Research on Hive Integration Application of Data Warehouse Based on Big Data Platform

Haixia Wang, Shuguang Cui*

Information Engineering Departement, Hunan Automotive Engineering Vocational College, Zhuzhou, Hunan, 412000, China

Keywords: Big data; Data warehouse; Hive

Abstract: Hive query data has a high delay because there is no index, so the whole database table needs to be scanned, and after HQL is converted into MapReduce program, the execution is delayed. Relatively speaking, the database latency is low; But if the data scale is very large, Hive's parallel computing can show its advantages. In this paper, the application research of Hive integration of data warehouse based on big data platform is launched. Based on the analysis and design of Hive-based online learning data warehouse, this paper puts forward a concrete implementation scheme of online learning data warehouse, which can provide high scalability by combining the virtualization technology of university cloud platform. According to different sources and formats, it is generally necessary to customize different data extraction and conversion tools, write data cleaning programs, check the consistency of data, and load complete data into the data warehouse environment on a regular basis through data loading. While ETL is implemented, this paper realizes the script of deleting fixed partitions according to the characteristics that all tables in Hive are stored by date partitions, and also uses Shell commands to execute it.

1. Introduction

Hive is a data warehouse tool based on Hadoop, which can map structured data files into a database table and provide simple SQL query function. Distributed computing and processing technology for large-scale and super-large-scale data has become an engineering research topic of great concern. The research and development community generally pays attention to Hadoop technology, which is widely used in the Internet field [1]. Hadoop implements a distributed file system, and HDFS is characterized by high fault tolerance. Hadoop consists of many elements. The data in the data warehouse comes from a variety of business data sources. These data sources may be on different hardware platforms and use different operating systems, so the data are stored in different databases in different formats [2-3]. How to load these large amounts and kinds of data into the data warehouse has become a key problem in building a data warehouse.

Hive data warehouse, as a new data warehouse architecture, makes use of the advantages of big data cluster, and can adopt common server cluster to meet the requirements of online learning platform for high efficiency, reliability and economy of data warehouse. Hive query data has a high delay because there is no index, so the whole database table needs to be scanned, and after HQL is converted into MapReduce program, the execution is delayed. Relatively speaking, the database latency is low; But if the data scale is very large, Hive's parallel computing can show its advantages.

2. Data warehouse tool Hive

The implementation of Hive is based on Hadoop ecosystem. Through Hive tools, data can be extracted, transformed and loaded, and massive data stored in Hadoop cluster can be stored, queried and analyzed. And provide SQL-like query interface, the essence of which is to convert SQL into MapReduce program, which makes the learning curve of Hadoop application for ordinary analysts slow down. Hive supports data ETL, and can also be used to store, query and analyze large-scale data in Hadoop. Hive can parse data only by specifying the column and row separators when creating the Hive table.

Using Hive for offline data analysis is more efficient than directly developing MapReduce programs. Because most data warehouse applications are based on the reality of relational databases, Hive reduces the barriers to porting these applications to Hadoop [4]. Spark SQL, as one of the main components of Spark ecosystem, is similar to Hive's query based on MapReduce. Spark SQL uses Spark as the calculation engine, and needs to be in the Spark environment when it is used. Hive architecture is shown in Figure 1:

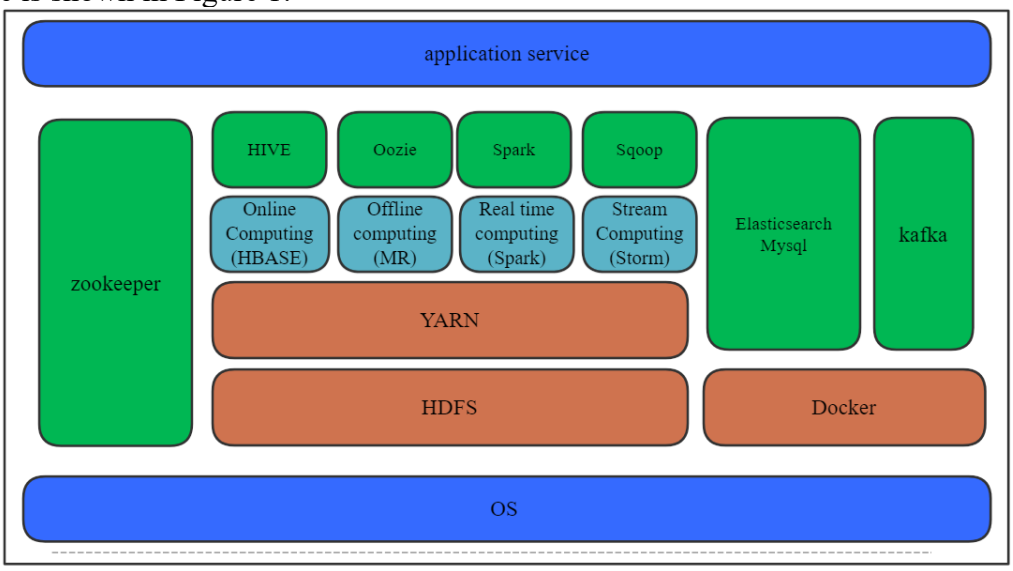

Figure 1 Hive architecture

Hive operation needs to start Hadoop service first, then start MetaStore service and hiveserver2 service. Metastore is the hivestore server, and its main function is to convert DDL, DML and other statements into MapReduce programs, submit them to hdfs cluster and run them [5-6].

On the Hadoop big data platform, the Hive data warehouse is built, and massive structured data is analyzed in the Hive data warehouse, which can meet the needs of big data analysis and processing. Get the related structured application data, and after preliminary processing, submit it to HDFS cluster, create Hive data table, and load the data, which can be analyzed, queried and counted through Hive data warehouse.

3. Hive integrated application of data warehouse based on big data platform

3.1. Overall architecture design

The online learning data warehouse based on Hive studied in this paper can provide background support for data import, storage and analysis of learning resources big data platform, and is an important foundation of this decision-making system. The data warehouse can regularly extract the data from Oracle data source to Hive through ETL for query and analysis, and finally save the result data to MySql database for display on the Web.

Data in data warehouse can be divided into historical detailed data, real-time detailed data and summary data according to different types [7]. Summary data can be subdivided into light summary data and high summary data. Light summary data refers to importing data from each database into data warehouse through data screening and cleaning, while high summary data refers to secondary mining and analysis of data from light summary data. The data sources are generally external business data headed by the government and companies, and internal user data. Import the underlying data from different information systems into the data source, clean, filter and integrate the data by ETL, transform it into unified structural data, analyze the data by the framework of data processing and analysis, and display the data analysis results by visualization technology [8].

As for the architecture of Hbase, the database of Hbase is also operated by the client, its data is stored in HDFS, and the cluster of Hbase is managed by Zookeeper. When Hbase operates data in real time, or periodically loads other data (such as data in source files or tables, etc.), because Hbase

loads data very quickly, it can directly query data through Hive, thus saving Hive from loading data. Before implementing integration, we need to pay attention to the version of both. The fields in Hive and columns in Hbase are maintained by StorageHandler, so the storage format of Hive table should be specified as this storage handler when creating it.

Hive data warehouse is composed of data layer and logic layer: the data layer is composed of Hadoop platform and its components, including Sqoop, Hive, MapReduce and HDFS. Original data enters the data layer through ETL and is stored on HDFS in the form of files as input data of logic layer [9]. During the whole Hive data warehouse work process, the code execution time will be saved as a log, and the automated program will analyze this log, and the execution start time and end time of different programs will be sent to the administrator by email. The online learning big data platform is divided into three modules: data source, Hive data warehouse and Web display. The overall architecture design is shown in Figure 2.

