2014

ISSN : 0974 - 7435 Volume 10 Issue 13

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(13), 2014 [7369-7375]

A study on mass data storage technology based on cloud computing

Lingxing Yang
Qujing Normal University, Qujing Yunnan 655011, (CHINA)

ABSTRACT

In recent years, because of the characteristics of high security, high efficiency and low cost, cloud computing gradually applied to various fields, causing huge benefits. With the constant development and ripeness of cloud computing technology, application of this technology to store mass data, reduce the storage and maintain the costs become a research focus in this field. This paper studies the distributed storage technology under the cloud computing environment and the methods to organize and manage the mass data. Firstly the paper introduces some basic concepts in the field of the current cloud computing, and then probes into mass data storage model and finally points out the problems that exist with the current technology.

KEYWORDS

Cloud computing; Mass data; Store; Distributed storage.

INTRODUCTION

As with the rapid development of economy, science and technology, the amount of information shows explosive growth. Social networking sites, mobile devices, astronomy and high energy physics constantly generate data every day. Only in 2011, the global data reached to the $1.8*10^{12}$ GB. According to estimation, the amount of global electronic form data is expected to be 35 zb in 2020. So there is a problem to be solved: with the more and more data, the traditional storage way shows more and more flaws, such as poor scalability, distributed storage, poor storage performance, failure in integrating different systems, and the higher and higher cost. The storage technology of cloud computing came under the backdrop of broadband, mobile internet, the internet of things, social networks and cloud computing. Through clustered application, grid technology, distributed file systems and other functions, a large number of various types of storage devices set up to work together. Common provide external data storage can greatly reduce dependence on its own hardware resources. [2]Therefore, the application of cloud computing technology in mass data storage field can greatly improve the utilization rate of resources. Now several major suppliers of cloud computing platform are as follows: the Google APP engine[3], the AT&T Synaptic Hosting [4], AppNexus [5], etc.

In cloud computing environment, mass data is stored in the same data center, but on different nodes. The location of the data and the organization is known by the user. The user only needs to store data through a simple interface to the data center. This study mainly introduces the structure model based on cloud storage system, the Map Reduce model,[6]the implementation of mass data storage technology of the Hadoop open source[7]distributed computing framework.

CLOUD COMPUTING AND CLOUD STORAGE

The concept of cloud computing

The increase in Internet related services, use and delivery mode, usually involves using the Internet to provide dynamically scalable and often virtualized resources. Cloud computing[8] is the product of traditional computer, such as distributed computing, parallel computing virtualization, network storage, and the development of network technology.

The concept of cloud storage

Cloud storage is a new concept based on cloud computing. For the consumer, the cloud storage is a data access service; the core is the combination of application software and storage devices. Behind the core is the cloud computing system configured with a large capacity storage space, that is, a storage layer is added on cloud computing, and at the same time, increase related functions such as data management. However, the user access and application interface is the same as the cloud computing.

