Research on Data Migration in Computer Cloud Storage

Pan Shanshan

Xi 'an Eurasia University, Xi 'an, 710065, China

Keywords: Data Migration; Cloud Storage; HDFS System

Abstract: In today's information society, it is a huge amount of data resources that support enterprises to make excellent decisions. In the fierce competition of data migration tools and services, it is crucial to seek better design concepts and to be able to take a longer-term view of product design. In view of the advantages and disadvantages of the current popular data migration tools in the industry, this paper has done a lot of research and learning work, and has a thorough understanding of the advantages and disadvantages of the existing tools. Based on the actual needs of the industry, a data migration cloud service system is designed.

1. Introduction

In recent years, with the rapid development of electronic information technology, computer networks and software have refreshed the world at an unprecedented speed. From cloud computing to big data, from machine learning to neural network and in-depth learning, one of the conditions for technological change is massive data [1][2]. Data migration is a matter of great importance to an enterprise, and at the same time it is a matter of high technical requirements. Because data is an invaluable asset, the storage and migration of data often have high requirements for the security and stability of data. [3][4]

In the fierce competition of Hadoop-related data migration tools and services, it is critical to seek better design concepts and to be able to look at product design in a longer term [5]. In current design, interfaces can be reserved for possible future changes. For example, providing support for the future Hadoop data security framework Apache Argus is an important aspect. Overall, the best long-term investment in Hadoop data migration is to understand existing tools and then combine their strengths to create better user satisfaction. [6]

In view of the advantages and disadvantages of the current popular data migration tools in the industry, this paper has done a lot of research and learning work. Understanding the strengths and weaknesses of existing tools. [7] Based on the actual needs of industry, a data migration cloud service system is designed.

Firstly, this paper introduces the relevant theoretical knowledge of data migration. Based on the research and analysis of this aspect, it provides the theoretical support of data migration for this article. Subsequently, this paper designs a data migration cloud service, which is closely related to the actual needs of enterprises and contains the core functions of data migration. For example, full data migration, i.e. one-time migration of all designated data from data sources to Hadoop ecosystem; incremental data migration, i.e. incremental migration of designated data from data sources to Hadoop processing platform in batches; streaming data migration, which is also the focus of this paper. Streaming data migration is similar to real-time data migration, that is, the data source changes in real time, and the system can automatically collect incremental data and migrate periodically to Hadoop platform. The whole data entered the Hadoop ecosystem efficiently in the form of water flow.

The traditional data migration, single-machine operation, to the level of cloud computing services. [8][9][10] That is, data migration is provided to users as a cloud service. Cloud computing will take advantage of the overall computing power of the cluster to improve the performance of data migration tasks, monitoring and management modules to a very high level. The performance is greatly improved. At the same time, it greatly simplifies the learning cost of users. Users can apply for tasks and fill in configurations through the web interface. Users can view the whole process

during the intermediate migration process, which is convenient and easy to use.

2. Analysis of Data Migration Related Technologies

HDFS is used as the main storage of Hadoop cluster. An HDFS cluster consists of two types of nodes: NameNode and DataNode. NameNode manages the namespace of the file system. DataNodes stores data files and divides large files into fixed-size blocks [11]. It guarantees the fault tolerance of DataNodes by each replication block (the default number of replicas is three).

The architecture of HDFS is shown in Figure 1.

