A Lightweight Experimental Platform for Big Data Based on Docker Containers

Gu Ruijun¹,²

¹Jiangsu Key Laboratory of Data Science and Smart Software, Jinling Institute of Technology, Nanjing 211169, China
²School of Computer Engineering, Jinling Institute of Technology, Nanjing 211169, China

*Corresponding author’s e-mail: guruijun@jit.edu.cn

Abstract. In recent years, many colleges and universities have set up the major of big data. The biggest problem in teaching is that there is no supporting basic experimental environment, and it is difficult to deploy and configure the big data environment at the same time. In addition, the lack of experimental data, experimental teaching plans and experimental manuals in the experimental process makes it difficult to carry out relevant teaching. In order to reduce the cost of laboratory construction and the difficulty of learning big data, a lightweight big data experimental platform was constructed based on virtualized container technology. Through this platform, we can create a big data cluster, provide various suitable experimental environments, focus on the technology itself, and greatly improve the learning efficiency.

1. Introduction

In recent years, many colleges and universities have set up the specialty of big data. The biggest problem in teaching is that there is no supporting basic experimental environment, and it is difficult to deploy and configure the big data environment quickly [1]. At present, there are three main solutions [2]: (1) Students build pseudo-distributed experimental environment through personal computers to simulate big data cluster. Due to the performance of personal computers, data processing capacity is limited, and cannot achieve better learning results; (2) Through OpenStack and other virtualization technology, big data cluster is constructed at the server side, and a dedicated person is assigned to be responsible for maintenance and management. The disadvantage is that the resource consumption is large and the cost of maintenance and management is high. (3) Rent a third-party cloud service platform (such as Aliyun) to deploy the experimental environment. The system is stable and has high performance, but it cannot customize the personalized environment, and the rental cost is high.

In order to reduce the cost of laboratory construction and the difficulty of learning big data, lightweight Docker containerization technology is used to provide users with a virtualized operating environment, which can ensure that the environment for users to carry out the same experiment is the same, and that each user's learning virtual environment is independent and does not interfere with each other. Because Docker uses virtualization technology at the operating system level, it has a higher utilization rate of system resources such as memory and CPU, and can support more users to use online at the same time under the same server cluster conditions.
2. Relevant works

2.1 Relevant Research
In the public cloud, many companies based on open source big data technology, provide big data platform solutions based on the cloud computing[3], including Amazon cloud AWS (Amazon Web Service), Microsoft cloud (Azure) Hadoop, Google cloud platform big data products, Aliyun big data service (MaxCompute), Tencent Big Data Suite and Baidu Big Data Base Suite. Solutions in the cloud are scalable and can meet the changing needs of users. For example, Aliyun provides real Alisyun environment through sandbox technology, and can experience and learn Aliyun products and services in depth. There are a large number of experimental projects on the platform, covering cloud computing, big data, cloud security and other fields, and providing detailed experimental documents and tips to guide. But relatively speaking, the cost is higher than the internal platform of non-profit organizations such as universities, and the intellectual property issues involved are more stringent, so it is more difficult to open and share data. Therefore, it is the most mainstream way to build virtual laboratories by virtualization technology and utilizing public resources of universities.

At present, most of the laboratories based on virtualization technology use OpenStack technology to virtualize the server resources, and then allocate them to students to build clusters. In Paper [4], OpenStack technology and IaaS layer architecture were adopted to centralize the hosting of resources, exchange resources, make full use of hardware resources, and provided expansible and invisible infrastructure for teachers and students. After clouding the laboratory environment, the hardware resources needed for various experiments were allocated on demand, which makes the teaching work focus on professional learning rather than on the construction and configuration of the environment. Similarly, Yao Qimin [5] used OpenStack to build a big data laboratory, which achieved the goal of open sharing of laboratory resources without time and space constraints. In addition, by optimizing laboratory management, the cost of laboratory construction was reduced and the utilization rate of laboratory teaching resources was improved. Wang Yongkun and others [3] try to build a flexible data platform using open source software design. The data platform was mainly built on the physical server to ensure the analysis performance. When resources are scarce, it can be dynamically extended to private cloud platforms. The bottom layer of private cloud platform adopts OpenStack framework to realize the functions of software definition network, elastic computing and software definition storage.

