The digital library cloud storage based on Hadoop

SHEN GE[1st]
Aviation University of Air Force
Changchun, China
e-mail: chuanshuo@vip.qq.com

HONGBO WANG [2nd] YUFEI WANG [3rd], CHEN LU [4th]
Aviation University of Air Force
Changchun, China
e-mail: chuanshuo@vip.qq.com

Abstract: This paper studies analyzes the relevant theories of library digital resource storage for its needs and characteristics, and constructs the digital model based on a"cloud storage" library, and makes an experiment and demonstration of the library "cloud storage" architecture model , at last deals with the library mass storage under the cloud storage and the solution to a large amount of data transmission problem.

Keywords: digital library;cloud storage; Hadoop

I. INTRODUCTION

With the rapid development of digital libraries, there are more and more video and native literature database, making the storage system more and more complicated[1]. Digital libraries need to build a good mass storage system with a high performance, a large capacity and good expansion to meet the requirements for all kinds of needs for security, scalability and I/O performance, but there is a huge challenge for building the mass storage system with a large amount of technology and cost.[2]

II. TRADITIONAL STORAGE SHORTCOMINGS OF THE DIGITAL LIBRARY

At present, the main storage technology used in digital library is the direct attached storage (DAS), network-attached storage (NAS) and storage area network (SAN)[3]. Under the condition of resource growth, only by increasing the hardware equipment can we really meet the demand of storage, at the same time a lot of problems will occur:

A. Weak disaster recovery

It is based primarily on whether the server uses the redundancy, online repair and cluster and other technologies. The use of DAS, NAS or SAN storage technology to build storage facilities will enable the information resource storage generally use a single array or server because of the limitation of structure or price. Once an array or a server fails to work, the user will not be able to store resources, also doesn't have access to resources. In addition, these storage facilities are usually concentrated in one room, once the room was damaged, the data will be lost.

B. Low data processing efficiency

Usually, the information resource is stored by a single point. When a large number of users have access to the data at the same time, it is easy to cause the failure of the storage device, affecting the performance of the system, such as database retrieval, document download, data analysis of readers and video database have the very strong abruptness, and high-performance storage platform system can efficiently handle the request.

C. Poor extensibility and high cost

In either way to store, data storage must use professional storage devices, because the professional storage equipment is expensive, making the cost of the resource storage increase. In addition, the storage capacity of each storage equipment is limited, when the storage capacity is insufficient, we will have to buy a new storage device. This greatly increases the cost of resource storage, not conducive to the integration and sharing of digital library information resources.

III. CLOUD STORAGE BASED ON HADOOP DISTRIBUTED FILE SYSTEM

HDFS is the Hadoop Distributed File System[4][5], the System adopts the master/slave structure, and this structure has the very high fault tolerance, and can be deployed on the low-cost hardware devices, which is suitable for the application for large data sets again at the same time, providing
the high throughput of data read-write.

Compared with the traditional storage devices, the structure model of cloud storage system is a complex system composed of not only a hardware, but a network device, storage device, server, application device, public access interface, access, and the client program and other parts. The parts is applied with a storage device as the core by applying the software to provide access to data storage and business services, such as table Ⅰ.

TABLE I. DIGITAL LIBRARY "CLOUD STORAGE" SYSTEM ARCHITECTURE

Data storage layer	Hadoop (storage virtualization,centralized management, state monitor,maintain promotion)
	Storage device（iSCSI、FC、NAS）
Base manageme nt layer	Distributed HDFS file system (content distribution, the data part, data disaster recovery) Distributed HBase storage system (real-time invitation)
Applicatio n interface layer	Network interface, user authentication, rights management
	Storage platform of library resources , a variety of Web services and public API interface in the library
Invitation layer	All kinds of storage applications of library users

© 2013. The authors - Published by Atlantis Press

IV. Digital library cloud model and function realization

Hadoop and HBase Hadoop cloud platform system are built based on digital library to realize the mass storage for library resources[5], HDFS is responsible for the actual data storage, HBase is used to improve the rate of data transmission[6], and the role of the Hadoop is virtual and it manages the equipment of each storage node, digital library cloud storage as shown in figure 1.

