Research on Big Data Security

Xufan Zhang [a], Qingyuan Liu [b], Wenhao Bai [c]
Xi'an station of surveying and mapping, Xi'an 710043, China.
[a]344518517@qq.com, [b]259415397@qq.com, [c]415972760@qq.com

Keywords: big data, data security, security strategy

Abstract. With the continuous development of data science, big data has become one of the strategic heights. competition increasing throughout time. The explosive growth of big data applications has created a rare opportunity for development, but there are still some difficulties and problems that exists, data security and privacy issues can't be avoided. Data security issues not only related to the characteristics of the data itself, but also involves data collection, transmission, storage, processing, destruction and other technical and data production, and the user's awareness. Need to start from the technical and legality level at the same time, providing an effective way to resolve the security issues of big data.

1. Introduction

Big data has become one of the strategic high ground of building a competitive nation, and the national competition is becoming more and more intense. The usage and mastering of big data has become an important factor of national competitiveness. Countries around the world are using big data as a national strategy, and using industrial development as the core development of the big data. The United States putted their attention onto the development and application of the big data, launched "Big Data Research and Development Initiative" in March 2012, seeing big data as an important national strategic resources for management and application. Moreover, they launched "federal data research and development Plan " in May 2016, continue to strengthen the layout and application of big data. In 2014, the EU launched a "data-driven economy" strategy, advocating European countries to seize the opportunity to develop big data. In addition, the United Kingdom, Japan, Australia, and other countries also introduced a similar policy to promote the application of big data, stimulating industrial development. Since 2013, China has shifted from its "speculation" period gradually and fall into the growth period of 2016, big data industry is experiencing an unprecedented transformation. Currently there is 84,000 websites that are available for the public to visit. They also went full out with the intelligence city, there are near 300 cities that has become the beta of intelligence city.

With the continuous development of big data industry, data information has penetrated into the political, economic, cultural, livelihood and other aspects of the activities. According to statistics, the current Google search average more than 2 million per second use, the number of people posted every day more than 340 million Twitter, every day there are 4 billion kinds of content sharing data generated by Facebook users. Nowadays, all industries are constantly producing a lot of data. Almost every industry will usher in the basic data construction and big data application system construction of two aspects of the explosive-development opportunities. Up to now, the world has officially have nearly 40 countries and regions to build their own data open portal[1]. The development of big data to expand the degree of openness of information, however the data security and privacy protection issues is kicking in.

2. The Security problems that big data is facing

2016 years can be described as a boomer year of the big data industry. There were a large-scale data growth of the field of data science. Most likely it will continue to grow into the next few years of development. With the Internet, the outbreak of big data, applications are getting more and more abundant, data security has become nowadays the most concerned problem for the people. According to Gemalto's recent data, the number of data leaks in the world in the first half of 2016 is as high as 974 cases, with a total of more than 554 million data leaks, increased by 15 percent from 2015. US National Revenue Service, Yahoo, Ecuador, the Philippines, Bangladesh and other national banks and China FACC have become the victims of data leakage.

2.1 Big data becomes the victim easily

Big data not only means that the amount of data is huge, but also means more complex data types and more sensitive data content, such data characteristics have become the coveted reasons for the attackers. With the deepening development of the network society, interoperability between networks are also deepening as well, the correlation between the various data is more closely, which not only increased the success rate of hackers successfully attack, let them get more data, reduced the cost of the hacker's attack[2]. It also allow the attacker to hide their attacks in the big data, so that the traditional protection systems can't detect it. Specifically for big data, APT (Advanced Persistent Threat) is a typical high-level continuous attack[3], it has the characteristic of long attacking duration, the attack process is complex and hard to be found. Attackers hide de APT attack code in Big data, using big data to launch botnet attacks, can simultaneously control a large amount of "puppets" to attack, making the attack more precise, causing a serious threat to the network security[4]. The continuous characteristic is the best cover for the attack, directly avoids the detection.

Big data's vulnerable security increase the risk of data security, program interface security mechanisms and data acquisition, storage, application, management and other aspects of the problem may cause data leakage[5]. Although the use of big data technology can be used to integrate computing and processing resources. It also can provide effective help to discover of network attacks and find the source of the attack[6], but the development of technology provides more strategies to the network attaching, and the attacker is also using data mining and data analysis to obtain the corresponding value information.

2.2 Technical development increases data leak risk

In the era of big data, the information security is facing the new normal state, and the whole process of big data life cycle in the process of data collection, transmission, storage, management, analysis, distribution, usage and destruction, there are security problems in every aspect of big data.

As a distributed system infrastructure, Hadoop can process a large amount of data, it is widely used in the big data processing applications with the advantage of low cost, scalability, high reliability, and high efficiency. But this technology also has security problems, with the continuous development of Hadoop, people also found a lot of security vulnerabilities. Such as the lack of access control mechanism, the lack of encryption mechanisms for data storage and transmission, and the lack of effective security authentication between users and servers. IBM, Yahoo, Google, and other Internet giants have taken precautionary measures against these security issues; government agencies have introduced role management to isolate Hadoop data to prevent illegal access; some companies, based on their data security considerations, simply gave up using Hadoop[7].

