# Research of Big data and data security

Qiong Ren [a], Zhongyuan Xu

School of Mathematics and Computer Science, Jianghan University, Wuhan, China

[a]qren@163.com

**Keywords:** Big data, Data security, Information security.

**Abstract.** With the rise of blogs, social networks, and the vigorous development of cloud computing and Internet technology, the data is increasing at an unprecedented speed and cumulative, the big data era has come. The basic concept of big data, key techniques and the use of there are a lot of doubt and controversy. In this paper, starting from the essence behind the big data problem, the existing big data research materials to conduct a comprehensive induction and summary, and data security of big data has carried on the further discussion.

## 1. Introduction

Wikipedia definition of big data refers to the use of commonly used software tools to capture, manage, and process the data by time than can tolerate data sets. Big Data to meet the three features: the scale (volume), diversity (variety) and recommend suite (velocity). At present, the development of the big data is still faced with many problems, security and privacy issues are one of the key issues of recognized. People's behavior on the Internet are all in the hands of merchants on the Internet, including shopping habits, contact friends, reading habits, retrieval, habits, etc. A number of actual cases show that even harmless after data is a large collection of, also exposed privacy. In fact, the big data security meaning more widely, people are faced with the threat is not limited to personal privacy leakage. As with other information, large data in storage, processing, transmission process faces many security risks, such as data security and privacy requirements. To achieve large data security and privacy protection, and other safety issues than ever before are more difficult. This is because in cloud computing, although service providers to control the data storage and operation environment, but the user is still some way to protect their data, for example by cryptography technology for data safety storage and computing, or by way of trusted computing environment safety, etc., under the background of big data, such as Facebook businessman is the producer of the data, and data storage, managers and users, therefore, simply by technical means to limit the businessman makes use of user information. Therefore, the user privacy protection is an extremely difficult matter.

## 2. Current situation of big data

The generation of large data. Database right from the start as the main way of data management, data mode of human society has experienced roughly three stages, and it was of great change in the way data that ultimately lead to the generation of big data. The first stage is the operating system. The emergence of database greatly reduce the complexity of the data management, mostly in the actual database for use by the operating system, as the operating system of the data management subsystem, such as supermarket sales record system, bank transaction records system, hospital patients' medical records, etc. This stage is the most important characteristics of the data is often accompanied by certain operations to generate and record in the database. The generation of this kind of data is a passive way. The next stage is the phase of user-generated content. The birth of the Internet makes human society of data appeared big leap forward, there are mainly two reasons: first is the blog, microblogging as a representative of the emergence of new social network and rapid development, make the user data will produce more strong; Second is represented by smartphones, tablets, the emergence of new mobile device, which is easy to carry, all-weather access networks mobile device makes people more convenient way to express his opinions on the Internet, the generation of data at this stage is active. The last stage is the phase of perception system. The human society the third data leap eventually led to the production of large data, today we are at this stage. The leap of perception system to which is widely used. These passive, active and automatic data constitute the source of data for large data, but the automatic data type is the most fundamental reason of big data.

The framework of big data. Applications of big data type has a lot, the main processing mode can be divided into stream processing (stream processing) and batch (batch-processing). After the first batch is stored in, and the stream processing is directly handle. Stream processing is the basic concept of the value of the data which will be reduced with the passage of time and continuously, so as soon as possible to the latest data analysis and the result is the common goal of all flow data processing model. Using the data processing flow of main application scenarios are hits real-time statistics, high-frequency trading in the sensor network, financial etc. When new data arrives immediately and return the desired results. Figure 1 is the basic data stream processing flow model:

Graphs is the most typical batch processing model, its core design idea is: 1) the problem to divide and rule;2) pushed the computation to the data rather than pushing data to calculate, effectively avoid occurring in the process of data transmission communication open pin. Simple graphs model, and a lot of problems in reality are available graphs model. So this model open immediately after is a lot of attention, and in the field of bioinformatics, text mining has been widely used. Both in stream processing and batch is feasible ideas of big data processing. The application of large data type many, in the actual data processing, often is not simply to use one of these, but combine the two. The Internet is one of the most important source of big data, many Internet companies according to their own business divisions according to the requirement of the processing time for online (online), near line (nearline) and offline (offline), including online processing time in the second level or even millisecond, so usually adopt the above stream processing. Offline processing time can as the basic unit, basic use batch mode, this way can make the most of I/O system. The last line of processing time average levels in minutes or hours, the processing model and no special requirements, can choose according to requirements, but in practice, use the batch mode.

