# Analysis and Research of Data Mining Technology in the Horizon of Cloud Computing

**Juan Peng**

Xi'an Eurasia University, Xi'an, Shaanxi 710065, China

**Keywords:** Cloud Computing technology; Big data visualization; Data Mining; Analytical research

**Abstract:** With the continuous development and progress of the economy and society, science and technology information technology is constantly improving in order to meet the needs of social development. As an emerging technology in the development of the Internet, Cloud Computing has gradually become an indispensable part of people's lives, and is widely used in the military, medical and financial fields. Based on the rapid development of Internet information technology, the number of users using the Internet is increasing day by day, generating a large amount of information data. The proliferation of information data has pushed society from the age of information scarcity to the era of information overload, so people began to explore a new technology to save and analyze the data, and then extract the hidden value behind the data. In order to achieve this, it is necessary to solve the storage problem, processing problem and mining problem of massive data information, and the birth of Cloud Computing provides a direction and way for solving this problem. This paper analyzes and explores the Data Mining technology in the Cloud Computing environment, and analyzes its scalability and versatility from the architecture of the application platform, which provides a reference for researchers engaged in Data Mining applications.

## 1. Introduction

The popularity of computer technology has expanded the scope of Internet information provision services, making people's production and life more convenient [1]. The development of modern information technology, both enterprises and individuals need to analyze and summarize from the massive data, in order to obtain useful information resources. The massive increase in data poses a great challenge to Data Mining systems. The emergence of Cloud Computing can effectively solve this problem. It can concentrate the data distributed on different computers in a unified cloud, and provide users with corresponding services from the scalability of the network, especially in the Cloud Computing. The sharing of resources makes it have the characteristics of unlimited resource services, which is more conducive to our acquisition and mining of data. As an effective technology for statistical analysis of various types of information, Data Mining technology is naturally favored by data managers [2]. Based on the characteristics of Data Mining technology, this paper uses Cloud Computing to build a Data Mining application platform, which provides an effective application reference for data managers.

## 2. Overview of Data Mining based on Cloud Computing

### 2.1 Related concepts of Cloud Computing.

Cloud Computing is a product of the integration of parallel computing, distributed computing, virtualization, and grid computing and utility computing, network storage and load balancing, and many more traditional computer technologies and network functions. It emerged as a dynamic fusion and reorganization of low-cost, low-utilization, and very large amounts of useful resources on the Internet, making it a highly virtualized, very powerful storage and computing power. A treasure trove of information technology resources, with the help of advanced business models, namely, Infrastructure-as-a-Service IaaS,Platform-as-a-Service PaaS,Software-as-a-Service SaaS, so that end users who have no knowledge of the internal expertise or details of the cloud are need to get the above

various services, including very powerful storage space, computing power and various software services [3]. Cloud Computing can make system virtualization reach the highest level. As far as the current development trend is concerned, the cloud technology industry still does not have a unified and unrecognized industry standard. However, there is a consensus in all aspects of the customer terminal and the industrial chain, that is: Cloud Computing is a service-oriented trend.

Cloud Computing has the following characteristics:

1) The server is huge. Cloud has a considerable scale. Google Cloud Computing already has millions of servers. Amazon, IBM, Microsoft, Yahoo and other "clouds" all have more than 500,000 servers. " cloud " can give users superior computing power.

2) Resource virtualization. Cloud Computing allows users to access services in different geographic locations and using various terminals. The requested resource is dynamic and intangible. The app runs somewhere in the "cloud", but in reality the user doesn't have to know its specific location.

3) High reliability. In order to use Cloud Computing more reliable, "cloud" uses data multi-copy fault tolerance and other measures to ensure its high reliability services.

4) Strong versatility. Cloud computing is not to specific applications, but can be constructed the ever-changing applications under the support of "cloud". Cloud Computing is aimed at changing applications. For different applications, the operation can be supported by the same "cloud."

According to the types and functions of services provided by Cloud Computing, they can be divided into three types: Infrastructure-as-a-Service,Platform-as-a-Service,Software-as-a-Service.