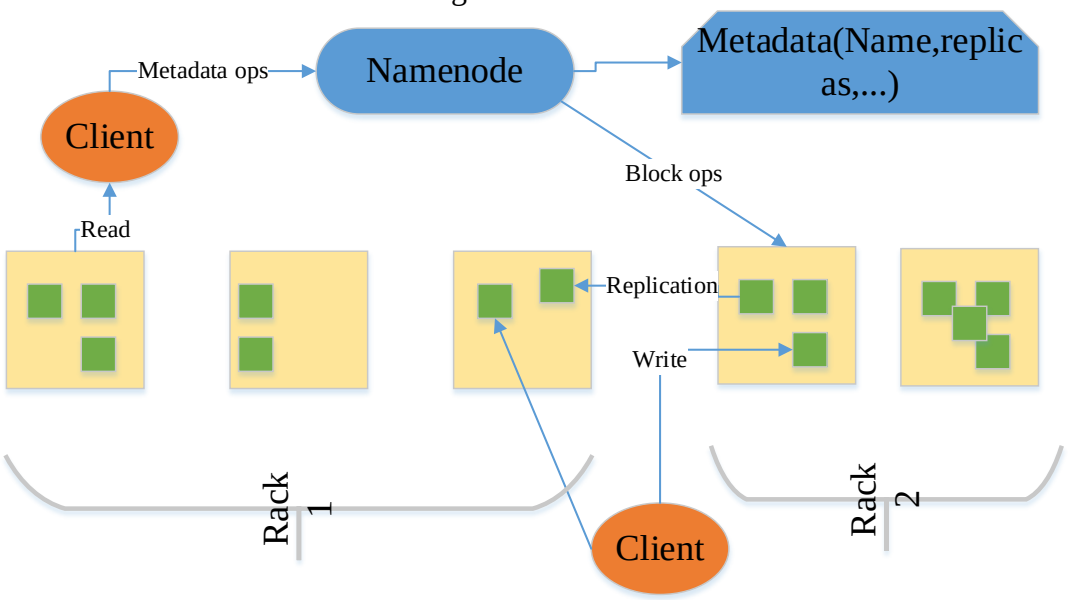

Fig.1. HDFS system structure diagram

NameNode manages the Namespace of the file system. It maintains the file system tree and metadata of all files and folders in the file tree. There are two files to manage this information: Namespace image and edit log. NameNode records the location information of the data nodes where the blocks are located in each file, but it does not persist the information because it will be reconstructed from the data nodes at system startup. NameNode is an HDFS namespace that stores user data in block data nodes based on file blocks. NameNode performs namespace operations such as opening, closing, and renaming files and directories. NameNode also maintains and finds the mapping of data blocks to data nodes. The responsibility of the data node to read and write requests from the client side of the file system.

DataNode is the working node of the file system, which stores and retrieves data according to the scheduling of the client or NameNode, and regularly sends NameNode a list of blocks they store. Each cluster node has a subordinate DataNode that serves read/write requests and performs data block creation, deletion and replication as directed by NameNode. A file is divided into one or more data blocks, which are stored in a set of data nodes.

3. Overall Design of Data Migration Cloud Service

Today, with the rapid expansion of data, the existing database and data query and analysis system of enterprises can not satisfy the high concurrent application scenarios. The original structured data such as relational database data, unstructured data CSV, EXCEL and other files need to be migrated to the distributed storage and processing platform for storage and analysis. The data migration cloud service designed in this paper is closely related to the actual needs of enterprises. In view of the shortcomings of the existing data migration system in use and maintenance, a more perfect architecture design is proposed. The purpose is to improve the high availability and reliability of data migration system, greatly simplify the operation configuration of data migration, and provide it to business departments in the form of cloud services. Its main functional requirements are as

follows:

1) Data migration implements data migration from traditional storage platform to Hadoop platform.

2) Design a high availability, high reliability and high throughput data migration cloud service system.

3) Reasonable and effective migration task management and scheduling system.

4) Scientific and Effective Monitoring and Fault Alarm Recovery System

The whole data migration cloud service system consists of several modules, including cloud service foreground task acceptance module, migration task management and storage module, load balancing module, data migration task execution module, monitoring and alarm module, etc.

The overall architecture is shown in Figure 2.

