Research on Database Massive Data Processing and Mining Method based on Hadoop Cloud Platform

Zhao Xiaoyong* and Yang Chunrong

Mathematics and Computer Science Institute, XinYu University, JiangXi, 338004, China

Abstract. This paper establishes the massive data processing mathematical model and algorithm of cloud computing, and the Hadoop distributed computing method is introduced to the database management system, to realize the automatic partition database data and master-slave node set. The master-slave nodes distributed algorithm is complied by using the MATLAB software realizing the data distributed computing function, and through the numerical simulation, we can compute the data processing speed, transmission rate, capacity and other system parameters. Compared with the Hadoop distributed processing algorithms and two kinds of traditional data processing algorithm, we can find that the data processing speed of Hadoop distributed computing algorithm is faster than the general algorithm, the amount of information storage, information transmission speed, which can satisfy the need of high data processing.

Keywords: Hadoop; Cloud computing; Massive data; Large capacity; Distributed computing; Master slave node

1. INTRODUCTION

With the development of computer internet communication computing, communication system needs to deal with a very large data. For the massive data processing, the server will consume a large amount of computer resources, and many traditional computing platforms cannot complete the mass data processing [1,2]. Combined with the calculation function of Hadoop distributed, a cloud computing data processing platform of high-speed processing mass data is designed by using cloud computing and cloud storage technology [3]. At the same time, combined with master-slave nodes automatic assignment, the platform uses the automatically partitioned database to realize a high data capacity and transmission speed data processing, which provides a new scheme for the massive data processing algorithms and data communication technology.

2. OVERVIEW OF HADOOP CLOUD PLATFORM DATABASE MASSIVE DATA PROCESSING AND MINING METHODS

The evolution process of the personalized internet causes a massive data, and the traditional single super server in the face of massive data has gradually fallen short, so the processing massive data has become a thorny problem [4]. The characteristics of open source Hadoop cloud platform developed by Apache foundation research brings endless possibilities for the huge amount of data processing methods, this paper uses the Hadoop data processing platform and combines with distributed data processing algorithm to establish the massive database processing data platform, the main structure is shown in Fig. (**1**).

Fig. (**1**) shows the schematic diagram of Hadoop massive data processing, it can be seen from the chart that the data processing of massive data Hadoop mainly consists allocation computing algorithm design and master-slave distributed, these two techniques can successfully achieve high speed processing of massive database data.

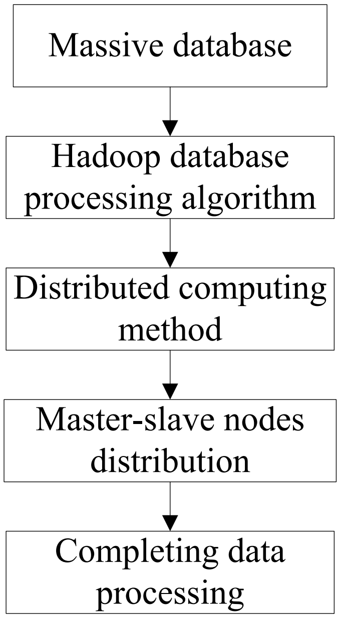

Massive database
Hadoop database processing algorithm
Distributed computing method
Master-slave nodes distribution
Completing data processing

Fig. (1). Hadoop massive data processing.

1874-4443/14 2014 Bentham Open

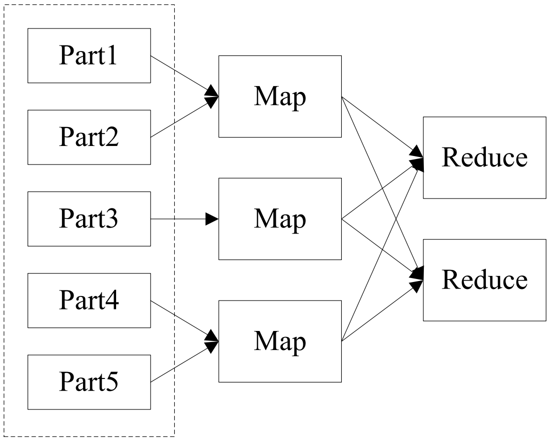


Fig. (2). Massive data cloud processing Hadoop platform.

