# Research on Cloud Storage Platform Technology Based on Hadoop

Hairong Hu¹, Lulu Fang², Chaohong Yang¹, Yang Zhang¹

¹Department of Information and Communication, Army Academy of Armored Forces, Beijing, China
²Center of Exercise and Training, Army Academy of Armored Forces, Beijing, China
890809long@163.com, etyk12@ykzq-it.com

Abstract. In order to solve the problem of high cost of traditional data calculation and storage, and difficulty in writing parallel programs, this article summarizes the core technology of Hadoop, and uses Hadoop distributed processing technology and virtualization technology to design and build a cloud computing storage platform. The experiment proves that, compared with the traditional single-machine computing method and the physical machine cluster computing method, the platform has higher performance, and the resource utilization rate has also been improved.

## 1. Introduction

With the development of Internet technology, digital information is increasing exponentially. According to the Digital Universe report released by the Internet Data Center, the amount of data generated in the next 8 years will reach 40 ZB, which is equivalent to 5200G of data per person [1] How to efficiently calculate and store these massive data has become a challenge for Internet companies. Traditional large-scale data processing mostly uses parallel computing, grid computing, distributed high-performance computing, etc., which consumes expensive storage and computing resources, and the effective allocation of large-scale data computing tasks and reasonable data segmentation require complex programming skills to be achieved [2]. The emergence of distributed cloud platform based on Hadoop has become a good way to solve such problems. This article reviews the core Hadoop technologies: HDFS and MapReduce, and then uses VMware virtual machines to build an efficient and scalable cloud data computing and storage platform based on Hadoop distributed technology, and verifies the advantages of distributed computing and storage by experiment results.

## 2. Hadoop and Related Technologies

Hadoop [3] is a product of the development of parallel technology, distributed technology and grid computing technology, and is a model architecture developed to adapt to large-scale data computing and storage. Hadoop is a distributed computing and storage framework platform of the Apache Company, which can store large amounts of data efficiently, and can write distributed applications to analyze and calculate massive data. Hadoop can run programs in a large number of cheap hardware equipment clusters, and provide reliable and stable interfaces for each application to build a highly scalable and highly reliable distributed system. Hadoop has the advantages of low cost, high reliability, high fault tolerance, strong scalability, high efficiency, strong portability, free and open source, and so on [4]. The Hadoop cluster is a typical Master / Slaves structure. The cloud computing and storage
architecture model based on Hadoop is shown in Fig.1.

![Cloud computing and storage architecture model based on Hadoop](image)

**Fig.1 Cloud computing and storage architecture model based on Hadoop**

### 2. 1. HDFS (Hadoop Distributed File System)

HDFS [5] is a distributed file system running on a large amount of cheap hardware. It is the underlying file storage system of the Hadoop platform. It is mainly used for data management and storage. It has good performance for large files’ data access. HDFS is similar to the traditional distributed file system, but there are also certain differences. It has the characteristics of hardware failure, large data sets, simple consistency, data streaming access, and the convenience of mobile computing [6]. The workflow and architecture of HDFS are shown in Fig.2.

![Workflow and architecture of HDFS](image)

**Fig.2 Workflow and architecture of HDFS**

One HDFS cluster has one Name Node and multiple Data Nodes. As shown in Fig.2, the Name Node is the central server, which is used to manage the metadata information of the file system and the client's read and write access to files, and maintain all files and directories under the file system tree and its sub-nodes. The information is saved on the disk in the form of edit log file and namespace image file. The Name Node also temporarily records the Data Node information where each block is located. Its functions mainly include: managing metadata and file blocks; simplifying metadata update operations; monitoring and processing requests [7].

There is usually one Data Node in per node in the cluster, the Data Node is used to store and search data blocks, respond to the commands issued by Name Node to create, copy, and delete data blocks, and periodically sends "heartbeat" to Name Node, and sends heartbeat information to report its own load status, and at the same time accepts the command information issued by the Name Node through the heartbeat information. The Name Node uses the "heartbeat" information to determine whether the Data Node is invalid, and it regularly pings each Data Node. If the Name Node does not receive the feedback from the Data Node, it will consider that the node is invalid, and then adjusts the load of the entire system. In HDFS, each file is divided into one or more data blocks and stored in different Data
Nodes. The data blocks are copied among the Data Nodes to form multiple backups.

