计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):47−56 [doi: 10.15888/j.cnki.csa.007663]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

高校大数据实验室及实验体系的规划与建设①

吴湘宁[1], 彭建怡[1], 罗勋鹤[2], 刘远兴[1], 李　敏[3]

[1](中国地质大学 (武汉) 计算机学院, 武汉 430074)
[2](中国地质大学 (武汉) 实验室与设备管理处, 武汉 430074)
[3](荆楚理工学院 计算机工程学院, 荆门 448000)
通讯作者: 彭建怡, E-mail: 547397787@qq.com

摘　要: 大数据产业已上升至国家战略, 建立大数据实验室及实验课程体系是培养大数据技术人才的必要条件. 本文对大数据的知识体系进行了梳理, 分析了“数据科学与大数据技术”专业和“大数据技术与应用”专业的培养目标及职业定位, 明确了大数据专业的学生应该掌握的关键知识和需要重点培养的专业技能, 介绍了主流的大数据生态系统, 选取了最通用的大数据架构, 提出了在单机环境、单机虚拟化环境、共享大数据集群环境、云计算环境下建设大数据实验室的不同方案, 并设计了大数据实验课程体系及实验项目.

关键词: 大数据实验室; 大数据实验课程体系; 大数据生态系统; 实验室建设; 云计算

引用格式: 吴湘宁,彭建怡,罗勋鹤,刘远兴,李敏.高校大数据实验室及实验体系的规划与建设.计算机系统应用,2020,29(11):47–56. http://www.c-s-a.org.cn/1003-3254/7663.html

Planning and Construction of Big Data Laboratory and Experiment System in Colleges and Universities

WU Xiang-Ning[1], PENG Jian-Yi[1], LUO Xun-He[2], LIU Yuan-Xing[1], LI Min[3]

[1](School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430074, China)
[2](Office of Laboratory and Equipment Management, China University of Geosciences (Wuhan), Wuhan 430074, China)
[3](School of Computer Engineering, Jingchu University of Technology, Jingmen 448000, China)

Abstract: Big data industry has risen to the national strategy. The establishment of big data laboratory and experimental curriculum system is necessary for training big data technical personnel. This paper combs the knowledge system of big data, analyzes the training objectives and career orientation of the major of “data science and big data technology” and the major of “big data technology and application”, and clarifies the key knowledge that big data students should master and the professional skills that need to be cultivated, introduces the mainstream big data ecosystem, selects the most general big data architecture, proposes different plans to build big data laboratory in single machine environment, single machine virtualization environment, shared big data cluster environment and cloud computing environment, and designs the big data experiment curriculum system and experiment projects.

Key words: big data laboratory; big data experiment curriculum system; big data ecosystem; laboratory construction; cloud computing

1　引言

自 2012 年以来, 我国开始步入全新的大数据时代. 海量数据的生产已经成为经济与社会生活中的一个普遍现象, 利用数据改善决策、合理配置资源已成为企业创造价值的重要方法. 党中央、国务院高度重视大数据在经济社会发展中的作用, 2015 年 11 月 3 日发布

① 基金项目: 中国地质大学 (武汉) 中央高校教改基金 (本科教学工程)(2019G51); 中国地质大学 (武汉) 实验技术研究项目 (SJ-201825)
Foundation item: The Central Universities Education Reform Fund of China University of Geosciences (Wuhan) (Undergraduate Teaching Project) (2019G51); Experiment Technique Research Project of China University of Geosciences (Wuhan) (SJ-201825)
收稿时间: 2020-03-31; 修改时间: 2020-04-24; 采用时间: 2020-05-10; csa 在线出版时间: 2020-10-29

的《中华人民共和国国民经济和社会发展第十三个五年规划纲要》第二十七章“实施国家大数据战略”提出: 把大数据作为基础性战略资源, 全面实施促进大数据发展行动[1]. 国务院印发《促进大数据发展行动纲要》, 全面推进大数据发展, 加快建设数据强国[2]. 工业和信息化部也印发了《大数据产业发展规划 (2016–2020 年)》, 指出数据是国家基础性战略资源, 是 21 世纪的“钻石矿”, 提出到 2020 年我国要建成技术先进, 应用繁荣, 保障有力的大数据产业体系[3].

“十三五”期间, 随着我国信息产业的迅速壮大, 积累了丰富的数据资源, 大数据技术创新取得了明显突破, 已从 2012~2013 年的启动期, 再经过 2014~2017 年的高速发展期, 逐渐进入了 2018 年至今的成熟发展期. 大数据技术的商业模式已经得到了市场印证, 并已进入市场细分的时代. “十三五”期间也正是全球新一代信息产业的加速变革期, 大数据技术和应用不断创新突破期, 国内市场需求集中爆发, 我国大数据产业面临着前所未有的发展良机. 据统计 (图 1), 2017 年中国大数据产业规模从 2016 年的 2840.8 亿元迅速增长至 2019 年的 5386.2 亿元, 预计 2020 年还会继续增长 22.6%[4,5]. 大数据领域的投融资规模也从 2014 年的 303.75 亿元升至 2018 年的 1581 亿元, 随着大数据在各行业的融合应用不断深化, 包括数据挖掘、机器学习、数据资产管理、信息安全等大数据技术及应用领域还将继续突破, 成为推动经济高质量发展的新动力[6].

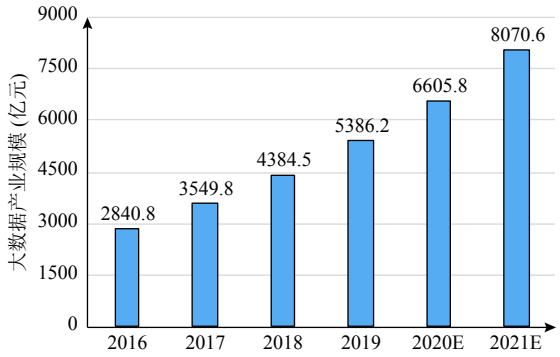

图 1 2016~2021 年中国大数据产业规模及预测

然而, 与大数据行业快速增长形成鲜明对比的是大数据人才的短缺. 根据猎聘网发布的“2019 年中国 AI&大数据人才就业趋势报告”, 中国大数据领域的人才需求呈现快速增长态势, 2019 年企业对大数据人才的需求约为 4 年前的 12 倍, 人才缺口高达 150 万.

