Design and implementation strategy of data migration system based on Hadoop platform

Xiaoshu Wang
Dalian Vocational and Technical College (Dalian Open University), Dalian City, Liaoning Province, 116037
20061215411@dlvtc.edu.cn

Abstract—This paper analyzes the Hadoop technical system related content, and combines the design points of the data migration system, including interface migration design, data synchronization design, data transmission process design, data consistency verification design, data storage scheme design. Through the research of data cleaning function, data conversion function, data loading function, data return function, unstructured data migration and other implementation strategies, the purpose is to improve the application value of data migration system, improve the application value of the collected data.

1. INTRODUCTION
In the background of increasingly rich types of data, the database system established in the past is difficult to meet the needs of data management, especially the integration of unstructured data, making data access, analysis, storage and other activities can not be smoothly carried out, need to use a more reliable platform to assist data processing. The use of Hadoop platform can carry out parallel processing on existing data, and also improve the transparency of underlying storage content, so that the database can have stronger cluster computing power, and then improve the application value of the data in the database.

2. CONTENT ANALYSIS OF HADOOP TECHNICAL SYSTEM

2.1 HDFS file system

Figure 1 HDFS file system diagram
In the process of Hadoop technology application system, belongs to the important content of HDFS file system, as shown in figure 1, the system in the application, classified as a high fault tolerance system, with the help of streaming data access pattern to obtain the corresponding data, including TB, GB, MB large files, and can also meet the demand of read and written many times many times, thus lays the foundation for the smooth deployment of the system server. Compared with other distributed processing systems, HDFS file system in the application process, also has the advantages of strong scalability, simple model application, strong data throughput ability, high fault tolerance ability, belongs to the key content of the system application.

2.2 Map Reduce model

![Map Reduce model flow chart](image)

In the application process of Hadoop technology system, the Map Reduce model is the core content of the system, which is also the basic condition to ensure the smooth parallel of the computing framework. The specific application process of this model is shown in Figure 2. From the perspective of practical application, its computing framework is derived from the Map Reduce structure of Google, which includes the run-time model and basic programming model. It adopts the shunt processing method to complete the parallel processing of data and information, and the Map Reduce model is used in the application. Its master-slave mode basically maintains a strong consistency with the HDFS file system, and its contents include Client, Task Tracker, Job Tracker, Task and other structures.

2.3 Hadoop data query system

![Schematic diagram of HBase system framework](image)
In the process of Hadoop data query, its system includes the following contents: Firstly, the HBase system, which is consistent with the master and slave Server system in the application, includes HRegion Server Server, HMaster Server, etc. Its main function is to apply as a computing node in the large-scale data deployment calculation, and its structural framework is shown in Figure 3. Secondly, the HBase storage structure belongs to the product of RDBMS and No SQL framework application in the practical application, so that the comprehensive management of sparse tables can be carried out in the application, and in the stored procedure, the hash structure will be used to complete the processing, and at the same time, the data retrieval will be carried out with the help of<row key,range> . To meet the application requirements of the system. Thirdly, query process analysis. In the HBase system, special directories such as META and -Root - are also retained, and the list of spatial regions is fully covered in the META table, which also lays a good application foundation for the smooth extraction of data.

3. DESIGN ESSENTIALS OF DATA MIGRATION SYSTEM BASED ON HADOOP PLATFORM

3.1 Interface migration design
In the design process of migration interface, it mainly includes three parts: interface machine, data exchange area and Hadoop platform. The specific division of labor is as follows: interface machine is mainly responsible for the collection, aggregation of interface data on the front machine and the distribution of data to the data exchange area; The acquisition and exchange area is mainly responsible for obtaining the data distributed by the interface machine, analyzing the data and adding the divider, then performing the record level verification, and finally compressing and transmitting the data to the Hadoop platform. Finally, the Hadoop platform is responsible for data storage, cleaning, processing, and data storage and management. The specific execution steps are as follows: Step 1: The interface machine collects and aggregates the verified interface data on the front machine, and generates a data distribution file to transmit to the acquisition and exchange area; Step 2: The acquisition and exchange area obtains interface data through the distribution file generated by the interface machine, and then decomposes and increases the separator operation; Step 3: Check the data files processed in the acquisition and exchange area at the record level, and then transfer the data to the interface layer of Hadoop platform and save it as the original data. Step 4: Clean, integrate and summarize the data, and save it to the data model layer or summary layer as the supporting data for indicators, reports and other applications.

