Research on Key Technologies of Data Warehouse Construction in Unified Data Center in Electric Power Enterprises

Ting Li\textsuperscript{1}, Ting Fu\textsuperscript{2}, Chen Hong\textsuperscript{2}, Jinhu Li\textsuperscript{3}, Haiyu Lin\textsuperscript{3} and Han Lin\textsuperscript{3}

\textsuperscript{1}State Grid Fujian Electric Power Company, Fuzhou, Fujian Province 350002, China
\textsuperscript{2}Information and communication branch of state grid Fujian power co., LTD., Fuzhou 350001, China
\textsuperscript{3}State Grid Yili Technology Co., Ltd. Fuzhou City, Fujian Province, 350003

Abstract. On the basis of studying the existing problems of power enterprise informatization at this stage, the paper introduces the application of data warehouse technology in power system, proposes to use enterprise-level data warehouse as the data centre, and describes the information of power enterprise based on data warehouse strategy. The overall design of the system model; Finally, specific solutions and application examples are given for the design of the dimensional model of the power enterprise data warehouse.

Keywords: Power system, data warehouse, dimensional modelling.

1. Introduction
With the deepening of the power reform process, the management and operation mechanisms of power energy production, transmission, sales, supervision and other aspects have been continuously optimized, making full use of the results of the previous stage of power informatization, and providing information technology support for power reform and corporate decision-making. Power companies set the overall goal of the integrated data platform. From the perspective of the overall planning of data resources, organize and manage the data of various departments in the form of a data warehouse, use various analysis and prediction tools and special analysis tools of the data warehouse to conduct in-depth data mining and analysis, and provide leaders with a global Proactive production and management decision support, adapt to the future large-scale information system project for the purpose of decision-making needs of enterprises.

In view of the problem of information islands in China's power companies, this paper proposes that the current power application system cannot meet the needs of sudden queries and the needs of decision-making, which greatly affects the efficiency of decision-making and causes the cost of obtaining necessary decision-making information. Very high. In order to solve these problems, the design plan of the power enterprise data centre platform based on data warehouse technology is proposed. The data centre data warehouse is deepened and improved. Lean management of fixed assets has been improved, customer core resource management has been improved, and asset overhaul analysis visualization has been completed. Design and implementation of 6 scenarios including real-time financial information reflection, equipment asset consistency monitoring, and quantitative management improvement of electricity consumption behaviour. This project is mainly based on the
data analysis domain, from data storage, data analysis, etc., to complete the operation and transformation analysis and application, asset management and diagnosis platform related functions. From the aspects of hardware platforms, key technical components, management workbenches and external applications, we have developed relevant monitoring and maintenance functions to enhance the technical analysis and maintenance capabilities of the data analysis domain [1].

2. Design of power enterprise data centre platform based on data warehouse technology

2.1. Overall architecture design

This time the power unified data centre data warehouse deepened and improved the design and implementation of the project. The project architecture design follows the overall architecture design results of the State Grid Corporation's full-service unified data centre and the data analysis domain architecture design results of the power-wide unified data centre. State Grid's full-service unified data centre is the further development and improvement of the company's existing data centre, which mainly includes three parts: data processing domain, data analysis domain and data management domain. Its overall architecture is shown in Figure 1:

Figure 1. The data warehouse of the unified power data centre deepens and improves the overall architecture.

The data processing domain is the key to ensuring data quality and the foundation for improving the level of data application; the data analysis domain is the core of mining data resource value and improving the level of data application; the data management domain is the key and guarantee to achieve data standardization, uniformity and security.

According to the overall architecture design of the unified data centre of the State Grid Corporation of China, the data analysis domain is a collection centre for full-service, full-type, and full-time data. It is a comprehensive data resource, efficient analysis and calculation capability for the company's various analysis and decision-making applications. The unified operating environment changes the situation of repeated extraction and redundant storage of analytical application data in the past, realizes the transformation from "moving data" to "moving computing", and supports the comprehensive development of enterprise-level data analysis applications.

The data analysis domain follows the unified data model, master data and other standard specifications, and is built on an enterprise-level big data platform. It includes three parts: unified storage service, enterprise data warehouse, and unified analysis service, and is deployed at two levels. The unified storage service provides unified storage and management of structured data, unstructured data, collection and monitoring data, and external data. Enterprise data warehouse supports the extraction, cleaning, storage and construction of multi-dimensional analysis models of structured data, and supports the application of multi-dimensional analysis. The unified analysis service provides calculation capabilities and application construction support for data analysis applications, and provides efficient and convenient access to data in the data analysis domain. The full-service unified data centre analysis domain is built on the integrated "State Grid Cloud" platform and has taken the lead in realizing data analysis on the cloud.
2.2. Business architecture design
The deepened and improved design and implementation project of the unified data centre data warehouse is based on the core functions of unified analysis services such as data mining, analysis and display services, and the main operation and monitoring analysis application scenario migration and transformation and asset management diagnosis platform migration and transformation work.

