Research on The Design of Power Enterprise Central Data Platform Framework Based on Big Data

Liang Shen1,*, Zhihua Cheng1, Lingchao Gao1, Yiwang Luo2, Yuxiang Cai3

1Big Data Center of State Grid Corporation of China, Beijing 100052, China
2State Grid Info-Telecom Great Power Science and Technology Co.,LTD, Fuzhou Fujian 350003, China
3State Grid Fujian Electric Power CO.LTD, Fuzhou 350003, China

*Corresponding author: liangshen@sgcc.com.cn

Abstract. With the rapid development of computer technology and Internet, the traditional data mining methods and technologies in power industry will face great difficulties, and it is difficult to carry out accurate data processing and analysis. How to mine valuable data from a large number of original data has become a research difficulty. Aiming at this problem, this paper establishes the framework design of power enterprise central data platform based on big data. In order to further improve the actual performance of the scheme, the defects of existing algorithms are analyzed by IM_Apriori improves the calculation method, simplifies the calculation steps, reduces the calculation times, and provides technical support for enterprise data analysis. Through the analysis of the test results, when the data peak reaches 100 m, the execution time is reduced by 25s, which is obviously superior to the traditional scheme. The test results show that the design scheme in this paper has a high comprehensive performance, compared with the traditional central data platform framework, the performance has been greatly improved. Through the analysis, the research in this paper has achieved ideal results, and has made a contribution to the research on the framework design of the central data platform of power enterprises.

Keywords: Data Center, Power Companies, Promotion of Information Technology, Framework Design.

1. Introduction
Big data contains the meaning of massive data, at the same time; it surpasses massive data in content. In short, big data is a collection of massive data and complex data [1-3]. In terms of data volume, big data includes all data sets of enterprises. However, in terms of its scale and complexity, the existing technologies such as data collection, data management and data calculation cannot be well processed and analyzed, so corresponding big data storage and computing technologies and schemes are needed [4-6]. We can also define the concept of big data as: big data = massive + diversified + fast + low information density.

Big data of power enterprises mainly includes power production and operation data and operation management data. The production and operation data of power enterprises include real-time data
collected from power generation, voltage stability, transformer lines, etc., while the operation management data of power enterprises includes power transaction data and ERP system data [7-8]. How to make full use of the big data information of power enterprises, carry out scientific and in-depth knowledge mining analysis, and establish the enterprise central data platform framework can bring a lot of high added value for power enterprises. The central data platform framework of power enterprises not only effectively improves the safety control and equipment detection level of power production and operation, but also enables power enterprises to carry out more scientific management on all aspects of the power grid, and also makes the market-oriented construction of power enterprises more perfect [9-10].

According to the actual situation of big data used in central data platform framework design of power enterprises, this paper analyzes the evidence collection and finds that there are still deficiencies in technology promotion and technical support with developed countries. Therefore, this paper establishes the framework design of big data in the power enterprise central data platform. Through the design strategy of power big data platform in this paper, the power big data platform is analyzed and designed from the overall architecture, platform function, technical level, module correlation and other four aspects. Through the analysis of the test results, this paper believes that the use of big data can optimize and improve the design of the central data center platform framework and achieve good results.

2. System Logic Architecture and Big Data Fusion Algorithm of Big Data Platform

2.1. System Logic Architecture of Big Data Platform

When users need data analysis, they can upload the company's existing data. Big data analysis platform provides access to multiple data sources. You can use common relational databases, such as MySQL, Oracle, sqlserver, hive, and other data warehouses, or you can use common data files, such as txt files and CSV files. Big data analysis platform mainly relies on Hadoop cluster and its components to provide efficient computing and mass storage support. Using the spring cloud framework combined with business to split the service, the microservice architecture is implemented, and the scalability of the system is improved. Even if the big data analysis platform is faced with the high concurrency scenario of massive users, it can also be easily expanded horizontally and vertically, to a large extent, to ensure the scalability and availability of the big data analysis platform.

