ISSN 2616-5767 Vol.6, Issue 4: 42-47, DOI: 10.25236/AJETS.2023.060407

Research on Key Technologies of Personal Information Security Protection in Big Data

Jiana Bi[1,*], Yonghong Guo[1], Ning He[1], Shuang Wang[2]

[1]*School of Software and Big Data, Changzhou College of Information Technology, Changzhou, China*
[2]*School of Physical Education, Bohai University, Jinzhou, China*
544099426@qq.com
[*]*Corresponding author*

Abstract: *Big data is composed of linear associated data and nonlinear data. Big data is characterized by multiple types, large data scale, fast data processing, high data value and low value density. As more information can be obtained through big data mining or web crawler, the leakage of users' personal information occurs from time to time, which seriously endangers the stability of social order. Therefore, protecting personal information from infringement in the era of big data has become an urgent problem to be solved. This paper constructs the technical framework of the life cycle and protection of personal information in big data, and summarizes the key technologies such as authorization, access control, security audit, traceability audit and data desensitization. In addition, in-depth research on these key technologies will be conducted to provide complete solutions for the protection of personal information in big data, crack down on illegal behaviors of user information trading, and create a positive network security environment.*

Keywords: *Big Data; Personal Information Security; Key Technology; Authorization; Access Control; Security Audit; Traceability Audit; Data Desensitization*

1. Introduction

Big data technology emerges at the historic moment in massive information processing, greatly promotes the interconnection between information and provides users with accurate information services [1]. Big data is composed of linear associated data and nonlinear data. Big data is characterized by multiple types, large data scale, fast data processing, high data value and low value density. Big data sources mainly include data processed by computer information systems, data generated by Internet activities, data generated by mobile Internet activities and data collected by other data acquisition equipment. In the big data environment, through the linear and nonlinear data generated by the network activities of individual users, users' age, occupation, behavior rules and interests can be clearly analyzed. Especially with the application and popularization of e-commerce and mobile network, the address, contact information and bank account information of individual users can also be obtained by means of big data mining or web crawler. As a result, the pressure of personal information security management is increased, and the leakage of user personal information occurs from time to time. Inexplicable sales calls, fraudulent phone calls and bank deposits are stolen and other illegal behaviors are repeatedly prohibited [2]. Therefore, it is necessary to study the key technologies of personal information security protection in big data to provide protection for personal information security.

2. Life Cycle and Protection Technology of Personal Information in Big Data

Personal information is a typical type of sensitive data. In terms of data security, boundary effect and the characteristics of data do not exist, and it is feasible to propose corresponding technical measures by analyzing the security risks in the data life cycle. The data life cycle includes the stages of data acquisition, data transmission, data storage, data processing, data exchange and data destruction. In the data acquisition stage, the data collector shall prevent unauthorized collection and illegal data acquisition by the personal information subject, and the personal information owner shall prevent unauthorized collection. In the data transmission stage, the implementation of security policies during personal information data transmission should be monitored to prevent sensitive data leakage and

ISSN 2616-5767 Vol.6, Issue 4: 42-47, DOI: 10.25236/AJETS.2023.060407

identity denial by both parties involved in data transmission that may occur in the transmission process [3]. In the data storage phase, data storage security is ensured. Data desensitization or encryption can be used to ensure data storage security and minimize sensitive data content. In the data processing stage, only legal personnel can see legal data, and unauthorized users cannot see unauthorized personal information data. In the data exchange stage, the transfer and sharing process of data should have clear logging and approval process, and understand the data flow process. In the data destruction stage, in order to avoid data leakage caused by incomplete information elimination, the data should be forcibly erased or desensitized, if the original data cannot be recovered. Protection technologies are required for networks at each stage of the data lifecycle. The relationship between life cycle stages and protection technologies is many-to-many, that is, one stage corresponds to multiple technologies, and one technology provides security protection at multiple stages. The corresponding relationship between the life cycle of personal information in big data and protection technology is shown in Figure 1.