The virtualization technology based on OpenStack cannot reach the level of real physical host in resource utilization, and the configuration of files is not flexible enough. As an open source lightweight application container engine, Docker can package applications into a portable container and then distribute them to Linux machines. It can also achieve virtualization without any interfaces between containers, so that the entire operating environment can be deployed quickly and automatically. For lightweight solutions based on Docker technology, some experts and scholars have begun preliminary exploration in telecommunications, power and other fields [6-8]. Using Kubernetes and Docker, an enterprise-level container big data cloud platform is constructed [6]. It is open to all kinds of PaaS capabilities, including big data processing capability components, cloud database and middleware components, data integration tool capabilities, container clouds, for provincial branches, subsidiaries and partners. Component capabilities, etc. In view of the large amount of data information contained in the actual operation of the communication network, Qu Rumping [8] proposes to establish a private cloud data platform based on Docker container technology, which can effectively manage and isolate the good multi-tenant resources. On the other hand, it is more convenient to start and destroy than the traditional virtual machine technology.

2.2 Relevant Technology
Cloud computing provides users with information services and storage space, and provides virtual computing services on the basis of virtualization technology and Internet infrastructure [9]. Cloud computing has four characteristics: large-scale, virtualization, scalability and stability. It only needs the Internet to access the cloud service platform and provide stable distributed file system services.
Under the running state of cloud platform, new servers can be added dynamically without affecting the normal operation of the cluster, which has good flexibility and scalability. Cloud computing can be divided into Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

Docker is an open source application container engine. The core problem to be solved is to use the Linux container (LXC) to realize the functions similar to the virtual machine (VM), so as to make use of the saved hardware resources to provide users with more computing resources. Container technology can effectively divide a single operating system resource into groups and meet the needs of resource utilization. That is to pack a process in a container, but it will not affect the normal operation of other processes [10]. By using scripts and management controls, Docker can be started anywhere, connected to ports and delivered to other parts of the application or service stack that need to use the service. Its working principle is shown in Figure 1.

3. Platform Architecture and Flow Design

3.1 Platform Architecture

The big data experimental platform adopts advanced cloud architecture, which can be built by using newly purchased servers or university own servers. The platform adopts B/S architecture, and the client configuration requirements are low. Both personal computer and thin client can be used. Under any condition of network connectivity, big data image system can be accessed through browser. The platform manages server cluster and divides resources by virtualization technology, loads big data experimental images on the divided virtual machine, and serves as the experimental system environment for students. Virtual machines can be created, started, shut down and other operations, but also through the system for overall cloud host management and overall cloud hard disk management.

3.2 Main Modules

The platform of big data experiment management mainly includes private cloud management system, experiment image system, experiment management system and experiment monitoring system.

(1) Private cloud platform architecture is adopted to manage the hardware resources needed for teaching. It can provide big data computing resources for teachers and students anytime and anywhere in the campus network. The system can create, start and close virtual machines, and manage cloud...
hosts and hard disks through the system.

(2) Provide stable experimental images, including complete experimental images and zero-based experimental images, to meet different needs at all study stages.

(3) Students can experiment online and submit experiment reports. Teachers can correct experiment reports online.

(4) The monitoring and control system of big data experiment is a set of monitoring and management system of hardware, software and system involved in the teaching of big data. The hardware monitoring mainly includes server cluster monitoring and switch monitoring. Software monitoring mainly focuses on the cluster environment of big data, including Hadoop, Spark, Storm, etc., and the operation of students' experimental image to obtain the data of students' experimental learning.

3.3 Flow Design
After users login on to web applications, they authenticate their information through MD5. If they pass the authentication, they enter the experimental system. If they fail to authenticate, they exit the application. Then start the Docker service, load the corresponding image according to the user's choice, and start the container. Finally, the application of big data experiment is started, including HDFS, Hadoop, Spark and so on, and then big data cluster is constructed for user to do big data experiments.

4. Application Evaluation
Next, according to the actual needs of a class (50 students), the corresponding hardware and software resources are selected to build the experimental environment, and the test and evaluation are made.

4.1 Platform Deployment
1) Basic environment
First, two 19-inch 2U rack servers are selected, one is the application server and the other is the support server. The application server has one 12 core CPUs, 64G memory, two hard disks of 2T and
the support server has one 24 core CPUs, 256G memory, four hard disks of 4T. Then two servers are interconnected through gigabit network card. Finally, centos7 systems where installed separately.

