Decision analysis of the weather log by Hadoop

Hao Wu
Department of Computer Science, JLUZH, Zhuhai, Jinwan519041, China
E-mail: haowu_mouse@hotmail.com
www.jluzh.com.cn

The data mining ability on Hadoop is much more rapider than that of traditional method, and coding is much simpler. It is mainly to quickly analyze the weather log using the parallel processing on Hadoop. The system realizes the analysis and processing of the weather log, utilizing the Hadoop core technology of Hadoop distributed file system (HDFS) and the MapReduce operation framework. Combined with Java Web programming techniques and ideas, it forms the B/S structure system and utilizes the Highcharts plug-in to generate charts to display to the user. According to the data graph, the user can predict the temperature change and make decisions in the future. This function is perfect, and its graphical interface is good, which can improve the operability and the interactivity, so as to reduce the burden on the user. Its usage is more convenient and much faster, and the operating is much simpler. It can objectively show the results of the analysis directly.

Keywords: Hadoop; MapReduce; Log; Decision Analysis.

1. Introduction

Any system can produce a lot of logs in the operation process, which can obtain more valuable information from the mass log data preserved on the server. Due to saving the log data recorded on increasing and accumulating every day, it is necessary to provide a system that can extend process logs to be more efficiently and more rapidly. Hadoop is a distributed storage system, which can store the data in different machines, and put the terabytes of data parallel processing, high efficiency and saving time. It will get higher commercial value in future, through the Hadoop distributed system platform finding the useful information on the system logs.

Making use of the decision analysis weather on the Hadoop log system, it can analyze the maximum and the minimum temperature in the annual. To help people understand the change of temperature and to predict the future tendency

can make decisions. The data mining ability on Hadoop is much more rapid than that of traditional methods, and its coding is much simpler. It is mainly to analyze the weather Log quickly by the parallel processing on Hadoop. It adopts the core technology of HDFS and MapReduce, and realizes the analysis and dealing with the weather log. This design is combined with Java Web programming techniques and ideas, forms the B/S structure and takes advantage of the Highcharts plug-in to generate charts to display to the user. According to the data graph, the user can predict the temperature change and make decisions in future. This function is perfect, and the graphical interface is good, which can improve the operability and the interactivity, and reduce the burden on the user. It is more convenient, much faster, and its operation is much simpler, that can show the results of the analysis objectively and directly.

2. Main Methodology

2.1. *Linux Fedora*

The work of the Fedora project leads the innovation and the free communication code, which is synonymous with interests in the world, and is used to establish friends in the free software community.

2.2. *Hadoop*

HDFS has the characteristics of a high fault tolerance, designed to deploy on cheap hardware. It provides a high transfer rate to access the application data that is suitable for the program of the very large dataset. Hadoop is able to handle with huge amounts of data. In addition, Hadoop is not a high requirement to the machine hardware and its cost is low, so anyone can use it.

2.3. *MapReduce*

MapReduce is one of the main framework of Hadoop, that is used to the parallel computing for the big data by adopting the idea of "divide and rule ". MapReduce is mainly to divide the data into Key/Value pairs < Key, Value >, then to put the same Key Value to merge. In the phase of Map the inputting data is splitted into< Key, Value > Key/Value pairs in a certain pattern, and then in the phase of Reduce the Value is merged on the same Key.

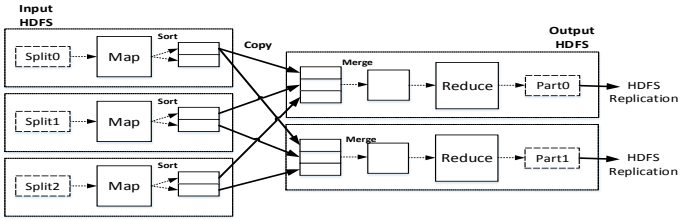

Fig. 1 MapReduce working principle

2.4. *HBase*

HBase is a high-performance, a high reliability, a column storage and a scalable distributed storage system. HBase can be used in low PC server to build a large structural storage cluster. In addition, the HBase data storage is loose. To put it simply, it makes use of the simply configuration of hardware, and can realize to process the large database consisting of huge amounts of rows and columns.

