ISSN (e): 2250 – 3005 || Volume, 05 || Issue, 06 || June – 2015 ||
International Journal of Computational Engineering Research (IJCER)

Big Data and Information Security

Gang Zeng
(Police Information Department, Liaoning Police College, Dalian, China)

ABSTRACT:

With the development of application of Internet/Mobile Internet, social networks, Internet of Things, big data has become the hot topic of research across the world, at the same time, big data faces security risks and privacy protection during collecting, storing, analyzing and utilizing. This paper introduces the functions of big data, and the security threat faced by big data, then proposes the technology to solve the security threat, finally, discusses the applications of big data in information security.

KEYWORDS: *big data, security risks, information security, information security technology*

I. INTRODUCTION

With the deepening of Internet applications, social networks and internet of things, produced a huge amount of data, which we called big data. It makes the analysis and application of the data more complex, and difficult to manage. These data, including text, images, audio, video, Web pages, e-mail, microblogging and other types, Among them, 20% are structured data, 80% are semi-structured and unstructured data.

big data is large and complex, so it is difficult to deal with the existing database management tools or data processing application. Why do we collect and analyze big data? Because we can get the benefit from it.

(1) to Acquire Knowledge.

Because of big data contains a large number of original, true information, big data analysis can effectively get rid of individual differences, to help people through the phenomenon, more accurately grasp the law behind the things.

(2) to Presume the Trend.

Using the knowledge, we can more accurately predict the natural or social phenomenon. Google predicted the occurrence trend of flu around the world, through the statistics of search for influenza information.

(3) to Analyze Personality Characteristics.

Commercial enterprise collect information on all aspects of customers for a long time, to analyze the user behavior law, more accurately portray the individual characteristics, to provide users with better personalized products and services, and more accurate advertising recommended.

For example, e-commerce sites now use Big Data technology record customer browsing and purchasing history, to guess his interest, and recommend products for him, this may be his interest.

(4) to Discern Truth by Analyzing.

In the network, data sources is diverse, type is rich, so the authenticity can't be guaranteed. At the same time, the spread of information on the Internet is more convenient, so the damage caused by false information on the Internet is greater. Due to the huge amount of data in the big data environment, to a certain extent, it can help discern truth by analyzing the data.

Big data bring the benefits to us, but also bring the questions of data security and privacy protection, since the emergence of big data technology, a large number of security incidents have been occurred, these incidents sounded the alarm for the society: We must study the security question of big data.

II. BIG DATA BRING THREATS TO THE INFORMATION SECURITY

Science and technology is a double-edged sword. Both of the security issues and the value brought by big data become the center of people's attention. Compared with the traditional information security issues, the challenge for big data security is mainly reflected in the following aspects.

2.1. Big Data Increase the Risk of Privacy Leakage

The collection of big data inevitably increases the risk of leakage of user privacy. Because it contains a large number of user information in the data, the development and utilization of big data is very easy to infringe the privacy of citizens, the technical threshold is greatly reduced. The privacy of citizens will be used poisonously.

The cause of information leakage is various, it can be summarized as follows:
(1) The abuse of data leads to the privacy leakage

Big data relates to several areas and departments, and relates to data collecting, storing, processing, analyzing, reporting and other processes, in this process, if the abuse of data will cause privacy leakage. For example, internal staff of data center do not abide by the professional ethics, abuse their functions and powers, release the data of the sensitive department, persons and events to the public, resulted in privacy leakage. When big data is shared among multiple relational departments, this should be also a cause of privacy leakage. Data has a life cycle, at the end of its life cycle, if it does not be effectively destroyed, sniffer can get the data by social engineering, results in privacy leakage.

(2) Analysis and use of data causes privacy leakage

One of the distinctive features of big data is a huge amount of data, the diversity of data source. When analyzing the data, data from different sources are integrated, we may get an unexpected result which can't be got from a single data source. A classic case of privacy leakage in big data environment, the owner of retail store know that his neighbor girl is pregnant early than her parents, and mailed advertising to her, All these are obtained from the analysis of the sales records.

