TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 3, March 2014, pp. 2021 ∼ 2027
DOI: http://dx.doi.org/10.11591/telkomnika.v12.i3.3864 ■ 2021

Layers Construct Design for Data Mining Platform Based on Cloud Computing

Yong-Zheng Lin
School of Information Science and Engineering, University of Jinan Jinan, 250022, China
e-mail: yzdynasty@163.com

Abstract

This paper studies the problem of design of layers construct for data mining platform based on cloud computing. First, the architecture of cloud computing is designed to deal with the data separately stored in network. Then, the layers construct for data mining platform based on cloud computing is designed, which includes data reduction tools, algorithm layer, application layer, and user layer. The Key-Points are introduced to design the plug-in unit system frame and open interface.

Keywords: *cloud computing, data mining, layers construct, data reduction*

1. Introduction

With the rapid development of modern society forces and compute science, large quantities of data will be produced and be saved in storage medias. It is worth noting that the technology of data mining supplies a useful tool for detecting the underlying knowledge and patterns from numerous, complicated, and heterogeneous data. However, with the application and development of the technology of computer network, especially the technology of cloud computing, more and more data is stored separately in different computers under network, the intense complexity to deal with such data has exceeded the limit of original data mining system and the scope of computing resources supplied by traditional single server. Thus, it's necessary to discuss the layers construct design for data mining platform based on the large scale parallel computing.

As upgrade for grid computing, cloud computing is viewed as a calculation mode, which can offer dynamic virtual resources to users via Internet[1]. What's more, during the design process for high performance program[2], cloud computing could be as a computing platform and dynamic resources pool with the characteristic of visualizations and high usability. Although, more and more heterogeneous and non-homogeneous data has been utilized in data mining based on the virtue of cloud platform. In order to adapt to the changes in the data and computing platform, it is necessary to implement data reduction during data mining process so as to visit heterogeneous and undefined data type, then to further design and achieve data mining system.

Since 2007, a considerable number of theoretical and experimental research effort has been paid more and more attention to the research, discussion and relevant products for cloud computing platform. At present, cloud computing platforms, such as Google App Engine(GAE)[3], Amazon Elastic Computer Cloud(EC2)[4], could offer preferable bottom architectures to support the realization of data mining system and make the computing resource on platform more convenient. As a kind of cloud infrastructure service based on virtual server technology, EC2 could provide large scale, reliable and elastic computing environment for users. By using the quality service afforded by EC2, users not only conveniently facilitate computing resource, but also customize special resource freely. As a new generation of network program development platform based on cloud computing, GAE allows users to develop and run network application program on Google foundation frame, which could be easily built and maintained. Developers could easily obtain corresponding service after unloading their application, so it is unnecessary to maintain servers for developers and users. For database, GAE provide powerful distributed data storage service – Big Table, which support structure query and update operation and provide transaction function to keep data consistency.

Received July 15, 2013; Revised September 23, 2013; Accepted October 12, 2013

The rest of paper is organized as follows. Section 2 presents the architecture of cloud computing. The design of layers construct for data mining platform based on cloud computing is given in Section 3, which includes data reduction tools, algorithm layer, application layer, and user layer. In Section 4, the Key-points are described to design the Plug-in unit system frame and open interface. Finally, we conclude our findings in Section 5.

2. Architecture of Cloud Computing

Cloud computing can offer a high-flexibility, high-reliability, transparent, and safe bottom structure with friendly monitoring and maintenance interface for users as shown in Figure 1, by which computing and memory resources could be distributed to users. Thus, based on the application of this structure, it is necessary to obtain the required resource according to respective interface rules. The cost is proportional to not the system throughput but the amount of resource used. As far as the realization of data mining based on cloud computing is concerned, users only need to pay more attention to the design of service logic layer instead of the details of bottom platform. Then, by running all kinds of algorithm under such cloud computing platform and setting reasonable response time, the satisfied results is obtained.

