

Research on Architecture Design of Data Warehouse System Based on Big Data

Xu Cheng^{*}, Zuobo Sun

China Automotive Technology & Research Center Co. Ltd, Tianjin, China

^{*}Corresponding author

Keywords: Data storage, Physical structure, Data processing

Abstract: The data warehouse system architecture can be designed to improve the intelligent level of the data system by means of large data technology. Based on this, this paper describes in detail the overall architecture model, physical structure, module function, model, data acquisition, data storage and data processing in the data warehouse system architecture under the big data. The research on the architecture design of the data warehouse system based on big data is realized.

1. Introduction

The main purpose of the data warehouse system design is to help the enterprise obtain the analytical report and the support basis of the decision-making activity. In this process, the application of the big data can deeply optimize the enterprise information utilization rate in the operation of the system, and strengthen the service performance of the data warehouse. Therefore, the designer should analyze the architecture design of the data warehouse system based on the big data to improve the internal management ability of the enterprise.

2. Overall Architecture Mode Design

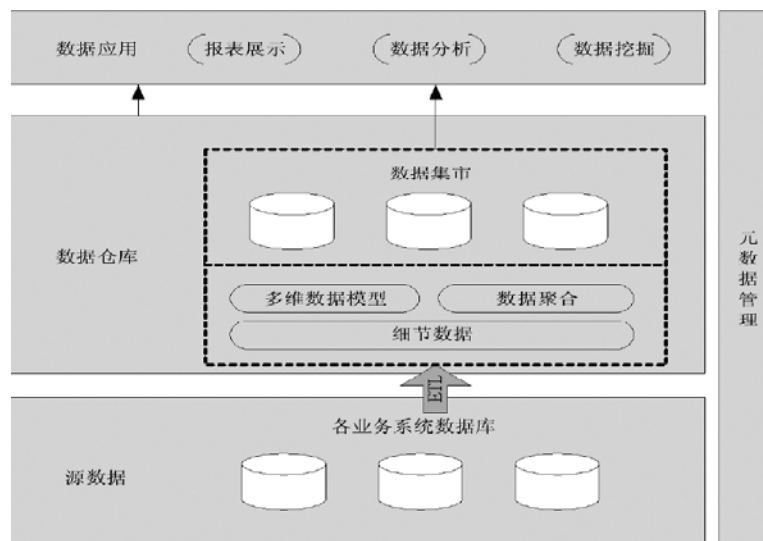


Fig.1 Architecture of a Traditional Data

In the design of data warehouse architecture, the traditional architecture design pattern mainly takes the database technology as the center, realizes the comprehensive application to the data presentation, the data mining, the metadata and so on technology, its architecture is as figure 1, but the big data-based data warehouse architecture has the stronger flexibility than the traditional architecture, the hierarchy division is also more fine and clear, as figure 2. In architecture pattern design, designers usually adopt big data processing technology based on Hadoop / Spark, and combine with traditional architecture to realize database architecture design based on big data. Modularization, component-based design concept and direction, to enrich the main content of the framework, combing the logical relationship between various levels, which can make the internal

operation of the software more orderly and fine, improve the effect of data analysis and processing, at the same time, because this architecture model can quickly grasp, gather, sort out and analyze various types of massive data. Therefore, it has strong adaptability and high efficiency in daily application, improves the big data processing effect of data warehouse system, and optimizes the design level of architecture scheme.