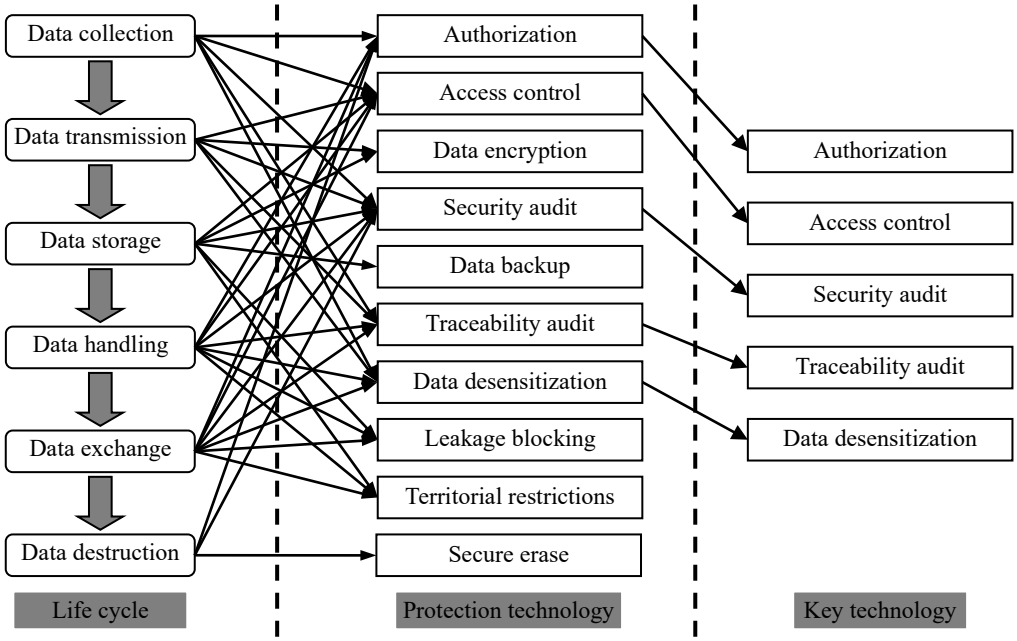

Figure 1: Life cycle and protection technology of personal information in big data

For the correspondence between the life cycle of personal information in big data and the protection technology shown in Figure 1, each data protection technology corresponds to one or more life cycle stages. "Security audit" corresponds to six life cycle phases. Access control" corresponds to five lifecycle phases." Authorization", "Traceability audit" and "Data desensitization" correspond to four lifecycle phases, respectively. Leakage blocking" and "Territorial restrictions" correspond to three life cycle phases respectively. Data encryption" corresponds to two lifecycle phases." Data backup" and "Secure erase" correspond to one lifecycle phase, respectively. Data protection technologies corresponding to more than or equal to 4 life cycle phases are defined as key technologies. Therefore, the key technologies of personal information Security protection in big data studied in this paper include "Authorization", "Access control", "Security audit", "Traceability audit" and "Data desensitization".

3. Key Technology of Personal Information Security Protection in Big Data

3.1 Authorization

A license is a business deployment in which an individual allows another individual to use its product, service, or certain information. An individual may be an individual, a unit, an organization, a company, an enterprise, etc. The specific content of the license is based on the consensus of both parties on the basis of equality and voluntalism, and also needs to adapt to the relevant provisions of the legal content [4].

ISSN 2616-5767 Vol.6, Issue 4: 42-47, DOI: 10.25236/AJETS.2023.060407

There are three kinds of licenses commonly used in the field of big data application: First, authentication technology. In the computer and network system to confirm the identity of the operator, so as to determine whether the user has access to a certain resource and use rights, so that the access strategy of the computer and network system can be executed reliably. Prevent attackers from impersonating legitimate users to obtain access to resources, ensure the security of the system and data, and the legitimate interests of authorized visitors. The second is the kerberos protocol, which provides authentication services to client/server applications through a key system. The implementation of the authentication process does not depend on the authentication of the host operating system, does not require trust based on the host address, does not require physical security of all the hosts on the network, and assumes that packets transmitted over the network can be arbitrarily read, modified, and inserted. Third, multi-tenancy, which is a software architecture technology that shares the same system or program components in a multi-user environment while still ensuring data isolation among users.

