Construction of Data Mining Platform in Cloud Computing Environment

Yu Wenwu*
Network engineering teaching and Research Office
Dalian University of Science and Technology
Dalian, China
E-mail: 264186958@qq.com

Wei Dan
Network engineering teaching and Research Office
Dalian University of Science and Technology
Dalian, China

Li Nan
Network engineering teaching and Research Office
Dalian University of Science and Technology
Dalian, China

Xu Peng
Network engineering teaching and Research Office
Dalian University of Science and Technology
Dalian, China

***Abstract*—In order to solve the problem of low accuracy of data mining in traditional data mining platforms, the construction of data mining platform in the cloud computing environment is proposed. Based on the data mining platform architecture design in the cloud computing environment, and the data collection and scheduling center design, completed the hardware design of the data mining platform in the cloud computing environment; Relying on the cloud strategy design of data mining and the parallel design of mining algorithm, the software design of the data mining platform in the cloud computing environment is realized and the construction of the data mining platform in the cloud computing environment is completed. Experimental data show that the proposed data mining platform, compared with the traditional mining platform, improves the mining accuracy by 47.29%, which is suitable for data mining in the cloud computing environment.**

Keywords-Cloud Computing; Data Mining; Platform Construction; Mining Strategy

I. INTRODUCTION

With the rapid development of the Internet, especially the mobile Internet and the Internet of things, we are now in the era of massive data and information overload. According to a research report by data research company IDC, the total amount of data created and copied globally was 1.8ZB in 2011, which was up by 1.0ZB compared with the same period last year, and the global total amount of information will double every two years. Users are faced with a huge amount of information but it is difficult to find the content they are really interested in. At the same time, as operators gradually launch mobile Internet products, such as app stores, reading, games and community products, how to analyze and mine the massive data generated by these products will become an urgent problem for operators to solve. With the emergence of cloud computing, the data mining platform has a new development direction and makes the new generation of data mining platform possible. Cloud computing is a computing platform that provides dynamic resources, virtualization, and high availability. Cloud computing platforms can be used to develop high-performance applications. However, for data mining, massive data itself has problems such as noise, heterogeneity, complex algorithm and complex technology, while the current cloud computing development platform does not provide data specification and other functions. Therefore, through the detailed description and analysis of data mining and cloud computing, this paper proposes a data mining platform based on cloud computing. The platform architecture is based on the fundamental capabilities of cloud computing and conforms to the design philosophy of cloud computing software as a service (SaaS). The platform can also greatly reduce the investment of operators and enterprises in data mining technology, accelerate the launch of their mining business, shorten the research and development cycle, and further improve product revenue.

II. HARDWARE DESIGN OF DATA MINING PLATFORM IN CLOUD COMPUTING ENVIRONMENT

A. *Data mining platform architecture design under cloud computing environment*

The distributed storage and computing of cloud computing have promoted the revolution of new generation data mining platform. Figure 2 is a cloud-based data mining platform architecture.

Considering the parallelization and distribution of mining algorithm and recommendation algorithm is a large professional topic, this paper does not include the parallelization and cloud-based content of specific algorithm for the time being[1].

As shown in figure 1, this platform is a data mining cloud service platform based on cloud computing platform. It adopts the idea of hierarchical design and component-oriented design, and is generally divided into three layers. The order from bottom to top is: cloud computing supporting platform layer, data mining capability layer, and data mining cloud service layer[2].

