Research and Application of Power Grid Data Blood Relationship Analysis

Ming Liu(✉), Shijin Liu, Lihua Sun, Hongyu Ding, Chuanrong Lu, Yang Zhang, and Tiancheng Qian

Nanjing NARI Information and Communication Technology Co., Ltd., Nanjing, China
{liuming,liushijin,sunlihua,dinghongyu,luchuanrong,zhangyang16,qiantiancheng}@sgepri.sgcc.com.cn

Abstract. State Grid Corporation of China implements the digital transformation strategy, integrates the resources of various data islands in various power fields under the traditional model, and quickly forms the capabilities of data governance and data services. The implementation process will encounter problems such as difficulty in obtaining big data metadata, data quality, and data traceability. This paper proposes a power grid data lineage analysis method, which performs metadata management, data extraction, data transformation, data calculation, etc. on the data at the source data end (especially for power grid big data) to generate bloodline data. See the data source link for the model object. The results show that the blood relationship analysis function can be beneficial to the tracking and positioning of data problems, to the diversified analysis of data, to the data governance of the big data platform, and to effectively solve the pain points of data display in the national network data center.

Keywords: power grid data center · power grid big data · data lineage · data governance · visualization

1 Introduction

In recent years, the State Grid Corporation of China has begun to implement the digital transformation strategy. With the large-scale digital construction, the scale of power grid data has gradually expanded, and the amount of data has increased year by year. The company has entered the era of big data. In recent years, companies in various Internet provinces have been building big data platforms to drive business with data [1].

State Grid Corporation adopts the idea of data middle platform [2], integrates the resources of various data islands in various power fields under the traditional model, and quickly forms the capabilities of data governance and data services, so as to meet the needs of horizontal cross-professional and vertical data sharing, analysis and mining between different levels and accommodation needs. However, many difficulties and challenges will still be encountered during the specific implementation.

(1) It is difficult to obtain metadata using a unified approach [3]: It is difficult to obtain accurate and complete metadata, because the most important metadata for blood

G. Guan et al. (Eds.): ICBDSS 2022, AHCS 8, pp. 277–286, 2023.
https://doi.org/10.2991/978-94-6463-064-0_32

relationship analysis is obtained in different ways. Especially for big data platforms, the big data platforms built by various power companies are closed-source, and it is difficult to obtain metadata related information. For some electric power big data, even if metadata is obtained, it may only obtain technical metadata information, lacking effective business metadata, which is not helpful for blood relationship analysis.

(2) The display form of data objects is complex: Usually, the data middle station adopts a data layered architecture [4]. According to the data middle station capability architecture, the power grid divides the middle station into the source layer, the sharing layer, and the analysis layer. The flow of data between levels is intricate and complex. As the scale of the project grows, the amount of data increases, and the data relationship becomes more and more complex, data managers cannot clearly and accurately view the data flow process of the entire professional and overall system from a global perspective.

(3) Data quality traceability is complex [5]: In the process of data access, data integration, and data calculation, there may be some inappropriate human processing, resulting in data quality problems eventually, but at this time there is too much data Layer processing, data managers cannot effectively locate the wrong location in a short time.

(4) The data structure is diverse and cannot be visualized [6]: In addition to traditional structured data, the power grid also has a lot of unstructured data, collected measurement data, E-files and message data of specific protocols. Therefore, it is necessary to support the high-performance display of massive complex data and a good interactive mode through the visualization ability of blood relationship analysis.

2 Data Lineage Function Structure

The construction of the power grid data center adopts a unified data resource, service platform and standard specifications. In data access, data integration, and data analysis, data quality problems may occur in each link. If there is no unified data quality processing in the system, accurate and effective information cannot be provided in subsequent business applications. Using the traceability of data blood relationship can solve the problem of data positioning and realize data traceability.

Data lineage analysis is a technical means [7], which is used to track the entire data processing process, so as to find any data object as a starting point, all relevant metadata and the relationship between these metadata objects.

Data blood relationship analysis is divided into four parts: data blood relationship extraction, data blood relationship analysis, data blood relationship display, and data blood relationship application (Fig. 1).

