Design and Implementation of Data Analysis System Based on Hadoop

Xiang Cui

Department of Public Basic Courses, Guangzhou Polytechnic of Sports, Guangzhou, 510650, China

Email: 2541275072@qq.com

Keywords: Data analysis system, Data mining, Hadoop

Abstract: With the increasing amounts of data, a single host can no longer meet the needs of computing and storage. At present, we mainly use distributed computing and storage methods to analyze and process large amounts of data, and tap potential value from it. Hadoop platform is the most widely used open source computing and storage framework. This paper analyses the functional requirements of data analysis system, and designs a data analysis system based on Hadoop, including data collection module, Hadoop module and HBase module. Experiment shows that, compared with the traditional database, the system has obvious advantages in dealing with massive data.

1. Introduction

In recent years, with the rapid growth of information and data, the traditional database system has been unable to meet the needs of the current society [1]. At the same time, with the development of Internet of Things and social networks and other emerging areas, the growth rate of data volume will be faster and faster. In addition to the explosive growth of data volume, the diversification of data structure and data mobility also put forward higher requirements for the management and analysis of large data. These requirements cannot be met by traditional relational database technology alone. Cloud computing has a super computing capacity of storage, but data mining has new opportunities. The data center of cloud computing platform not only has the characteristics of huge data storage, but also can dynamically distribute related resources according to the actual data mining application requirements, so as to ensure that the data mining algorithm is more scalable. Hadoop framework not only has the advantages of high efficiency and good parallelism, but also has the advantages of economy, reliability and scalability. It is the most widely used and the best cloud computing platform in the future. The design and implementation of data analysis system based on Hadoop is an important development trend of data analysis system in the future. In the era of big data, decision-making in business, economy, administration and other fields will rely more on data and analysis than on experience and intuition. For the massive structured or unstructured data in specific application areas, because of the huge amount of data, complex data structure and the need for timely response, it is impossible to rely on manual statistics and analysis. The information lag brought about by manual methods will result in fatal consequences. Therefore, it is necessary to study data management and analysis system based on large data in order to quickly count and analyze large amounts of structured or unstructured data in specific fields and provide support for decision-making [2].

2. Hadoop: Key Technology of Data Analysis System

Nowadays, Hadoop has grown into a huge ecosystem. Figure 1 is a map of the Hadoop ecosystem.

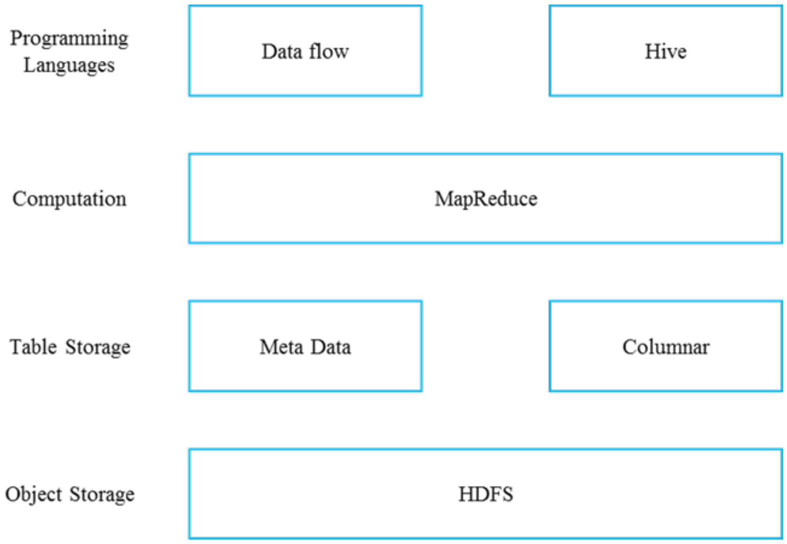

Figure 1. Hadoop ecosystem

In cloud platform framework, Hadoop is currently the most widely used, with powerful functions such as diversity and flexibility. HDFS (Hadoop Distributed File System) and MapReduce parallel programming model are the core technologies of Hadoop. Compared with other file systems, HDFS is excellent in fault tolerance and data throughput. It has the functions of security mode and space recovery. It is very convenient to recover files. Therefore, it is not only used in low-cost hardware design, but also suitable for some applications with large data sets. MapReduce is a distributed parallel programming model and framework. It is mainly used to process large-scale data in parallel. It usually consists of several Map and Reduce tasks. It can divide large-scale data into several Map tasks and assign them to other machines to form intermediate files, then merge these intermediate files to obtain output files. Play the role of data warehouse in Hadoop. Hive adds data structures in HDFS and allows data queries using a similar SQL syntax. Like Pig, Hive's core functionality is extensible. Hive is suitable for the task of data warehouse. Hive is mainly used for static structure and work that needs frequent analysis [3].