2.2. Improved Algorithm of Apriori Algorithm

Aiming at the performance bottleneck problem of Apriori algorithm, many scholars have improved the Apriori algorithm from the nature of the algorithm, the times of scanning transaction database, reducing the database capacity, converting the storage mode of database and combining with other technologies. Apriori algorithm needs to scan the database many times, which greatly affects the performance of the algorithm. In the improved Apriori algorithm, the improvement of database storage mode can solve this problem well. For example, to convert a database into a Boolean matrix, as long as the Boolean matrix is calculated and calculated, it is not necessary to scan the original database. In this paper, the improvement of Apriori algorithm is also based on matrix, but different from the previous improved algorithm based on matrix, the IM_Apriori algorithm based on iterative matrix puts forward the concepts of k-candidate set matrix and k-frequent set matrix for the first time. Through the construction of these two matrices, in the calculation process, it improves the shortcomings of the previous improved algorithm based on matrix, which requires a lot of operation and matrix multiplication, and improves the efficiency of the algorithm.

In order to understand IM_Apriori improved algorithm more clearly, we first understand some concepts and related properties of the algorithm before introducing the algorithm. Support count of item set \( I_j \) is the number of 1 in Boolean matrix, and the calculation formula is as follows:
\[ \text{Support}_c(I_j) = \sum_{i=1}^{n} d_{ij} \]  

(1)

Where \( d_{ij} = 0 \) or \( 1 \) (\( 0 \leq i \leq n, 0 \leq j \leq m \)).

Support count of item set \( I_j \) and \( I_j \) is the number of corresponding columns with 1 at the same time. The formula is as follows:

\[ \text{Support}_c(I_i, I_j) = \left| D_i \times D_j \right| = \sum_{k=1}^{n} d_{ik} \times d_{jk} = d_{i1} \times d_{j1} + d_{i2} \times d_{j2} + \ldots + d_{in} \times d_{jn} \]  

(2)

Logic of matrix and \( D_i \land D_j \),

\[ H_{ij} = D_i \land D_j = \begin{bmatrix} d_{i1} & d_{i2} & \cdots & d_{in} \\ d_{j1} & d_{j2} & \cdots & d_{jn} \end{bmatrix} = \begin{bmatrix} h_{i1} & h_{i2} & \cdots & h_{in} \end{bmatrix} \]  

(3)

Where \( h_{in} = d_{in} \land d_{jn}, \ h_{in} = \begin{cases} 1 & \text{if } d_{in} = 1 \text{ and } d_{jn} = 1 \\ 0 & \text{else} \end{cases} \)

3. Big Data Platform Environment and Test Results Analysis

Improved Apriori algorithm based on matrix is to scan the transaction database once and convert it into 0-1 matrix. When calculating candidate sets and frequent sets, most of them perform logical and or matrix multiplication operations on matrices. Although the time of logical multiplication is much less than that of scanning transaction database, the performance of the improved algorithm will be improved with the increase of transaction number, and the advantage will not be obvious.

To verify Im_Apriori the efficiency of Apriori algorithm, in the single machine environment and cluster environment, respectively, for Apriori, MC_Apriori, Im_Apriori for comparison.

In this paper, Apriori and Im are verified by the operation time and test results of Apriori and MC_Apriori algorithms under different data (n). According to the test results in Table 1, the efficiency of improved matrix Im_Apriori algorithm is higher than that of MC_Apriori algorithm. Moreover, the larger the amount of data, the greater the difference of running time between them. Apriori algorithm is more suitable for large data mining calculation, and has high scalability.