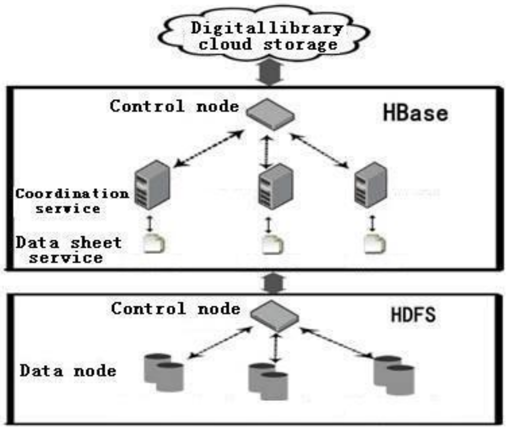

Figure 1 Digital library cloud model based on Hadoop

A. Function implementation

In this paper, the Hadoop distributed file system is used for the concrete introduction of file storage process under the cloud environment and the output class diagram and the methods, which focuses on the inherited rewritten Hadoop Cloud File system classes[7][8]: because this class is internal interacted with API interface with a distributed system, and then it makes the analysis of the performance of the system.

B. Output class for cloud file

Starting from the architecture model of the current design, it rewrites the File System class and forms the new file system. Because the first programming language in a cloud environment is only Java, so in this paper, the algorithm implementation uses Java language to describe, the task on the early stage is to find specific files in a cloud environment, making Cloud File System mainly realized by API in the following two ways:

Protect the static Cloud File System Get File (Configuration conInstance) throws IOException

Protect the static Cloud File System Get File (WEBURL WEBURL, Configuration conInstance) throws IOException

In this paper, a Configuration is used to store the configuration storage on the client and the server side, in which the settings comes from the Configuration file on the path. Among them, the first function is used to extract the Settings in cloud file system , if there is no configuration, then the default configuration scheme is adopted and the second function is to use the cloud storage file permissions set under the WEBURL , if there is no set in advance, then, the default configuration scheme will be used.

When a specific file in a cloud environment is got, Open Files () is used to obtain the file stream:

Public Cloud Data Input Stream open Path (f) throws IOException

Public abstract Cloud Data Input Stream open (Path Path, int buffer Size) throws IOExceptio

The first function of buffering the file size is 4 KB by default.

Cloud File System is used to operate files in the cloud environment, and the unified output is made with a standard format. and the flow is shown in figure 2.

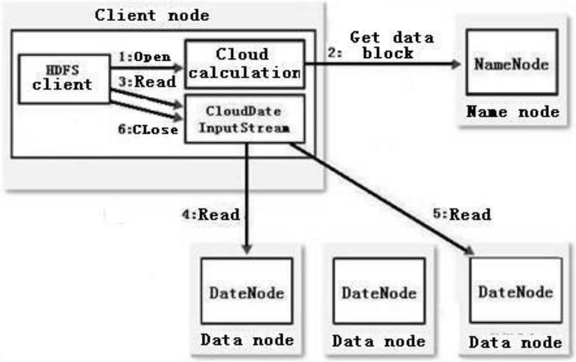

Figure 2 Flow chart of output class

C. Storage class of cloud file

In a cloud environment, the documents are created, create (Path Path) method in Cloud File System is mainly used , and the function prototype is:

Public Cloud Data Output Stream create Path (Path Path) throws IOException

Important attributes about the files (for example: copy number block, buffer capacity, capacity, file access permission, whether i tcan cover or not, etc.) can be set. It is common in one type: a file depending on the directory may not be in the default, so you need to determine beforehand the existence in directory (using function exists), and then create () is used to set.

There is also an important method of Cloud Progressable, mainly it is to provide the data storage progress of each data node for specific instances.

Public interface Cloud Progressable {
Public void Cloud progress ();
}

If the file has been created successfully and want to add content,

Public Cloud Data Output Stream append (Path Path) throws IOException

should be used.

The writing of data is at the end of an existing File, as a result, borderless File will appear, take one of the most common in the daily log as an example, this File is added content constantly and updated again and again at the end.

The design of local file, file transfer and progress in the storage center are as follows, and the flow is shown in figure 3.

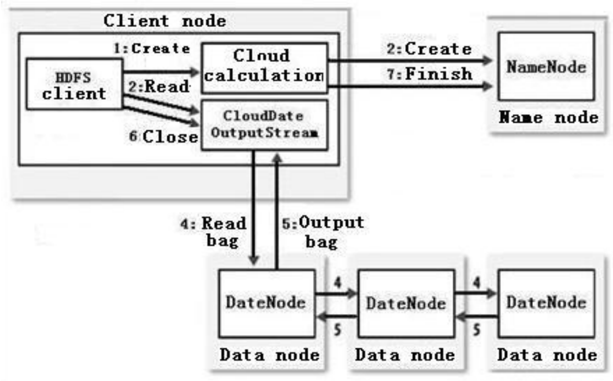

Figure 3 Flow chart of storage class

V. THE TEST SCHEME

Single node patterns: general test system files, not including the cloud file system.

Cloud structure test: It is tested under the cloud storage architecture, and the read speed comes to an average of about 2 MB.

The data amount is made a efficiency comparison of that increases from 1 MB to 1GB, at the same time, the result records of running time are shown in tableⅡ.