为了有效缓解大数据人才供给的缺乏, 教育部加快大数据人才培养布局, 2016 年增设“数据科学与大数据技术”本科专业及“大数据技术与应用”高职专业. 全国各地院校也积极做出响应. 根据教育部发布的历年“普通高等学校本科专业备案和审批结果”, 截至 2019 年底, “数据科学与大数据技术”专业新增备案学校数量达 631 所, 尤其是 2017 年通过审批的学校数量同比增长近 10 倍[7,8]. “大数据技术与应用”高职专业的增长势头同样迅猛, 有时甚至是成倍增长, 截至 2020 年底, 已有 1354 所高职院校获批该专业[9], 具体可见图 2、图 3.

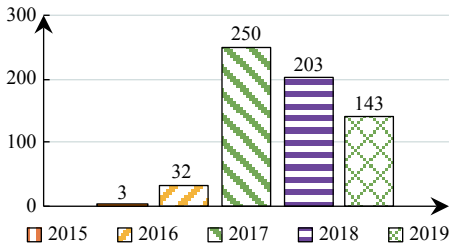

图 2 教育部审批的开设“数据科学与大数据技术”专业高校数目

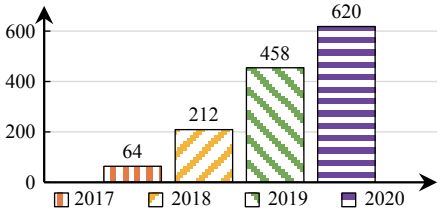

图 3 教育部审批的开设 “大数据技术与应用”专业职业院校数目

虽然我国高校的大数据人才培养经过几年的发展已经有了一定基础, 但是仍然存在着一些问题, 如专业师资紧缺、缺少系统化的权威教材等, 而其中最为紧迫的问题是许多院校还没有专业的大数据实验室, 甚至只是将普通的计算机实验室改造, 临时充当大数据实验室, 此外, 也没有比较系统的实验体系和实验项目, 这在很大程度上影响了大数据专业毕业生的质量, 并造成高校对大数据人才的培养与企业对人才专业技能需求的脱节. 尤其是随着第一批入学的学生开始进入专业课程和实习环节, 这个问题变得愈发严重. 因此, 建设大数据实验教学和实训环境, 并构建结构合理的大数据实验课程体系就成为了开设大数据专业的院校急需解决的问题.

2 大数据专业的培养目标、职业定位及知识结构

2.1 大数据专业的培养目标

“数据科学与大数据技术”本科专业是一个软硬件结合、以计算技术为基础、以数据科学与大数据技术为特色的宽口径专业, 以计算机科学、数学和统计学为三大基础支撑 (图 4), 并向经济、农业、生物、医学、地质、环境、社会、管理等应用领域拓展的, 典型的多学科交叉的新工科专业[10].

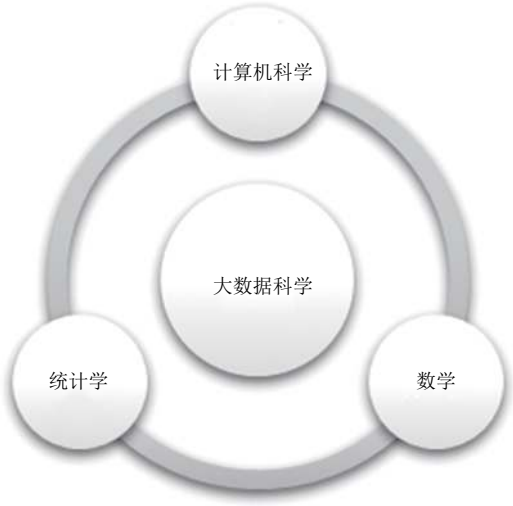

图 4 大数据科学的三大基础

虽然不同高校“数据科学与大数据技术”本科专业的人才培养目标侧重不同, 但仍可归纳出一些共同点: 培养适应社会与经济发展需求的大数据相关领域的, 具备较扎实的计算机科学、数学和统计学基础知识, 掌握大数据相关技术, 能够从事一定应用领域 (如金融、商务、农业、公安) 的大数据采集、存储与管理、分析处理、服务等相关工作, 具有扎实专业基础、很强的实践能力, 并具有广泛适应性的高素质理学或工学人才[11].

出于对师资力量和教师知识结构的考虑, 高校通常将“数据科学与大数据技术”专业放在计算机、信息科学、数学等学院, 或新兴的大数据科学学院. 由于各高校的优势学科、学科结构不同, 在专业的人才培养目标上自然也各有侧重. 部分偏理学的高校强调毕业生应具有一定的大数据科学研究能力及适应数据科学家岗位的基本能力, 但是大多数的高校更倾向于培养能够满足不同行业需求的应用型、复合型工学人才.

“大数据技术与应用”高职专业人才的培养更重视技术技能型人才的培养, 其培养目标可以归纳为: 培养具有精益求精工匠精神, 具有较强就业能力和可持续发展能力, 掌握大数据技术基本理论和基本技能, 面向软件与信息服务行业, 能够从事大数据平台运维、数据采集、数据清洗、数据加工、数据可视化、系统开发、系统实施等工作的高素质技术技能型人才[12,13].

可见, “数据科学与大数据技术”本科专业与“大数据技术与应用”高职专业在人才的专业知识和专业技能上具有一定的差异, 但是也具有明显的互补性.

2.2 大数据专业的职业定位及知识结构

大数据相关的职位可以分为: 数据产品经理、数据架构师、算法设计及数据挖掘工程师、大数据应用系统研发工程师、数据分析师/工程师、运维工程师 (如表 1 所示). “数据科学与大数据技术”专业的毕业生比较适合前 4 种职位. “大数据技术与应用”专业的毕业生则比较适合后 3 种职位, 但是在累积一定的工作经验和知识以后, 亦可胜任前 3 种职位. 有时, 这些职位之间的界限并不明显, 比如算法设计及数据挖掘工程师也可从事大数据应用系统的研发工作.