2) Installing Docker environment

(1) Setting install source

```
sudo yum-config-manager --add-repo https://download.Docker.com/linux/centos/Docker-ce.repo
```

(2) Install Docker

```
sudo yum install Docker-ce
```

(3) Start Docker

```
sudo systemctl start Docker
sudo systemctl enable Docker
```

(4) Create Docker groups

```
sudo groupadd Docker
sudo usermod -aG Docker $USER
```

(5) Testing whether the installation was successful, if “Hello from Docker!” printed Indicates that the Docker has been installed correctly.

```
Docker run hello-world
```

3) Build Open VPN server

In order to ensure that the client can connect the container in the server in the LAN, it is necessary to build an Open VPN server on the server side, then the client can access the server through Open VPN. The main steps include download and installation, CA certificate making, OpenVPN configuration and network configuration. Its core is to make server certificate, key and encryption files on the server side.

4) Deploy application

After JDK8, Tomcat 7 server and MySQL 5.7 database were installed successively, Java Web application war package was deployed on the platform. Then server was started to test whether it can be accessed normally.

4.2 Platform Test

![Main UI of platform](image)

Figure 3. Main UI of platform
Figure 4. Cluster and configuration

Figure 5. Resource share analysis

Figure 3 shows the main UI of the platform for student role. Click "Create" to create a cluster with three nodes in 5-10 seconds, including one master and two slaves, as shown in Figure 4. After configuring Open VPN, you can successfully connect to the server by clicking “Configuration Download”. Then configure and start the cluster, you can enter the experiment. After the experiment, click "Release" to release all the resources allocated in 40-60 seconds. Figure 5 is the monitoring situation of one month. Practice shows that the experiment runs smoothly and the performance index meets the requirements.

5. Conclusions

Building a low-cost and lightweight big data laboratory is an urgent need for training big data talents. By installing Docker software on Linux system, using OpenVPN routing and forwarding, and using Java web technology, the rapid construction of big data cluster environment in LAN is realized. On the one hand, without requiring high-performance hardware or new equipment, the cost of laboratory construction is saved. On the other hand, the integrated computer system greatly reduces the learning cost of students. It provides an all-round experimental and training platform for the training of big data talents in colleges and universities. In the future, the key work will be to enrich curriculum resources in different categories, and further optimize the platform to improve operational efficiency.

Acknowledgments

The authors acknowledge the financial support from the Major Project of Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 18KJA520003) and the Key Project of Modern Educational Technology Research in Jiangsu Province of China (No. 2018-R-63369) and the Initiation Fund for High-level Talents in Jingling Institute of Technology of China (No. jit-b-201505).

References

[1] Lin Ziyu. (2018) Experience Sharing of Big Data Technology Principles and Application Course Construction. Big Data, 4 (06): 29-37.
[2] Mao Shaofeng. (2019) Design and Implementation of Online Experimental System Based on Docker Container. Huazhong University of Science and Technology.
[3] Wang Yongkun, Luoxuan, Jin Yaohui. (2018) Design and Implementation of Hybrid Large Data Platform Based on Private Cloud and Physical Machine. Computer Engineering and Science, 40(02): 191-199.
[4] Shang Saihua. (2018) Education Cloud Implementation Scheme Based on OpenStack Architecture. Wireless Interconnection Technology, 15 (16): 62-63+73.
[5] Yao Qimin. (2019) Research on OpenStack-based Shared Cloud Platform for University Laboratories. Laboratory Science, 22(01): 76-79+82.
[6] Wang Zhijun. (2019) Exploration and Practice of China Unicom Containerized Big Data Cloud Platform. Information Technology and Standardization, 2019 (05): 66-69.
[7] Liu Shengqiang, Wang Jing. (2018) Docker's Hadoop Platform Architecture Analysis. Automation and Instruments, 2018 (10): 192-195.
[8] Qu Runping. (2019) On the Private Cloud Data Platform Architecture of Communication Network Based on Docker Container Technology. China New Communications, 21 (09): 37.
[9] Luo Junzhou, Jin Jiahui, Song Aibo, Dongfang. (2011) Cloud Computing: Architecture and Key Technologies. Journal of Communications, 32 (07): 3-21.
[10] Zhang Jian, Xie Tianjun. (2014) Research on Docker-based Platform as a Service Architecture. Information Technology and Informatization, 2014 (10): 131-134.