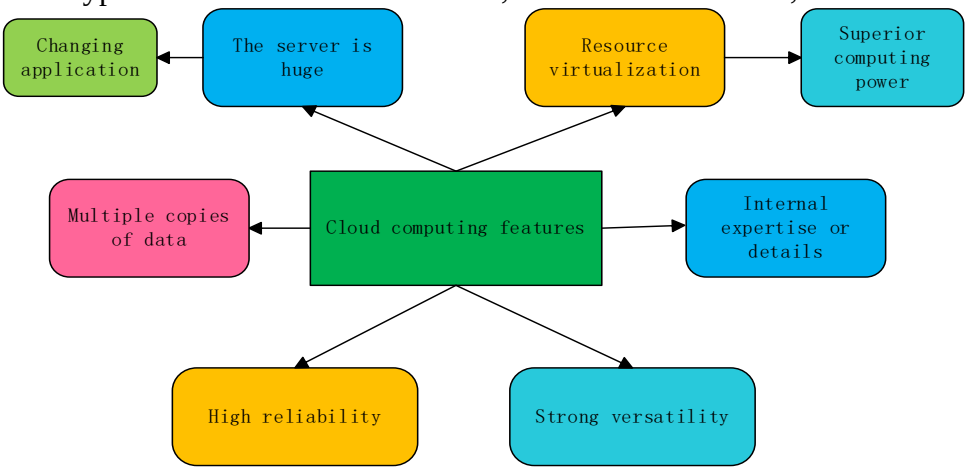


Figure 1. Characteristics of Cloud Computing

## 2.2 The development status and core technology of Cloud Computing.

Cloud Computing is a data-intensive, data-centric supercomputing method. It uses many technologies, including virtualization technology, data storage and management technology, and parallel programming mode [4].

The purpose of Cloud Computing is to enable us to make use of computing, services, applications, and other computer resources as easy and fast as using a public facility such as water and electricity resources or gas resources and telephones. Whenever we need resources or services, we Can be at your fingertips. An important part of Cloud Computing is computing centers and data centers. The business model of Cloud Computing is a new form for us. By implementing a range of technologies including WEB2.0, SOA, and virtualization, etc., it forms a new computing platform that is distributed. Cloud Computing applications, including running web applications and web services, rely on the use of powerful servers and large-scale data centers. In recent years, the world's growing number of information technology industry customers have successively entered the Cloud Computing service. On May 10, 2008, IBM established the first Cloud Computing center in Wuxi's Taihu New City Science and Education Industrial Park. On November 25 of the same year, the Expert Committee on Cloud Computing was established by the Chinese Institute of Electronics. Alibaba established China's first e-commerce Cloud Computing center in Nanjing in early 2009. China

Mobile Research Institute's research on Cloud Computing has begun quite early [5]. Through the massive data storage and distribution calculation of Cloud Computing, it provides a new method and means for massive Data Mining in Cloud Computing environment, effectively solving the problem of distributed storage and efficient computing of massive Data Mining. The continuous development and improvement of Cloud Computing technology has brought tremendous opportunities and challenges to the development of the entire information technology industry in China.

### 2.3 Advantages of Cloud Computing-based Data Mining technology.

Data Mining (DM) refers to the process of extracting potentially valuable information and knowledge from a large number of random, noisy, fuzzy and incomplete data. With the rapid development of contemporary information technology, we have accumulated more and more data. How to find valuable knowledge from a large amount of data is the primary problem [6] we face. The goal of Data Mining is to transform massive data into valuable information and knowledge. . The most essential difference between Data Mining technology and data analysis in the usual sense is two points. First, it is carried out without any assumptions, focusing on potentially unknown information inside the data. In addition, when it is in statistics to propose a link between the corresponding decision and the data. There are seven main tasks of Data Mining: class/concept description, association analysis, classification, prediction, clustering, deviation, and evolution. Analysis). In general, the Data Mining process has the next five steps: the first step, data selection. Select data that is relevant and suiTable for mining for Data Mining. Processing data according to different Data Mining objectives can improve mining efficiency. The second step is data preprocessing. Data preprocessing includes data integration, reduction, cleanup, and transformation, as well as conceptual layering and discretization. The third step is pattern discovery. Model discovery refers to finding patterns that are popular with users from the data, and is also the main process of knowledge discovery. The fourth step is the mode assessment. That is, through a certain measure, the model that truly represents knowledge is obtained. The fifth step is the representation of knowledge. It is to use the technology such as visualization to give the discovered knowledge to the end user.

