ISSN : 0974 - 7435 Volume 10 Issue 19

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(19), 2014 [11465-11470]

Research of storage technology for large data volume based on cloud computing

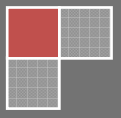
Zhang Zhiqiang, Hu Yibo
Xi'an International University, Xi'an, 710077, (CHINA)

ABSTRACT

Internet is now becoming increasingly popular, and there is more and more data generated on the Internet. With respect to the storage problems of large amount of data, at present, the three most commonly used storage methods are open system's direct storage, network attached storage and storage area network. With the development of network, there is the emergence of a new type of storage technology, which is the storage technology for large data volume based on cloud computing. Technically in cloud computing, the reach is based hadoop and virtualization technology, focusing on the storage technology for large data volume of cloud computing. First, we have a preliminary idea on how to access the large amount of data based on cloud computing. Secondly, we need to select the algorithm and model, and we choose the MapReduce model in this regard, and conduct the corresponding programming based on this model, so as to achieve processing of large data according to the Hadoop framework. Finally, according to the above ideas, we have completed storing a large amount of data based on the cloud computing. At last, we make outlook and summary for the storage technology for large data volume based on cloud computing. As the next generation of computing model, cloud computing is widely used in many areas of scientific computing and commercial computing. But it is still in its infancy, and there are still a lot of unclear problems to be solved in the field of cloud computing and there are still many open questions requiring further research and exploration.

KEYWORDS

Cloud computing; Storage technology for large data volume; Mapreduce algorithm; Virtual resources.



INTRODUCTION

With the progress of time, the computer is becoming increasingly popular, and the development of information technology has also greatly contributed to the social and scientific development and progress, and relatively, various industries are constantly advancing information technology, which gives IT a huge boost, making information technology continue to develop. In this process, as science, business and everyday life continues to deepen its level of information, the amount of data generated is increasingly expanding, especially in high-energy physics, biology, astronomy, weather forecasting and earthquake prediction and other fields that require a lot computing, as well as Web search and social networks, and other business computing fields, where the problem become particularly prominent. Accordingly, the data storage becomes an urgent problem to be solved.

Currently, the three most commonly used storage methods are open system's direct storage, network attached storage and storage area network. When the volume of stored data is very large, these three storage methods can be a perfect solution to storage problems, but with the development of network, the amount of data increases rapidly, their shortcomings have exposed. The method of direct storage has poor scalability, low system performance, and dispersed storage. Network-attached storage is easy to use, has low cost, but its storage performance is the worst. Excellent performance of storage area network storage allows effectiveness of work and data transmission efficiency to increase dramatically, but its architecture is a closed one, which cannot be integrated with different systems and requires large scale and high cost.

In these methods mentioned above, in order to improve the efficiency of information exchange, data is usually locally stored and processed in a concentrated manner. If enterprise establishes its own set of IT system, they not only need to purchase hardware, bandwidth, and other facilities, but need specialized IT staff in charge of IT system maintenance. However, this method is usually not enough. With the development of computer technology, the amount of data that needs to be stored and processed continues to increase, and the enterprise data storage space is increasingly scarce. Enterprises still need to continue to spend a lot of money on the purchase of a variety of data storage and processing equipments, and need to afford increasingly higher data center management costs.

With the development of the network, there is the emergence of a new type of storage technology, which is the storage technology for large data volume based on cloud computing. This technology allows large amounts of data can be stored on a non-local computer or a remote server, enterprise or individual users do not need to spend money in buying expensive hardware devices, instead they just need to buy or lease data storage and computing power through a network, which makes enterprise's access to a computer and storage system according to the needs possible. This research is about a new storage technology for large amounts of data based on cloud computing, and its core calculating method is MapReduce, and it also has a corresponding programming. The storage technologies and cloud computing will be carried out to achieve distribution storage for a large amount of data.

CONCEPT OF CLOUD COMPUTING

Definition of cloud computing has a variety of claims. For what on earth is cloud computing, you can find at least 100 interpretations. Currently the widely accepted definition is that of the U.S. National Institute of Standards and Technology (NIST), i.e. cloud computing is a pay-per-use model that provides available, convenient, on-demand network access into configurable and shared pool of computing resources (resources include networks, servers, storage, applications and services), and these resources can be quickly provided, just require a few management works, or a little interaction with service providers.

