Research Journal of Applied Sciences, Engineering and Technology 6(3): 433-436, 2013
DOI:10.19026/rjaset.6.4097
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: August 17, 2012 Accepted: September 27, 2012 Published: June 15, 2013

Research Article

Research on Data Integration Based on Cloud Computing

[1]Yanxia Wang and [2]Huijun Liu
[1]College of Computer and Information Science, Chongqing Normal University, Chongqing 400047, China
[2] College of Computer Science, Chongqing University, Chongqing, China

Abstract: In this study, we give some strategies for selecting the data source and several methods for data integration after analyzing the problems of sharing information resources among universities. According to the characteristics of various types of information resources in the university website, we propose the data integration framework model which combines virtual view method with data warehouse method.

Keywords: Cloud computing, cloud computing architecture, data integration, data warehouse, federated database, middleware model

INTRODUCTION

With the rapid development of Internet, websites become important medium for spreading and exchanging of information. Abundant information can be found on the web, which realizes the information sharing and improves the utilization of resources in some sense. But it also forms "Information Island" in some ways. In our country, most of colleges and universities build their own teaching resources with high expenses. So these resources are open to all of the teachers and students in the same institution and partly available to other institution because of some reasons. These resources are often similar since the students learn the same lessons such as English, mathematics, computer application and have similar learning materials for the similar lessons. Cloud computing provides a new pattern for efficient sharing of information, hardware and software resources.

Cloud computing can be defined as "a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers" (Buyya *et al*., 2008). Current examples of Cloud computing include Microsoft Azure (Chappell, 2008), Amazon EC2, Google App Engine and Aneka (Chu *et al*., 2007). Clouds refer to the data center hardware and software and divide into Public Cloud and Private Cloud. When a Cloud is made available in a pay-as-you-go manner to the public, we call it a Public Cloud; the service being sold is Utility Computing. Private Cloud refers to internal data centers of a business or other organization that are not made available to the public. Clouds aim to power the next generation data centers by architecting them as a network of virtual services (hardware, database, user-interface, application logic) so that users are able to access and deploy applications from anywhere in the world on demand at competitive costs depending on users QoS (Quality of Service) requirements (Calheiros *et al*., 2009). Thus, Cloud Computing is the sum of SaaS and Utility Computing, but does not normally include Private Clouds. According to the understanding of the cloud computing, the data resources on the Internet can be shared by data integration based on cloud computing platform.

Data integration (Malcolm *et al*., 2003) is to integrate a variety of heterogeneous data to provide a unified view to users. For users, data source is transparent, namely, it shields the difference of underlying data source, let users feeling data from a large data source. At present, the relatively mature data integration methods are federated database, based middleware model and data warehouse. Federated database system for data sharing uses data-exchange format to construct the mapping between data sources and provides access interface among the various data sources, but with the increase of integrated systems, the cost will be doubled. Therefore, the federal database integration system is suitable for autonomous database of less. Data warehouse (Diego *et al*., 2001) makes more data sources convert a unified model to store the integration data in accordance with requirements of a unified view. Data warehouse is suited to applications of decision-making. Middleware model (Yangjin, 2009; Sun *et al*., 2007) provides unified logical views to hide

Corresponding Author: Yanxia Wang, College of Computer and Information Science, Chongqing Normal University (Huxi Campus), Chongqing, 400047, China, Tel. : 023-65910270

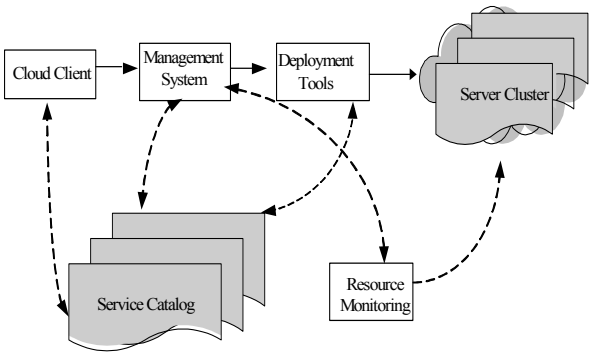

Fig. 1: Cloud computing architecture

the implementation details of the underlying data and makes users to consider integrate data sources as a unified whole, but not the physical integration of data, it is mainly to receive the user's data request and transmit the final receiving results to the user. The method is applicable to Web data integration. Because of the Web data coming from various organizations or individuals, there is no fixed data model. Even if the same semantics uses different data types. Hence, the integration of the Web data resource is very difficult. The study mainly analyzes the content and ways of network data access and designs Web data integration solution based on the cloud service platform to provide a unified portal and personalized website.

