DOI:10.16515/j.cnki.32-1722/n.2019.04.001

基于Docker容器的轻量级大数据实验平台构建研究

谷瑞军

(金陵科技学院计算机工程学院,江苏 南京 211169)

摘 要:高校大数据专业教学中最大的问题就是没有配套的基础实验环境,无法同时对大数据环境进行部署,配置难度较高。另外,在实验过程中缺少实验数据、实验教案和实验手册,难以很好地开展相关的教学活动。为降低实验室建设成本,减少大数据的学习难度,基于虚拟化容器技术,构建了轻量级的大数据实验平台。通过该平台,可一键创建大数据集群,提供各种与课程匹配的实验环境,将重心聚焦到技术本身,大幅提高学习效率。

关键词:虚拟化技术;大数据;Docker容器;实验平台

中图分类号:TP311 **文献标识码**:A **文章编号**:1672-755X(2019)04-0001-05

Research on Construction of Lightweight Experimental Platform for Big Data Based on Docker Containers

GU Rui-jun

(Jinling Institute of Technology, Nanjing 211169, China)

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 large 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 the 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.

Key words: virtualization technology; big data; Docker container; experimental platform

近年来,很多高校都开设了大数据专业,教学中最大的问题是没有配套的基础实验环境,无法对大数据环境进行快速部署,配置难度较高。实际教学中,大数据的实验环境一般要求运行在Linux上,且需要多台机器构建集群,搭建过程涉及Linux系统、虚拟机软件、JDK、Hadoop、Spark、HBase、Hive等多种开源软件的安装,安装过程繁琐、配置工作量极大。而且有时同样的操作方法,在某台计算机上可以成功,在另一台计算机上却会失败[1]。因此,传统的PC机房无法满足实验环境的要求。目前主要有三种解决方案:1)学生通过个人电脑构建伪分布式的实验环境,模拟大数据集群。因受限于个人电脑的性能,数据处

收稿日期:2019-09-15

基金项目:江苏省高等学校自然科学研究重大项目(18KJA520003);教育部产学合作协同育人项目(201801094010);江苏省现代教育技术研究课题(2018-R-63369,2019-R-69581);江苏省高校自然科学研究面上项目(19KJB520007);金陵科技学院博士科研启动基金(jit-b-201505)

作者简介:谷瑞军(1979—),男,山东菏泽人,副教授,博士,主要从事数据挖掘和商务智能方面的研究。

理能力有限，无法获得较好的学习效果。2)通过 OpenStack 等虚拟化技术，在服务端构建大数据集群，安排专人负责维护管理。缺点在于资源消耗较大，维护管理成本高。3)租用第三方云服务平台(如阿里云)部署实验环境。此方案系统稳定、性能高，但很难定制个性化的环境，且租用成本高。

为降低实验室建设成本，减少大数据的学习难度，利用轻量级的 Docker 容器化技术为用户提供虚拟化的操作环境，能够保证用户进行相同实验时的环境是一样的，确保每个用户的学习虚拟环境都是独立的，相互之间不会发生干扰。因为 Docker 采用的是操作系统层面的虚拟化技术，所以对内存、CPU 等系统资源有更高的利用率，在相同的服务器集群条件下可以支持更多的用户同时在线使用[2]。

1 相关工作

1.1 相关研究

在公有云方面，国内外很多厂商基于开源的大数据技术，提供了建立在云上的大数据平台方案[3]，包括国外的亚马逊云的 AWS(Amazon Web Service)、微软云(Azure)上的 Hadoop、谷歌云平台大数据产品，国内有阿里云大数据服务 MaxComputer、腾讯大数据处理套件(Tencent Big Data Suite)、百度大数据基础套件(鲁班)。云上的解决方案可扩展性强，可以满足用户不断变化的需求。例如：阿里云通过沙箱技术，提供真实的阿里云环境，可深度体验和学习阿里云产品和服务。平台上线了大量的实验项目，涵盖云计算、大数据、云安全等领域，并提供了详细的实验文档和提示引导。但是相对而言成本也比学校等非盈利机构的内部平台要高，而且涉及的知识产权问题等较严格，数据开放共享的难度较大。因此，通过虚拟化技术，利用学校的公共资源搭建虚拟实验室，是当前最主流的方式。