The advantages of Cloud Computing-based Data Mining technology are mainly reflected in the following aspects: First, Cloud Computing-based Data Mining can realize distributed mining of data information, and realize real-time and efficient data information mining. At the same time, it can adapt well to organizations of different sizes. For example, for large enterprises, Cloud Computing-based Data Mining for certain specific data information mining will greatly reduce the dependence on large high-performance machines, and for small and medium-sized enterprises, can greatly reduce the data of small and medium-sized enterprises. Mining costs. Secondly, the Data Mining based on Cloud Computing has the advantage of easy development, so users do not need to consider dividing data, allocating data, loading data and scheduling computing tasks. Thirdly, cloud-based Data Mining can realize the utilization of the original equipment, improve the processing capability of large-scale data information, and undoubtedly become more convenient and free in terms of adding nodes, and greatly improve its own fault tolerance. Sex. Finally, cloud-based Data Mining has greatly reduced the threshold of application Data Mining technology, and can fully meet people's demand for massive data information mining.

### 2.4 Cloud Computing Data Mining research methods

1) It is a data association mining method. In the detailed analysis and value extraction of massive data information, associative Data Mining can centralize divergent network data information. The associative Data Mining method is generally divided into three steps: First, determining the scope of the data to be mined, and collecting the data objects to be processed, so that the attributes of the association research are clarified. Second, preprocessing the massive data to ensure the authenticity and integrity of the data, and the pre-processing results will be stored in the mining database. Third, shape the Data Mining of training. Entity threshold analysis is performed by means of permutation combinations.

2) It is a data ambiguity learning method. The principle is to first assume that there are a certain number of information samples under the Cloud Computing platform, then perform index description

on any one of the information samples, perform standard deviation calculation on all information samples, and finally realize Data Mining value information operation and high compression. Faced with the mining of massive data, the key to applying data ambiguity learning method is to screen and determine the fuzzy membership function, and finally realize the fuzzy operation of the massive Data Mining value information based on Cloud Computing. However, it should be noted here that the collection of node information of network data can be realized under the condition of activation.

3) Apriori algorithm. The Apriori algorithm is an algorithm for mining association rules. It is a basic algorithm designed by Agrawal. It is a two-stage mining idea and is based on multiple scan transaction databases. Different from other algorithms, the Apriori algorithm can effectively avoid this problem in the face of the tedious and complex nature of massive data, which leads to the poor convergence of Data Mining algorithms. Under the premise of saving input costs as much as possible, the use of computer simulation will greatly improve the mining speed of massive data.

## 3. New Construction of Data Mining Platform in Cloud Computing Mode

### 3.1 Data Mining platform framework under cloud service mode

The realization of Cloud Computing technology is based on the network platform, and realizes the storage, calculation and application expansion of data through virtual technology [7]. The combination of Data Mining technology and Cloud Computing technology, its topology structure is shown in Figure 2:

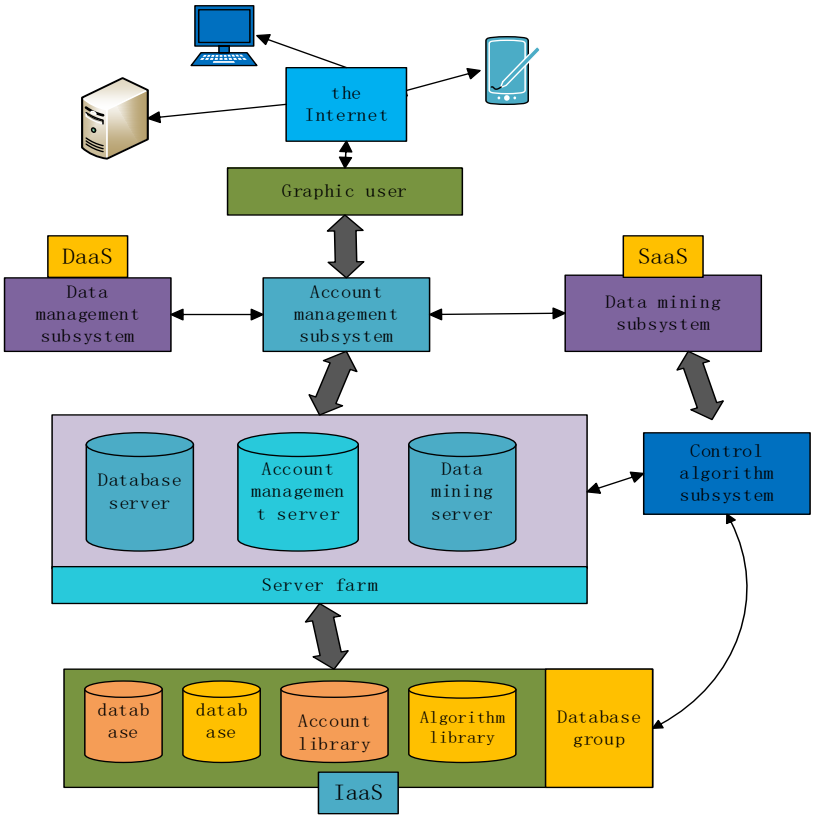


Figure 2. Schematic diagram of Data Mining framework based on cloud service model

