Design of Forum Log System Based on Big Data Analysis

Guanghua Yu(✉), Linan Sun, and Yongjuan Wang

Heihe University, Heihe City, Heilongjiang Province, China
Ygh2862@163.com

Abstract. Clicking on the log data saved by the website, using the big data means to analyze and mine the information stored by massive data, many crucial information of website operation can be learned. This paper adopts Hadoop distributed platform, uses HDFS data storage and Hive to analyze log big data, designs a Web log analysis system and expounds the design process of the system.

Keywords: Big data · Hadoop · The log

1 The Introduction

In the age of information and amount of data that can be accessed from web logs is getting larger and faster. However, with the traditional stand-alone analysis approach to log data, reading one T data will take a few hours, and the amount of data we need to process is much larger than this. Thus it can be seen, such a long waiting time can no longer meet the daily requirements. Based on the above problems, the big data technology can be well solved. It adopts a cluster to process massive data in parallel, compared with a single server to process data, which undoubtedly provides technical support of each forums and saves log processing costs [1].

2 Technical Introduction

2.1 Introduction to Hadoop

Hadoop is an open-source distribution framework developed by the Apache Software Foundation, which uses clusters to compete, analyze and store data. The core design of Hadoop framework is HDFS and Mapreduce. HDFS provides distributed storage for massive amounts of data, while Mapreduce provides analysis and calculations for data. At the same time, Hadoop can read the stored data quickly, while HDFS uses the method of data stream to read the data [2].

Foundation item: Philosophy and social science project of heilongjiang province (16EDC04);School-level topics(KJY202002)

X. Jiang and P. Li (Eds.): GreeNets 2020, LNICST 333, pp. 29–35, 2020.
https://doi.org/10.1007/978-3-030-62483-5_4

2.2 Introduction to Hive

Hive is an open source tool based on Hadoop for storing and processing large amounts of structured data that is presented as forms in Hive. Compared with traditional databases, Hive enjoys a larger scale of data processing and gives a support on using the collection of data such as map, struct and array. In addition, it will search for data with minimal header addressing, so it is fast in processing data over TB and PB, eliminating the energy consumption disadvantage of traditional database when searching data [3].

2.3 Introduction to Flume

Flume is a reliable, distributed log collection system that collects, aggregates and transfers data from large logs. It is a component of the Hadoop, with high ease of use and an open source tool. According to the need to modify the configuration file, system can achieve the receipt of different data sources, also can do some simple processing of the data, then the data will be transferred to the receiving place(HDFS receiving in this system).

3 The Overall Design of the System

The system is divided into three parts: data collection, data processing and data display. The overall function module is shown in Fig. 1.

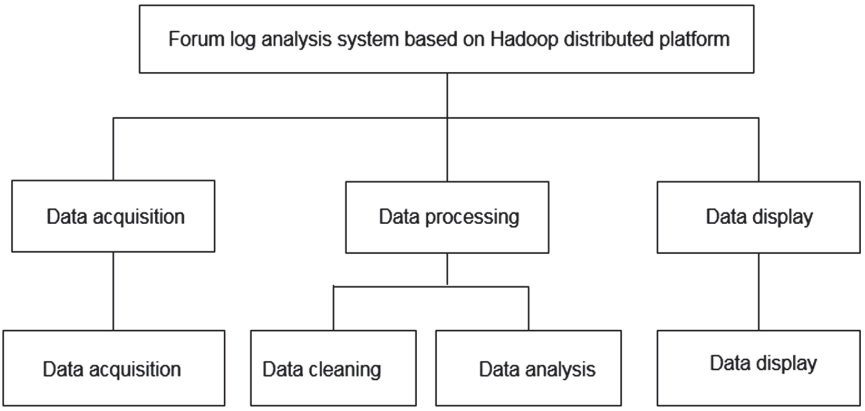

Fig. 1. Overall functional module diagram

4 Main Function Design

4.1 Data Acquisition Design

The system needs to adopt data once a day. For the convenience of classification, the system uses time as a catalogue to classify collected data and configure flume. Hadoop has the disadvantage of not being good at handling large numbers of small files. The system edit and control flume to make flume pass the data every once in a while or wait for the data to reach a certain size to pass the data once. The data acquisition process is shown in Fig. 2:

