Research on Data Resource Standardization Management Method based on Enterprise-level Data Analysis Domain

Weijie Shen[1,*], Ming Zeng[1], Liang Shen[2], Xing Wang[1] and Zhihua Cheng[2]

[1]School of Economy and Management, North China Electric Power University, Beijing, China

[2]State Grid Big Data Center, Beijing, China

* 2367254472@qq.com

Abstract. Due to the accumulation of information construction work, the State Grid has collected and stored a large amount of data resources such as structured data, unstructured data, power grid GIS data, collected measurement real-time data, and data from various management business departments. The lack of unified standards in data middle courts, scattered data storage, difficult to cross departmental barriers and data barriers, making it difficult for companies to share data internally. In the context of the company's ubiquitous electric power Internet of Things(IoT), this paper proposes the corresponding big data technology architecture based on the enterprise data middle platform, studies the data analysis domain and the corresponding data resource standardization management methods in the enterprise data middle platform, including the analysis domain overall architecture, the analysis domain data flow path, the data unified aggregation and the standardization management method, realizes the unified management and sharing of data, speeds up the application of decision analysis and the whole business The integration and transformation of analysis domain provides the company with efficient analysis and calculation ability and solid data foundation for cross business data analysis and decision-making.

Keywords: big data technology; IoT; management and sharing of data; cross business data analysis.

1. Introduction

The State Grid Corporation of China proposed that it will focus on building world-class energy Internet companies, keep up with innovation and play a role, build a "hub-type, platform-type, and shared-type" enterprise, build and operate a "strong smart grid, and ubiquitous power IoT " in the 2019 working conference. That is, the" three types and two networks "development strategy [1]. Although the data development and application of China's power grid companies have a certain foundation, they need to further tap the value of the data and quickly improve the service level of the data. The ubiquitous electric power Internet of Things requires data construction to be guided by data value mining and the development of the digital economy. It removes the “department-level” barriers to system construction from management, upgrades resources, systems, and data to “enterprise-level”, and simultaneously promotes the enterprise. Organizational structure change; technically process the common data of the enterprise into a service and deposit it in the data middle court to form a flexible and powerful shared service capability for direct application of front-end business application construction or data analysis. It can be seen that building an enterprise-level data middle court is an effective way to improve the data sharing service level of electric power enterprises.

Data middle court is a relatively new concept in the big data industry. At present, research on data middle court in power companies is mainly focused on big data, databases, business models and other fields. The literature[2] proposes power model design and optimization methods for power companies, including conceptual models. logical model and physical model; the literature[3] researches on the development of information technology in electric power enterprises based on big data, and puts forward the challenges and applications of informatization development of electric power enterprises by big data; the literature[4] introduces the SDN network architecture, and believes that the architecture can be simplified Power enterprise network management, reducing operation and maintenance costs. Finally, the deployment scheme of SDN in the power grid cloud terminal system is proposed. The literature[5] analyzes how the random matrix theory is applied to the smart grid big data processing. The proposed scheme can improve the visual display of smart grid big data. Literature The research on the grid line loss analysis is made, and the key technologies required for the big data platform to solve the grid line loss problem are identified. Literature has proposed a unified modeling method for smart grid monitoring and operation of big data analysis systems, which can build application models based on business needs, but the issue of standardized management of data resources has not been discussed in depth.

In summary, the current research on data middle court is mainly focused on the key technology direction of big data, the construction of specific business models, etc. There is less research on the field of data analysis and the standardization of corresponding data resources in enterprise data. The data storage in the data middle court is scattered, and the analysis of departmental barriers is not enough. In order to solve the above problems, this article discusses the method of standardized management of data resources based on the analysis domain of enterprise-level data in order to make full use of enterprise-level data in parallel computing, data analysis and mining, full-dimensional data fusion, full-type data support, etc. In terms of advantages, it will support the company's ubiquitous power IoT strategy.

2. The Overall Architecture of the Data Analysis Domain

Based on the investigation and analysis of the information system and data status of power companies, the overall architecture and technical route of the data analysis domain based on the enterprise-level data middle court are designed. At present, the technical architecture of the analysis domain is summarized as "two areas and four layers": "two areas" refers to the application area and exploration area; "four layers" refers to the basic data layer, integration detail layer (SG-CIM), The light summary layer and the data mart layer; the basic data layer includes the source history area and the vertical history area; the exploration area includes an open experimental environment and a temporary working environment. The overall architecture of the enterprise-level data analysis platform is shown in Figure 1.