Cloud computing platform is dynamically flexible platform. An application program can be run only on a virtual machine when it need less resource. However, with the increase of resource demand, the computer power of current running environment will, first, become system bottleneck. When system monitor detect excessive load, cloud computing platform will automatically and dynamically request new virtual machines to join current running environment from cloud computing resource pool. Then, the current computer power will be encouraged until satisfying the resource demand of application program. However, with the decrease of system resource demand, virtual machines will be retrieved to resource pool so as to be used by other application program with high resource demand. Then, the process above mentioned of system dynamic could be expanded and contracted without user intervention, the system will automatically run. In contrast with the development of the local application, there is no great distinction for developers on their platform except in accordance with the norms and procedures followed easily extend transversely of the principles. It is extremely convenient for both the system developers and users.

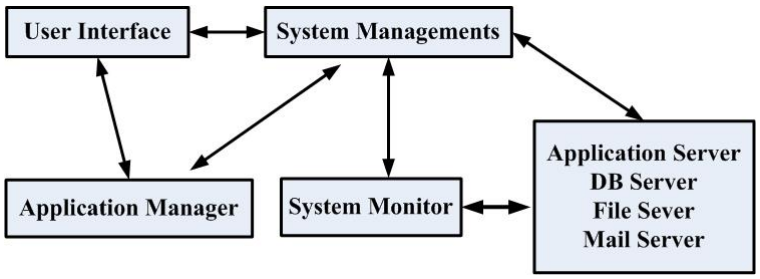

Figure 1. Architecture of cloud computing

3. Layers Construct for Data Mining Platform Based on Cloud Computing

During the process of designing data mining system based on cloud computing, hierarchical design thought is put forward, namely, the platform bottom-up is divided into two layers: algorithm layer, task layer and user layer. As is shown in figure 2, each layer servers for its top layer, the bottommost layer provides cloud computing platform with application program interface, and the topmost layer is user interface and opening interface which is used to share data set, call data cleaning, and mine algorithms so that the platform become more opening and could be easily integrated into user application. Target system, which is built on cloud computing, could be used by users not only directly by means of all kinds of terminals but also indirectly by opening interfaces of other application programs. The, users only need to pay more attention to adopt-

ing the algorithm dealing with data so as to get data mining results. In addition, internal module data mining platform provides services through the user interface and open interface, each layer opens to the outside world through the REST interface, which can be embedded into the application developers. In the process of algorithm design, the multi-layer plug-in framework structure is designed in order to increases the flexibility of implementation and maintenance.

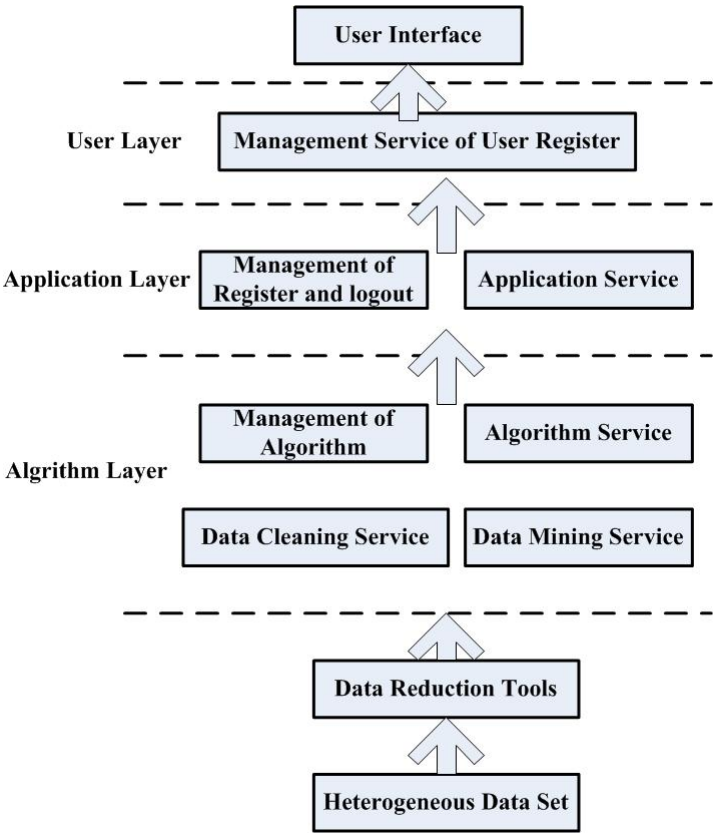

Figure 2. Layers Construct for Data Mining Platform Based on Cloud Computing

3.1. Design of Data Reduction Tools

As the object of data mining, data set is a set of same type data, and data set instance is an concrete data set which includes visiting address. The target of data reduction tools is to expand cloud computing platform so as to visit heterogeneous data and unidentified data. Therefore, data reduction tools could be used by data mining system to visit heterogeneous data.