3. Functional Requirements of Data Analysis System

At present, a large-scale system usually consists of hundreds of servers [4]. At this time, we need to monitor the data of each system at all times. We may need to know how many users are using each software system every day, how many new users are installed every day, and summarize, analyze, count and report. Typical applications are collecting logs generated by hundreds of systems, analyzing and counting the data, mining and forecasting. In terms of data storage and computation, we adopt and frame, which can solve the problems of massive data storage and massive data calculation and analysis. Deploy monitoring scripts on these clusters to monitor the status of the cluster and provide guarantee for good service. Data analysis system should have the functions of data collection and aggregation, including automatic collection of logs generated by various application servers and system servers, automatic extraction of data from the database of data servers, etc. Data analysis system should have data statistics function. Users are the foundation of a product and an enterprise. Only a better understanding of users can better seize users and serve users. User information mainly includes user basic information, product information installed by users and user behavior information. Products are the basis of enterprise development. Only with products suitable for users can more users be attracted. Product information includes product category, version information, channel information, etc. At the same time, the data analysis system should also have the functions of report presentation and data query. Data items that are queried periodically as required are displayed as reports, which can reduce the workload of querying each time. It mainly includes daily, weekly and monthly newspapers. Report query personnel select the data items needed for data query.

4. Design of Data Analysis System Based on Hadoop

4.1 Data Collection Module.

Data collection refers to the collection of data from various application servers and data servers, and then processing. For an enterprise, there are dozens of hundreds of application servers, recording a wide variety of data, formats are different, the amount of data generated every day is very large. If these data are centralized on a server for analysis, it will cause a great burden on the server, so we use distributed file system to replace the traditional stand-alone mode to save the generated data. The same server in the log may record multiple types of logs. The same type of log data is recorded on different servers, and the same type of log needs to be transmitted to the same file. For the way we use the file system. The name of the log has a corresponding specification manual, and the same type of log is named as the file beginning with the same log identifier. For each server, multiple logs are recorded, and for the same server, multiple logs are configured to accept different types of data files. Generating the same kind of log in pairs and neutrals, each server has a separate acceptance, which will be aggregated into a new one. Finally, the data will be transferred to the file system in a unified way. The central server will save the information to the final file system. When the storage system is unavailable. When the stored machine is down, the data will be written to the local disk. After the storage system returns to normal, the log will be reloaded to the storage system. In order to improve. In the efficiency of data collection, when the local server sends the collected data to the central server, we compress the data transmission and multi-threaded transmission. This can reduce the network overhead in the process of data transmission.

4.2 Hadoop Module.

In Hadoop module, we mainly improve the performance of Hadoop, including using NameNode hot backup mode to provide more stable services, compressing the data collected by Scribe, and using MapReduce for data calculation and analysis. We use HDFS to provide data storage function and MapReduce to provide data summary module. Traditional Hadoop uses a single NameNode to provide services, and NameNode is the most important node in Hadoop, responsible for the file space management of the entire HDFS. Once NameNode has problems, HDFS will be unavailable and data storage will be lost. When NameNode fails, we use Scribe framework to collect data. The collected data will be cached on the local hard disk of the Scribe central server instead of on the original HDFS which provides data storage function. When NameNode is restored, the Scribe central server will write the data cached on the local disk into HDFS again. From this point of view, we can see that the data collection system designed by us, with good fault tolerance, will not lead to the loss of data. In order to make our data mining analysis system more stable, the stability of data storage and data calculation is the prerequisite of system stability. We change the unique NameNode to two NamNodes in the system. The two NameNodes are each other's hot standby. If one NameNode has a problem, it can be switched immediately, and the other one can replace the original NameNode. Of course, if the number of files in HDFS is very small, you can choose to restart NameNode mode, progressive recovery service. We can use MapReduce to distribute the computing tasks of each sub-module in the data mining system with large computational load to each node in the cluster to achieve parallel computing. MapReduce has good scalability and extensibility. It shields the underlying layer and enables us to quickly implement parallel methods of various algorithms by providing programming interfaces.