TABLE II. TABLE 4-1 TEST PATTERN

FILE SIZE /MB	Test patten /10^4ms	
	SINGLE NODE MODEL	*CLOUD STORAGE MODEL*
1	05 234	46 950
4	08 874	56 939
16	14 621	153 741
64	45324	547 924
256	215 874	2 384 100
1024	837 141	11 960 675

Because of the occurrence of Hadoop architecture calculation model, it has broken the speed limit from traditional database system to mass data processing by concurrent access to a large amount of data, which effectively reduces the query computing time of data, making the user's query response more quick. In this paper, by comparing the operating speed between stand-alone mode and the cloud computing model, it is concluded that under the cloud storage model, the operation efficiency is several times higher than that of single mode. Hadoop architecture model can make full use of different machines at the same time and participate in the concurrent operation and improve the operation efficiency. In this paper, it will continue to make the centralized research for the scheme of distributed arithmetic algorithm under the Hadoop framework and according to the characteristics of it and apply it to specific scenes.

The simulation experiment of this article is in view of the architecture model of the input and storage capacity, and it mainly aimed at all kinds of files, the files shall be carried out in accordance with the size based on the order of the input and storage. The experiment is divided into two types, the first type is under the environment of the parallel test and the composition of only one file. The second type is under the environment of parallel and file structure is composed of two files with the same size. The test result is shown in figure 4.

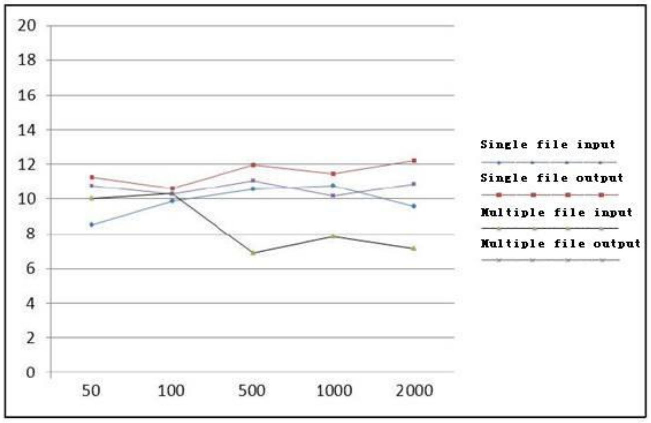

Figure 4 File read-write speed

It is clearly seen from the picture, when a file is gradually increasing, the read-write rate is a linearly growing, which also has a lot to do with the file input, storage rate and the rotate speed of the hard disk as well as the network environment , this test was conducted under the old hard disk with slow rotational speed, and then within the LAN network, the stability is bad, if it is switched to the new hard disk, the result will be improved greatly.

VI. SUMMARY

This paper analyzes the construction of digital library "cloud storage" model, sets up a test environment and makes a transmission performance test of the distributed cloud storage mode, showing that the cloud storage model is advanced compared with the traditional mode of transmission speed. Hopefully, through the research in this article, it will further speed up the construction of digital library in our country, solve the problem of mass storage requirements for cloud library, reduce the operating costs of the library, and at last improve the information service level of the digital library to the greatest extent and service quality.

REFERENCE

[1] cloud storage [EB] . http : // baike. baidu. com/ view/ 2044736. htm ,2010.

[2] John Gantz, David Reinsel. The Digital Universe Decade –Are You Ready?[R].EMC sponsored IDC survey. 2010. 5: 2-4.

[3] Saracevic, Tefko. 2010.Digital library evaluation: Towardan evolution of concepts [J]. Library Trends 49 (2): 350-369.

[4] Apache Hadoop[EB/OL]. (2010-03-25). http://hadoop.apache.org.

[5] Intel.Optimizing.Hadoop*Deployments[EB/OL].(2010-05-23).http://communities.intel.com/servlet/JiveServlet/downloadBody/5645-102-1-8759/Optimizing%20Hadoop_2010_final.pdf.

[6] Xie Jiong, Yin Shu, Ruan Xiaojun, et al. Improving MapReduce Performance Through Data Placement in Heterogeneous Hadoop Clusters[C]//Proc. of IPDPSW'10. Atlanta, USA: [s. n.], 2010.

[7] Polo J, Nadal D, Carrera D, et al. Adaptive Task Scheduling for Multijob Mapreduce Environments[EB/OL].(2010-03-25).http://adaptive-task-scheduling-for-multijob-mapreduce-environments.

[8] Kambatla K,Pathak A,Pucha H.Towards Optimizing Hadoop Provisioning in the Cloud[C]//Proc. of the 1st ACM Symposium on Cloud Computing. New York, USA: ACM Press, 2010: 137- 142.