Design of Data Fusion Technology for Energy Enterprises

Di Wang[1, a ,*] and Guang Chen[2,b]

[1]State Grid Energy Research Institute Co, LTD; CHINA

[2]State Grid Energy Research Institute Co, LTD; CHINA

[a]164433103@qq.com; [b]chengguang@sgeri.sgcc.com.cn

*corresponding author

Keywords: Data fusion; Energy enterprises; Technical solutions

Abstract. Energy enterprise groups have a huge amount of data assets, diverse types and sources, but there are some difficulties in cross-disciplinary multi-source data integration. Based on the data characteristics of energy enterprises, this paper studies the big data fusion scheme to lay a good foundation for enterprises to develop the application of big data.

1. Introduction

Due to energy companies more business in a professional division, the data show "chimney" state, cross major of multi-source data fusion has certain difficulty, and corresponding to the human resources and technology of data fusion are insufficient, so the research on data fusion scheme suitable for energy companies is of great significance, to break the barrier of the single system application, restore data collection channels and patterns, condensation and scattered fragments of the data, give full play to the data value.

2. Energy Enterprise Data Fusion Technology Scheme Design

2.1 Design Principles

(1) Multi-source isomeric integrability

Data fusion technology solutions not only integrate structured and unstructured data from multiple sources within the company, but also integrate external data.

(2) High efficiency and reliability

The overall fusion technology solution should have high efficiency and reliability to ensure that real-time data and batch data are completed within the specified time window, so as to avoid backlogs or delays of interactive information.

(3) Standardization

The design of big data fusion technology scheme should follow the standardization principle, fully follow all kinds of mature external standards, and adopt them in accordance with the order of national standards, national standards and industry standards.

(4) Extensibility

The technical scheme should have enough flexibility and expansibility, adopt the design idea of high cohesion and low coupling to adapt to the continuous adjustment and increase of business, and adapt to the new type of business in the way of self-expansion, so as to facilitate the expansion and upgrading and extend the life cycle of the system.

(5) Long-term governance

The data fusion technology scheme should have long-term governance ability to meet the requirements of multi-source heterogeneous data fusion and ensure the effective data fusion.

(6) Economic principle

The technical scheme should be practical and economical. It should make best use of existing resources and insist on reasonable implementation under the premise of advanced and high

  DOI: 10.25236/ciais.2019.002

performance, so as to obtain the maximum economic and social benefits under the premise of the best cost.

(7) Intelligent principle

With the continuous development of computer technology, various algorithms and machine learning methods continue to emerge. Data fusion technology solutions need to use artificial intelligence methods as far as possible to improve the automatic discovery and automatic solving ability of problems, so as to maximize the automation rate and reduce the degree of human intervention.

2.2Technical Architecture Design Ideas

This paper designs the technical architecture according to the layered architecture design and Kimball data modeling theory. As the data sources of energy enterprises are heterogeneous and diverse, and the fusion process is complex, layered architecture design is adopted to achieve the purposes of decentralized attention, loose coupling, logic reuse and standard definition. Enterprise data integration is a process of gradual, application oriented, implementing conditions consistent with Kimball data modeling theory, therefore, in accordance with the Kimball data modeling theory method, the data points in the master data (or dimensional data) of two kinds of data and facts, the technology architecture of the layers of the mapping model and relational mix way for processing and storage.

2.3Data Fusion Technology Scheme Design

This paper designs the technical architecture according to the layered architecture design and Kimball data modeling theory. The technical architecture is divided into six layers, including data source layer, data acquisition layer, data integration layer, fusion degree check and repair layer, fusion model layer and data application layer. The data source layer describes the data source of the data fusion technology scheme. The data acquisition layer is responsible for collecting data in the data source layer; The data integration layer is responsible for integrating data into the fusion model layer. The fusion degree checking and repairing layer is responsible for the master data fusion degree checking and completing the data quality checking and repairing. The fusion model layer implements the fusion model and contains the fusion data itself and its correlation. The data application layer is a concrete implementation of operational monitoring across business applications.

The data source layer describes the data source of the data fusion technology solution, which can be taken from the unified data center of the enterprise, or from the existing business system. With the development of enterprises and the great enrichment of big data, data sources will expand from traditional business systems to larger and more diversified external data. Since typical data warehouse is divided into at least ODS (operational data) layer, model DW (data warehouse) layer and DM (data mart) layer, considering that the checking and automatic repair of data fusion need complete and original data of the business system, the main data sources should choose ODS layer data. For applications built on the data fusion model, the data can also be obtained from DW layer or DM layer according to the characteristics and requirements of the application.

The data acquisition layer is responsible for collecting data in the data source layer. According to the different data types of the source system and the characteristics of inflow aging, different processing modules are designed to deal with it, including batch structured data acquisition module, batch unstructured data acquisition module, stream data acquisition module and real-time structured data acquisition module. Batch structured data acquisition module for traditional structured data acquisition, using ETL tool Informatica or SQL script to achieve; Bulk unstructured data acquisition module for bulk acquisition of unstructured data, using Informatica plug-in, Java program or Python script to achieve; Stream data acquisition module collects stream data, using Kafka, Flume system and combining stream data processing framework Spark, Storm, Flink, etc. The real-time structured data acquisition module aims at the data acquisition with high real-time requirement in traditional structured data sources and USES OGG technology to realize it.

