MATEC Web of Conferences **61**, 07015 (2016) DOI: 10.1051/matecconf/20166107015
APOP2016

Data Mining Based on Cloud-Computing Technology

Ren Ying[1], Lv Hong[1], Li Hua-wei[2], Zhou Li-jun[1] , Wang Li-na[1]

[1]Naval Aeronautical and Astronautical University,Yantai 264000,China;

[2]Shan dong Business Institute,Shandong Yantai 264001,China)

Abstract. There are performance bottlenecks and scalability problems when traditional data-mining system is used in cloud computing. In this paper, we present a data-mining platform based on cloud computing. Compared with a traditional data mining system, this platform is highly scalable, has massive data processing capacities， is service-oriented, and has low hardware cost. This platform can support the design and applications of a wide range of distributed data-mining systems.

1 Submitting the manuscript

With the rapid development of mobile Internet and the Internet of things, huge amounts of data are produced in every minute. Data has penetrated into each field of industry and business functions. In the age of big data,if we want to excavate implicit useful information from nonholonomic, massive,noise and random data,wo must improve the efficiency data mining algorithm. Cloud computing is to provide dynamic resource,virtualization and high available computing platform.Cloud computing into data mining can solve the efficiency problem of massive data mining.

2 Data mining and cloud computing technology

Data mining[1-3] is the technology finding the valuable information from large amounts of data by the analysis of the data.The data mining process base on cloud computing technology is basic consistent with the traditional data mining ,which are made up of data preparation,data mining,evaluation results of three stage and explain the composition.With the information age development resulting from the "big" data,data mining tasks will bring forth the new through the old,to emphasis on a large database of efficient and scalable data mining technology.

Cloud computing[4-5]distributed the tasks to a large number of computers resources pool,so that all applications can access computing power, storage space and information service according to the needs of. At the same time, cloud computing is the development of distributed computing and grid computing, parallel computing[6].Cloud computing usually consists of the following 3 levels of service: Saas,Iaas,Paas. Service model as shown in Figure 1.

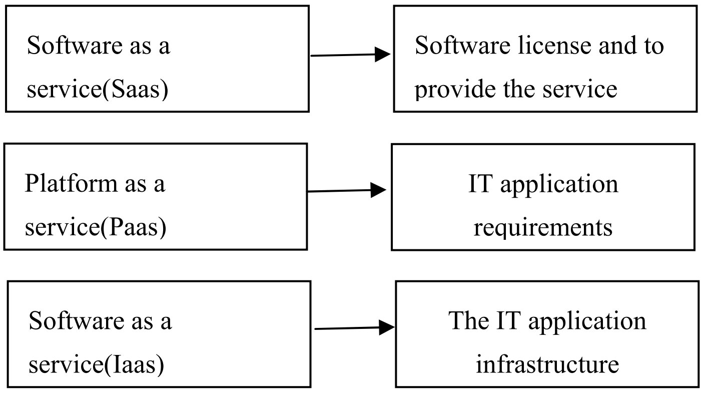
Software as a service(Saas)
Software license and to provide the service
Platform as a service(Paas)
IT application requirements
Software as a service(Iaas)
The IT application infrastructure

Figure 1. The service model of cloud computing

Cloud computing technology requirements to put compute nodes and storage nodes together.Task scheduling assigns and executes tasks on the preservation of equipment corresponding input file blocks as far as possible. This method makes the most of the parallel tasks to read the input data on the local machine, reducing the network data flow effectively[7].

Distributed computing is one of the effective means to solve the mass data mining tasks and improve the mass data mining[8]. Cloud computing platform provides a distributed file storage and parallel computing ability,which is a good solution to the distributed memory contained in distributed computing and parallel computing in two levels of content[9]. A good framework for the construction of cloud computing data mining platform core support ability of distributed file system and distributed parallel computing[10].

The popular distributed file system has Google file system (GFS), distributed file system(HDFS),the file system(KFS),which can effectively solve the problem of massive data storage.