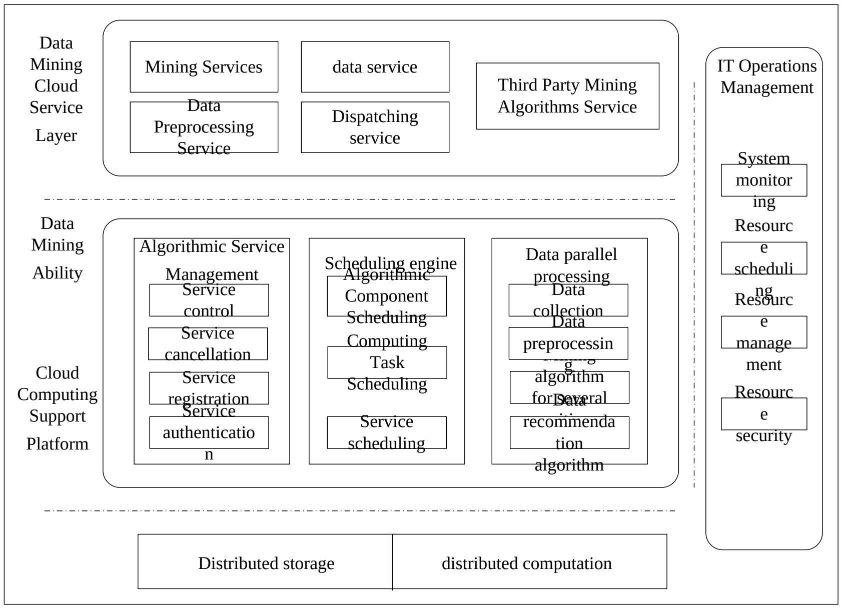

Figure 1. Data mining cloud service platform

The cloud computing supporting platform layer mainly provides distributed file storage, database storage and computing power. Zte has its own cloud computing platform, which can be based on the cloud computing platform independently developed by enterprises or the cloud computing platform provided by a third party.

The data mining capability layer is mainly to provide the basic ability of mining, including the framework of algorithm service management, scheduling and data parallel processing, and to provide the ability support for the cloud service layer of data mining. This layer can support the access of third-party mining algorithm tools, such as Weka, Mathout and other distributed algorithm libraries, as well as provide internal data mining algorithms and recommendation algorithm libraries. Data mining cloud service layer[3].

The cloud service layer mainly provides external data mining cloud services, and the interface forms encapsulated by the service capability can be diverse, including Webservice, HTTP, XML or local application programming interface (API) based on simple object access protocol (SOAP). The cloud services layer can also support access based on structured query language (SQL) statements and provide a parsing engine to automatically invoke cloud services. Each business system can invoke and assemble data mining cloud services according to data and business needs.

Compared with the traditional data mining system architecture, the data mining platform based on cloud computing proposed in this paper has the advantages of high scalability, mass data processing capacity, service orientation, low hardware cost, etc., which can support the design and application of large-scale distributed data mining.

B. Data collection and scheduling center design

The data collection and scheduling center realizes the collection of business data connected to the platform, which can solve the protocol problem of different data and support various source data formats. The source data format supports online transaction processing system (OLTP) data, online analytical processing system (OLAP) data, various log data, crawler data, etc., and provides multiple data synchronization modes.

Online transaction processing (OLTP) is an application for data mining, which is mainly used for data entry and transaction data retrieval of a certain number of industries. The most widely used OLTP product is probably IBM's user information control system (CISS). OLTP software USES client/server processing mechanisms and intermediary software that allows transactions to run on different computer platforms in the network. Reflect the current running state of data, complete the daily tasks of data management mining database application. In online transaction processing, transactions are executed immediately, as opposed to batch processing, where a batch of transactions is stored for a period of time and then executed. Most batch processing takes place at night. The results of OLTP are immediately available in this database, assuming these transactions can be completed. Online transaction processing occurs in real time.

Executing transactions in a single-user, single-database environment is simple because there is no conflict or need for synchronization between databases. Maintaining the integrity of multiple databases is another issue in a distributed environment. Traditionally, most online transaction processing systems have been implemented on large computer systems due to the complexity of their operations and the need for fast input/output, suppression, and management. If a transaction must be modified in more than one locale, then a management mechanism is needed to prevent overwriting the data and provide synchronization. Other requirements include the ability to rollback failed transactions, provide security features, and, if needed, provide data recovery. This is handled through a transaction handler. This monitor ensures that the transaction is fully completed or rolled back, thus guaranteeing the correctness of the database state.

In a distributed environment, writes often occur in parallel on multiple database servers. Such concurrent transaction processing requires a "rollback" mechanism to guarantee the integrity of the database even if the system fails in a single write operation. Transactions are either confirmed together or aborted. If one or more transaction-related system responses are inconsistent, this means that the system or communication may fail and a transaction may be abandoned. As you can see, conflicts occur when multiple users try to change the same piece of data at the same time. In addition, writes to multiple databases must be processed synchronously and must be guaranteed to be processed by all databases. A monitor program is required to ensure data integrity. There are four requirements for transaction processing in a distributed environment, collectively known as "ACID", in various ways such as database real-time synchronization, socket message synchronization, file transfer protocol synchronization, and so on, as shown in figure 2.