2.1 Data Bloodline Extraction

Data bloodline extraction is responsible for collecting the original data of the power grid, generally in the following ways.

Data replication: refers to the incremental capture mechanism based on database logs, which realizes real-time synchronization of data from the source database to the data center.

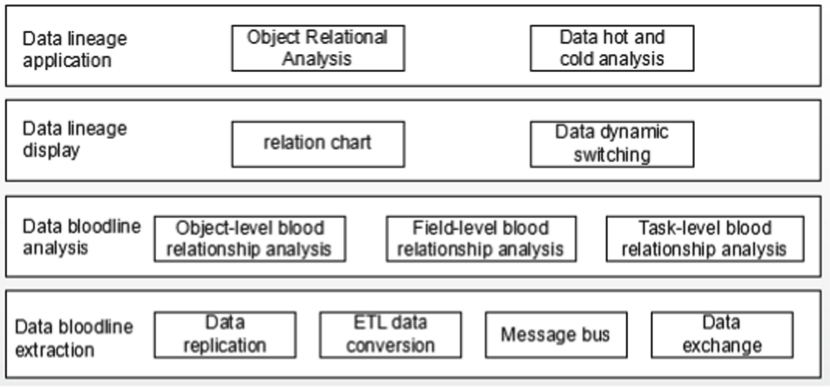

Fig. 1. Data lineage function structure

ETL data conversion: refers to extracting, converting, and finally loading and saving data from the source business system to the data center through ETL tools.

Message bus: refers to the real-time subscription and real-time consumption of source business system data.

Data exchange: refers to the data exchange used to realize the data exchange between the headquarters and the provincial and municipal companies.

2.2 Data Bloodline Analysis

Data bloodline analysis is responsible for analyzing the extracted data and converting it into bloodline data for storage. At the same time, it needs to analyze the data flow and trace the source of the data resources. It generally includes the following three aspects [8].

(1) Object-level blood relationship analysis: It is used for the data link relationship between relational data tables, non-relational data file objects, etc., and is generally used to reflect object relationship scenarios between different levels.

(2) Field-level blood relationship analysis: It refers to the blood relationship of metadata, mainly including table attributes and the data source relationship between file attributes.

(3) Task-level blood relationship analysis: It refers to the data processing link of task processing (including data extraction, data cleaning, data transformation, etc.), so that users can see the data processing process of any object, the execution time of the task, and the successful execution of the task.

2.3 Data Lineage Display

It Refers to graphically displaying the parsed blood relationship data through visualization, including object overview, field identification, blood relationship diagram between objects, object sample data, etc., to view the blood relationship data of objects from different perspectives.

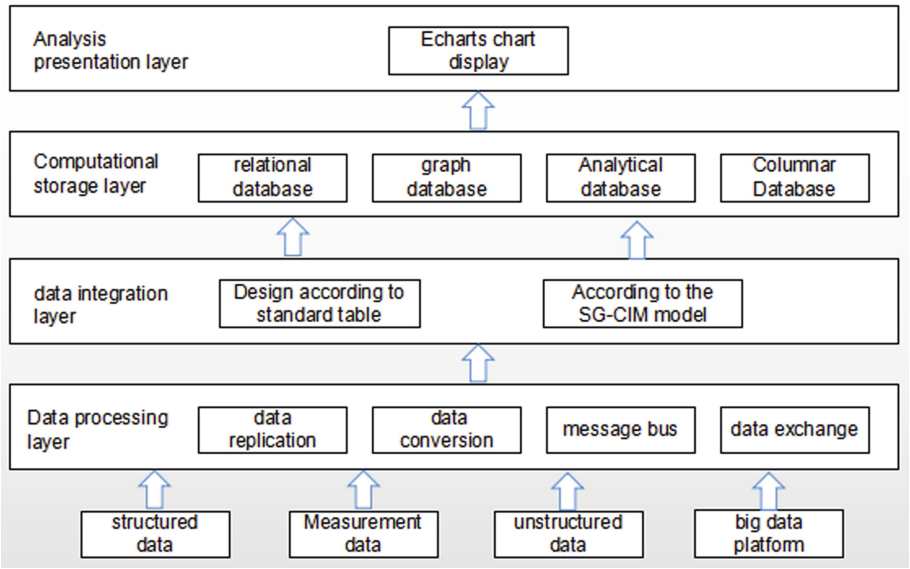

Fig. 2. Data lineage technology structure

2.4 Data Lineage Application

Object lineage analysis can clearly reflect the data flow between objects such as tables, files, fields, tasks, etc., and can clearly show the data flow link of the entire system to the user.