BASIC INTRODUCTIONS

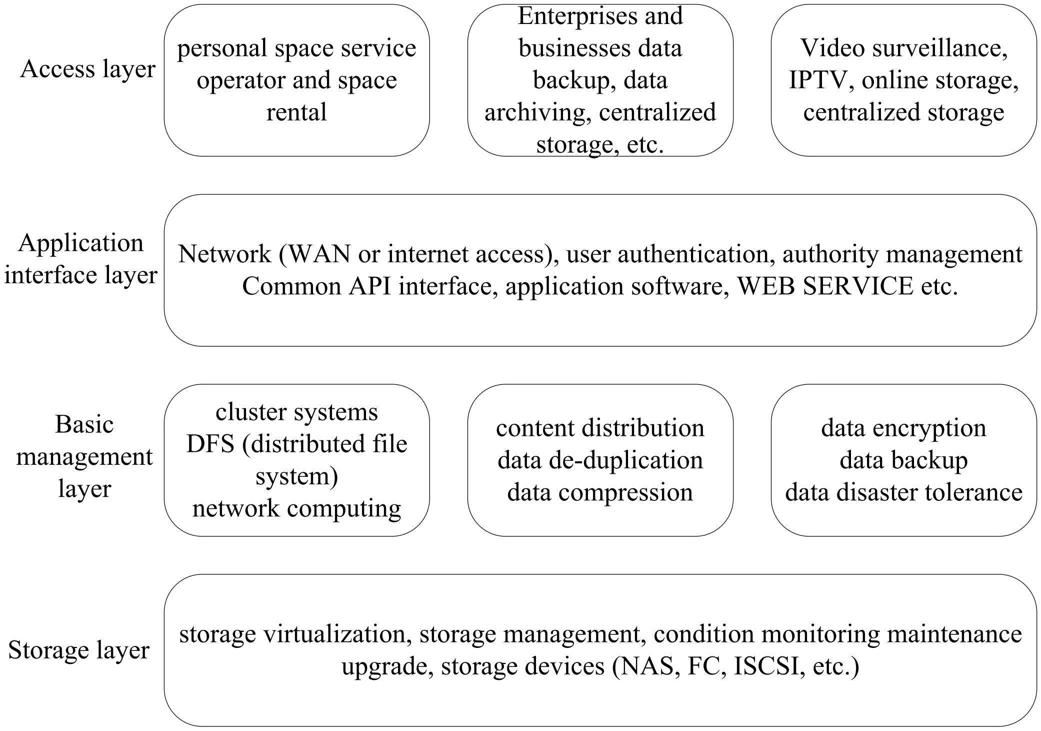
Access layer
personal space service operator and space rental
Enterprises and businesses data backup, data archiving, centralized storage, etc.
Video surveillance, IPTV, online storage, centralized storage
Application interface layer
Network (WAN or internet access), user authentication, authority management Common API interface, application software, WEB SERVICE etc.
Basic management layer
cluster systems DFS (distributed file system) network computing
content distribution data de-duplication data compression
data encryption data backup data disaster tolerance
Storage layer
storage virtualization, storage management, condition monitoring maintenance upgrade, storage devices (NAS, FC, ISCSI, etc.)

Figure 1 : The structural model of cloud storage

Mass data brings a big challenge to the storage architecture. First of all, mass data must be organized via distributed data organization and management strategy. Secondly, mass data keeps sustained growth and accumulates gradually. This process generally takes a long time, so the storage organization mode and indexing mechanism must guarantee the size and performance extension at the same time. Next is the introduction of the basic structure model and distributed storage model.

The structure model

There are four layers of the structure model of cloud storage system: the storage layer, the basic management layer, the application interface layer and the access layer. The structural model is shown in Figure 1.

Storage layer is the most basic part of cloud storage, the storage devices it used are not only in large quantity but also in wide distribution. The storage devices can be IP storage devices, DAS storage devices, the FC fiber channel storage devices. They can connect each other by WAN, Hu Liang net or FC fiber channel devices, and then through the unified storage management system, monitor equipment state and perform maintenance.

The basic management layer is not only the core part of the cloud storage, but also the toughest part. Through cluster, distributed file system and network computing technology, collective work among multiple storage devices can be achieved so as to make them provide powerful data access functions.

The application interface layer is the most flexible part. According to realistic business types, different cloud storage operation units can develop various application service interfaces and provide different services.

The access layer: Any authorized user can log into storage system via common application interfaces and enjoy the cloud storage service. However, different operating units also have different types of access and means.

Map Reduce mode

Map Reduce mode, the heart of the cloud computing and computational model, [9]is a kind of distributed computing technology. It is used to solve the problematic program development model and process or produce huge amounts of data sets.

The main idea of Map Reduce model is automatically to decompose problems into the perform Map and Reduce. The user can specify a map function to process key/value pairs, which will produce a series of intermediate key/value pairs, and then reduce function is used to merge all the same key value part of the intermediate key/value pairs. The specific flow chart is shown in Figure 2.