At present, more than 80% of the various types of big data are unstructured data[8]. NoSQL database came along with this, because it can solve the massive data storage problems, with a high concurrency, high scalability, and other advantages to occupy a rapid development. Because NoSQL technology to solve the massive data in the data source and the diversity of the problem, it makes enterprises difficult to locate and protect the confidential information[9], which is because NoSQL lack of confidentiality and integrity caused by The inherent security mechanism is imperfect. On the other hand, NoSQL associates data from various systems, applications, and behaviors, and increases the risk of data leakage.

2.3 Data junks that can't be ignored

At present, the data generated by human activities are growing exponentially. people often only concerned about their own interest in the relevant data, other data will be low or have a negative effect, resulting in creating a lot of data junk.

Data waste, like domestic waste, is the waste that people produce in the application process, which directly affects the day-to-day use of the data, not only to the impact of storage efficiency, but also to the efficiency of the data that the data manager must clear or abandoned[10]. But in the big data era, any data is recognized as valuable, which contains data garbage. Data garbage will generally show fragmented, small, and unimportant features, these fragmented, will add up to the formation of big data after sorting. "The true value of the data is like an iceberg floating in the ocean, at first glance, only the beginning, and most of it is hidden beneath the surface." (Big data, A Revolution That Will Transform How We Live, Work, and Think)

In the era of big data, we need to change the way we know and deal with data garbage. The re-mining of any data and the re-use of the results are unpredictable. Data junk may be dormant most of the time, but we cannot put the data aside, they should be kept under effective supervision as long as possible, preventing data leaking due to the cause of ignoring data garbage.

3. Big data application security policy

Security is the prerequisite for development, development is the base security protection. The core resources for big data applications are data, data regulation and security protection become the primary task of big data application security. At the same time, big data operating environment involves network, host, application, computing resources, storage resources and other aspects, it needs to have a deeper security means. Facing with all aspects of security challenges, big data security protection needs to go from the technical and legal protection of two-pronged approach.

3.1 Increase big data application’s ability to protect privacy

Through the reconstruction of hierarchical access control mechanism, deconstruct sensitive data association, the implementation of data lifecycle security protection, and enhance the application of big data privacy protection. "We should try to figure out several important rights and obligations in the big data area, including data ownership.” (Douglas Laney, Gartner)

Part of the big data applications to further analyze the PII and UI-related information, and thus targeted to the application of precision marketing, such applications have the greatest impact on privacy violations, so the relationship between PII and UL information is the focus of big data privacy protection, and because PII direct Related to all types of user information, it is also the focus of big data privacy protection.

In the big data privacy protection, it should be based on the sensitivity of PII and UL data classification, and then reconstruct the data security access control mechanism. Sort the original data, UL data, PII data and PII and UL related data according to the security level from low to high classification, and according to the security needs of the implementation of user identity access control, encryption and other different levels of security policy to limit the scope of data access. At the same time in the large data operation should be as much as possible to achieve PII data and personal attribute data deconstruction, store the PII data and UL data separately, add index for the PII data, connect the UL and PII through the index table, so if the hackers were to obtain the UL information, they can’t obtain user’s PII information and its corresponding information. Encrypt the index at the same time, so if the hackers were to obtain the index, they still can’t obtain the user’s PII info.

In the data classification and deconstruction on the basis of the data should also be implemented throughout the life cycle of security protection. Focus on strengthening the data interface control, the data batch export interface for approval and monitoring of the data interface for periodic audit and evaluation, standardize the data interface management, and data out of the sensitive data desensitization. At the same time using secure communication protocol to transmit data, such as SSL / TLS, HTTPS, SFTP, etc., and the transmission of important data per the need for encryption. When data is destroyed, all copies of the data should be cleared to ensure that the storage space for the user's authentication information, files, directories, and database records is released before being released or reallocated to other users.

Strengthen the development of big data security technology products. Focus on the research of unified account, authentication, authorization and auditing system and big data encryption and secret management system under big data environment, breaks through key technologies such as differential privacy technology, multi-party security calculation, data flow monitoring and traceability. Promote anti-leakage, anti-theft, anonymity and other big data protection technology, research and development of big data security products and solutions.

3.2 Improve Big data computing environment security defense

Make a good big data application computing platform, distributed probe, network and host infrastructure security protection. Big data computing environment including the network, host, computing platform, distributed probe, etc., for all aspects of the security risks facing.

First, strengthen the big data computing platform to enhance the computing environment security. Enhance the client / server application authentication service function, establish a security authentication center, deploy multiple server nodes, avoid single point defects; store all the metadata encryption; in the case of performance allows HDFS raw data transparent encryption, Also configure the Web console and MapReduce between the random operation using SSL for encryption, configure the HDFS file transfer for the encrypted transmission.

