Research on the Construction of the "Hive Big Data Warehouse" Experimental Teaching Environment

Huatao Zhou[1a] , Yuan Yao[1b], Fang Tu[1*],

[a] zhouhuatao@163.com, [b] 30038238@qq.com, *382935224@qq.com

[1]School of information engineering, Wuchang Institute of Technology, Wuhan, Hubei

Abstract. The construction of experimental teaching environment for Hive course in big data warehouse was studied. Taking into account the limitations of existing traditional computer rooms and current technical conditions, it is proposed to install Linux system and Hadoop high availability cluster in a VMware virtual machine software environment. Then, a big data warehouse Hive is installed on the Hadoop high availability cluster platform, and a construction plan for a big data warehouse Hive experimental teaching environment based on Hadoop+Hive is implemented. The experimental environment of the Big Data Warehouse Hive course helps to improve the theoretical teaching effect of the Big Data Warehouse Hive course, and helps students better master the data warehouse tool Hive for data analysis and data mining from practice, achieving the goal of cultivating applied and innovative talents.

Keywords: experimental teaching; big data; Data warehouse; Hadoop; Hive

1 Introduction

Today, big data technology, as a new information technology that combines computer technology and network technology, has been widely applied to people's various work and life, profoundly affecting people's clothing, food, housing, and transportation activities [1][2][3][4]. With the arrival of the big data era, big data has become an important force driving the upgrading of existing industries and the birth of new industries. As the main force in cultivating big data talents, universities have a certain foundation in the construction of big data majors and talent cultivation. As an applied undergraduate private university in Hubei Province, our school attaches great importance to students' ability to apply computer and big data technology to solve practical problems. Therefore, experimental teaching is a very important part of our school's teaching. Through experimental teaching, teachers can guide and help students complete experimental content face-to-face and hand in hand in the laboratory. Through experiments, students can exercise their ability to analyze and solve problems, Improve the level and ability of big data application.

"Big Data Warehouse Hive" is an elective course for the data science and big data technology majors of information engineering College. The experimental teaching environment of the course is based on the Hadoop and Hive platforms. Hive[5][6][7][8]is a data warehouse built on the Hadoop[9][10][11][12][13]file system. It provides a series of tools to extract, transform, and load (ETL) data stored in HDFS A tool for querying and analyzing large-scale data stored in Hadoop. Hive defines a simple SQL like query language (HQL) that can map structured data files into a

ICEKIM 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2337303

data table, allowing users familiar with SQL to query data, and developers familiar with MapReduce[14][15][16][17][18]to develop mappers and reducers to handle complex analysis work. Compared to MapReduce, Hive has more advantages[19].

2 Experimental Environment Design

The Hive data warehouse is developed based on the Hadoop platform, so the experimental environment is Linux, Hadoop, and Hive software. The computers in the laboratory room of our school are installed with teaching software based on the Windows operating system, which does not meet the experimental teaching environment of this course. Considering the practicality, ease of use, and operability of the experimental software, this experiment first installed VMware Workstation virtualization software on the Windows operating system of the computer room, then installed Linux operating system on the VMware virtualization software, and finally installed software such as Hadoop and Hive on the Linux operating system, solving the problem that the existing experimental environment in the experimental room cannot meet the experimental teaching environment, Students are able to comprehensively and quickly grasp the usage methods and development techniques of the big data warehouse Hive in experimental teaching, and use Hive tools for data analysis and data mining, improving their hands-on ability and cultivating their practical spirit, which is conducive to the cultivation and output of applied undergraduate talents in schools.

2.1 Software Version Selection

The software used in this experiment is stable, reliable, and widely used in actual production environments. The software list is shown in Table 1.

Table 1. Software Version Table

Software Name	edition	purpose
VMware Workstation	16 Pro	Virtual machine software
CentOS	six point seven	Linux operating system
SecureCRT	nine point one	SSH remote connection
Xftp	seven	Windows uploads files to Linux
JDK	8u161(Linux x64)	Java Development Kit
Zookeeper	3.4.10	Distributed Application Coordination Service
Hadoop	2.7.4	Distributed Storage HDFS and Distributed Computing MapReduce
MySQL	five point seven	Hive metadata storage
Hive	2.3.7	Data Warehouse Tools

2.2 Host Network Configuration Planning

Three Linux virtual machines can be installed using Vmware Workstation virtual machine software. These three virtual machines and laptops are in the same network segment, and the network is interconnected. In this experiment, three computers were used to form a Hadoop high

availability cluster. The laptop, as a client, can connect three virtual machines to upload installation software and test the Hadoop high availability cluster function. The host network configuration plan is shown in Table 2.

