Application of Hadoop Distributed Cluster in Big Data

Yuanyuan Liu¹*
¹Qingdao Technical College, Shandong, China, 266555
*Corresponding author e-mail: liuyuanyuan@qtc.edu.cn

Abstract. Hadoop is the core technology of large data storage. Distributed data storage based on Hadoop technology plays a central role in large data storage. This paper introduces the basic concepts of Hadoop platform, focusing on the analysis of HDFS and MapReduce. Then, the process of building Hadoop cluster environment is described in detail, and Hadoop is applied to a file publishing system. Finally, time-consuming comparisons of file upload operations are made for different orders of files in different clusters [1].

Keywords: Hadoop Distributed Cluster, Large Data, Big Data

1. Introduction
With the explosive growth of network information, the way to improve data processing ability only by upgrading computer hardware equipment has become "inadequate". In order to meet the needs of high-speed growth of information and efficient processing, Hadoop is regarded as an effective solution to high-performance processing of large data, and has become a hot research object of cloud computing at home and abroad. By studying the principle of Hadoop's distributed storage framework HDFS (Hadoop Distributed File System) and MapReduce distributed computing framework, Hadoop is applied to the actual big data storage in the cluster environment, and the superiority of Hadoop in big data processing is verified by experiments. In today's information age, distributed data storage based on Hadoop technology can meet the needs of modern enterprises and individuals. The development of this technology not only helps the rapid development of information technology in China, but also helps the popularization of information technology for the whole people [2].

2. Hadoop basic principles
Hadoop is a distributed system infrastructure developed by Apache Foundation and a software platform for processing large-scale data. It has the characteristics of mass storage (able to process data at PB level), low cost, high efficiency and reliability. It mainly includes two core functions: HDFS and MapReduce. HDFS is responsible for distributed storage of large-scale data, MapReduce is responsible for distributed computing of large-scale data [3].

2.1. Hadoop distributed file system
HDFS is a distributed file system running on ordinary PC. It uses Master/Slave structure to build a cluster system, usually consisting of one Master node and several Slave nodes (Facebook has amended this by using two nodes to act as Master, one is Active and the other is Standby, which
automatically switches to Standby node when Active node fails). Master node is called Name Node, which is usually served by high-performance servers. It is mainly responsible for the management of metadata, including the namespace management of file system and the access operation of client to files. Slave node is called Data Node, which is usually composed of cheap PC in cluster environment. It is mainly responsible for data storage management and response to client's read and write requests [4]. In HDFS, the directory structure of files is stored on NameNode independently, and every actual data file is split into several Block. These Block redundancy is stored in the data of DataNode collection. NameNode is responsible for storing the directory information, file information and corresponding block information of the entire file system, as shown in Figure 1.

Figure 1. HDFS storage principle.

2.2. Map/Reduce
Map/Reduce is a computing model used by Hadoop platform for parallel operation of massive data. The data processing process of the model is divided into two stages: the first stage of Map and the second stage of Reduce. The approximate calculation process is to divide the data into several parts, call each processing process Map, and then merge each processing result. The merged data processing is called Reduce, as shown in Figure 2.

Figure 2. MapReduce working process.

3. Application of Hadoop distributed cluster in large data processing
Taking a Web-based information publishing system as an example, the application of Hadoop core API is described in detail with the function of file upload. Users can upload text information, pictures and videos in the system. With the increase of system users and running time, it is inevitable to accumulate large data [5]. How to store and retrieve large data effectively is the key to the successful operation of the system. For data storage system, the traditional relational database and HDFS are
combined to realize. The database stores basic information, such as the name of the video file, the uploader, the upload time and so on. Real video files are stored in the cluster environment by HDFS. When the number of cluster nodes is 1 (stand-alone), 2 and 4, files with capacity of about 500MB, 2GB and 10GB are uploaded respectively. In order to reduce errors, the average time of five times is taken. The results are shown in Table 1. Figure 3 shows the data in Table 1.

| File size | Average operating time |
|-----------|------------------------|
|           | Stand-alone  | 2 nodes   | 4 nodes   |
| 500MB     | 74.3        | 295.5     | 384.5     |
| 1.5GB     | 302.7       | 220.2     | 310.4     |
| 10GB      | 2251.1      | 2136.2    | 1544.4    |

4. Conclusion
Hadoop implements a distributed storage and computing framework. Using Hadoop to process large-scale data is an effective way to deal with the explosive growth of data volume. Hadoop distributed cluster is applied to file upload function, and different data level files are tested by single node and cluster environment. The experimental results show that the larger the amount of data, the more the number of cluster nodes, the stronger the data processing ability of Hadoop cluster. How to apply MapReduce to solve some practical problems, how to make Hadoop achieve load balancing in cluster environment, and how to select the appropriate number of clusters for different orders of magnitude, will be the next research direction.
References
[1] Guo Song, Zhao Guohai. Research on cloud computing platform based on Hadoop [J]. 2013, 31 (Natural Science Edition), 2013, 31 (1): 86–87.
[2] Qiang Baohua, Shilong. Large scale data index construction and cluster performance analysis based on Hadoop MapReduce [J]. J, 2012, 32 (4): 310–311.
[3] Wang Hua. Massive data storage technology research on Hadoop platform [D]. Master's Degree Thesis of Northeast Petroleum University, August-9, 2012.
[4] Lin Qingfu. Cloud Computing Model Based on Hadoop [J]. Modern Computer, 2010, 7: 114-115.
[5] Xia Chun he, Li Xiaojian, A computer network defense strategy language, computer programming technology and maintenance, 2012, 10.
[6] Huang Zhenkui. Statistics and application of massive data based on Hadoop. (Master's thesis), Nanjing: 2004: 162-165.