2. 2. Map / Reduce Programming Framework
Map / Reduce [8] is a programming framework used by Hadoop to process massive data in cloud computing. It is simple and easy to be used. Programmers can write programs to process massive data without understanding the underlying implementation details. Using Map/Reduce technology, it is possible to simultaneously carry out tasks such as advertising business and web search on thousands of servers, and can conveniently process terabytes, petabytes, and even EB-level data. The Map / Reduce frame is composed of Job Tracker and Task Tracker. There is only one Job Tracker; it is the master node which is used for task allocation and scheduling, and managing several Task Trackers. One node has one Task Tracker, which is used to accept and process tasks sent by Job Tracker. MapReduce performs distributed operations on large data sets in the cluster. Its entire framework consists of Map and Reduce functions. When processing data, Map function is executed first and then Reduce function is executed. The specific implementation process is shown in Fig.3. The input data is sliced before the Map function is executed; then different fragments are assigned to different Map execution, after the Map function is processed, outputs in the form of <key, value>; before entering the Reduce phase, the Map function first divides the original <key, value> into multiple groups of key-value pairs and then sends them to a Reduce for processing. Finally, the Reduce function merges the values with the same key and outputs the result to the disk.

3. Cloud Storage Platform Design Based On Hadoop
At present, when building a Hadoop cluster system, due to the widespread use of multi-core computers, the multiple tasks assigned to each Data Node will cause competition for resources, such as memory, CPU, input and output bandwidth, etc. The received resources will be in an idle state, resulting in a waste of resources and a prolonged response time. The increase in resource overhead will eventually lead to a decrease in system performance. To solve the problem, this article proposes a cluster environment model based on the combination of VMware virtual machine and Hadoop, as shown in Fig.4. It builds multiple virtual operating systems in one computer. The advantage of this approach is that Data Node and Task Tracker nodes can be added, and physical resources can be fully utilized, the efficiency of computing and storage can be improved.
4. Experimental Platform Constructions

4.1. Hardware Environment Configuration
We prepare 3 dual-core computers, install 2 VMware in each computer, and then install Linux OS in VMware. Thus, the 3 computers are expanded into 6 computers, and the 3 computers have the same configuration. The configuration is shown in Table 1.

Table 1 Software and hardware configuration information

| Hardware | CPU: Intel Core i7 |
|----------|--------------------|
|          | Memory: 4G         |
|          | Disk: 20G          |
| Software | OS: Red Hat Linux OS5.0 |
|          | JDK version: jdk_1.6 |
|          | Hadoop version: hadoop-1.0.3 |

Table 2 Hadoop cluster device configuration information

| Host | IP address | Node type          | Mem | Disk |
|------|------------|--------------------|-----|------|
| master | 192.168.1.100 | NameNode, JobTracker | 2G  | 20G  |
| node1  | 192.168.1.101 | DataNode, TaskTracker | 1G  | 20G  |
| node2  | 192.168.1.102 | DataNode, TaskTracker | 1G  | 20G  |
| node3  | 192.168.1.103 | DataNode, TaskTracker | 1G  | 20G  |
| node4  | 192.168.1.104 | DataNode, TaskTracker | 1G  | 20G  |
| node5  | 192.168.1.105 | DataNode, TaskTracker | 1G  | 20G  |

The Hadoop cluster includes 1 Name Node server and 5 Data Node servers. The configuration information is shown in Table 2.

4.2. Hadoop Environment Constructions
The Hadoop environment construction process is: configure the cluster hosts list, install the JAVA JDK system software, configure environment variables, generate login keys, create user accounts and
Hadoop deployment directories and data directories, configure hadoop-env.sh, configure core-site.xml, hdfs-site.xml, and mapred-site.xml. After the configuration is complete, the file will be formatted and the command is that: /opt/modules/hadoop/hadoop-1.0.3/bin/hadoop namenode format and then start all nodes, enter the command: start-all.sh. To check whether the cluster is deployed successfully through the interface, we firstly check whether the Name Node and Data Node are normal, open the browser and enter the network address: http://master:50070. If there are 6 Live Nodes, it means that all nodes have been successfully started. Then check the Job Tracker and Task Tracker nodes, enter the URL: http://master:50030, if there are 6 Nodes, the node startup is successful.

