Research on Privacy Protection of Power Users Based on Big Data Desensitization Technology

Xin Jiang, Ran Chen, Dongyin Chen, Yuxiang Cai, Jia Liu, Jiemin Zhang
Information and Communication Branch of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350001, China

Keywords: Big data, Power user information, Privacy protection

Abstract: With the advent of the Internet and big data era, the value of data has become more and more significant, and its role in various fields has become more and more important. It has also spawned a new research field, namely data privacy and security. Grid companies have collected and stored massive amounts of electricity customer data, and it has become particularly urgent to solve the problem of how to ensure the privacy and security of customer data during use. By analyzing the characteristics and application scenarios of power customer data, we propose a power customer data privacy protection system construction plan covering business scenario classification, privacy classification, protection strategy and protection method library, and build an active privacy protection mechanism for power customer data to achieve Active recommendation of power customer data privacy protection strategies and technical methods. The actual case shows that the established power customer data privacy protection system has good practicability.

1. Introduction

Compared with the current status and implementation prospects of data applications in the power industry, the current power customer data privacy protection mechanism has the following shortcomings: Different types of data protection goals are unclear and protection strategies are single. There is a lack of differentiated protection for different business scenarios and privacy protection goals, and it is impossible to implement privacy protection with appropriate strength for different types of data in combination with data application characteristics. One-sided attention to security and lack of consideration for data availability. At present, most power data is protected based on the principle of non-disclosure, and the protection strategy is one-sided, resulting in a prominent contradiction between data availability and security. One-sided dependence on desensitization and encryption, especially lack of adaptability to the big data environment. The existing privacy protection methods unilaterally rely on access control, encryption and data desensitization. Privacy and security protection can indeed be achieved in a closed system, but multi-source data fusion analysis has become the norm in the big data environment, and isolated data desensitization or encryption is difficult to resist multi-source data cascading attacks, and there is a hidden danger of privacy leakage.

2. Data Desensitization Concept

Data desensitization is to change its value while preserving the original characteristics of the data, thereby protecting sensitive data from unauthorized access, and at the same time performing related data processing, which can preserve the meaning and validity of the data while maintaining the data Security and compliance with data privacy regulations. With the help of data desensitization, the information can still be used and is associated with the business without violating relevant regulations and avoiding the risk of data leakage. The first is how to identify sensitive data, determine the definition of sensitive data, and study the dependence of sensitive data. The application is very complex and complete. It is very difficult to know where the sensitive information is and to know which data refers to or applies the sensitive data. Static desensitization is generally used in non-production environments. In situations where sensitive data cannot be
stored in a non-production environment, the production data is converted through a desensitization program, so that the correlation between data content and data can meet the needs of testing and development problem investigation, and data analysis, data mining and other analysis activities. Dynamic desensitization is usually used in the production environment. Sensitive data is desensitized only when it is accessed by authorized individuals, and corresponding desensitization methods can be implemented according to the strategy. The difference between static desensitization and dynamic desensitization lies in whether real-time desensitization is performed when sensitive data is used. According to statistics, the current global data desensitization technologies are mainly put into practical applications in three categories: data distortion technology, data encryption technology and anonymization restriction release technology (including: k-anonymity, L-diversity, data suppression, data disturbance and differential Privacy).

For the dynamic desensitization system, the data source-related configuration is generally pre-configured by the administrator user and does not need to be involved by ordinary users; while for the static desensitization system, it must be configured on demand according to the characteristics and requirements of each task. This process is due to Limited by business requirements, it is generally difficult to achieve automation or configuration-free. The automatic identification of sensitive data is generally divided into two parts: (1) identification of sensitive data; (2) identification of the relationship between sensitive data in different tables and different fields. The sensitive data discovery and association recognition of traditional data desensitization systems are generally achieved through manual configuration and regular expression matching. The accuracy of its recognition mainly depends on whether the rule setting of the regular expression is accurate and reasonable. But in general, the accuracy and performance of traditional regular expression matching sensitive data are often not compatible: the simpler the regular matching rule setting, the better the system performance, but simple rules will lead to relatively poor recognition rate; and the setting is relatively complicated. Although the regular rules can improve the recognition rate to a certain extent, it often has a greater impact on system performance, and the experience and technical requirements for manual configuration are relatively high, which is difficult for ordinary business personnel. In this regard, it is urgent to apply artificial intelligence technology to the identification stage of sensitive data and association relationships, which can effectively solve the pain points that traditional regular recognition rules cannot have both performance and accuracy, and can also save professionals from maintaining regular recognition rules. Achieve the realization of configuration-free and automatic functions, and bring value enhancement for users.