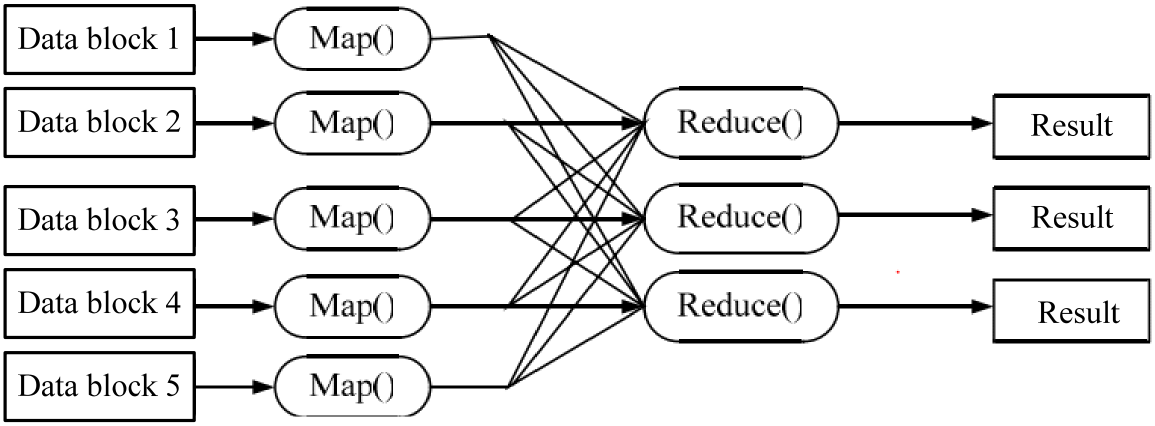

Figure 2 : Map Reduce processing flow

The prototype of the map reduce process is as follows.

Map：(Initial Key, Initial Value)→[(Inter Key, Inter Value)]

Reduce (Inter Key, Inter Values Iterator) →[(Inter Key, Inter Value)]

The map function is the list of some concepts of independent elements. It is operated separately for each element. And the original data has not been modified, in other words, the Map can be highly parallel operation. Reduce operation is a proper merger of a list of elements. Although not as good as the map function in parallel, it always has a simple answer; and its large-scale computing is relatively independent, so the reduce function is also useful in a highly parallel environment. The map and reduce functions are known as higher-order functions, because they can put the other function as their parameters, so that they can be combined very well with other functions.

Map Reduce distributed processing can deal with all kinds of query, greatly improving the efficiency of query, and at the same time releasing the data stored on disk. Except the difference shown in TABLE1, the distinct difference between the relational database and Map Reduce is the quantity of structured data in their operation data sets.

TABLE 1 : The comparison between the relational database and map reduce

	Traditional relational database	Map Reduce
Data size	GB	PB
Visit	Interactive and batch	Batch
Update	Multiple read and write	Write once read many
Structure	Static mode	Dynamic mode
Integration level	High	Low
Elasticity	nonlinear	linear

Hadoop framework

Hadoop, an open source distributed computing framework developed by Apache, shows excellent performance in distributed computing and data storage. The biggest characteristic is to run the application by running mass cheap hardware facilities and provide a set of stable and reliable interface so as to build a high reliability of distributed system. [10]Hadoop can provide the HDFS and HBase, so it can also store data or deploy on each compute node.

Hadoop framework technology is shown in Figure 3

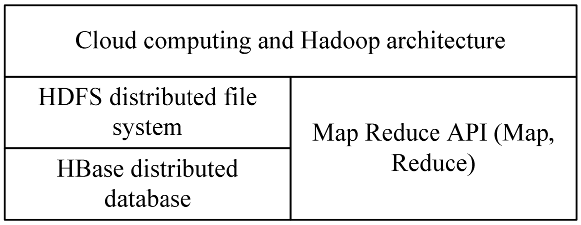
Cloud computing and Hadoop architecture
HDFS distributed file system
HBase distributed database
Map Reduce API (Map, Reduce)

Figure 3 : Hadoop frame structure

HDFS is Hadoop distributed file system. With powerful data storage capacity, it is suitable for mass data storage systems. In the interior of the Hadoop framework, the HDFS and HBase are used to store or deploy on each compute node, and eventually the Map Reduce mode are used to deal with the data within the Hadoop framework. With the help of the Hadoop framework and map reduce, data computing and storage can be implemented; cloud computing can process distributed, parallel computing and storage and the cloud computing platform possesses the ability of dealing with mass data.

