Open Access

Research and Application of Redundant Data Deleting Algorithm Based on the Cloud Storage Platform

Long-long Song[1,*], Tai-yong Wang[1,2,], Lan-ying Zhang[3] and Xiao-wen Song[4]

[1]School of Mechanical, Electronic and Control engineering, Beijing Jiaotong University, Beijing 100044, China; [2]School of Mechanical engineering, Tianjin University, Tianjin 300192, China; [3]School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China; [4]R&D Center, CSR Qingdao Sifang Co., Ltd., Qingdao 266111, China

Abstract: The management of information resource has become the key point to the large complicated system with the development of information technologies. Problems of storing pressure and reading efficiency became more serious because of data explotation and large scale of redundant data. A novel redundant data deleting algorithm based on the cloud storage plalform was proposed in this paper to deal with the problem of mess data store. Space division and role division techonlogy was used to insure the security of data sharing in the system. A novel educational administration information system was used to validate the feasibility and effectiveness of the algothrisms proposed in this paper.

Keywords: Cloud storage platform, data explosion, data sharing, redundant data deleting, space and role division.

1. INTRODUCTION

With the increasing application of information technology, the data and information increasing exponentially in large complicated administration information systems, which has brought up high storing pressure and reading pressure [1]. Existing data centers faces some serious problems such as dynamical expansibility, data security and the openness and sharing of the data resources [2, 3]. The cloud service provides a virtualized, dynamical and scalable model for organizing and distributing computing resources on demand [4].

Cloud storage has become more and more popular since the successful simple storage service (S3) of Amazon [5]. Some commercial cloud storage platforms provide reliable storage service and can be friendly accessed through the Internet. Most of them, such as Amazon S3, Microsoft Live Mesh, Mozy and Symantec's Protection Network (SPN) [6-8], focus on providing interfaces of storage similar to traditional file systems. A trend of new educational administration information system based on the cloud technology formed in recent years. Since early 2006, Amazon cloud services have supported universities and colleges by providing small grants to supply computing power, storage and other services for research and other educational purpose [9]. ICES@BUPT is a system developed on the cloud storage platform (consulted Dynamo of Amazon [10], which relied on consistent hashing to distribute the load across multiple storage hosts [11]) and this system solved the issues of data storing and data sharing between different users. All manuscripts must be in English, also the table and figure texts, otherwise we cannot publish your paper.

2. SYSTEM ARCHITECTURE BASED ON CLOUD STORAGE PLATFORM

A novel educational administration information running on the cloud storage platforms, with the structure of B/S, can provide services such as secure courseware sharing in the class, secure homework sharing to the specified teacher who has the right of reading/writing, on-line score management and so on. The architecture of this system includes cloud storage platform layer, storage management layer, supporting layer and service interface layer.

2.1. Cloud Storage Platform Layer

Cloud storage platform layer relies on the consistent hashing to distribute the load across multiple storage hosts and the extensibility in adding/removing hardware nodes. When a node fails, another node would automatically replace this node to provide service. All the process is transparent to users.

2.2. Storage Management Layer

In the storage management layer, management module could delete redundant data in the local host. And the data is sent to appropriate nodes of cloud storage platforms. All the data is classified into structured data and unstructured data. The structured data, stored in the database, including tables of educational information and that of users' information, provides basis for roles and space division. The unstructured data, stored in the cloud storage platform, includes private and sharing data.

