Research of cloud computing based agriculture virtualized information database

Zhang Haihui, Zhao Chunjiang, Wu Huarui, Yang Feng, Sun Xiang
National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
E-mail: zhanghh@nercita.org.cn Tel: +86-010-51503619

***Abstract*—Agriculture information resources such as multimedia files, remote sensing images and monitoring data stream, are characterized with distributed and massive storage. This paper presents virtualized information database which organize heterogeneous resources distributed in multiple storage nodes based on open source cloud computing platform. With three virtualized database access interface, the transparent user resource access had been realized by XML metadata service, resource information service and resource server monitoring service. The database is scalable and high fault tolerance, which can automatically handle failed storage node.**

I. Introduction

With the development of modern agriculture techniques and the generation and utilization of massive agriculture information resources, the management and sharing of the distributed, massive and heterogeneous resources has become the urging problems. Taking remote sensing image data as an example, which has been widely used in disease and pest control, yield forecast, crop quality analysis etc, the dynamic monitoring of the crop growth has been realized. However, as the large quantity of remote sensing data which is more than hundreds of GB everyday, and as the image resolution increases geometrically [1]. Different agriculture applications can not reach the unified management and sharing of the agriculture data, farmland monitoring data, climate data and soil investigation data which have distributed in isolated organizations.

Grid technology is an attractive distributed computing paradigm, which mainly focuses on resource sharing and coordinated problem solving in dynamic, multi-institutional virtual organization over the network [2]. Typical grid application like Globus[3]、DataGrid[4]、ChinaGrid[5] etc, uses service to virtualized the heterogeneous resources, collects and manages information through hierarchical control system[6]. The present grid uses GridFTP as the underlying transport protocol, which is deficient in large transparent information segmentation, multi-copy distribution and storage. Cloud computing is come forward with evolvement of parallel computing, distributed computing and grid computing which features for virtualization, highly reliable, commonality and serve to the demand[7,8]. At present, Google、IBM、Yahoo、Amazon、Microsoft have conducted extensive research and have come out with each solutions[9,10]. Hadoop, as an open source framework, is suitable for the distributed data storage and management in cheap computers [11].

On the basis of the study on open source Hadoop, this paper developed virtualized information database (VIDB) using cloud storage. Through the distributed storage framework of Hadoop, the information resources of several agriculture research institutes across the nation had been integrated. The transparent user database access had been realized by XML metadata management, resource attribute information management and users' data access interface.

II. The Architecture of VIDB

The design target of VIDB is to realize the unified organization and management of massive, distributed and heterogeneous agriculture information resource. Fig.1 shows the architecture of the database. It is composed by agriculture resources, physical storage layer, logical storage layer and database access layer.

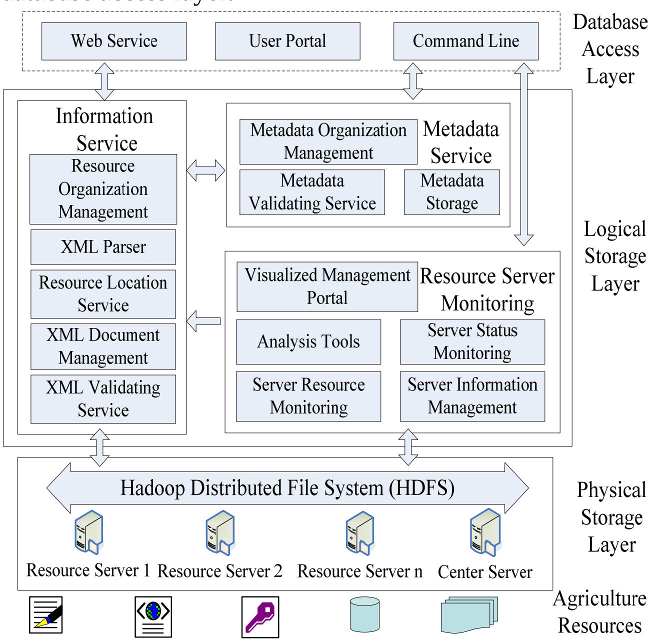

Fig. 1 Architecture of VIDB

The physical storage layer realizes the distributed storage of massive information resources. The cloud storage platform is constituted by the central server in Beijing and 32 resource

Proceedings of the Second APSIPA Annual Summit and Conference, pages 835–838,
Biopolis, Singapore, 14-17 December 2010.

servers in 7 provinces and municipalities. The cloud storage platform provides interfaces like resource access, platform management and status monitor etc. After submitted to the cloud platform, various information resources like remote sensing images, videos, text can be stored separately in resource servers.