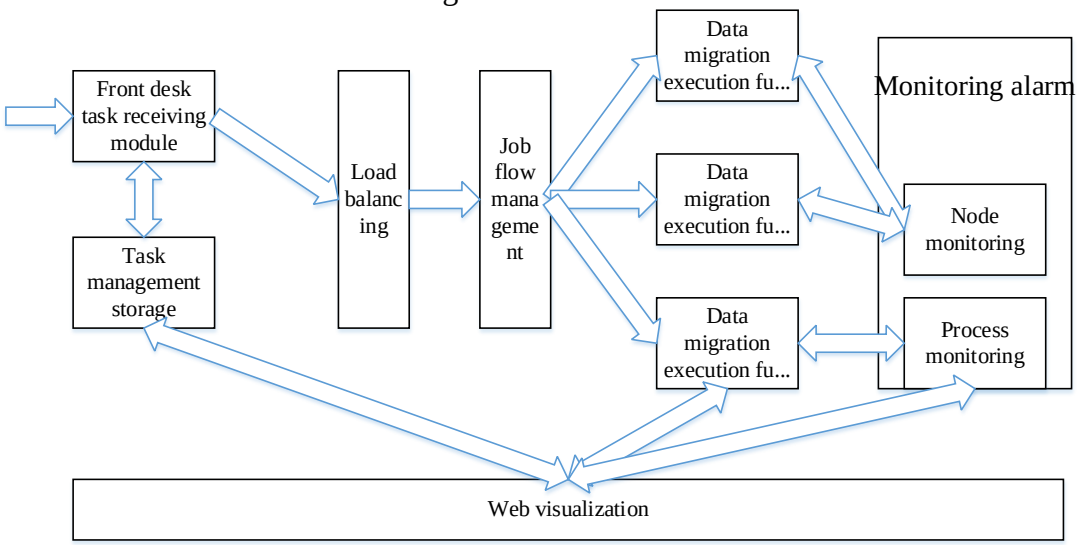

Fig.2. Data migration overall framework

The whole cloud service architecture design is based on the actual needs of users and can meet the functions of multi-user submission, multi-task operation scheduling management, task monitoring and cluster fault alarm.

4. Experiment

Reasonable migration and placement of data is an important factor to ensure the performance of hierarchical storage system. This part verifies the feasibility of migration strategy and the impact of data migration on system performance through experiments. The migration strategy proposed in this paper, combined with the migration model based on data importance, evaluates the importance of all data files. And start the migration process, focus on the changes of water level, accuracy of data migration and hit rate of data access before and after data migration. Finally, compare the cost of memory Cloud Architecture and this architecture, and compare the performance-price ratio.

Migration strategy can dynamically migrate and place data according to user's I/O access. This section tests the effectiveness of migration strategy. First, the data of the system is initialized without considering any attributes of the data, as shown in Table 1. The memory cloud is 30G, the SSD is 20G, the HDD is 10G, and the backup of the memory cloud and SDD is 60G. Then the cycle parameter Z = 3 is set to trigger the migration condition and start the migration process to initialize the data placement.

Tab.1. Data changes before and after system migration

Medium	RAM	SSD	HDD
Data size	30/16.43	20/40.32	60/53.24
Space water saturation	53.57%/29.34%	8.93%/18%	10.71%/9.51%

Figure 3 depicts the start-up of migration process, the change trend of system data flowing up and down in all levels of media. Generally speaking, the water level of memory decreases and the amount of data decreases significantly. Similarly, the water level of SDD and HDD also shows a rising and falling curve and tends to be sTable. Compared with the data flow before and after migration, the spatial water level saturation of memory decreases by 24.23%, the spatial water level saturation of solid-state hard disk increases by 9.07%, and the spatial water level of mechanical hard disk decreases by 1.2%.