3. Deploying the Server Environment

It adopts the Hadoop distributed completely that there are 3 machines and their user names are all Hadoop. Fedora version is more convenient to be installed, and its graphical interface and installation are the same as windows. Hadoop cluster includes three nodes. One is Master, the other two are Salve. Nodes are connected by the local area network (LAN), and can ping each other. The node information is as follows.

Tab. 1 The cluster information table

Machine name	IP address
Master.Hadoop	192.168.1.67
Salve1.Hadoop	192.168.1.65
Salve3.Hadoop	192.168.1.66

The Master is mainly on two roles of NameNode and JobTracker, to undertake the main distributed data and decomposition of the task execution. Three Salve machines deploy the role of DataNode and TaskTracker, to be responsible for the distributed data storage and the task execution.

4. The System Analysis

4.1. *System requirements*

The requirement of the weather log analysis system on Hadoop is to combine small files into a large file and to process, that can analyze the annual maximum and the minimum temperature. The graphical interface is good and can improve the operability and interactivity, which is to reduce the burden of operators and the operation to be risen more convenient, faster and simpler. It can display the analysis results objectively and directly, and can be downloaded to view.

4.2. *The system function analysis*

It is listed the system function as Figure 2:

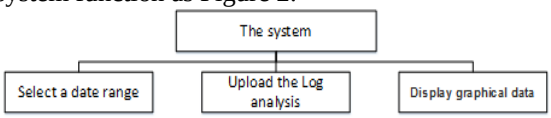

Fig. 2 The system function analysis diagram

5. Overall Design

5.1. *The system structure*

At the client the user uploads the data wanted to view on the Hadoop cluster, and the backend code analysis is to be stored in HBase. The background code reads the data from that database to return the client that is showed in the form of chart at the client. Such as Figure 3.

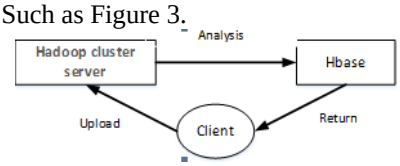

Fig. 3 The system strcuture diagram

5.2. *The overall functional design*

The overall goal of the system development is to use the Hadoop platform analysis and process the weather Log, to find out many years and the temperature corresponding to the maximum and the minimum temperature in annual and to display the data objectively. It realizes the systematic, standardized, scientific and automatic objective, so as to improve the temperature prediction. The function is comprehensive, and the interface is simple, easy to be operated.

5.3. *The algorithm thought*

The data is cut according to certain rules, and then to take out the year and the temperature. The year is set as a key, and the temperature is set as a value, that is <year, temperature> key/value pair as the value of temperature. The value of the same year will be merged before the output is reduced, and it compares the maximum and the minimum value in Reduce at this time.

5.4. *The idea of uploading and merging data*

That HDFS provides API can be used to realize to upload local files to HDFS. When uploading, it can read the data of each small file, and then write to one file. So it can put the small file merged into a large file.

5.5. *The Database design*

HBase is a column database and need not set the data type. The line key in HBase can be sorted automatically, and so as MapReduce, that is shows that HBase can be combined with Hadoop very well, so the year value is to be the line key which will be sorted from small to large and these data is intuitive. The highest temperature is to be column name and the lowest one is to be value. That data is sent back to customers in a short time can reduce the network I/O request, thus to improve the transmission speed.

Tab. 2 The table structure

Line key	Column cluster	
Year	The highest temperature	The lowest temperature

6. The Detailed Design and Implementation

6.1. *Data preparation*

To open the file to check the data format, the characteristics of the data can be found that "year-month-day hour-minute-second temperature", and it is to record a temperature in every two hours. Such as Figure 4. The graphic design uses plug-in Highcharts generate mainly on the page references. Such as Figure 5.