3.2 Access Control

Access control policies are security measures against possible illegal operations, and are the basis for blocking, controlling, and alerting illegal operations. Different from the access standards of traditional control methods, big data access control has certain flexibility and can quickly adjust the standards in different access environments to meet the sensitivity requirements of the access control model [5]. The access restriction policy specifies the time range and IP address segment that the platform allows access to. All access within the range that is not allowed is denied. Access control is implemented by agents. Proxy services are a very important part of the big data platform and the only way for users to access the big data system. The authentication, control, authentication, and operation records of the access are all completed by agents.

API access agent is a component that actually sends operation requests to the big data platform. All operations of users on the big data platform in the portal are forwarded by the agent, and the agent determines whether the operation is authorized before execution. Without authorization, the operation is not performed and the user is prompted through the portal [6]. At the same time, the agent analyzes whether the content of the operation contains sensitive data. If the content contains sensitive data, the agent blocks or permits the request according to the preset sensitive data access policy. The agent service receives the user permission information configured by the management platform, judges the user permission according to the permission information, and controls the user's access to the big data platform. Users access the big Data platform through the unified access portal, and the unified view forwards the specific operation request to the proxy service. If the user has no operation permission, the proxy service blocks access requests. If the user has relevant operation rights, the agent service forwards the access operation to the corresponding big data platform component and returns the access operation result to the user.

3.3 Security Audit

Big data audit means that audit institutions follow the concept of big data, use big data technology methods and tools, and use huge amount of economic and social operation data with dispersed sources and diverse formats to carry out cross-level, cross-region, cross-system, cross-department and cross-business in-depth mining and analysis, so as to improve the ability of audit to discover problems, evaluate and make macro analysis. Compared with data audit, big data audit uses more heterogeneous data sources, uses more complex and advanced technical methods, and has more keen and profound insights into data.

In the era of big data, security audit has developed fundamental changes, from sampling analysis to full data analysis, from finding causality of things to using correlation of things, from pursuing accuracy of data to improving efficiency of data use [7]. In the process of security audit using big data technology, it is necessary to analyze and collect a large amount of data to ensure the quality of heterogeneous data processing. Using big data technology to build a security audit system, the following points have been optimized [8]: Improved storage, collection, and analysis capabilities. Improved ability to handle unstructured data. More in-depth analysis of historical information data, from a large number of data mining valuable information for users.

The log-based audit method mainly combines Hadoop component logs and metadata for combined analysis. In the situation that the big data platform does not perceive, restore the user's operation. The corresponding log file collector is deployed on the server of the big data platform without in-depth

ISSN 2616-5767 Vol.6, Issue 4: 42-47, DOI: 10.25236/AJETS.2023.060407

analysis of the log data and imposes relatively small load on the server. However, this approach depends on the accuracy of logging.

Network traffic data-based audit restores user operations by collecting, filtering, and analyzing network traffic data entering and exiting the server. Network traffic data can be collected and analyzed by deep packet detection technology, which is a traffic detection and control technology based on application layer. When the IP packet, TCP or UDP data stream passes through the bandwidth management system based on DPI technology, the application layer information of OSI seven-layer protocol is reorganized by reading the IP packet load content, so as to get the content of the entire application program. Then traffic shaping is performed based on the management policy defined by the system. Because this kind of mode requires network packet capture and analysis, the implementation is very difficult, but also increases the server load.

3.4 Traceability Audit

Traceability audit is to find out the evolution process of data products to form metadata that can accurately express data characteristics and object history, so as to provide structured guidance for the analysis and understanding of complex data in data-intensive systems. Traceability audit focuses on data source detection, data creation and dissemination processes in data-intensive systems. By recording the derived process information and workflow evolution process of data products, metadata containing historical information of objects and accurately expressing data characteristics can be formed, and structured data can be presented to promote information disclosure and derivation of operable information. It is of great value to analyze and understand complex data in big data system.