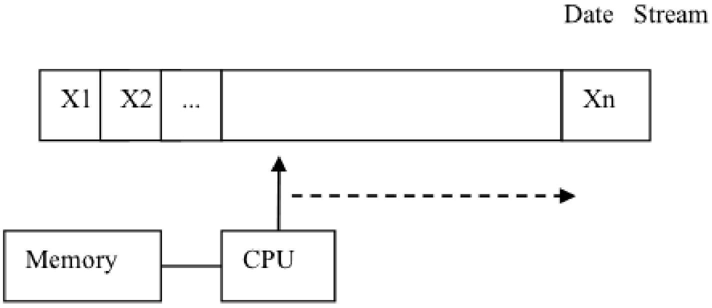

Fig. 1 Basic data stream model

## 3. The challenges for the data security

It is these very different with traditional data management features, making the data security is facing new challenges of the era of large data.Information technology, the security and privacy has always been a key problem. Big data era, along with the increase in data, data face more serious security risks, the traditional data protection methods is not suitable for large data, large data security in the face of challenges.

The big data era of big data privacy.data privacy issues include two aspects: on the one hand, is a personal privacy protection, with the development of data mining technology, the user can't detect, personal interests, habits, physical characteristics such as privacy information can be more easily access; On the other hand, even if obtained permission from the user privacy data in the process of storage, transfer and use, also have leaked risk. Big data analysis ability, leading to seemingly simple information can be excavated the privacy, so in the face of big data privacy protection will become the new proposition of The Times.

Tata qualityData quality affects the use of big data.low quality of the data is not only a waste of the transmission and storage resources, and even can't be used.Factors on the quality of the data has a lot of, in the process of generation, collection, transmission and storage, can affect the data quality, data quality specific displays in: accuracy, completeness, redundancy and consistency. Although there are a lot of measures to improve data quality, but the data quality problem is impossible to completely eradicate. Therefore, it is needed to study a kind of method to automatic detection of data quality, and can repair some quality problems of the data.

The big data security mechanism.Big data in terms of data size and data types, brought challenges to data encryption. In view of the small and medium-sized encryption method in performance before can't meet the requirements of large data, need to study and efficient large data cryptography. According to different structure of the structured, semi-structured, and not knotStructure data In addition, under the mode of multi-tenancy, needs on the premise of guarantee efficiency, realize the tenant isolation, confidential dataSex, integrity, availability, controllability and traceability.

Big data applications in the field of information security. Big data not only brings challenges to information security, has injected new impetus for the development of information security. For example, through to the invasionDetection system of the log file for big data analysis, can find potential security vulnerabilities and advanced sustainability threatened (advancedpersistent -- kyoui, APT). In addition, the information such as virus, vulnerability characteristics and attack were also more likely to is mastered by big data analysis.

## 4. Big data security and privacy protection key technologies

Data released anonymous protection technology. For large structured data in the data (or data), data released on condition of anonymity to protect key technology is the core to realize its privacy and basic means, is still in the stage of continuous development and improvement. In big data scenario, data protection issues are more complex: anonymous attacker can get data from multiple sources, not just the same source. For example, the applications of Netflix, it was found that the attacker can pass the data as opposed to a publicly available imdb ratio, so as to identify the target in Netflix account. Accordingly get the user's politics and religious belief and so on (by the user to watch the history and analyzing some film reviews and ratings), such problems remain to be further research.

