Analysis of storage requirements of big data system for radio and television users

Huang He
Guangdong Innovative Technical College, Dongguan, Guangdong, 523960

About the author: he Huang, Guangdong Innovative Technical College, date of birth: August 1973, nationality: Han nationality, household registration: Guangzhou, Guangdong Province, education: Master of software engineering, lecturer in software, main research direction: big data technology, software development, etc. Author email: Dynaudio8@126.com

Abstract: through the application of big data, grasp the characteristics of radio and television user groups and viewing behavior habits, understand the actual characteristics and needs of customers, and provide personalized, accurate and intelligent recommendation services. Provide users with a more direct, convenient and personalized user experience to retain customers and reduce the loss of customers. Based on the demand analysis of media content and user service, combined with the big data platform architecture and data governance process, this paper introduces the typical applications of radio and television big data platform and the mechanism of continuous data rolling, and thinks about how to promote the digitalization of radio and television industry.

Keywords: big data; Big data storage; Distributed system

1 Introduction

According to the data of the 2019 China Radio and television cable network technology and annual development report, by the end of the third quarter of 2019, the number of cable TV users nationwide had reached 210million, a slight decrease compared with 2018. The actual number of users in the interim period was 190million, a decrease of 10million compared with 18 years ago, a decrease of 5%; The actual number of digital TV users accounted for 90.5% of the actual number of radio and television users, a year-on-year decrease of 0.4%. In the third quarter of 2019, the number of payment users fell to 150million.

2 demand analysis

The rapid development of new generation information technology and Internet has profoundly changed the wired network industry, but also brought unprecedented challenges and opportunities. The new generation 5g network has also been promoted all over the country, and users' demand for differentiated, diversified and personalized radio and television network services is increasingly urgent. With the rapid development and application expansion of Internet technology, the state has officially promoted the policy of three networks integration. Through technological upgrading and transformation, the three major networks tend to have the same functions, and their business scope is becoming more and more similar. They are interconnected and interconnected, enabling the sharing of business resources, and providing end users with data, voice, television and other services.

With the rapid development of new media business, the advantages of radio and television industry relying on scarce resources have been lost. In the complex and fierce competitive environment, the problem of customer loss of radio and television has become extremely prominent. How to reduce the loss of customers, retain customers and tap the potential needs of customers is an urgent problem for radio and television companies.

Through the network terminal equipment and background system to collect user basic information data, user viewing data, user order data, user bill data and other information, the radio and television industry has gradually formed a demographic characteristic data, terminal usage records, terminal customer behavior trace records, terminal search and demand information records, customer consumption behavior recordsReal huge amount of data such as customer social interaction and suggestion record data. Using this user information database, radio and television companies can analyze and mine user data from the dimensions of people, time, place, products and payment methods according to the characteristics of users, and make a comprehensive portrait of users. Through the application of big data, grasp the characteristics of radio and television user groups and viewing behavior habits, understand the actual characteristics and needs of customers, and provide personalized, accurate and intelligent recommendation services. Provide users with a more direct, convenient and personalized user experience to retain customers and reduce the loss of customers.

3 system architecture selection

The key technologies of big data include big data storage, processing and application. According to its processing process, it can be divided into five links: data collection, data pre-processing, data storage, data analysis and mining, and specific application. In the radio and television cable network, big data technology plays an important role, extending to many businesses such as operation and production, customer service, management and operation.

Use the latest big data technology to associate with external data and application data to provide background support for business and personalized service recommendation. Finally, based on the classification of product classification labels and customer user portraits, we study and judge customer preferences from the classification information, and analyze and estimate the potential behavior, and then

recommend the favorite content to customers with artificial intelligence algorithm.

Before analyzing the mass data related to customers of radio and television companies, it is necessary to consider what storage technology should be used to save the data for subsequent data query and analysis.

In big data applications, after massive data collection and cleaning, it is necessary to determine the storage method that can save the data for a long time. At the same time, it is also necessary to consider a scheme to organize and manage the data for business query and use. Finally, it is also necessary to weigh whether memory storage and processing methods are needed to improve performance. At present, there are many big data storage solutions, including commercial AWS S3 and EMC products, as well as open-source HDFS, swift and alluxio.