Before accessing data set instance, users add a data set example definition file firstly by management module of data reduction tool to describe the data set instance, and then use the definition file ID to call conventions tool data access module for data reading and writing as shown in Figure 3.

In internal model of data reduction tools, the data access module provides reading service from data set, definition management module responding to the data definition service, registration and logout of data set definition service, definition analysis module including data analysis service, data set analysis service, and data set instance service. The definition file, heterogeneous data sets, and the relationship between them are described in Figure 3, in which the arrow means calling.

Definition analysis module can achieve to visit heterogeneous data sets undefined data type by means of abstract and extended metadata definition, data set template definition, and layer-by-layer abstract data set definition. Among them, the metadata definitions file, which store

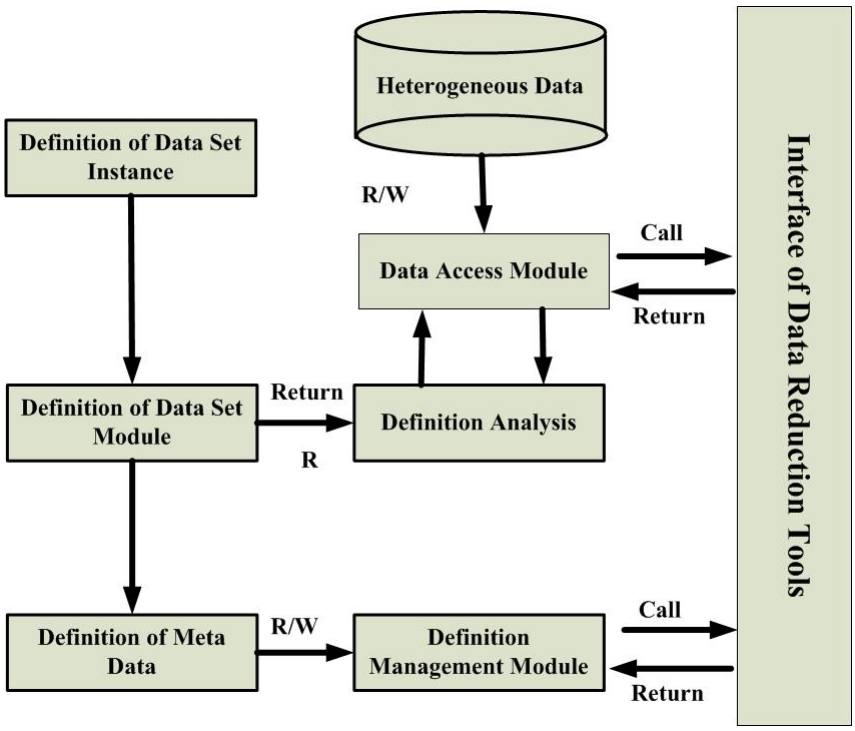

Figure 3. Internal model of data reduction tools

the definition of each data type, is maintained by service that is only open to the administrator, and provides atomic definition for data set template definition. Data set template file store user defined data set template, which is composed of data set definition service maintenance, data set registration service, and the actual data set corresponding and instantiation for data sets the instance file. Data set instance file is the basis of data access, in which data access components call service. By using this service, the specified data engine access to the specified location data is selected, and the meaning of data set used by various algorithms as parameters is returned. In other word, when the data set is accessed, the user can not only obtain the data content, but also the data set definition. Then, it will be resolved to be various atomic data type combination and be recognized by a variety of algorithm so as to satisfy the data layer abstract objective. Specific call process is defined as follows:

1) When user define the data set, the system will check whether the data set type defined is following the defined data templates or not. If the template is defined, the data set registration and cancelation of services, the data set stored position, and a data engine registration information to the data set example files are called. When templates are not meeting the requirements, data set definition services and use data analysis services is provided by a data element definition to define the target data set template. Finally, call data set registration and cancelation of services are completed.

