

# Research on Data Storage Technology in Cloud Computing Environment

**Caiyun Xu\***

School of computer and Information Engineering, Wuhan Institute of Bioengineering,  
Wuhan, China

\*Corresponding author e-mail: 1845503683@qq.com

**Abstract.** With the rapid development of information technology, cloud computing, large data, mobile interconnection and other technologies, the number of users is increasing, and the amount of data generated by users is increasing exponentially. How to store large scale and massive data, the traditional distributed storage technology will face challenges. Distributed storage technology can use multiple servers to store data and share storage load. However, in the cloud computing environment, it looks pale. In order to improve the storage efficiency of massive data, data blocking algorithm is studied to shorten the response time of data backup. Effective application of data duplication technology improves the storage performance of data.

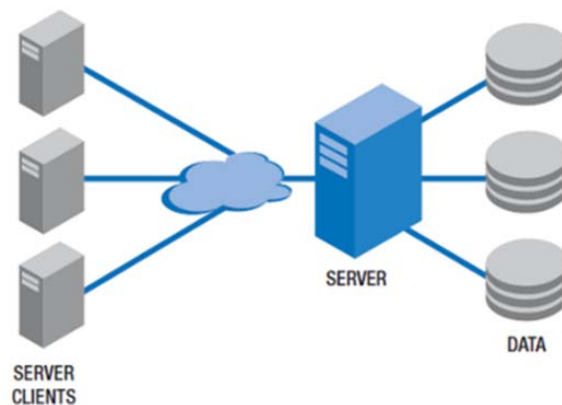
## 1. Introduction

With the rapid development of cloud computing, big data and mobile Internet technology, cloud computing technology as the core of the new generation of IT technology [1], it is exerting a subtle influence on people's way of life and learning. The concept of cloud computing was first put forward at the search engine conference held in 2006. So far, it has penetrated into all walks of life. Cloud computing is to automatically split a huge computing process into numerous smaller subroutines through a network, and then make a large system composed of multiple servers to assemble a large number of heterogeneous storage devices in the network and return the results to users after search, calculation and analysis. In this environment, the volume of data generated by enterprises and individual users is increasing exponentially. It is estimated that the total data of 40ZB will be generated worldwide by 2020 [2]. How to store large scale, massive data, data processing and storage technology is facing new challenges. Traditional data storage technologies such as network storage, centralized storage, and distributed file systems can not efficiently complete the task of storing and processing mass data in the cloud computing environment. Therefore, this paper makes detailed research on data storage technology under the cloud computing environment, improves the data partitioning strategy, shortens the response time of data backup, and effectively applies duplicate data deletion technique to improve the efficiency of data storage. As the core component of the new generation of IT technology, cloud computing will become a new hot spot of the new revolution of information technology in the world after the personal computer and the Internet [3].



## 2. Distributed data storage technology

Cloud computing is a novel computing model, which consists of a large number of servers, mainly contains the cloud computing platform management technology, the programming model, virtualization technology, data management technology, data storage technology, etc. Cloud computing system data storage adopts distributed storage technology, and the reliability of data is guaranteed by redundancy storage [4]. The core idea of distributed storage system is to store data in multiple independent storage devices, using extensible system structure, using more than one storage server. Shared storage load, using location information stored in the server location, not only improves the system reliability, availability, and access efficiency, but also is easy to expand. The storage structure is shown in Figure 1. Common distributed storage technologies are:



**Figure 1.** Distributed structure diagram

### 2.1. Network storage technology

Network storage is a special private data storage server, which can provide cross-platform file sharing function. Network storage usually occupies its own node on a LAN without the need for application server intervention to allow users to access data on the network. In this configuration, network storage centrally manages and processes all data on the network, unloads the load from the application or enterprise server, effectively reduces the total cost, and protects the user's investment.