Social network anonymous protection technology. Social network in the typical anonymous protection requirements for anonymous user identity anonymous and attributes (also known as anonymous), in the release of the data hiding the user's identity and attribute information; As well as the relationship between user anonymity (also called anonymous), on the relationship between the release of the data hiding users. The various properties of the attacker tried to use the node (degree, tag, some specific connection information, etc.), to identify the identity of the node information in the graph. Important problem facing social network anonymous scheme is that the attacker may infer by other public information anonymous users, especially if there is a connection between users. According to the existing social structure on the hierarchy of the crowd to recover and speculation; For weibo type of composite relationship between social network analysis and prediction; Limit based on the random walk method, speculated that the probability of different connection relations, and so on. Research has shown that the agglomeration characteristics of social networks for the accuracy of the prediction method has an important influence, social network and local connection density growth, cluster coefficient increases, the connection to further enhance the accuracy of prediction algorithm. Therefore, the future of anonymous protection technology should be effective against such speculation attack.

The data watermark technology. Digital watermarking is the identity information embedded in imperceptible way within the data carrier and does not affect its using method, see more at copyright protection of multimedia data. There are also some watermarking scheme against the database and text files. By the disorder of data, such as dynamic characteristics, embed the watermark in the database, document method and multimedia carrier are quite different. The basic premise is redundant information in the data or can tolerate a certain accuracy.

Data source technology. As mentioned earlier, data integration is one of the early stage of the big data processing steps. As the data source of diversification, it is necessary to record the data source and transmission, calculation process, provide auxiliary support for mining and decision-making in the late. As early as the former of the concept of big data According to the source (DataProvenance) technology is widely in the field of database research. The basic starting point is to help people to determine the source of all the data in data warehouse, such as understanding what they are made in the table which operation data items, therefore the correctness of the result can easily check, or with minimal cost to update the data.

The role of mining. Role-based access control (RBAC) is the current widespread use of an access control model. Through the assigned roles for the user, the characters related to privilege set, realize the user authorization, and simplify permissions management. Early RBAC authorization management adopts the pattern of "top-down", namely according to the position of enterprise set up the roles. When applied to large data scenarios, face to a lot of artificial participation role division, authorization problem (also called role engineering). Later researchers begin to pay close attention to the "bottom-up" mode, namely according to the existing authorized "user - object", the role of automatic extraction and optimization design algorithm, called role mining. In simple terms, the role of is how to set up reasonable. Typical work includes: in the form of visual, through user permissions two-dimensional figure sort merge the way of character extraction; through the subset enumeration and clustering method to extract the role of formal methods, such as; are based on the formal semantic analysis, through the methods of extracting roles to more accurate hierarchical mining. In general, dig to generate the optimal algorithm of minimum character set time complexity is high, the more belongs to NP complete problems. And therefore researchers focus on the heuristic algorithm in polynomial time.

Risk of adaptive access control. In big data scenario, security administrator may lack enough professional knowledge, can not accurately for the user to specify its can access the data. The risk of adaptive access control for this kind of scenario discussion more a method of access control.

## 5. Summary

Big data era has arrived. The correct use of big data has brought great convenience to people's life, but at the same time also to the traditional approach to data management has brought great challenges. In this paper, the large data related research results at home and abroad in recent years a comprehensive review and summary, this paper introduces the basic concept of big data, analyzes the present situation of big data management in detail, and face the challenge.This article introduces will give big data system background of the research provides a good reference.

## Acknowledgments

2014 Wuhan Municipal colleges and universities teaching and research project (No.2014088)

## References

[1] 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)

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

[3] Liu Yu-Chao, Ma Yu-Tao, Zhang Hai-Su, et al. A method for trust management in cloud computing: Data coloring by cloud watermaking. International Journal of Automation and Computing, 2011, 8 (3):280-285

[4] Gantz J, Reinsel D. Extracting value from chaos. IDC iView, 2011:1-12

[5] Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. Communications of the ACM, 2008, 51 (1): 107-113