Strengthen the security of the probe device, set a secure login account and password, update the system patch from time to time; set the anti-virus and intrusion detection; remote operation for strict access control, limit the specific IP address access; run fine-grained audit against probe login and its operations; set encryption protection on the stored local data; implement abnormal traffic monitoring and DDoS attack protection in the probe public network.

Finally, strengthen the big data system network, host, terminal and other infrastructure operating environment security protection. It should adopt the traditional security protection means to construct the deep security protection system, divide the security domain at the network level, deploy the border access control, intrusion detection / defense, abnormal traffic monitoring, DDoS attack defense, VPN and other security means; place deployment intrusion detection, Vulnerability scanning, virus protection, operation monitoring, patch management and other security means at the host level; Place the deployment of access control, terminal security management, vulnerability scanning, virus protection and other security means in the terminal level.

In addition, it should enhance the support of big data on the security of network information. Comprehensive the use of multi-source data, to strengthen big data mining analysis, and enhance network information security risk perception, early warning and disposal capacity. Build a big data unified security control, component monitoring, resource monitoring and other basic security service facilities. Data monitor and analysis big data platform hosts, network, big data components and tenant applications.

3.3 Improve supervision and punishment measures

Today, data ownership, privacy and other relevant laws and regulations and information security, open sharing and other standard is not perfect, it has not yet established a balance of security and development of data open, management and information security system. In the State Council issued the "thirteen five" national strategic emerging industry development plan mentioned: comprehensively promote the key areas of big data efficient collection, effective open sharing and application development, improve the supervision and management system, strengthen security. "Big data security issues were also highlighted in the Ministry of Industry issued Big Data Industry Development Plan (2016-2020). The EU has just introduced a new privacy protection law for the use of big data and the establishment of analytical models. During process of collection and using big data, it may involve the national security information and may violate the privacy of citizens, which is the problem that can’t be avoided throughout the development of big data, not only for the needs of enterprises and related institutions of self-restraint, but the country also need to be clear about the principle of big data collection at the legislative level. Need to consider the personal privacy issues that is brought along by opening of the data, as well as the protection of business secrets, the relationship between national information security and social data needs. Not only to strengthen the awareness of self-restraint, but also to clearly regulate the data collection, storage, data transmitting and the process of application.

On the other hand, protect the information through the information system hardware and software investment. The relevant departments of the country should pay attention to the security of the big data platform itself, through the development of big data technology standards and operational norms to strengthen the construction of information security system, but also to strengthen the key areas of sensitive data regulation. Not only in the data and information collection and application, but also in the collection of storage, management and other processes to strengthen the system. Application of explicit data Open sharing must be bordered, regular, step-by-step, and supervised on the use of open object data in accordance with applicable laws and conventions to achieve a balance between data open demand, privacy requirements and security requirements. Making data open, mobile, usage and sharing, can further reduce the cost of governance, improve governance efficiency, so as to further enhance the effectiveness of governance.

4. Conclusion

According to Jim Gray, data-intensive science, represented by big data, is the fourth largest paradigm behind experimental science, theoretical science, and computational science, and is an inevitable trend for future scientific development. Today, the challenges and opportunities that the development of big data is facing, needs to consider design, construction, operation and other stages at the same time to develop a protection technology and solutions. Construct a sophisticated big data security system, and thus continue to promote the development of big data applications[11].

References

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

[2] Yang JianChun. Research on Data Security Control Technology in Network Environment [J]. Gansu Science and Technology,2011(16), pp:22-24.

[3] Pan ZhuTing. High end information security and big data[J]. Information Security and Communication Confidentiality,2012(12), pp:19-20.

[4] Guo Sanqiang, Guo Yanjin. Resear ch of Data Security Based on Big Data Era [J]. Technology Square,2013(2):28-31.

[5] Zhou DongSheng. Information Security Risk and Countermeasure in Cloud Computing and Big Data era. Research Report, computer science and Technology, Jiangnan Institute of Computing Technology. 2014

[6] Feng Wei. Information Opportunities and Challenges for Big Data era[J]. Expert forum 50-53.

[7] Hong Sha, Xiang SenYuan. Research on Cloud Computing Key Technology and Hadoop Based Cloud Computing Model[J]. Software Guide, 2010,9(3):31-33.

[8] The big data security gap: protecting the Hadoop custer challenges and opportunities with big data. http://www.zettaset.com/info-center/datasheets/zettaset_wp_security_0413.pdf.

[9] Gao ChunYan. Big data’s security essential point [N].. China Computer News,2012,8(20): 1-4.

[10] Fu Guo. Be aware of data junk leaks in big data era[J]. Information Security, 2015(10):55-56

[11] Gong Xueqing, Jin Cheqing, Wang Xiaoling elt. Data-Intensive Science and Engineering: Requirements and Challenges[J]. Chinese Journal of Computers,2012,8(35):1563-1578.