Cloud computing architecture: Cloud computing makes full use of network and computer technology to share the resources and services. According to the service mode, computing clouds can be divided into three service types (Cheng and Bing, 2009; Rao and Vijay, 2009), namely Infrastructure as a Service (Iaas), Platform as a Service (PaaS), Software as a Service (SaaS). Infrastructure as a Service offers hardware, software and equipments for users. In the case of a particular service constrains, IaaS provides an intermediate platform to run arbitrary operating systems and software. Platform as a Service provides a high-level integrated environment to design, build, test, deploy and update online custom applications over virtualization resources and it is the middle part between infrastructure resource and upper application (SaaS). Software as a Service (SaaS) will integrate all application software and data resources in the cloud for providing software application, resource library and user interaction interface and so on for users, which mainly provides service by Web Service and offers services access using a Web browser.

Cloud computing is a huge service network constituted parallel grids, expands the service capability of cloud client by virtualization technology and provides supercomputing and storage capacity by the cloud computing platform centralizing cloud client resources. General cloud computing architecture shows in Fig. 1 (Changcheng, 2010).

In the cloud computing architecture, users select the desired service from the service catalog through cloud clients; the request schedules corresponding resources by the management system, distributes requests and configures the web application through deployment tools. For now, many data resources are distributed as well as heterogeneous. For example, educational management systems and scientific research management systems in the same university adopt different data structures; Data resources among universities are not shared among universities because of distribution in different colleges and universities, so that there exists duplicate data. Therefore, the scheduling data resources in cloud computing platform firstly need to integrate. As data integration in traditional information management system is mainly on the mapping between heterogeneous databases, it is relatively simple; Web information data integration is very difficult because of no fixed data model, data organization arbitrariness, dynamic changes of data contents and representations. According to the content type on the website, this study takes the university web site's content as an example to study data integration ways to provide specific pages, display specific content for a particular user. University website contents broadly include text, courseware, video, data, etc. By analyzing the characteristics of querying information and web-side information, the cloud system separates data integration ways into different types. The first type is that links to all the university web site in cloud system, when users search for information to display the entire website. In this way it is clear that the cloud system does not play any role and there is no difference from a Google search, but in some cases, this approach is the simplest and most effective, such as when a user queries a college introduction, recruiting information and so on. Second, cloud system integrates data inside. Cloud system inside links several representative universities, when a user searches for information through the cloud platform, cloud system integrates information according to the website in cloud system. The information provided in this way in some cases, relatively speaking, than the first way is better, such as a user queries courseware, video and so on. However, this way has some limitations; because users want to search for information which websites within cloud system do not have necessarily the best or incomplete. The third way is that cloud system within integrates information referring also to external network resources. Integration of information in this way is relatively good, but the cloud system is difficult to determine referencing websites and integration efficiency is relatively lower than the second way. According to different integrated ways, the functions of service catalog and server cluster of the cloud computing architecture have a greater change and others change small. The following we give in detail the function of each part of the cloud computing architecture.

Cloud client: Provide service interaction interface, users can register, login, customize services, configure and manage users through the Web browser; Offer access interface and fetch user requirement in Web Services.

Service catalog: Provide list of services. Cloud users after obtaining the appropriate permissions (pay or other restrictions) are allowed to choose or customize the list of services and also cancel the existing services. And the corresponding icons or a list is generated on the interface of the cloud client to display related services. According to the analysis of website information integration ways, the different integration way, the content of the service list is different. For the first integrated way, the service catalog lists directly university websites; and for the second and third integration ways, the service catalog displays lists after integration.