HDFS system architecture, a kind of typical client-server architecture, is composed of a control node and multiple data nodes. As shown in Figure 4, the system architecture of HDFS uses control node design, which greatly simplifies the structure of the file system. Although control node maintains the namespace of the entire file system, the actual data is not stored in the control node, but stored in the data node.

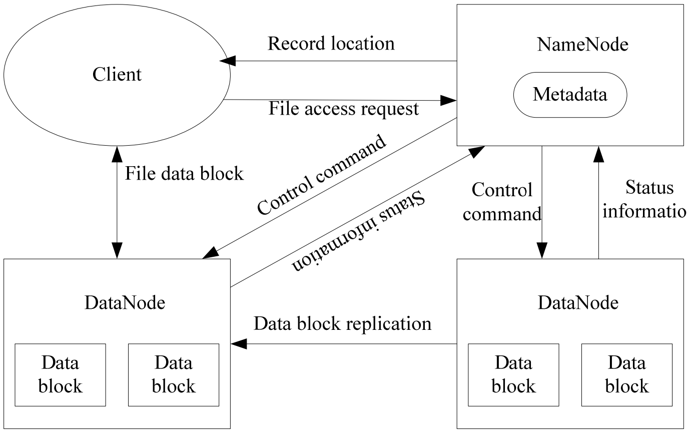

Figure 4 : HDFS system architecture

THE DESIGN OF MASS DATA STORAGE PLATFORM

Mass data storage model

After understanding the Hadoop framework and Map Reduce distributed computing, mass data storage model [11]of cloud computing, which contains Hadoop framework will be shown in Figure 5.

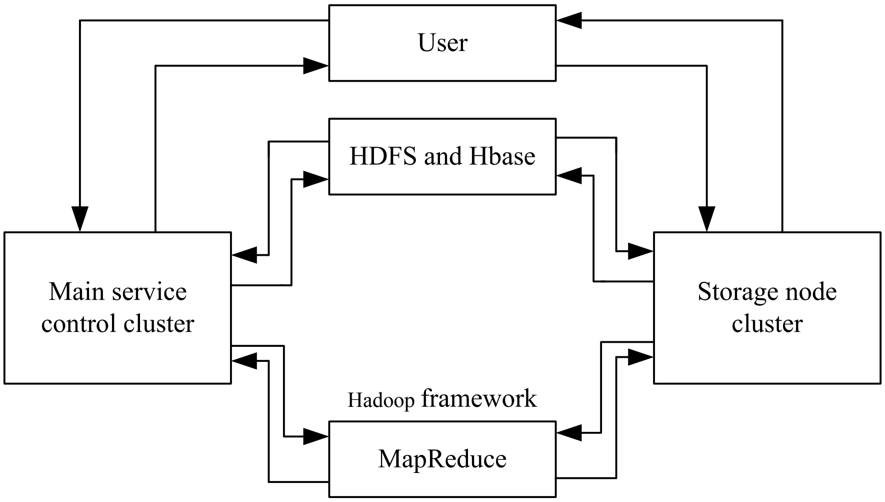

Figure 5 : mass data storage model of cloud computing

From Figure 4 it can be seen that the main service control cluster as the part of controller, is responsible for receiving all kinds of application request and replying according to the request type. The storage node cluster, equivalent to a memory, is a system with a lot of disk array or a cluster system with mass data storage capacity. Its main function is to process the storing and retrieving data resource.

Data will be stored or deployed to the various computing nodes by HDFS and Hbase. Master server of the Hadoop is used to manage and schedule the rest of the computers. The master server can be run on each computer in the cluster while the others are responsible for performing tasks. However, the master server must be run on data storage nodes (compute nodes). The master server assigns the Map tasks and Reduce tasks to spare computers, makes these tasks run in parallel, and monitor the operation of tasks. If any problem should appear on those performing computers, the master server will transfer the task to other free computer to complete.

