Design of Space Remote Sensing Data Storage Platform Based on Distributed File System

Di Li[1,2], Yizhun Peng[1,2*], Ruixiang Bai[1], Zhenjiang Chen[1,2], Lianchen Zhao[1,2]

*[1]College of Electronic Information and Automation, Tianjin University of Science and Technology,
Tianjin, 300222, China;*

*[2]Advanced Structural Integrity International Joint Research Centre, Tianjin University of Science and
Technology, Tianjin, 300222, China
E-mail: * pengyizhun@tust.edu.cn
www.tust.edu.cn*

Abstract

Due to the large space remote sensing data, a space remote sensing data has seven or eight hundred megabytes or more, and a large amount of space remote sensing data is generated every day for hundreds of GB, TB or even more, so a large amount of space is needed for storage. Space remote sensing data. In order to solve such problems, this paper prepares for the analysis of the subsequent space remote sensing data and combines the distributed file system to store the space remote sensing data. The installation of the CentOS 6.5 virtual machine through VMware to build an HDFS cluster, through a Namenode, three Datanode nodes to achieve access to space remote sensing data. Through the upload of the server, the space remote sensing data can be uploaded to the client, and the space remote sensing data can be downloaded through the client.

Keywords: Namenode, Datanode, HDFS, Space remote sensing data, Hadoop

1. Introduction

With the continuous development of the information age, data has shown an exponential growth, and the era of big data has emerged as a result. With the continuous progress of the times, the continuous development of science and technology, the rapid development of remote sensing technology is also quite rapid. Multi-sensors, multi-temporals, high spatial resolution, high spectral resolution of remote sensing data are increasing, and the data types are becoming more and more complex. . Among them, satellite remote sensing data is a layered digital image matrix, a data type with spatial characteristics, and an unstructured data, which cannot be represented by key-value pairs. Because traditional relational databases store structured data in the form of tables, they cannot store unstructured data such as remote sensing data[1].

The amount of space remote sensing data is huge, and the amount of data in a certain period of time in a certain area will reach 900 million or more. If comprehensive analysis and processing of space remote sensing data for multiple time periods are performed in a certain area, the amount of data will increase in the form of geometric multiples, and the data will reach hundreds of GB or even more. Capacity cannot afford the storage of massive data. Therefore, in recent years, with the continuous development of big data, distributed storage technology has gradually matured and improved, and distributed systems such as Hadoop and Spark have emerged as the times require. The distributed system uses multiple independent server nodes connected together to form a physically distributed, logically unified computer cluster distributed data storage system under the unified scheduling of the master node server, which can solve the multiple disadvantages of single-machine storage. Provides an effective and secure method for mass data storage.

2. Introduction of related technologies

2.1. Distributed Architecture Hadoop

Distributed architecture Hadoop can be built into clusters with common computer configuration. As the basic platform of cloud computing, Hadoop mainly consists of three modules: HDFS, MapReduce and Yarn. HDFS is a distributed file system, mainly serving as distributed storage in clusters. Function to achieve distributed storage of big data. MapReduce is a distributed computing programming framework whose main function is to implement distributed parallel computing in a cluster. Yarn distributed resource scheduling platform, the main function is to help users call a large number of MapReduce programs, and allocate computing resources reasonably. HDFS provides support for reading and writing files during MapReduce task processing[2]. MapReduce implements task distribution, tracking, execution, and collection of results based on HDFS. The two functions interact with each other to complete the core tasks of Hadoop cluster. Hadoop can freely organize computer resources, build a distributed cloud computing platform, and make full use of the computing and storage capabilities of the cluster to complete the storage of massive data.

The advantages of Hadoop clusters are as follows:

- Hadoop clusters can be scaled horizontally. When the data is too large and too large, the cluster can not bear the pressure. The cluster storage capacity can be directly expanded by adding nodes. Dynamic data movement can reduce the pressure on each node.
- Hadoop cluster adopts master-slave architecture. Nodes are divided into two categories: Namenode is the main node responsible for storing cluster metadata, which plays a role in supervising the execution of MapReduce. Datanode is the child node responsible for storing specific data. Perform specific tasks to keep the heartbeat with the primary node.

