2014

ISSN : 0974 - 7435 Volume 10 Issue 19

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(19), 2014 [11354-11360]

Research of data mining technology based on hadoop platform

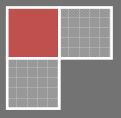
Dezhi An, Guangli Wu, Jun Lu
School of Information Engineering, Gansu Political Science and Law Institute, Lanzhou, 730070, (CHINA)

ABSTRACT

With the constant development and improvement of information technology, the Internet technology has become the essential ingredients of people's daily life in current days. Due to the traditional computer architectures can not meet the needs of the majority of people in front of deal with big data, the cloud computing has been come up with to provide an effective solution of handling and processing a huge data. Hadoop platform is a major project developed by the Apache Foundation, and is used in a cluster of general business computers commonly, And its most significant characteristics are of a super calculating power, flexible storage capacity and various scheduling capacity. On this basis, the data mining technology supported by Hadoop platform, after being deep processed on its model, has entered the development phase of the information age. The present paper, based on the research of Hadoop platform and in the study and application of the platform, makes an in-depth research on the algorithm. And finally a corresponding operation platform is set up, and a Hadoop version which can run well is offered, the purpose of which is to provide an effective basis for data mining personnel in the application of this platform.

KEYWORDS

Cloud computing; Data mining; Hadoop.



CLOUD COMPUTING TECHNOLOGY

The definition of cloud computing

Cloud computing is a new thing in today's information age, and is currently in the early stages of development. As for the definition of cloud computing, scholars do not provide an unified the answer for the moment. But related enterprises and research institutes, according to the nature of their company's products as well as their own understanding towards cloud computing, summarizes and generalizes the several representative definitions of cloud computing, namely:

(1) In the understanding of U.S. national institute of standards and technology, cloud computing is a model, which is ubiquitous in the information age, and is convenient shared platform allocated by space requirements. It is also a configurable resources space provided by network access. These resources space can be distributed and released quickly, and do not participate in the interactive sessions of corresponding service. The above definition is put forward by Mell and Grance in 2011[1].

(2) In the white paper of IBM cloud computing, cloud computing is defined as a computer terminology used for platform description and application. It can, according to the specific needs, dynamically provide the appropriate server, and can according to demands of the users, configure or cancel the operation of the server. Among them, the "cloud" server has two forms, namely physical machines and virtual machine[2].

(3) In Wikipedia, the cloud computing is defined as a computing method over internet, whereby shared resources are provided to computers and other devices according to one's needs.

Hadoop platform in apache project

Hadoop platform is a major project developed by the Apache foundation's, which has relatively large information resources and contains many components. Using Hadoop platform, although without particularly understanding towards the details of the system is not, one still can undertake the development and application of a simple small program and make a process and handle of data resource stored in the computer in the cluster[3]. One of the typical features of Hadoop platform is that it has high fault tolerance, which makes Hadoop platform can be adapted to the low price of desktop computer, as well as installed in expensive mainframe. Hadoop platform scale has diversified characteristics, being able to be used from a single mode of computer to thousands of servers. And a computer based on Hadoop platform could combine every local machines and the memory at each node into a larger cluster. The ecosystem of the Hadoop platform is shown in Figure 1.

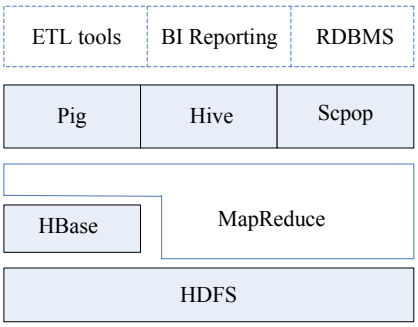


Figure 1 : The ecosystem of the hadoop platform

From the ecological system of Hadoop platform in Figure 1, it is figured out that HDFS, graphs and HB existed throughout the Hadoop platform as fundamental parts. Among them, the Hive in the software of the whole system is a kind of high level language.