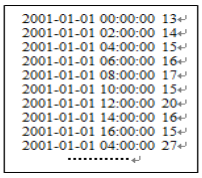

Fig. 4 Data Content

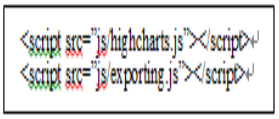

Fig. 5 References plugin highchair

6.2. *The algorithm design*

According to the log format, this code will add the space into the inputting data value as "year-month-day hour-minute-second temperature" that is deposited into an array. The first part is to be the year cut up "-", so the year can be got. The third can get the temperature, then the year and the temperature composite the new key/value pair <key, value> will be passed to Reduce. For example, the inputting data is 2005-01-01 00:00:00 13, 2005-07-02 12:00:00 29, 2005-12-01 02:00:00 6, 2006-03-01 00:00:00 12, 2006-06-01 08:00:00 16, 2006-09-01 14:00:00 30 which is sent into Map will become <2005,13>, <2005,29>, <2005,6>, <2006,12>, <2005,16> and <2005,30>.

These six groups of data sent to Reduce will be merged into two sets of data of <2005, 13 29 6> and <2006, 12 16 30>, which is a mechanism of graphs framework. According to the same key, it is merged into the value automatically. The temperature in 2005 will not appear in 2006, so as to the same that which in 2006 doesn't appear in 2005. The Reduce will get the value from that two groups of data and compare their value, so as to determine the maximum and the minimum value. So the outputting data after Reduce will become <2005,6 29> and <2006,12 30>. At this time it need only put the value into the database one by one.

```
Public void map(Object key, Text value, Context context)
        throws IOException, InterruptedException{
    String[] str=value.toString().split(" ");
    If(str[0].matches("\\d{4}-\\d{1,2}-\\d{1,2}")){
        String st=str[0].split("-",2)[0];
        Context.write(new Text(st),new Text(str[2]));
    }
}
```

Fig. 6 Map algorithm

```
Public void reduce(Text key,Iterable<IntWritable> values,Context context)
        throws IOException, InterruptedException{
    int maxvalue=Integer.MIN_VALUE;
    int minvalue=Integer.MAX_VALUE;
    while(values.iterator().hasNext()){
        String s=values.iterator().next().toString();
        String str="";
        Pattern p=Pattern.compile("\\s*|\t|\r|\n");
        Matcher m=p.matcher(s);
        str=m.replaceAll("");
        int a=Integer.parseInt(str);
        if(maxvalue<a){
            maxvalue=a;
        }
        else if(minvalue>a)
        {
            minvalue=a;
        }
    }
```

Fig. 7 Reduce algorithm

6.3. *The running process and results*

The System will merge the content according to the data accepted. In the file of Log the user can select the date range of merger. In the log folder the time interval files will be uploaded to HDFS and be handled with MapReduce, to analyze the annual maximum and the minimum temperature. The data in HBase will read in the background, and be sent to the front desk. That result will display in the form of line chart, to let it objectively and readable.

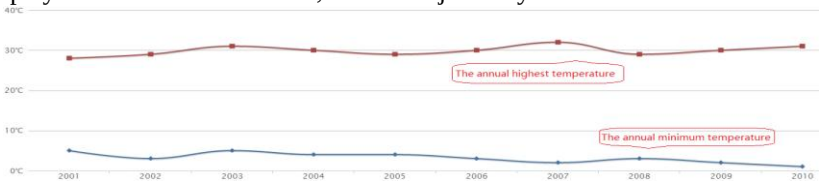

Fig. 8 Displaying the system data diagram

7. The System Testing and Implementation

7.1. *The system running environment*

The system is suggested to adopt the following configurations. The minimum configuration is that CPU is above CII300, the memory is 1GB and Linux. The suggested configuration is that hat CPU is above G3220, the memory is 2GB and Linux.

