ‘ Big Data: The End of Privacy or a New Beginning?”, International Data Privacy Law Advance Access published January 25, 2011.

[10] R. Agrawal, R. Srikant, “Privacy-preserving data mining”, In: Proceedings of the 2000ACM-SIGMOD on management of data, Dallas, TX, USA, May 15-18, 2000.