Research Strategies for Privacy Security and Protection in The Context of Big Data

Yinuo Wang *

Portland Institute, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China

* Corresponding Author Email: p23000109@njupt.edu.cn

Abstract. The use of big data has brought intelligence and convenience to people, but it has also raised concerns about security and privacy issues. The article lists five main processes of big data processing and analyses the potential risks involved, which may adversely affect the data. These include issues such as internal and external threats, missing and duplicated data during data acquisition. To address these challenges, the article describes three privacy protection strategies, namely encryption techniques, privacy-preserving algorithms, and access control. These methods help to effectively ensure user privacy and provide security for big data, which meets the need for space saving and protects user privacy. With the continuous development of big data technology and the expansion of the application market, the issue of privacy security is becoming more and more prominent, and there is a need to establish sound privacy protection techniques and raise the awareness of privacy protection in the whole society.

Keywords: Risk point; Big data processing; Cryptography; Privacy-preserving algorithms; Access control.

1. Introduction

With the continuous development of science and technology, society has entered the Internet era. Big data has also gradually entered people's vision, and big data has a significant impact on people's lives. Big data makes people's lives smarter and more convenient at the same time, its security and privacy issues are also increasingly concerned [1].

This paper analyses the security and privacy protection of big data. Firstly, this paper introduces the five main processes of big data processing and analyses the risk points, including the threats to data security from internal and external sources in the process of data acquisition, as well as the problems of missing data and data duplication. Secondly, this paper introduces three strategies in privacy protection, including encryption techniques, privacy-preserving algorithms, and access control.

2. Process and risk point analysis of big data processing

Big data can quickly collect, store, process, transmit, and analyze huge amounts of data when faced with it so that the potential value of the data can be explored. The big data processing process is mainly divided into five steps. They are data collection, data storage, data processing, data transmission, and data analysis. The big data processing flow is shown in figure 1.

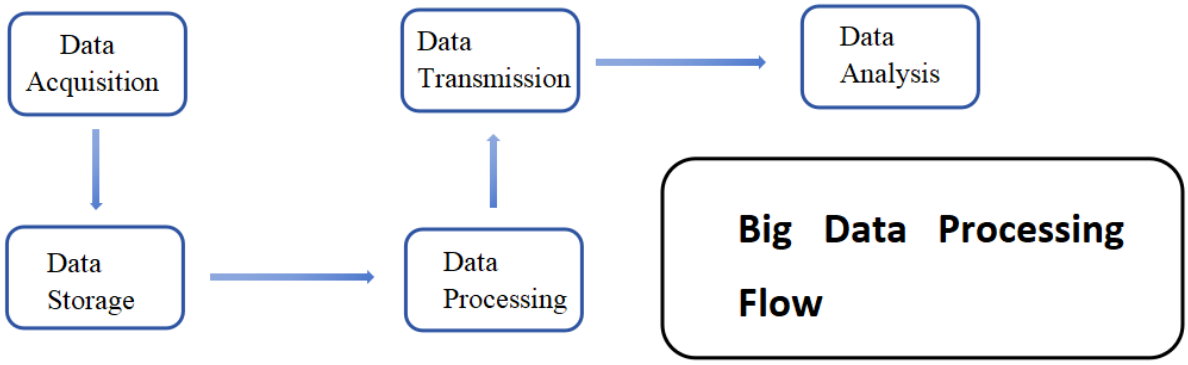


Fig. 1 Big Data Processing Flow (Photo/Picture credit: Original)

2.1. Data Acquisition

Data collection is also known as data acquisition. RFID technology, sensors, social networks, and various software will generate a large amount of data every moment, which is complex and structured, through data acquisition tools, such as Cloudera's Flume, Facebook's Scrible, etc., to obtain specific data from external systems and transform it, and then stored in the system. The key to this technique is how to obtain specific, more accurate, and effective data according to the different needs of users and software respectively [2]. With the continuous development of social science and technology, the way of data acquisition is also changing, and the innovation and development of intelligent data acquisition technology is also promoting the development of the whole big data field. In today's big data field, the intelligent perception layer plays an extremely important role, which includes the network communication system, data sensing system, intelligent identification system, software, and hardware resource access system, etc., which can carry out specific identification, positioning, tracking, transmission, conversion, and regulatory processing according to the collected data. Through these steps, the data becomes easier to manage and analyze. At the same time, there are some potential risk points in the process. In the process of data collection and storage, there will be the risk of data leakages, such as unauthorized by others or the system, and there will be the problem of privacy leakage. Not only that, in the process of collection, it will be accompanied by the problem of missing data and data duplication.

2.2. Data Storage

After data collection in different fields, a huge amount of data is created. This makes how to effectively store and manage these data an urgent problem. In this era of big data, data storage has come to tens of billions of megabytes, and at the same time, data in the process of storage is also facing the risk of being illegally accessed, stolen, and tampered with, so that the data loses its original integrity. Hardware failure and human error may also lead to the risk of data loss. With the development of technology, today's cloud storage services provide support for this, cloud storage services can quickly expand the storage capacity and achieve data backup and recovery, to ensure data security, scalability, and high performance.