Mass data read operation can cause system congestion. In order to avoid this phenomenon, cloud computing service provider has carried on the design which will bar the user from reading via the Hadoop framework and accessing data via HDFS and Hbase. After sending message from the Hadoop architecture to the service control machines, a user can perform interactive read operation on the storage nodes.

Importing DFS

Importing the DFS before starting the Hadoop rule calculation, this operation is very simple, only input DFS-put command.

Rule evaluation

Rule the core of Map/Reduce calculation model. Through the above discussion of the Map/Reduce calculation model, it is known that it can efficiently handle mass data information of the text.

Exporting DFS

The result exported via DFS will be stored in the specified NFS or local disk. In order to enjoy an easy data mining after rule evaluation and make user pay more attention to the implied message in the data analysis, the standard output format is generally adopted. Its operation is also simple, using the command: hadoop dfs-get.

Storage solution test

The functions used for mass data storage solution are as follows: the Input Format, reduce, the map and the output Format, etc. Here are a few major functions:

Input Format：The input function cut the text into small pieces and records these main functions.

MAP：This function performs map operations, deals with the strings in small pieces like spaces and generates the < key, value > pair. The key is for the word and the value for the number.

Output Format：This function checks the output directory and then writes the result in related documents.

Setting up a test environment requires: hardware environment: one master, two clients; Software environment: centos5, Hadoop0.21.0, jdk1.6.0_22, VMware workstation and Eclipse3.2.

Test steps:

(1)install and start the SSH service;

(2)configure the SSH service

(3)install JDK1.6；

(4)install Hadoop；

(5)start Hadoop；

(6)install eclipse；

After completing the steps above, environment configuration completed.

Realization of the storage platform

The main function of mass data storage platform consists of the following several modules: user management, folder management and file management. Specific functional division is shown in Figure 6

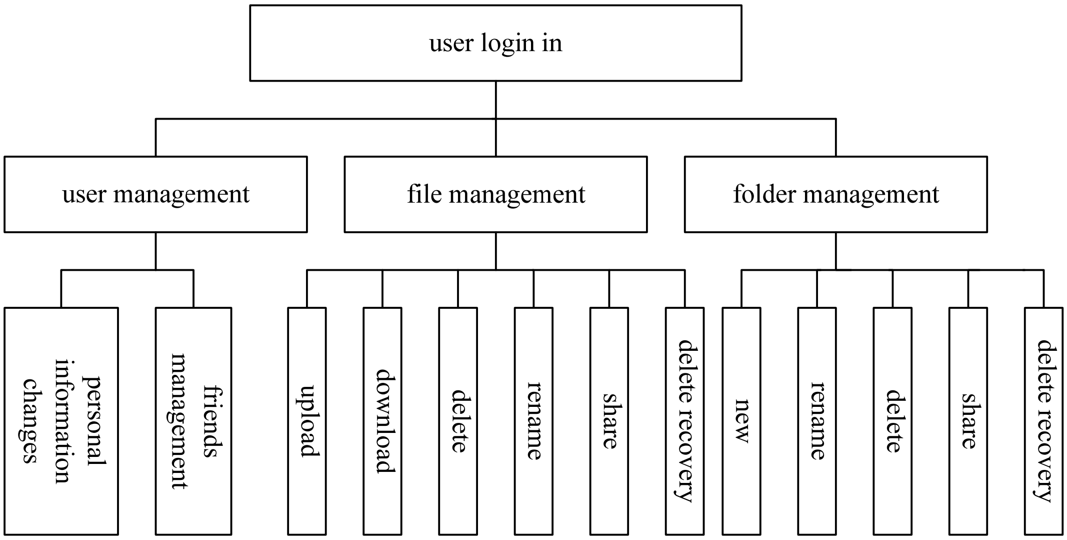

Figure 6 : The platform function list

File upload: wile uploads files to HDFS cluster, the key lies in the background to write code. By using a small file processing module and at the same time calling HDFS API, the copyFromLocalFile in HDFS is used to upload the file to the HDFS cluster according to the given file path.