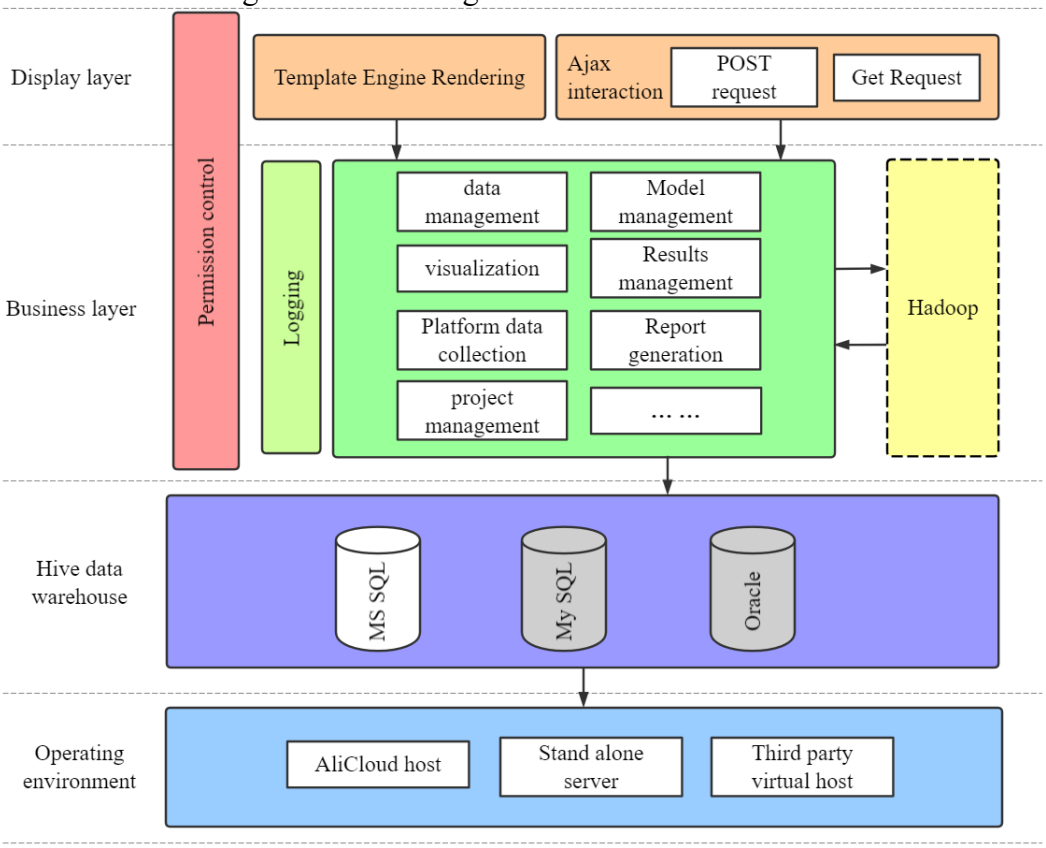

Figure 2 Overall architecture design

Big data all-in-one machine is composed of standardized integrated server, database management system, Sql-like data query software and visualization software. The biggest advantage of this kind of storage mode lies in its excellent stability and vertical extensibility. The main force of batch processing framework is Hive, HBase and other computing frameworks under Apache Foundation. Generally, Apache Spark is the head of streaming framework, which mainly deals with real-time data flowing into the warehouse.

In terms of data collection, we use Kafka, Flume and Sqoop transmission frameworks, and HDFS distributed storage structure under Hadoop version is used for data storage. Hive is selected as the offline data calculation framework for data calculation, and Baidu's E-charts is used for data visualization.