**Table 1.** Performance test results of different support degrees under different data volume

| Operation method | \( N = 1000 \) | \( N = 10000 \) | \( N = 100000 \) |
|------------------|----------------|----------------|----------------|
| Apriori          | 180000         | 22000          | 27000          |
| IM_Apriori       | 120000         | 14000          | 17000          |
| MC_Apriori       | 140000         | 17000          | 22000          |

4. Discussion

4.1. Analysis of Time Performance Test and Data Storage Run Time Test Results Under Minimum Support

According to the analysis of test results in Figure 1, the lower the support of Apriori, IM_Apriori and MC_Apriori algorithms, the higher the Apriori time value is, the greater the gap is. This shows that the smaller the support degree is, the higher the efficiency is.
In addition, this paper further tests the system performance; the results are shown in Figure 2, through the analysis in Figure 2: when the amount of data is less than 50, the improved algorithm does not have much advantage in execution time. When the amount of data is larger than 100, the execution time of IM_Apriori algorithm is obviously less than that of BM_Apriori algorithm. The test results of this set of data sets prove that the improved algorithm has a good performance improvement.
4.2. Design Strategy of Power Big Data Platform

(1) Power big data platform is analyzed and designed from the aspects of overall architecture, platform function, technical level and module correlation. The overall architecture analysis is to explore the overall functions and characteristics of the power big data platform from the perspective of the relationship and interaction between the system, subsystem and module functions, so as to improve the overall effect and pursue the optimization of the overall goal.

Platform function analysis is to find out the reasonable structure of each level function of power big data platform, find out the integrity, correlation and hierarchical characteristics of power big data platform in functional composition, so as to optimize the functional composition and correlation of power big data platform.

Technical analysis is to analyze the power big data platform from the perspective of technical implementation. Hierarchical structure is an important guarantee for the stability and continuity of power big data platform. Power big data platform is divided into data source layer, acquisition layer, storage and processing layer, service layer, application layer, etc.

Module association analysis is to analyze the relationship and interaction between subsystems and modules of power big data platform. In the power big data platform, each module is not isolated, but connected through the structure of the system, interdependent and interactive.

(2) Power big data platform provides the data processing ability of the whole process and the whole cycle to meet the diversified needs of big data acquisition, storage, management, analysis and presentation. All activities in the platform are scheduled and processed through workflow.

(3) Platform functions include efficient data acquisition, storage and processing, data management, analysis and mining, visualization, etc.; data acquisition: it has the ability of massive multi-source heterogeneous data collection and preprocessing, realizes the collection of relational databases, data file and real-time data stream, and has the ability to collect and preprocess structured, semi-structured and unstructured data.

Data storage and processing: it has the capacity of massive heterogeneous data storage and processing, including distributed file system, NoSQL database, distributed queue, distributed object database and other storage methods, and has the ability of real-time computing, offline computing and memory computing.

Comprehensive data management capabilities: including data standardization, data quality, data security.

Statistical analysis mining: multi-dimensional statistics, with multi-dimensional data statistical function; analysis mining, based on data mining technology and machine learning model, combined with the actual scene of electric power enterprises, establishes analysis model.

Visualization ability: it provides a variety of data display forms and rich human-computer interaction modes to realize data interaction, filtering, drilling, brushing, correlation, conversion, etc., and multi-level and multi-dimensional display of data.

(4) Considering the requirements and characteristics of the power big data platform, the following requirements must be met: high performance: the platform adopts load balancing strategy to distribute business requests to multiple servers to improve the processing capacity of the whole platform; distributed cluster and cloud ETL Technology are used to realize the collection and preprocessing of massive multi-source heterogeneous data, and distributed storage and computing technology is adopted to improve the system performance Management ability, avoid data processing performance bottleneck.

High reliability: in addition to having a good system development architecture and deployment architecture, providing highly reliable equipment configuration and optimized program code after repeated tests, it also comprehensively considers the realization of high reliability of the system, application system and system security, system management functions from the perspective of network, host, database, etc.

High availability: there is no single point device in hardware and software structure. LVS is used to realize the load balance of the server and support the linear expansion of the server. Processing
capacity can be improved by adding machines. It supports online addition, modification and deletion of nodes, and does not need to shut down the cluster when upgrading and maintaining.

High compatibility: the platform has high compatibility and can be deployed across platforms without relying on specific hardware resources. The user-oriented page can be compatible with various browsers to ensure the stability of users in different system platforms, terminals and browsers. High compatibility: the platform has high compatibility and can be deployed across platforms without relying on specific hardware resources. The user-oriented page can be compatible with various browsers to ensure the stability of users in different system platforms, terminals and browsers.