2.2. Big Data Becomes the Carrier of Advanced Persistent Threat

APT is a kind of advanced persistent threat, It's a long time of the attack, the attack process is complex and difficult to find. The main features of APT is that the space to attack is very wide, long duration, single point concealment ability is very strong.

Traditional protection policies is difficult to detect hacker attacks behind big data. Traditional threat detection is based on a real-time feature matching detection on a single point, but APT is an ongoing process, without obvious feature to be detected in real-time, so it can't be detected in real time. Meanwhile, APT code hidden in big data is also difficult to find. In addition, an attacker can also use social network and system vulnerabilities to attack.

Hackers can use big data to expand the attack effect, this is mainly reflected in three aspects:

(1) Hackers can launch a botnet attack using big data server, they may control millions of puppet machine and attack, a single point attack has not such aggressivity.

(2) Hackers can enlarge attack effect by controlling the key nodes;

(3)Hackers can hide data for attacking in large data, which makes difficulty to analysis of security vendors. Any misguiding hacker set, will lead to deviation from the proper direction of safety monitoring.

2.3. Big Data Is Not Necessarily Credible

Because big data is original, the general view is that the big data is true and reliable, the data itself is a fact. In fact, this is not necessarily true, as people can't always believe their eyes.

An important factor affecting the credibility of big data is the correct level of data, If the data comes from the real and reliable production processes, then these data is credible, but if these data is fabricated for a special purpose, so these data is untrusted, because incorrect data will lead to erroneous conclusions .

For example, in some review sites, the real reviews information and bogus reviews mixed together, the user is difficult to distinguish between true and false, and sometimes make wrong judgments based on false reviews, to choose inferior products and services.

Another factor is the gradual distortion in the data dissemination. Data collection process under manual intervention may introduce errors, data distortion due to errors and deviations, and it ultimately affects the accuracy of the results of data analysis. Changes of versions of the data is another factor of data distortion, because of the different versions of the data, users have different understand to the same data, finally, errors will occur.

2.4. How to Achieve Access Control for Big Data

Access control is an effective method to realize data controlled sharing, it is divided into discretionary access control, mandatory access control and role-based access control. While in big data environment, it is difficult to preset the role, to realize the role and to predict the actual authority of each role. Discretionary access control is unable to meet the diversity of the permissions due to the diversity of users, mandatory access control is unable to meet the dynamics of authority, role-based access control is not able to effectively match the role and the corresponding permissions. Therefore, a new security access control mechanisms must be adopted to protect data in big data environment.

III. SECURITY TECHNOLOGY IN BIG DATA ENVIRONMENT

For the security risks of big data, we need to address the security issues of big data from the following points: data privacy protection technology; data integrity and trusted technology; access control technology.

3.1. Data Privacy Protection Technology

To protect the privacy of big data, even if the data with privacy leak, the attacker can’t obtain the effective value of data. We can use data encryption and Data anonymity technology

(1) Data Encryption Technology

Data encryption technology is an important means to protect data confidentiality, it safeguards the confidentiality of the data, but it cut down the performance of the system at the same time. The data processing ability of big data system is fast and efficient, which can satisfy the requirements of the hardware and software required for encryption. So the homomorphic encryption has become a research hotspot in data privacy protection.

The homomorphic encryption is a model for the calculation of the cipher text, avoiding the encryption and decryption in the unreliable environment, and directly operation on the cipher text. Which is equivalent to the procedure of processing the data after decryption, then encrypting it. Homomorphic encryption is still in the exploratory stage, the algorithm is immature, low efficiency, and there is a certain distance away from practical application.