Logical storage layer realizes the unified management of agriculture resource descriptive information, establishes the link between users' transparent access and the physical storage. Information service realizes the management of underlying servers' registration information, legal validation, status monitor and visualization. Metadata service maintains the unified resource descriptive limitation, realizes the consistent translation of the grammar and semantic of the shared information between different divisions and departments. Resource server monitoring realizes the category of agriculture information, validation and management of resource descriptive information, resource retrieval etc. The categorization and description uses XML, and realizes the distributed storage of resource through calling storage interfaces.

The access layer adapts web service, user portal or command line to provide location transparency access to agriculture information. Normal user can register, retrieve and access to the agriculture information. The admin user can dynamically monitor the node of resource server monitoring, thus real-time control the database system.

III. THE PHYSICAL STORAGE OF RESOURCE BASED ON CLOUD STORAGE

Cloud storage is a kind of virtualized resource storage pool. Combined with one or more data center's software and hardware, it has the ability of dynamic allocation, smooth extension storage and communication. Cloud storage deals with the integration of resources distributed separately and forms a virtualized storage space providing to every user through unified interface.

There are big files in massive agriculture information resource. Take remote sensing image for example, with the size commonly above 1G, it could cause overload to individual server node, so that the responds to the requirements wouldn't be in time. In this paper, a Hadoop Distributed File System was established with characteristic of scalable and high fault tolerance, which can automatically handle failed storage node. The original big file was break down into blocks, and the storage management strategy of the block was designed to strengthen the parallel dataset read and write ability and facilitate the data compression. Based on the profile of "once storage, multi-read" of such large data resources like remote sensing, the consistency validation and parallel data access control were simplified thus improved the service efficiency.

There are two ways to break down a big file into distributed file system in a long term, one is to save each block as a new file and take the filename as the index key name, the other one is to save the blocks of the big file as a whole, separately store the blocks in different distributed storage node through the scheduling strategy of Hadoop and set the according storage copy rate. This strategy needs additional definition and index key name. When the dataset is large enough, the former method will generate large amount of small files and increase the load of the file system but the later method won't.

To each big file, every block data is recorded in <Blk_ID，MetaData> form. The Blk_ID is the serial number of the blocks and the MetaData is a binary data of the block. Thus the record data type is <int，byte[]>. Given a certain Blk_ID, the byte data of the block can be determined. Figure 2 shows the storage method of big files.

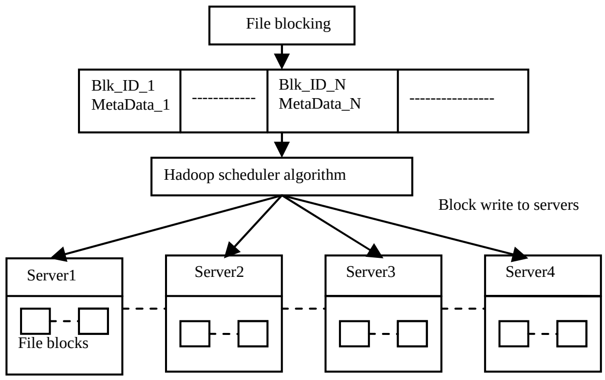

Fig.2 File block storage

Every file is saved in block sequence. Except for the last block, every block is the same size. The block size and the replication factor can be set. The resource in storage cloud is saved in different resource node. The high efficient data transmission is achieved by distributed storage scheduling algorithm. The aim of this algorithm is to homogeneously distribute the copies of the block in each resource node, and maximize the parallel data transmission, fully take advantages of the upload and download band width between data nodes to avoid the data bottle neck. The dynamic priority strategy, minimum number of block copies strategy, random block chooses strategy can be choose based on demand.

IV. RESOURCE LOGICAL STORAGE BASED ON XML

The realization of the metadata resource service complies with XML schema based metadata expression standard and the schema file organization accordingly. The data is categorized and the dynamic management of metadata classification, registration and retrieval functions is achieved. The main functional modules include the metadata layered organizational management, metadata storage, resource information validation etc.

Every resource metadata is assigned with global unique hierarchical category. Metadata Organization Management (MOM) manages pairs of <Category, Schema>, this ensures unique agriculture information representation, which is the precondition of global resource sharing.