Table 2. Host Network Configuration Planning Table

Host Name	IP address	host name	Subnet mask	gateway	DNS1
hadoop01	192.168.121.134	hadoop01	255.255.255.0	192.168.121.2	8.8.8.8
hadoop02	192.168.121.135	hadoop02	255.255.255.0	192.168.121.2	8.8.8.8
hadoop03	192.168.121.136	hadoop03	255.255.255.0	192.168.121.2	8.8.8.8
client	192.168.121.5	client	255.255.255.0	192.168.121.2	8.8.8.8

2.3 Hadoop High Availability Cluster Node Configuration

Hadoop high availability cluster consists of three virtual machines. The high availability cluster is reflected in that both the NameNode node and the ResourceManager node are composed of two nodes, one as the primary node and the other as the standby node, to avoid single point of failure. Both hadoop01 and hadoop02 serve as NameNode and ResourceManager nodes, with one as the primary node active and one as the standby node standby. The three virtual machines serve as DataNode, NodeManager, JournalNode, and Zookeeper nodes respectively. Both hadoop01 and hadoop02 run the ZKFC process to monitor the status information of the primary nodes of the NameNode and ResourceManager of the two virtual machines. Once the primary node fails, the standby node immediately acts as the primary node, Avoid single point of failure, as shown in Table 3.

Table 3. Hadoop High Availability Cluster Node Configuration Table

host name	Name Node	Data Node	Resource Manager	Node Manager	Journal Node	Zookeeper	ZKFC
hadoop01	√	√	√	√	√	√	√
hadoop02	√	√	√	√	√	√	√
hadoop03		√		√	√	√	

2.4 Hive Data Warehouse Functional Topology Map

Hadoop01 installs MySQL database and Hive software as the Hive data warehouse server. The Hive data warehouse functional topology diagram is shown in Figure 1.

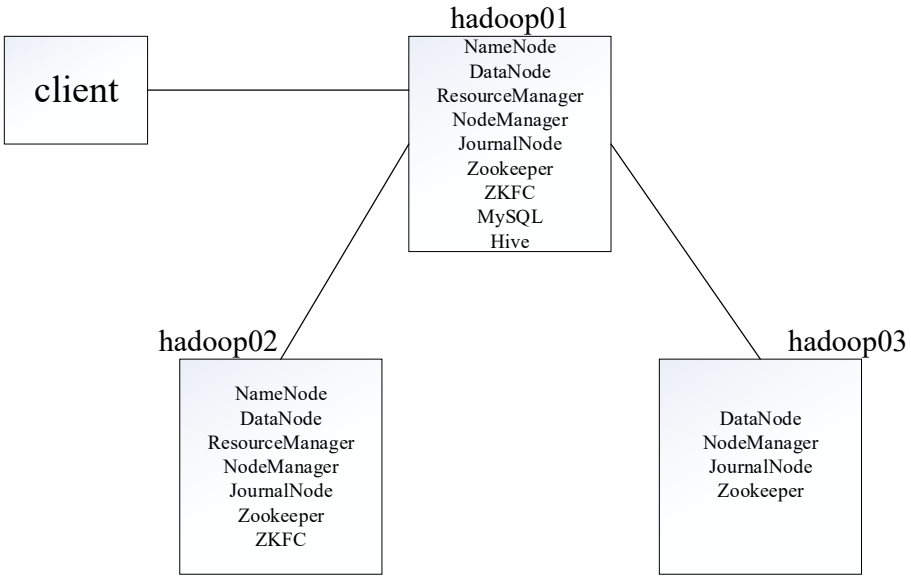

Fig. 1. Hive Data Warehouse Functional Topology Diagram

3 Preparation of experimental environment

3.1 Linux environment setup

Install VMware Workstation 16 Pro software on the Windows operating system environment of the computer, use this software to create three virtual machines, install CentOS 6.7 operating system respectively, configure network parameters according to Table 2, and ensure that the three virtual machines can connect to the internet through the laptop in NAT mode, as three Linux virtual machines are standby.

3.2 JDK deployment

1. Download JDK

2. Upload JDK installation package

3. Install JDK

4. Configure JDK environment variables.

5. JDK environment validation

6. Distribute JDK related files

3.3 Zookeeper Deployment

Zookeeper is an open-source framework for distributed coordination services, implemented by Google's Chubby open-source. Zookeeper is mainly used to solve the consistency problem and single point of failure problem of the application system in the distributed cluster. It can solve

the single point of failure problem of the NameNode and ResourceManager in the Hadoop cluster in this experiment, improve the high availability of the cluster, and provide fault tolerance and backup mechanisms.

4 Hadoop High Availability Cluster Construction and Testing

There are three deployment methods for Hadoop clusters, namely standalone mode, pseudo distributed mode, and cluster mode. Independent mode and pseudo distributed mode are mainly used for learning and debugging, and fully distributed mode is usually used in actual production environments. In order to improve the high availability of Hadoop clusters, ZooKeeper is usually used to provide automatic failover and data consistency services for Hadoop clusters. The architecture diagram of Hadoop high availability clusters is shown in Figure 2.