3. DESIGN OF DATABASE MASSIVE DATA PROCESSING AND DATA MINING MATHEMATICAL MODEL

Massive image processing belongs to a large image data processing, and it not only has rather high processing requirements, but also is very harsh on the storage space requirements, if we use the general storage space, it is in difficult to meet the image processing capacity requirements [5, 6]. This paper adopts Hadoop distributed processing algorithm, the data space is partitioned to realize high-speed mass data processing. The database the data is divided into $[w_i, w_{i+1}]$, the establishment of two-dimensional database Hadoop data processing function is shown in formula (1).

$$\begin{cases} \mathrm{M}(m,n) = x_1 + x_2 m + x_3 n + x_4 mn \\ N(m,n) = y_1 + y_2 m + y_3 n + y_4 mn \end{cases} \tag{1}$$

Assumed that memory allocation value is

$$E^T = \left[E_1, E_2, E_3, E_4\right] \tag{2}$$

Then formula (1) and formula (2) can be written in matrix form:

$$e(m,n) = \alpha x \, . \tag{3}$$

Among them,

$$\alpha = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}. \tag{4}$$

For each storage node, it is

$$\overline{Z} = ZR \, . \tag{5}$$

Assuming the generalized coordinate R is

$$F = H^{-1}\overline{x} \, . \tag{6}$$

We can get the distributed node calculation formula that is

$$X(x,y) = a(x,y)\overline{X} \, . \tag{7}$$

In order to realize the cloud computing and distributed computing, the master-slave nodes' distributed computing program is set by using MATLAB software, the main procedures are as follows [7]:

```
function dy=vdp100000(t,y)
 dy=zeros(2,1);
 dy(1)=y(2);
 dy(2)=1000*(1-y(1)^2)*y(2)-y(1);
t0=0， tf=8000
t0=0， tf=8000
function dy=rigid(t,y)
 dy=zeros(3,1);
 dy(1)=y(2)*y(3);
 dy(2)=-y(1)*y(3);
 dy(3)=-0． 51*y(1)*y(2);
t0=0， tf=12
 [T,Y]=ode45('rigid',[0 12],[0 1 1]);
 plot(T,Y(:,1),'-',T,Y(:,2),'*',T,Y(:,3),'+')
……
```

4. DESIGN OF HADOOP CLOUD PLATFORM DATABASE MASSIVE DATA PROCESSING SYSTEM

In order to verify the validity and reliability of massive data processing mathematical model and the algorithm in second section, this paper uses MATLAB software to carry on simulation for the massive data processing, and designs the Hadoop cloud data processing platform, in which the main platform frame is shown in Fig. (**2**).

As shown in Fig. (**2**), the Hadoop cloud computing platform is mainly composed of three technologies, including distributed parallel computing framework MapReduce,

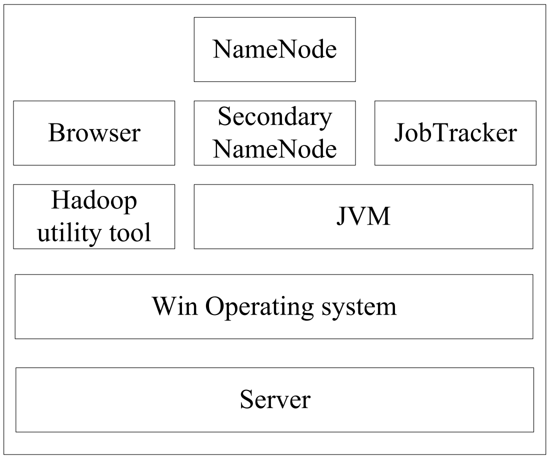


Fig. (3). The schematic diagram of Hadoop master node design.