(2) Data Anonymity Technology

Data anonymity is another important technology for privacy protection, Even if the attacker gets the data that contains the privacy, he can't get the original exact data, because the value of the key field is hidden. However, in the background of big data, the attacker can obtain data from multiple sources, then associate the data from one source with another source, then will find the original meaning of the hidden data.

(3) Generalization Technology

The third technology of privacy protection is generalization technology, which is to generalize the original data, so that the data is fuzzy, so as to achieve the purpose of privacy protection. For example: I live in 3-13-2 No. 187 Guangyuanli Lane Xuanwu District Beijing. This address is very detailed, now we change the address to the city of Beijing, so that the value of address has become vague, so as to achieve the purpose of privacy protection.

3.2. Access Control Technology

Big data contains a wealth of information resources, all professions and trades have great demand of the data, so we must manage access rights of big data carefully. Access control is an effective means to achieve controlled sharing of data, but in big data environment, the number of users is huge, the authority is complex, and a new technology must be adopted to realize the controlled sharing of data.

(1) Role Mining

Role-based access control (RBAC) is an access control model used widely. By assigning roles to users, roles related to permissions set, to achieve user authorization, to simplify rights management, in order to achieve privacy protection. In the early, RBAC rights management applied "top-down" mode: According to the enterprise's position to establish roles,

When applied to big data scene, the researchers began to focus on "bottom-up" mode, that is based on the existing "Users - Object" authorization, design algorithms automatically extract and optimization of roles, called role mining.

In the big data scene, using role mining techniques, roles can be automatically generated based on the user's access records, efficiently provide personalized data services for mass users. It can also be used to detect potentially dangerous that user’s behavior deviates from the daily behavior.

But role mining technology are based on the exact, closed data set, when applied to big data scene, we need to solve the special problems: the dynamic changes and the quality of the data set is not higher.

(2) Risk Adaptive Access Control

In big data scene, the security administrator may lack sufficient expertise, Unable to accurately specify the data which users can access, risk adaptive access control is an access control method for this scenario. By using statistical methods and information theory, define Quantization algorithm, to achieve a risk-based access control. At the same time, in big data environment, to define and quantify the risk are more difficult.

3.3. Data Provence technology

Due to the diversification of data sources, it is necessary to record the origin and the process of dissemination and calculation, provide additional support for the latter mining and decision.

Before the emergence of the concept of big data, Data Provence technology has been widely studied in database fields. Its purpose is to help people determine the source of the data in the data warehouse.

The method of data provence is labeled method，through the label, we can know which data in the table is the source, and can easily checking the correctness of the result, or update the data with a minimum price.

In the future, data provence technology will play an important role in the field of information security. But data provence technology for big data security and privacy protection also need to solve the following two questions: 1, The balance between privacy protection and data provence;2, to protect the security of data provence technology itself.

IV. APPLICATION OF BIG DATA TECHNOLOGY IN INFORMATION SECURITY

4.1 Threat Discovery Technology Based on Big Data

Compared with the traditional security system, the big data technology used in the security system, can find the security threat earlier. For example: big data technology can detect abnormal behavior in the network, predict the attack behavior, and analyze the source of the attack. By analyzing the email, social network information, you can analyze the disgruntled employees in your enterprise, and make timely plans to prevent security incidents.

Security analysis techniques using big data, has the following characteristics:

(1) Big Data Is More Suitable For Security Analysis

The core technology of safety analysis is the analysis of safety behavior and security incidents, this often come from the analysis of the log, and big data technology is more suitable for the data collection, storage, analysis of the log and other unstructured data. The result of the analysis is very valuable for the security protection, and can also be visualized.

(2) The Range of Content Analysis Is Broader

The security monitoring system using big data technology can collect social network data, personally identifiable information, track information of travel and driving, financial information, call ticket, e-mail information, shopping information, medical information, enterprises business information. According to these information, some abnormal incidents related to security can be analyzed, and steps are taken to prevent security incidents happening.