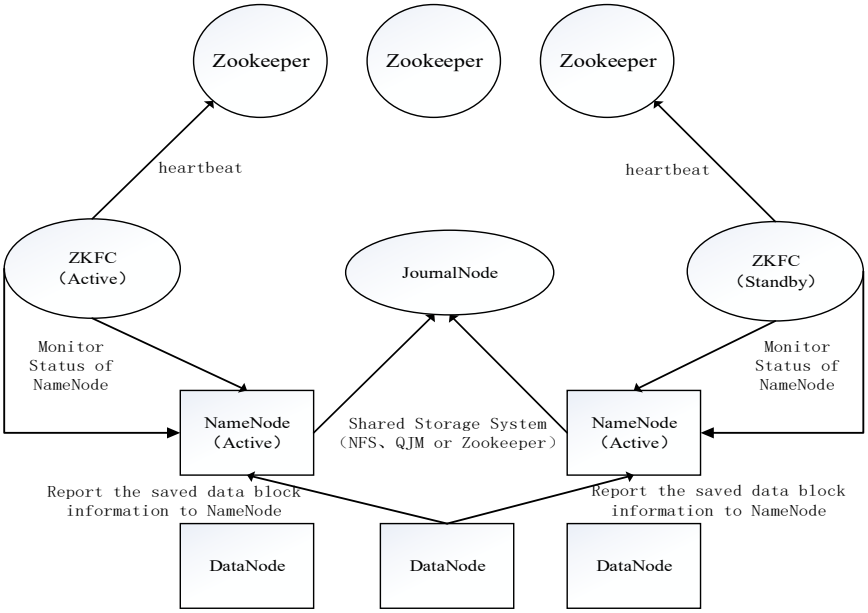

Fig. 2. Hadoop High Availability Cluster Architecture Diagram

4.1 Installing Hadoop

Upload Hadoop installation package and install Hadoop.

4.2 Configuring Hadoop High Availability Clusters

Set the Hadoop configuration files separately, as shown in Table 4

Table 4. Hadoop Configuration Files

configuration file	Function Description
hadoop-env.sh	Configure the environment variables required for Hadoop operation (ensure the normal operation of the HDFS cluster daemon)
yarn-env.sh	Configure the environment variables required for Yarn operation (ensure the normal operation of the daemon of Yarn, ResourceManager, NodeManager)
core-site.xml	Hadoop core global configuration file, which can be referenced in other configuration files
hdfs-site.xml	HDFS configuration file, inheriting the core-site.xml configuration file
mapred-site.xml	MapReduce configuration file, inheriting the core-site.xml configuration file
yarn-site.xml	Yarn configuration file, inheriting the core-site.xml configuration file

5 Deployment of Hive Data Warehouse

There are three deployment modes for Hive data warehouse: embedded mode, local mode, and remote mode. Let's take local mode as an example to build an experimental environment for Hive data warehouse.

Local mode deployment essentially replaces Hive's default metadata storage medium with an embedded Derby database and a standalone database, namely a MySQL database. Deploying Hive in local mode requires both MySQL and Hive to be installed on a virtual machine. Next, taking virtual machine hadoop01 as an example, we will deploy Hive in local mode.

5.1 Install MySQL

To install MySQL version 5.7 online, it is necessary to ensure that the virtual machine can connect to the external network.

5.2 Start MySQL service

After the installation of MySQL is completed, execute the "system ctl start mysqld. service" command to start the MySQL service. After the MySQL service is started, execute the "system ctl status mysqld. service" command to view the running status of the MySQL service.

5.3 Install Hive

Install Hive by decompressing and installing it in the directory such as “/export/servers/”,where the application is stored.

5.4 Configure Hive

Configure two files of hive env.sh and hive site.xml.

6 Conclusion

The major of Data Science and Big Data Technology is an emerging field in today's universities, and many fields and contents require teachers to discover and understand. At the same time, many practical problems such as the limitations of traditional computer rooms will also arise. This article aims to build a complete and usable experimental teaching environment for the big data warehouse Hive by installing Linux, Hadoop, and Hive software in a VMware virtual machine environment, Enable teachers to apply theoretical knowledge to experimental teaching, enable students to combine theory and practice in experimental teaching, deepen their understanding and mastery of Hive theoretical knowledge in big data warehouses, improve students' engineering practical abilities, and cultivate more and better application-oriented undergraduate talents for the motherland.

Acknowledgment.2021 Hubei Provincial Department of Education Philosophy and Social Science Research Guidance Project (21G155); Supported by the Doctoral Research Initiation Fund of Wuchang Institute of Technology (2021BSJ02)

References

[1] Tan, R.H., Chen, Y., Cai, W.T., Dian, Y.Y.(2023) Research on Building a Big Data Application Platform for Rural Landscape Resources. Journal of Huazhong Agricultural University: 1-9 . http://kns.cnki.net/kcms/detail/42.1181.S.20230322.1339.002.html.