表 1 大数据相关职位的职责及知识要求

职位名称	角色	主要职责	需具备的关键知识
数据产品经理	管理团队	项目管理、质量管控.	大数据技术、软件工程、成本管理、人际沟通.
数据架构师	设计、创建数据管理系统, 整合、集中数据资源	大数据平台架构设计、创建和优化. 数据相关应用系统架构设计.	数据建模、数据治理、分布式计算框架、系统备份与恢复、负载均衡、信息安全、统计学理论方法、机器学习及人工智能理论.
算法设计及数据挖掘工程师	业务逻辑建模、数据建模、数据挖掘及预测	业务理解并深挖、高级算法设计与优化、API设计与实现.	统计学理论方法、大数据技术、机器学习及人工智能算法、高级语言编程.
大数据应用系统研发工程师	应用系统开发	用编程方式实现业务逻辑.	数据库、数据仓库及OLAP(联机分析处理)、大数据技术、可视化技术、高级语言编程.
数据分析师/工程师	数据采集、加工、预处理及管理	运用工具采集、提取、分析, 初步呈现数据商业意义.	数据库、数据仓库及BI(业务智能)、大数据技术、统计学理论方法、数据ETL(抽取、转换、加载)、数据清洗、可视化技术.
运维工程师	技术支持及运行维护	大数据平台搭建、运维.	计算机硬件及操作系统、数据库、数据仓库及OLAP、大数据技术、云计算技术、系统备份与恢复.

图 5 是大数据相关的知识体系, 不同的大数据职位, 所需具备的关键知识会有所不同, 但是均属于此知识体系的子集. 院校可以有针对性地开设不同的课程来为学生补充相应的专业知识, 并设置不同的实验项目来提高学生的实践技能. 例如: 对“数据科学与大数据技术”本科生除了要具备计算机科学、数学、统计学知识以外, 需有意识地加强大数据、数据库及数据仓库、机器学习与数据挖掘、可视化等方面知识的学习和技能训练. 而对于“大数据技术与应用”专业的学生除了需要具备必要的计算机科学、统计学知识以外, 需要有目的地加强大数据、数据库及数据仓库、云计算、可视化等方面知识的学习和技能训练.

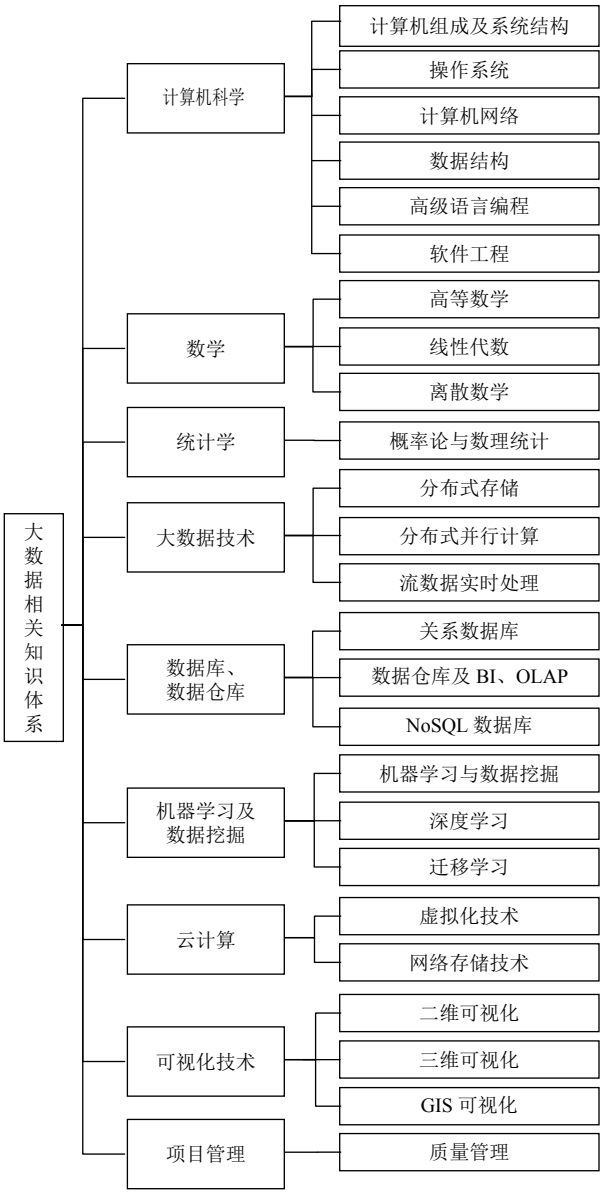

图 5 大数据相关知识体系

3 大数据实验室的架构

3.1 主流大数据技术生态环境

在大数据领域, Hadoop 已经成为当前的主流框架, 它是 Apache 旗下开源分布式计算平台, 提供了底层细节透明的分布式大数据存储基础架构, 基于 Java 语言开发, 具有很好的跨平台特性, 并且可以部署在廉价的计算机集群中. Hadoop 以其实用、成本低廉、开源等优点受到了业界的欢迎.

Hadoop 经过不断完善和发展, 已经形成一个种类丰富的生态系统 (如图 6 所示). 其核心是 Hadoop 分布式文件系统 HDFS (Hadoop Distributed File System) 和分布式计算架构 MapReduce. 同时也提供了资源管理调度器 YARN、ETL 工具 Sqoop、日志采集工具 Flume、分布式 NoSQL 数据库 Hbase、分布式数据仓库 Hive、类 SQL 语言 Pig Latin、分布任务协调工具 Zookeeper、分布式消息发布订阅系统 Kafka、流计算框架 Storm、工作流管理系统 Oozie、机器学习库 Mahout、大数据集群部署及管理工具 Ambari 等组件. Hadoop 可通过 Java、R、Python 等语言的 API (Application Programming Interface, 应用程序接口) 访问.