AWS S3 storage form can be easily expanded horizontally to adapt to the scenario of high concurrent access by a large number of users, but it does not support random position read and write operations, and can only operate the entire file uniformly. HDFS is an easy to expand distributed file system. Based on the design concept of "mobile computing is more economical than mobile data", it can be built on a large number of ordinary PCs, saving unnecessary investment, and has reliable data fault tolerance, effectively reducing operation and maintenance costs. HBase is more suitable for the business scenario of random reading and writing of massive data, and is suitable for storing massive sparse data. EMC provides high-end products and solutions that support pb~zb level data storage, with excellent data protection and security. However, because it is a commercial product, the application cost is high. Swift supports multi tenant mode, reliably stores a large number of files of different sizes, but optimizes for large files. Alluxio is a memory centric virtual distributed storage system. The core idea is to separate storage and computing, so that spark and other frameworks focus more on computing, so as to achieve higher execution efficiency.

The main characteristics of the user data of radio and television companies are that the number of users is huge, and the relevant information files are also very large. Once the basic data is written, it will not be frequently modified, so it is more appropriate to choose HDFS distributed file system of Hadoop open source framework as the data storage platform.

3.1 introduction to hive, a big data storage tool

Hive is a data warehouse tool based on Hadoop. Its advantage is that the learning threshold is low. Ordinary data personnel can structure script statements to achieve rapid MapReduce statistics without using java language cloud to develop corresponding programs, making the use of MapReduce easier. Therefore hive is a very suitable tool for data statistics and analysis in data warehouse applications.

The main working principle of MapReduce in Hadoop is to divide computing tasks into multiple small units to reduce costs and improve scalability. However, the personnel using MapReduce have high requirements, and they must master Java and other languages to program for MapReduce API programming interface, so as to be more proficient in processing. In addition, data is stored in HDFS of Hadoop, and schema information is not stored. If data is migrated from traditional relational database to Hadoop HDFS for application, schema information will be lost. At this time, hive becomes a bridge between traditional data architecture and Hadoop MapReduce.

3.2 hive principle architecture

Hive defines a simple SQL like query language that uses a script like language to query data. At the same time, it allows developers familiar with MapReduce to handle complex analysis work that can not be completed by the built-in mapper and reducer through the customized mapper and reducer.

Hive's architecture design follows the design pattern of master-slave architecture, as shown in Figure 1.

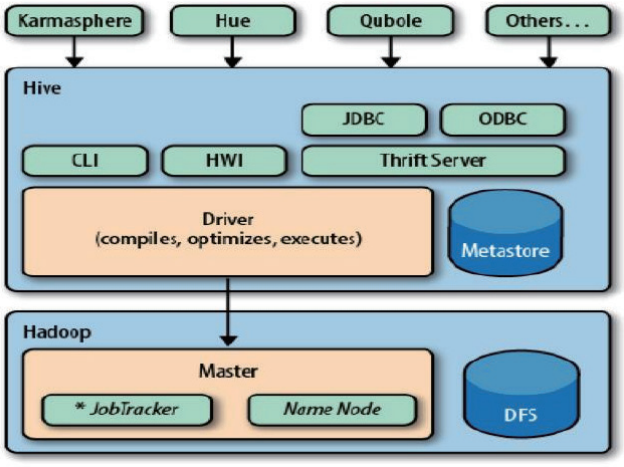

Figure 1 hive architecture

3.3 hive access interface

Hive provides three access modes: command line interface (CLI), hive web interface (HWI) and thrift server client connection. Details are as follows.

CLI is the abbreviation of command line interface, that is, the command line interface of the console. It is the basic connection mode and uses the "hive" command to connect. When cli is started, a hive copy will be started at the same time, which is equivalent to "hive -service cli".

Thrift server provides JDBC (Java database connectivity, JDBC) and ODBC (open database connectivity, ODBC) access capabilities for the development of extensible cross language services. Hive integrates this service, allowing different programming languages to call hive's interface.

Hive web interface is a web access method of hive command line interface (CLI).

3.4 metastore

Metadata service component. This component stores hive metadata. Hive metadata is stored in a relational database. Hive supports relational databases such as derby and mysql.