Hierarchical category system improves platform extensibility and maintainability and represents logical relationship between resource schemas. Application domain experts define category hierarchy with domain knowledge. A well defined category system design is important to system extensibility. An example of schema block is list as fig3. Every node element has four attributes: Text, ID, FID, and MaxRe. The root's MaxID attribute represents the largest number of the category. Dom4j is used to parse and manage XML document.

```
<?xml version="1.0" encoding="gb2312" ?>
- <root MaxID="37">
   <node Text="data" ID="0-1" FID="0-0" MaxRe="0" />
   <node Text="database" ID="0-1-16" FID="0-1"
MaxRe="4" />
   <node Text="document database" ID="0-1-16-2" FID="0-
1-16" MaxRe="1" />
   <node Text="relational database" ID="0-1-16-3" FID="0-
1-16" MaxRe="0" />
   <node Text="other database" ID="0-1-16-4" FID="0-1-16"
MaxRe="0" />
   <node Text="text file" ID="0-1-5" FID="0-1"
MaxRe="26" />
   <node Text="structure text" ID="0-1-5-17" FID="0-1-5"
MaxRe="2" />
```

Fig.3 Example of hierarchical metadata category system

When users add agriculture information, whose schema has been registered, the category should be specified. Information Service would validate the resource information with schema according to the category. Before a new type of grid resource is added to the system, the schema should be registered and a category ID will be returned. MOM supports run-time adding and updating schemas that endow users with ability to extend information model flexibly to meet the needs of various agriculture applications.

The follow diagram is the resource processing flow chart of the resource information management platform. Each metadata class corresponds with a metadata description. The information is stored by XML schema files. In a certain classification, whenever a user upload a resource, he needs to fill in the required resource descriptive field, so as to generate a XML file in the catalog, and write the field information into the main server database, the download address is written into the database at the same time.

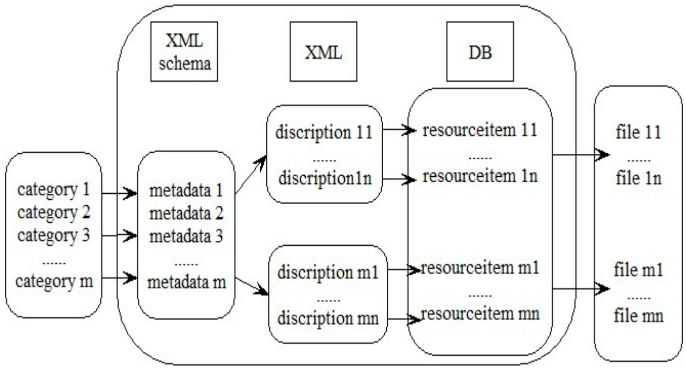

Fig.4 Resource processing flow chart

The resource registered is keyword labeled by resource labeling management system with agricultural concept database, thus each resource has some keyword labels, forming the basis of concept processing. Resource retrieval service uses keywords labels to search the information. When provided with one or more concept, the system will find the information and return it to the upper layer agriculture applied system. Agriculture domain knowledge structure is used to organize the resource information. As Figure 5 shows, FID corresponds with different agriculture domain category, through the check up of FID, the architecture of different resource can be determined promptly, thus increase the validity of search results.

```
<?xml version="1.0" encoding="gb2312" ?>
- <root MaxID="75">
  <node Text="AgricultureDomain" ID="0-1" FID="0-0"
MaxRe="0" />
  <node Text="Food" ID="0-1-2" FID="0-1" MaxRe="0" />
  <node Text="Fruit" ID="0-1-3" FID="0-1" MaxRe="0" />
  <node Text="Vegetable" ID="0-1-4" FID="0-1"
MaxRe="0" />
... ...
  <node Text="AgricultureNature" ID="0-11" FID="0-0"
MaxRe="0" />
  <node Text="Soil" ID="0-11-12" FID="0-11" MaxRe="0"
/>
  <node Text="Weather" ID="0-11-13" FID="0-11"
MaxRe="0" />
  <node Text="Hydrology" ID="0-11-14" FID="0-11"
MaxRe="0" />
```

Fig.5 Example of hierarchical resource information organization

Node server management includes node information management and node resource management. A distributed node dynamic monitoring platform based on Hadoop cloud storage was developed by open source software Ganglia. It can achieve dynamic monitoring on child node of underlying physical storage platform, including the hardware system resource of each distributed node server, data resource, service resource and knowledge resource. Ganglia monitoring software mainly inspect on system performance like cpu, memory, hard disk size and use ratio, I/O load, network flow. With the curves, the performance of each node is clearly presented, which contribute a lot to the system configuration adjustment and distribution of system resource thus improve the overall performance. Fig.6 shows the performance monitoring of resource node.