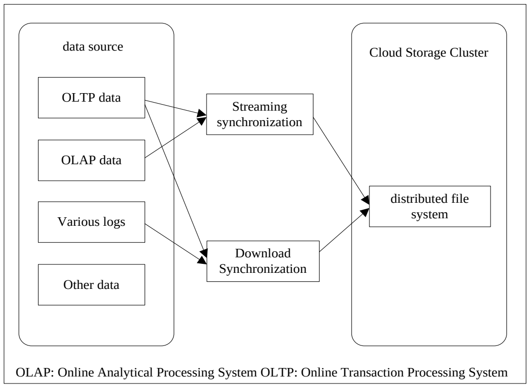

Figure 2. Data collection and scheduling center structure

Data collection and scheduling center is mainly to complete the collection of different types of data. The data collection and scheduling center adopts the template design technology to support the template and metadata configuration of new data to achieve the unified collection and specification of different business data.

III. SOFTWARE DESIGN OF DATA MINING PLATFORM UNDER CLOUD COMPUTING ENVIRONMENT

A. Data mining cloud strategy design

The cloud strategy of data mining firstly selects the data set, which is the basic program of data mining. All data mining is based on reasonable data set selection. In order to adapt to the collaborative mining of big data in the ring network under cloud computing, the data set is selected based on hardware support and software logic by connecting the microprocessor through the cloud computing interface. The basic block diagram of data set selection is shown in FIG. 3[4] :

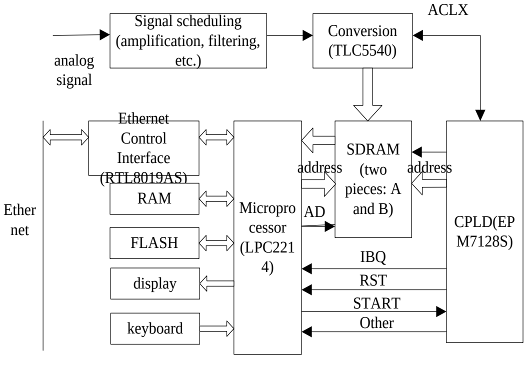
analog signal
Signal scheduling (amplification, filtering, etc.)
Conversion (TLC5540)
ACLX
Ethernet Control Interface (RTL8019AS)
RAM
FLASH
display
keyboard
Ether net
Micropro cessor (LPC221 4)
SDRAM (two pieces: A and B)
address
address
AD
CPLD(EP M7128S)
IBQ
RST
START
Other

Figure 3. A basic block diagram for data set selection

In the context of big data collaborative mining platform of ring network under cloud computing, data sets should be selected with a certain representativeness. Therefore, data separation technology and feature extraction technology are used to select data sets. n sets of data are set to satisfy the equation $y_i^* = V_{i0}^{-1/2} / (y_i - x_i\beta_0)(i = 1, 2, ..., n)$. Where i stands for data correlation coefficient, x and y for coordinate parameters of data representability, β for data fuzziness, and V for data identity.

Then, data with the same value are divided into the same class by data separation technology to obtain the same data set equations, as shown in formula (1) [5] :

$$x_n = x_i^{'} V_{i0} / y_i^* \beta_0 \tag{1}$$

That is, the data value of the same feature can be represented by x1, and the features of x1 can be extracted based on the feature extraction technology. Similarly, for x2，x3，...，xn is used to extract numerical features and establish a data set, namely X = { x1，x2，x3，...，xn } to realize the selection of data sets.

B. Parallel design of mining algorithm

The parallelization of mining algorithm is based on data preprocessing, and a more refined mathematical model will be obtained. Based on this mathematical model, data relations will be defined according to the mining wizard, data integration processing will be carried out, and the collaborative preparation mining of big data will be realized.

Data integration, data set selection and data preprocessing are all data filtering processes. However, different from data set selection and data set preprocessing, data integration is data filtering or fitting in the process of data mining. Data set filtering and data preprocessing are the data preparation stage in the data mining process, and the data integration process is shown in figure 4[6].

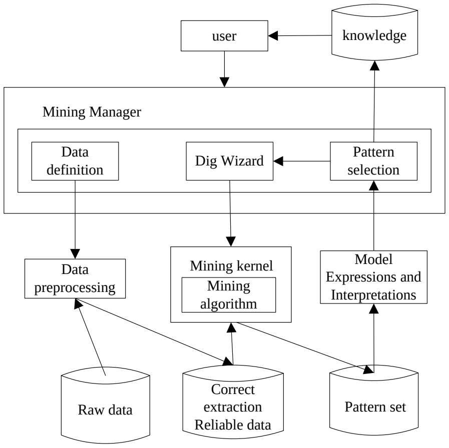

Figure 4. Schematic diagram of data integration process

Suppose that during the integration process, the data quantity obtained by data preprocessing is t, the data size is M, the data length is I, and the randomness of data generation is k, then the import equation of data integration can be expressed as follows:

$$f = \sum_{i=0}^{M-1} (t_i \times 2^{M-i-1} x_n) \bmod 2^M \ / \ k \qquad (2)$$

Through the integrated processing of different characteristic values, we can get different digger relations, and establish the matching index function, so that it can carry out data mining on digger points, so that it can cover the data surface related to digger points.