Through the cloud platform, not only the corresponding Data Mining tasks can be completed as a whole, but also the individual user service requirements can be realized by some functions [8]. The user completes the login of the platform with its own account information by means of different terminal devices, and uses the platform resources under the supervision of the account management subsystem. For example, the user can access the database group and the server group through the IaaS service, or select the Data Mining algorithm through the PaaS service, and apply the standard algorithm to write the corresponding mining method, or call the Data Mining subsystem through the SaaS service, or Through the DaaS service, the corresponding resource exchange between the data

service and the platform is implemented to meet the completion requirements of the corresponding mining task.

### 3.2 Data Mining platform construction and results analysis

The Hadoop cluster supports three modes of operation: stand-alone mode, pseudo-distributed mode, and distributed mode. Among them, the stand-alone mode runs on a single machine, there is no distributed file system, Hadoop is configured as a stand-alone Java process running in non-distributed mode. This mode is mainly used to understand the Hadoop model principle and development test, only suiTable for The function of the program is tested and is the default mode [9]. From a distributed application perspective, the nodes in the cluster consist of a Jobtracker and a number of Tasktrackers, the Jobtracker is responsible for scheduling tasks, the Tasktracker is responsible for parallel execution, and the Tasktracker must be running on the Datanode for data reading and local computing, Jobtracker and The Namenode does not have to be on the same machine. This mode is suiTable for debugging and testing in the middle of program development.

This article builds a Hadoop multi-node distributed environment, the process is as follows:

(1) Install the JDK;

(2) Set the CLASSPATH and JAVA_HOME system environment variables, under ubuntu, add CLASSPATH=(the lib directory under JDK) via sudo/etc/environment, JAVA_HOME=(JDK directory);

(3) Each node must have the same username, such as test;

(4) The path of Hadoop should be the same, such as home/test/Hadoop;

(5) Install Hadoop on home/test/Hadoop and give test permission, update the Hadoop environment variable to the value of JAVA_HOME set in the second step

(6) Install SSH, and configure passwordless SSH, all processes on the DataNode (slave node) The NameNode (master node) is completed by SSH; in the state of no input password, the master node can freely access and control other slave nodes;

(7) Modify the configuration files conf/core-site.xml,conf/hdfs-site.xml,and conf/mapred-site.xml in the Hadoop file directory, specifically: conf/core-site.xml:

<configuration><property><name>fs.default.name</name><value>hdfs://NameNodeIP:9000</value></property></configuration>Conf/hdfs-site.xml:<configuration><property><name>dfs.data.dir</name><value>/home/test/hdfs/data</value></property></configuration>conf/mapred-site.xml:<configuration><property><name>mapred.job.tracker</name></configuration>

Start hadoop:<HADOOP_HOME>/bin/start-all.sh

Stop hadoop:<HADOOP_HOME>/bin/stop-all.sh

Description:

(1) After the Hadoop process is started, three java processes are started on the master server, namely NameNode, SecondNameNode, and JobTracker [10]. Two files are generated in the LOG directory, corresponding to the running log of the NameNode and the running log of the JobTracker. The slave server will start two Java processes, DataNode and TaskTracker. In the LOG directory, two files will be generated, which correspond to the running log of DataNode and the running log of TaskTracker. You can check whether the startup of Hadoop is correct by checking the log.

(2) Browse files in the distributed file system through IE: visit http://hdfs1:50030 to view the running status of the JobTracker; visit http://360quan-1:50060 to view the running status of the TaskTracker; visit http:/ /360quan-1:50070, you can view the status of the NameNode and the entire distributed file system. You can easily develop and debug Hadoop parallel programs in the Eclipse environment [11]. This article uses the MapReduce for Eclipse plug-in that comes with Hadoop. This Eclipse plugin simplifies the process of developing and deploying Hadoop parallel programs. Based on this plugin, you can create a Hadoop MapReduce application in Eclipse. The plugin also provides some wizards for class development based on the MapReduce framework. You can package it into a JAR file and deploy a Hadoop MapReduce application to a Hadoop server (local and remote). You can view the status of the Hadoop server, Hadoop Distributed File System (HDFS), and currently running tasks through a dedicated view (perspeetive).