2.3. Application architecture design
According to the overall architecture requirements of the unified data centre data analysis domain of State Grid Corporation, combined with the actual construction of the power data centre and big data platform, the application architecture of this project follows the system architecture of the power data analysis domain, including data access, data storage, and data calculation, Unified analysis service, system management in 5 levels.

2.3.1. Data access. Data access includes functions such as extraction and cleaning of structured data, collection of measurement data, and access to external data collection. Data access mainly connects various types of data to the data analysis domain for the process of calculation and analysis [2].

2.3.2. Data storage. Data storage is mainly for the company's structured data, collecting measurement data, providing a carrier for centralized storage and query, characterized by mass storage at low cost and fast and efficient query reading. According to the overall architecture requirements of the full-service unified data centre analysis domain, enterprise data storage is generally divided into five parts: data buffer, data warehouse, data mart, measurement data storage area, and hot data storage area.
2.3.3. **Data calculation.** Data computing provides distributed operation engines and collaborative computing functions for the data analysis domain, and provides offline computing and real-time computing capabilities to meet the data computing needs of various business applications with different timeliness.

2.3.4. **Unified analysis service.** Unified analysis services provide data mining tools and self-service analysis services for company business personnel to meet easy-to-use, fast, and flexible "DIY" business reports and big data mining applications, as well as provide data routing, data gateway and other functions to achieve external provision Unified data service and presentation service.

2.3.5. **System Management.** System management provides functions such as metadata management, system monitoring, interface monitoring, operation scheduling, and operation and maintenance assistance for the data analysis domain, and provides technical support tools for system management and system real-time monitoring for management and operation and maintenance personnel.

3. **Data architecture design**

Combined with the design results of the data analysis domain architecture of the unified data centre of the entire power business, the design and planning of the data storage and data circulation of the data involved in the deepening and perfecting of the design and implementation of the data warehouse of the unified power data centre.

3.1. **Storage area design**

According to the analysis domain's overall architecture design and application architecture design analysis, the data storage area is divided. The analysis domain data storage area mainly includes: data buffer, data warehouse, data mart, collection and measurement data storage area, and hot data storage area.

3.1.1. **Data buffer.** The data buffer stores the structured data mirror of the source business application system to provide original basic data for subsequent circulation. After the source data of the source business system enters the data buffer, it integrates and processes the source data according to business requirements such as data warehouse and data analysis, and enters the corresponding subsequent storage area. This area is a temporary data storage area. Data is generally stored for 6 months to 1 year. Overdue data will be returned to the warehouse periodically and cleared.

3.1.2. **Data warehouse.** The data warehouse stores historical data that is divided and categorized according to business themes, and provides a low-level and fine-grained data environment for supporting enterprise-level data deep mining, big data applications, and decision analysis. Data storage in the data warehouse area follows the data warehouse model standard. The stored data is mainly enterprise-level data and historical information. The online storage period of data is generally longer, and the data is generally stored for 5 to 10 years. However, after the data is stored for a period of time, the data should be archived regularly considering the carrying capacity of the data warehouse.

3.1.3. **Data mart.** The data of the data mart is formed by the data of the data warehouse after conversion, and directly supports the application requirements of the front end. The data of the data mart is usually used as the data input of the analysis service. Data set the data stored in the urban area is mainly enterprise-level or department-level theme analysis data. The data storage cycle is generally longer, and the data is generally stored for 5 years. However, after the data is stored for a period of time, the data should be archived considering the carrying capacity of the database.

3.1.4. **Storage area of collected measurement data.** The collection and measurement data storage area is used to store the measurement point data, and it can be accessed in real time from the transmission
and transformation equipment status monitoring system, power information collection system, distribution automation system, power quality monitoring system and other systems through the message queue + flow calculation method. Collect measurement data and store it according to SG-CIM specifications. Business applications can directly access this area to obtain measurement point data [3].

3.1.5. Hotspot data storage area. Hotspot data storage aims to temporarily store some data with high access frequency in the data warehouse, data mart or collection measurement data storage area for quick reading and application.