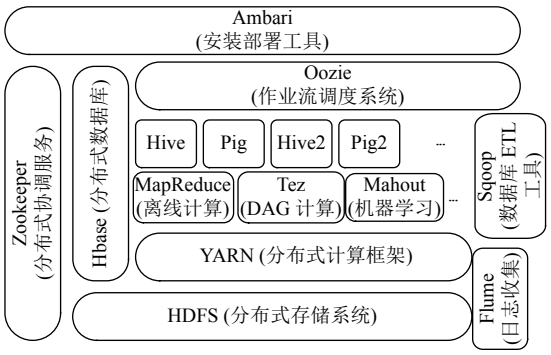

图 6 Hadoop 的生态系统

除 Hadoop 之外, 近年来还比较流行大数据快速并行计算引擎 Spark. Spark 可以同 HDFS 很好地结合, 也可以同 Amazon 的 S3 等云平台结合. 不同于 MapReduce 的是, Spark 计算作业的中间输出结果保存在内存中, 不像 MapReduce 那样需要频繁读写 HDFS, Spark 采用 RDD(Resilient Distributed Dataset, 弹性的分布式数据集) 作为数据交换结构, RDD 代表一个不可变、只读的, 被分区的数据集, Spark 会根据 RDD 的依赖关系生成 DAG (Directed Acyclic Graph, 有向无环图), 并从 DAG 的起点开始优化并执行. 正是由于这些特殊机制,

使得 Spark 比 MapReduce 计算性能更佳, 甚至能够提升百倍, 可用于替代 MapReduce 来实现一些需要迭代的数据挖掘与机器学习算法. Spark 也拥有自己独立的生态体系 (如图 7 所示), 包括流计算 Spark Streaming、机器学习库 MLlib、数据库查询语言 Spark SQL、数据仓库 Shark、图计算 GraphX 等一整套分布式计算组件.

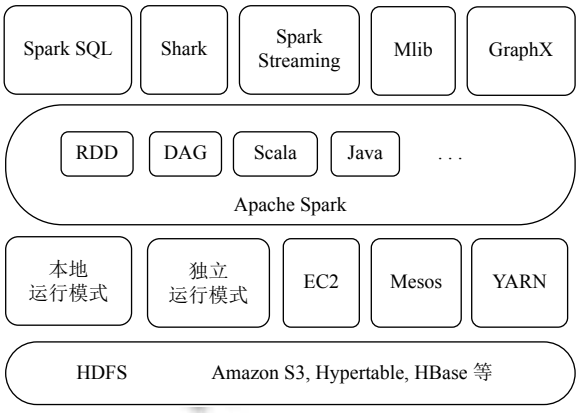

图 7 Spark 的生态系统以及与其他平台的关系

Spark 和 Scala 语言紧密集成, Scala 可以像操作本地数据一样轻松操作分布式数据集. 但是 Scala 语言的面向对象和函数式编程的混搭风格, 使得学习 Scala 语言存在一定的门槛.

高校的大数据实验课程及实验项目大多可以基于 Hadoop 生态系统及 Spark 生态系统这两种主流框架来设计, 挑选其中的一些常用的核心组件的使用来设置实验内容. 例如: Hadoop 环境下对 HDFS、HBase、Hive 的访问, 基于 Flume 的日志数据采集、基于 Storm 的流式数据处理、基于 Sqoop 的数据抽取、基于 Spark 的大数据并行分析等.

3.2 Hadoop 社区版与发行版

Apache Hadoop 社区版 (也称原生版) 虽然完全开源免费, 社区活跃. 但是也存在版本管理混乱、部署过程繁琐、升级过程复杂、运维难度大、组件之间兼容性差、安全性低等不足, 手工安装和配置一个 Hadoop 集群往往耗费大量时间.

Hadoop 发行版衍生自社区版, 由第三方简化并提供 Hadoop 部署、安装、配置工具, 大大提高了集群部署的效率, 可以在几个小时内完成集群的部署. 而且提供了配置修改、监控、诊断的管理工具, 管理配置方便, 定位故障快速准确, 使运维工作变得简单有效. 发行版经过大量测试和众多部署实例验证, 在兼容性、安全性、稳定性上有所增强, 可部署到生产环境. 目前常用的开源发行版主要有 Cloudera 的 CDH (Cloudera's Distribution including Apache Hadoop), 以及 Hortonwork 的 HDP (Hortonworks Data Platform).

院校的大数据实验环境建议使用 Hadoop 发行版, 让学生将更多精力放在分布式大数据应用的开发上, 而不是放在大数据系统的安装配置上, 此外, 让学生在实验中使用业界认可的商业版本, 也有利于将来毕业后很快适应工作岗位.

3.3 单机环境大数据实验平台

开展大数据实验不一定只能在具有一定规模的计算机集群环境中才能实现, 其实大多数大数据实验在单机环境下就可以完成. Hadoop 已经考虑了大数据开发环境、测试环境和生产环境的不同, 分别提供了 3 种运行 (启动) 模式:

(1) 单机模式 (Local/Standalone Mode, 也称为独立模式): 不需要对配置文件进行修改. 程序运行时使用本地文件系统, 而不是 HDFS. Hadoop 不会启动 NameNode、DataNode 等守护进程, 此模式主要用于对 MapReduce 程序的逻辑进行调试, 确保程序的正确.

(2) 伪分布式模式 (Pseudo-Distrubuted Mode): Hadoop 的 NameNode、DataNode 等守护进程运行在本机上, Hadoop 使用的是 HDFS, 用来模拟大数据集群, 是完全分布式模式的一种仿真, 此模式常用来测试开发的 Hadoop 程序执行是否正确.

(3) 全分布式模式 (Full-Distributed Mode): Hadoop 的守护进程运行在由多台主机搭建的集群上, 是真正的生产环境.

同样, Spark 也支持 3 种分布式部署方式:

(1) 独立模式 (Standalone): 自带完整的服务, 可单独部署到一个小型集群中, 无需依赖任何其他资源管理系统.

(2) 基于 Mesos 模式 (Spark on Mesos): 利用 Mesos 做资源管理, Spark 在开发之初就已考虑了和 Mesos 的兼容, 所以两者存在天然的血缘关系, Spark运行在 Mesos 上会比运行在 YARN 上更加灵活、自然.

(3) 基于 YARN 模式 (Spark on YARN): 利用 Hadoop 平台的 YARN 做资源管理. 好处在于可以将 Spark 与 Hadoop 兼容. 又可进一步细分: 生产环境可选择 yarn-cluster 模式, 调试程序则选择 yarn-client 模式.