INTRODUCTION OF HDFS FUNCTIONS

Cluster structure of HDFS

The main feature of HDFS architecture is to provide effective solution for data management in the system. The main function of HDFS is to store information of the users, and provide corresponding service for the components in the system[4]. The main part of a perfect effective HDFS cluster consists of four components, namely:

(1) NameNode, in HDFS cluster, is called control node. Its main task is to synchronously process the scheduling and management of the file in system. There can be only one control node in the cluster.

(2) Secondary NameNode, which is the backup node of NameNode, mainly makes backup for the control node in the system. If anything unusual happens to NameNode, the cluster will switch to the node of Secondary NameNode. And its purpose is to ensure the effective work of the whole system.

(3) DataNode, is defined as work node in the model. And its primary function is to store data of the users. In this node, the data is stored mainly in the form of blocks, the default storing size of which is 64m. But it be changed through the file configuration.

(4) Client is defined as the user interface in the model. The role of the user interface is to interact with the cluster system, and to undertake the operations of reading and writing system files. It is a work way formed between the NameNode and DataNode, which is called as the Master - Slave structure[5].

The workflow of HDFS

File reading process in HDFS architecture is divided into the following several steps, namely:

(1) Access the control node through the client, and then inquire the concrete location of the document in the system, namely which work node in the system that the document is deposited in. For the reason that the document may correspond to multiple blocks, it needs to return to addresses of all the blocks after the final inquiring.

(2) The client obtains the required data information in the work node of the system in the order of sequence, starting from the first one, then connecting the next block, until the last one. In this way, it makes an effective access to all the required data information[6].

(3) After reading the data information needed by the client, the node needs to be shut down. By now, the file reading task has been completed smoothly. And information reading process in the HDFS architecture is shown in Figure 2:

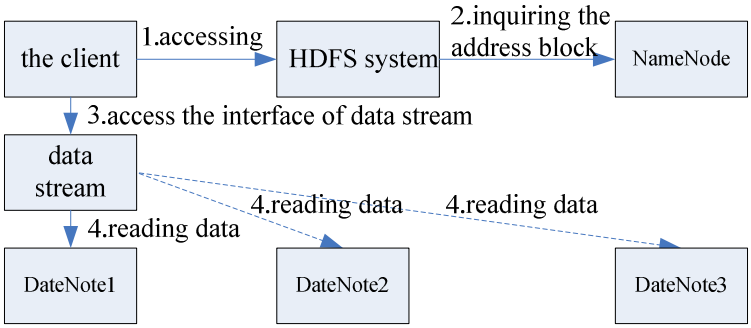


Figure 2 : Information reading process of HDFS

In this model, the process of writing data information is relatively complex, with the specific operation steps as follows:

(1) Send the request of writing data information by the control node of the client;

(2) By way of comparing fingerprints in the system, the control node will determine wheter the inquiring file exists. If not, the control node will create a new file identifier or eventually the system will prompt that the state is abnormal[7];

(3) While writing the information data, the client will divide the file into multiple sub-blocks. And the default space size of each sub-block is 64M. After division, the request of using available space will be submitted to the control node. At last the control node will make feedbacks of the specific address of the work node;

(4) Then, the client will write the data information into the work node mainly in the form of flow. Firstly, write the data into the first work node, then the second work node till the last one;

(5) After all writing tasks of work nodes in the system is completed, the system will return to the client a confirmation package, which marks the termination of information writing. The HDFS file writing operation is shown in Figure 3.

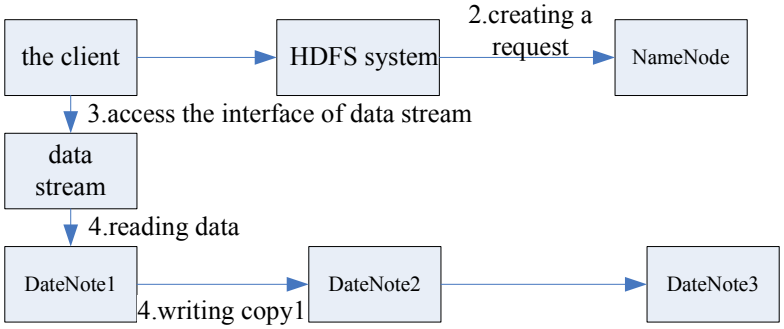


Figure 3 : File writing operation of HDFS

DATA MODEL OF THE HBASE SYSTEM

The storing way of the traditional data model is in the form of row. The storage unit is the row, That is to say, the two columns of data in one row are also adjacent to each other in physical storage. This way of storage process has many defects in the concrete operation, which leads to its availability being effected to some degree during handling the huge data