3.2. Realization and optimization

During the installation of Hadoop, it doesn't matter whether password-free login is available or not. However, if password-free login is not configured, every time Hadoop is started, you need to enter a password to log in to DataNode of each machine, and JAVA needs to be installed independently on each server. The installation directory and environment variables can be set to be

the same. hadhoop can only be installed on centos 1 server, but when the server directory is the same, you can copy the installation directory to centos 2 and centos 3 after centos 1 is installed.

For the current data stored in the operating environment or the historical data stored in the slow equipment, according to different sources and formats, it is generally necessary to customize different data extraction and conversion tools, write data cleaning programs, check the consistency of data, and load the complete data into the data warehouse environment regularly through data loading. Figure 3 shows the overall framework of Hive integration of data warehouse.

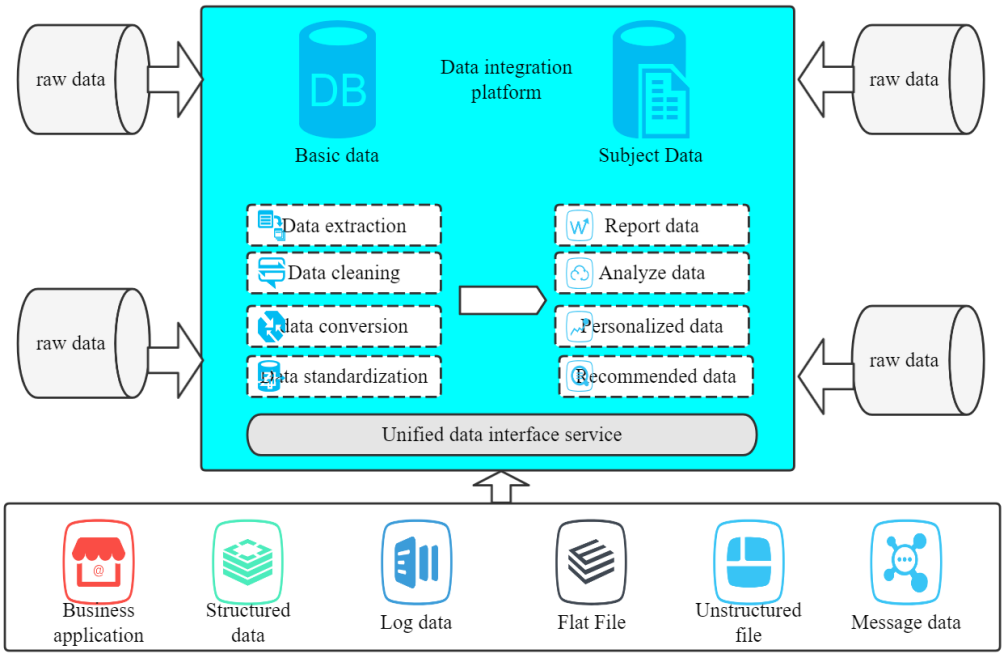

Figure 3 Overall framework of data warehouse Hive integration

The data warehouse builder has to make many decisions, one of which is the key decision to choose which data to load into the data warehouse from the operating system. The builder has to analyze the operating system and decide which data are useful for decision support. Once the source data are selected, the builder will start to consider how to load them into the data warehouse. For example, Oracle's SQL∗Loader can be used to load the data in the input file (including the external data to be loaded into the database) into multiple tables of the database in one operation through the operation control file.

What is needed for data display is the data on Hadoop cluster. If the data in the file needs to be displayed correctly, the format of its data must match the metadata of the data table, otherwise the data cannot be displayed correctly. The underlying service is supported by Hadoop, and it is necessary to configure MetaStore service and Hive service. The underlying service of MetaStore is supported by Mysql, which stores the metadata of Hive [10]. Hive service provides data information, and the bottom layer relies on Hdfs.

MapReduce program optimization is related to each submission of Hive job, and the setting of some specific values will greatly affect the efficiency of MapReduce task execution. One principle of Map Task and Reduce Task optimization is to reduce the amount of data transmission, use memory as much as possible, reduce the number of disk IO, and increase the number of parallel tasks. Besides, it should be optimized according to the actual situation of your own cluster and network [11].