### 2.2. The file system

The distributed file system is a single name space which combines the geographical location of the files on different computers, and makes a single, hierarchical multiple file server on the network. Files distributed on multiple servers are in the same location as users in front of the network. Users are more convenient to access and manage data. For example, GFS, Google File System, is an extensible distributed file system for large-scale, distributed applications that access large amounts of data. The design idea of GFS is different from the traditional file system. It is designed for large-scale data processing and Google application features.

### 2.3. P2P storage technology

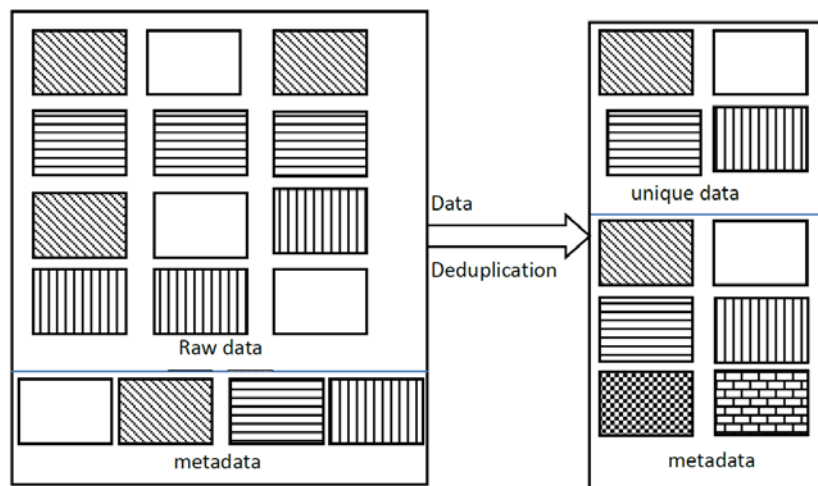
Peer-to-peer technology (P2P), also known as peer interconnected network technology, is a new network technology that relies on the computing power and bandwidth of the participants in the network, rather than relying on a small number of servers. A pure point-to-point network has no concept of a client or server, it only has equal peer nodes, and ACTS as a client and server for other nodes on the network. P2P networks can be used for many purposes, such as various file sharing software, real-time media business, etc [5]

New features based on cloud computing environment, such as s mass data, super large-scale user scale and high availability, traditional data storage and management pose new challenges. The traditional data storage and management is facing new challenges. The distributed storage system and its related technologies are flexible and adaptable to the complex data storage and management tasks. However, it is still unable to meet the requirements of the massive data scale and user scale.

### 3. Data deduplication Technology

With the rapid development of large data technology and mobile Internet, how to efficiently store and manage the huge amount of data generated by various organizations and individuals using distributed data storage technology is the focus of current research.

In recent years, Data deduplication technology has been used as an effective solution for massive data storage and management. The main idea is to delete duplicate data in the data set, and only one of them is retained, thus eliminating redundant data. Data deduplication is shown in Figure 2:



**Figure 2.** Data Deduplication

In cloud computing and big data environment, massive data makes the problem of data storage more prominent. Using "deduplication" technology can reduce the storage of data to the original 1/20, save more backup space, keep the backup data longer and reduce the resource consumption of data center.

After deleting data, it does not affect the speed of data backup. How does the data deleting technology include how files are segmented? How does the data block fingerprint be calculated? How to carry out data block retrieval? It is particularly important.

#### 3.1. Fixed-length block

The fixed-length block algorithm is used to synopate the file with the predetermined block size, and the weak check value and the md5 strong check value are performed. The weak check value is mainly to improve the performance of the differential coding, first calculate the weak check value and perform the hash lookup, and then calculate the md5 strong check value and make further hash lookup. Because the computation of weak check value is much smaller than md5, it can effectively improve the coding performance. The advantages of the fixed-length block algorithm are simple and high performance, but it is very sensitive to data insertion and deletion, which is very inefficient and cannot be adjusted and optimized according to the content changes.