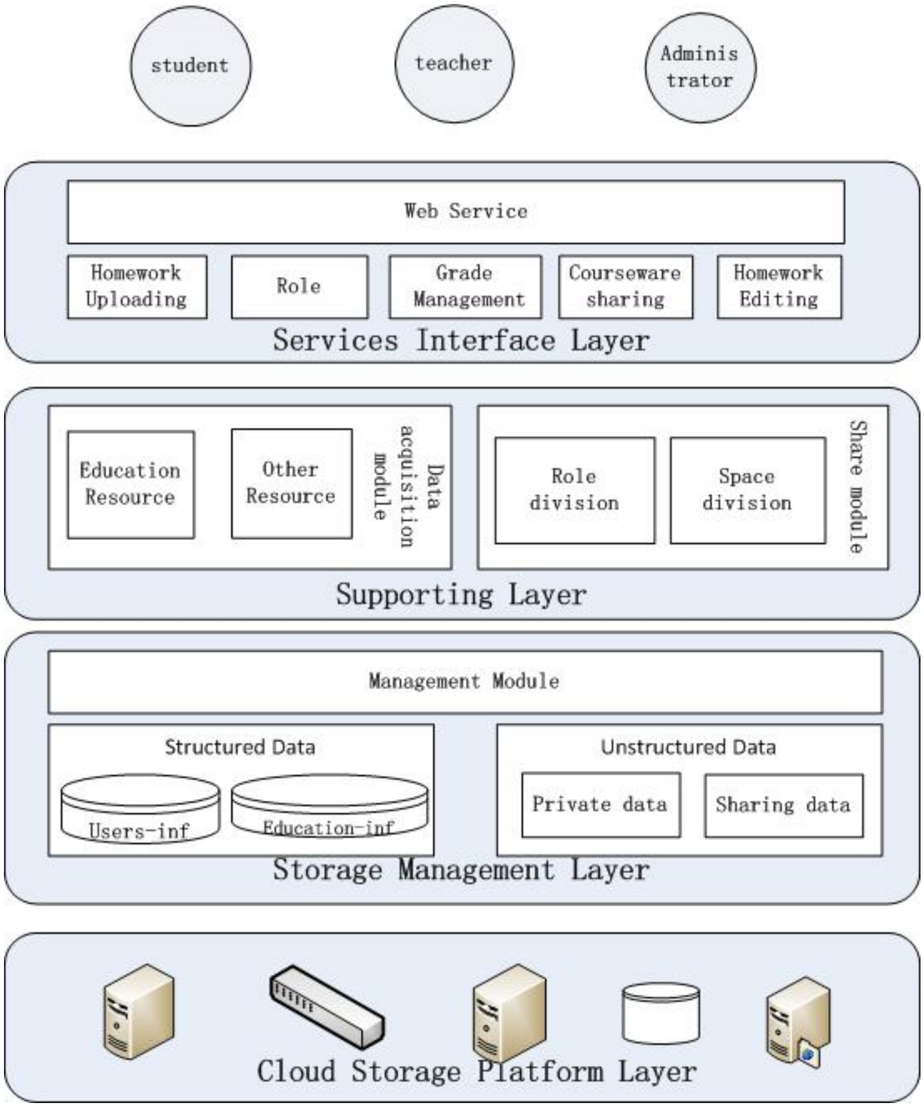

Fig. (1). System architecture based on cloud storage platform.

2.3. Supporting Layer

Supporting layer, based on the storage management layer, including data acquisition module and sharing module, provides technical support for the upward layer. Data acquisition module is responsible for the format conversion. Sharing module, including role and space division, assures the safety of data sharing. The functions involve five aspects as followed.

(1) Format conversion: The structured data (information of student, course and class) is stored into database by the data acquisition module.

(2) Role division: Roles of the users include teacher, student and administrator. Different roles have different rights.

(3) Space division: The storage space is departed into class-partition, course-partition and public-partition. When users create/delete classes or courses, system would automatically create/delete the virtual directory.

(4) Data information management: When the users upload/delete/remove data, system would automatically store/delete/move and do mapping records for the data.

(5) Services interface layer: In this layer, all the users get web service on their roles. All the web services provided by this system include safe courseware sharing among the course class, safe homework sharing to the special teacher who has rights to read/write this data, on-line score management, roles division and so on.

3. KEY TECHNOLOGY

The information administration system aims to assure the safety of data storing and data sharing. The cloud storage platform relies on the consistent hashing cloud storage technology. In the management module, redundant data could be deleted in the local host. The support layer could provide space and role division to deal with the safety of data sharing.

3.1. Cloud Storage Technology

Cloud storage platform needs to distribute the load by the use of multiple storage hosts and extensible adding/removing hardware nodes. When a node failure occurs, another node would automatically replace this failed node and provide service.

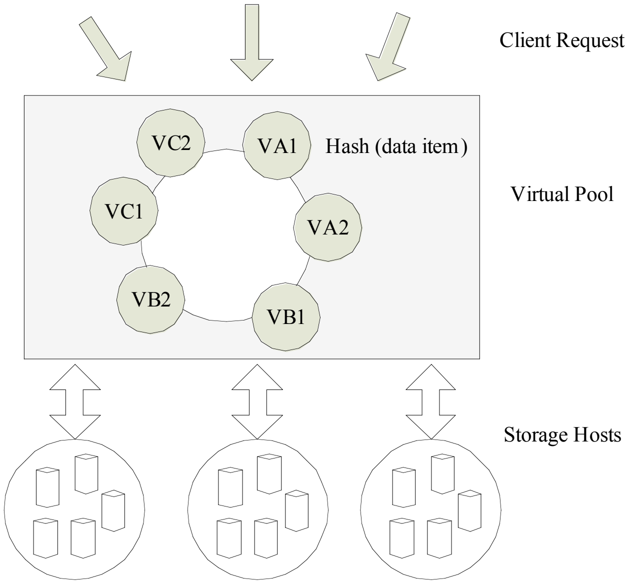

Fig. (2). Architecture of cloud storage platform.