7.2. *The system performance testing*

The purpose of the system performance testing is that the efficiency of processing and analyzing files in the same size and different copies. The design and result of the testing case is as follows.

Tab. 3 Designing the testing case

Cases	The size of files Uploaded	The copy of files Uploaded
1	5M	5
2	25M	1

7.3. *The testing process and methods*

Accessing the system can view the status of the cluster, as well as the information detailed in each of MapReduce, including the running time, the size

of the file, the utilization rate of CPU and the memory usage, etc. Using its own regulatory mechanism, it can be used to detect the efficiency of the system on the same file size and different copies.

(1) Usecase1: To upload 5 files of the weather log as 5M, it is to check the processing time of runing and opening the web page.

(2) Usecase2: To upload 5 files of the weather log as 25M, it is to check the processing time of runing and opening the web page.

To compare with these two operational time, it shows that the efficiency of the usecase2 is higher than the usecase1 obviously. The results are as follows.

Tab. 4 Table testing reslults

Usecases	Time/Second
1	2.8
2	1.2

After testing, it is found that the processing speed of the less copies of file is faster than these in the same file size and different copies. The conclusion is that Hadoop is not suitable for processing a large number of small files, but to process large files, which spends less time and the efficiency is higher. It is because that the Hadoop is chunking for data processing, which default is 64M. If there are lots of small data files, such as a file of 2-3M, that a small data file is far less than the size of a block of data has to process as a block of data. Storing a large number of small files occupies storage space, so its storage efficiency is not high and the retrieval speed is slower than a large file. Such small files consume the calculable capacity in MapReduce, because it is to allocate the Map tasks in a block.

8. Conclusion

In the process of the study, it uses Hadoop technology in the era of the cloud. After building a Hadoop platform, it adapts the core technology of HDFS and the operational framework of MapReduce to handle the analysis of the weather log. Combining with the programming techniques and ideas of Java Web, it forms a B/S structure and uses the plug-in Highcharts to generate charts.

After testing it finds that Hadoop is not good at handling small files, especially a large number of small files. The size of the file is bigger than HDFS block size, because of MapReduce computing in such a small file consumes the calculable capacity, which default is allocated on the Map tasks in a block. This research is mainly aimed at the weather Log, which is single in the content and in the format. The format must be the "year-month-day hour-minute-second

temperature", so it does not apply to others logs. If to analyze other types of logs, it has to modify the algorithm. Such as analyzing and statistics the count of the wrong class in the using process and in a certain period of time, it can analyze the seriousness of the Bug. To let programmers modify them in purpose, it will make the system perfectly.

Acknowledgment

Thank Zhuhai College of Jilin University for providing the Funds on the training project of both the teaching and the research for young teachers. Thank Zhuhai Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education.

References

1. M. Kaplan; M. Hudaverdi, “The correlations of space weather and satellite anomalies: RASAT, ” 2013 6th International Conference on Recent Advances in Space Technologies (RAST), 711-715, 2013.
2. C. Naik; A. Kharwar, “Knowledge discovery of weighted RFM-QR sequential patterns with multi time interval from customer sequence database,” (IC4), 2015 International Conference on Computer, Communication and Control, 1-8, 2015.
3. S. Sharma; V. Mangat, “Technology and Trends to Handle Big Data: Survey,” 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 266-271, 2015.
4. M. Pardini; A. Cantini; F. Kugler; K. Papathanassiou; F. Lombardini, “Monitoring dynamics in time of forest vertical structure with multibaseline PolInSAR data,” 2014 IEEE Geoscience and Remote Sensing Symposium,366-3369, 2014.
5. S. Sharma; V. Mangat, “Technology and Trends to Handle Big Data: Survey,” 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 266-271, 2015.