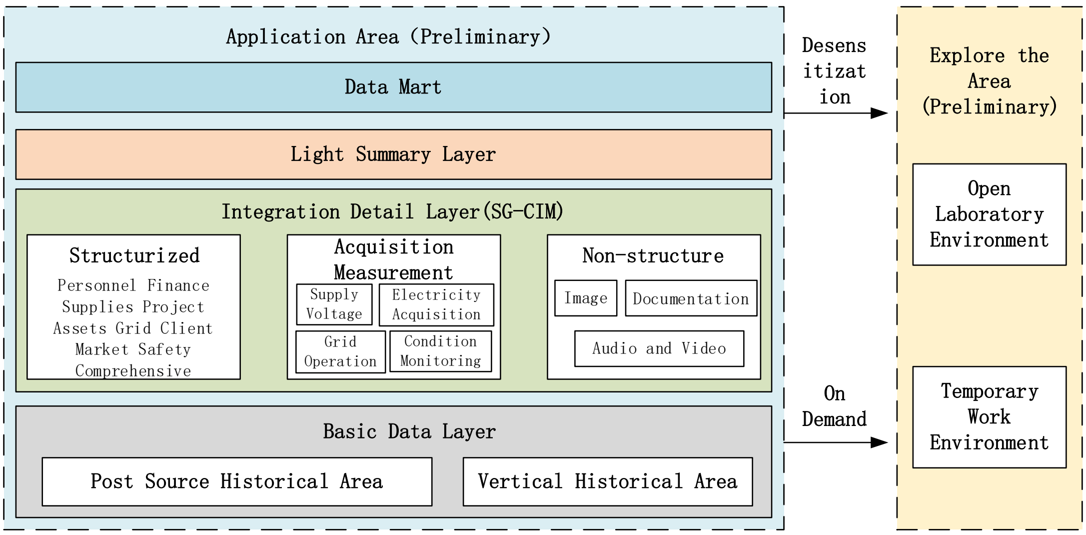

Figure 1. The overall architecture of the data analysis domain

2.1 Application Area

(1) Basic data layer

The basic data layer of the enterprise-level data includes the source history area and the vertical history area. The source history area mainly stores business data from the source. It is required to keep the same with the source business system table structure, and keep historical data for a long time. Trace the data source and shield the impact on the source system. The vertical history area mainly stores the headquarters and provincial and municipal units. Interactive data, such as centralized deployment system data, primary deployment system data, etc.

The source history area and the vertical history area both use the state-owned relational database independently developed by the State Grid based on the open source MySQL version, namely SG-RDB, which provides a storage medium for structured data in the basic data layer. The characteristics of the SG-RDB database are that the kernel is consistent with the open source database, which is easy to integrate into the advanced technology of the open source community and maintain the advanced of independent products. It also improves security performance, functional performance, compatibility optimization, and columnar storage clusters on the basis of open source. And other functions, which have passed the functional, performance, and security tests of the Chinese Academy of Sciences, and meet the national security level 3 security standards; SG-RDB (MySQL) inherits the features of native MySQL transaction security and transaction integrity, and is suitable for more demanding online transaction processing application, has the characteristics of fast speed, small size and simple use, and can provide complete transaction submission and rollback functions.

(2) Integration detail layer

The integrated detail layer stores three types of data for the enterprise: structured data, collected measurement data, and unstructured data. The diversity of data types determines the differences in storage mechanisms.

1) Structured data storage mechanism

The integration of detailed layered structured data is completely implemented based on the SG-CIM full model. A data warehouse is established from an enterprise perspective, and the data is integrated and unified by domain to ensure data uniqueness.

The integrated detailed layered structured data storage uses the industry-renowned domestic database Nanda Universal large-scale distributed parallel database cluster system, referred to as: GBase8 a mpp cluster. Nanda's GBASE8A is the leader in MPP database. The architecture adopts a flat architecture without master and peer-to-peer, and a fully parallel MPP + SharedNothing concept, which has functionally unmatched advantages of traditional databases. ① Columnar storage MPP architecture, supporting horizontal and dynamic expansion of nodes; ② Optimizing column and column hybrid storage using column storage, adaptive compression, horizontal partitioning, and intelligent indexing, taking into account statistics and content querying; ③ supporting standard SQL language, taking into account Friendly and high-performance query method; ④ It also has high-speed data exchange function with Oracle traditional database and HADOOP big data platform. It is a perfect solution for PB-level data warehouses, and is widely used to build very large data warehouse systems and BI analysis displays systems and leadership decision support systems.

