Survey of Research on Information Security in Big Data

Zhang Hongjun[1], Hao Wenning[1], He Dengchao[1], Mao Yuxing[1]

[1]PLA university of Industry and Technology
Nan Jing, China

hdchao1989@163.com

Abstract. *With the fabulous development of information technology, big data application prompts the development of storage, network and computer field. It also brings new security problems. This security challenge caused by big data has attracted the attention of information security and industrial community domain. This paper summarizes the characteristics of big data information security, and focuses on conclusion of security problems under the big data field and the inspirations to the development of information security technology. Finally, this paper outlooks the future and trend of big data information security.*

1. Introduction

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], such as Amazon, DangDang know our reading habits, and Google, Baidu knows our search habits. A number of actual cases show that personal privacy will be exposed even after harmless data being collected[1]. In fact, the meaning of big data information security is much extensive. The threat person facing with is not only personal privacy leak, but also the protection of big data itself and knowledge acquired from it.

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, based on the status of big data research, we analyzed the current security challenges by big data, and elaborated the current information security protection method of big data.

2. Threats of Big Data Security

Just as Gartner said: "big data information security is a necessary fight"[3]. 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 potential 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.

2.1 Data Acquisition

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 confidentiality of information refers that according to a 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.

2.2 Storage of Data

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.

2.3 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, anonymous 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[4].

Currently collection, storage, management and use of user data is short of specification, and regulation[5][6]. 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[7]. In fact, if not carefully screened, the data can deceive people, just as people can sometimes be deceived by their eyes.

one of the threat of big data credibility is counterfeit or deliberately 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 propagation of false information in network community is becoming more and more easy, its effects should not be underestimated and simply using information security technology to identify the authenticity of all sources is impossible.

3. Reason Analysis

With the development and progress of information technology, the security of sensitive data is facing with unprecedented challenges. This is a serious impediment to the spread of new applications. Safety problems mainly displays in the following respects.

3.1 Lack of world recognized laws and regulations for data security and privacy protection

Privacy is not a new problem, but with the development of network technology, privacy has also been gradually amplifier, especially e-commerce (Electronic Commerce, EC) privacy issues, which has become one of the most important issues in the network

economy. However, for existing privacy regulations and policies, there are still somewhere to improve[8].

First of all, because of the different of specifications and law cultural of different countries, privacy law only applies to certain territorial limits which impact limited on the global network. Secondly, many countries are not willing to weaken the economic rise of the Internet brought by the economic boom, so they try to avoid joint intervention with other countries. Moreover, because of the long-term and stability of the law, legal measures can't meet the needs of the rapid development of the Internet.

3.2 The cloud infrastructure has not a uniform and reliable authentication, which cannot prove its credible

With the rapid development of cloud storage, more and more users choose to use the cloud storage to store information. The key characteristic of cloud storage is stored as a service. Users can upload their data to the public API in the cloud. But due to the loss of the users' absolute control of data, some hidden danger of data security arises: (1) Rely on customer management of the certificate too much. (2) The granularity of data storage protection is not enough.(3) Do not consider the perfect data sharing requirements.(4) The lack of an effective regulatory pathway to ensure that the storage of data would not be lost, leak, or abuse.

Nowadays, many cloud storage service provider provides cloud storage services with a very low price or even free. Because of the loss of control of data caused by cloud storage, user is difficult to check the data integrity and confidentiality in cloud storage environment. In the worst case data is stored in the an unknown "corner" of service pool, which lead to the poor cloud storage environment disaster resistance[10].

3.3 Lacking of Creditable Authentication in Cloud Computing Service

While bringing convenience, there are problems in cloud computing, among which security issues are the most critical ones and the main factors enterprises users worry about. CSA(Cloud Security Alliance) puts forward the risks cloud computing faces, including data center security, event responding security, application security, key management security, authentication and access control security, virtualization layer security, backup for disaster recovery and business alignment. At the same time, people have realized there are differences between cloud computing security and traditional security. In traditional IT systems, the owner and the user of the fundamental facility are identical. When it comes to cloud computing, CSP(Cloud Service Provider) owns the fundamental facility which offers computing service, while users have the access to it. This makes adversarial relationship between CSP and users. Cloud computing is a trusted model in its nature, CSPs prove the creditability of its service and users build up confidences in it through CSPs' proof[12].

