Data LifeCycle Management Method Research Based on Traceability Technology

Zhihua Cheng[1], Lingchao Gao[1], Zhouchun Lei[1], Zhenyu Chen[1], Xiangzhou Chen[1], Jiakai Wang[1], Jiasong Sun[2]

1 State Grid Corporation of China Big Data Center, Beijing, China
2 E. E. Department, Tsinghua University, Beijing, China
zhihuacheng1990@126.com

Abstract: Manual management is the bottleneck of data asset sorting. The traceability has been one of the important contents in the data center construction. The relationship between traceability tables was found manually and data inconsistency is easy to occur in existing systems. This paper proposes the data Traceability Method for power-grid data resource life cycle management. This vectored compressed data traceability method achieves accurate traceability with low computational cost by constructing an inverse function, which can record the changes of data in the whole life cycle and identify multiple sources. The traceability performance test on the existing data of data center shows that the method is robust to the volume and layer changes of data, and the traceability time meets the practical application requirements. Compared with other data traceability technologies, this method is more suitable for life cycle management of large-scale business data in power grid industry.

Keywords: Data Traceability; Whole Life Cycle Management; Compression Traceability; Data Asset Carding.

I. Introduction

With the rapid development of informatization, the size of the information volume within the enterprise is becoming larger and larger, and the management business is complicated and complicated. The rapid introduction of new business has further caused the dynamic adjustment of the table model. In order to better cope with the big data planning of the State Grid, it is necessary to further enhance the intelligent analysis capabilities of data resource services and fully realize the standardization and intelligence of data management and data scheduling. Establish the concept of "speak with data, use data management, use data decision, and use data innovation." Starting from a unified data center, analyze and mine data value, and fully realize the application and promotion of data business.

Data traceability technology has been widely used in many scientific fields [1]. Traditional database technologies provide effective means in terms of data sharing, data and application independence, maintaining data consistency and integrity, and data security, and can support online transaction processing well. However, the tasks of extracting, retrieving, and querying from a large amount of data involve a large amount of data for decision-making. Traditional database management systems have been difficult to meet the requirements of modern databases for data management due to their limitations. With the continuous deepening of data analysis and the rapid development of software and hardware technologies, data tracing has begun to become a cutting-edge research direction in data management [2, 3]. Product life cycle data plays an important role in the process of product production, development, maturity and demise [4].

Through the full life-cycle management of power grid data, basic tasks such as access, query, and modification of all stored data can be realized. Feature extraction, intelligent judgment, and reasoning can be performed in real time. A variety of service methods can be used through customized data

Copyright © (2020) Francis Academic Press, UK DOI: 10.25236/ictmic.2020.002

analysis services. Utilize and access the data platform to provide statistical basis for business personnel's analysis and decision-making, provide accurate and detailed statements and conclusions for data center personnel's scheduling and management work, and to fully realize the full life cycle and entire process of power grid data Management provides a theoretical paradigm and framework, promotes intelligent and lean power grid data management, and promotes the further development of smart grids.

This article proposes a method for tracing the life cycle management data of the grid data resource business. It is mainly used to record the evolution information and evolution processing content of the original grid business data throughout the life cycle, and to identify multiple sources and prove the data source and location. Credibility. Through this data tracing technology, the credibility of network nodes can be established, and the tracing metrics are used in the data analysis system to construct a prediction model of key attributes of the data, which is helpful to infer the data attributes based on its tracing. Compared with other data source tracing technologies, this method is more suitable for the full life cycle management of large-scale business data in the power grid industry. Not only can it alleviate the problems in the multi-level secure database, but it can also retain each change in the data flow path. Logging to ensure traceability of results, as well as data recovery, replay, auditing and evaluation.

The structure of this paper is as follows: first introduce the intelligent extraction method of data resource characteristics, then give the basic architecture and principle of data traceability technology, and propose a method and implementation scheme of data traceability for the whole life cycle management of power grid data, and then apply the scheme to the data center grid business The data set is tested and finally summarized and prospected.

II. Data resource feature intelligent extraction

At present, the main data resource feature extraction uses mainstream methods of machine learning or deep learning [6,7,8]. Feature extraction using machine learning requires expertise in feature extraction methods, such as HOG / SIFT features in computer vision, MFCC features in the speech domain, and so on. After selecting these features, further use machine learning methods for classification and recognition, such as MLP, support vector machine, hmm, k-means, etc. The main method for deep learning feature extraction is to design the data and extract the hierarchical representation of the neural network, as shown in Fig.1, which is also the most commonly used method for processing large amounts of data. When deep learning is used for feature extraction of data resources, its neural network structure, such as a restricted Boltzmann machine or auto encoder, consists of a bunch of feature extractors.