Then, data storage based on the cloud computing is as follows: in a cloud computing environment, a large amount of data will be stored on different computing nodes in the same data center or even in on computing nodes in different data centers, but the relative information such as the location and organization method of the stored data is transparent to user, and the user can realize the data storage, organization, management and other activities through some simple excuse. In this process, the reliability and availability etc. of the data are the responsibility of cloud provider. In cloud computing data storage, users do not need to build their own data centers, which greatly reduce the user's cost. They only need to pay the appropriate fee, then they can quickly store the data on the data center and get the corresponding results.

STORAGE OF LARGE AMOUNT OF DATA BASED ON CLOUD COMPUTING

In this process, we have mainly carried out four steps. First, we have a preliminary idea on how to access the large amount of data based on cloud computing. Secondly, we need to select the algorithm and model, and we choose the MapReduce model in this regard, and conduct the corresponding programming based on this model, so as to achieve processing of large data according to the Hadoop framework. Finally, according to the above ideas, we have completed storing a large amount of data based on the cloud computing.

Ideas of storage of large amount of data based on cloud computing

We divide the contents in this regard into two parts, in which the first part is the idea to store data, and the second part is the idea to take data. We first introduce the idea to store data.

When storing data, the data we need to store and some service-related information shall be sent to the main service control machine group, and then the data is sent to the Hadoop framework. Data will be cut into blocks in the process for calculation, and then the data cut into blocks will be distributed to different storage nodes. The information on these nodes will be sent by the main service control machine group to the user as feedback. For each storage node, the user will establish a queue of data block and at last, the data blocks are uploaded in parallel to the corresponding storage nodes.

When obtaining data, the download address of needed information will first be transmitted to the main service control machine group, finding the appropriate information of the file according to the download address, and found information will be sent to the main service control machine group as feedback. The main service control machine group transfers information to the user, and the user needs to create download thread for these storage nodes based on the received information. At last, the file blocks are downloaded in parallel to a temporary file folder on the local computer, and finally the there will be an integration of these downloaded files into one complete file, and the file blocks will be deleted.

Mapreduce

MapReduce is a programming model for parallel computing of large-scale data sets (greater than 1TB). The concept "Map" and "Reduce" has the main idea borrowed from the functional programming language and have the features borrowed from the vector programming languages. It greatly facilitates the programmers if they do not know distributed parallel programming. In this case, they can run their programs on distributed systems. The current software implementation is to specify a Map function, which is used to map a set of key/value pairs into a new set of key/value pairs, and designate concurrent Reduce function to ensure that each of all key/value pairs in mapping shares the same set of keys.

The most prominent feature of MapReduce is the reliable distribution. MapReduce achieves reliability by distributing large-scale operations on data sets to each node on the network; each node will periodically return its completion of work and the latest status. If a node remains silent over the preset time interval, the master node will record this node status as dead, and send data assigned to this node to other node. Each operation use named file's atomic operation to ensure that the conflict between parallel threads does not occur; when a file is renamed, the system may probably copy them to another name other than the name of task.

Its essence is a programming model. MapReduce cloud computing is the very core computing model, and it is not only a distributed computing technology, but a simplified distributed programming model as well. In general, MapReduce is a model and method to solve problems of developing program. The main idea of MapReduce is to dismantle the executing issue and turn it into a Map and Reduce consisting contents. The detailed process of MapReduce is shown in Figure 1.

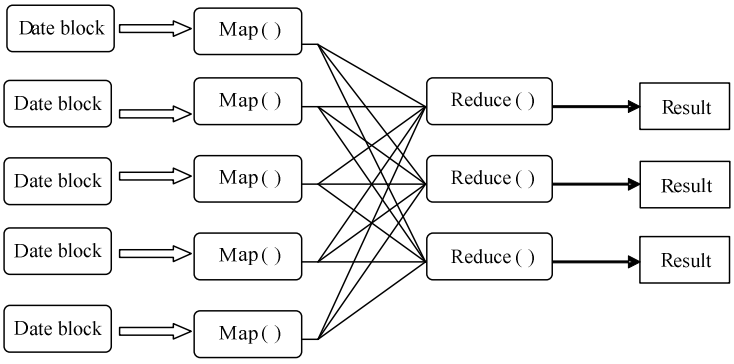


Figure 1 : Detailed process of mapreduce

In the Figure 1 above, map function is mainly map divided data blocks into different blocks and then carry out processing via Reduce function, finally summarize and output the results.

In fact, in MapReduce module, its function to be achieved is to specify a map function, make the key/value pairs mapped into a new key/value pairs, and these new key/value pairs form a series of key/value pairs in the form of intermediate results, and finally Reduce function processing is carried out on these key/value pairs, combining the value with the same intermediate form key. Its specific contents are shown in TABLE 1.