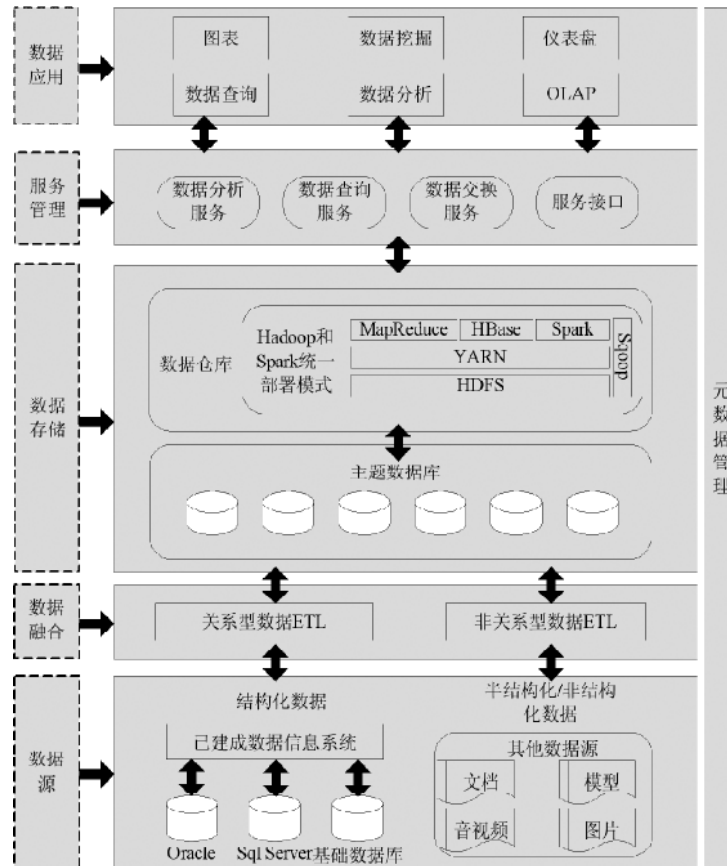


Fig.2 Architecture of Data Warehouse System Based on Big Data

3. System Physical Architecture Design

As an important part of the data warehouse system design based on big data, the designer should strengthen the rationality and science of the design scheme and provide the guarantee for the normal operation of the system. In the physical architecture design, the designer needs to divide the structure into three parts, namely, the source system environment, the development test environment and the production environment, in which the source system environment is an important part of the hardware facilities supporting the operation of the system, which is generally based on data backup, the server and switch of the original system data. At the same time, the designer should connect the source system environment, the production environment and the test environment respectively with the help of the internal switch. Ensure smooth internal operation. In the physical architecture design scheme, the test environment should include the server of terminal, database, application and so on, which is the main platform for the development and testing work. In addition, in the design of data warehouse production environment, designers should set up all kinds of devices, such as application server, portal server, management user access terminal and so on, in this environment, at the same time, they should do a good job in isolating firewall and data exchange equipment between each environment, improve the security in the process of system operation, and perfect the design of data warehouse architecture based on big data.

4. Functional Design of Each Module in the Architecture

In the design of data warehouse based on big data, the design of each module function is the key

factor to ensure its later use effect. Therefore, the designer should refine its functions and improve the effectiveness of the architecture design scheme according to the design goal of the system. In the module function design, generally speaking, the system will include data model management, metadata management, coding management, index management and other aspects. The designer should first make clear the function points of each module, then make a detailed function description according to the function points, and then optimize and perfect the module configuration design according to the use requirements of the enterprise for the system, so as to improve the later stage of the design scheme Implementation effect. Taking the metadata management module as an example, the designer should first clarify the metadata management function point of the module, and in the function description, the module mainly describes the definition and rules of metadata, including its changing configuration, process approval, synchronous management and release, and then in order to meet the regulatory needs of enterprises. In the design of unified application structure data, improve the data warehouse architecture design effect.

5. Model Design in Architecture

In general, the structure of the data warehouse system based on large data mainly relies on the establishment of a reasonable data model to meet the needs of different levels of the user, and therefore, in the design of the system architecture, the designer should make the model design plan. And the flexibility of the data warehouse system architecture is enhanced. In the model design, the designer generally adopts the entity relation model design and the dimension model design, in which the main advantage of the entity relation model design is that the data redundancy can be eliminated, and the operation processing performance of the system architecture can be improved, but because it is structurally too complex for the query function of the data warehouse system, There is a flaw in the user's query and understanding. and the dimension model is designed as a logical design technique, and compared with the entity relation model, it can keep the plane as much as possible and replace the coding with the text attribute, so that the use of the code in the dimension table is greatly reduced, the storage space is reduced. However, the designer should design the model in the data-based data warehouse architecture based on the user's specific needs.