As Fig. (**2**) illustrated, the architecture of the platform includes the storage hosts layer and the virtual pool layer. The storage hosts layer, the basis for the cloud storage platform, is constituted by a number of physical hardware. The virtual pool layer is designed as a ring which relies on the consistent hashing to distribute the load across multiple storage hosts. In the rest of the paper, node is stand for the storage host. Data item represents the smallest unit of simple read and write operations. N stands for the number of storage hosts. Virtual node represents the replica of node. VA1/ VA2 represent the virtual node.

Key K1 = hash (node) % N (1)

Key K2 = hash (data item) (2)

Map (node A) → (VA1, VA2) (3)

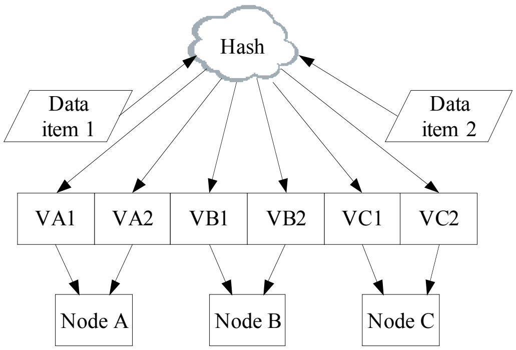

Fig. (3). Consistent hashing.

In equation 1, each node is assigned a key value which represents its "position" on the ring. To address load distribution issues, each node get assigned to multiple points on the ring in equation 3. As Fig. (**3**) showed, every node is mapped into two virtual nodes. Each data item is assigned to a virtual node by the data items' key (in equation 2) to yield its position on the ring, and then goes through the ring clockwise to find the first virtual node with a position larger than item's position (Fig. **2**, VA2). Thus, each virtual node could be responsible for the region between itself and its predecessor on the ring. Only its immediate neighbors can be affected when a new virtual node is added /removed. Other nodes remain unaffected.

3.2. Delete Redundant Data

Because the sharing information involves information about courseware, homework and examination papers, there are a lot of redundant data in the storage platform. The system could delete the redundant data at local machine and send data to the cloud storage platform. Fig. (**4**) showed the process of deleting the redundant data during the process of homework sharing.

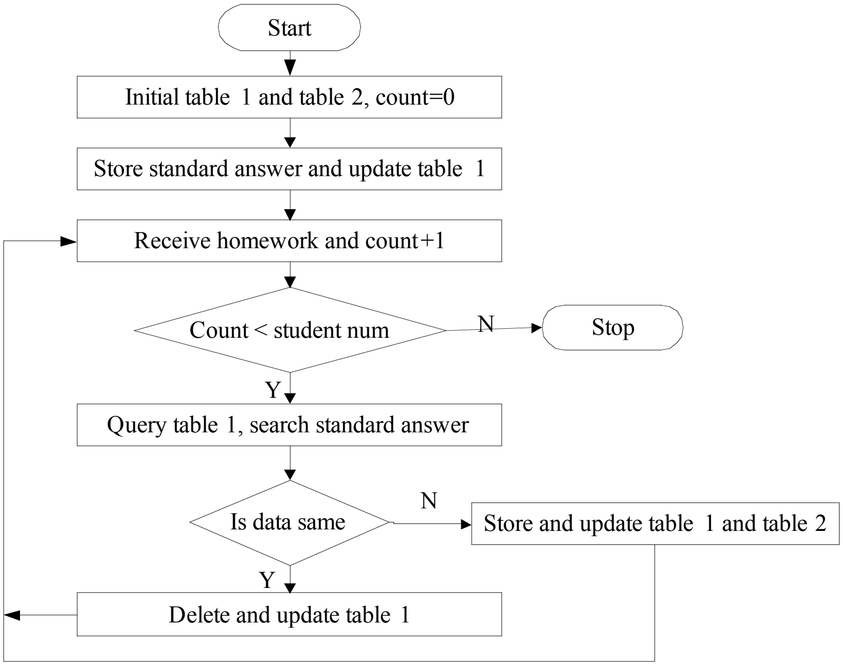
Start
Initial table 1 and table 2, count=0
Store standard answer and update table 1
Receive homework and count+1
Count < student num
N
Stop
Y
Query table 1, search standard answer
Is data same
N
Store and update table 1 and table 2
Y
Delete and update table 1

Fig. (4). Process of deleting redundant data.

Table 1. Role and space division.

Roles	Class-Partition	Course-Partition	Public-Partition
Administrator	Read/Write	Read/Write	Read/Write
Student	Related class member Read/Write	Related course member Read/Write	Read/Write
Teacher	Related class member Read/Write	Related course member Read/Write	Read/Write

In Fig. (**4**), table 1 records the address of the standard answer and the compared result between the students' homework and the standard answer. table 2 records the address of the wrong part of the students' homework.

3.3. Role and Space Division

The system classifies all the users into teachers, students and administrators. Different roles have different rights and can implement different functionalities.