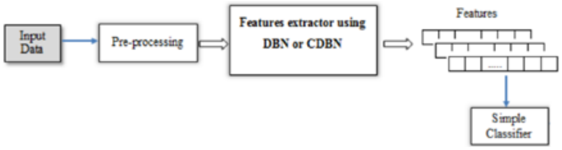

Figure 1. Example of data feature extraction using deep learning

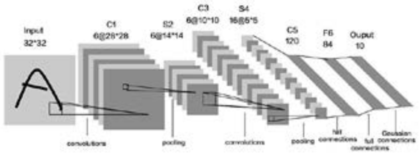

Figure 2. Convolutional neural network structure

These mainstream deep learning structures are used for unsupervised feature learning, such as convolutional neural networks, deep belief networks, and convolutional Boltzmann machines. As shown in Fig.2, by using the back-propagation method, abstract features that can better represent the characteristics of data are learned iteratively at each layer of the network. The unsupervised pre-training method is further used on the convolutional neural network, and it has been well verified in applications such as handwriting recognition.

III. Basic architecture and principle of data traceability

With the development of Internet and cloud services, the volume of data is increasing, and many data centers manage more than one million tables. The rapid introduction of new services has further intensified the dynamic adjustment of the table model. At the same time, there are also irregular development processes and non-uniform standards. , Incomplete source code, etc. Especially in the identification of data resources, due to the fluidity of the data, it is necessary to consider the type, responsibility, source and path of these data, and it is necessary to rely on a large amount of manual combing, which has low work efficiency and cannot guarantee accuracy. When the source system redefines the data or changes the information, the data resource directory cannot be synchronized in real time. Manual identification is required, which often causes the resource directory information to be disconnected from the actual environment. For all tasks of accessing, querying, and modifying all stored data in the data center, it can perform feature extraction, intelligent judgment, and reasoning in real time. Through customized data analysis services, it uses multiple service methods to access and access the data platform for analysis of business personnel. And decision-making work provides a statistical basis. Only when sufficient and effective data traceability is achieved, can the machine learning automatically recommend the background data mode associated with the front-end interface to provide training data and form key data for the life cycle management of the data center data table.

Data tracing is an emerging research field. It was born in the 1990s. It is defined as recording the evolution information and evolution processing content of the original data throughout its life cycle (from generation, transmission to extinction), and reproduce the history of the data according to the tracking path Status and evolution process to achieve data history traceability. There are two types of traceability information exchange models: parallel and master-slave. The parallel information exchange model does not have a central database. It establishes its own database at each link of the supply chain to record the source and flow of the product. When the downstream link understands the relevant information of the product, it needs to search from the database of the upstream link, and it cannot search beyond. The master-slave information exchange model uses a central database as the main database, and the data link database is a slave database. The slave database records the product information in this link, and regularly uploads the information closely related to the traceability to the master database [9]. Data traceability can be classified according to the direction of information activities. Backward traceability refers to the ability to find the source and characteristics of a product according to one or more given criteria at each point in the data chain [10].

Data traceability is used to record the evolution information and evolution processing content of the original data throughout the life cycle, as well as identify multiple sources and prove the credibility of the data source and location. Data traceability can establish the credibility of network nodes, and use the traceability metrics in the data analysis system to construct a prediction model of key attributes of the data, which can help infer data attributes based on its traceability. Furthermore, it is also necessary to store and manage sensitive data of different security levels through a database system, while maintaining data security through autonomous access control or mandatory access control mechanisms, to provide multiple levels of granularity, ensure consistency and completeness, and implement inference control. , Prevent sensitive aggregation, conduct hidden channel analysis, support multi-execution concurrency control and other principles, and solve various key issues of multi-level secure databases. In addition, by keeping a log record of each change in the data flow path, the results can be traced, and data can be recovered, replayed, audited, and evaluated.

IV. Data tracing method for power grid data full lifecycle management

With the increasing scale of power grid data centers, data security management has become a top priority. It is urgent to start from the "data life cycle management" idea of the data center and adopt a data life cycle management model to realize the full life cycle management of data from generation to use, migration, cleaning, and destruction. Data life cycle management can effectively control the data scale of the production system and improve the efficiency of data access, thereby improving the overall efficiency of the system operation.