### 3.2. The CDC algorithm

The CDC is variable length block algorithm, which is based on the network file system LBFS [6], puts forward repeated data deletion based on different content block ideas. First, the CDC algorithm uses the Rabin fingerprint algorithm [6] to divide the file into data blocks with varying degrees of growth. Second, the hash algorithm is used to calculate the fingerprint of each data block using a fixed-size sliding window. Finally, the obtained data block fingerprint is compared with the existing data block fingerprints in the storage system, and the duplicate data blocks with the same fingerprint value are deleted, and there is only one piece of data block in the file system. Thus, the storage space occupied by the data in the disk is effectively reduced, and the utilization efficiency of the existing storage space is improved. However, during the execution of CDC algorithm, the size of data blocks is difficult to determine. The granularity is too detailed and the overhead is too large. In the process of dividing the data blocks, a large number of large data blocks and small data blocks may be generated, and these two extreme cases are not a better solution at present.

### 3.3. Sliding block algorithm

The sliding block algorithm combines the advantages of fixed length segmentation and CDC segmentation, and the block size is fixed. It first calculates the weak check value for the fixed length block, and then calculates the MD5 strong check value if the match is applied. Both matching is considered as a data block boundary. The data fragment in front of the data block is also a block of data. It is indefinite long. If the distance between sliding windows and a block size is still not matched, it is also identified as a data block boundary. The slider algorithm is very efficient for inserting and deleting problems, and can detect more redundant data than CDC, which is not enough to produce data fragments easily. After optimized block partition strategy, the number of data blocks can be reduced. The number of data blocks is reduced. On the one hand, it can improve the query and contrast efficiency of the data block fingerprint. On the other hand, it also reduces the storage cost of data blocks and improves the overall performance of the system.

## 4. Conclusion

In large scale, massive data, virtualization, and high scalability cloud computing environment, how to optimize the performance of the distributed storage system and ensure its high reliability is a key research problem. Based on data deduplication technology, this paper studies the data partitioning technology, compares the advantages and disadvantages of the fixed long block algorithm and the variable length block algorithm, and uses the slider algorithm to optimize the storage space utilization, shortens the data backup time and improves the distributed data storage performance based on the cloud computing environment.

## Acknowledgments

This work was financially supported by Wuhan municipal higher education research project fund (NO.2017179) and Wuhan Institute of Bioengineering teaching research project (NO. 2017J25).

Introduction: Caiyun Xu (1978. 09- ) Female, Master, lecturer, Inner Mongolia people, Main research direction: Database Technology, Address: Wuhan Yangluo Economic Development Zone Han Shi Lu 1, zip code: 430415, E-mail: 1845503683@qq.com

## References

- [1] Merrikh-Bayat F, Bagheri-Shouraki S. Mixed analog-digital crossbar-based hardware implementation of sign–sign LMS adaptive filter [J]. Analog Integrated Circuits and Signal Processing, 2011, 66(1): 41-48
- [2] J. Abawajy, M. Deris. Data replication approach with data consistency guarantee for data grids [J]. IEEE Transactions on Computers, 2014, 63 (12): 2975–2987
- [3] B. Liskov, J. Cowling. Viewstamped replication revisited [R]. Cambridge: DSpace@MIT, July 23, 2012

- [4] H. Howard, D. Malkhi, A. Spiegelman. Flexible paxos: quorum intersection revisited[J]. Distributed, Parallel, and Cluster Computing, 2016: 1–20
- [5] P. Li, D. B. Gao, M. Reiter. Replica placement for availability in the worst case [C]. 2015 IEEE 35th International Conference on Distributed Computing Systems (ICDCS), Columbus, 2015, 599–608
- [6] Wood T, Ramakrishnan K K, Shenoy P, et al. CloudNet: dynamic pooling of cloud resources by live WAN migration of virtual machines [J]. IEEE/ACM Transactions on Networking, 2011, 46(7): 121-132.