3.2 Data synchronization design
Data migration only covers part of the data with the same business logic in the original system, but the application involved in this part of the data is intersected with other modules. In order not to affect the normal operation of the application after relocation to the Hadoop platform, these unmigrated cross-data need to be synchronized to the Hadoop platform. The specific scheme is as follows: firstly, the applications that need to be synchronized and the tables involved should be listed, and the corresponding model should be established on the Hadoop platform according to the table-building specification of the Hadoop platform; Secondly, the interface number is set for each model table that needs to be synchronized, and the interface synchronization service is established between the original database and Hadoop platform. Then, according to the scheduling cycle of each application that needs to be synchronized, the data files of the table to be synchronized are extracted from the original database and transferred to the Hadoop platform at regular intervals. Finally, the Hadoop platform loads the data files to the corresponding entity tables.

3.3 Design of data transmission process

3.3.1 Data cleaning process
In the process of system application, the main function of data cleaning is to deal with the errors and redundant parts in the data, so that the correct rate of data migration can be controlled above 99%. In the specific cleaning process, the cleaning steps are as follows: (1) Clean the incomplete data, such as missing name, coding error, coding loss, etc., which accounts for 5%-10%. In this case, it is necessary to complete the data before application. (2) Carrying out format content cleaning, such as date format error, numerical content error, redundant characters, etc., which accounts for 7%-12% of the data, and these contents need to be adjusted uniformly, so that all data can maintain a unified format during migration and improve the application value of data. (3) Logical error data cleaning, such as repeated data, unreasonable data, contradictory data, etc., accounting for 2%-10% of the data, it is necessary to comb again according to the logical relationship, merge repeated data, remove unreasonable data, modify contradictory data, so as to improve the practical value of data analysis results.

3.3.2 Data conversion process

From the point of view of data migration application, there are some differences in the content of the data warehouse when it is defined. In order to adjust the data application value to the unified format, it is necessary to transform the data according to the requirements, so that it can better meet the migration requirements. Normally, in carries on the transformation processing, can use a Java program to secondary data transformation, the existing database SQL data, according to the conversion principle replace it into Java file, which is then input to the Java program, which will increase the data utility ratio to 95%, reduced the rate of fault tolerance in the process of data transfer.

3.3.3 Data verification process

Normally, the data validation process, need to adopt cluster calibration principle to deal with it, it is using the data interface machine is introduced in the cluster, with a cluster of high quality resources and technical structure, check for inputting data content, regular checking and records including file content respectively. In the application of the former, the contents to be checked include whether the data naming is standard, whether the data date is correct, whether the file size is compliant, etc.; In the application of the latter, the content to be checked includes whether the data format is correct, whether the data is NULL, whether the data value is in a reasonable range, whether the data logic is reasonable, etc. After the checked data is in compliance, it will be directly input into the transfer system to ensure the rationality of the transferred data.

3.3.4 Data loading process

When the data is loaded, the data will be applied to the loading link according to the requirements, which also lays a good application foundation for the data transfer. In practical application, structured data needs to be migrated from the existing Teradata database to the corresponding Hive database, and some unstructured data will be directly transferred from Teradata database to the HDFS system, so as to improve the rationality of the data processing process. In the process of data information transfer, its corresponding structure and content will change accordingly. At this time, it is also necessary to supervise the integrity and compatibility of the transferred data, and make a good backup in advance as required to avoid the occurrence of data loss. In addition, reasonable control of system performance is also required during loading to optimize the number of schedulers and maps and improve the task execution efficiency by 50%-70%.

3.3.5 Data return process

In addition to the application content mentioned above, data transmission process also needs to be handled well in practical application. The main reason is that there is a certain cross relationship between the new system and the original system. In order to ensure that the original system can continue to run after being relocated to the Hadoop platform, cross data transmission also needs to be done well. So as to ensure the stability of the original system working state. During the specific application of the system, firstly, the interface number of the system return interface needs to be set to
ensure that the Teradata database can establish synchronous service with the Hadoop platform, so that
the accuracy of information transmission can be improved to more than 95%. Second, the scheduling
cycle of the returned data is determined. Meanwhile, the integrity of the returned data is checked.
After meeting the requirements, it enters the next use stage [1].

3.4 Data consistency verification design
In the verification design of data consistency, the following contents should be paid attention to: First,
interface data verification, which completes data statistics with corresponding dimensions, and then
compares it with the data in the Teradata database to check the consistency of the content. Second,
check the intermediate summary data. According to relevant requirements, make statistics on the
number of recorded lines and compare them with the summary data to see whether the content is
consistent. Third, check the index statement, which also needs to summarize the existing index content,
and then compare the data to check the consistency of the data content. In this process, logical
relations can be used to assist content sorting to improve the accuracy of data analysis results [2].