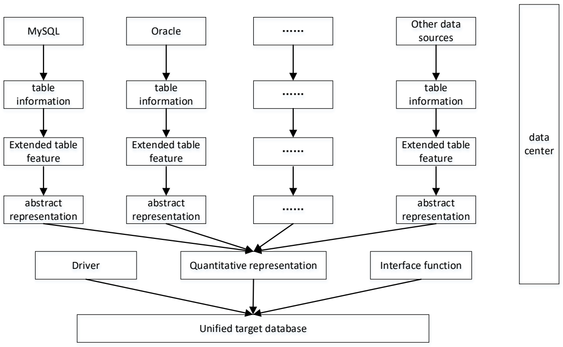

Figure 3. Structure chart of data source lifecycle management for data traceability

The core of a general data tracing method is to record all the data of the application state change, and then realize the final state of the application in a data-driven manner. The traditional development model modifies data at the business logic processing layer and stores the business status in the business database; business development based on data traceability sends the modified data to the database after the business logic processing is completed and then the corresponding data processor receives the data Process, generate new data, and modify business databases. The traditional model pays more attention to state storage. It cannot trace the change process of the application. It is inflexible to expand the business and cannot guarantee consistency in multiple business scenarios. The data traceability enables the entire data to be traced through the storage database. It is convenient to add new logic to all business State data can be recovered by backtracking. Data tracing splits the entire business logic from the traditional logic layer, and separates the business logic into two parts: the business data source and the data controller. The business data source is responsible for generating application data, and the data controller is responsible for data processing. The controllability and stability of new data generation; data tracing adds data storage and data controllers, all services are distributed, and each layer needs to recover data from failures.

The purpose of data tracing technology is to reproduce the historical state and evolution of data according to the tracking path, to achieve the tracing of historical data archives. This paper proposes a set of quantitative source tracing technology based on data tables. It traces the power source data for the whole life cycle management, and sorts out and controls the correspondence information between the front-end business functions of the system and the back-end database tables. The main methods include: first obtaining the user's target data table information; then further expanding the easy-to-query table name characteristics based on a variety of external information; then sampling the data in the table to generate a more abstract representation of the data; and finally synthesizing the target data table Data features and content features to achieve a vectorized representation of the output. When operating different types of heterogeneous databases, you need to call the functions supported by the database access interface, dynamically link to the driver, and then form a unified target database through data conversion. Data traceability information can be passed to the target database through this method. . Each source tracing system follows the VTML Schema mode, adopts XML files as the information carrier for data exchange, and implements data exchange by reading and writing XML files.

Because users have different requirements for traceability accuracy, data traceability can be divided into vectorized fine-grained traceability and coarse-grained traceability. Coarse-grained source tracing can use the traceability labeling method to trace the historical state of the data by recording and processing related information, and use the labeling method to record some important information of the original data, such as background, time, provenance, etc., by viewing the labeling of the target data. Obtain the traceability of the data. Use the annotation method to mark the source in advance and carry the traceability information to complete the data traceability model. It is further possible to use compressed source tracing information to achieve data source tracing. First, separate the source data from the metadata, and establish the association between the two through indexing; then store the independent labeled content in columns, changing the traditional row storage method. The same content needs to be stored only once when storing, and the other only needs to save the row. No.

Compared with traditional data source tracing technology, this vectorized compressed data source tracing method inverts query results by constructing an inverse function. Data tracing. This quantitative traceability technology helps the business end to obtain the maximum value at the lowest cost in each stage of the data life. It sorts out and controls the correspondence information between the front-end business functions of the system and the back-end database tables, and provides support for more accurate data table full lifecycle management.

V. Experiment and test results

The data model is the key to data tracing technology. According to the model, the general steps and basic ideas of data tracing can be determined initially. The quantitative data source tracing model for the full life cycle management of power grid data in this paper considers the heterogeneous distribution characteristics of the data. By introducing a heterogeneous layered model, data tracing information is stored in different databases to form a heterogeneous database with tracing information. Converged into a unified target database through database interfaces and data conversion tools. At this time, the target database carries data traceability information. The path of the reverse process of this process can realize various operations of data tracing, such as: data tracing, information evaluation, process reproduction, etc., to complete the task of data tracing of heterogeneous data. Fig.4 shows the traceability diagram of the result data generation process in the fault maintenance query example.

