SUGISAKA MASANORI

Journal of Advances in Artificial Life Robotics
Vol. 1(4); March (2021), pp. 214–218
ON LINE ISSN 2435-8061; ISSN-L 2435-8061
https://alife-robotics.org/jallr.html

Research Article

Design of Distributed Remote Sensing Data Storage System based on Hadoop Framework

Lianchen Zhao[1,2], Yizhun Peng[1,2], Di Li[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*

ARTICLE INFO

Article History
Received 09 November 2019
Accepted 16 February 2021

Keywords
Hadoop
Name node
HDFS
Space remote sensing data

ABSTRACT
With the progress of the times, remote sensing technology has become more and more mature, and remote sensing data has become more and more complex with the continuous progress of technology. Remote sensing data has the characteristics of unstructured and large amount of data, so the distributed system can be used to store it. In this paper, according to the characteristics of space remote sensing data, we build a distributed storage system for space remote sensing data. According to the unstructured characteristics of space remote sensing data, distributed storage is used to store the data in virtual environment Hadoop distributed cluster is built to realize the distributed storage of space remote sensing data.

1. Introduction

With the continuous progress of the times, remote sensing technology is developing with the development of the times. The amount of multi-sensor, multi temporal, high spatial resolution and high spectral resolution remote sensing data is also increasing, and the types of remote sensing data are becoming more and more complex with the development of remote sensing technology. Remote sensing technology started in the 1960s. The theoretical basis of remote sensing technology is electromagnetic wave theory. It uses sensing technology to radiate the surface of the object, then collects information, processes the radiation information, and finally images the remote sensing image.

Remote sensing image includes non-imaging hyperspectral reflectance data, satellite remote sensing image data, research area attribute data and so on. The non-imaging hyperspectral reflectance data belongs to structured data, which can be represented by K-V key value pairs. However, satellite remote sensing data is a layered digital image matrix, and its data characteristics are high-dimensional spatial characteristics, so it cannot be represented by K-V model, and its data type belongs to unstructured data. The traditional relational database is mainly used to store structured data, such as Oracle, My SQL, SQLServer and other traditional relational databases for remote sensing data, which is an unstructured data type, storage efficiency is very low.

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

Corresponding author's E-mail: pengyizhun@tust.edu.cn URL: www.tust.edu.cn

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 [1] 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 cannot 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: Name node is the main node responsible for storing cluster metadata, which plays a role in supervising the execution of MapReduce. Data node 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

HDFS distributed file system is the core of the whole Hadoop. It is a distributed file system with high fault tolerance and can be built on cheap machines. HDFS provides efficient data access for the whole distributed system, so HDFS is suitable for remote sensing data processing.

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 hang. 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 pairs, which is an unstructured data. The Figure 1 shows the composition of space remote sensing data.

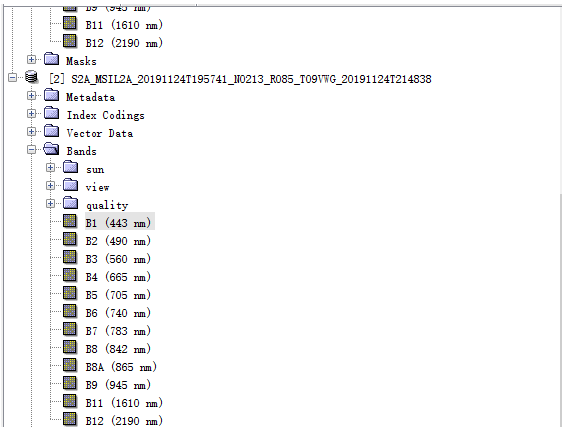

Fig.1. Composition of space remote sensing data

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 in the Figure 2:

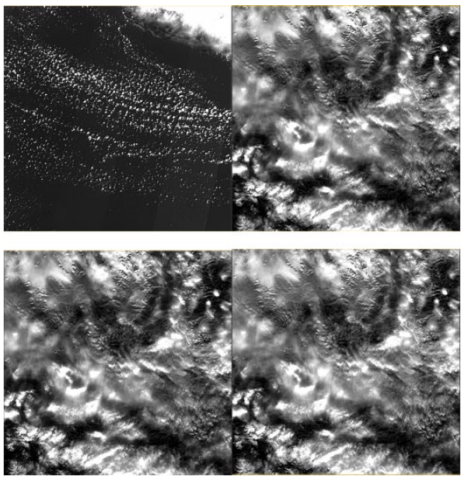

Fig.2. Specific display of space remote sensing data

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 Name node node and three child nodes data node 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]. The distributed remote sensing data storage system is mainly divided into five layers, which are object layer, computer virtual layer, storage layer, data access and operation layer, and data display layer from bottom to top.

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 3:

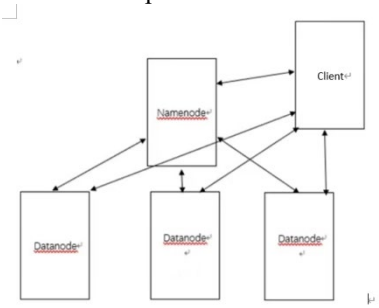

Fig.3. Block diagram of Hadoop cluster architecture

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 Name node node, and the other three are used as child node data node nodes. The Master node IP is 191.168.220.30 and the Name Node and Secondary name node are installed. The Slave1 node IP is 191.168.220.31 and the data node is installed. The Slave2 node IP is 191.168.220.32. The data node 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

name node storage metadata, and the local directory where the data node storage folder is specified.

- Start HDFS. First, you need to initialize the metadata directory of the name node.

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 space data folder by the instruction hdfs -put /S2A_MSIL2A_20191124T195741_N0213_R085_T09VWG_20191124T214838.zip /space data. It is shown in the Figure 4 where the remote sensing data is located.

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 /space data under the client's Browse Directory to see the stored space remote sensing data, as shown in the Figure 5 :

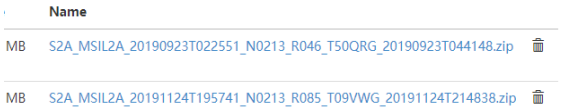

Fig.5. Remote sensing data is stored in hdfs

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, it is shown in the Figure 6.

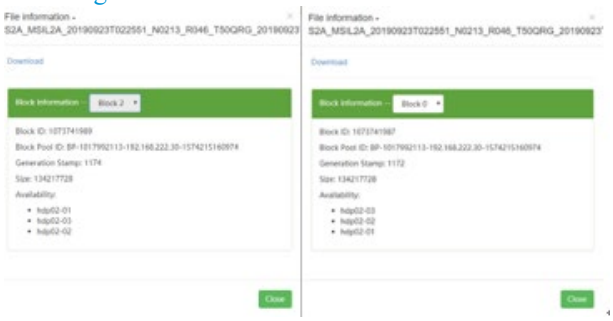

Fig.6. Block information of remote sensing data

5. 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. This paper first introduces Hadoop framework, and then uses desktop computer as the basis of hardware based on virtualization technology. Based on the above conditions, the distributed remote sensing data storage system under Hadoop framework is successfully built.

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, Distributed storage of AIS data based on HBase. *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

Authors Introduction

Mr. Lianchen Zhao

He is currently the master course student in Tianjin University of Science & Technology. His research field is artificial intelligent. His research direction is object detection and deep neural network.

Dr. Yizhun Peng

He is an Associate Professor in Tianjin University of Science & Technology. He received a doctor's degree in control theory and control engineering from the Institute of Automation Chinese Academy of Sciences, in 2006.His research field is intelligent robot and intelligent control.

Mr. Di Li

He is a master graduated in Tianjin University of Science & Technology. His research field is artificial intelligent. His research direction is object detection and data handling.