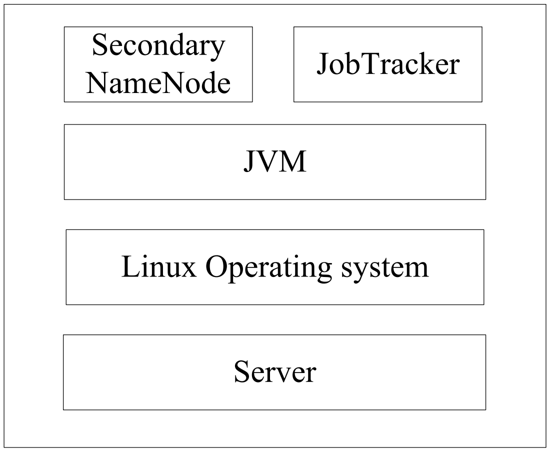
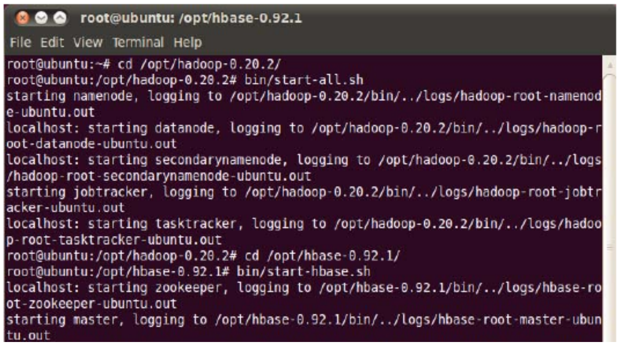


Fig. (4). The schematic diagram of Hadoop slave node design.



```
root@ubuntu:~# cd /opt/hadoop-0.20.2/
root@ubuntu:/opt/hadoop-0.20.2# bin/start-all.sh
starting namenode, logging to /opt/hadoop-0.20.2/bin/../logs/hadoop-root-namenod
e-ubuntu.out
localhost: starting datanode, logging to /opt/hadoop-0.20.2/bin/../logs/hadoop-r
oot-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /opt/hadoop-0.20.2/bin/../logs
/hadoop-root-secondarynamenode-ubuntu.out
starting jobtracker, logging to /opt/hadoop-0.20.2/bin/../logs/hadoop-root-jobtr
acker-ubuntu.out
localhost: starting tasktracker, logging to /opt/hadoop-0.20.2/bin/../logs/hadoo
p-root-tasktracker-ubuntu.out
root@ubuntu:/opt/hadoop-0.20.2# cd /opt/hbase-0.92.1/
root@ubuntu:/opt/hbase-0.92.1# bin/start-hbase.sh
localhost: starting zookeeper, logging to /opt/hbase-0.92.1/bin/../logs/hbase-ro
ot-zookeeper-ubuntu.out
starting master, logging to /opt/hbase-0.92.1/bin/../logs/hbase-root-master-ubun
tu.out
```

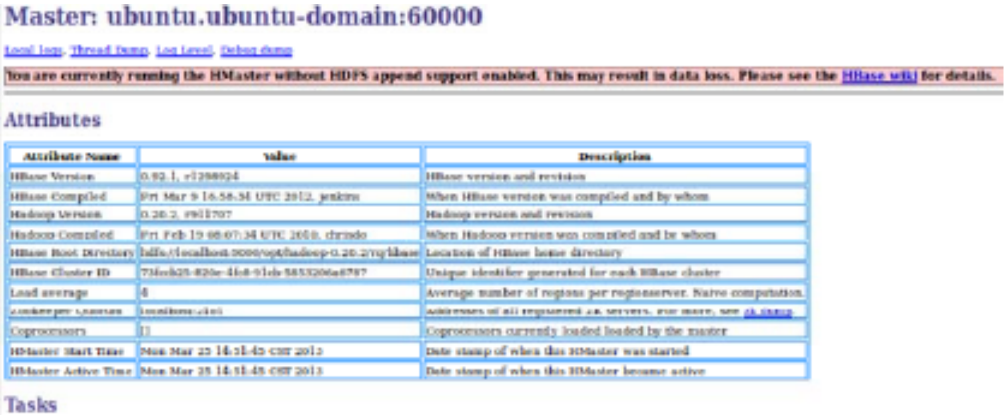
Fig. (5). The schematic diagram of Hadoop and Hbase running.