After the data is successfully imported into Hive, the bolScheduler.sh script will be executed immediately, which will call the corresponding business processing scripts for each different business according to the business requirements. These business scripts are stored in the bol folder, which belongs to the same directory as bolScheduler.sh, and they all belong to the subdirectories of the stats folder. The long-term backlog of historical data will inevitably affect the performance of the cluster, and some old invalid data need to be cleaned regularly. Therefore, while ETL is implemented, according to the characteristics that all tables in Hive are stored by date as partitions,

this paper implements the script of deleting fixed partitions, which is also executed by Shell command.

4. Conclusions

Hive data warehouse, as a new data warehouse architecture, makes use of the advantages of big data cluster, and can adopt common server cluster to meet the requirements of online learning platform for high efficiency, reliability and economy of data warehouse. In this paper, the application research of Hive integration of data warehouse based on big data platform is launched. This paper expounds the importance of data integration in data warehouse, and puts forward the overall framework of Hive integration in data warehouse. The realization of Hive-based graphical interface for data management solves the graphical problem of Hive data operation, can directly talk to Hive, and has the functions of data query, data deletion and database management. In this paper, according to the characteristics of Hive that tables are stored in partitions by date, the script of deleting fixed partitions is implemented, and the Shell command is also used to execute it.

Acknowledgements

Hunan Province vocational college education and teaching reform research project, Research on the Construction of Big Data Technology Major Blended Curriculum System under the Background of "New Engineering Disciplines"—Take *Hive Data Warehouse* online course as an example, (ZJGB2019318)

References

[1] Barkhordari, M. , & Niamanesh, M. . (2017). Atrak: a mapreduce-based data warehouse for big data. Journal of Supercomputing, 73(10), 4596-4610.

[2] Malysiak-Mrozek, B. , Wieszok, J. , Pedrycz, W. , Ding, W. , & Mrozek, D. . (2021). High-efficient fuzzy querying with hiveql for big data warehousing. IEEE Transactions on Fuzzy Systems, 2021(99), 1-1.

[3] Jason, L. , Sarah, T. , & Anthony, B. . (2018). Time series classification with hive-cote. Acm Transactions on Knowledge Discovery from Data, 12(5), 1-35.

[4] Bimonte, S. , Gallinucci, E. , Marcel, P. , & Rizzi, S. . (2021). Data variety, come as you are in multi-model data warehouses. Information Systems, 2021(3), 101734.

[5] Subramanian, G. H. , & Wang, K. . (2019). Systems dynamics-based modeling of data warehouse quality. Journal of computer information systems, 2019(1/4), 59.

[6] Naeem, M. A. , Khan, H. U. , Aslam, S. , & Jamil, N. . (2020). Parallelisation of a cache-based stream-relation join for a near-real-time data warehouse. Electronics, 9(8), 1299.

[7] Wrembel, R. , Abello, A. , & Song, I. Y. . (2019). Dolap data warehouse research over two decades: trends and challenges. Information Systems, 85(10), 44-47.

[8] Zhang RUI. (2017). Research and Design of Logistics Big Data Platform Based on hive Data Warehouse. Electronic Design Engineering, 2017(9), 5.

[9] Xu Yuewei,&Xia Lingyun. (2021). Design and implementation of university stream tracing system based on wlan big data and hive data warehouse. Microcomputer Application, 2021(011), 037.

[10] Chen Ken. (2018). Application Research of Data Rights Center Based on Unified Portal of Big Data Platform. Modern Scientific Instruments, 2018(4), 4.

[11] Li Yang, Li Hongxia, Liu Fei, Qiao Xinhui, Li Nan, & Huang Cunqiang. (2021). Research on power system integration and application based on spatio-temporal big data. Electronic Design Engineering, 29(14), 5.