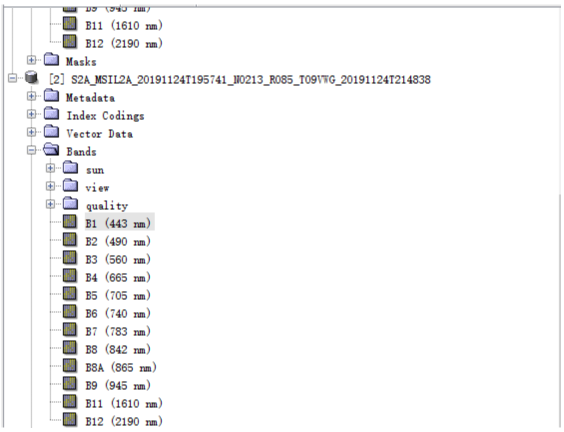
2.2. Distributed File System HDFS

The distributed file system HDFS has the same characteristics as the ordinary file system, and (1) has a directory structure. (2) All files stored in the system are files. (3) The system provides functions such as copying, moving, creating, deleting, modifying, and viewing files. The distributed file system and the stand-alone file system are different. The file system stored in a single machine is only in the operating system of one machine, and the distributed file system spans multiple machines. A single file is placed on a single machine's disk, while a distributed file system stores files on multiple machines. The working mechanism of the distributed file system is: when the client stores a file to the distributed file system, the distributed file system cuts and blocks the stored file, and stores the diced in the child nodes in the cluster. On the disk; once the file is cut and stored in the distributed file system, there is a mechanism for recording the dicing information of each file stored by the user, and dicing the specific storage path; in order to ensure the security of the data Sex, to ensure that data will not be lost, the distributed file system will store multiple backups of each file in the cluster to prevent data loss when a server hangs. In general, a distributed file system consists of a primary node server and N child node servers.

3. Space remote sensing data storage

3.1. Characteristics of space remote sensing data

Aerospace remote sensing data is taken by space satellites. Remote sensing satellite data is used by remote sensing satellites to detect the reflection of electromagnetic waves on the Earth 's surface objects in space and the electromagnetic waves emitted by them, so as to extract information about the object, complete the identification of objects at long distances, and convert these electromagnetic waves. The visible image is recognized as the satellite image[3]. Space remote sensing data is difficult to represent with key-value key-value pairs, which is an unstructured data. The following figure shows the composition of space remote sensing data.



```
        B9 (945 nm)
        B11 (1610 nm)
        B12 (2190 nm)
    Masks
[2] S2A_MSIL2A_20191124T195741_N0213_R085_T09VWG_20191124T214838
    Metadata
    Index Codings
    Vector Data
    Bands
        sun
        view
        quality
        B1 (443 nm)
        B2 (490 nm)
        B3 (560 nm)
        B4 (665 nm)
        B5 (705 nm)
        B6 (740 nm)
        B7 (783 nm)
        B8 (842 nm)
        B8A (865 nm)
        B9 (945 nm)
        B11 (1610 nm)
        B12 (2190 nm)
```

Fig.1. Composition of space remote sensing data

The above figure shows the space remote sensing data at a certain time in a certain area[4]. The specific space remote sensing data is shown below:

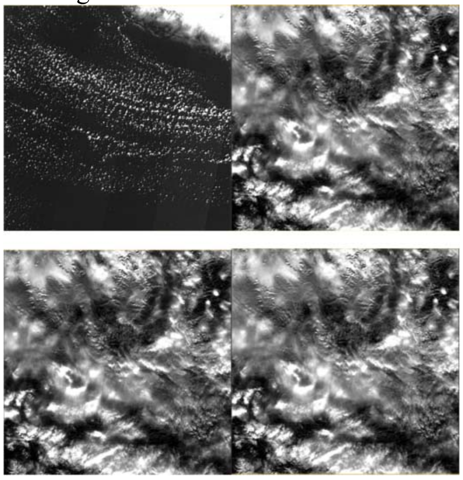


Fig.2. Specific display of space remote sensing data

3.2. Distributed file system setup

According to the characteristics of space remote sensing data, relevant space remote sensing data will be stored in the HDFS distributed file system. The system will build a distributed file system based on Hadoop cluster to store space remote sensing data. The distributed file system has four nodes, which are a master node Namenode node and three child nodes Datanode nodes, among which the master node It is responsible for recording the location of the stored file partition and the node where the chunk backup is located. The main task of the child node is to store the specific block[5].

3.3. Cluster architecture design

Using four ordinary computers, using a local area network to form a Hadoop cluster, you can use a common computer as a client for the client to log in to the client to access the cluster. The specific architecture is shown in the figure:

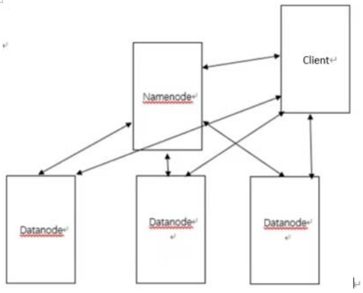


Fig.3. Block diagram of hadoop cluster architecture

The configuration of each computer is the same, the processor AMD Ryzen 3 2200G with Radeon Vega Graphics 3.50Ghz, memory 16g, hard disk 1T.

3.4. Building a clustered software environment

Install VMware Workstation Pro15 on the Windows host and four virtual machines on the VMware Workstation Pro15. The operating system is all Centos6.5, the JDK version is jdk1.8.0_212, and the Hadoop version is hadoop-2.8.5.

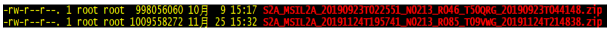
In a Hadoop cluster, one virtual machine is used as the primary node Namenode node, and the other three are used as child node Datanode nodes. The Master node IP is 191.168.220.30 and the NameNode and Secondarynamenode are installed. The Slave1 node IP is 191.168.220.31 and the DataNode is installed. The Slave2 node IP is 191.168.220.32. The DataNode is installed. The Slave3 node IP is 191.168.220.33[6].

