Research and design of a data desensitization system

Wei Li¹ and Chuan Li*¹
¹School of Control and Computer Engineering, North China Electric Power University
*Corresponding author’s e-mail: 120192227057@ncepu.edu.cn

Abstract. With the extensive implementation of SG186 Project and SG-ERP Project, State Grid has accumulated abundant data resources, and the hidden information in these data resources can create great value. But at the same time, the security of massive data has become more important. From the perspective of system design, workflow and key technology of power information desensitization system, this paper analyzes and studies the sensitive level and data type of sensitive information, matches the corresponding desensitization strategy and completes a series of operations of desensitization task, so as to realize the protection of sensitive data in power information under big data environment.

1. introduction
At present, with the development of smart grid, various monitoring devices in the power industry have obtained a large amount of data information, with an average increase of 46TB data per month. The data generated is four orders of magnitude higher than that in the past, and the amount of data increases rapidly, from TB level to PB (1000TB) or even to ZB (1000PB)[1]. These data contain great value. However, the leakage of sensitive information will bring huge risks and losses. How to ensure the security and privacy of the data is still an urgent problem to be solved while using the power big data to extract value.

Data desensitization refers to the bleaching, masking or deformation of sensitive information in data to protect private data or business sensitive data. Although the traditional firewall technology protects the data security to a certain extent, it also limits the scope of application of data, which is not conducive to the application of data mining. Using data desensitization technology, we can formulate corresponding data desensitization scheme for data users with different permissions, so as to ensure that the sensitive information in data can be protected while making full use of data information to create value[2-3].

In the traditional research of data desensitization, more emphasis is put on the implementation and improvement of the algorithm, lack of systematic design. According to the characteristics of electric power information data, this paper carries out the algorithm configuration of authority decision, sensitive identification and desensitization rules for sensitive information, and proposes a desensitization system which can be applied to different business scenarios[1,4].

2. Data desensitization
Data desensitization refers to changing the display of some sensitive information to users through some desensitization rules. These sensitive information are generally user ID number, mobile phone number, bank card number, salary, address, etc. In the process of data desensitization, it is necessary to retain the original characteristics of the data and transform the data. The basic principle is that people
can not infer the original data from the desensitized data. For desensitized data, it is generally required that the data still conform to people's authentication rather than meaningless string. For example, for an 11 digit mobile phone number, the desensitized display should still be 11 digits, which is consistent with people's cognition of mobile phone number. For the data desensitization system, it is generally required to complete the automatic desensitization, because in the actual production environment, the generation of data is continuous, for the new data, desensitization is also an important demand.

3. System structure design

3.1. System framework
As shown in Figure 1, according to the characteristics of highly structured power information and fixed types of information contained in the data, the overall framework of the system adopts the form of building desensitization proxy server between the database and the client. All the interaction processes between the client and the data storage server go through the proxy server, the requested content is processed by the desensitization server and then sent to the client to ensure that the client can only get the desensitized data.

![Figure 1. Overall framework of the system.](image)

3.2. Structure design of desensitization system
The desensitization system is composed of five modules, which are user management module, request forwarding module, sensitive information discovery module, data desensitization module and data output module. The data desensitization module includes desensitization rule management and desensitization algorithm management. Data through the four layer structure of data desensitization system, through the management of data resources, desensitization rules and desensitization algorithm configuration management, desensitization scheme is formulated and desensitization operation is implemented. The system has the characteristics of convenience, adaptability and scalability. And it can audit user desensitization operation to ensure data security. The structure is shown in Figure 2.

![Figure 2. Structure design of desensitization system.](image)
1) User management module
The module is responsible for user account registration, deletion, user information verification and user rights management, and provides support for data dynamic desensitization. The module establishes the corresponding relationship between the permission level and the sensitive data, encapsulates the permissions of different users and the sensitive range of different permissions. After the user information is verified, it will keep connected with the desensitization server, and the user management module will provide the current user’s permission range to the sensitive data discovery module.

2) Request forwarding module
The request forwarding module is responsible for receiving the data request from the client, obtaining the client information, modifying the target request address and the target address in the replacement request, and achieving the acquisition of the original data and the forwarding of the desensitized data through the data desensitization server.