Spark 计算框架的一般开发模式为: 为了快速开发, 先不需要考虑服务 (如 master/slave 环境) 的容错

性, 直接在 Standalone 模式下开发, 之后再开发相应的 wrapper, 将 Stanlone 模式下的服务原封不动地部署到资源管理系统 YARN 或者 Mesos 上, 由资源管理系统来负责服务本身的容错.

在单机环境下要开展大数据实验, 只需要配置 Hadoop 的单机模式就可以开展 MapReduce 程序的逻辑调试. 如果想看一下 HDFS 的实际效果, 以及分布式环境下多个守护进程协调运行的效果, 则可以在单机的 Hadoop 伪分布式模式下进行测试. 但是如果数据量很大, 且需要验证分布式处理的效果, 则需要采用 Hadoop 全分布式模式, 理论上 Hadoop 生产环境至少需要 1 个 NameNode, 1 个 Secondary NameNode, 以及 3 个 DataNode.

如果想在单机上开展 Spark 的实验, 可以在 Spark Standalone 模式下在主机上同时启动 Master 进程和 Worker 进程, 也就是在主机上实现 Spark 伪分布式部署, 便可以开展 Spark 程序的逻辑调试. 但是如果要在 Hadoop 分布式环境下测试程序, 则必须采用 Spark on YARN 模式实现 Spark 与 Hadoop 平台的对接.

3.4 基于虚拟化技术的单机大数据实验平台

在开展 Hadoop 全分布式模式实验时, 需要数台计算机构成集群来实现大数据平台. 然而, 为每个学生配备几台物理机来做实验并不现实, 主要是因为成本过高、维护不易, 而且资源利用率很低. 此时, 可以在单机上采用虚拟化技术, 从一台物理机中虚拟出好几台虚拟机, 并将这些虚拟机通过虚拟网络连接成大数据集群.

图 8 是基于虚拟化技术的单机大数据实验平台架构. 虚拟化技术就是指将宿主物理主机上内存、CPU、存储、网络等硬件资源通过虚拟化管理程序 (hypervisor) 统一调度, 并分配给多台虚拟机使用, 虚拟机之间虽然共享宿主机上的硬件资源, 但是相互之间却互不干扰, 在逻辑上相互独立. 虚拟化技术实现了物理资源到逻辑资源的转化, 解决了物理资源使用效率低、成本高等问题.

单机上常用的虚拟化软件有商业软件 VMware workstation, 以及开源的 Oracle VM VirtualBox 等. 虚拟机可以通过拷贝文件或文件夹的形式在物理机之间相互复制, 因此只需要制作一套虚拟机即可在所有的物理机之间共享, 可以节省大量的系统安装时间.

由于所有虚拟机实际上都在使用宿主机的资源, 因此如果想在宿主机上仿真一个集群, 宿主机的配置必须高于所有虚拟机配置的总和, 例如: 如果同时启动 5 台虚拟机, 每台平均分配内存 4 GB、硬盘 50 GB、1 个 CPU 核, 则宿主机的配置应是内存不少于 24~32 GB、硬盘不少于 300 GB、CPU 不少于 6~8 核.

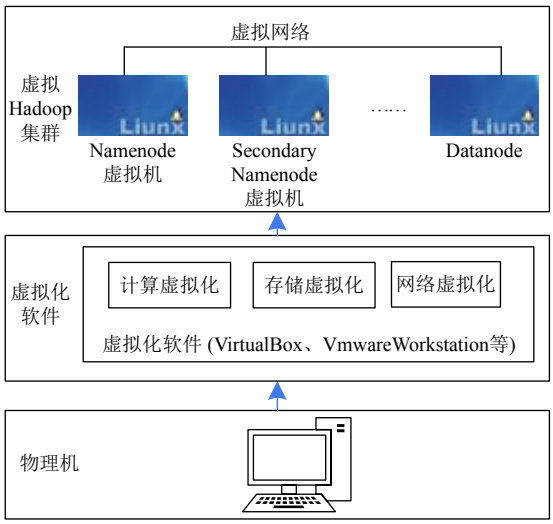

图 8 基于虚拟化技术的单机大数据实验平台架构

3.5 共享大数据集群的实验平台

如果实验的内容不是侧重于大数据平台的安装、配置和运维, 而是侧重于大数据平台的使用及分布式应用程序的开发, 就没有必要为每个学生配置一套大数据集群, 只需要所有学生共享一套大数据集群即可 (如图 9 所示). 大数据集群向所有学生终端机提供包括 HDFS、NoSQL 数据库、Spark 计算的各类大数据服务, 学生终端机只需要安装客户端软件及 IDE (集成开发环境), 用指定的用户账号登录, 即可通过 Hadoop CLI (Command-Line Interface, 命令行界面) 或 API 访问大数据集群.

3.6 基于云计算的共享大数据实验平台

虽然单机上的虚拟化技术可以很好地用物理机虚拟出大数据实验集群, 但是在这种方式中, 所有的负载均由宿主机承担, 因此对宿主机的配置要求比较高. 然而, 许多高校使用的电脑并没有那么高的配置, 无法在单机上实现虚拟机集群, 此时, 应采用云计算技术来实现共享大数据实验平台.

云计算的核心技术也是虚拟化技术, 但是与单机虚拟化不同的是, 其后台是由云平台管理系统统一管理的物理机集群共同充当提供资源的宿主机, 生成的虚拟机也是在物理机集群上运行, 客户端通过网络来访问云平台中的虚拟机 (如 DataNode 虚拟节点), 或访问云平台提供的服务 (如 NoSQL 数据库).

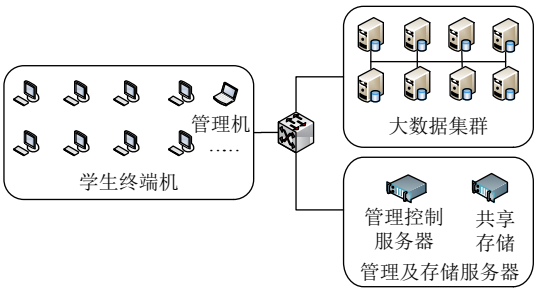

图 9 共享大数据集群的实验平台

云计算可以分为公有云、私有云和混合云. 公有云通常指第三方提供的、一般需通过 Internet 访问的云. 私有云是企业内部单独建立和使用的云, 通常部署在企业数据中心的主机托管场所, 位于防火墙内. 私有云极大地保障了云平台的安全, 但是需要稳定的云平台部署场地, 以及一支比较专业的云平台硬件、软件维护队伍.