(1) Administrators are responsible for the maintenance of the database, the management of all the users and the division of the storage space.

(2) Teachers are in charge of the management of the educational resource including courseware sharing, homework editing and grade managing.

(3) Students have the right of getting the courseware from the teachers, uploading homework and taking part in course discussion etc.

The storage space is divided into Public-Partition, Class-Partition and Course-Partition.

(1) Public-Partition: This partition could be visited by all of the teachers and the students. All the users could share their own resources and access other information stored in this partition.

(2) Class-Partition: This partition is served for the administrative classes. Only the members in the class can be allowed to access the data stored in this partition.

(3) Course-Partition: If a teacher shares data in this partition, for example the courseware information, only the students who have selected this course could access this data. When students share information, only the related teacher has the rights to read/write this sharing information. Table **1** shows this division strategy.

4. IMPLEMENTATION AND APPLICATION

The functions that have been implemented include role division which defined all the rights as different roles (administrators, teachers and students), space division (Public-Partition, Class-Partition, and Course-Partition), safe data sharing (homework/courseware safely shared between teachers and students), and students' performance management (teachers could edit students' homework and give scores online). System operating environment could be described as followed.

4.1. Cloud Storage Platform

IP SAN/NAS storage system;

Hard disk: SATA 2TB;

Database: MySQL15.15;

Server: Tomcat6.0/ jdk6.0/ngnix:0.8.36.

4.2. The Client Operation Environment

CPU: 2.93GHz;

Memory: 2.00GB;

Hard disk: 250G;

Browser: IE browser/ Firefox browser.

4.3. Test Case

Duration: two semesters;

Participants: five hundred students and two teachers.

4.4. Result

The effect of cloud storage and deleting redundant data: there are over 500 students and each of them has 2GB virtual storage space. So the total storage space should be equal to 1000GB. But in our system, the actual space was just 3575MB in the cloud storage platform after deleting the redundant data.

The test of other functions is described as followed.

(1) Format conversion function: Information of the students, the courses and the classes were stored into the database. All of the users were divided into different roles and got their storage space automatically in our system.

(2) Role assignment function: The default user of the system is the administrator and he/she could divide other users into teachers and students.

(3) File operating function: All of the users could upload, download, move, share, copy, delete and rename the files.

(4) Data sharing function: Students could share their homework to the specified teacher. And the teachers could share courseware to a course class.

(5) Online editing function: Teachers could open and edit homework and submit a score for every student on-line.

(6) Query function: Teachers could get the homework according to the course_name or the submit_ID. Students could get courseware by the course_name or the teacher_name.

(7) Scores management function: Teachers could edit the students' scores and computes average scores in a specified weight on-line.

CONCLUSION

This paper presented an educational system based on the cloud storage platform which was very convenient for different users to share the information and data safely. By redundant data deleting algorithm and cloud storage technology, the system relaxed the pressure on network bandwidth and reduced the requirement of storage resource. For the data sharing, it provided role and space division to ensure different users get their legal information.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The research is supported by excellent doctoral candidate innovation research foundation of Beijing Jiaotong University (No. 2014YJS113). We also thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available key-value store", *J. ACM SIGOPS Operating Systems Review - SOSP '07 Homepage*, no. 41, pp. 205-220, 2007.

[2] Information on http://aws.amazon.com/ec2

[3] Y. Khmelevsky, and V. Voytenko, "Cloud computing infrastructure prototype for university education and research," Canada: University of British Columbia Okanagan, WCCCE - *Proceedings of the 15th Western Canadian Conference on Computing Education*, 2010.

[4] Information on http://www.amazon.com/s3

[5] Information on http://www.mesh.com

[6] Information on http://www.spn.com

[7] G. Li, and G. Chen, "A novel enhanced education application of cloud computing," In: *CCIS - Proceedings IEEE International Conference on Cloud Computing and Intelligence Systems*. Beijing, PR, China, pp. 526-529, 2011.

[8] B. Wang, and H. Xing, "The apping in education informatization," In: *CSSS - Proceedings International Conference on Computer Science and Service System.Nan jing*, P.R. China, pp. 2673-2676, 2011.

[9] L. P. Li, and W. Toderick, "Cloud in cloud: approaches and implementations," In: *Proceedings of the ACM Conference on Information Technology Education*, New York, NY, USA, pp. 105-110, 2010.

[10] L. M. Vaquero, "EduCloud: PaaS versus IaaS cloud usage for an advanced computer science course", *J. IEEE Trans. Educ.,* vol. 54, pp. 590-598, 2011.

[11] Y. Cheng, D. Zhao, A. R. Hu, Y. L. Luo, and L. Zhang, "Multi-view models for cost constitution of cloud service in cloud manufacturing system", *J. Commun. Comput. Inform. Sci.*, vol. 202, pp. 225-233, 2011.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014