4. Data Security Protection Technique

Key technologies in Security protection fields are in great demands to face the security challenges. In this section, we introduce important relevant fields.

4.1 Individual User

As with individual users' information in big data environment, the core and basic techniques to provide privacy protection are still in developing period. Take typical K-anonymity scheme as an example, its early version[13] and optimized version divide quasi-identifiers into groups through tuple generalization[14] and restraining method. When an equivalence class has identical value on some sensitive attribute, attackers are able to confirm its value. In response to this issue, researchers proposed 1-diversity[15] anonymity.

Current edge anonymity schemes are mainly based on adding and deleting of the edges. Edge anonymity can be effectively achieved by adding, deleting and exchanging edges randomly[16]. There are problems in such methods that noises randomly added are exiguity, and protections to anonymous edges are insufficient. An important method is to perform division and aggregation operations to super nodes such as node aggregation based anonymous method, genetic arithmetic based method and simulated annealing method based method.

4.2 Internet Enterprise

Information security is critical important for Internet enterprises. System security adopts techniques such as redundancy, network separation, access control, authentication and encryption [18]. Security issues are caused by openness, boundless, freedom of the networks, the key to solve such issues are making network free from them and turning network into controllable, manageable inner system. As network system is the foundation of application system, network security becomes principal issue. Ways to solve network security issues are network redundancy, system separation and access control

4.3 Cloud Service Provider

CSPs provide following measures to prevent security issues in cloud environment. In order to prevent CSPs from peeping users' data and program, separating power and hierarchical management are needed to control access to data in cloud. Provide different authority in accessing data to service provider and enterprise to ensure data security. Enterprise should have total authority and limit authority to CSP.

In cloud computing environment data separation mechanism prevents illegal access to data, however, we should take care of data leakage from CSPs. Mature techniques as symmetrical encryption, public key encryption are available to encrypt data and then upload data to cloud environment. In cloud environment data division is often used with data encryption i.e. encrypted data are scattered in user end and spread in several different clouds. In the way, any CSP is not able to gain complete data.

5. Conclusion and Prospect

Information security in big data environment is a promising fields in information security. This paper introduces impact to information security from two aspects of big data and cloud computing. In general, improving system efficiency and provide general cloud storage functions on premise to ensure user data and access authority are the

research direction of future safe cloud computing. At present, more things need to be done in cryptograph searching and reduplicate data removing.

After all, there is an urgent need of improved solutions concerning the users to control the use of their data and more research should be done in this field and there is also a need for more robust approaches in key management limitation, which could extend traditional approaches to Cloud computing.

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]Chen Mingqi, Jiang He. USA Information Network Security New Strategy Analysis in Big Data [J]. Information Network Security. 2012(8):32—35

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

[5]Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation.//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval(SIGIR'11), Beijing, China, 2011: 325-334

[6]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

[7]http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[8]Study Finds Web Sites Prying Less: Shift May Reflect Consumer Concerns[EB/OL]. http://www CNN.com, 2002-03-18

[9]A survey of data disclosing in 2010 by Verizon[EB/OL].[2012-05-10].

[10]Bessani A, Correia M, Quaresma B, et al. DEPSKY: Dependable and secure storage in a cloud-of clouds [C] //proc of the 6thConf on Computer System. New York: ACM, 2011:31-46

[11]Sweeney L..k-anonymity: a model for protecting privacy. InternationalJournal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10 (5): 557-570

[12]Sweeney L..k-Anonymity: Achieving k-Anonymity Privacy Protection using Generalization and Suppression.

[13]AshwinMachanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1):1-52

[14]Ying X. and Wu X.. Randomizing social networks: a spectrum preserving approach. //Proceedings of the SIAM International Conference on Data Mining (SDM'08), Georgia, USA, 2008: 739-750

[15]Lei Zou, Lei Chen and M. Tamer zsu. k-automorphism: a general framework for privacy preserving network publication. // Proceedings of the 35th International Conference on Very Large Data Bases (VLDB'2009), Lyon, France, 2009: 946-957