Data reduction and discretization is the core program of data mining. Based on data integration, data reduction and discretization are used to realize data mining[8].

Data reduction process before dealing with the collection of data are identified, prevent due to the large amount of data, the data reduction process there is a big error, therefore according to the variable state, the characteristic identification, the identification process according to the data of the variable degree is designed the Gn(t), in turn, the size of the sort, according to the variable degree is designed the Gn(t) the sort order, conditional reduction, the reduction treatment available formula 3[9]:

$$Q = \sigma(1 + a\frac{\partial}{\partial t})E\varepsilon fG_n\left(t\right) \qquad (3)$$

In the formula, σ represents the predetermined standard range. ε is a persistent variable for the data. Eε represents the data-driven inheritance state, i.e. the relationship between data, and a represents the adjustment coefficient.

According to the formula 3 will arrange good data reduction, to get a basic characteristic of value, which is a feature point in the process of mining, calculated according to repeat multiple feature points and discrete calculation of multiple feature points, realize the cloud ring network data mining together, the discrete calculation, formula 4[10] available:

$$F = (Q + aE\frac{\partial}{\partial t})\lim_{L \to 0}(2^{M-i-1}S) \qquad (4)$$

Where, E represents the discretization determined before mining. The larger the discretization, the larger the data mining scope and the larger the error. S stands for the characteristic characteristics of data. Different data characteristic characteristics will result in different discrete calculation results. Based on data set selection, data preprocessing, and data integration data reduction and discretization. Data mining is conducted by parallelization of mining algorithm. This paper explains the parallelization technique of mining algorithm through the parallel computing framework of k-means clustering algorithm.

The main idea of k-means algorithm is based on minimizing clustering performance index. The clustering criterion function used here is the sum of the square distance between each local point in the cluster and the center point of the cluster, and it is minimized. The processing flow of k-means algorithm is as follows:

First, k objects were randomly selected, and each object represented the initial mean and center of a cluster. For each remaining object, it is assigned to the most similar cluster according to its mean distance from each cluster. Then calculate the new mean of each cluster. This process is repeated until the criterion function converges. Generally, the square error criterion is adopted, and its definition is as follows:

$$E=\sum_{i=1}^{k}\sum_{p=C_n}\left|p-m_i\right|^2 \qquad (5)$$

Where, E is the sum of squared errors of all objects in the data set, p is the point in the space, represents the given object, and mi is the mean of cluster Ci. For each object in each cluster, we first need to square the mean of the objects to the center of the cluster, and then sum.

Clustering is divided into clustering centers. Once k clustering centers are determined, clustering can be completed immediately. Therefore, this paper focuses on how to update the clustering center in parallel. After randomly initializing k clustering centers, each task will update the value of the current clustering center. In the mapping stage, for each sample OS, it is necessary to calculate its nearest clustering center Oi, and then output the key value. In the Reducer stage, the framework will collect values belonging to the same key, which is equivalent to the clustering center Oi (0 I k-1), and the samples closest to it will be collected as values. In this way, k clustering centers can be reestimated by using these samples in the Reducer, as shown below:

$$O_i=\sum_{j=1}^{n}O_j\times\frac{1}{n} \qquad (6)$$

Thus, after the completion of a round of MapReduce, the new clustering center has also been calculated. By comparing the difference degree between the clustering center of this round and that of the previous round, the convergence of the algorithm can be determined and the construction of the mining platform can be realized.

IV. Instance analysis

In the process of the experiment, different data types are used to analyze the mining effectiveness according to the changes of data amount, and the proposed data mining platform research is verified. The traditional data mining platform is used as the experimental comparison object for the experimental analysis.

