

Docker 在大数据课程实验教学环境构建中的应用研究

汪文彬

浙江科技大学理学院 浙江杭州 310023

摘要: Docker 作为一种具备轻量级、高性能、易迁移的虚拟化容器,已在软件开发、部署和运维领域普遍应用。本文探讨了利用 Docker 容器技术构建出高效、便捷、安全、可重复的标准化实验环境,为数据科学与大数据专业课程实验教学提供新的思路和方法。

关键词: Docker; Dockerfile; Docker Compose; 实验教学; 环境构建

1. 引言

“新工科”背景下高校教育改革的重要目标是培养学生创新实践能力,使其成为适应社会需要的工程应用型人才^[1]。数据科学与大数据技术作为“新工科”背景下应运而生的专业,其特点是多学科融合、技术更新快、实践应用广,实验在培养学生解决复杂问题能力和创新能力方面起着至关重要的作用^[2]。传统实验教学中,学生在部署实验环境(操作系统、中间件、依赖库),不仅费时,还容易因操作失误导致部署失败。此外,学生实验环境的多样化常常出现团队实验结果不一致。实验管理方面,教师需要花费大量时间维护和更新多个实验环境,加大了管理和维护量^[3]。如何高效、便捷地构建标准化的学生实验环境,成为实验教学过程中亟待需要解决的问题。

2. Docker 容器技术及其特点

Docker 容器是一种基于操作系统的轻量级虚拟化技术,在同一实例中可同时运行多个独立的容器。与传统虚拟机相比, Docker 具有启动快、资源占用低、轻量化、易迁移等优点^[4]。容器还能实现资源共享与封装,无论基础设施如何变化,都能确保应用在一致的环境中运行^[5]。将 Docker 容器应用于实验环境构建具有以下优点:

(1) 标准化: 教师将实验相关的系统、软件及所需依赖打包成一个镜像,并实行统一版本控制。

(2) 快速部署: 学生拉取镜像到本地即可快速启动实验环境,降低环境部署难度。

(3) 资源利用高: Docker 容器共享宿主机内核,在同一台机器上可同时运行多个实验环境,大大节约硬件资源^[6]。

(4) 一致性: 通过提供统一版本的 Docker 镜像,确保

团队协作实验结果一致。

(5) 易管理: 借助 Docker Compose 或 K8S 等编排工具,实现高效地部署与扩展多容器环境,简化实验管理流程^[7]。

3. Docker 容器在实验环境构建中的应用

应用 Docker 容器构建实验环境包括: 构建镜像文件(Dockerfile)、编写容器编排文件(Docker Compose)、镜像构建与分发、拉取启动镜像。

3.1 构建镜像文件

教师在 Dockerfile 文件中定义实验所需的基础运行环境、工具软件及依赖,配置环境变量和运行参数。例如,在部署 Hadoop 集群实验中,可以通过 Dockerfile 为学生提供一个预装 JDK、Hadoop 以及相关依赖的基础镜像。以下是适用单节点或多节点的 Hadoop 实验环境的 Dockerfile 模板:

```
FROM ubuntu:20.04
RUN apt-get update && \
    apt-get install -y wget unzip default-jdk-headless && \
    rm -rf /var/lib/apt/lists/*
ARG HADOOP_VERSION=3.3.1
ARG URL=https://dlcdn.apache.org/hadoop/common/hadoop-${HADOOP_VERSION}/hadoop-${HADOOP_VERSION}.tar.gz
RUN wget $URL && \
    tar -zxvf hadoop-${HADOOP_VERSION}.tar.gz && \
    mv hadoop-${HADOOP_VERSION} /usr/local/hadoop3
ENV JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
ENV HADOOP_HOME=/usr/local/hadoop3
ENV PATH=${HADOOP_HOME}/bin:$PATH
WORKDIR /app/
RUN mkdir -p /root/.hadoop && \
    echo "export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64" >> /root/.bashrc && \
    echo "export HADOOP_HOME=/usr/local/hadoop3" >> /root/.bashrc
```

3.2 编写容器编排文件

团队协作实验可以使用 Docker Compose, 教师定义一个多容器化的实验环境, 不仅能保证学生的实验环境一致,

还可以简化版本控制和部署流程。例如，Hadoop 集群通常包含 HDFS、YARN 等组件。使用 Docker Compose 可以将这些角色定义为不同的服务，并在同一网络中运行。以下是适用于部署 Hadoop 集群的 Docker Compose 模板：

```
version: '3'
services:
  namenode:
    image: namenode
    container_name: nn
    ports:
      - "8020:8020"
      - "50070:50070"
    volumes:
      - /var/hadoop/namenode:/namenode
    networks:
      - hadoop-network
  datanode1:
    image: datanode
    container_name: dn1
    ports:
      - "8021:8021"
    volumes:
      - /var/hadoop/dn1:/datanode
    networks:
      - hadoop-network
  datanode2:
    image: datanode
    container_name: dn2
    ports:
      - "8022:8022"
    volumes:
      - /var/hadoop/dn2:/datanode
    networks:
      - hadoop-network
networks:
  hadoop-network:
    driver: bridge
```

3.3 镜像构建与分发

镜像通常存储在 Docker Hub 等外部公共仓库中。为提高镜像拉取速度，方便镜像管理，基于 Registry 搭建私有化

的 Docker 镜像仓库，用于存储和管理实验镜像^[8]。

(1) 在 Dockerfile 文件所在目录中构建镜像：docker build -t hadoop_exp:1.0 .