3. Analysis of Data Desensitization Technology under the Background of Big Data Intelligence

Sensitive data recognition is an important prerequisite for data desensitization. For example, pattern recognition is used to realize automatic recognition of desensitization information. For different types of data, the detection methods of its sensitive features will be different. After the feature data is obtained through the training set, combined with the pattern matching of sensitive information and the importance of the source business system, the sensitivity level value is manually set for sensitive data grading. Perform feature extraction on the preprocessed target data, and match the extracted feature value with the feature value of the sensitive data. When the match is hit, the system automatically records the sensitivity level value of the current sensitive data. Finally, the misclassification is corrected through the recognition quality assessment, and the sensitive data that has not been recognized is supplemented. Perform classification training research on sample data and metadata of some business systems, and finally classify and establish sensitive data sets to achieve accurate identification of sensitive data.

For structured data, especially data sets with very complex relationships between different data elements, there is often a corresponding relationship between a field in the same data table and another field. Generally speaking, the corresponding relationship before and after data desensitization should not be destroyed, otherwise the use value of the field will no longer exist. Generally speaking, when data statistics and reference quantities are needed, the data relevance requirements are relatively high. Applying machine learning algorithms to research on the
identification and maintenance of association relationships between large-scale data.

At present, traditional data desensitization systems often require users to configure desensitization strategies one by one for all identified sensitive data. In addition to certain requirements for the technical and business understanding of operation and maintenance personnel, this method also requires a large amount of manual operation when faced with massive amounts of data. If it is planned to be based on the above-mentioned artificial intelligence sensitive data automatic discovery and database association relationship recognition technology, it is necessary to study an intelligent adaptive desensitization algorithm based on user usage model. When dealing with static desensitization scenarios, the system automatically selects the appropriate algorithm for desensitization after the user has limited the characteristics of the desensitization results, eliminating the problem of user configuration policy field by field; and when dealing with dynamic desensitization scenarios, the system learns the user configuration strategy for a period of time, the automatic strategy configuration and configuration-free function of sensitive data can be realized. This significantly improves the ease of use and availability of the desensitization system, and reduces the deployment and application costs of enterprise-related systems.

4. Application Analysis of Power User Privacy Information Protection

The main purpose of data exchange and sharing within power companies is to evaluate management analysis, audit processing, and training. There is still a risk of privacy information leakage in data exchange and sharing. Through desensitization processing, the security of various information data can be effectively improved, and the security of user information can be guaranteed. Development and testing scenarios are analyzed by importing original information data during R&D and testing. If the information data is directly affected, problems such as information leakage will be caused. In this regard, the original information data should be desensitized and then applied rationally. External release scenarios the most critical form of data business in electric power companies is external data transfer, which involves various customer access requests and transaction information data. These outsourcing processes involving users’ personal information should strengthen privacy leakage risk control. The characteristics of related factors such as private information are desensitized through big data technology.

Based on sensitive data, it can be divided into static and dynamic data desensitization types. Static is abbreviated as SDM, while dynamic data desensitization is called DDM. 3.2.1 Dynamic data desensitization the application of dynamic data desensitization in the power business will not affect the original data in the database. It mainly realizes the desensitization by inputting and outputting data, thereby reducing the hidden danger of information leakage. This form of desensitization is mainly to carry out dynamic access and systematic retrieval of production data, mainly combined with reasonable application of permissions. In the process of dynamic data desensitization, first enter the data access request in the user terminal of the electric power company, and then the system uses the data access method to realize the conversion, and analyzes the query statement under the support of SQL to analyze whether it is safe and compliant. The unsafe statement requires feedback warning. The rewriting process of the safe SQL statement can be realized through the desensitization rule; at the same time, the rewritten SQL statement can be used for data retrieval and query, and the desensitization engine can be used for processing. After processing, the obtained data will be fed back to the user terminal in time as a result. Strengthen the processing of key links through the power operating system, audit based on requirements, and save it in the security log. Static data desensitization Static data desensitization can realize online desensitization processing and can be applied in a non-production environment. After the desensitization is completed, the development test can be implemented based on the operation process. In the static desensitization process, according to different power enterprise business demand scenarios, data use request parameters are proposed, according to the data desensitization engine, comprehensive desensitization processing, through desensitization processing of key information in the original database, storage and batch applications are realized Export and other related operations. In the process of data application and export, security approval processing must
be done to ensure that it meets the requirements. Static data desensitization realizes batch processing through desensitization according to offline requirements, truly realizes internal analysis and data sharing, and provides basic support for business research and development and the development of typical business activities.

5. Conclusion

Data desensitization technology is based on the reasonable analysis of data desensitization, and realizes deformation optimization by analyzing sensitive information, and then achieves the purpose of hiding sensitive data. Data desensitization is to achieve the purpose of shielding data sensitive information through data desensitization technology. The data obtained after processing can maintain the original format and attribute characteristics, thereby providing technical support for the stable development of applications and desensitized data.

References