When an abnormality occurs in a certain data, the user can analyze the association relationship of the object to quickly find out the cause of the data problem. For example, when the user finds that the line loss of the analysis layer is abnormal during the same period, he can analyze the statistical index data in the upstream shared layer of the topic table, track it layer by layer, find out the detailed list of the problem, and locate the cause of the problem.

3 Data Lineage Technology Architecture

The implementation architecture of data lineage visualization technology is shown in Fig. 2. Combined with the concept of data middle platform layering, from the perspective of data processing sequence, the entire platform is divided into data processing layer, data storage layer and data presentation layer.

3.1 Data Access Layer

Data access is the basic service capability for aggregating various types of company data into the data center. At present, company data mainly includes structured data, unstructured data, collection and measurement data, and message data with specific protocols.

The data mainly comes from the company's business systems, terminal equipment and external third-party service provider systems. Real-time data access and timing access are realized through data replication, data extraction, data exchange and other access methods.

For structured data, according to the frequency of data calculation, it can be divided into offline processing and real-time processing.

Offline processing: After data access, data integration, logic processing or analysis model calculation, service packaging and publishing, etc., it can meet the application T + 1 or H + 1 data analysis requirements.

Real-time processing: After real-time data access and stream computing processing, real-time data is provided for applications in the form of service subscription.

For unstructured data, after unstructured object storage, unstructured content extraction and analysis, and structured data conversion, it can meet the application's analysis and processing requirements for unstructured data.

For data from big data platforms, relational databases are generally used to store metadata information. Integrate the data in the table according to the correlation of different tables, so as to obtain technical metadata such as table fields and table views, and then store the metadata information of power big data through the business table model and business field model. When the big data platform actually processes data information, due to the large amount of data, the workload of analysis is also very large. At this time, various processing methods such as log segmentation method and multi-thread processing method can be used to improve the efficiency of data information analysis and processing, so as to meet the needs of data processing in the power industry.

3.2 Data Integration Layer

Data integration and transformation means that the source layer data is stored in a standard table or SG-CIM model table in the data center shared layer after cleaning, filtering, encoding conversion, and data integration, and updated regularly.

The data integration transformation includes two aspects: Based on the source table, designed according to the standard table, after cleaning and conversion, a standard table is generated in the shared layer; Based on the source table, according to the SG-CIM model, after integration and transformation Generate physical model tables at the shared layer.

(1) Standard table integration conversion

Based on the standard table physical table structure, create a physical table in the data center shared layer, and create a standard table according to the standard table development specification.

Combined with the field mapping of standard tables and the field association standard, from the business meaning, carry out the source business system table and field traceability, and combine with the actual business, configure and complete the field mapping of the source table and the model table or field association processing logic and other transformations rule.

Based on the configuration of the conversion rules, by regularly executing the data integration conversion script, the full data table of the posted source layer is integrated and implemented in the shared layer.

(2) Model table integration conversion

Based on the physical table structure designed by the SG-CIM unified data model, the physical model table is created in the data center shared layer according to the model table naming convention. Missing or changing is not allowed for fields in the model.

Combined with the field mapping and field association standards of the physical model, from the business meaning, carry out the source business system table and field traceability, and combine the actual business, configure and complete the field mapping of the source table and the model table or field association processing logic and other transformations rule.

Based on the configuration of the conversion rules, by regularly executing the data integration conversion script, the full data tables of the source layer are integrated and implemented in the shared area of the shared layer.

3.3 Computational Storage Layer

Storage computing is the ability to process, calculate and implement storage capabilities for various types of data that have been connected to the data center in accordance with rules such as model conversion or business processing. Data storage includes structured data storage and unstructured data storage, among which structured data is mainly stored in relational databases, distributed columnar databases, distributed data warehouses, analytical databases, graph databases, etc.; unstructured data is mainly stored in Distributed file system or object storage; data computing methods mainly include batch computing, stream computing, memory computing, etc.