2) Collection measurement data storage mechanism

The integrated detailed layer collection measurement data storage uses the State Grid SG-BDP as the storage medium. SG-BDP is a HBase-based distributed column database system that is self-encapsulated by the State Grid. It is the best for quasi-real-time collection of measurement data. The solution uses a distributed column database of the big data platform (based on the open source column database HBase optimized packaging) as the storage medium.

The distributed column database of the big data platform is based on the mainstream big data open source technologies HBase and Hadoop optimized packaging. HBase is a column database based on Hadoop. It can provide quasi-real-time read and write operations for big data and can use distributed file systems. (HDFS) handles distributed transactions, and obtains powerful offline processing capabilities based on customizable MapReduce computing tasks. At the same time, it uses the key and value pair storage mode to achieve efficient query capabilities. HBase is a columnar database with excellent distributed storage, horizontal expansion, and column-oriented characteristics. It can use inexpensive PCs to build large-scale clusters, and use the cluster's powerful computing power to process massive data.

3) Unstructured data storage mechanism

The integrated detail layer unstructured data storage uses the State Grid big data platform HDFS as the storage medium. HDFS is a Hadoop distributed file system, which provides a cross-server elastic data storage system to provide highly reliable file storage on a cluster of ordinary PCs.

Solving the problem of server or hard disk damage by storing multiple copies of blocks. To ensure data consistency, HDFS uses a lease mechanism when appending operations, that is, authorizing block write operations to the primary chunk server, and the other copy is called the secondary chunk server. When multiple clients perform concurrent write operations, the primary block server caches the write order, and then contacts the secondary server for additional operations to ensure consistency of the HDFS data supply version.

(3) Mild summary layer

The light summary layer uses the MPP database (GBASE8A) developed by Nanda General. It is a large-scale distributed database cluster based on the GBase8a column storage database that conforms to the Shared Nothing architecture. It has high expansion, high performance, and high availability, can provide a cost-effective universal storage and computing platform for pet a byte data management. The design of the light summary layer is mainly oriented to actual needs. It provides a cross-business and cross-region data cleaning and conversion place, stores the final data after cleaning and conversion, strives to avoid repeated loading of data, and efficiently supports applications. It is a key area to promote data fusion.

(4) Data mart

The data mart is implemented based on the State Grid SG-RDB (PostgreSQL) database, which is an adapted and improved version of the open source database PostgreSQL's State Grid. State Grid SG-RDB (PostgreSQL) integrates the excellent characteristics of PG, has high operation stability, supports concurrent reading and writing, can provide powerful query capabilities, supports rich data types, and is particularly prominent in GIS data support. PostgreSQl's unique lock-free technology perfectly supports multi-version parallel operation, providing strong technical support for the construction of data analysis platforms.

2.2 Exploration Area

(1) Open experimental environment

The open experimental environment of the exploration area is based on the construction of a big data platform. The Physics Department of the Joint Research Institute uses the national network data desensitization technology to store the desensitized data and provide data support for scientific research experiments on the premise of ensuring data security.

(2) Temporary working environment

The temporary working environment of the exploration area is based on the SG-RDB (MySQL) of the State Grid. As a temporary external window for the enterprise-level data analysis platform, it is mainly open to various business departments and provides data for data mining exploration and cross-business monitoring activities service.

3. Analysis Domain Data Flow Path

Enterprise-level data middle court uses ETL, distributed message queue (kaka) real-time data replication (OG), data interface services, and other technical means to integrate structured type data through source-side business systems, collect measurement real-time data, non- Structure type data, etc., and classify data storage and management according to data type, time dimension, and functional requirements. The data integration path of enterprise-level data is shown in Figure 2:

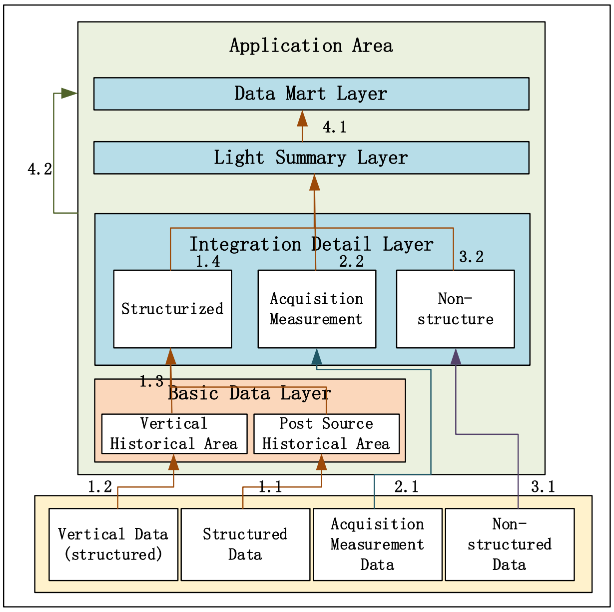

Figure 2. Enterprise data integration route

There are the following data transfer paths.

1) Structured data transfer link (1.1 in Figure 2): The structured data of each business system is replicated in real-time through the OGG method, and then the historical data is exported through the offline DMP method, and the historical data is restored to the basic data layer pastes the source historical area. At the same time, the basic data layer pastes the source historical area, integrates the detail layer structured data, the light summary layer, and the data mart data to clean and transform to meet the data requirements for front-end unified analysis services.

2) Structured vertical data transfer link (1.2 in Figure 2): The structured data issued vertically by the headquarters is connected to the basic data layer of the vertical historical area, and the detailed historical layer of the basic data layer is integrated with the structured data of the detailed layer. , Mild summary layer and data mart data are cleaned and converted to meet the data requirements for front-end unified analysis services.

3) Collection measurement data transfer link (see 2.1 in Figure 2): The measurement data access usually adopts the collection measurement data access component to connect the collection measurement data to the integration detail layer to collect the measurement data. The storage area is then cleaned and converted to a light summary layer and a data mart by TL for unified analysis services at the front desk.

4) Unstructured data transfer link (see 3.1 in Figure 2): Use unstructured data access components to collect unstructured data and store it in the unstructured data area of the integration detail layer. It is then cleaned and converted to a light summary layer and a data mart through unstructured conversion components for unified analysis services at the front desk.

3.1 Structured Data Acquisition and Access

(1) Structured historical data initialization

Structured historical data initialization is usually carried out online and offline. Online method: Online data migration tools, such as Informatica, Kettle and other tools commonly used by State Grid, are read logically from the source database table. The data is extracted into the basic data layer of the data middle court; offline method: the structured historical data is exported to an offline data file using the database's own export tool, and restored to the basic data layer of the data middle court through the offline file. The advantages and disadvantages of the two methods are very obvious, you need to choose the appropriate method according to the actual situation. The online method is straightforward and easy to operate, but the essence of online extraction is to query the database, which places a lot of pressure on the source database and is suitable for system data migration with a small amount of data. The offline method has many steps and complex operations, but the source-side database has less pressure and is suitable for large-scale data volume data migration.

(2) Structured incremental data synchronous replication

Structured incremental data synchronization replication is carried out using the current mainstream data synchronization replication tool OGG (Oracle Golden Gate). Oracle Golden Gate software is a structured data synchronization replication software that captures data changes based on database logs. Archive logs or online logs of the source database to capture data changes, and then transfer these data changes to the target database in the form of files, easily synchronizing the data of the source and target database. Oracle Golden Gate supports a very wide range of IT environments and enables near real-time data replication and backup.

OGG monitors and captures the source system database logs and applies the data change log to the structured data incremental database in real time to form a complete data operation log to maintain consistency with the current data of the source database.

(3) Merge incremental data with historical data

The combination of structured incremental data and historical data is divided into two steps. The first is to smooth the data. Before the incremental data and historical data are merged, it is necessary to perform a smoothing operation on the departments where the historical data and the incremental data are duplicated. The subsequent data merge work has an impact; after the data smoothing operation is completed, start the data merge work, write the corresponding script, read the operation log on the data in the incremental database, and insert or update the incremental data to the history according to the type of operation In the database, the merge of incremental data and historical data is finally completed.

3.2 Acquisition of Measurement Data

The acquisition of measurement data is based on the improved big data platform message queue based on Katka, and it is implemented through programmatic expansion to quickly capture and store the measurement data collected by the State Grid Corporation of China.