5. Experimental Content and Result Analysis

Experiments are conducted on the deployed Hadoop cloud data computing and storage platform to verify that the method based on distributed data computing and storage has advantages in data computing and storage.

1) Experiment 1: Running Hadoop's internal Monte Carlo seeking PI program to verify the efficiency of distributed cloud computing based on Hadoop. The calculation tasks are set to 10, and the calculation amount is $10^3$, $10^4$, $10^5$, and $10^6$. Environment 1: Running in a single machine; Environment 2: Running in a cluster system built by 3 physical machines; Environment 3: Running in a cluster system built by 6 virtual machines.

Each group of experiments runs 5 times to calculate the average of the required time. The results of the calculation execution time are shown in Fig.5. The vertical axis is time (s), and the horizontal axis is calculated amount (power). It can be seen from Fig.5 that the computing time in a single-machine environment is much longer than that in a distributed system, and the more nodes in the cluster system, the faster the computing speed.

![Fig.5 Performance comparison of Monte Carlo program running in 3 environments](image-url)
2) Experiment 2: Testing the efficiency of distributed cloud data reading and writing based on Hadoop by running the character counting program (wordcounter.jar) to verify its storage performance. There are 4 sets of data; the sizes are 400MB, 600MB, 1GB and 1.5GB. In this group of experiments, the Hadoop block size is set to 16M (64M by default), the redundancy backup parameter is set to 3 (default), the experimental environment is the same as experiment 1, the program runs 5 times, we record the time and calculate the average.

The running result is shown in Fig.6. The vertical axis is execution time (s), and the horizontal axis is data volume (MB). It can be concluded from Fig.6 that the data reading and writing speed in a stand-alone environment is significantly lower than that in a distributed environment, and the more nodes there are, the faster the reading and writing speed are.

It can be seen that, compared with traditional data computing, reading and writing methods, the cloud computing and storage platform based on Hadoop distributed technology built in a virtualized environment proposed in this article effectively improves the speed of massive data analysis and reading and writing. The use of virtualization technology to build clusters is more efficient and faster than physical machine clusters, which greatly improves the utilization of resources.

6. Conclusions
This paper studies the Hadoop distributed file system (HDFS) and MapReduce programming framework, uses VMware virtual machine to build a Hadoop-based cloud data computing and storage platform, and verifies that it is more efficient and faster compared to traditional data processing methods through experiment results, and satisfies related requirements in the field of cloud computing; and through the application of virtualization technology to expand the number of nodes, it not only improves operating efficiency but also improves the utilization of hardware resources, laying the foundation for future cloud computing research directions.

References
[1] Zhang Bo. Research and optimization of file storage under HDFS [D]. Guangzhou: Guangdong University of Technology, 2013.
[2] Cui Jie, Li Taoshen, Lan Hongxing. Design and implementation of a massive data storage platform based on Hadoop [J]. Computer Research and Development, 2012 (49): 12-18.
[3] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, et al. The Hadoop Distributed File System [Z]. Sunnyvale, California USA, IEEE 2010: 1-10.
[4] Taylor RC. An overview of the Hadoop/MapReduce/HBaseframework and its current applications in bioinformatics [C] //BMC Bioinformatics. Washington: BMC Bioinformatics, 2010.
[5] Wang Hongyu. Application of Hadoop platform in cloud computing [J]. Software, 2011, 32(4):
36-39.

[6] Huang Xiaoyun. Research on cloud storage service system based on HDFS [D]. Dalian: Dalian Maritime University, 2010.

[7] Wang Yongzhou. Research on HDFS-based storage technology [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013.

[8] Gao Jichao. Research and optimization of storage strategy on Hadoop platform [D]. Beijing: Beijing Jiaotong University, 2012.