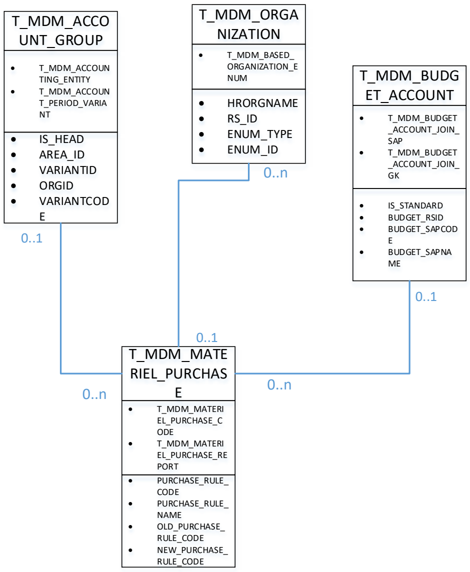

Figure 4. Example of data traceability diagram

The traceability tracing method used in this article is to use the labeling method to track the specific source of a single data item. Through the inverse function reverse query method, the inverse query is reversed through the inverse query or the construction of the inverse function, or the reverse derivation is based on the conversion process. The process of tracing the results back to the original data. Its implementation is mainly based on the traceability files generated during the execution of the model, and the traceability files establish a connection through an association identifier. The traceability tracing method uses a recursive method to track data items. The difficulty lies in the need for additional Storage space.

The key of the reverse query method is to construct an inverse function. The quality of the inverse function construction directly affects the query effect and the performance of the algorithm. Compared with the annotation method, it is more complicated, but requires less storage space than the annotation method. At the same time, you need to trace the source, such as (Zhang modified, 1, 5, 8, 1001), tuples 1, 5, and 8 record the same label content (such as: Zhang modified), then only the rows of tuples need to be recorded (1, 5, 8) without storing the same content (such as: Zhang modification), this will save a lot of unnecessary storage space. When the same information exceeds a predefined data block, this expansion method needs to be referenced to recursively achieve the same content compression to achieve the purpose of saving space.

VI. Conclusion

The traceability analysis technology based on the intelligent extraction of data resource characteristics is an important and basic work of the data center, which has significant application and promotion value. This paper proposes a method for tracing the source life cycle management data for power grid data resource services, and performs a tracing performance test on the existing data in the data center. The results show that the single data tracing track changes the data volume and number of layers under this method. Robust, traceability time meets actual application needs. The next step is to perform research on performance and speed optimization for multi-model and multi-data source tracing.

Acknowledgment

This paper is supported by the State Grid Big Data Center Self-built Technical Project: Research on intelligent data resource catalog construction and business feature Traceability Technology (5500/2019-83001B).

References

[1] Cheney J, Chiticariu L, Tan W C. Provenance in databases:Why, how, and where[M]. Now Publishers Inc, 2009.

[2] Cao J, Jarvis S, Saini S, et al. Gridflow: Workflow management for grid computing[C].Cluster Computing and the Grid, 2003. Proceedings. CCGrid 2003. 3rd IEEE/ACM International Symposium on. IEEE, 2003: 198-205.

[3] Yu G E, Zhao P, Di L, et al. BPELPower—A BPEL execution engine for geospatial web services[J]. Computers & Geosciences, 2012, 47: 87-101.

[4] Zhao Fei. Master Data Management Based on the Whole Life Cycle: Detailed Explanation and Practice of MDM [M]. Tsinghua University Press, 2015.

[5] Song Xinyi. Research on detection Data Management for product LifeCycle [D]. Master's thesis of Shenyang University of Aeronautics and Astronautics, 2018.

[6] Ali A, Sharma S. Content based image retrieval using feature extraction with machine learning[C]. International Conference on Intelligent Computing & Control Systems. 2018.

[7] Yu Y T, Lin G H, Jiang H R, et al. Machine-learning-based hotspot detection using topological classification and critical feature extraction[C].Design Automation Conference. 2013.

[8] Zhu P, Isaacs J, Bo F, et al. Deep learning feature extraction for target recognition and classification in underwater sonar images[C]. 2017 IEEE 56th Annual Conference on Decision and Control (CDC). 2017.

[9] Xinting Y , Jianping Q , Chunjiang Z , et al. Construction of information description language for vegetable traceability based on XML and its application to data exchange[J]. Transactions of the Chinese Society of Agricultural Engineering, 2007, 23(11):201-205.

[10] Aung M M, Chang Y S. Traceability in a food supply chain: Safety and quality perspectives[J]. Food Control, 2014, 39(1):172-184.