Research on Security and Privacy Protection Based on Big Data

Yan Yang

Officers College of Pap, Chengdu, Sichuan, 610213, China

Keywords: The big data, Privacy protection, Security problems

Abstract: Big data is the current research hotspot in academia and industry, which is affecting people's daily life, work habits and thinking patterns. However, at present, big data faces many security risks in the process of collection, storage and use. The privacy leakage caused by big data has caused serious problems for users, and false data will lead to erroneous or invalid big data analysis results. This article analyzes the technical challenges facing big data security and privacy protection, and summarizes some key technologies and their latest developments. The analysis points out that while big data introduces security problems, it is also an effective means to solve information security problems. It has brought new opportunities for the development of the field of information security.

1. Introduction

Big data refers to a collection of data that cannot be captured, managed, and processed with conventional software tools within an affordable time frame. It is a massive and high-growth model that requires new processing models to have greater decision-making power, insight, and process optimization capabilities. Rate and diversified information assets. User data security and privacy protection are undoubtedly one of the most important issues in the context of big data. Although there are various methods for realizing big data security and privacy protection, the most thorough method is to realize user data security and privacy protection through encryption. However, an ensuing question is: how to implement the same big data processing technology in the ciphertext domain as in the plaintext domain? The most important problem is not only to solve the functional problems such as ciphertext calculation, ciphertext access control and ciphertext data aggregation, but also the more important problem is to solve the new processing mode problems corresponding to these problems, that is, like The problem of big data processing mode in the plaintext domain is the same.

2. Big Data Sources and Characteristics

The general view is that big data refers to a data set that is large and complex, making it difficult to process with existing database management tools or data processing applications. Common characteristics of big data include large-scale, high-speed and diversity. According to different sources, big data can be roughly divided into the following categories: (1) from people. People generated in the process of Internet activities and the use of mobile Internet All kinds of data, including text, pictures, videos and other information; (2) From the machine. The data generated by various computer information systems exist in the form of files, databases, multimedia, etc., and also include automatically generated information such as audits and logs. (3) From objects. Data collected by various types of digital devices, such as digital signals generated by cameras, various characteristic values of people generated in the medical Internet of Things, and large amounts of data generated by astronomical telescopes.

People have been analyzing data for a long time. The first and most important purpose is to acquire knowledge and use knowledge. Because big data contains a lot of original and real information, big data analysis can effectively discard individual differences and help people grasp the laws behind things more accurately through phenomena. Based on the mined knowledge, it is possible to predict the natural or social phenomena more accurately. A typical example is Google's

Google Flu Trends website. It counts people's search for influenza information and queries the IP address of the Google server log to determine the source of the search, thereby releasing forecasts of influenza conditions around the world. As another example, people can predict stock prices based on Twitter information.

While individual activities meet certain group characteristics, they also have distinctive individual characteristics. Like the slender tail in the “Long Tail Theory”, these characteristics may vary widely. Through long-term, multi-dimensional data accumulation, enterprises can analyze user behavior laws, more accurately depict their individual profiles, provide users with better personalized products and services, and more accurate advertising recommendations. For example, Google uses its big data products to analyze the habits and hobbies of users, help advertisers evaluate the efficiency of advertising campaigns, and estimate that there may be a market size of hundreds of billions of dollars in the future. Error information is worse than no information. As the spread of information in the network is more convenient, the harm caused by false information on the network is also greater. For example, on April 24, 2013, the Twitter account of the Associated Press was stolen and false information was issued saying that President Obama was injured by a terrorist attack. Although false news was banned within a few minutes, it still triggered a brief dive in the US stock market. Because of the wide source and diversity of big data, it can help to realize the anti-counterfeiting of information to a certain extent. At present, people are beginning to try to use big data to identify false information. For example, social commenting website Yelp uses big data to filter false reviews and provide users with more authentic commentary information; Yahoo and Thinkmail use big data analysis techniques to filter spam.