2.3. Data Processing

Data processing has a non-negligible position in the field of big data. The application scope of big data processing is very wide, covering not only industries such as finance, medical care, and e-commerce, but also playing an important role in system engineering and automatic control in social life. It allows raw data, such as large-scale, diverse, and high-latitude data, to be analyzed, organized, calculated, and edited in a process of processing for subsequent data transmission and data analysis. Data processing involves data cleansing, integration, and quality assessment to extract useful information from massive amounts of information and discover potential patterns and trends to support innovation and decision-making processes. The risk points are mainly found in the technical

challenges. Big data processing requires intensive computing resources and professional technical support, which may lead to system crashes or data loss when these technical difficulties are not solved effectively.

2.4. Data Transmission

Big data transmission has become a crucial part of today's modern technology. With the continuous expansion of data volume and the increasing complexity of data structure and requirements, effective data transmission means have become a key part of the current big data processing system. In the process of data transmission, it is necessary to ensure the integrity and security of data and prevent data loss and data leakage, at the same time, it is also necessary to consider the speed of data transmission and the bandwidth consumption in the transmission process. In today's world, based on blockchain technology The data transmission of big data is a bridge for the flow of data in various links, and it is an important link to ensure the integrity and security of data.

2.5. Data Analysis

In today's era of digitization and informatization, big data analytics has become an important engine for driving social development and economic growth. Big data analytics deals with massive amounts of data and then mines the laws and information behind the data to maximize its use. Big data analytics helps enterprises better understand market demand and optimize products and services through data collection, cleaning, storage, processing, and mining. It provides important decision-making services for enterprises to achieve precision marketing and personalized services. Big data has gradually become an important tool for accurate decision-making, improving efficiency, reducing costs, and promoting the development of society in the direction of intelligence and digitalization. There is a risk of misinterpreting data when analyzing data, and incorrectly correlating different data can lead to wrong conclusions and decisions. It can also cause problems such as data discrimination and misleading information.

3. Privacy and Security Protection Strategy

The challenges of big data storage are becoming more and more prominent nowadays. One of the core issues is the balance between data compression and user privacy protection. System designers are often caught in a dilemma between compression and encryption to save space and protect user privacy at the same time. The Privacy and Security Protection Strategy is shown in figure 2.

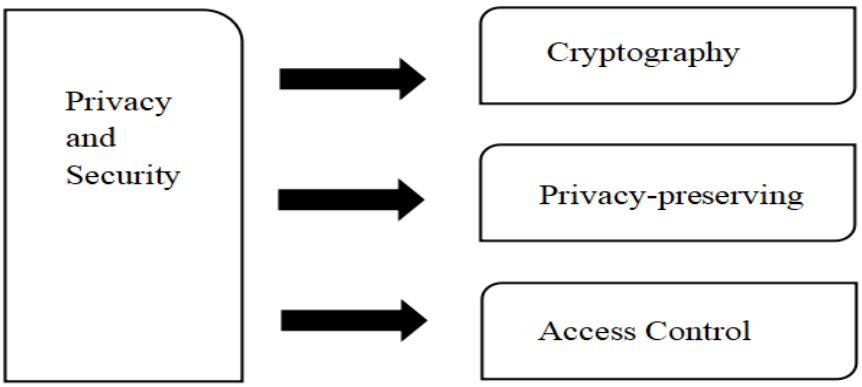


Fig. 2 Privacy and Security Protection Strategy (Photo/Picture credit: Original)

3.1. Cryptography

Big Data Encryption refers to the methods and techniques for encrypting data in a big data environment [3]. The original big data is converted into the corresponding ciphertext form by specific

encryption algorithms, thus ensuring that the data will not be tampered with and stolen by unauthorized visitors during transmission and storage. There are two compromise methods between big data compression and encryption, namely convergent encryption, and homomorphic encryption [4].

Convergent encryption is a very forward-looking solution. By using the Hash value of the data as the encryption key, the ciphertexts derived from the data encrypted by different users are identical, so this method can not only achieve the sharing of data among different users but also effectively de-emphasize the data, thus meeting the requirements of privacy protection and the convenience of data sharing, thus saving space.

In addition, homomorphic encryption is widely used in cloud storage systems. The uniqueness of this technology is that it allows encrypted data to be processed without the need to decrypt the data first, which helps to protect the privacy of the data, and the result of the processing is the same as the result of processing unencrypted data after decryption. This means that the user's private information will not be leaked when data is searched and processed in the cloud, thus realizing a win-win situation for both efficient data management and privacy protection. By introducing these advanced technological means, the cloud storage system achieves the dual goals of saving storage space and efficiently utilizing data resources while safeguarding data privacy. This balanced design concept not only improves the reliability and practicality of the cloud storage system but also brings users a more secure and convenient data management experience.

3.2. Privacy-preserving Algorithms

Differential Privacy Techniques play a key role in today's data-driven society, facilitating the balance between data privacy and data analysis [5]. In the era of big data, the collection and processing of personal data has become the norm in all industries, however, the ensuing privacy leakage and abuse problems are becoming increasingly prominent. The emergence of differential privacy techniques provides an important way to address this challenge.