TABLE 1 : Function description in mapreduce module

Function	Input	Output	Description
Map	<k1,v1>	List (<k2,v2>)	Small data set is further analyzed into a group of <key, value> pairs, enter them into the Map function for processing;

Reduce	<k2,List(v2)>	<k3,v3>	② Each input <k1, v1> pair outputs a group of <k2, v2>. <k2, v2> is the intermediate result of calculation The List (v2) in intermediate result of input <k2, List (v2)> means a group of value belonging to the same k2

Hadoop concept

Hadoop is a software framework capable of processing large amount of data in distributed manner. The core design of Hadoop framework is HDFS and MapReduce. HDFS provides storage for massive data, and then MapReduce provides calculation for massive data.

Hadoop mainly consists of HDFS (Hadoop Distributed File System) and MapReduce engine. HDFS is at the bottom, and it stores all the files on the storage nodes it the cluster. Layer above the HDFS is MapReduce engine, which consist of JobTrackers and TaskTrackers. Hadoop also includes a distributed database (HBase, Hadoop Database), which is used to store data to the underlying computer. Hadoop framework is shown in TABLE 2.

TABLE 2 : Hadoop framework

Cloud Computing and Hadoop Architecture	
HDFS distributed file system HBase distributed database	MapReduce API

Hadoop is reliable, because it assumes that computing elements and storage will fail, so it maintains multiple copies of work data, ensuring the failed node can be re-distributed for processing. Hadoop is efficient because it is working in a parallel manner, and speeds up processing through parallel processing. Hadoop is also scalable, and able to process PB-level data.

Hadoop is featured by high reliability, high scalability, high efficiency, high fault tolerance, and low cost, etc. For us, the most useful feature of Hadoop here is to achieve MapReduce computation model. We can make use of Hadoop in programming, and programs written can be used on a computer cluster in order to process large amount of data.

Based on Hadoop frame and combined with MapReduce for calculation and storage of data, we combine the HDFS distributed file system and the HBase distributed database, integrate these into the cloud computing, whereby we can achieve cloud computing and storage, and it has good processing capacity with respect to large data amount. Here we attach the simplified diagram of Hadoop seen as Figure 2.

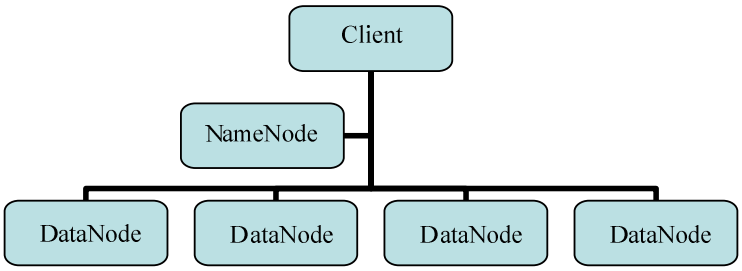


Figure 2 : Simplified diagram of hadoop

Storage model for large amount of data

Based on the features of large data amount and combined with the characteristics of cloud computing, after the previous idea analysis and understanding of MapReduce and Hadoop, we, based on the above, have proposed the model for storage of large amounts of data based on cloud computing. The Storage Structure of the Large Amount of Data is shown as Figure 3.