高校在建设大数据实验室的时候需根据自己的经费预算和实际需要, 来选择是使用公有云还是自建大数据私有云. 如果是一两个月就可结课的短期课程, 有移动性需求 (如疫情期间开展网络实验), Internet 速度和稳定性能够得到保证的情况下, 可以从公有云租借 Linux 服务器或 Hadoop 集群开展大数据实验. 公有云服务商有阿里云、百度云、腾讯云、华为云、亚马逊云等. 而有固定实验场所以及服务器机房的院校可搭建大数据私有云.

云平台的虚拟化通常采用应用容器引擎 Docker 来实现, Docker 容器是当前最主流的云服务解决方案. 与 KVM、Xen 等云平台生成虚拟机时会包括完整操作系统不同, Docker 容器是建立在操作系统上的轻量级虚拟化技术, 直接和宿主机的操作系统内核交互, 性能损耗较少, 容器的创建和启动都很迅速. 使用 Docker 容器打包和快速运行大数据集群, 可以节省大量的安装、配置系统、设置参数及运行的时间. 为了便于管理, 常采用大规模容器编排管理框架 Kubernetes 来统一规划和部署 Docker 容器.

图 10 是大数据云平台的体系结构, 底层是云平台物理集群的物理硬件资源, 包括 CPU 运算资源、阵列存储资源、网卡网络资源. 再上一层便是虚拟化管理程序, 负责将所有的硬件资源虚拟化并放入统一管理和分配的虚拟资源池, 然后通过虚拟机来对用户需要的计算资源、存储资源、网络资源进行定制和封装, 并通过网络提交给终端用户使用.

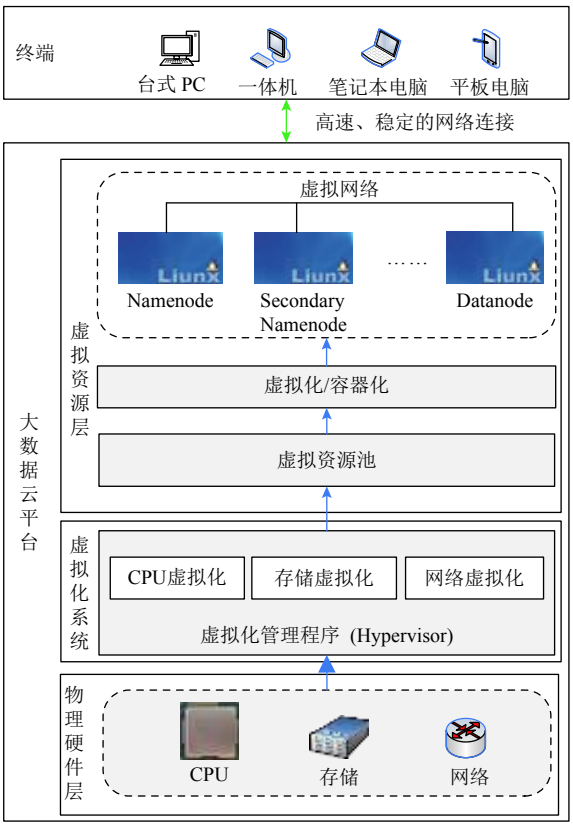

图 10 大数据云平台的体系结构

相关的软件及特定的配置以容器镜像的形式存储, 在上实验课之前, 可快速地从容器镜像中克隆出大量的虚拟机并供学生使用, 每个学生可根据需要得到一个实验虚拟机集群, 集群间相互隔离、互不干扰, 实验结束后, 所有虚拟机被回收, 其资源被重新放回云平台的资源池供下次分配使用.

云计算从低到高分为 IaaS (Infrastructure as a Service, 基础设施即服务)、PaaS (Platform as a Service, 平台即服务)、SaaS (Software as a Service, 软件即服务) 3 层. IaaS 是最底层云服务, 提供一些基础资源服务, PaaS 提供软件研发平台, SaaS 将软件的开发、管理、部署都交给第三方, 用户不需要关心技术问题.

大数据实验云平台可以提供纯的 Linux 云服务器 (IaaS 级, 需学生自己进一步安装 Hadoop、Spark 各类组件), 或是已经预装不同组件的大数据云服务器 (PaaS 级, 已经预装了各类组件, 直接提供 HDFS、Hive 等服务), 或具体的大数据应用 (SaaS 级, 如大数据实训项目).

大数据私有云由一台管理服务器、若干计算服务器, 以及众多学生实验终端机构成. 表 2 是 60 人大数

据实验教学云平台的主要硬件设备配置示例 (不含网络设备、机柜、软件). 计算服务器的数量与实验人数成正比, 若满足 150 人上机, 大约需要管理服务器 2~3 台、计算服务器 15 台及 150 台学生终端机 (设备数量及配置仅供参考, 具体实施时需根据实际负载来确定).

表 2 大数据云平台主要硬件设备配置示例

设备名称	数量	作用	配置
管理服务器	1	管理云平台所有节点的资源分配、管理云平台中容器的元数据、用户、镜像及生命周期	双路Intel E5 6核CPU/64 GB DDR4内存/1块240 GB SSD固态/2块4T SATA硬盘/RAID卡/2~4个千兆网口/PCI-E3.0总线/双电源800W
计算服务器	6~8	提供当前节点的容器运行环境、监控并向管理节点汇报容器生存情况	双路Intel E5 6核CPU/64 GB DDR4内存/1块240 GB SSD固态/1块4 TB SATA硬盘/RAID卡/2~4个千兆网口/PCI-E3.0总线/双电源800 W
学生实验终端机	60	作为客户端访问大数据虚拟集群, 安装IDE	单CPU/2~4 GB以上内存/200 GB以上硬盘/1个千兆网口

图 11 是一个基于大数据云平台的典型实验过程. 实验用户可以申请不同配置的 Hadoop/Spark 集群. 申请成功后, 云平台会从资源池中划出资源, 分配给从容器镜像中克隆出来的 Hadoop 容器集群. 此时实验用户可将实验数据从网络存储 (如果实验数据和容器分开存放的话) 中加载到 Hadoop 集群的 HDFS 中. 然后实验用户可以在学生终端机上, 通过 Hadoop CLI、spark-shell 命令、或通过装有 Java、Scala、Python、R 等语言模块的 IDE, 输入大数据操作的各种命令, 或开发大数据应用程序. 在实验过程中, 可从分析挖掘算法库中调入事先做好的大数据分析挖掘算法, 最后的分析结果以报表、图表等可视化形式展示. 为了方便数据共享, 容器镜像库、实验数据库、分析挖掘算法库可存储在网络云盘或对象存储服务器中.