(1) Structured data

The platform analyzes and calculates the model data of the paste source layer and the shared layer, and uses the packaged data analysis service and algorithm model service to integrate and synchronize the data to the analysis layer relational database or analytical database to form statistical index results and broad topic correlations. Tables, etc., are used by upper-layer applications.

When performing data analysis and calculation, reasonably select data analysis and calculation tools (including Spark, Python, or SQL) according to the complexity of the model data and business requirements. SQL is suitable for data development scenarios where the processing logic is clear, and multiple iterative calculations are not performed for a certain row or multiple rows, which can usually be realized by one cycle. For scenarios that require multiple iterations or dynamic query and calculation, use Spark or Python for data development combined with your own development capabilities. By writing a data analysis calculation script or program, the calculation is summarized, and then the summary data is synchronized to the analysis layer relational database or analytical database through the synchronization tool.

(2) Measurement data

The "message queue + real-time calculation" component is used to complete the real-time calculation of collection and measurement. The real-time calculation engine completes the real-time calculation through the association of the stream processing platform. The types of real-time calculation dimension tables include distributed columnar database, distributed data warehouse and relational database. In the real-time calculation result data, the measurement business application is supported on demand.

Table 1. Core fields of table structure

English field name	Chinese field name
SOURCE_TYPE	source
SOURCE_OBJ_ID	source table unique identifier
PROCESS_ID	task id
DESTINATION_ID	target table unique identifier
CTIME	Generation time

SOURCE_TYPE represents the bloodline type, such as table level, field level, task level, etc.; SOURCE_OBJ_ID represents the unique identifier of the source object, representing the upstream object before data processing; PROCESS_ID represents the task ID in the table processing process; DESTINATION_ID represents the unique identifier of the destination object, representing the downstream after data processing Object; CTIME represents the bloodline generation time

Real-time calculation results are output to a distributed database), and a unified data service is built to support business applications.

The real-time calculation results are output to the distributed data warehouse, and the subsequent offline data analysis and calculation are supported by the data warehouse.

The real-time calculation results are output to the message queue, which supports real-time publishing, subscription and data sharing through the message queue.

(3) Big data platform metadata information

It is necessary to use the log method to solve the acquisition and arrangement of bloodline information in technical metadata.

Extract the useful data information that needs to be processed in the log file through HIVESQL.

The key information in the data is divided into blocks to form logical blocks of information.

Obtain the source, logic, field, target, etc. of the information to be processed from the logical block, and then integrate these information to form the basic structure of the metadata, and record the link information of the data to form the information Relevance.

The blood relationship data adopts different storage methods according to the requirements of different scenarios. For example, for the blood relationship analysis that simply displays the upstream and downstream layers, a relational database can be selected to store the processed blood relationship data. The core fields of the table structure are shown in Table 1.

Based on the above-mentioned core storage structure can implement one-level blood relationship analysis scenarios at the table level, field level, and task level.

3.4 Analysis Presentation Layer

The development of the analysis display layer is based on the open source big data chart component Echarts. Echarts provides a graph component that displays the data relationship structure. By integrating and parsing the data in the computing storage layer, the corresponding component relationship graph is rendered.

