Research on Web Data Integration Framework Based on Cloud Computing

Yanxia Wang
College of Computer and Information Science, Chongqing Normal University, Chongqing, 400047, China
Email: cloudwyx@163.com

***Abstract*—According to the sharing of information resources among universities, web data integration framework is proposed . The analysis of the current using situtation of the network information resource as well as the characteristics of various types of information resources of the university website, the paper presents the several ways for data integration according to the characteristics of the different resources, sets the data source selection strategy.**

Keywords Cloud Computing; Data Integration; Cloud Computing Architecture; Middleware Model; Data Warehouse.

I. 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 and a portion of them are analogous to each other. 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. 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 is a new business computing model and it is developed from distributed computing, parallel computing, and grid computing. It builds data center or super computing infrastructure though distribution and virtualization technologies and provides the technology developers and enterprise clients with data storage, analysis and science computing without charge or with payment according to the requirement. Generally, cloud computing also means that it provides all kinds of clients with different services such as software, hardware, and data by web server clusters. And thousands of computers in the internet will send the required resources back when the local computer sends a requirement through the internet. A simple example of cloud computing is Yahoo email, Gmail, or Hotmail etc. A consumer does not need a software or a server, all he would need is just an internet connection and he can start sending emails. The server and email management software is all on the cloud (internet) and is totally managed by the cloud service provider Yahoos, Google etc. The consumer gets to use the software alone and enjoys the benefits.

Data integration [1] 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 [2] 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 [3,4] provides unified logical views to hide 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. However, because the Web data is not controled by any department or organization, but 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 paper mainly analyzes the network data access content and model, and design Web data integration solution based on the cloud service platform to provide a unified portal and personalized website.

The remainder of this paper is organized as following. Section 2 introduces cloud computing architecture and the function of each part combining education field. Section 3 gives data source selection strategy and data integration method of netwok education resources. Conclusions and future work is in Section 4.

II. Cloud computing architecture

Cloud computing makes full use of network and computer technology to share the resources and services. According to the service mode, clouds computing can be divided into three service types [5,6], 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,

978-1-4577-1415-3/12/$26.00 ©2012 IEEE

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) concentrates 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 with WebService 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 Figure 1 [7].

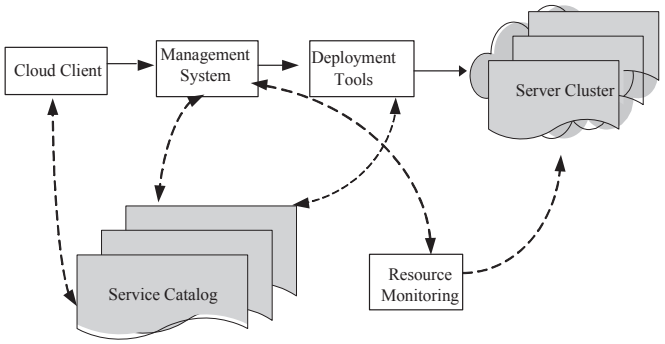

Figure 1. Cloud computing architecture

In the cloud computing architecture, users select the desired service from the service catalog through cloud clients, the request scheduls corresponding resources by the management system and distributes requests, configures the web application through deployment tools. For now, many data resources are distributed, heterogeneous, for example, educational management systems and scientific research management systems adopt different data structures in a university so that there exists duplicate data; data resources are also not shared among universities, and the waste of resources is more serious. Therefore, the scheduling data resources in cloud computing platform firstly needs to integrate. Data integration in traditional information management system is mainly on the mapping between heterogeneous databases; 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 paper 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 inquiring information and webside information, the could 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 all the 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 inquires 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, is better than the first way, such as a user queries courseware, video and so on. However, this way has some limitations, because users want to search for information, wedsites within cloud system do not have necessarily the best or incomplete information. The third way is that cloud system within integrates information refering to external network resources. Integration of information in this way is relatively good, but referencing websites the cloud system are difficult to determine, 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 folowing 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 uers, such as user authorization, authentication, login; And be also responsible for the management and distribution of service resources, receive requestes sent by the user. According to user requests, transpond requests to corresponding application, scheduling of resources, and dynamically deploying, configuring, recycling resources. The main task 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 makes 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, 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 to store.