3.2. Data link design
Through various technical means such as distributed message queues, ETL, Sqoop, API, etc., extract structured non-real-time data, collect measurement data, and external data; at the same time, standardize and correlate various types of data according to a unified data specification, and according to different timeliness. The calculation and application requirements are classified, and the data storage, circulation and management are classified, and the work of analysis scenarios, operation monitoring analysis application scenarios migration and asset management diagnosis platform migration are supported. The data architecture is shown in Figure 4 below.

![Data architecture design of power grid design and development implementation project](image)

**Figure 4.** Data architecture design of power grid design and development implementation project.

Data link 1: The non-real-time structured data of the business system in the data processing branch centre is loaded into the data buffer through the initialization of the inventory data and the synchronous replication of the incremental data. At the same time, the data buffer, data warehouse, and data mart are the data is cleaned and converted between the time for the front desk data calculation and unified analysis service.

Data link 2: Connect the real-time data collection to the storage area of the collected measurement data for the front desk data calculation and unified analysis service.

Data link 3: Store the external data in the data buffer, and then clean and convert the external data in the data buffer according to the demand for the front desk data calculation and unified analysis service.
Data link 4: Temporarily save some hotspot data in the enterprise data warehouse and the collection and measurement data storage area to the hotspot data storage area for data computing and unified analysis services to quickly read and apply.

3.3. Database design

The database design of this project is in accordance with the relevant requirements of the "National Grid Corporation Database Design Specification", and the database design of the project is completed in accordance with the naming convention, design principles and design steps. Databases generally include objects such as tables, views, procedures, functions, triggers, sequences, and indexes. All objects are named in a mixed naming manner, such as prefix + object type.

By naming, you can know the meaning and role of data tables and data fields. By naming, you can know the relationship between different data tables. By naming and organizing, you can basically know the data filling rules of the data table. It is not recommended to use database keywords as object names (such as ID, FUNCTION, etc.). If there are many subsystems, when the database structure is designed, different business subsystems can be distributed under different data users, and they can also be combined in the same one as needed. Data users, if the data needs to be integrated into the unified information platform, can correspond to the original business system data table on the unified information platform according to the original table name [4].

The naming of data tables of different business subsystems should be globally unique, and no data tables of the same name should appear. Through the prefix of the data table, distinguish different subsystems, different business groups, as shown in Table 1 for the object naming comparison table.

| Object name       | Prefix         | description                                                                 |
|-------------------|----------------|-----------------------------------------------------------------------------|
| Name database     | no             | Use the English abbreviation of the software engineering name               |
| Table space name  | database_      | The table space is named using the database name functional description    |
| Table space data  | no             | Use table space naming                                                     |
| Table space data  | no             | In principle, the length of the table name should not exceed 25 letters    |
| Database name     | no             | In principle, the name of the view should be able to express the table associated with the view and the meaning it represents |
| Database view     | V_             | In principle, the index should bring the table name or table name abbreviation plus the index field or field abbreviation |
| General index     | IDX_           | In principle, the index should bring the table name or table name abbreviation plus the index field or field abbreviation |
| Bitmap index      | BNX_           | Same as general index                                                       |
| Unique index      | UQ_            | Same as general index                                                       |
| Sequence          | SEQ_           | If there is only one sequence in a table, SEQ_table name is used directly; otherwise, SEQ_table name (or abbreviation of table name) _field name (or abbreviation of field name) |
| Primary key       | PK_            | If the primary key of a table has only one field, PK_table name; otherwise, the naming rule is PK_table name (abbreviation of table name) _field name (abbreviation of field name) |
| Foreign key       | FK_            | Same primary key                                                            |
4. Power data mining and processing algorithms

4.1. Association rule mining algorithm

Association rule data mining algorithm has been a research hotspot in recent years, and its research has never been interrupted. Association rules mainly express the interdependence between transactions. In the interrelationship of transactions, if there is an association relationship between a thing and other things, some attribute values in these transactions can be inferred from other things [5].

(1) Using the accuracy of the obtained classification model prediction, the accuracy of the prediction affects the ability to predict and classify the data set. (2) The execution efficiency of the classification model, which also includes the complexity and calculation speed of the model itself, the implementation details of the algorithm and model, and the hardware environment will have a greater impact on the execution efficiency. The data sets faced are often very large, and sometimes can reach the PB level, which requires the classification model we use to achieve optimal in terms of time complexity and space complexity, which is worth more in-depth classification model the study. (3) The ease of use and understanding of the model adopted. With regard to the application of data mining, we have been optimizing data mining algorithms so that the final data mining algorithm can be both efficient and easy to use. (4) The scalability of the classification model. For the processing of data sets of different sizes and types, the classification model needs to be changed to make the classification of the data set have better accuracy, which requires the classification model to have a certain scalability.