distributed database system HDFS and distributed file system Hbase.

As shown in Fig. (**3**), the master-slave node design is the main calculation mode of Hadoop implementation distributed computing and data storage, which has a master node and multiple slave nodes in each database cluster, and the master node running daemon includes NameNode, Secondary NameNode and JobTracker.

As shown in Fig. (**4**), there are many salvee nodes in the Hadoop massive data cloud processing system, which can realize the data distributed processing function [8]. In the slave node running, the daemon is DataNode and TaskTracker.

Fig. (**5**) shows the Hadoop and HBase run simultaneously, it can layout Hbase information and data partition, this



Master: ubuntu.ubuntu-domain:60000

Local logs, Thread Dump, Log Level, Debug dump

You are currently running the HMaster without HDFS append support enabled. This may result in data loss. Please see the HBase wiki for details.

Attributes

Attribute Name	Value	Description
HBase Version	0.92.1, r1298924	HBase version and revision
HBase Compiled	Fri Mar 9 16:58:34 UTC 2012, jenkins	When HBase version was compiled and by whom
Hadoop Version	0.20.2, r911707	Hadoop version and revision
Hadoop Compiled	Fri Feb 19 08:07:34 UTC 2010, chrisdo	When Hadoop version was compiled and by whom
HBase Root Directory	hdfs://localhost:9000/opt/hadoop-0.20.2/tmp/hbase	Location of HBase home directory
HBase Cluster ID	73fecb25-820e-4fc8-91eb-5853206a6787	Unique identifier generated for each HBase cluster
Load average	4	Average number of regions per regionserver. Naive computation.
Zookeeper Quorum	localhost:2181	Addresses of all registered ZK servers. For more, see zk dump.
Coprocessors	[]	Coprocessors currently loaded loaded by the master
HMaster Start Time	Mon Mar 25 14:31:45 CST 2013	Date stamp of when this HMaster was started
HMaster Active Time	Mon Mar 25 14:31:45 CST 2013	Date stamp of when this HMaster became active

Tasks

Fig. (6). Master node opening.

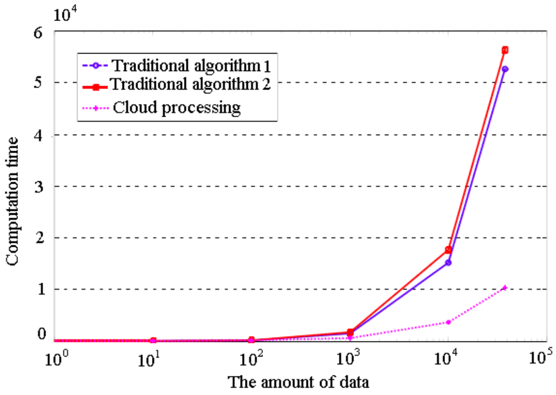


Fig. (7). The massive data processing computing time simulation curve.

paper uses the browser Web log to deal with massive data, the master node opening is shown in Fig. (**6**).

Fig. (**6**) shows the schematic diagram Master node opening, it is the key of implementing Hadoop database massive data processing [9,10]. In order to verify the validity and reliability of the system, this paper goes through simulation that can obtain the calculation performance table as shown in Table **1**.