Traceability information collection technology includes three aspects: traceability capture, traceability elements and traceability model. Traceability capture focuses on how to collect traceability information from the object system, including capture intensity, capture method, collection time and data version control. After collecting reliable data from different terminals, traceability elements mainly solve what key data should be included in traceability information, and summarize data from different sources to form standardized data. A traceability model is a formal description of normalized data. With the help of the traceability model, the correlation between data can be mined from the normalized data in sequence, the mapping between normalized data and structured data can be realized, the structured data can be presented and persisted, and the information disclosure and derivation of operable information can be promoted.

3.5 Data Desensitization

Data desensitization refers to the extraction of large amounts of unstructured, random and uncertain data with the analytical ability of computers, and the accurate identification and estimation can be achieved by obtaining relevant information [9]. Data desensitization is to transform some sensitive information through desensitization rules to achieve reliable protection of sensitive privacy data. The essence is to collect massive original samples in large databases and screen them. In the case of customer security data or some commercially sensitive data, under the condition of not violating the rules of the system, real data should be transformed and provided for test use, including personal information such as ID number, mobile phone number, card number and customer name. Data desensitization is accomplished by the data desensitization management platform, and the architecture is shown in Figure 2.

Data desensitization technologies can be broadly divided into two categories [10]: static data desensitization and dynamic data desensitization. Static data desensitization is generally performed on non-real-time access data. Before desensitization, the desensitization strategy is set uniformly and the desensitization results are imported into a new file or database. The data desensitization tool carries out full scanning of static data and uses the sensitive data signature database formed after sampling to conduct matching desensitization of data. Dynamic data desensitization generally refers to the desensitization of the data or data stream accessed. The desensitization rules can be modified in real time. The desensitization is only for the data passed by the data desensitization tool, and the desensitization results are displayed to the user. On the basis of static desensitization, it explores timely desensitization technology, extends structured data, and explores dynamic desensitization of unstructured data, including large database platforms and text files.

Data desensitization will lead to increased operation and maintenance costs. Enterprises need to set realistic desensitization goals according to the actual situation. Desensitization technology includes

ISSN 2616-5767 Vol.6, Issue 4: 42-47, DOI: 10.25236/AJETS.2023.060407

sensitive information fields, sensitive information names, sensitive levels or sensitive types, etc., which must be determined during data desensitization to ensure better service for customers [11]. Desensitization strategies are referred to as desensitization rules, norms, desensitization methods and desensitization restrictions. First of all, users need to develop desensitization rules for sensitive data, which can be achieved with the help of data and user global, as well as individual Settings. Desensitization specifications actually require users in desensitization work, must follow the relevant specifications and relevant laws, and ultimately make the management more convenient, or further improve the security.