The main steps in building a distributed file system cluster:

- Modify the machine's host name and specific IP address to configure the machine's host name to the Windows local domain name mapping file.
- Configure the basic software environment of the Linux server. For example, turn off the firewall and disable it. Install the JDK to configure its environment variables and the domain name mapping configuration of the hosts in the cluster.
- Modify the configuration file, specify the default file system as: hdfs, the primary node that specifies hdfs is the machine, the local directory that specifies the namenode storage metadata, and the local directory where the datandoe storage folder is specified.
- Start HDFS. First, you need to initialize the metadata directory of the namenode.

4. Experimental results

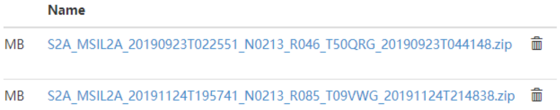
By looking at the server root directory to know that there is aerospace remote sensing data, it is uploaded to the distributed file system and stored in the spacedata folder by the instruction hdfs dfs -put /S2A_MSIL2A_20191124T195741_N0213_R085_T09VWG_20191124T214838.zip /spacedata.



```
-rw-r--r--. 1 root root  998056060 10月  9 15:17 S2A_MSIL2A_20190923T022551_N0213_R046_T50QRG_20190923T044148.zip
-rw-r--r--. 1 root root 1009558272 11月 25 15:32 S2A_MSIL2A_20191124T195741_N0213_R085_T09VWG_20191124T214838.zip
```

Fig.4. Where the remote sensing data is located

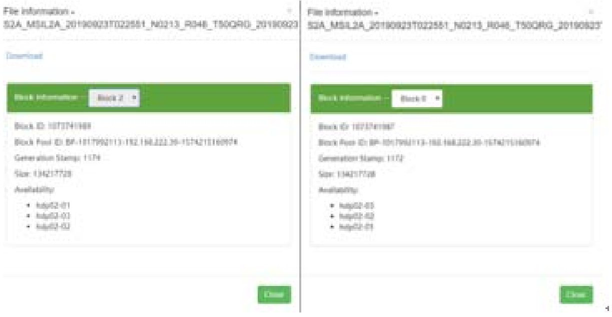
Can check the storage location and backup status of the aerospace remote sensing data through the client. Enter the client IP address and log in to the /spacedata under the client's Browse Directory to see the stored space remote sensing data, as shown in the following figure：



	Name
MB	S2A_MSIL2A_20190923T022551_N0213_R046_T50QRG_20190923T044148.zip
MB	S2A_MSIL2A_20191124T195741_N0213_R085_T09VWG_20191124T214838.zip

Fig.5. Remote sensing data is stored in hdfs

Click on one of the data to observe the size of the data and the location of the server where the backup is located, and the data can be downloaded through this page.



File information - S2A_MSIL2A_20190923T022551_N0213_R046_T50QRG_20190923

Download

Block information -- Block 2

Block ID: 1073741969

Block Pool ID: BP-1017992113-192.168.222.35-1574213160974

Generation Stamp: 1174

Size: 134217728

Availability:

- hdp02-01
- hdp02-03
- hdp02-02

Close

File information - S2A_MSIL2A_20190923T022551_N0213_R046_T50QRG_20190923

Download

Block information -- Block 0

Block ID: 1073741967

Block Pool ID: BP-1017992113-192.168.222.35-1574213160974

Generation Stamp: 1172

Size: 134217728

Availability:

- hdp02-03
- hdp02-02
- hdp02-01

Close

Fig.6. Block information of remote sensing data

Summary

Space remote sensing data is very important data and has important research significance for national geomorphology. Space remote sensing data is unstructured data so traditional relational databases cannot satisfy its storage. The storage of HDFS-based space remote sensing data is good for future data retrieval. The foundation is convenient for future research work.

References

1. Cheng Li, *Research on Distributed Storage of Remote Sensing Data Based on Hadoop*. (Shandong, Shandong Agricultural University, 2018)
2. Dazhi Wang, Research on Cross-Cluster Distributed File System Based on HDFS. *Information Technology and Informatization*, 2019 (8).
3. Zhongyi Chen, Distributed File System Based on Hadoop. *Electronic Technology and Software Engineering*, 2017 (9): 175-175.
4. Huan Yan, *Research and Implementation of Parallel Index Technology for Aerospace Information System*. (Xi'an, Xidian University, 2018)
5. Fanjun Meng, Wei Cao, Zhiqiang Guan, HIS-based Distributed Storage of AIS Data. *Information and Communications*, 2016 (5): 172-174.
6. Lixuan Chen, Shiyu Du, Chenlin Huang, et al. Design and implementation of teaching cloud platform based on distributed file system. *Wireless Internet Technology*, 2019 (9): 94-9