The collection measurement data collection component based on Kafka message queue is a distributed high-performance messaging system. It is currently widely used for log collection, online and offline message distribution, etc. It is mainly used for data collection and transmission in big data clusters. The message queue component of the big data platform has strong message processing capabilities. It uses zero-copy and instant consumption mechanisms for batch processing of messages, and has the ability to consume millions of messages per second.

3.3 External Data Acquisition Access

The external data required by the enterprise, such as macroeconomic data and weather data, are mainly obtained in two ways. One is to obtain offline data files in the form of offline export, and then import the offline files to the data middle court. The second is through external data. Get data from interfaces developed by your department;

(1) Import big data platform in file form: Export offline files from the website or database of the department where the external data is located, and import the data files to the data middle court. This method has a large delay, the timeliness of the data cannot be guaranteed, and the work amount is very large.

(2) Interface acquisition: Use the external data interface developed by the department where the external data is located to obtain data, such as weather data. This method requires less work and guarantees timeliness of data.

4. Unified Data Aggregation and Standardization

4.1 Business Data Post Source Access

During the construction of enterprise data , in order to completely solve the problems of "data accuracy, poor real-time performance, and low data quality" and other issues, in terms of data access, the method of business data post source access was adopted, and real-time through OGG Copy, access data quality verification and other methods to ensure that the data connected to the enterprise data middle court is consistent with the data of the source business system, and historical data is stored for a long time, which provides a source of traceability for data quality problems in subsequent applications. Minimize the impact on the source business system, and provide strong data support for future data association analysis and data value deep mining.

4.2 Centralized Storage of Business Data

In the development of electric power informatization, a large number of information systems have emerged at the historic moment. Based on different business directions and different functional requirements, they generate massive amounts of data with their own distinct characteristics. These data do not have a unified data model and rules, and there are relatively strict departmental barriers between decentralized storage and various departments, forming data silos. The data middle court adopts a centralized data storage method for data storage. It collects all structured, unstructured, and measurement data in the power system and stores it in the basic data layer. It physically breaks down the data isolation for the follow-up. Data cleansing transformation provides a high-quality data foundation.

4.3 Data Cleaning Conversion based on the Standard Model

In the construction of traditional data middle courts, it is often only concerned with the storage of data, and there are few achievements in data integration. The enterprise data middle court focuses on establishing a standard data model covering the entire business scope. Based on the standard data model, sort out the business trend of the data flow, clean and transform all collected data according to the data's full life cycle and the business's full process link, so as to form a full amount of data that conforms to the unified data model and complies with standard business processes. Data cleansing and conversion based on the standard model is the key means to solve the inconsistency of the power data caliber, break the barriers of the data department, eliminate the repeated storage of data, promote the integration of business data, and promote the sharing of data resources.

4.4 Unified Information Model SG-CIM3.0

Achieving the unified storage, management, and service of the company's entire business range, all data types, and all-time dimensions is the ultimate goal of the construction of enterprise data . The Unified Information Model (SG-CIM3.0), as one of the core components of the company's information construction, is also necessarily the basic support for the construction of enterprise data . During the construction of enterprise data , whether it is the centralized storage of data or the cleansing and conversion of data, it fully complies with the company's unified information model. Through the design of a unified information model, problems such as inconsistent data standards, inability to share information resources, and difficulty in business integration can be effectively solved during the construction and application of State Grid informatization, thereby establishing a sound information resource management system and sharing mechanism. Effective ways and methods to improve the level of information construction and application. Through the application of the unified information model, it is of great significance to further expand the application scope of data assets, deepen the construction of corporate data platforms, improve data fusion and resource sharing capabilities, and break the system of data partitioning and department isolation.

5. Typical Application of Data

One of the most typical applications of data is through optimization and improvement of management and distribution. Based on enterprise-level data, China and Taiwan deeply explored the results of camp allocation, built a multi-scenario application of camp allocation, strengthened lean management internally, realized visual services externally, and comprehensively improved the intelligence level of mobile operations.