6. Design of Data Acquisition Part in Architecture.

In the operation of data warehouse architecture based on big data, the data often comes from various business processing systems, so there are great differences in form. Therefore, before processing the data, the acquisition function of the system will capture the data from the non-homologous data system according to certain rules, after cleaning and conversion, and then transmit it to the warehouse system. Therefore, the designer should do a good job in the design of the data acquisition part, ensure that the information can be loaded into the data warehouse smoothly, and improve the effect of architecture design. In the design of data acquisition part, designers should construct high quality transcoding and filtering system, effectively remove redundancy and repetition, and realize the number of data. According to the specification of the format, at the same time, the incremental extraction system is set up to improve the ability of the warehouse system to grasp a large number of data. In addition, the designer should construct the data quality inspection function in the data acquisition part, detect the data problems such as format error, lack of integrity, logic inconsistency and so on in time, help the system to complete the cleaning and filtering work accurately, and lay the foundation for the follow-up operation of the architecture.

7. Design of Data Storage Part in 6 Architecture

As an important part of the design of data warehouse architecture based on big data, the design of data storage part mainly relies on the subject database to classify and store the input information and provide the data source for the warehouse, so the designer should ensure the running effect of classification, calculation, scheduling and other functions in the design scheme, and optimize the

design level. In the design of the storage part, the general designers should use Hadoop and Spark to deploy the internal mode architecture uniformly, give the storage system loose coupling and high modularization characteristics, and based on this, make up for the limitation of big data storage processing in the traditional data warehouse. In the process, the design The TARN resource management framework can be used to run and deploy MapReduce, Spark, to achieve the scalability of data resources on demand and share formation storage, which can ensure the accuracy of data classification and storage to a great extent. In addition, designers can also build Sqoop components to realize the data exchange between subject database and Hadoop, and realize the relational data migration within the platform [1].

8. Design of Data Processing Part in Architecture

The data-based data warehouse architecture design is not limited to the structured data processing scope compared with the traditional system architecture, but can realize the processing of both structured and unstructured data, Therefore, the designer should improve the non-structured processing mechanism in the design of the data processing part, and guarantee the structure of the assurance system to meet the application demand of the user under the condition of large data. In the design of data processing part, the designer is to construct a data mart function, so that when the user is inquiring the information, the inquired data can be converted into the multi-dimensional data to be stored in the database to be convenient for the user to use, The main application of the architecture is reflected in the decision-making analysis and data mining. Therefore, the designer should perfect its on-line analysis system and the data mining system to realize the report function of the system, so that the system can play the human-computer interaction capability of the system normally, and in the operation process, Users can meet the needs of multi-dimension analysis, flexible query and key performance index monitoring, and enhance the operation utility of the system architecture [2].

9. Conclusion

To sum up, the architecture design of data warehouse system based on big data can promote the development of enterprise. in the design, the designer can improve the data processing effect of the system by using the large data, improve the architecture design of the system, ensure the later use effect of the system and the reasonable design model, lay a foundation for the follow-up operation of the system architecture, realize the classification and storage of the data, and enhance the utility of the architecture of the system, So as to optimize the design level of the data warehouse system architecture.

References

- [1] Yao, Pengfei. (2019). Design of data warehouse architecture based on big data technology. Digital Technology and applications, vol. 37, no. 03, pp. 141 ≤ 143.
- [2] Yuan, Shiye. (2018). Architecture design of financial management information system in colleges and universities based on real-time data warehouse method. Economic and Trade practice, no. 12, pp. 126.