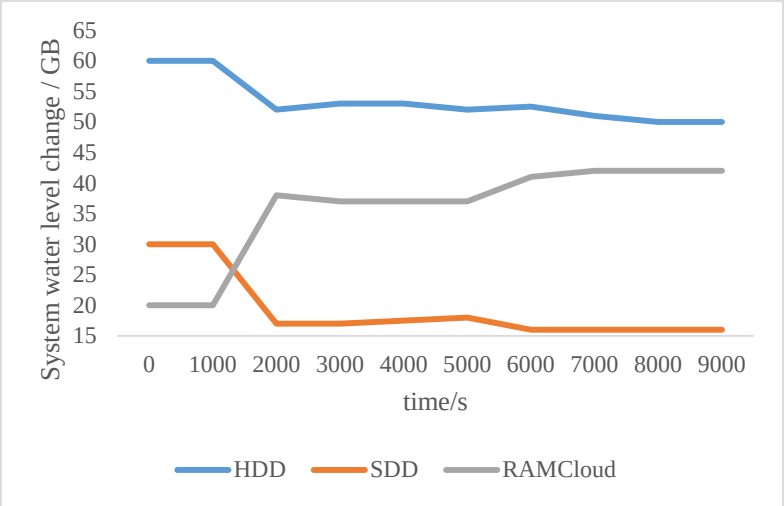

Fig.3. Storage capacity at all levels over time

Experiments show that the data of the system are migrated by the migration strategy in this paper, which shows that MMDS can evaluate the importance of data files according to their attribute parameters. The migration process can migrate the data adaptively according to the importance of the data, complete the hierarchical placement of the data, and satisfy the preliminary data migration. However, the selection of the migrated data set is affected by the attributes of the data itself and user access.

5. Conclusion

This paper mainly studies the migration mechanism of dynamic data migration in different servers in cloud computing environment. By analyzing the architecture of HDFS architecture, based on the existing research results and practical experience of industry, a high-availability, high-performance, fault-tolerant and self-recovery streaming data migration cloud service is designed according to the actual needs of enterprise data migration. Finally, the experiment proves the validity, accuracy and feasibility of this migration strategy. It reduces the storage pressure of memory servers, reduces the cost of building clusters, enlarges the storage capacity, and basically achieves the desired goal.

References

[1] Fonseca A, Cabral B. Prototyping a GPGPU Neural Network for Deep-Learning Big Data Analysis [J]. Big Data Research, 2017, 8:50-56.

[2] Yaghmazadeh N, Wang X, Dillig I. Automated migration of hierarchical data to relational Tables using programming-by-example[J]. Proceedings of the VLDB Endowment, 2018, 11(5): 580-593.

[3] Simonsen I, Lazzari R, Jupille J, et al. Optical response of supported particles[J]. Physics, 2017, 61(11):7722-7733.

[4] Forys U, Kheifetz Y, Kogan Y. Critical-point analysis for three-variable cancer angiogenesis models [J]. Mathematical Biosciences & Engineering Mbe, 2017, 2(3):511-525.

[5] Guerrero C, Lera I, Juiz C. Migration-Aware Genetic Optimization for MapReduce Scheduling and Replica Placement in Hadoop[J]. Journal of Grid Computing, 2018(2):1-20.

[6] Hu H, Wen Y, Chua T S, et al. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial [J]. IEEE Access, 2017, 2(1):652-687.

[7] Grm K, Štruc V, Artiges A, et al. Strengths and weaknesses of deep learning models for face recognition against image degradations[J]. Iet Biometrics, 2018, 7(1):81-89.

[8] Bunt C M, Castro-Santos T, Haro A. Reinforcement and validation of the analyses and conclusions related to fishway evaluation data from Bunt et al.:'performance of fish passage structures at upstream barriers to migration'[J]. River Research and Applications, 2016, 32(10): 2125-2137.

[9] Krishna J V, Naidu G A, Upadhayaya N. A Lion-Whale optimization-based migration of virtual machines for data centers in cloud computing[J]. International Journal of Communication Systems, 2018, 31(1):e3539.

[10] Chang V. Towards a Big Data system disaster recovery in a Private Cloud[J]. Ad Hoc Networks, 2015, 35: 65-82.

[11] Ma K, Dong F. Live data migration approach from relational Tables to schema-free collections with mapreduce[J]. International Journal of Services Technology and Management, 2015, 21(4-6): 318-335.