information by database system. With a growing number of fields in data table and data information in the system, the usual way is using dividing tables and depots to share the pressers of the data information in database. At this time, the workers of database introduce the column-oriented storage model into database[8].

The main feature of column-oriented storage method is storing information based on field and its unit is column. Two rows of adjacent data in one column are physically adjacent to each other, which is same with row storage. In column-oriented storage, the columns can be dynamically increased. In the operating process, if one column is empty, it is allowable to store no data information; If a column need to be increased, only adding an identifier during storing data can do the job. The most outstanding characteristic of the above way of storing is that it can save a lot of resources and space for the system.

For example, in this case of inquiring the data in a certain field, if the traditional storing way is used, the first stop is to read information of each row, then do the corresponding operation according to the specific requirements of the fields. In this situation, all the data of every row should participate in the data inquiring process. If using column-oriented approach to inquire, one just needs to make an effective extraction of the data information in this column, and then the machine and process the data. The whole processing process does not involve fields of the entire column.

In addition, the biggest advantage of the column-oriented storage is the high feasibility in concurrent reading and writing operations. When the data in the table pass a certain threshold, the system will automatically carry out a segmentation process, promoting the scalability of data storage, and improving the flexibility of data information inquiring. The recorded content of a specific data table IS shown in TABLE 1.

TABLE 1 : A data table and one piece of the recorded information

Number	Property 11	Property 12	Value label
row1	11	122	Value

In the process of HBase storage, as shown in TABLE 2, the value of timestamp t2 is smaller than t3. There is only one row of data information in the following table with the label as row1. That is to say that every logical modification of the data is corresponding with a timestamp. There are four columns in this table: column:l1(t1), column:l2(t2), column:l2(t3)label: type(t4). The column-oriented storing method is used in the following table, and only one piece of data is store each time. In the table below, as t2 is smaller than t3, the value of column:12,namely the 122, needs to be extracted from the HBase system.

TABLE 2 : The storing method of HBse system (column-oriented storage)

Number	Timestamp	Column	Value label
row1	T1	column:l1=”l1”	label: type=””value
	T2	column:l2=”l21”	
	T3	column:l2=”l22”	
	T4		

NAÏVE BAYES

The main purpose of making classification algorithm on the data in the system is to, based on the relevant features of the data and predefined data categories, construct a classifier, whose function is to directly map the sample of an unknown data type into a given category[9].

The main stages of the training and testing process is the core of the composition of the classifier., In the training process, the system, through the analysis of the training data set, makes the corresponding construction of the model. And the expression of a specific training sample is $(\mu_1, \mu_2, ..., \mu_n; c)$. Each training sample is labeled correspondingly. And the monitoring and learning processes are included in this process. In the test process, the data of the model needs to be tested using the naïve Bayes, with the specific calculation processes shown below:

(1) Assuming that the total number of samples is L, the property number of each sample is n, and the clarification number of the samples is m, namely $c_1, c_2, ..., c_n$,then each sample can be described with n+1 property vector, namely $X = (x_1, x_2, ..., x_n, C)$.

(2) If the sample-X which has not been clarified needs to be forecast, then the following condition needs to be meet:

$$P(C_i \mid X) > P(C_j \mid X) \quad 1 \le j \le m, j \ne i$$

Wherein, because the probability of the situation that X belongs to C_i is bigger than that in other category, $P(C_i \mid X)$ needs to be maximized. According to the bayesian formula:

$$P(C_i \mid X) = \frac{P(X \mid C_i) \times P(C_i)}{P(X)}$$

SPECIFIC BUILDING PROCESS OF EXPERIMENTAL PLATFORM

Hardware configuration

The used computing cluster is constructed by IBM blade. There are 34 nodes in total, one control node, one input and output node and 32 computing nodes. There are three networks in the cluster: gigabit management network, 40GB high-speed network and gigabit storage network. This study mainly aims at personal computer, so the program running and test on the basis of the personal computer. And the operation is of high performance and high feasibility.