4 大数据实验课程体系及实验项目设计

在大数据实验的硬件环境基础上, 还需要科学设计大数据实验课程体系及实验项目, 必须以满足未来用人单位对大数据人才专业技能的需求为导向, 分析大数据专业理论课的知识点分布, 科学、合理地构建实验体系并设计实验项目. 图 12 是大数据实验课程体系, 表 3 是大数据实验项目的列表.

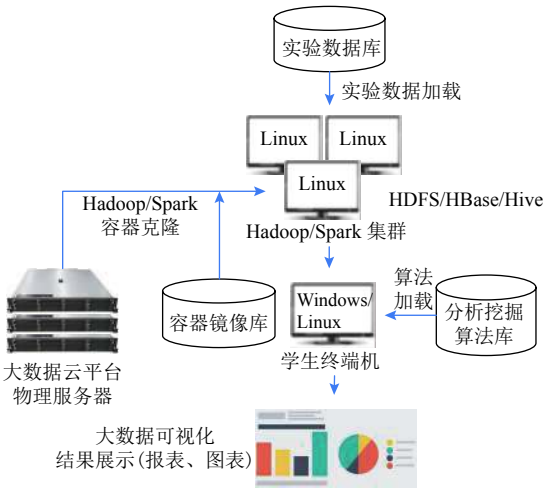

图 11 基于大数据云平台的典型实验过程

为了全方面培养大数据人才的专业实践技能. 大数据实验体系应包括以下 4 类实验项目:

(1) 基础实验: 包括计算机科学基础、编程语言、数据采集及预处理、数学统计分析等实验.

(2) 大数据基础实验: 包括主流大数据平台 Hadoop 中的各类组件实验, 如 HDFS、Storm、HBase、Hive、Redis、MongoDB、Flume、Sqoop、Kafka、YARN、Oozie、Zookeeper 等实验.

(3) 大数据进阶实验: 包括主流大数据分布式计算框架 Spark、Mapreduce 的编程实验, Mahout、MLlib 等机器学习组件实验, 及数据仓库和联机分析处理实验.

(4) 大数据实战项目: 结合不同行业数据的实战项目, 如电子商务、物联网、农业、环境监控、金融等领域的大数据分析实战项目.

5 结语

建设一个软硬件搭配合理、性价比高的大数据实验室, 必须考虑众多因素. 本科院校、专科院校大数据专业的培养目标互不相同, 不同院校及专业需要考虑毕业生将来的职业定位以及所需要的知识结构, 有针对性地开设相关课程及实验项目, 以充实学生的专业知识、提高学生的大数据实践技能. 同时, 还要考虑业界主流的大数据技术的生态系统, 选取被业界认可的通用框架作为实验平台. 根据预算及现有的实验条件, 合理制定大数据实验室的建设方案. 院校可以建立基于单机工作站的实验环境, 也可以建立扩展性和可管理性更高的、基于私有云或公有云的实验环境. 除硬件环境外, 大数据实验体系及实验项目等软环境的建

设也是大数据实验室建设的重要内容. 随着大数据技术的不断发展, 院校在大数据的实验环境、实验内容方面要不断推陈出新, 才能够源源不断地向社会输送合格的大数据专业人才.