Management systems and deployment tools: Provide management and services. Namely, manage cloud users, such as user authorization, authentication and register; and be also responsible for the management and distribution of service resources, receive requests sent by the user. According to user requests, transponder requests to corresponding application, scheduling of resources and dynamically deploying configure, recycling resources. The core is load-balancing.

Monitoring: Monitor and measure the usage of the cloud system resource in order to make rapid response, complete the node synchronization configuration, load balancing configuration and resource monitoring to ensure that resources are well allocated to the appropriate users.

Server cluster: In addition to a general cloud architecture has the features, server clusters perform different functions according to different integration ways of the university website information. For the first integrated way, server cluster does not do any treatment, but for the second and third way, it processes not only information adopting different ways by various datum, but also make a decision whether to store the result of integration according to the integrating algorithm complexity. For example, the integrating algorithm is complicated and even needs human intervention and then the integration results are sent to the service catalog as well as stored in servers. Using relatively simple algorithms to integrate information, the integrated data needs only to send the service catalog, not store.

According to different integration ways and website information types, the following are different methods of data integration and the related technologies.

CONCLUSION

The study analyzes in detail information resources sharing among universities, the difference of data integration between the traditional database and web information integration and describes the cloud computing architecture combining education field. Data integration ways and the data source selection strategy based on cloud computing system are proposed according to various types of website information resource and we propose the data integration framework model which combines virtual view method with data warehouse method, at the time, the framework of the execution model which is very crucial in the data integration model is given. This study is of great significance for the network information resources sharing. How to concretely implement data integration will be our future work.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This study is supported by Natural Science Foundation Project of CQ CSTC (No: cstc2011jjA40027), Scientific Research Program of Chongqing Municipal Commission of Education (No. KJ120634), Dr. Research Funds of Chongqing Normal University (No: 10XLB19).

REFERENCES

Buyya, R., C.S. Yeo and S. Venugopal, 2008. Market-oriented cloud computing: Vision, hype and reality for delivering IT services as computing utilities. Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications, University of Melbourne, Australia, pp: 5-13.

Calheiros, R.N., R. Ranjan C.A.F. De Rose and R. Buyya, 2009. Cloudsim: A novel framework for modeling and simulation of cloud computing infrastructures and services. Technical Report, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory, the University of Melbourne, Australia.

Cheng, Z. and L. Bing, 2009. Rearch on the stack model of cloud computing. Microel. Ectronics Comp., 26(8): 22-27.

Chu, X., K. Nadiminti, C. Jin, S. Venugopal and R. Buyya, 2007. Aneka: Next-generation enterprise grid platform for e-science and e-business applications. Proceedings of the 3rd IEEE International Conference on E-Science and Grid Computing, pp: 151-159.

Changcheng, Q., 2010. The research of active architecture based on cloud computing. Wuhan University of Technology, pp: 17-18.

Chappell, D., 2008. Introducing the Azure Services Platform. Retrieved from: http:// download. microsoft. com/ download/e/ 4/3/343bb 484-3b52-4fa8-a9f9-ec60a32954bc/Azura_services_platform.dox.

Diego, C., D. Giuseppe, Giacomo, L. Maurizio, N. Daniele and R. Riccardo, 2001. Data integration in data warehousing. Int. J. Coop. Info. Sys., 10(3): 237-271.

Malcolm, P.A., D. Vijay, G. Leanne, W. Paul, S. Toney and P. Dave, 2003. Grid Database Access and Integration: Requirements and Functionalities [J/OL]. http:// www. gridforum. org/ documents/ GFD.13. pdf.

Rao, M. and S.Vijay, 2009. Cloud Computing and the Lessons from the Past. Proceeding of the 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises. IEEE Computer Society Washington, DC, USA, pp: 57-62.

Sun, Y.C., C.L. Song and R.Z. Li, 2007. A middleware of heterogeneous data integration based on Web service. J. Xi, Uni. Sci. Technol., 27(2): 284-287.

Yangjin, 2009. Design and implementation of a data integration model based on DDS and XML. Ph.D. Thesis, Beijing University of Posts and Telecommunications, pp: 5-8.