3.5 Data storage scheme design
After completing the above design work, it enters the design link of the data storage scheme, which
also creates strong conditions for the smooth extraction of subsequent data. In the specific application,
firstly, it is necessary to do a good job in the design of the data model, which can be assisted by the
Hive model. Its contents include internal tables, external tables, partition tables and bucket tables,
which are responsible for storing different types of data information to meet the actual application
requirements. Second, the data format is designed uniformly. The commonly used data formats include
Text File format, Sequence File format, RC File format, ORC File format, etc., which can be selected
according to the actual situation to meet the corresponding application requirements [3].

4. ANALYSIS OF IMPLEMENTATION STRATEGY OF DATA MIGRATION SYSTEM

4.1 Data cleaning function
As mentioned above, the data contained in the Teradata database has low regularity when used, which
will also bring interference to the data processing activities, and it is necessary to clean the collected
data. In the realization process of this function, the Map Reduce model can be relied on to assist the
cleaning work, during which the Map mode will also be used to clean the data, and each group of
cleaning tasks will adopt the matching Map mode. At the same time, the Key/Value values obtained in
the Map Reduce model are read with the help of the Context variable, and then transmitted to the
Reduce stage. The obtained information will also be directly input into the HDFS system to complete
the data cleaning of one stage [4].

4.2 Data conversion function
In the realization process of data conversion function, its main work is to sort out the data in the source
data table, and then scientifically convert the application type, so as to get the required data type. In
the process of function implementation, format conversion of existing source data is needed to obtain
the content of Java entity class. Then, the Hive table structure is transformed smoothly with the help of
the reflection mechanism of Java and the Teradata table structure, so as to obtain the required Java
objects and output Key/Value values. This also provides reliable help for the smooth operation of
obtaining values in the Map Reduce model [5].

4.3 Data loading function
To enable data loading, you need to ensure that data can be read from the Teradata database and
written to the corresponding location in the Hive model. In function realization process, needs to be
between the two kinds of database mapping specifications to determine, on this basis, to complete the
data type conversion processing, be able to smooth application of SQL statements in the Hive model,
and load the data in the form of simple string said, and in particular database directory, will also set up a corresponding statement data table, Write relevant data to it as required to obtain the required application content [6].

4.4 Data return function
After the data to be processed is successfully migrated to the Hive database, in order to ensure the stable operation of the original business system, part of the statistical data needs to be directly returned to the database. Then, with custom functions and UDF format, data return processing can be completed smoothly. This kind of implementation approach is relatively simple. In the application, the contents of UFD can be inherited smoothly, and at the same time, the Evaluate function will be organized, so that it and the server can establish a stable Hive application function, and then in the command issued by the Hive function, the query statement can be smoothly operated to get the required output results [7].

4.5 Unstructured data migration
Besides the function of the above mentioned content, in practice, also need to do a good job in unstructured data migration from the point of practical application, the realization of the function of this kind of way, is on the FTP mode, the unstructured data according to certain requirements, directly transferred to the HDFS system, during the data processing and transmission, Flume tool in Hadoop platform will also be used to help with data processing, during which the migration path will be optimized so that it can establish stable connection, so as to speed up data migration and meet the requirements of system functions [8].

5. CONCLUSION
To sum up, in the era of big data, the scale of data to be processed every day is also increasing, and the traditional data migration system is difficult to meet the development needs at the present stage. Further development of the system is needed to meet the needs of data processing. The improvement of the existing system based on Hadoop platform has a positive significance for enhancing the value of data utilization and improving the efficiency of information utilization.

REFERENCE
[1] Hou Qianhao. Research on Data Migration in Communication Resource System -- An Automatic Repair Algorithm Logic for the Under-support Segment of Optical Fiber Cable [J]. Information Communications, 2020(11):223-226.
[2] Gao Xing. Design and implementation of data migration system based on Hadoop platform [D]. Shandong University, 2020.
[3] Cui Jian, Yao Ning. Application of data migration technology in social security system reform and upgrade [J]. Network Security Technology and Application, 2020(03):116-118.
[4] Xie Mengyi. Design of Distributed Big Data Asynchronous Migration System under Hybrid Cloud Storage Architecture [J]. Electronic Design Engineering, 2019, 27(23):45-49.
[5] Chen Xiaowen, Chen Wencho, Liu Zhen. Application of ETL technology in business data migration of power grid information system [J]. Automation & Instrumentation, 2019(08):215-218.
[6] GU Jin. Design and Implementation of Data Migration System Based on Hadoop Platform [D]. Hubei University of Technology, 2018.
[7] Jin Shisheng, Li Jue. Design and Implementation of Data Migration Processing System Based on Spring Batch [J]. China New Communications, 2016, 18(20):54-55.
[8] Xiao Wanwu, Xiang Ning. Design and Implementation of Data Transfer Model for Website Content Management System [J]. Modern Communication, 2011(01):30-31.