In order to verify the computational performance and computational advantages of the MRD_Apriori algorithm, the paper tests the algorithm from the aspects of data volume, number of nodes, and different support degrees, and analyzes the results of Data Mining. In this paper, the algorithm is tested and compared in terms of changing the database size, increasing the number of computing nodes, and different support degrees, and analyzing the results effectively. In order to verify the computational efficiency and scalability of the improved Apriori algorithm, 256M, 512M, and 1G data were used in experiments, respectively, at 5, 10, 15, and 20 operating nodes with support degrees of 5%, 10%, and 15% run. The first set of experiments is the performance of 256M, 512M, and 1G data on 5, 10, 15, and 20 compute nodes.

The future will be a world where Data Mining and knowledge discovery are ubiquitous. There are also problems in Data Mining that are difficult to process. Experiments show that the improved algorithm is more efficient in processing massive data than traditional ones. Data Mining technology has a good prospect in the research and application of Cloud Computing environment. Data Mining algorithm based on Cloud Computing environment is a key step in dealing with massive Data Mining, and it is also the core problem of massive Data Mining.

## 4. Conclusion

With the continuous deepening of research in the field of science and technology in China, Cloud Computing technology has become more and more perfect, and its application scope and degree have also been developed to varying degrees. This paper first expounds the definition of Cloud Computing, Cloud Computing technology and Cloud Computing architecture, and introduces the related concepts and related algorithms of Data Mining technology. After understanding the Cloud Computing environment, and organically integrating it with Data Mining technology, it can achieve a series of operations such as rapid analysis, calculation and storage of massive data through the complementary advantages of the two, making the efficiency and quality of Data Mining work. Both can be greatly improved. Different types of data use different algorithms in Data Mining. These algorithms can also be combined and distributed to various nodes of the Cloud Computing framework, so that they can play their role in Cloud Computing technology. In short, I believe that with the development of the times and the passage of time, the future Cloud Computing technology and Data Mining technology will be further optimized and improved to better provide more convenience for Data Mining.

## References

[1] Zhang L, Wu C, Li Z, et al. Moving Big Data to The Cloud: An Online Cost-Minimizing Approach [J]. IEEE Journal on Selected Areas in Communications, 2013,31:2710-2721.

[2] Willcocks L, Venters W, Whitley E A. Cloud Computing as Innovation: Studying Diffusion[J]. Lecture Notes in Business Information Processing, 2013,163:117-131.

[3] Zhu X, Yang L T, Chen H, et al. Real-Time Tasks Oriented Energy-Aware Scheduling in Virtualized Clouds[J]. IEEE Transactions on Cloud Computing,2014,2:168-180.

[4] Zheng K, Meng H, Chatzimisios P, et al. An SMDP-Based Resource Allocation in Vehicular Cloud Computing Systems[J]. IEEE Transactions on Industrial Electronics, 2015,62:7920-7928.

[5] Huang H, Song G, Peng L, et al. Cost Minimization for Rule Caching in Software Defined Networking[J]. IEEE Transactions on Parallel & Distributed Systems,2016,27:1007-1016.

[6] Tran T, Yazdanparast A, Suess E A. Effect of Oil Spill on Birds: A Graphical Assay of the Deepwater Horizon Oil Spill's Impact on Birds[J]. Computational Statistics, 2014,29:133-140.

[7] Shiraz M, Sookhak M, Gani A, et al. A Study on the Critical Analysis of Computational Offloading Frameworks for Mobile Cloud Computing[J]. Journal of Network & Computer Applications, 2015,47:47-60.

[8] Joshi N, Baumann M, Ehammer A, et al. A Review of the Application of Optical and Radar Remote Sensing Data Fusion to Land Use Mapping and Monitoring[J]. Remote Sensing, 2016,8:70.

[9] Gaggero M, Caviglione L. Predictive Control for Energy-Aware Consolidation in Cloud Datacenters[J]. IEEE Transactions on Control Systems Technology, 2016,24:461-474.

[10] Wang Z J, Mujib A B M M. The Weather Forecast Using Data Mining Research Based on Cloud Computing.[J]. Journal of Physics: Conference Series, 2017,90:012-020.

[11] Koubaa A, Qureshi B. DroneTrack: Cloud-Based Real-Time Object Tracking Using Unmanned Aerial Vehicles Over the Internet[J]. IEEE Access, 2018,6:13810-13824.