(1) The relevant configuration of personal computer:

Hardware configuration: laptop, 1.73 GHz CPU, memory of 2G.
Operating system: Windows XP
Basic software: J2SE6.0, Eclipse3.5.
Blade interactive software: Putty 0.6, X - manager, F-Secure SSH File Transfer.

(2) The configuration of platform environment is as follows:

Single node hardware configuration: Intel Xeon E5504 2.00 GHz CPU, memory of 8 G, hard drive of 120G.
Operating system: RedHat Linux ES5.
Basic software: J2SE 6.0, OpenSSH4.3, Eclipse 3.5.
Platform software: Hadoop 0.20.2, HBase 0.90.5, Hive 0.8.1.

Accessing without password has been configured between each node in the platform, and the storing way between the various nodes is Sharing storage.

The construction of hadoop platform

There are mainly three models of Hadoop in a computer environment, namely stand-alone mode, pseudo-distributed mode and fully distributed mode. The stand-alone mode refers to the independent launching operation made by Hadoop platform to a procedure in a single node; Pseudo-distributed mode can launch a single node when the platform runs multiple independent processes. The common characteristic of these two modes is that they make a distribution process towards the system on single machine. And being analyzed from the point of the computing, they cannot be called as a completely distributed computing. So the major usage of the two modes is for learning and testing. However, the main characteristic of the fully distributed mode is to construct cluster for the node in multiple systems, which is a substantial computation and owns all the features of cloud computing platform. And the fully distributed mode is used in the presented paper for the data construction.

In the operation of Hadoop platform, the primary node needs to protect the corresponding belonging node. And after launching the Hadoop platform, the control node needs to start and stop the procedures in the work node through SSH.

The configuration of the names of the machine

In the integrating process of Hadoop platform, in order to make an effective and convenient management of each node by the program, the IP addressed in the system is replaced by the names of the machine in all the configuration files. The main characteristic of this approach is that mapping the name of the machine into the IP address in the system. And the main method is adding the following text into the/etc/hosts file at each node, with the specific configuration method is shown in TABLE 3.

TABLE 3 : The configuration file/etc/hosts

Configuration File-1/Ect/Hosts	
	192.168.10 console
	192.168.101 node01
	192.168.102 node02
	192.168.132 node32

The configuration process of hadoop

The default configuration form of Hadoop platform downloaded from the official website is stand-alone mode. And the biggest defect of this mode is that it could not be completely distributed, so the mode needs to be changed accordingly, with the changing steps are as follows: add the file-conf/hadoop-env.sh into the installation directory of export JAVA_HOME=java, and add console to the file-conf/masters.

The hadoop startup and the method of checking operating information

After completing a local configuration of the Hadoop platform, the information of each node needs to be confirmed, and a backup operation needs to be done to the node information in the console, and node01 ~ node32.

(1) Doing the related operations at console node in the system. Namely, run the file- bin/start-all.sh, which starts the related procedures in Hadoop system, and then run the file-bin/stop-all.sh, ending the Hadoop process. TABLE 4 shows the related information of the file of configuration.

TABLE 4 : The information of configuration file-conf/core-site.xml

Configuration File-2:Conf/Core-Site.Xml
<configuration> <property> <name>Hadoop.tmp.dir</name> <value>/tmp/Hadoop </value> <description> < Hadoop location for temporary files </description> </property> <property> <name>fs.default.name</name> <value>hdfs://console:9000 </value> <description>hdfs's Data management port </description> </property> </configuration>

(2) What need to be stressed is that not all nodes in the experiments are involved in the process, and usually only 2 to 4 nodes are needed. That is to say, after ending the process of Hadoop system, file-conf/slaves in the system needs to be modified, and the needed node name should be saved. Then, restart the Hadoop platform. For example, change the content of file-slave into that of TABLE5-6. After restarting Hadoop, there are only two nodes: node01 and node02. And the TABLE 5 is an example of a configuration file.