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

III. Data integration

A. Data source selection strategy

Web information data sources include a variety of topics of data resources, and there are many data sources on a topic, although these sources belong to the same topic, the quality of data varies widely: some out of date, inaccurate or inconsistent, and something to update timely, accurate and consistent. And the data source covers each other, even some even completely contains the other data source. In the educational field, there are thousands of colleges and universities, almost all universities have set up basic courses such as english, mathematics, basic computer, etc., when the user searches for related resources, how to select the data source to provide users with high quality resources to become one of the key problems. According to the characteristics of the network education resources, we put forward a data resource selection strategy. First, resources are classified according to the content of network education resources, and then based on the classification of resources universities are classified.

There are mainly the related information about universities, course descriptions, teaching outlines, courseware, course video, discipline construction and so on in university websides. Coures and disciplines are the national, provincial or university elite courses and key disciplines respectively. Data sources are selected by the rank of courses and disciplines,or else according to well-known colleges and universities. The strategy can guarantee to provide users with high quality results of the inquiry, and that can reduce the number of the visiting database, so that the redundancy of query results is low, meanwhile it reduces the querying cost and improves the efficiency of data integration.

B. Data integration method

In general, depending on the data distribution location and the coupling degree of function interfaces, information integration mechanism can be divided into two methods [8]:virtual view and data warehouse methods. In the virtual view way, the data is not stored in local, still remaining in their original system. When the client sends search requests, the integrated system will send request to appropriate data sources, and finally merge returned search results. Data warehouse method is that the sharing information extracted from various data sources is stored in a central database before users put forward inquiry requests. Integrated methods of the two traditional structure have different advantages. Virtual view method can provide the latest datum to users, so it is more suitable for the integration of quick update Web information. Data warehouse method has the fast inquiry advantage, suitable for not always change data source. According to the content of the university website resources and characteristics of resources integration, we give the data integration framework model which combines virtual view method with data warehouse method, as shown in Figure 2.

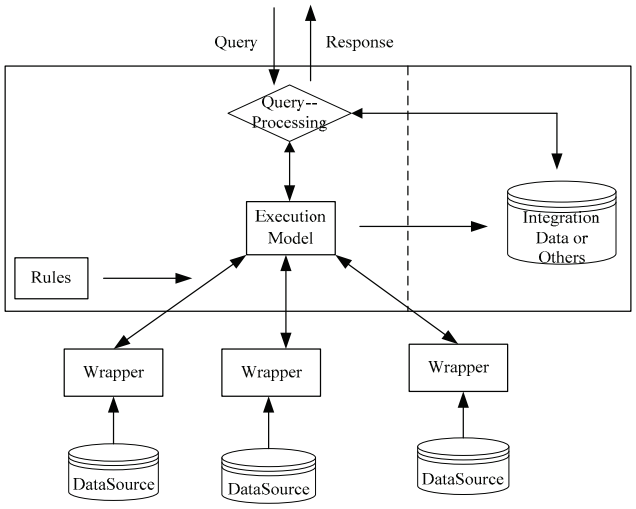

Figure 2. Data integration framework model combining virtual view method with data warehouse method

IV. CONCLUSION

The paper 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 the framework model of data integration is gived. This paper is of great significance for the network information resources sharing. the execution model of the framework model of data integration is key, so how to design execution model and concretely implement methods will be our future work.

Acknowledgment

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

References

[1] Malcolm P A, Vijay D, Leanne G, et al. Grid Database Access and Integration: Requirements and Functionalities [J/OL]. http://www.gridforum.org/documents/GFD.13.pdf.

[2] Diego C, Giuseppe de G, Maurizio L, et al. Data integration in data warehousing [J]. International Journal of Cooperative Information systems, 10(3): 237-271, 2001.

[3] Yangjin. Design and Implementation of a Data Integration Model Based on DDS and XML. Beijing University of Posts and Telecommunications. Thesis, 2009.2. (Chinese)

[4] Sun You-cang, Song Cai-li, Li Run-zhou. A middleware of heterogeneous data integration based on Web service. Journal of Xi an University of Science and Technology, 27 (2), 2007. (Chinese)

[5] Z.Cheng, L.Bing, Rearch on the Stack Model of Cloud Computing. Microel Ectronics&Computer , Vol.26, No.8, 2009.8. pp22-27.

[6] Dr. Rao Mikkilineni, Vijay Sarathy. Cloud Computing and the Lessons from the Past. 2009 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009, pp57-62.

[7] Changcheng Qiu. The Research of Active Architecture Based on Cloud Computing. Wuhan University of Technology, Thesis, 2010. (Chinese)

[8] Fang Wei. Research on Key Technologies of Ontology--Based Deep Web Information Integration.Soochow University, Doctor, 2009.(Chinese)