Before the deployment of data , you can establish the SG-CIM data mapping table to ensure the standardization and portability of the distribution management data, and add verification functions such as line loss calculation and visual repair in the unified data middle court of the entire business. With data verification. After the deployment of the data middle court, an enterprise-level shared service center that forms a data resource integration and business capability can be constructed, and the mapping table can be smoothly migrated to the data middle court to achieve the governance data storage in the form of a standard model and pass the multi-scenario deployment application test Management and distribution data, forming a data preservation mechanism, and consolidating the results of management and distribution. For example, build a multi-dimensional labeling system and analysis model at the data level of the marketing side, realize 14 core data capabilities such as energy consumption analysis, promote digital transformation of marketing, and support the improvement of corporate customer satisfaction. In the new installation phase of power consumption, through the support of customer centers, service centers, etc., push the electricity service package, guided by the intelligent customer service to handle the new installation business online, and generate diversified power supply solutions. Support, automatically generate and push integrated energy service solutions such as photovoltaic construction and energy efficiency management; during the commissioning phase of the enterprise, through the support of measurement centers, product centers, etc., regularly push energy analysis reports, and smart customer service guides to handle power package changes. Save electricity costs; in the business development stage, through the support of payment centers, marketing centers, etc., to help companies expand the market, increase corporate exposure, and bring more orders to enterprises.

The establishment of enterprise-level data in China and Taiwan to achieve the optimization and improvement of business allocation mainly includes the following means:

(1) Design the overall structure of the camp. In order to break through the barriers between systems in the marketing and power distribution fields, we will use the data middle court to fully share and distribute business data, to achieve ubiquitous interconnection of equipment-to-equipment, human-to-equipment, intelligent fault analysis, active power outage push, and visual display of repairs, etc. , To improve the efficiency of fault repair and customer service support, to achieve close integration of data and channel business ends and shared support at the front desk. Figure 3 shows the overall structure of the Taiwan-based management and distribution system based on data.

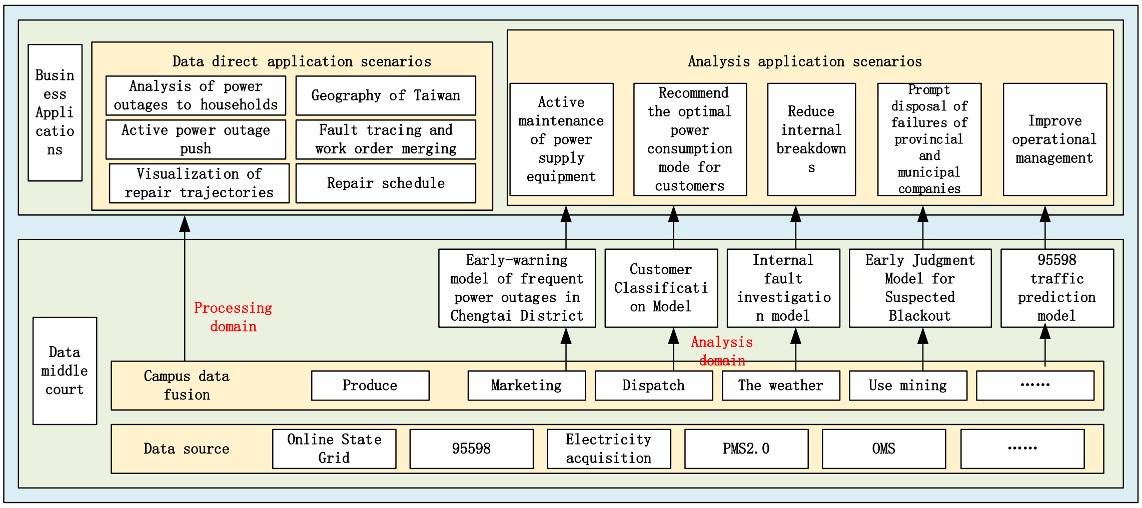

Figure 3. Overall structure of operation and distribution

(2) Improve the basic data model of operation and distribution. Based on the actual needs of the distribution business, based on the SG-CIM4.0 standard of the State Grid Corporation of China, iteratively improve the basic data model standards for the distribution of assets and users.

(3) Establish a mobile operation verification mechanism. Relying on the mobile operation tools of the deployment and verification to carry out on-site manual verification, to achieve real-time on-site maintenance and recording of coordinates and registration data.

(4) Establish the preliminary application of battalion distribution. Based on the results of the optimization and improvement of operation and distribution, we set up preliminary applications of operation and distribution through visualized customer maintenance and mobile operation integration.

(5) Complete the smooth migration of business distribution data. After the data middle court is completed, the managed distribution data is smoothly migrated to the data middle court, and it is stored uniformly in accordance with the standard data model of operation and distribution that complies with the requirements of the SG-CIM 4.0 standard.