3 Cloud based data mining algorithm

The algorithm of data mining is the soul,only the most efficient data mining algorithm in order to better accomplish the task of data mining. But because of a variety of data mining algorithms, there are also many data types, the requirements of different types of data mining algorithm is not the same.The most commonly used data mining algorithm has the following categories.

3.1. Classification algorithm

The main purpose of the classification algorithm is based on the existing data sets for mining to find the other data, and analysis of existing data sets and the discovery of new data, and then find the principle of data classification. This principle can be used to classify the data after adding. Classification algorithms are suitable for relational data consisting of tuple.

3.2 Cluster analysis

The main purpose of cluster analysis is find meaningful data distribution pattern of new from the potential data.The process is that the existing data is not specified grouping rules in advance,while is divided into different groups to miningn accordance with the data itself characteristics. Cluster analysis is also used for relational data consisting of tuple.

3.3 Association rules

The main purpose of the association rules is to find the interesting association or correlation between sets of items in large amounts of data. Association rules for data type is relatively more,mainly is suitable for the transaction type, transaction type and relation type data.Association rule best suited for handling the variable type is Boolean and numerical type.

Parallel mining algorithm is one of the key technologies which can effectively use the basic ability provided by loud computing platform.The general process of parallel data mining algorithms such as shown in figure 2.

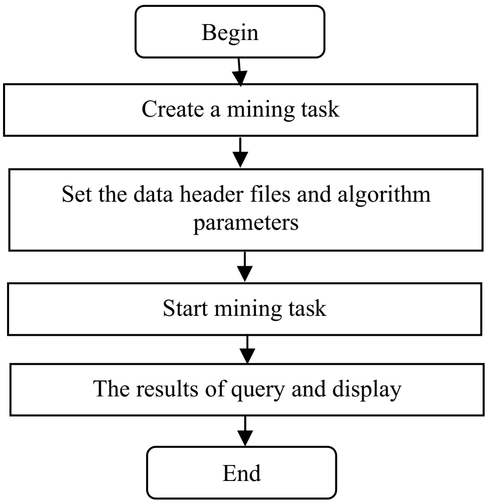
Begin
Create a mining task
Set the data header files and algorithm parameters
Start mining task
The results of query and display
End

Fig 2 Parallel data mining algorithms perform general process

4 Based on the data mining system based on Cloud Computing

The data mining system based on cloud computing is built on the "cloud", transparent providing interface services for a variety of terminal users,providing an open interface for program based on for the development of the system.The user can use indirect the various service by calling the open interface provided by system.The user does not need to know how the system is to achieve, no need to worry about the computing and storage capacity of the system,only need to select the appropriate algorithm to process data,and ultimately to the way the task for system deployment area executed to get the data mining results[11].

The data mining system based on cloud computing can adopts on-demand pay way.Enterprises or individuals for a service can be directly through this platform to obtain,which do not have to buy expensive software.The data are mostly stored in storage cloud after the arrival of the cloud Era so that the mining tool based cloud computing platform based data become possible.

5 Cloud computing system architecture based on Data Mining

Cloud is a computational model based on the Internet, public participation, whose computational resources include computing power, storage capacity expansion and is virtualization, and is to provide services to the user. With the massive data increasing, diversification, and personalized data mining to strong demand,the traditional centralized data mining methods cannot adapt. Cloud computing become efficient way to solve the problem of massive data mining because of its huge storage capacity and computing ability of elastic changes.

Hadoop cloud computing framework is an open source distributed system architecture widely used[12]. Users can easily build private cloud platform. Because not needing to understand the development of distributed applications distributed the underlying details of the case, the user can make full use of the ability of cluster computing and high-speed storage.

The current cloud computing data analysis and processing widely use distributed development framework for dealing with similar MapReduce. It can execute in parallel massive data collection and analysis tasks in a large number of PC machine.This model can highly abstraction the complex operation in large -scale cluster parallel computing on process of to two functions: Map and Reduce[12].