The data in hive is divided into two parts. One part is real data, which is generally stored in HDFS; The other part is the metadata of real data, which is stored separately in the relational database.

3.5 driver

Driver is the core component of hive and the core of the whole hive. This component includes parser (interpreter: convert hive SQL into abstract syntax tree), compiler (compiler, compile syntax tree into logical execution plan), optimizer (optimizer, optimize the logical execution plan), Executor (executor, which cuts the logical plan into the executable physical plan of the corresponding engine and calls the underlying execution framework for execution).

3.6 applicable scenarios for hive

Batch operation of large data sets is the best place to use hive tools. The most typical application is network log analysis. In addition to encapsulating Hadoop at the bottom, hive also uses hiveql language similar to SQL to implement data query, which is essentially a data warehouse processing tool. Therefore, hive data will be stored in Hadoop compatible file systems: Amazon S3, HDFS, etc. During the process of loading data, hive will not make any changes to the data, but will only move the data to the preset hive directory in HDFS, which indicates that it does not support adding and rewriting data, and the data is determined during loading.

Differences between 3.7hive and traditional database

Before the birth of Hadoop, most data warehouse applications were based on relational databases, while data warehouse applications were data applications based on data warehouse, including report display, ad hoc query, data analysis, data mining, etc. Data warehouse data comes from database and is different from database. The main difference is that data warehouse is suitable for online analytical processing, which usually analyzes the historical data of some topics; The database is suitable for online transaction processing. It is usually used to add, delete, modify and query business data when the database is online. Hive is designed as a data warehouse. Its previous version or new version does not support transactions by default (system default state), and is generally not suitable for OLTP.

Hive is also different from RDBMS in other aspects. For example, in RDBMS, the schema of the table is determined when the data is loaded. If it does not conform to the schema, the loading will fail; Hive does not validate the data during loading, but simply copies or moves the data to the directory corresponding to the table.

4 Conclusion

This paper describes the demand background of radio and television big data user portrait, and introduces several common big data storage and analysis technologies in the current market. It focuses on the big data storage and analysis architecture technology of Hadoop and hive integration, and makes an in-depth discussion on hive from the aspects of hive's development, principle architecture, main characteristics, and comparison with traditional DBMS.

References:

[1] the 19th national symposium on the development of Internet and audio and video broadcasting and the 28th Annual Conference on the development of China's digital radio, television and network [j]China media technology, 2020 (03): 1

[2] Capriolo et alHive programming guide [m]Translated by kun Cao Beijing: People's Posts and Telecommunications Press, 2013

[3] Shuai Sun,Meijia Wang Hive programming technology and application [m]Beijing: water resources and Hydropower Press, 2018

[4] dark horse programmerJava foundation case tutorial [m]Beijing: People's Posts and Telecommunications Press, 2021

[5] dark horse programmerHive data warehouse application [m]Beijing: Tsinghua University Press, 2021

[6] Li Yang Big data hive offline computing development practice [m]Beijing: People's Posts and Telecommunications Press, 2020

[7] Hongzhi Wang,Chunjing Li Hadoop cluster program design and development [m]Beijing: People's Posts and Telecommunications Press, 2018

[8] Hong Mi,Ling Zhang Hadoop platform construction and application [m]Beijing: People's Posts and Telecommunications Press, 2021

[9] Fang Xiao,Liangjun Zhang Spark big data technology and application [m]Beijing: People's Posts and Telecommunications Press, 2018

[10] Junjie Li,Zhiming Xie Big data technology and application foundation project tutorial [m]Beijing: People's Posts and Telecommunications Press, 2017

[11] Yingnan Li Design and implementation of purchase and sales data warehouse system based on hive [d] Chongqing: Southwest University, 2020

[12] Hongsheng Yang Design and implementation of user personalized recommendation system based on big data [d] Nanjing: Nanjing University of Posts and telecommunications, 2020

[13] Yingying Wang Design and implementation of user behavior analysis system based on Hadoop [d] Beijing: Beijing University of technology, 2015

[14] Dazhou Li Design and implementation of user behavior log system based on big data [d] Nanjing: Nanjing University of Posts and telecommunications, 2020

[15] Qinqin Hu Research and application of data visualization technology based on Hadoop [d] Beijing: Northern University of technology, 2016