(6) Supporting the deployment of collaborative deployment scenarios. Based on the data in the platform, based on the standard data model to design multi-scenario camp allocation applications, timely find the camp allocation problem data in actual application, form a camp allocation data preservation mechanism, consolidate the results of camp allocation data governance, and promote business distribution business process changes . Use the data middle court to interactively query, process and analyze offline data such as customer files, historical loads, and real-time data such as fault information and equipment status to deeply dig the results of operations and distribution, support power supply service command, simultaneous line loss analysis, and industry expansion A series of typical application scenarios such as installation and control, power failure analysis, households and customer portraits.

① Power service command. Through the analysis of customer files and historical data of work orders, accurate and automatic distribution of service work orders and repair orders; the analysis of customer energy consumption habits and environmental factors to achieve the forecast of the load in the Taiwan area; through the blackout information and equipment monitoring information of various channels Analyze and realize fault diagnosis and active repair of distribution network.

② Line loss analysis over the same period. Based on the precise "station-line-variation – household" topological relationship, combined with the collected information, it automatically calculates the line loss for the same period and assists in checking the distribution data. Ancillary analysis of abnormal line loss is combined with data such as power grid load and power inspection, and abnormalities such as grid topology errors, energy meter failures, and suspected power theft are discovered in a timely manner.

③Industrial reporting and installation control. The overall analysis of the data of the marketing system, the operation inspection system, and the infrastructure system will support the seamless expansion of the industry expansion process across disciplines, and realize customer demand, project reserve, and construction line circulation. Monitor the progress of key links such as the progress of the industry expansion plan preparation, the supporting project progress, and the completion and acceptance progress.

④ Analysis of power outages to households. Based on the basic data of a distribution network, combined with the power outage (transmission) plan and failure power outage information, the automatic analysis of the equipment and user lists affected by the power outage is realized. And based on the complete customer profile information, the power outage information is accurately pushed to the households through WeChat public accounts, SMS and other channels to improve the customer's power consumption experience.

⑤ Customer portrait. By building general models such as energy efficiency models, risk models, and evaluation models, as well as dedicated models such as complaint propensity identification models and channel preference analysis models, customers' explicit and implicit characteristics are mined to form customer labels. Providing customer label analysis services based on the data middle court can support applications such as quality service, precision marketing, risk prevention and control, and lean management.

6. Conclusion

This paper describes the technical architecture of the data analysis domain of the full-service unified data middle court based on the data middle court, and analyzes the data transfer path of the analysis domain and sharing method, it is committed to overcome the shortcomings of the traditional data middle court's poor horizontal expansion capability, low hardware obsolescence, and computational capacity that cannot meet the growth of data volume. The proposed method is helpful to realize the unified aggregation and centralized storage of all business fields, all time dimensions, and all data types of power companies, and promote the comprehensive sharing and application of data resources significantly.

Acknowledgements

Science and Technology Project of SGCC(Research on key technology of data center design and realization) SGFJXT00WLJS1900278.

References

[1]. Editorial Department of this journal. State Grid Corporation of China fully deployed the construction of ubiquitous power Internet of things [J]. Rural electrification, 2019 (04): 1.

[2]. Wang Jiakai. Optimization and improvement method of unified data model of electric power enterprises based on data [C]. Professional Committee of electric power information of China Institute of electrical engineering, people's post and Telecommunications Press, 2019:213-215.

[3]. Dong Li, Yu Liang. The exploration and research of big data on the informatization development of electric power enterprises [J]. China informatization, 2015 (12): 78-81 + 83.

[4]. Xing Ningzhe, Wu Shun, Wan Ying, et al. Exploration and Research on the SDN Technology in Electric Power Companies [J]. China informatization, 2015 (12): 78-81 + 83.

[5]. Wang Yifei, Li Xian, Lai Ji, et al. Design of Big Data Architecture of Smart Grid Based on Random Matrix Theory [J]. Computing Technology and Automation, 2019,38 (03): 17-21.

[6]. Tian Dan. Research on Key Technologies of power grid line loss analysis based on big data platform [J]. Science and technology innovation, 2019 (26): 92-93.

[7]. Leng Xiwu, Li Ping, Huo Xuesong, et al. Research on unified modeling method of big data analysis system for monitoring operation of smartpower grid [J]. Power Systems and Big data, 2019,22 (07): 48-54.

[8]. Yang Huafei. Method and research of data resource unified aggregation and sharing based on the analysis domain of full service data center [C] 2017 smart grid conference proceedings of "electronic technology application", 2017:514-518.