[2] Wang, J.L., Gao, P.J., Zhang, J., Wang, L.H.(2023) Overview of Manufacturing Big Data Analysis: Connotation, Methods, Applications, and Trends. Journal of Mechanical Engineering: 1-16. http://kns.cnki.net/kcms/detail/11.2187.TH.20230309.1714.038.html.

[3] Luo, T.Y., Xie, K., Liu, Y.(2023) Big data-driven interactive innovation recommendation system for enterprises and users and its application. Journal of Beijing Jiaotong University (Social Science Edition): 1-13. DOI: 10.16797/j.cnki.11-5224/c.2020303.006.

[4] Ma, R.K.(2023) Design and Application of a New Distribution Network Power Supply Service Command System Based on Big Data Technology. Journal of Power Systems and Automation: 1-10. DOI: 10.19635/j.cnki.csu-espa.001230.

[5] Chen, L., Chen, X.S., Luo, Y.G., Yang, L., Yuan, D.H.(2022) Hive data operation compliance analysis method based on subgraph isomorphism. Journal of Electronics and Information Technology, 44 (12): 4367-4375.

[6] Tang, D.Y., Han, W.L.(2022) SELive: Hive Mandatory Access Control Model and Implementation Based on Type Enhancement. Computer Application and Software, 39 (07): 281-286+294.

[7] Xu, Z.H., Wang, Y.Z., Wang, L.Q., Dong, Y.F.(2020) Hive based method for mining the origin and destination of massive bus passenger flow. Science and Technology and Engineering, 20 (20): 8300-8309.

[8] Wang, H.J., Li, J.H., Shen, Z.H., Zhou, Y.C.(2018) Hive Join query reducer load balancing method based on ORC metadata. Computer Science, 45 (03): 160-166.

[9] Liu, Y., Wang, J., Tang, M., Zhang, Y.D.(2023) A Hybrid Neural Network Load Classif ication Model Based on Hadoop Distributed Computing. Science and Technology and Engineerin g, 23 (04): 1549-1556.
[10] Chang, W.P., Yuan, Q.(2022) An Intelligent Detection Method for Abnormal Nodes in Dynamic Networks Based on Hadoop. Computer Simulation, 39 (11): 402-405+462.
[11] Yu, X.R., Fan, J.J.(2022) Hadoop based recommendation algorithm design for heterogene ous network collaborative filtering. Information Network Security, 22 (10): 91-97.
[12] Tian, B.B., Tian, C., Zhou, Y.H., Chen, G.H., Dou, W.C.(2022) Scheduling algorithm fo r reducing network Head-of-line blocking in Hadoop cluster. Computer Science, 49 (03): 11-22.
[13] Cai, Y.N., Bao, X.Y., Lin, Y.K., Peng, J.X., Peng, Z.B., Lin, Y.Q., Li. J.L., Guo, Y.(20 22) Research on a Parallel Named Entity Recognition Model Based on Hadoop. Experimental Te chnology and Management, 39 (02): 7-12+39. DOI: 10.16791/j.cnki. sjg. 2022.02.02.
[14] Feng, Y.H., Wu, K.H., Huang, Z.H., Feng, Y.Z., Chen, H.H., Bai, J.C., Ming, Z.(2023) Set similarity self connection algorithm based on FP tree and MapReduce. Computer Research a nd Development: 1-18.http://kns.cnki.net/kcms/detail/11.1777.tp.20230310.0925.004.html.
[15] Liu, W.M., Cui, Y., Mao, Y.M., Liu, W.(2022) Parallel K-means algorithm based on Ma pReduce and MSSA. Computer Application Research, 39 (11): 3244-3251+3257. DOI: 10.19734/ j.issn.1001-3695.2022.04.0149.
[16] Du, J., Zhang, Z., Cao, J.C.(2021) MapReduce load balancing method using fast unbiase d stratified graph sampling algorithm. Computer Application and Software,38 (11): 288-294+313.
[17] Huang, X.Y., Xiang, C., Tao, T.(2021) A Partitioning and Clustering Algorithm Based o n MapReduce and Improved Density Peaks. Computer Application Research, 38 (10): 2988-2993 +3024. DOI: 10.19734/j.issn.1001-3695.2021.03.0093.
[18] Tao, T., Mao, Y.M.(2021) Parallel partition clustering algorithm based on MapReduce an d improved artificial bee colony algorithm. Science and Technology and Engineering,21(21): 898 9-8998.
[19] Black, H.P.(2019) Principles and Applications of Hadoop Big Data Technology. Tsinghu a University Press, Beijing.