4.2. Implementation of parallelized Apriori algorithm based on MapReduce

4.2.1. Data sets and data items. Let all data sets be \( I = \{I_1, I_2, I_3, \ldots, I_m\} \), and each \( I_k \ (k = 1, 2, \ldots, m) \) attribute in this set is a value pair, which is a data item, abbreviated as \( I \), and transaction database set as \( D \), then there is \( T \subseteq I \).

4.2.2. Support. The support of rule \( A \Rightarrow B \) refers to the number of transactions that contain both \( A \) and \( B \) in the database transaction set.

The ratio to the total number of transactions. Suppose \( I_i \subseteq I \), the \( I_i \) item set's support on database \( D \) refers to the proportion of transactions containing \( I_i \) in total transactions.

\[
supp(A \Rightarrow B) = \frac{|\{T : A \cup B, T \in D\}|}{|D|}
\]  

The association rule \( A \) means: in the transaction database set \( D \), it also contains the proportion of transactions \( X \) and \( Y \) and all transactions in the transaction database \( D \). The support degree is an important indicator in judging whether the association rule has meaning. If the support value is relatively large, it means that the association rule is likely to occur. On the contrary, if the support value is too small, it means that the association rule is changed. May happen only by accident [6].

4.3. Algorithm application

Using the parallelized association rule Aprion algorithm in the power cloud data analysis platform, the above data is processed, and then the results of the load forecast affected by the weather are obtained, and the results are analysed. The following table is a comparison of the load value affected by the temperature and the load value not affected by the received temperature, as shown in the following table:
Table 2. Comparison result table.

| Numbering | Affected value | Actual value |
|-----------|----------------|--------------|
| 01        | 49187.6928     | 52300.4172   |
| 02        | 53497.2420     | 52234.8813   |
| 03        | 55148.6522     | 53421.9352   |
| 04        | 51134.4917     | 54310.3988   |
| …         | …              | …            |
| 96        | 72358.8236     | 72313.4196   |

In order to display the load forecast results more intuitively, the comparison results are displayed in the form of graph rows, as shown in Figure 5:

![Figure 5. Comparison of power load forecast results.](image)

5. Conclusion
The power enterprise data centre platform based on data warehouse technology performs data analysis and processing on the basis of the data warehouse to provide effective data services for management and decision-making layers. An online analysis system (OLAP) has been implemented for power planning departments and production management departments., Marketing management department, material management department, financial management department, etc. provide free and flexible query function, report generation function and auxiliary decision-making function; at the same time, it also realizes the centralized data collection of major application systems such as dispatching production, planning management, marketing management system, etc. of the entire power enterprise And management; it establishes a data release system that can meet the needs of different levels, including data open to different users; it provides accurate, convenient, and convenient for enterprise leaders to conduct a variety of key indicators analysis, decision analysis, and a series of advanced management activities. Fast detailed data and flexible reports; the data centre platform also proposes unified specifications and standards for future application development to support all new application development for power companies in the future.

References
[1] Cui, H., & Yao, S. [key technology and quantity control of wearable medical devices]. chinese journal of medical instrumentation, 39 (2) (2015)113-117.
[2] Xie, H. P., Gao, M. Z., Gao, F., Zhang, R., & Wang, Y. W. Strategic conceptualization and key technology for the transformation and upgrading of shut-down coal mines. Meitan
[3] Gaggero, M., & Caviglione, L. Predictive control for energy-aware consolidation in cloud datacenters. IEEE Transactions on Control Systems Technology, 24 (2) (2016) 461-474.

[4] Chen, H. X., Cai, C., Liu, J. Y., Zhang, Z. G., & Li, W. M. Discriminatory power of variable number on tandem repeats loci for genotyping mycobacterium tuberculosis strains in china. Zhonghua Liuxingbingxue Zazhi, 38 (6) (2017) 794-799.

[5] Xiong, Y., Shi, J., Yang, Y., Lv, Y., & Rouskas, G. Lightpath management in sdn-based elastic optical networks with power consumption considerations. Journal of Lightwave Technology, 36 (9) (2018) 1650-1660.

[6] Zhu, Z., Bo, Y., Wu, P., Zhao, G., & Zhu, J. Short-term power load forecasting model based on qosqo-rbfnn. Nanjing Li Gong Daxue Xuebao/journal of Nanjing University of Science & Technology, 40 (1) (2016) 97-101.