After the data acquisition layer pushes the data to the data integration layer, the data integration layer is responsible for integrating the data into the fusion model. The data integration layer consists of three main business modules: graph master data building module, relational master data building

module and fact data building module. Graph master data construction module USES Cypher, a graph database query and operation language supported by Neo4j, to build a data fusion graph model with multiple business perspectives. The relational master data construction module obtains the results of operation statement transformation of Cypher graph according to the data fusion graph model, and stores the final processed master data and original data into relational master data. The fact data construction module obtains the fact data, converts and standardizes the master data in it, forms the standardized fact data and stores it, and realizes the complete integration of the business data from each business line.

The fusion degree inspection and repair layer is responsible for the fusion degree inspection of master data, data quality inspection, and intelligent automatic identification and repair of data found problems according to rules. For unsolvable data problems, it will be uniformly displayed to business personnel, so that users can solve data fusion problems quickly and ensure data quality. This layer focuses on the four themes of integrity, legitimacy, accuracy and consistency to check and repair. Firstly, cross-checking of integrity, legitimacy, accuracy and consistency between subject domains is carried out. In fact, cross-checking of integrity, legitimacy, accuracy and consistency within subject domains is carried out. Finally, field-level checking is carried out. Fusion degree is mainly detected from the following dimensions. Integrity - whether or not there is a missing condition of the data information. The missing condition may be the missing record of the whole data or the missing record of some field information in the data. Incomplete data can be used for reference value will be greatly reduced, is also the most basic data quality evaluation criteria. The fusion degree check and repair layer checks the integrity of the data setting field after fusion to check whether there are missing information items. Legitimacy - whether the field itself conforms to the expected data standard. The fusion degree checking and repairing layer sets the field of data after fusion, sets the legitimacy check according to the data standard, and finds out the data that does not conform to the data standard and data specification. Accuracy - whether there are anomalies or errors in the data recorded. Common data accuracy errors such as garbled code, abnormal large or small data is not eligible data. The fusion degree inspection and repair layer establishes some rules for the inaccurate inspection data of the fused data and finds out the problems that obviously cannot accurately reflect the business conditions. Consistency - does the data follow a uniform specification, and does the data set maintain a uniform format?

The fusion model layer implements the fusion model and contains the fusion data itself and its associations, including graph master data, relational master data and fact data. In the fusion model layer, Neo4j graph database technology is mainly used to establish the data fusion map model. Neo4j is a database based on Property Graph Model. When modeling Graph data, Neo4j follows the following modeling points: use nodes to represent entities, i.e. things in the domain; The domain is built by using relationships to represent relationships between entities and establishing semantic context for each entity. Use directed association to further clarify semantic association; Use node attributes and any necessary entity metadata to express entity attributes; Use the association attributes and any necessary relationship metadata to express the strength, weight, or quality of the association.

The data application layer is the place where data value is released. The ultimate value of data lies in data mining and data analysis after data fusion, which are all reflected in the final data application layer. After the full sharing and integration of data, the theme of operation monitoring can be completed rapidly and the purpose of "speaking with data, analyzing with data, controlling with data and making decisions with data" can be realized across business applications. Data application layer based on the fusion model layer, can be integrated by adopting the technology of data mining and statistical analysis, machine learning methods, complex network analysis methods, such as large data visualization technology, implementation across business line operation and monitoring functions, for intelligent decision, improve business efficiency and precision marketing support, focus on comprehensive support core business intelligence operations, full service Internet ecological energy, promote the management and business transformation.

3. Summary

In conclusion, a set of mature and flexible data fusion technology scheme is of great significance for energy enterprises to implement data fusion. First of all, there is no practical data fusion scheme for large-scale application and implementation, so as to form large-scale, standardized and systematic data fusion application. Based on the characteristics of energy enterprise data technology, this paper designs the principles of data fusion technology scheme. Secondly, data fusion technology scheme is constructed to realize the whole life cycle process from data acquisition to final application, and the complete data acquisition, integration, check, repair and fusion process is realized to support the subsequent data fusion application practice of energy enterprises.

References

[1]Yixin Sun,xiaobao Yu, Zhongfu Tan, Xiaofei Xu, Qingyou Yan. Efficiency Evaluation of Operation Analysis Systems Based on Dynamic Data Envelope Analysis Models from a Big Data Perspective[J].Applied Sciences,2017,7:624.

[2]Qingyou Yan, Yixin Sun, Chao Qin,Zhongfu Tan.The Pricing model for Transmission and Distribution Tariff Under Different Voltage Levels Based on the Long-run Marginal Cost Method[J]. The Open Electrical & Electronic Engineering Journal, 2015, 9:347-354.

[3] Canbas S, Cabuk A, Kilic S B. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case [J]. European Journal of Operational Research, 2005, 166(2):528-546.

[4] Swan M. Blockchain: Blueprint for a new economy. California: O’ Reilly Media，2015.

[5] Nakamoto S. Bitcoin: a peer-to-peer electronic cash system [Online], available: https://bitcoin.org/bitcoin.pdf, 2009

[6] Wang Fei-Yue. Computational experiments for behavior analysis and decision evaluation of complex systems. Journal of System Simulation, 2004, 16(5): 893－ 897

[7] Brito J, Shadab H, Castillo A. Bitcoin financial regulation:securities, derivatives, prediction markets, and gambling.The Columbia Science & Technology Law Review, 2014, 16:144－ 221