4.3. Technical Architecture of Power Big Data

Power big data platform needs to collect, store and calculate massive data. It is necessary to carry out statistical analysis and deep mining of massive data to find out the law contained in it. Traditional database mode, single data processing mode, has been unable to meet the challenges brought by big data. In order to improve the performance of data processing, reduce the cost and enrich the modeling formula, new technologies, new ideas and new strategies are needed.

Platform adopts the current industry-leading data processing technology and mode, combined with a variety of data analysis models, to achieve the industry-leading big data processing and analysis platform. It adopts hybrid architecture, integrates data warehouse, Hadoop, spark, mahout, R and other technical components to realize unified scheduling and control.

Platform realizes the unified management and control of various computing modes. Using the mature big data processing technology and mode of the current industry, build a big data statistics, analysis and mining platform for specific business.

5. Conclusions

This paper analyzes and designs the power big data platform from four aspects: the overall structure of the power enterprise, platform function, technical level and module correlation. Through the test and analysis of the system logic architecture of big data platform and the improved Apriori algorithm based on matrix mentioned in this paper, big data optimizes and improves the design of power enterprise central data platform framework, which makes power enterprise practitioners free from various complicated and time-consuming work. Therefore, the value of big data has been recognized by enterprises. In the end, this experiment is tested under the big data platform. In the test, a number of experiments including run time test, data memory test and so on are carried out. Through the analysis and test results, we can see that when the peak data volume exceeds 100m, the running time of the improved scheme is shortened by 25s, and the running time has obvious advantages over the traditional scheme. When the data volume exceeds 100m, it will reach a higher level in the same industry, which can meet the production demand of enterprises. This research has achieved ideal results and provided technical support for the design and research of power enterprise central data platform framework.

Acknowledgments

This work was financially supported by the project fund: Headquarters Science and Technology Project of State Grid Co., Ltd., “Research on key Technologies for design and implementation of data center”
References

[1] Lee, S. B, Ahn, S, & Jang, I. G. (2015). Development of the optimization framework for low-power wireless power transfer systems. IEEE Transactions on Microwave Theory & Techniques, 63(3), 813-820.

[2] Sinha, P, & Sharma, S. (2016). Determinants of bank profits and its persistence in indian banks: a study in a dynamic panel data framework. International Journal of System Assurance Engineering & Management, 7(1), 35-46.

[3] Kelejian, H. H, & Piras, G. (2016). An extension of the j-test to a spatial panel data framework. Journal of Applied Econometrics, 31(2), 387-402.

[4] Akbar.S Basha, Shanmug.N anathan, Rat.P hish, & Sath.S ish. (2018). Need for comprehending ground water nature utilizing geological data framework- a study report. International Journal of Civil Engineering, 5(3), 24-29.

[5] Kumar, N, & Karhana, A. (2019). Data security framework for data-centers. INTERNATIONAL JOURNAL OF COMPUTER ENCES AND ENGINEERING, 7(1), 451-456.

[6] Hong, H, Chu, H, Zhang, J, & Carlin, B. P. (2016). Rejoinder to the discussion of "a bayesian missing data framework for generalized multiple outcome mixed treatment comparisons," by s. dias and a.e. ades. Research Synthesis Methods, 7(1), 29-33.

[7] Kelejian, H. H, & Piras, G. (2016). An extension of the j-test to a spatial panel data framework. Journal of applied econometrics, 31(2), 387-402.

[8] Xu, W, Zhou, H, Cheng, N, Lyu, F, Shi, W, & Chen, J, et al. (2018). Internet of vehicles in big data era. IEEE/CAA Journal of Automatica Sinica, 5(1), 19-35.

[9] Zou, Y, Peng, Y, Deng, L, & Jiang, T. (2015). Monitoring infectious diseases in the big data era. Science Bulletin, 60(1), 144-145.

[10] Li, C, Hu, Y, Liu, L, Gu, J, & Li, T. (2015). Towards sustainable in-situ server systems in the big data era. Acm Sigarch Computer Architecture News, 43(3), 14-26.