TABLE 5 : The configuration file- conf/hdfs-site.xml

Configuration File-3:Conf/Core-Site.Xml
<configuration> <property> <name>dfs.replication</name> <value>3 </value> <description> The number of copies of each file saved </description> </property> </configuration>

(3) Inquiring information requires inputting the corresponding website addresses at console node in the system. And the specific address is: http: // localhost: 50030. The address of inquiring HDFS information is http: // localhost: 50070. Figure 4 shows the showing picture of the Web in the operating state of MapReduce.

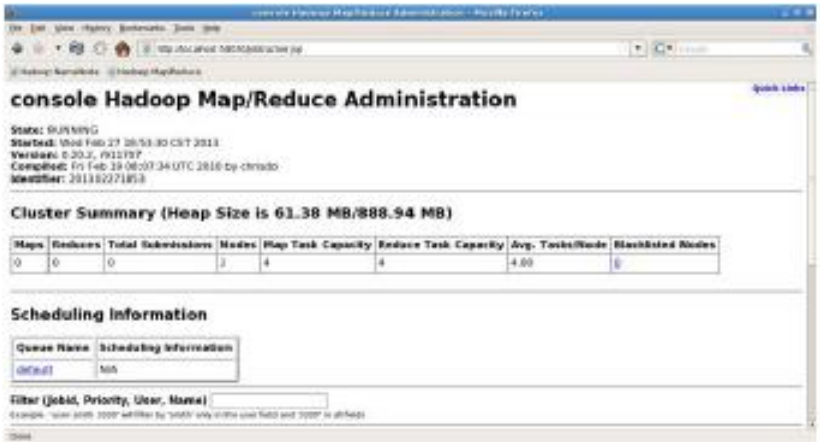


Figure 4 : Web in the operating state of MapReduce

CONCLUSION

This study is mainly to make a simple introduction of cloud computing, then introduces Hadoop platform which is the core of this And several core components of the platform are analyzed as well. Form the point of view of the data digging in cloud computing, the present situation is analyzed. And on this basis, make a deep computation aiming at naïve Bayes. The present paper, based on the research of Hadoop platform and in the study and application of the platform, makes an in-depth research on the algorithm. And finally a corresponding operation platform is set up, and a Hadoop version which can run well is offered, the purpose of which is to provide an effective basis for data mining personnel in the application of this platform. Hadoop platform is a major project developed by the Apache Foundation, and is used in a cluster of general business computers commonly, And its most significant characteristics are of a super calculating power, flexible storage capacity and various scheduling capacity. On this basis, the data mining technology supported by Hadoop platform, after being deep processed on its model, has entered the development phase of the information age.

ACKNOWLEDGEMENT

The national natural science foundation of China, Research of the key technology of large-scale depth calculation group social dimension under the Web environment sensitive (61363024).

REFERENCES

[1] Sun Mu; Dumbo in the Cloud [J], Programmer, **2011(10)**, 100-102 **(2008)**.

[2] China Internet Weekly; The graph of data service of facebook [J], China Internet Weekly, **10**, 28-29 **(2012)**.

[3] Li Chao, Zhang Mingbo, Xing Chunxiao, Hu Jinsong; Survey and review on key technologies of column oriented database systems [J], Computer Science, **(12)**, 1-7 **(2010)**.

[4] Liu Hongyan, Lu Hongjun, Chen Jian; A scalable classification algorithm exploring database technology [J], Journal of Software, **(6)**, 1075-1081 **(2013)**.

[5] Zhang Cheng, Guo Yi; Data mining and cloud computing——an interview with dr.he qing from the institute of computing technology, Chinese academy of science [J], Digital Communication, **38(3)**, 5-7 **(2011)**.

[6] Zhang Jianxun, Gu Zhimin, Zheng Chao; Survey of research progress on cloud computing [J], Application Research of Computers, **27(2)**, 429-433 **(2010)**.