3. Security Challenges Brought by Big Data

At present, the collection, storage, management and use of user data lack standards, and even lack supervision, mainly relying on the self-discipline of enterprises. Users cannot determine the purpose of their private information. In a commercialized scenario, users should have the right to decide how their information is used to achieve user-controllable privacy protection. For example, users can decide when and how their information is disclosed and when it is destroyed. Including: (1) Privacy protection at the time of data collection, such as data accuracy processing; (2) Privacy protection at the time of data sharing and publishing, such as anonymous processing of data, manual scrambling, etc.; (3) Privacy protection at the time of data analysis; (4) Privacy protection of data life cycle; (5) Trusted destruction of private data, etc. A common view about big data is that the data itself can explain everything, and the data itself is the truth. But the reality is that without careful screening, the data will be deceived, just as people are sometimes deceived by their own eyes. One of the threats to the credibility of big data is forged or deliberately created data, and wrong data often leads to wrong conclusions. If the data application scenario is clear, some people may deliberately create data and create some kind of “illusion” to induce analysts to draw conclusions that are beneficial to them. Because false information is often hidden in a large amount of information, it makes people unable to distinguish between authenticity and false judgment. For example, some fake reviews on review sites, mixed in with real reviews, make users indistinguishable, and may mislead users to choose some inferior goods or services. Since the generation and dissemination of false information in the current online community is becoming easier, its impact cannot be underestimated. It is impossible to identify the authenticity of all sources with information security techniques. The second threat to the credibility of big data is the gradual distortion of the data during its propagation. One of the reasons is that manual intervention may introduce errors in the data collection process. The errors cause data distortion and deviation, which ultimately affect the accuracy of the data analysis results. In addition, data distortion also has the factor of data version changes. During the communication process, the reality has changed, and the data collected in the early stage can no longer reflect the reality. For example, the restaurant phone number has been changed, but the early information has been included by other search engines or applications, so users may see conflicting information and affect their judgment. Therefore, users of big data should have the ability to understand the credibility of various data

based on the authenticity of the data source, data dissemination channels, and data processing processes, so as to prevent analysis from obtaining meaningless or erroneous results.

4. Big Data Services and Information Security

Due to the emergence of big data analysis technology, enterprises can go beyond the previous "protection-detection-response-recovery" (PDRR) model and more proactively discover potential security threats. For example, IBM has launched a new security tool called IBM Big Data Security Intelligence, which can use big data to detect security threats from inside and outside the enterprise, including scanning emails and social networks, marking employees with obvious dissatisfaction, reminding enterprises to prevent their disclosure of corporate secrets. The "Prism" project can also be understood as a success story of applying big data methods for security analysis. By collecting various types of data from various countries, using security threat data and security analysis to form a systematic method to discover potentially dangerous situations and identify threats before an attack occurs. Compared with traditional technical solutions, the threat discovery technology based on big data has the following advantages.

The traditional threat analysis mainly targets various security incidents. An enterprise's information assets include data assets, software assets, physical assets, personnel assets, service assets, and other intangible assets that support business. Due to the limitations of traditional threat detection technology, it cannot cover these six types of information assets, so the threats that can be found are also limited. By introducing big data analysis technology in threat detection, attacks against these information assets can be discovered more comprehensively. For example, by analyzing the instant messaging data and Emilia data of enterprise employees, it can be found in time whether the personnel assets are facing the threat of attack by other enterprises. As another example, through the analysis of the order data of the customer department of the enterprise, it is also possible to find some abnormal operation behaviors, and then judge whether it harms the company's interests. It can be seen that the expansion of the analysis content makes the threat detection based on big data more comprehensive. Although the threat discovery technology based on big data has the above-mentioned advantages, there are still some problems and challenges in this technology, mainly focusing on the accuracy of the analysis results. On the one hand, the collection of big data is difficult to be comprehensive, and the data is the basis of analysis. Its one-sidedness often leads to the deviation of the analyzed results. In order to analyze the threats faced by enterprise information assets, it is necessary not only to comprehensively collect data within the enterprise, but also to collect data outside some enterprises, which is a big problem to some extent. On the other hand, the lack of big data analysis capabilities affects the accuracy of threat analysis. For example, the New York Investment Bank will have 5,000 network events per second, and will capture 25TB of data from it every day. If there is not enough analysis ability, it is necessary to accurately find very few events that indicate a potential attack from such huge data, and then analyze the threat is an almost impossible task.

5. Conclusion

Big data brings new security problems, but it is also an important means to solve the problem. This article reviews the key technologies related to big data security and privacy protection from the perspective of privacy protection, trust, and access control of big data. But generally speaking, the current research on big data security and privacy protection at home and abroad is still insufficient. Only through the combination of technical means and relevant policies and regulations can we better solve the problems of big data security and privacy protection.

References

[1] Mi Nan, Liang Yu. Research progress based on big data security and privacy protection. Construction Engineering Technology and Design, no. 19, 2017.

[2] Cao Zhenfu. Two explanations for the article “Research Progress of Big Data Security and Privacy Protection”. Computer Research and Development, vol. 52, no. 12, 2893-2894, 2016.

[3] Zhang Xiaoxuan. Research on Big Data Security and Privacy Protection Methods. Private Technology, vol. 22, no. 9, pp.151, 2018.

[4] Zhong Libo. Research on privacy protection based on big data processing flow. China Information Security, vol. 10, no. 5, pp.110-113, 2014.

[5] Tong Wei, Mao Yunlong, Chen Qingjun, et al. Privacy protection against big data analysis: research status and progress. Journal of Network and Information Security, vol. 2, no. 4, pp. 44-55, 2016.