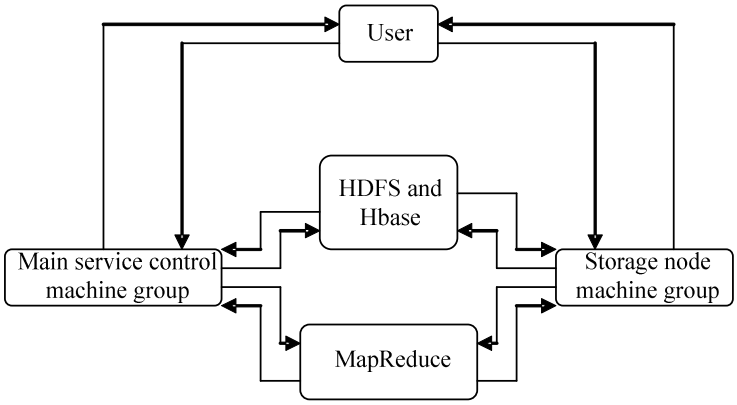


Figure 3 : Storage structure of the large amount of data

In Figure 3, can be seen, data processing of information, primarily through control of the primary cluster to control the service, which is equivalent to a controller that is responsible for applying the received request to give an appropriate response. The computer used in each underlying layer is an equivalent of the storage node machine group and its main function is to store, like a memory, which has huge disk array or the cluster system with storage capacity of a large amount of data. Its main function is to access data resources. HDFS and Hbase in them is to store the needed data separately in each calculation point. In Hadoop, there is a server called JobTracker, it plays the role of scheduling and management of other computers. JobTracker has no fixed running place, and it can run on any computer in the machine group. During the execution of scheduling and management, it is necessary for it to run in the data storage node, i.e. the DataNode, which is also known as compute nodes. JobTracker processes Map tasks and Reduce tasks, and it distributes them to the idle task block, and makes these programs in parallel, and its operation will be monitored during execution. If one task execution fails, then this task will be dispatched to another idle task block for execution. In this process, the user does not need to conduct data reading and storing operations on HDFS and Hbase via Hadoop, so that they can avoid a lot of reading operation, which may cause the system blockage. After the user pass control information to the main service control machine group, the data reading operation can be carried out directly with the storage nodes.

PROSPECTS

Combined with cloud computing technology and data storage, challenges we are now facing as well as research focus and application prospects, the author believes that future research may be carried out from several aspects.

First is the study on the data center network architecture and routing strategies for storing large amount of data. Different data access will have different characteristics, for example for the application characteristics of key/value pairs for frequently access, or for the access to storage applications, etc., we can design the appropriate network data structure, accordingly.

Second is how to reduce the cost of data center network. With expansion of data center scale and increasing amount of data, data center will require higher building cost. With the constant improvement in computer hardware, hardware process and performance has achieved tremendous development, which makes the replacement of the low-end switch with high-end switch possible, and finally reduces the cost of building data center network.

Third is research on erasure code placement technology. For the current technology, the data stored under the cloud computing environment has relatively simple placement strategies, but most of these strategies are fault-tolerant technology based on replication, in fact, the fault-tolerant technology based on erasure codes has a great influence on fault tolerance and access efficiency in different strategies, but there are few studies in existing works based on this.

Fourth is the energy saving research of erasure code. Software energy-saving technology is currently a hot spot of data storage in cloud computing environment, although there have been some studies, they are not yet mature, and most of the existing researches are carried out with respect to the fault-tolerant technology based on replication. For how to place block erasure codes in order to reduce the cost of data migration, there is still not a more in-depth study that can be conducted in this regard.

Fifth is the development and deployment of data storage system. To date, the cloud computing storage platforms that have been deployed and applied are built by some large companies, such as Google’s GFS and Amzon’s and so on. But the most influential in academia is Hadoop system, and we can further explore this system so as to promote further theoretical research.

CONCLUSION

As the next generation of computing model, cloud computing is widely used in many areas of scientific computing and commercial computing, Its data center physical network topology construction technology as well as the technology to improve data fault tolerance and various energy-saving technologies to reduce energy consumption still need continuous research.

Generally speaking, the research in the field of cloud computing is of course still in its infancy, still lacks a unified and clear framework, and there are still a lot of unclear problems to be solved. For the large amount of data storage technology research for cloud computing, the meaning and value is enormous. In the current studies, they are more focused on the cloud architecture, cloud storage, cloud data management, virtualization, cloud security and other technologies, and there are still many some open-ended questions in cloud computing field requiring further study and exploration.

REFERENCES

[1] He Xueqing, Wu Jinghai; Research on digital library resources storage based on cloud computing [J], Information Research, **15(12)**, 92-95 **(2010)**.

[2] Tuo Shouheng; Research on cloud computing and cloud data storage technology [J], Computer Development & Applications, **15(9)**, 1-4 **(2010)**.

[3] Cheng Jingjing; Study and design based on Hadoop distributed cloud computing/cloud storage solutions [J], Data Communication, **28(5)**, 14-19 **(2012)**.

[4] Jiang Wuxue, Zhang Jing, Wang Zhiming; Research on mapreduce parallel programming architecture [J], Microelectronics & Computer, **5(06)**, 168-171 **(2011)**.

[5] Zhang Diankui; Research on virtualization storage technology based on cloud computing [J], Communication of Science and Technology, **23(16)**, 218-219 **(2013)**.

[6] Li Yumin, Zhang Caineng, Xie Jie; Data storage in cloud computing environment [J], Computer Knowledge and Technology, **05(05)**, 1032-1035 **(2010)**.

[7] Li Xiaofei; Research on big data processing system based on cloud computing technology [J], Journal of Changchun Institute of Technology (Natural Science Version), **15(05)**, 136-140 **(2014)**.