(2) 登录私有仓库：docker login

(3) 分发镜像：docker push hadoop_exp:1.0

3.4 拉取启动镜像

拉取镜像：docker pull hadoop_exp:1.0

启动镜像：docker run -it --name my-hadoop hadoop_exp:1.0

4. 应用效果测试

4.1 测试环境

以部署 Hadoop 集群实验环境为例，比较虚拟机和 Docker 在启动时间、资源占用率、网络性能、集群扩展方面的性能：

4.1.1 宿主机配置：

CPU：Intel i7，内存：16GB，磁盘：1TB，操作系统：

Ubuntu

4.1.2 虚拟化软件：

虚拟机：VirtualBox，Docker 容器：Docker Engine

4.1.3 其他软件：

JDK，Hadoop

4.2 性能比较与分析

4.2.1 启动时间

方法：测试集群所有节点的启动时间。

结果：

环境	时间（秒）
虚拟机	210
Docker	55

分析：每个虚拟机都需要加载完整的操作系统和硬件模拟层，启动时间较长；Docker 使用轻量级隔离机制，启动速度更快。

4.2.2 资源占用率

方法：测试运行相同的 MapReduce 计算任务，监控宿主 CPU、内存使用情况，使用 top 和 docker stats 进行资源分析。

结果:

环境	CPU 占用 (%)	内存占用 (MB)
虚拟机	23	2400
Docker	12	900

分析: 每个虚拟机独占资源, CPU 和内存消耗较高; Docker 共享宿主机资源, 显著降低资源占用。

4.2.3 网络性能

方法: 测试数据上传、下载到 HDFS 上的传输速度。使用 iperf 测试虚拟机与 Docker 容器间的带宽。

结果:

环境	上传速度 (MB/s)	下载速度 (MB/s)
虚拟机	82	65
Docker	103	89

分析: 虚拟机依赖额外的网络模拟层, 导致一定程度的带宽损失; Docker 采用轻量级的网络虚拟化实现, 网络性能更优。

4.2.4 集群扩展

方法: 测试新增集群节点的响应时间和资源分配情况。

结果:

环境	时间 (秒)	资源分配效率
虚拟机	485	低
Docker	183	高

分析: 虚拟机需要独立配置和启动, 新节点扩展时间较长; Docker 新增节点更快速, 扩展性更出色。

5. 结语

Docker 容器为课程实验教学环境的构建与管理提供了有利支撑。利用 Docker 技术构建出标准化的实验环境, 不

仅有效降低了复杂实验环境搭建的门槛, 学生可以更加专注于完成实验任务和实践目标, 提高了学生的实验成就感, 同时还减轻了教师在管理实验资源的工作量。随着云计算、人工智能等技术的快速发展, Docker 在教学中的应用也将更广泛深入, 必将在提升教学质量和科研效率方面发挥重要作用^[9]。

参考文献:

- [1] 周杰, 王伟, 李明. 新工科背景下工程教育改革的实践与探索 [J]. 高等工程教育研究, 2020, 38(5): 1-6.
- [2] 张伟, 李娜. 新工科背景下数据科学与大数据技术专业建设研究 [J]. 计算机教育, 2018, (5): 45-50.
- [3] 李明, 周军. 传统实验教学中的资源管理和维护成本分析 [J]. 教育技术研究, 2017, (4): 65-70.
- [4] 周军. 容器技术与虚拟机技术的对比研究 [J]. 计算机应用研究, 2017(3): 98-102.
- [5] 王旭, 王伟, 陈渝. 基于容器技术的轻量级虚拟化资源隔离机制研究. 计算机研究与发展, 2017, 54(6): 1314-1325.
- [6] 刘洋, 陈刚, 赵海. 容器与虚拟机性能对比及其在云计算教学中的实践. 实验技术与管理, 2020, 37(5): 156-161.
- [7] 周俊, 刘振宇, 黄薇. Docker Compose 在微服务实验环境编排中的优化研究. 实验室研究与探索, 2022, 41(4): 89-94.
- [8] 徐明, 赵磊, 周航. 基于 Docker Registry 的私有镜像仓库构建与性能优化研究. 计算机科学, 2020, 47(S2): 312-318.
- [9] 王芳, 陈强. 容器技术在教育信息化中的创新实践研究——以 Docker 为例 [J]. 现代教育技术, 2020, 30(3): 78-82.

基金项目:

浙江省高校实验室工作研究项目 (YB202251)