(3) The time span of the analysis is longer

Analysis of security incidents tend to have a very long time span, it requires a lot of computing power and a large amount of data. The traditional threat analysis system can not meet the requirements of computing power, but the threat analysis system using big data has a variable resource pool, and can analyze vast amounts of data, so as to predict the security incidents as soon as possible, such as APT attack.

(4) The prediction of threats

We always want to predict the security incidents as early as possible, and prevent them happening. The traditional threat analysis system has small amount of data, small range, short time span, and can't predict the security incidents within a long time span and a wide range. The threat analysis system using big data technology can predict the impending threat, according to the characteristics of the security incident.

(5) Detection of unknown incidents

Based on historical experience, we can detect security incidents happened before, but the world is constantly changing, and new security incidents may occur. The analytical work in traditional threat analysis system are carried out by experienced analysts, they mostly make judgments based on historical incidents, so that they can't detect the unknown incident. The security incidents are often interrelated, the threat analysis system using big data technology can detect the unknown security incidents, according to the relation between security incidents, and not a causal relationship, for example: a web site, it has been running very well, recently the ability to provide services constantly reduced, there were no security incident that have been occurred in history, then we can predict that the web site may be attacked by unknown way.

4.2 The Identity Authentication Technology by Behavior Characteristic Based on Big Data

Traditional authentication technologies rely on three ways for certification: Who I am (biometric authentication), what do I know (password authentication), what I have (ID card). In the big data environment, users can be authenticated by behavioral characteristics. This is different from the previous authentication technology, it is difficult for user to pass authentication by feature forgery, at the same time the burden on the user will be reduced, the user needn't remember the complex password, or with id card.

But, this authentication method requires a training data set of user's behavioral characteristics, in the initial stage, it is necessary to collect data and form the training data set which authentication needs, this may impact on the operation of the system.

4.3 Authenticity Analysis Based on Big Data

Above, we discussed that big data is not necessarily true, the real data on the security will have a certain impact, untrue data will impact on security. We can analyze the authenticity of the data by big data technology, and remove false data. For example: using big data, we can distinguish which mail is spam, which comment in comment network is false comment, which data collected in a production system is wrong.

V. CONCLUSION

This paper first introduces the function of big data, and then introduces the security threat faced by big data, and proposes the technology to solve the security threat, finally, discusses the applications of big data in information security. Of course, with the development of big data technology, new security threat may appear, we need to find new solutions and technologies to solve it.

ACKNOWLEDGEMENT

This work was partly supported by general scientific research project of education department of Liaoning Province of China (No. L2013490).

REFERENCES

[1] FENG Deng-Guo, ZHANG Min, LI Hao. Big Data Security and Privacy Protection[J]. Chinese Journal of Computers, 2014,37(1):246-258.

[2] MA Li-chuan, PEI Qing-qi, LENG Hao, LI Hong-ning. Survey of Security Issues in Big Data[J]. Ｒadio Communications Technology, 2015,41(1):1-7.

[3] Hu Kun, Liu Di, Liu Minghui. Research on Security Connotation and Response Strategies for Big Data[J]. Telecommunications Science, 2014(2):112-117,122.

[4] WANG Yu-long，ZENG Meng-qi. Big Data Security based on Hadoop Architecture[J]. Information Security and Communications Privacy, 2014(7):83-86.

[5] Big Data Working Group. Big Data Analytics for Security Intelligence[EB/OL]. https://www.cloudsecurityalliance.org/research/big-data.

[6] Guillermo Lafuente. The big data security challenge[J]. Network Security,2015.(1):12-14.

[7] Hrestak D, Picek S. Homomorphic Encryption in the Cloud［C］‖2014 37th International Convention on Infor-mation and Communication Technology，Electronics and Micro electronics(MIPＲO)，2014: 1400-1404．

[8] L.D. Cohen. NOTE On Active Contour Models and Balloons. Computer Vision and Image Processing: Image Understanding, 53:211-218, March 1991.