4.3 HBase Module.

In the data visualization module, we are aiming at the problem of massive data query. Generally, we use relational database for small data query, but we are not competent for hundreds of millions of records of data every day. We use HDFS-based non-relational Hbase to solve this problem. HBase is a subproject of Apache's Hadoop project. Unlike general relational databases, HBase is a database suitable for unstructured data storage. Another difference is that HBase is column-based rather than row-based. HBase is a scalable, highly reliable, high performance, distributed and column-oriented dynamic schema database for structured data. Unlike traditional relational databases, HBase uses data model: an enhanced sparse sort mapping table, where keys are composed of row keys, column keys and timestamps. HBase provides random and real-time access to large-scale data. Meanwhile, the data stored in HBase can be processed by MapReduce, which integrates data storage and parallel computing perfectly. The HBase cluster runs on Zookeeper, which manages a Zookeeper instance by default. If there is a server crash in the area allocation process, Zookeeper is needed to coordinate the allocation. When the client reads and writes data in HBase, it also needs to visit Zookeeper first to understand cluster attributes. For data with small result sets, we import it directly into relational databases, and the data is finally displayed in the form of reports.

5. Implementation of Data Analysis System Based on Hadoop

There are thirty nodes in the Hadoop cluster used in the formal environment of the system. One server serves as the Master node, the NameNode node of HDFS and the Jobtracker of MapReduce model are deployed on it. One server serves as the Secondary NameNode node, and the Secondary NameNode node backs up the data of the Master node to prevent the failure of the Master node. The other twenty-eight servers are Slaver nodes for storing HDFS data and performing computational tasks. First, you need to configure SSH services. There is no key access between the master node and the slave server in the Hadoop cluster. First, the SSH trusted certificate is generated on the master node, and then copied to the slave node. Then the master node can access the slave node without password. Secondly, we need to configure the hosts file, which is used to resolve the corresponding relationship between the host name and IP address. For NameNode node, it is necessary to add IP address and host name of all nodes in the cluster in its hosts file. For DataNode node, it is necessary to add IP address and host name of NameNode node and local machine in hosts file at least. In order to facilitate communication between computers in the cluster, information of other nodes is added in DataNode node's hosts file. For example, the following hosts file is partially configured. Hive needs a database as a place to store Hive's basic information. We use Mysql database. The metadata of Hive is stored in the database, including field information, index information, physical location of data storage and a series of whole Hive metadata. Configure Hive by configuring a hive-site.xml file. Mainly configure Mysql database connection URL and database name, database connection driver, user name and password and other related information. The test method is to compare the efficiency of relational database and cluster in data statistics. The data in both statistics are completely consistent. By controlling the amount of data, the performance of distributed processing is compared with that of single-machine processing of database. Data testing mainly tests grouped data statistics and data association statistics, and achieves the desired results. By using the query system, the data query requirement is more convenient. At present, the system has been put into use formally.

6. Conclusion

This paper analyses the problems of collecting, storing, analyzing and querying massive data in massive data processing, designs and implements a data analysis system based on Hadoop. The actual data validation shows that the system has a great advantage over traditional relational database in Hadoop distributed computing, and has obvious advantages in processing large data. The system is simple to deploy, scalable and can expand nodes whenever necessary.

References

[1] Liao Liang, Yu Hongxiao. Research and Design of Medical Mega Data Analysis System Based on Hadoop [J]. Computer Systems & Applications, 2017, 26(4): 49-53.

[2] Yang Chenjun, Zhang Xin, Yang Zhuodong. Traffic Data Analysis System Based on Hadoop [J]. Electronic Science and Technology, 2017, 30(4): 156-158.

[3] Li Yanmei. Analysis and Design of Data Mining System Based on Hadoop Platform [J]. Computer and Information Technology, 2018, 26(2): 20-22+58.

[4] Yu Shuyun, Lin Shumin. Research and design of user behavior data analysis system based on Hadoop [J]. Journal of Beijing Information Science & Technology University, 2018, 33(5): 65-70.