2) Once user access to the data set, the system call the data access service, which can get instance file through the data access component service and call corresponding service to connect to the data source. Then, according to concrete data type recorded by the instance file, the data content will be returned. At the same time, the data access component service call the data set analysis service, and ultimately through data analysis services data set definition and return to the user.

3.2. Algorithm Layer

Algorithm layer achieve to call a variety of algorithms and manage interfaces by using uniform data source provided by next layer. According to the three kinds of algorithm execution order and returning result, algorithms is separated as follows:

1) Data cleaning algorithm: calling corresponding interface according to preprocessing method, that is implemented by the noise data set before data mining algorithms, and data after being cleaned will be deposited through the data layer in cloud computing platform storage space for the next data mining.

2) Data mining algorithm: using data having been cleaned or not be cleaned data as the unified call interface of data mining.

3) Visual algorithm: showing the results of data mining as tabular, graphic, or other styles.

4) Algorithm registration and logout: Algorithm management module, which manages kinds of algorithms modules by means of plug-in units.

3.3. Application Layer

Application layer can describe data and algorithms, which are involved in data mining process. Their relation and the order as task, moreover, can provide the call and the maintenance interface by application.

1) Application service: provide call interface having been registered.

2) Application registration and logout service: application management module, which manage various task definition files by way of plug-in units.

3.4. User Layer

User identity authentication and authorization function is provided in this layer.

User registration, authentication and authorization service will provide user identity authentication and authorization interface. Authorization information will be the passport that can call lower service so as to ensure the security of the platform. User management interface is also provided in this layer.

Services among layers mentioned above are described in the XML as the communication language. Based on the representation of the state change of Web service form internal call to better support the scalability, the services among layers will be ended with an open interface. The user may develop from any layer and load the existing services into the system. Then, the system's openness and ease of use is enhanced greatly, and is preceded by a data mining platform architecture.

4. Design of Key-points

4.1. Plug-in Unit System Frame

According to certain application development interface, the plug-in is a kind of program developed, whose structure is shown in Figure 4. Each plug-in is composed of three parts: expansion point which provides service for upper layer, business logic layer, new extension point which calls the lower layer. The above three parts are composed of a charge module management binding package with various services. By using bind package containing a service interface and various service interface, return the service calling method and specification for upper the caller could be returned to provide specific service parameter information. Bind package interface conforms to uniform standard. Then, once the plug is placed on the platform specific directory, plug-ins can be dynamic identification and loaded. At the same time, the algorithm plug-in service interface function parameter is atomic data type combination mentioned before the current algorithm. In other word, algorithm instead of a specific number, according to the specific arrangement of data to achieve specific, but to meet the algorithm under the premise of using the mentioned before abstract data to provide as much compatibility. Although the difficulty of algorithm realization is relatively increased, but the algorithm greatly enhances the reusability, and can be used for other users on platform to deal with various data processing.