In the stage of Map,the Map/Reduce framework will split the input data into a large number of data segments,and each data fragment is assigned to a Map task. Each Map task will be to its assigned to Key-Value to calculate, to generate an intermediate result,then all intermediate results with the same Key value of the Value pass to the Reduce function after the calculation.

In the stage of Reduce,each Reduce task take the two tuples Key-Value as input.Two tuple will call the Reduce function to merge with Value, and set the formation of a smaller Value,and each Reduce function call has only 0 or 1 value output. Each stage of the task execution are supporting fault tolerance.If one or more nodes appear error in the calculation of the process will be automatically re allocation of tasks to other nodes.

This paper designs the data mining system based on cloud computing technology,the overall structure as shown in Figure 3.The nodes in the system are divided into two categories:MainCtrlNode and WorkNode. MainCtrlNode in the system consists of NameNode, data warehouse,JobTracker,SecondaryNameNode,data mining algorithms library.WorkerNode consists of Task-Tracker,DataNode,which is responsible for actual storage and computational work.NameNode manages file system metadata,which is the main server of distributed file system and implement open,closed,operation,rename of the file system.DataNode is responsible for handling customer read and write requests,to store the actual data,in accordance with the NameNode command, performs the data block copy,delete,create work.We apply data mining to be used in the data set to uploaded to the data warehouse,NameNode will automatically block files and data redundancy storage to each DataNode. SecondaryNameNode assisted NameNode processing image files and transaction log.