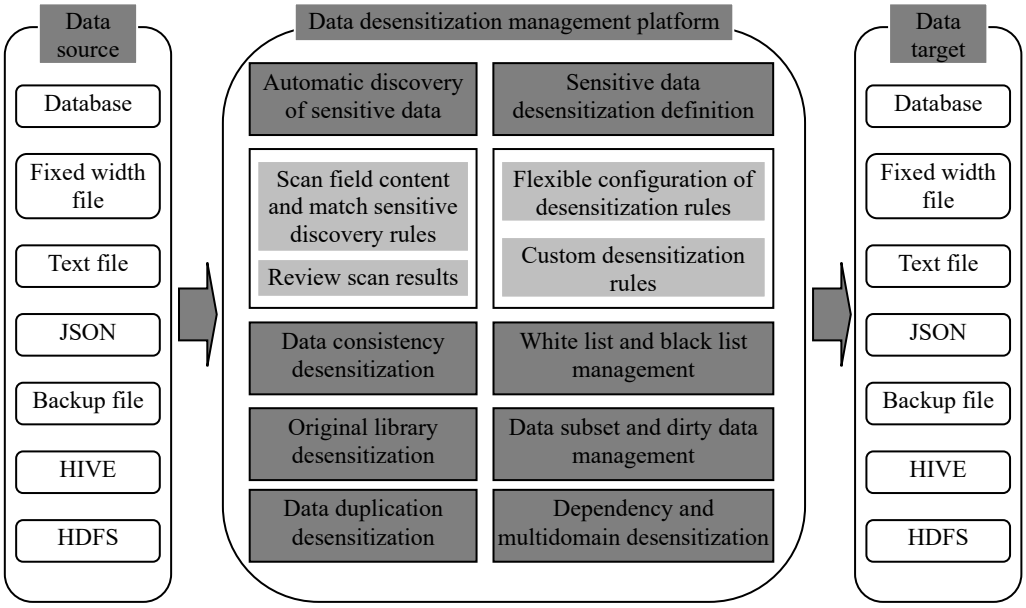

Figure 2: Data desensitization management platform architecture

4. Conclusions

The era of big data has come, and the realization of digitalization is the trend of the fourth Industrial revolution. In the new era of complete data of personal information, security issues are becoming increasingly prominent with the development of big data technology. The more individual activities in the network world, the more data will be generated, and the more resources and resource values can be mined in the personal information set. Research on the key technologies of personal information security protection in big data, improve the network security factor, ensure the security of users' personal information, crack down on illegal behaviors of user information trading, create a network security environment clean and positive, and promote the implementation of the strategy of cyber power.

Acknowledgements

This work is supported by Jiangsu innovation and entrepreneurship doctor project: Research on personal information security protection protocol in big data; Natural Science Foundation of Changzhou College of Information Technology (CXZK202004Y): Research on the model and service discovery mechanism for the future distributed social network; Scientific Research and Development Center for Colleges and Universities of the Ministry of Education, China University Innovation and Research Fund (2021LDA06008): Steel product defect detection based on big data and industrial vision.

References

[1] J. Li, J. Zhang, L. Wang. Research on Personal Information Security Protection from the

ISSN 2616-5767 Vol.6, Issue 4: 42-47, DOI: 10.25236/AJETS.2023.060407

perspective of Big Data[J]. Satellite TV & IP Multimedia, 2019, 16(18): 75-76.
[2] W. Zhao, M. W. Shui. Research on Personal Information Security Protection Measures under the background of Big data[J]. Computer Programming Skills & Maintenance, 2018, 25(07): 86-87+93.
[3] Y. M. Bao, P. J. Song, G. Q. Yu, et al. Research on Technical architecture of Personal Information Protection based on Data security[J]. China Security & Protection, 2022, 13(Z1): 42-48.
[4] L. P. Gu, S. Fan. The Licensing Management of Data Flow and Data Migration[J]. Frontiers of Data & Computing, 2019, 10(02): 51-60.
[5] W. Y. Li. Big Data Access Control Method Based on Machine Learning[J]. China Computer & Communication, 2021, 33(13): 30-32.
[6] Y. J. Lu, Y. H. Li. Research and lmplementation of Access Control Method for Big Data Platform[J]. Journal of Information Security Research, 2016, 2(10): 926-930.
[7] W. J. Liu, Q. J. Rong, Y. Wang. Analysis on Internet security audit in the era of big data[J]. Auditing and Finance, 2022, 39(02): 50-52.
[8] R. Tao, S. J. Zhang. Research and analysis of security audit system under Big data technology[J]. Chinese informatization, 2020, 17(04): 90-91.
[9] Z. T. Chen, B. Jin, M. Y. Guo, M. M. Li. Research on data desensitization technology of big data platform[J]. Communication & Information Technology, 2022, 52(04): 31-33+42.
[10] D. Tang, J. Gu, K. Y. Zhang, X. Gu. Development trend of data desensitization technology[J]. Secrecy Science and Technology, 2021, 22(04): 4-11.
[11] B. He. Research on data desensitization technology of big data platform[J]. Information Recording Materials, 2020, 21(09): 134-135.