目前基于虚拟化技术的实验室大多采用 OpenStack 技术，对服务端资源进行虚拟化，然后分配给学生端搭建集群。文献[4]采用 OpenStack 技术和 IaaS 层架构，对资源进行集中托管，资源互通，充分利用了硬件资源，为师生提供可扩展、不可见的基础设施。对实验室环境云化后，各种实验所需的硬件资源按需分配，使教学工作聚焦在专业学习上而不是环境的搭建和配置上。与之类似，姚琦敏[5]采用 OpenStack 构建大数据实验室，实现了实验室资源不受时间和空间限制、开放共享的目标。另外，通过优化实验室管理，降低了实验室建设成本，提升了实验室教学资源的利用率。王永坤等[3]尝试完全使用开源软件设计搭建一个弹性的数据平台。该数据平台主要构筑在物理服务器上，保证了分析性能。当资源紧张时，可以动态扩展到私有云平台上。私有云平台的底层采用了 OpenStack 框架来实现软件定义网络、弹性计算以及软件定义存储等功能。

基于 OpenStack 的虚拟化技术在物理主机的资源利用率上无法达到真实物理主机的水平，对于文件的配置不够灵活。Docker 作为开源的轻量级的应用容器引擎，可以将应用包到一个可移植的容器当中，然后发布到 Linux 机器上，同样可以实现虚拟化，且容器之间不会有任何接口，从而可以快速、自动化地部署整个运行环境[5]。对基于 Docker 技术的轻量级解决方案，部分专家学者在电信、电力等领域已开始探索[6-8]。文献[6]采用 Kubernetes+Docker 的主流技术路线，建设了企业级容器化大数据云平台，面向各分公司、子公司及合作伙伴，一点开放各类 PaaS 能力，包括大数据处理能力组件、云数据库及中间件组件能力、数据集成工具能力、容器云组件能力等。瞿润萍[8]针对通信网实际运行当中蕴含着的大量数据信息，提出基于 Docker 容器技术建立通信网私有云化大数据平台，一方面将良好的多租客资源进行有效的管理和隔离；另一方面，与传统的虚拟机技术相比具有启动销毁方便、占用资源少等优势。

1.2 相关技术

云计算为用户提供信息服务和存储空间，在虚拟化技术和互联网技术的基础上，提供虚拟的计算服务[9]。云计算具有大规模、虚拟化、扩展性和稳定性等四个特点，只需要互联网就可以接入云服务平台，并且提供稳定的分布式文件系统服务。在云平台运行状态下，可以动态加入新的服务器，且不会影响集群的正常运行，具备良好的弹性和可扩展性。根据服务类型云计算可分为将基础设施作为服务(IaaS)、将平台作为服务(PaaS)和将软件作为服务(SaaS)。

Docker 是一个开源的应用容器引擎，要解决的核心问题是利用 Linux 容器(LXC)来实现类似虚拟机

(VM)的功能,从而利用节省的硬件资源提供给用户更多的计算资源。容器技术可以有效地让单个操作系统资源划分到各个组中,并且满足资源使用的需求。也就是将一个进程打包在集装箱之中,但不会影响其它进程的正常运行[10]。即便是容器需要启动,也不会影响整个系统,只是在进程间进行调度。Docker 现阶段已成为业界设计和开发的主要技术。通过使用脚本和管理控制,可以在任何地方启动 Docker,连接端口并将其传达到需要使用该服务的应用或服务堆栈的其他部分。其工作原理如图 1 所示。

2　平台架构和设计

2.1　平台架构

大数据实验平台采用先进的云架构,可利用新购服务器或高校自有服务器进行搭建,支持目前国内外主流的服务器品牌。平台采用 B/S 架构,客户机配置要求低,PC 和瘦客户机或工作站等形式均可。在任何网络连通的条件下,都可通过 IE 访问大数据镜像系统。该平台通过虚拟化技术进行服务器集群的管理与资源划分,在划分出的虚拟机上加载大数据实验镜像,作为学生大数据的实验系统环境。可以对虚拟机进行创建、启动、关闭等操作,也可以通过该系统进行整体云主机管理和整体云硬盘管理。其架构如图 2 所示。