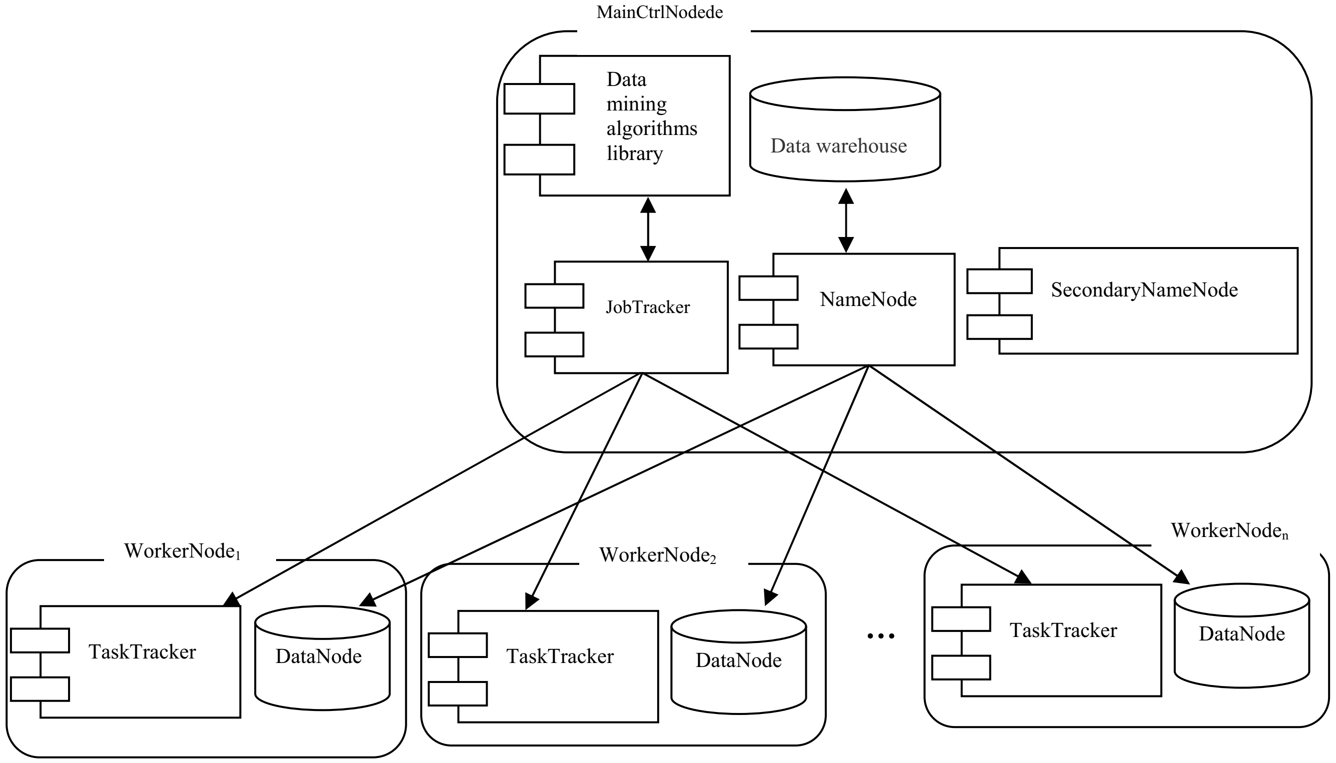

Fig 3 The overall architecture of data mining system diagram based on cloud computing technology

This paper designs the data mining system based on cloud computing technology,the overall structure as shown in Figure 3.The nodes in the system are divided Into two categories: MainCtrlNode and WorkNode. MainCtrlNode in the system consists of NameNode,data warehouse,JobTracker,SecondaryNameNode,data mining algorithms library.WorkerNode consists of Task-Tracker,DataNode,which is responsible for actual storage and computational work.NameNode manages file system metadata,which is the main server of distributed file system and implement open,closed,operation,rename of the file system. DataNode is responsible for handling customer read and write requests,to store the actual data,in accordance with the NameNode command, performs the data block copy,delete,create work.We apply data mining to be used in the data set to uploaded to the data warehouse,NameNode will automatically block files and data redundancy storage to each DataNode. SecondaryNameNode assisted NameNode processing image files and transaction log.

6 Conclusion

Massive data information and powerful computing and data processing capabilities of cloud computing provide powerful support for data mining. Through the analysis of the data mining and the cloud computing technology, this paper proposes the architecture of data mining platform based on cloud computing,for enterprise and individual user data mining task provides a good solution.

References

1. J Han,M Kamber. Data mining concepts andtechniques[M].Third Edition.San Francisco, CA,USA:Morgan Kaufmann Publishers,2012.
2. Shao feng-jing,Yu zhong-qing. Principle and algorithm of data mining[M].Beijing: Science Press,2009.
3. Shang Lin,Luo Bin. A data mining system based on Data Warehouse Framework[J]. Application Research of computers,2000,17(9):63-65.
4. Yang Yong,Dong zhen-jiang,Lu Ping. With the characteristics of cloud computing service delivery platform and its key technology research[J]. ZTE Communications,2011,17(5):55-57.
5. Wu zhu-hua. The analysis of the core technology of cloud computing[M].Bei Jing: People's Posts and Telecommunications Press,2011.
6. Mell P,Grance T． The NIST Definition of Cloud Computing ［R］.Gaithersburg,MD: National Institute of Standards andTechnology,2011．
7. Zhang jian-xun,Gu zhi-ming,Zheng chao. Review on research progress of cloud computing. 2010,27(2)：429-433.
8. Chen Quan,Deng qian-ni. Cloud computing and its key technology[J]. The computer applications,2009,29(9):2562-2567.
9. Li jian-jiang,Cui jian,Wang pin. MapReduce parallel programming model of review[J]. Chinese Journal of Electronics,2011(11):2635-2642.
10. Wang yi-jie,Sun wei-dong,Zhou Song. The key technology of distributed storage in cloud computing environment[J]. Journal of software,2012,23(4):962.
11. Wang Cong,Wang cui-rong,Wang xing-wei. The design of data center network architecture for Cloud Computing[J]. Research and development of computer,2012,49(2):286-293.
12. Hang He,Yi xiao-dong,Li shan-shan. Realization and evaluation of massive data processing platform for high performance computer[J]. Research and development of computer,2012,49:357-361.