Fig. 3. Overview of blood relationship analysis

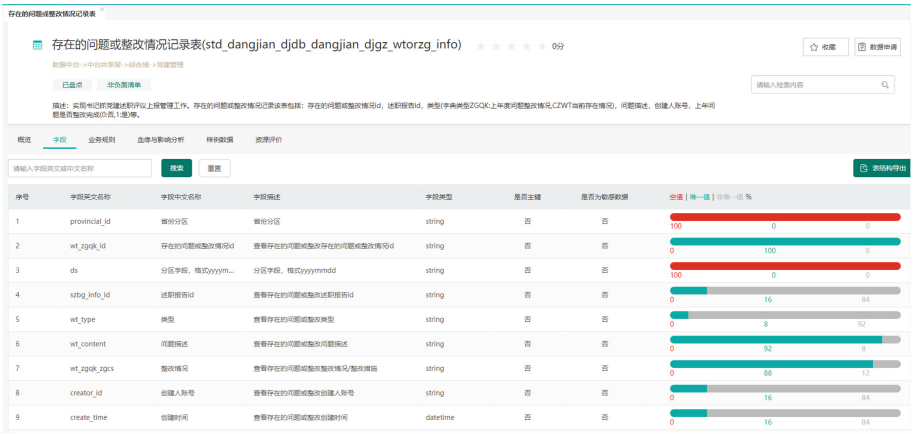

Fig. 4. Bloodline Analysis Field Diagram

4 Application and Practice

Based on the above functional structure and technical architecture, and relying on the State Grid's digital capability open platform, the data lineage analysis function based on the State Grid Data Center has been developed and fully implemented and displayed on the State Grid Intranet (Figs. 3, 4 and 5).

The data blood relationship analysis module is based on the data center of the State Grid, and provides a wealth of innovative value, as follows:

(1) Conducive to problem tracking and positioning: In the process of data development and processing, the processing of data fields can be clearly seen. The data lineage module can visually see the data source link of the model object, including source data

Fig. 5. Power grid data lineage map

table, source field, target data table, target field and other information. When the application runs in error, the error location and error information can be quickly displayed in the link, which is convenient for users to quickly locate the problem.

(2) Conducive to data diversification analysis: Through the display of the whole process link, you can intuitively see the relationship between data objects and indicators. In addition, when an indicator is abnormal, it can help users determine the location of the abnormality, find out the influencing factors by querying the upstream nodes, and conduct further analysis.

(3) It is beneficial to quickly adjust the data processing logic: when the data processing level is relatively deep or the processing time is long, if the processing logic of the first half of the processing nodes needs to be adjusted, through the visual analysis of the data lineage, you can quickly find out the adjustment node. The scope of influence makes R&D personnel only need to adjust the processing logic of some nodes, so as to solve the problem quickly and effectively.

(4) Different from structured data, power grid big data uses metadata management to ensure data quality and avoid the impact of changes through blood relationship analysis, providing companies with a more effective and controllable data management method, guiding and supporting the data of enterprises operation.

5 Conclusions

This paper expounds the characteristics of power grid data, and points out the problems of obtaining metadata, data quality, data positioning, and data display based on some power big data encountered by the State Grid Corporation in the process of digital transformation. Then the concept of data blood relationship is introduced, and the functional structure of data blood relationship, data blood relationship extraction, data blood relationship analysis, data blood relationship display, and data blood relationship application are described. Based on the functional structure of data lineage, the technical framework of data lineage is proposed. Relying on the State Grid's digital capability open platform, the data lineage analysis function based on the State Grid Data Center has been developed and fully implemented and displayed within the State Grid. The results show that the blood relationship analysis function can be beneficial to problem tracking

and positioning, data diversification analysis, rapid adjustment of data processing logic, data governance of the big data platform, and effective solution to the pain points within the State Grid.

References

1. Jia Fuqing. Make persistent efforts to promote the construction of the "three sets and five" system [J]. State Grid. 2013(2): 5051.
2. Fu Dengpo, Jiang Min, Ren Yinan, et al. Data Center: Making Data Useful[M]. Machinery Industry Press, 2020: 104–132.
3. Chang Siyuan. Metadata management and application in massive network data environment [D]. Beijing University of Posts and Telecommunications, 2017.
4. Sun Xiaomin, Ren Guangwei. Summary Design of DW Multidimensional Hierarchy [J]. Digital Technology and Application, 2010(07):13-14.
5. Jin Yong. Research on data lineage management based on data warehouse [J]. Light Industry Science and Technology, 2019,000(004):81-82.
6. Ye Tianqi, Shen Chunfeng. Research and application of data lineage visualization analysis platform [J]. Information Technology and Standardization, 2020,11:17-20.
7. Li Xufeng, Luo Qiang. Analysis of blood relationship for data fields [J]. China Financial Computer,2016,31(7):14-21.
8. Jiang Zhenhua, Zhang Xiaolei. Establishment of data analysis method based on blood relationship [J]. Science and Technology Information Development and Economy, 2015, 25(4): 141-142.