File download: in addition to download user’s upload file, user can also download shared files via the cloud storage platform. HDFA API system can be used to download large files and realize the process of the download files from the HDFS cluster to local.

File list browse: browsing the list, users can access and retrieve files quickly at the same time. By calling the list status method in the Distributed File System class, the specific list of files information in the directory can be obtained. Other file or folder information can also be seen, such as time, founder, size, etc.

File sharing: invite other users to join the Shared folder by setting the sharing function so as to share files between each other, realize efficient resource sharing and more convenient collaborative work. After setting the Shared folder, program writes the file ID of the folder into a shared folder database table. This process is not a new file creation but a shared index setting up.

PROBLEMS IN CLOUD STORAGE

According to the above content, it is easy to see that the mass data storage technology based on cloud computing has comprehensive application value. However, many technical problems still need to be more widely discussed and processed at the time being.

Security and usability

In most of the cloud storage market, data security is based on trust. Users place their data in the cloud, thus causing the security and usability of the user data depends on cloud storage server providers. Once confidential data leakage, loss and other problems appear, trust between user and supplier will be completely destroyed. This worry is the maximum factor that hampers individuals and businesses from using cloud storage mode.

Technology of WEB2.0

The main function of web technology 2.0 is sharing. Only with this technology can users in the cloud computing store mass data and share information via PC, mobile phones and other mobile devices.

Broadband bottleneck

In the past 20 years, broadband upgrade has been unable to follow the rate of data growth, and the gap becomes bigger and bigger. Probably broadband could become the biggest obstacle for cloud storage commercial popularization.

CONCLUSIONS

The rapid development of Internet makes the advantages of mass data storage technology based on cloud computing become the focus of the industry. This paper discusses mass data storage structure mainly based on the Hadoop distributed platform and map reduce computing mode of the cloud computing. Using the Hadoop distributed platform can effectively improve the data processing speed and provide good solution for mass data processing. While the map reduce provides different descriptions, corresponding modification and corresponding test for different problems, greatly enhances the security and practicability of the system.

Generally speaking, studies in the field of cloud computing is still in its infancy stage; and a lot of problems need to be solved. For instance, aiming at highly complex data structure, the processing capacity of the distributed storage model is not very perfect; complex data structures handle capacity needs to be strengthened in the future. However, for mass data storage based on cloud computing technology, it still has good application value and research prospects.

REFERENCES

[1] Huang Hao; Wave of mass data [Z], Informatization of China, **1**, 34-35 **(2012)**.

[2] Wang Lisha; Data storage technology based on cloud computing [J], Times report, **7**, **(2012)**.

[3] Budavari, T.Szalay; SkyQuery-A prototype distributed qeery web service for the virtual observatory: Proceedings of ADASS Ⅻ, 31, **(2003)**.

[4] chang, F.Dean, J.Ghemawat, et al; Bigtable:A distributed storage system for structured data **(November 2006)**.

[5] Deanand; Mapreduce:Simplified data processing on largeclusters:In Proceedings of OSDI December, 137-150 **(2004)**.

[6] A.Szalay, A.Bunn, J.Gray, et al; The Importance of Data Locality In Distributed Computing Applicationgs[M],[s.l.]: NSF Workflow Workshop, **(2006)**.

[7] Abraham Silberschatz; The concept of database system [M],Beijing: Mechanical Industry Press, 5-8 **(2000)**.

[8] J.Han, H.Pei, Y.Yin; Mining Frequent Patterns without Candidate Generation[M], New York :ACM Press, **200**.

[9] Xu Xiaolong, Wu Jiaxing, et al ; A study on mass data processing system based on the massive link computing platform, [J], Computer application study, **29(2)**, 582-585 **(2012)**.

[10] Chen Yong ; Design and implementation of communication data distributed query and algorithm based on the Hadoop platform [D],Beijing Jiaotong University: **(2009)**.

[11] Zhang Hesheng, Zhang Yi, Hu Dongcheng; Study on mass data management structure and its method, [J],Computer engineering and its applications, **40(11)**, 26-29 **(2004)**.