The differential privacy-preserving algorithm is a technique used to protect the privacy of an individual, which is based on the principle of analyzing the data by adding a certain amount of random noise, which makes it difficult to infer the specific information of a single individual data, thus ensuring the accuracy of the data analysis results while protecting the user's privacy [6]. This approach fundamentally changes the limitations of traditional privacy-preserving techniques that only protect private data through anonymization. By ensuring the accuracy and validity of data processing results while minimizing the risk of leakage of individual private information, differential privacy achieves a balanced development of data privacy protection and data analysis.

Differential privacy technology is widely used in areas including, but not limited to, healthcare, financial services, intelligent transport, social networks, and so on. In the medical field, anonymized patient data is used for medical research while protecting patient privacy; in the financial industry, differential privacy technology can be used for the analysis of transaction data, ensuring that the user's private data is not leaked; and in the field of intelligent transportation, differential privacy can safeguard the privacy of vehicle trajectory data while supporting road congestion analysis. As an innovative privacy protection technology, differential privacy technology not only provides a new guaranteed mechanism for data security and privacy but also provides reliable technical support for data-driven decision-making and social progress.

3.3. Access control

Access control is the process of managing and controlling user access to resources in a computer system. It mainly includes determining user identity, verifying user identity, authorizing user rights, and monitoring user access. After the system identifies the user's identity through information such as username and ID, it further verifies whether the user is a legitimate user through passwords and biometrics. Then the system assigns appropriate access rights to the user based on his/her role or attributes. Monitoring user access means that the system records and monitors user access and

operations on resources to detect potential security problems or risks in time and safeguard system security and data integrity.

In the overall framework of access control, the importance of the data privacy management module cannot be overstated. The privacy of user data can be effectively protected through techniques such as data encryption, data desensitization, and data masking. Data encryption utilizes symmetric encryption and asymmetric encryption algorithms to safeguard the confidentiality of data during transmission and storage. Data desensitization, on the other hand, reduces data sensitivity by means such as data generalization and protects user privacy while ensuring data availability. Meanwhile, data masking technology can partially hide or replace data to further enhance the level of data privacy protection. On the other hand, the function of the access control module is also crucial. Through the role-based access control (RBAC) and attribute-based access control (ABAC) and other access control policies can accurately manage and control users' access rights to data, thus ensuring that only legally authorized users can access and operate the data [7-10].

RBAC technology achieves the management of user access rights through the correlation of roles and permissions, while ABAC technology is more flexible, providing access control based on user attributes and environmental conditions, which improves the adaptability of the system to dynamic access control needs. In addition, the security authentication module ensures the legitimacy and trustworthiness of user identities through various identity verification mechanisms, effectively preventing the intrusion of malicious users. The audit log module, on the other hand, records and monitors the accesses and operations of user data, which provides an important security guarantee for the system operation and helps detect potential security threats in time and take corresponding measures.

4. Conclusion

The analysis of risk points within the big data processing framework is systematically conducted through a detailed examination of its constituent phases: data collection, storage, processing, transmission, and analysis. This discussion then transitions to a focused examination of key strategies in privacy security protection, specifically highlighting encryption technologies, privacy protection algorithms, and access control mechanisms.

As big data technologies continue to advance and their application domains expand, the challenge of safeguarding privacy has become increasingly acute. This situation necessitates not only the development of robust privacy protection technologies but also the enhancement of societal awareness regarding the importance of privacy security. Such dual focus is essential for maintaining the integrity of privacy in an era dominated by extensive data utilization.

References

[1] Zhan Yuqi. A preliminary study of data security risk and response in cloud computing. Network Security Technology and Application, 2024(04): 80-82.

[2] Li Xinping. Research on Big Data Processing Process and Storage and Management Technology. Computer Programming Skills and Maintenance, 2023(03): 97-100+133.

[3] Yang J. E. A double-layer structure blockchain financial data encryption system based on cloud computing. Information Technology, 2022, 46(10): 97-101.

[4] Zhu Yurui. Exploration of privacy protection technology in the context of big data. Network Security Technology and Application, 2021(09): 76-77.

[5] Song Jiasan. Research on statistical inference of big data based on differential privacy. Shandong University, 2024.

[6] Xie Yaqi, Yang Geng. Differential privacy preserving algorithm for polynomial regression. Computer Technology and Development, 2022, 32(08): 103-109+128.

[7] Wang Jun, Wei Wenliang. Construction and Research of RBAC-based Information Technology Resource Sharing Platform. Electronic Testing, 2021(10): 70-71.

[8] Cheng Xiangran, Chen Xiyuan, Zhang Bin, et al. Attribute-based access control policy model. Computer Engineering, 2010, 36(15): 131-133.

[9] Xiong Houren, Chen Niyuan, Fei Xiaofei, et al. Hybrid Extended Access Control Model Based on Attributes and RBAC. Computer Applied Research, 2016, 33(7): 2162-2169.

[10] Liu Qiangbing. Data privacy protection and access control in cloud computing environment. Information Record Material, 2023, 24(11): 207-209.