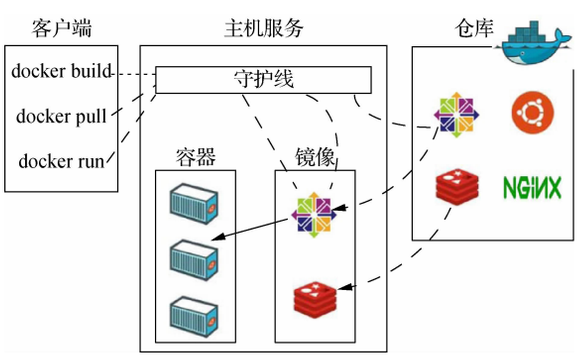


图 1　Docker 工作原理

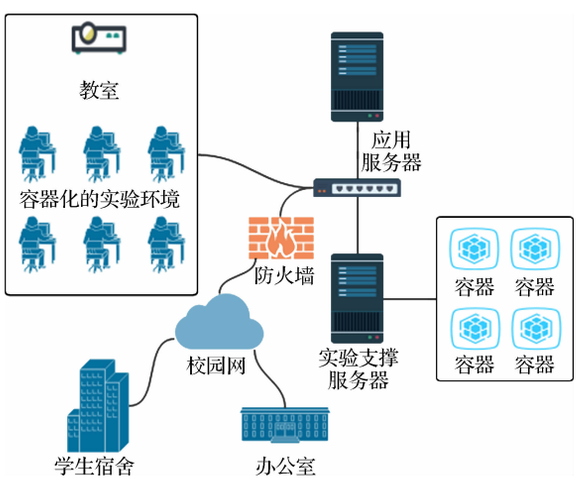


图 2　平台架构

2.2　主要模块

大数据实验管理平台主要包括:1)私有云管理系统。采用私有云平台架构,对教学所需的硬件资源进行统一管理,在学校内网中可随时随地为教师、学生提供所需的大数据计算资源。该系统可以对虚拟机进行创建、启动、关闭等操作,也可以通过该系统进行云主机和云硬盘管理。2)实验镜像系统。提供稳定的实验镜像,包括完整实验镜像和零基础实验镜像,满足各阶段的课时管理需求。3)实验管理系统。学生可在线实验,并提交实验报告,教师可在线批改实验报告。4)大数据实验监控系统。其是对大数据教学涉及到的硬件、软件等监控的一套系统。其中硬件监控主要包括对服务器集群监控和交换机监控;软件监控主要关注大数据的集群环境,包括 Hadoop、Spark、Storm 等,以及学生实验镜像的运行情况,获取学生实验学习情况数据。

2.3　大数据实验的加载流程设计

用户登录 Web 应用后,对获取用户的信息通过 MD5 进行鉴权认证,认证通过,则进入实验系统,认证失败则退出应用。接着启动 Docker 服务,根据用户的选择加载对应的镜像,并启动容器。最后,启动大数据实验的应用,包括 HDFS、Hadoop、Spark 等,进而构建大数据集群,用户即可以进行大数据实验了。

3　应用评测

以下根据一个班级(50 人)的实际需求,选取相应的软硬件资源来搭建实验环境,并作了测试和评价。

3.1 平台部署

1)基础环境。首先选择两台 19 英寸 2U 机架服务器,一台为应用服务器,一台为支撑服务器。应用服务器配置为:12 核心 CPU、64 G 内存、两块 2 T 硬盘。支撑服务器配置为:24 核 CPU、256 G 内存、四块 4 T 硬盘。两台服务器间通过万兆网卡互联。然后分别安装 Centos7 系统。

2)安装 Docker 环境

①设置源

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

②安装 Docker

sudo yum install Docker-ce

③开启 Docker,并设置开机启动

sudo systemctl start Docker

sudo systemctl enable Docker

④创建 Docker 组,避免普通用户使用 Docker 命令时需要 sudo

sudo groupadd Docker

sudo usermod-aG Docker $USER

⑤测试是否安装成功,如果打印出 Hello from Docker! 表示 Docker 已正确安装成功。

Docker run hello-world

3)搭建 Open VPN 服务器。为保证客户端能够在局域网连接支撑服务器里的容器,需要在服务端搭建 Open VPN 服务器,客户端则通过 Open VPN 访问服务器。主要步骤包括下载安装、CA 证书制作、Open VPN 配置和网络配置。其核心是制作 Server 端的 Server Certificate、Key 和 Encryption Files。