During the experiment, the experimental data in *.sdr, *.skz and *.asa formats are firstly prepared. The amount of experimental data to be prepared is 1GB to 17GB, and it can run normally in the cloud computing environment and execute the cloud computing program. As shown in table 1:

TABLE I. Experimental preparation data

Project	Parameter/scope of execution	Remarks
data type	*.sdd，*.skz，*.asa	Data volume from 1GB to 17GB
Calculation method	cloud computing	Data analysis
Network Topology	Ring Topology	Connecting equipment needs more than 50 sets
Data Communication Protocol	IP/ICP Communication Protocol	Data transmission between ring topologies

In order to ensure the objectivity of the experiment, an internal LAN is firstly established. The local area network is a ring topology structure and is effectively connected to the cloud computing program. It is only used for experimental test and analysis, and no other commercial activities are carried out. The simulated data volume is typical laboratory value, non-repeated value and random value, and cloud computing behavior is performed in the annular topology as required.

Then, access to the traditional data collaborative mining platform, and use the data provided for data mining in 1GB, 4GB, 7GB, 11GB, 14GB and 17GB. Compare the mining results with the real values to obtain the validity of data mining and record it in the experimental chart.

Second, perform the same local area network, use the same ring topology, in the data volume of 1GB, 4GB, 7GB, 11GB, 14GB, 17GB cloud computing, access to the proposed big data collaborative mining platform, data mining, data mining effectiveness, and record in the experimental chart.

Finally, the experimental results were sorted out, the experimental results were analyzed, and the experiment was completed. According to the validation results of data mining under different data volumes, the comparison curve of experimental results was drawn, as shown in figure 5:

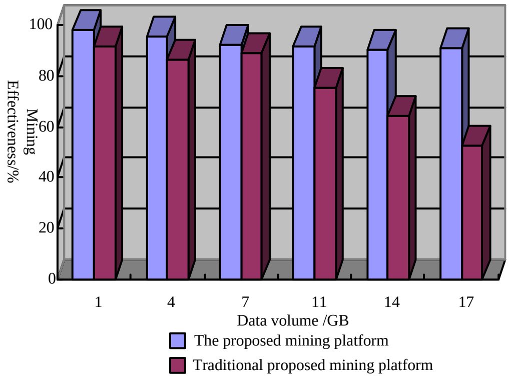

Figure 5. Simulation experiment results

According to the statistical calculation, the proposed average mining effectiveness is 94.51%, and the traditional data mining platform average mining effectiveness is 47.22%. The proposed big data collaborative mining platform is 47.29% more effective than the traditional mining platform, which is suitable for big data collaborative mining in the cloud computing environment.

V. CONCLUSIONS

This paper proposes the construction of the data mining platform in the cloud computing environment. Based on the design of the hardware and software of the data mining platform in the cloud computing environment, the construction of the data mining platform in the cloud computing environment is completed. Experimental data show that the proposed data mining platform is highly effective. It is hoped that the research in this paper can provide theoretical basis for the construction of data mining platform.

REFERENCES

[1] Chen J, Li K, Member S, et al. A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment[J]. IEEE Transactions on Parallel and Distributed Systems, 2017, 28(4):919-933.

[2] Noraziah A, Fakhreldin M A I, Khalid A , et al. Big Data Processing in Cloud Computing Environments[J]. Advanced Science Letters, 2017, 23(11):11092-11095.

[3] Wang Y, Zhao Y W. Transplantation of Data Mining Algorithms to Cloud Computing Platform when Dealing Big Data[J]. 2017.

[4] Karimi M B, Isazadeh A, Rahmani A M. QoS-aware service composition in cloud computing using data mining techniques and genetic algorithm[J]. The Journal of Supercomputing, 2017, 73(4):1387-1415.

[5] Zhao Y, Calheiros R, Gange G, et al. SLA-Based Profit Optimization Resource Scheduling for Big Data Analytics-as-a-Service Platforms in Cloud Computing Environments[J]. IEEE Transactions on Cloud Computing.

[6] Zhu J. Research on data mining of electric power system based on Hadoop cloud computing platform[J]. International Journal of Computers & Applications, 2017(1):1-7.

[7] Kune R, Konugurthi P K, Agarwal A , et al. XHAMI - extended HDFS and MapReduce interface for Big Data image processing applications in cloud computing environments: XHAMI - EXTENDED HDFS AND MAPREDUCE FOR BIG DATA PROCESSING[J]. Software—practice & Experience, 2017, 47:43-51.

[8] Parra-Royon M, Atemezing G, J. M. Benítez. Data Mining definition services in Cloud Computing with Linked Data[J]. 2018.

[9] Lee D, Park N, Kim G, et al. De-identification of metering data for smart grid personal security in intelligent CCTV-based P2P cloud computing environment[J]. Peer-to-Peer Networking and Applications, 2018.

[10] Basu S, Pattnaik P K. Maintaining Consistency in Data-Intensive Cloud Computing Environment[J]. 2018.