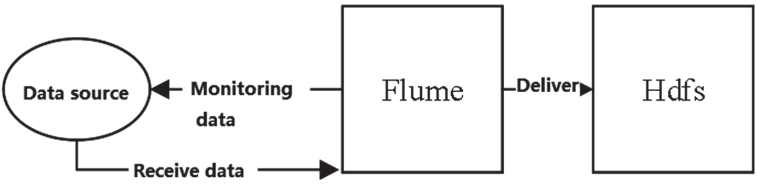

Fig. 2. Flow chart of data acquisition module

4.2 Data Processing Design

The first step: the log data in HDFS is cleaned with the written MR algorithm,and then cleaned data stored into HDFS.

The second step: Analyze the data in hdfs using Hive's hal statements. Write hiveudf code to analyze some data that cannot be analyzed with HQL statements, and then store the analyzed data in HDFS.

The third step: sqoop is used to transform the analyzed data and transfer to the MySQL database storage. The overall flow chart of data processing is shown in Fig. 3:

4.2.1 Algorithm Design of Cleaning and Analyzing Data

This system applies MR algorithm, which has two functions of mapping and reduction (map function and reduce function). Map function is to use the form of key-value pairs to carry out preliminary processing on the transmitted data and use the key to mark each row of log data. The value is the log data [5].

The Reduce function is the further processing of the data processed by the map function and the marked data is processed by Reduction. Combine the same data into a single line and merge once to count a number. Then combine the combined data with the corresponding count times to form an array and mark the array. Finally, the array is analyzed according to the code in reduce to get the required data and the data is stored in HDFS. The flow chart of MR algorithm is shown in Fig. 4:

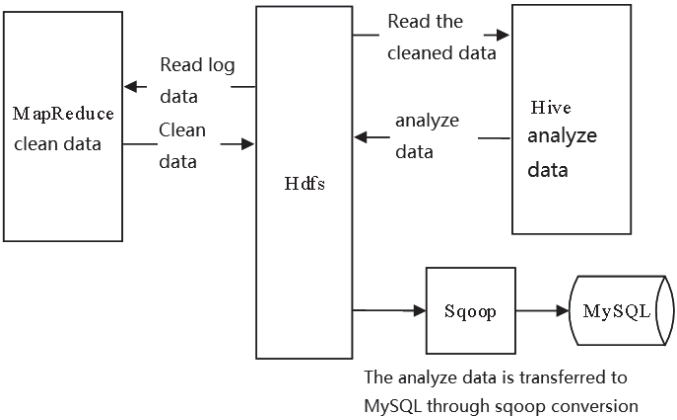

Fig. 3. Overall flow chart of data processing

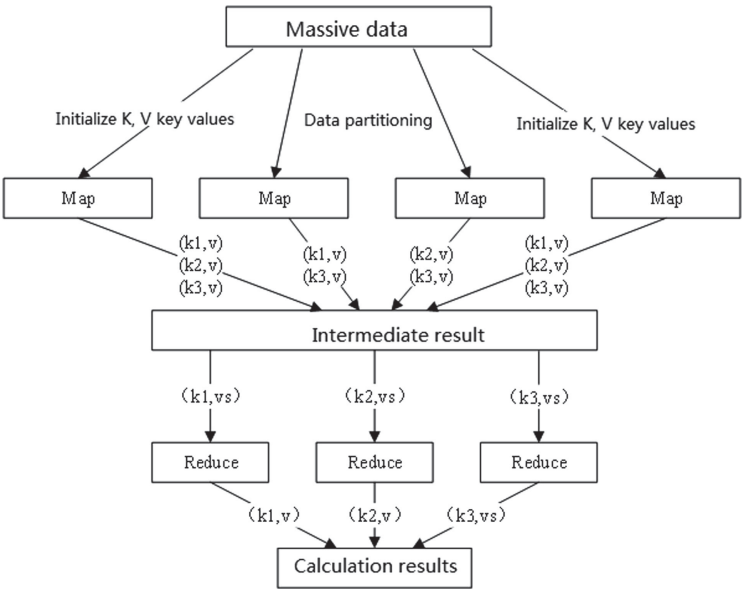

Fig. 4. Flow chart of MR algorithm

4.2.2 Detailed Design of Data Cleaning

Data cleaning is written by MapReduce to read the log files on HDFS. Then the log file is parsed into a Txt file. Clean up log data by writing map and reduce to get rid of useless fields, clutter, and clean up the data. Type the prepared MR program into a jar package, run the jar package, clean the data, and transfer the cleaned data to HDFS. The data cleaning process is shown in Fig. 5:

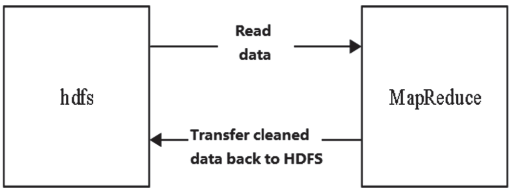

Fig. 5. Data cleaning flow chart

4.2.3 Detailed Design of Data Analysis

The first step in data analysis, read the cleaned data of MR in HDFS through Hive. The second step, design analysis methods based on requirements. The third step,The cleaned data is further analyzed using HQL statements and written hiveudfs. Get the data that is useful for the development of BBS, and store the analyzed data on the hdfs. The step 4, Sqoop is used to extract the data which analyzed by Hive and convert the data into a format that MySQL can recognize and store in the MySQL database. Data analysis flow chart is shown in Fig. 6:

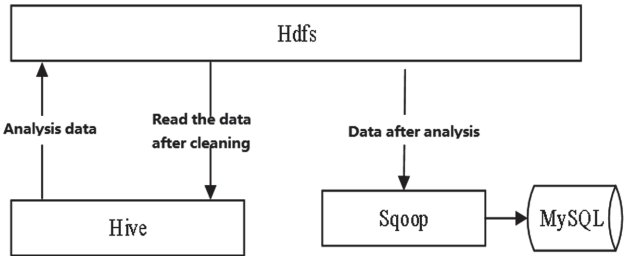

Fig. 6. Data analysis flow chart

Processing in Hive: the first step, according to the log data cleaned by MR in hdfs, establishing a Hive temporary storage framework. The second step is to analyze the log data by using HQL statements. The data that cannot be analyzed with HQL, we need to write hiveudf and put the compiled hiveudf into a jar package in hive's lib. Use the hql statements to run the self-built hiveudf jar package.

4.2.4 The Data Visualization Design

Use JSP technology and Struts2 framework to design different statistical analysis pages after obtaining the analysis results. The page provides the user with the query results. Page function module design can be based on different business needs, personalized business development. The data presentation process is shown in Fig. 7:

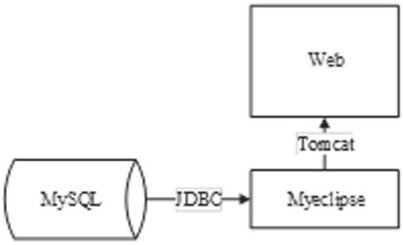

Fig. 7. Data display flow chart

5 Experimental Results

In order to verify the high efficiency of Hadoop for log analysis processing, an experimental comparison was made between a stand-alone machine and an HDFS cluster. Web logs of different file sizes were processed separately in the experiment and calculate the execution time to obtain the following data. The result is shown in Fig. 8.

Fig. 8. The result

6 Conclusion

Data processing is the comparison of processing time between Hadoop distributed platform and stand-alone platform. As the data set grows, the processing time of the Hadoop

platform becomes shorter. Since the time consumed by the system to start the MapReduce task is negligible in the case of large data sets, the computing efficiency is relatively high.

References

1. Yanhui, M.: Big Data Technology Foundation. Tsinghua University Press, Beijing (2016)
2. Wbite, T.: Hadoop: The Definitive Guide. O'Reilly Media, Sebastopol (2015)
3. Rutherglen, J.: Hive Programming guide.People Post Press (2013)
4. Shenzhi, S.: Research on Web Log Data Analysis System Based on Hadoop. Xidian University
5. [Apache Flume] Official document http://flume.apache.org/