Table **1** shows the simulation parameters of the cloud massive data processing Hadoop platform [11]. It can be seen from the table that Hadoop data processing calculation can achieve a higher level, in which the minimum value of data calculation is 0.001Mb/s, the information capacity of the minimum value is 100Gb and the speed is 1800 frames /s, they all meet the needs of design.

In order to compare different algorithms on the processing effect of the massive data, this paper selects two kinds of traditional algorithm and Hadoop cloud processing algorithm to carry on comparison, the amount of data is from the initial dozens of million to several hundreds of megabytes [12, 13]. As shown in Fig. (**7**), the MATLAB numerical simulation calculation can be found that the computing speed of Hadoop cloud processing computing platform is significantly higher than two traditional algorithms, especially when the data is reached 105 level, data processing speed will be faster, it can save several times data processing time, which is a kind of high efficient data processing algorithms.

Table 1. The massive data processing cloud platform simulation parameters.

Main Simulation Parameters	Numerical
Calculated data Min（Mb/s）	0.001
Calculated data Max（Mb/s）	0.01
Information capacity Min（Gb）	100
Information capacity Max（Gb）	100000
Speed (frame /s)	1800

5. SUMMARY

Massive image processing technology requires very high for the processor and memory, which need to adopt high performance processor and large capacity memory, because the traditional single or single core processing and ordinary memory already can not satisfy the need of image processing. In this paper, cloud computing function is introduced to mass data processing system, and then through the Hadoop distributed computing function, combined with the automatic distribution method of master slave node, the high speed data processing can be realized. Finally, compared with the Hadoop distributed processing algorithms and two kinds of traditional data processing algorithms, the data processing speed of Hadoop distributed algorithms is faster and its information capacity is also higher, it is a massive data processing algorithm, which provides a new method for the study of large data, high capacity and high transmission rate data treatment scheme.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Wang Wei, Jiang Lina. Research on cloud computing security requirements analysis [J]. Information network security, 2012(1): 35-37.

[2] Sun Fuquan, Zhang Dawei, Cheng Xu, Liu Chao. Construction of enterprise private cloud storage platform based on Hadoop [J]. Journal of Liaoning Technical University, 2011(3): 112-113.

[3] Zhou Zichen, Shen Zhenning. The power supply integrity design method in high speed embedded system [J]. Microcontroller and embedded application, 2010(3): 35-37.

[4] Kuang Shenghui, Li Bo. Analysis of cloud computing architecture and its application case [J]. Computer and digital engineering, 2011, 3(6):61-63.

[5] Xin Jun, Chen Kang, Zheng Weimin. Research on virtual cluster management technology [J]. Computer science and exploration, 2011, 4(23):325-327.

[6] Yu Fei. Research on data preprocessing technology in Web log mining [J]. Computer technology and development, 2011, 20(5): 47-50.

[7] Jinhai. On cloud [J]. China Institute of computer communication, 2012, 5(6): 22-25.

[8] Wu Jiyi, Ping Lingdi, Pan Xuezeng, Li Zhuo. Cloud Computing: from concept to platform [J]. Telecom science, 2011,12 (4): 25-27.

[9] Xue Zhiqiang, Liu Peng, Wen AI. Research on distributed file system management strategy [J]. Computer knowledge Technique, 2011(1): 112-113.

[10] Li Cheng, Cheng Xiaoyu. Analysis of high-speed DSP signal integrity simulation based on Hyperlynx [J]. Electronic devices, 2011, 33(2): 45-47.

[11] Zhu Zhijun, Le Zichun, Zhu ran. Research and implementation of OBS network simulation platform based on NS2 [J]. Journal of communication, 2012, 30(9): 128-134.

[12] Zhang Jian. The concept of cloud computing and its influence analysis [J]. Telecommunication technology, 2012, 1(12):15-16.

[13] Chen Quan, Deng Qianni. Cloud computing and its key technology [J]. Computer application, 2011, 29 (9):263-264.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014