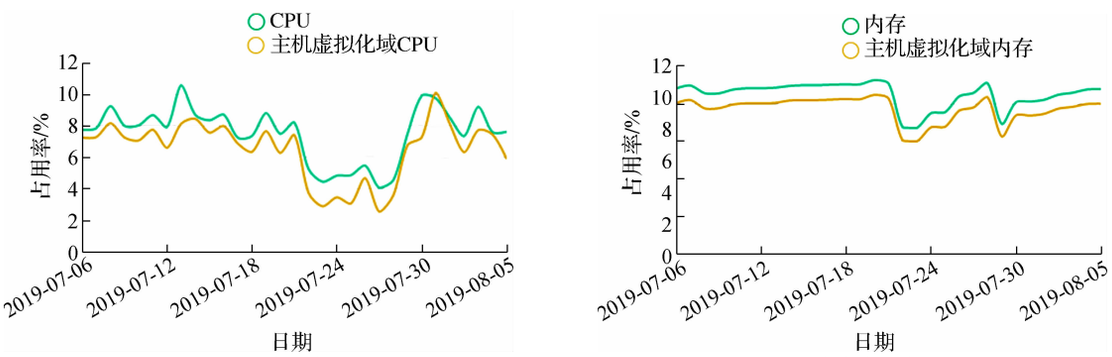
4)应用部署。先后安装 JDK8、tomcat7 服务器和 mysql5.7 数据库,将 java web 应用 war 包部署到平台上,并启动服务器,测试是否能正常访问。

3.2 平台测试

图 3 为平台学生端主界面,单击“创建”,5～10 s 可创建一个包含 3 个节点的集群,包括 1 个主节点(master)和 2 个从节点(slave),如图 4 所示。单击“配置下载”,配置好 Open VPN 后,可成功连接到服务器。然后配置并启动集群,即可进入实验。实验完毕后,单击“释放”,可以在 40～60 s 将所分配的资源全部释放。在大数据监控系统的主界面,可显示 CPU、内存、JVM 等主要指标的即时占有率。图 5 为一个月的监控情况。实践表明,实验运行流畅,性能指标满足要求。

图 3 平台主界面

图 4　大数集群及配置信息



(a)CPU 占用率　　(b)内存占用率

图 5　资源占有率情况

4　结　语

建设低成本轻量级的大数据实验室是培养大数据人才的迫切需求。通过在 Linux 系统上安装 Docker 软件,借助 Open VPN 的路由转发,采用 Java web 技术,实现了局域网内的大数据集群环境的快速搭建。一方面,在不需要高性能硬件或新购设备情况下,节省了实验室建设成本。另一方面,集成化的一体机系统,大大降低了学生的学习成本。这为高校大数据人才培养提供了一个触手可及、随时可用、秒级启动、用完即走的全方位实验实训平台。未来的重点工作将结合实际课程,分门别类地丰富课程资源,并进一步优化平台,提升运行效率。

参考文献:

[1] 林子雨. 大数据技术原理与应用课程建设经验分享[J]. 大数据,2018,4(6):29 - 37
[2] 毛少枫. 基于 Docker 容器的在线实验系统设计与实现[D]. 武汉:华中科技大学,2019
[3] 王永坤,罗萱,金耀辉. 基于私有云和物理机的混合型大数据平台设计及实现[J]. 计算机工程与科学,2018,40(2):191 - 199
[4] 尚赛花. 基于 OpenStack 架构的教育云实现方案[J]. 无线互联科技,2018,15(16):62 - 63
[5] 姚琦敏. 基于 OpenStack 的高校实验室共享云平台的研究[J]. 实验室科学,2019,22(1):76 - 79
[6] 王志军. 中国联通容器化大数据云平台的探索与实践[J]. 信息技术与标准化,2019(5):66 - 69
[7] 刘胜强,王晶. Docker 的 Hadoop 平台架构分析[J]. 自动化与仪器仪表,2018(10):192 - 195
[8] 瞿润萍. 试论基于 Docker 容器技术的通信网私有云化大数据平台架构[J]. 中国新通信,2019,21(9):37
[9] 罗军舟,金嘉晖,宋爱波,等. 云计算:体系架构与关键技术[J]. 通信学报,2011,32(7):3 - 21
[10] 张建,谢天钧. 基于 Docker 的平台即服务架构研究[J]. 信息技术与信息化,2014(10):131 - 134

(责任编辑:湛　江)