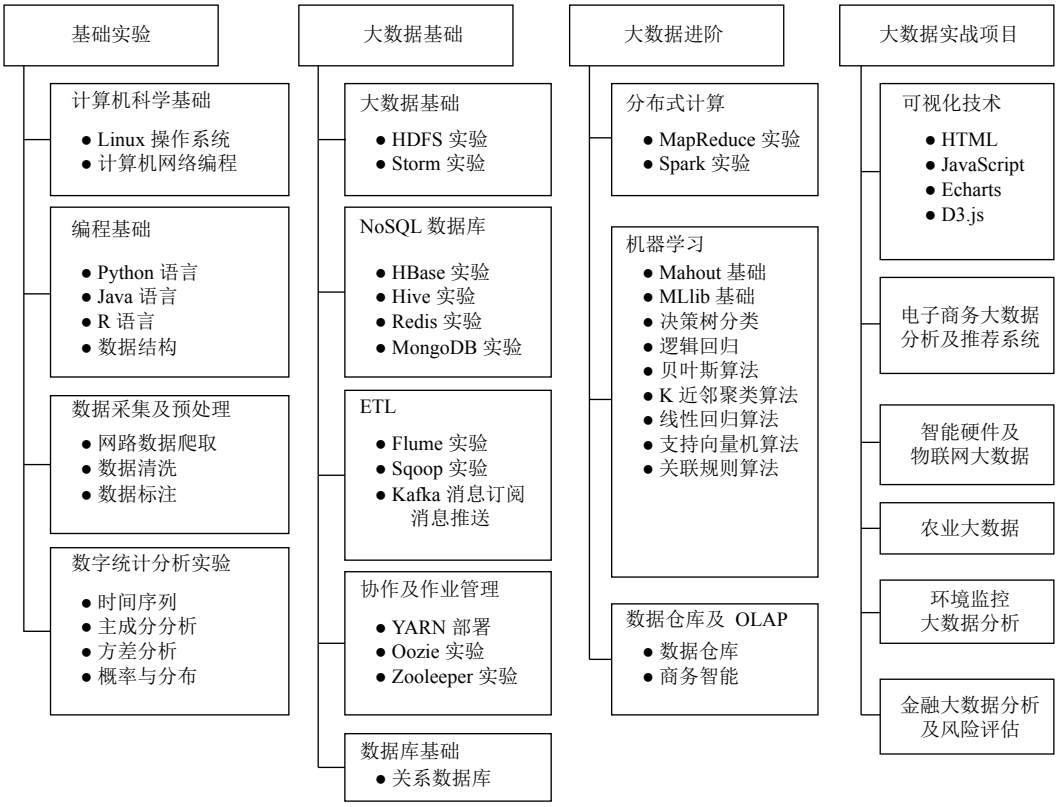

图 12 大数据实验课程体系

表 3 大数据实验项目

实验分类	实验项目
计算机科学基础	Linux操作系统实验: Linux常用命令、文件操作、文本编辑器vi、shell命令、主机及进程状态检测、telnet远程访问、awk文本处理工具、正则表达式. 计算机网络编程实验: 基于TCP协议的socket编程、RPC通信原理.
编程语言	Python实验: 列表、元组、字典、函数和程序包、Numpy数组操作、Matplotlib数据可视化、Pandas数据框操作. Java实验: 类的定义及对象的实例化、文件操作等. R语言实验: 控制语句、数据类型、常用函数、程序调用、文件操作、绘图等.
数据采集与预处理	网络数据爬取实验: 网络爬虫爬取电影、小说信息. 数据清洗实验: Kettle工具的使用, CSV、Excel、JSON、XML文件的抽取, 数据平滑、数据筛选与汇总、数据转换与重构, 数据增量更新, 数据脱敏. 数据标注: 遥感影像数据标注, 医疗数据标注, 人脸、车牌、物品的标注.
数学统计分析	基于SAS、SPSS、Matlab等平台的时间序列分析、主成分分析、因子分析、方差分析、概率与分布分析等.
大数据基础	HDFS实验: HDFS读写、WebHDFS访问. Storm实验: Strom部署、实时WordCount Topology设计、Storm app设计.
关系数据库	关系数据库实验: MySQL, 使用SQL创建数据库以及记录的插入、修改、删除、查询.
NOSQL数据库	HBase实验: HBase部署、建表、读写数据, HBase Java API实验. Hive实验: Hive部署、新建Hive表、基于HQL的查询与统计. Redis实验: Redis部署及读写等简单操作. MongoDB实验: 文档数据的读写.

续表 3

实验分类	实验项目
ETL	Flume实验: 使用Flume实时获取日志数据, 并向HDFS导入. Sqoop实验: Sqoop部署、Sqoop实现MySQL与HDFS的数据互导. Kafka实验: Kafka消息订阅及推送.
协作及作业管理	YARN实验: 部署YARN集群. Oozie实验: Oozie部署、Oozie任务流程设计. Zookeeper实验: Zookeeper部署、进程协作.
分布式计算	MapReduce实验: WordCount、SecondarySort、Join等. Spark实验: Spark部署、SparkDemo、Spark-SQL、Spark-streaming、GraphX等.
机器学习	Mahout实验: Mahout部署、基于Mahout实现朴素贝叶斯分类算法、逻辑回归算法、决策树分类算法. MLlib实验: 基于Spark实现SVM(支持向量机)分类算法、K-means聚类算法、FP-growth关联规则算法、ALS协同推荐算法等.
可视化技术	页面编程及可视化实验: HTML、JavaScript、Echarts、D3.js编程等.
大数据实战项目	综合实战: 网络日志分析、电商广告推荐、智能硬件及物联网大数据、农业大数据、环境监控大数据分析、金融大数据分析及风险评估、图片分类等系统.

参考文献

1 中共中央. 中华人民共和国国民经济和社会发展第十三个五年规划纲要 (2016 两会十三五规划纲要). 北京: 人民出版社, 2016.

2 中国政府网. 国务院关于印发促进大数据发展行动纲要的通知国发〔2015〕50 号. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm, 2015-09-05.

3 工业和信息化部网. 工业和信息化部关于印发大数据产业发展规划 (2016–2020 年) 的通知工信部规 [2016]412 号. http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5464999/content.html, 2017-01-17.

4 大数据产业生态联盟 赛迪顾问. 2019 中国大数据产业发展白皮书 (上). 中国计算机报, 2019-12-16(08).

5 大数据产业生态联盟 赛迪顾问. 2019 中国大数据产业发展白皮书 (下). 中国计算机报, 2019-12-30(08).

6 中商情报网. 大数据产业十四五规划展望: 加快关键技术研发, 促进大数据与行业深度融合. https://baijiahao.baidu.com/s?id=1648249458116413373&wfr=spider&for=pc. (2019-10-24)

7 中华人民共和国教育部网. 教育部关于公布 2018 年度普通高等学校本科专业备案和审批结果的通知教高 [2019] 2 号 . http://www.moe.gov.cn/srcsite/A08/moe_1034/s4930/201903/t20190329_376012.html. (2019-03-25)

8 中华人民共和国教育部网. 教育部关于公布 2019 年度普通高等学校本科专业备案和审批结果的通知教高函 [2020] 2 号 . http://www.moe.gov.cn/srcsite/A08/moe_1034/s4930/202003/t20200303_426853.html. (2020-02-25)

9 中华人民共和国教育部网. 教育部关于公布 2020 年高等职业教育专业设置备案和审批结果的通知: 教职成函 [2020] 1 号 . http://www.moe.gov.cn/srcsite/A07/moe_953/202001/t20200122_416286.html. (2020-01-17)

10 汪中, 施培蓓. 数据科学与大数据技术专业建设研究. 安庆师范大学学报 (自然科学版), 2019, 25(1): 117–120.

11 陈江林, 姚继美, 孙永香. 数据科学与大数据技术专业人才培养目标与定位研究. 电脑知识与技术, 2018, 14(35): 166–167.

12 杨灿, 李尹. “1+1+1”高职大数据技术与应用专业人才培养模式研究. 教育现代化, 2019, 6(90): 15–16.

13 任泰明. 高职院校大数据技术与应用专业人才培养方案的探索与实践. 电脑知识与技术, 2019, 15(7): 169–170.