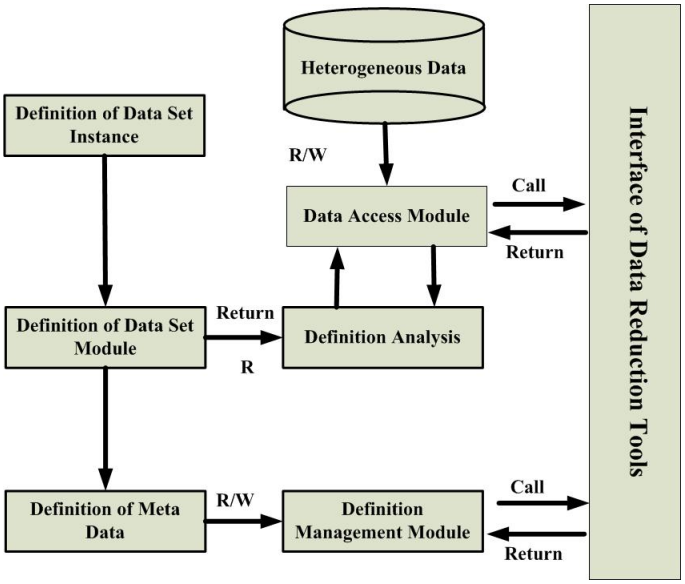

Figure 4. Internal model of data reduction tools

The advantages of the plug-in system framework include good system scalability, facilitates modular development, and open development team. This plug-in system architecture presented in this paper is composed of core layer, core plug-in layer, and user-defined layer. Excepted to the core layer, all plug-ins call to cloud computing platform resources to achieve the underlying architecture. Moreover, user-defined plug-in layer is opened to the outside and maintained by users. Then, the system scalability and openness of this system is enhanced greatly.

4.2. Design of Open Interface

The function of open interface is convenient for other applications on data mining platform to use a variety of resources and services provided by data mining platform. In order to achieve high performance, simplicity, intuitive open interface, and high scalability purposes, platform interface design is presented based on the REST (Representational State Transfer) [5] in this paper.

The resources in REST is the combinations of data and forms that is classified into different resource according to their different forms of expression. All resources are uniquely identified by URL (Uniform Resource Identifier). REST is based on the Http protocol. Thus, any operation on resources is achieved by the Http protocol, which put on an operational resources including GET, POST, PUT, DELETE, and realize creating , reading, updating, deleting, and other operations. It should be noted that the resource and URL are one-one corresponding. Then, in contrast to previous web development and greatly simplifies the Web development, URI will be not changed when these operations are executed. The URI could be designed to reflect the resource structure more precisely. REST can improve the scalability of the system, because it requires that all operations are stateless. In the absence of contextual constraints, it is more simple for distributed system and cluster to make the system more efficient by using the buffer pool. Because the server does not need to record the client a series of visits. Then, the load of server is reduced.

5. Conclusion

Cloud computing is an computing platform that can provide dynamic resource pools, visualization and high-availability. Because of some problems, such as the noise data and heterogeneous structure, data mining solutions by means of cloud computing have been put forward. In view of the above question, the layers construct for data mining platform based on cloud comput-

ing was put forward in this paper.

During the data mining system design process based on cloud computing, based on the hierarchical design, the platform level bottom-up algorithm is divided into: algorithm layer, task layer and user layer, wherein, the bottom layer transparently server for upper layer services, the upper layer call the lower layer by the opening interface, which makes functions among the layers relatively independent and be convenient to two times the development of system. Interface of each layer is REST interface which is open to the outside and can easily be embedded into the application programs. In design process of the algorithm design, multilayer plug-in framework structure is designed in order to increases the flexibility of implementation and maintenance.

Acknowledgement

This work was supported by A Project of Shandong Province Education Science Program (ZK1101322B021) and Science Foundation of University of JinanXKY1020.

References

[1] Buyya R, Yeo CS, Venugopal S, Tamboli JA, and Joshi SG. Market-Oriented Cloud Computing: Vision, Hype and Reality for Delivering IT Services as Computing Utilities. *Proceedings of the 2008 10th IEEE International Conference on High Performance Computing and Communications.* 2008: 5-13.

[2] Qi JJ, Liu AJ, Lei Y, and Xu HF. Research on XML schema based manufacturing information integration specification. *Jisuanji Jicheng Zhizao Xitong/Computer Integrated Manufacturing Systems*. 2005; 11(4): 565-571.

[3] Chang MF. An Introduction to Cloud Computing Service Platform-Google App Engine. *Computer and Communication*. 2008; 126: 24-33.

[4] Xu W, Li Z, Cheng C, Zheng T. Data mining for unemployment rate prediction using search engine query data. *Service Oriented Computing and Applications*. 2013; 7: 33-42.

[5] Fabian G, Van Eyck J, Meyfroidt G. Predictive data mining on monitoring data from the intensive care unit. *Journal of Clinical Monitoring and Computing*. 2013; 27: 449-453.