3) Sensitive data discovery module
According to the pre-configured sensitive data discovery rules, this module discovers the sensitive data fields and types in the data table returned by the data storage server. According to the user authority scope provided by the user management module and the pre-set sensitive data classification table, the sensitive information is labeled with the sensitive level. Complete the task of data sorting before data desensitization, prepare for the development of desensitization strategy, and improve the efficiency of desensitization.

In order to effectively prevent resource waste and unnecessary trouble caused by over protection, the classification table of sensitive data should be configured in the module based on the in-depth analysis of business scenarios, and the boundary of sensitive information classification should be clear.

Data desensitization module
In the desensitization algorithm management, the commonly used traditional desensitization methods such as hiding, hashing, replacement, offset, interval floating, truncation and encryption are built in, and the desensitization algorithms such as k-anonymity algorithm, l-diversity algorithm and t-closeness algorithm are also included for desensitization service call [5-6].

In the desensitization rule management block, different desensitization methods are configured according to the data type and sensitivity level of sensitive information, including single desensitization method or multiple desensitization algorithms, with the goal of achieving the best desensitization effect. After the desensitization strategy is formed, the data desensitization module completes the desensitization task according to the desensitization strategy to form the desensitized data.

4) Data output module
The module uses the client information obtained by the request forwarding module to send the desensitized data formed by the data desensitization module to the client.

3.3. Work flow of desensitization system
The workflow of the whole desensitization system is as follows: firstly, the client establishes the connection and sends the data request, so that the desensitization proxy server can obtain the original data that needs desensitization. Then, the original data is identified with permission decision and sensitivity level, and the desensitization scheme is obtained according to the identification and desensitization rule configuration, and the desensitization processing is carried out, and finally the desensitized data is output to the client. As shown in Figure 3, desensitization data is formed by desensitization of original data, which generally includes four steps: data application, sensitive data identification, desensitization execution and controlled output.

1) The client logs in to the desensitization proxy server, establishes the connection after the authentication, sends the data request, and forwards the data request to the data storage server through the desensitization proxy server.
2) After the desensitization proxy server obtains the original data, it identifies the sensitive information according to the user’s permission range and desensitization data level, and identifies the data type of the sensitive data. Then according to the pre-configured desensitization rules, the desensitization strategy was formulated.

3) According to the desensitization strategy, the desensitization task was performed and the desensitization data was generated.

4) The desensitization data is sent to the client.

3.4. Data desensitization process

The data desensitization process is completed by the sensitive data discovery module and the data desensitization module (as shown in Figure 2). The logical location is between the client and the data storage server. The task is to desensitize the data returned by the data storage server to ensure that the data received by the client has no sensitive information. The data desensitization process mainly
includes: desensitization target confirmation, desensitization strategy formulation, data desensitization realization and other operations. The specific process is shown in Figure 4.

4. summary
The problem of data security cannot be ignored. The leakage of sensitive information will cause losses to both enterprises and users.

The desensitization system proposed in this paper separates the desensitization work and modularizes the desensitization system. When faced with high concurrency, the desensitization agent cluster can be deployed to achieve high availability and load balancing, providing users with data resources and protecting sensitive information in the data. The desensitization strategy and algorithm in the desensitization system are scalable and flexible, and can adapt to different desensitization needs. Through the research and configuration of desensitization rules, the desensitization data can keep the relationship characteristics of the original data and retain the value of data mining.

Reference
[1] Yang, T., Xue, Z., Shi, Y. (2016). Research on key technology of privacy preservation based on k-anonymity. Information Technology: 6-9.
[2] Li, Z. K., Ding, M. J. (2015). Big data environment data security research. Information Security & Technology: 7-9.
[3] Fang, Z., Cheng, Q., Pei, X. B. (2019). Data security protection of electricity marketing information system. Computer and Modernization: 111-116.
[4] Ke-He, W. U., H. Zhu, Wei, L. I., Cui, W. C., Zhang, X. L., Cheng, R. (2019). An improvement of t-closeness technology based on sensitive information metric. Information Technology: 5-9.
[5] Xing, Y., Zhang, B., Mao, Y. (2017). Application of data desensitization in massive data system. Telecommunications Science: 8-14.
[6] Zhang, P. D., Ming, L. I. (2016). The hurst index estimation of dna series based on fgn. Information Technology: 11-13.