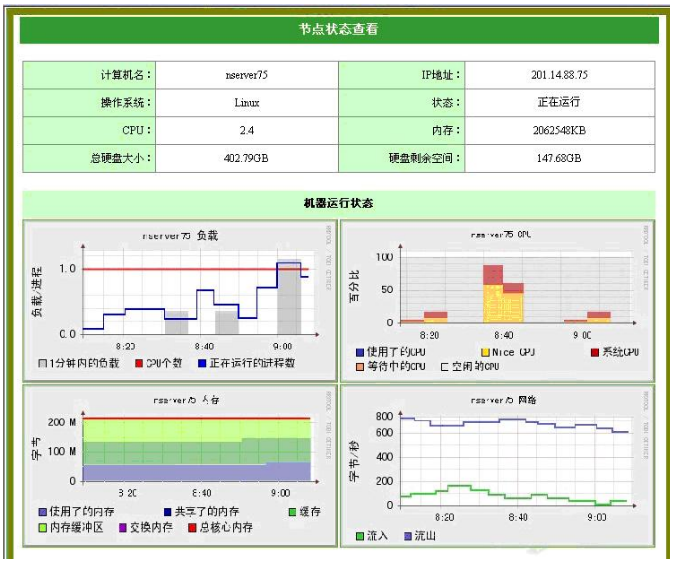

Fig.6 performance monitoring of resource server

V. THE REALIZATION OF VIDB AND TEST

The VIDB operating platform is constituted by the central server in Beijing and 32 resource servers in 7 provinces and municipalities. The hardware configuration of each resource server is as follows: CPU Pentium 2.2GHz; RAM 2GB; hard disk 400GB SATA-II 7200RPM; Ethernet 100Mbps; operating system Redhat 9.0; local file system Ext3. The time consuming on reading large files from virtualized database of different number of users was tested. Fig. 7 shows the relation between the download rate and the number of Client.

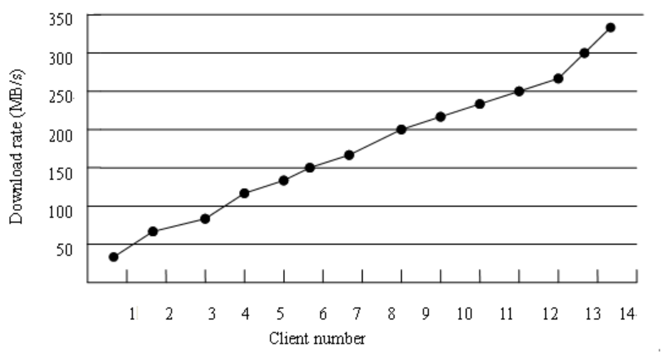

Fig.7 Platform client-side file read test result

VI. CONCLUSION

At present, the agriculture virtualized information database has been developed and implemented in agriculture research institutions in Beijing, Tianjin, Hebei, Shaanxi. It is evident that the database can efficiently solve the problem of the blocking of massive agriculture information, muti-copy storage and concurrency transmission etc, providing services to the multi-institutional organization, management, sharing and transmission of agriculture resource information. The exsising problems of VIDB lie in: (1) the WAN band width of each research institution is different. So single server storage volume and bandwidth can not reflect the distributed service quality, the assessment model of service quality is needed as well as the real service ability by dynamic monitoring of each resource server. (2) the agriculture resource security in each research institution should be considered.

REFERENCES

[1] Jianting Shi, Yinan Chen, Chunyuan Liu. Database system for archiving and managing remote sensing images. First International Workshop on Database Technology and Applications. 2009: 536-538.
[2] I. Foster, C. Kesselman, S. Tuecke, The anatomy of the grid enabling scalable virtual organizations, Int. J. Supercomput. Appl. 2001, 15 (3): 115-128.
[3] The Globus alliance. Globus Toolkit 4.0. Available at http://www.globus.org/toolkit/docs/4.0/info.
[4] Groep, D., J. Templon,C. Loomis. Crunching real data on the Grid: practice and experience with the European DataGrid. Concurrency and Computation: Practice and Experience. 2006. 18(9): 925-940.
[5] Wu, S., W. Wang, M. Xiong, et al.. Data management services in ChinaGrid for data mining applications. Proceedings of the 2007 international conference on Emerging technologies in knowledge discovery and data mining, Springer-Verlag. 2007: 421-432.
[6] Zanikolas, S.,R. Sakellariou. A taxonomy of grid monitoring systems. Future Generation Computer Systems. 2005. 21(1): 163-188.
[7] Armbrust, M., A. Fox, R. Griffith, et al.. Above the clouds: A berkeley view of cloud computing. EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28. 2009.
[8] Buyya, R., C.S. Yeo, S. Venugopal, et al.. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems. 2009. 25(6): 599-616.
[9] Wang, L., G. von Laszewski, A. Younge, et al.. Cloud Computing: a Perspective Study. New Generation Computing. 28(2): p. 137-146.
[10] Hakan Erdogmus. Cloud Computing: Does Nirvana Hide behind the Nebula? IEEE Software, 2009, 26(2): 4-6.
[11] Borthakur, D. The hadoop distributed file system: Architecture and design. Hadoop Project Website. 2007. http://hadoop.apache.org/common/docs/r0.18.0/hdfs_design.pdf