

基于 Hadoop 的《大数据开发技术》课程实践教学研究

王英瑛 乔小燕 柳伟 曲子芳

山东工商学院 山东 烟台 264005

摘要: 随着时代的不断发展,我国教育模式也发生了不同程度的变化,这需要教师在教学过程中,既要提高学生对基础知识内容的掌握能力,又要培养学生的逻辑思维及动手能力。《大数据开发技术》是我校数据科学与大数据技术专业专业的专业核心课程之一,需要教师改善教学模式,拓展教学维度,尤其要加强实践教学环节,提高教学质量。本文主要对基于 Hadoop 的《大数据开发技术》课程实践教学进行分析,提出该课程的实践教学策略。

关键词: Hadoop; 大数据开发技术; 实践教学; 研究

Research on the practical teaching of Big Data Development Technology course based on Hadoop

Yingying Wang Xiaoyan Qiao Wei Liu Zifang Qu

Shandong Technology and Business University, Yantai, Shandong 264005

Abstract: With the continuous development of The Times, different changes have taken place in our educational mode, which needs teachers in the teaching process to improve the students' ability of basic knowledge content, as well as develop students' logical thinking and practical ability. "Big Data Development Technology" is one of the core courses of data Science and Big Data technology major in our school, which requires teachers to improve the teaching mode, expand the teaching dimension, especially strengthen the practical teaching link, and improve the teaching quality. This paper mainly analyzes the practical teaching of "Big Data Development Technology" course based on Hadoop, and puts forward the practical teaching strategy of this course.

Keywords: Hadoop; Big data development technology; Practice teaching; research

信息技术背景下,大数据技术迅猛发展,面对这种发展趋势,我校 2019 年开始招收数据科学与大数据技术专业学生。然而在当前实际教学中,存在开设时间短、资金投入不充足,学生学习质量不佳等问题,因此该课程建设尤其是实践教学环节,更应该充分分析课程实际,改善教学方式,提高教学质量。

一、基于 Hadoop 的《大数据开发技术》实践教学内容

为提升《大数据开发技术》实践课程教学质量,首先需要对 Hadoop 知识体系进行分析。基于 Hadoop 的《大数据开发技术》课堂教学内容有以下六部分: Hadoop 平台安装、Hadoop 分布式文件系统、Hadoop 分布式计算框架、YARN 架构、HBase 分布式数据库、Hadoop 综合实战。教师在教学中,需对其进行多层次分析,明确教学目标,制定合理的实践教学方案。基于 Hadoop 的《大数据开发技术》实践教学,重点需要学生在学习期间认识到计算和存储分离的架构,感受到并行与计算架构在可扩展性指引的优势。学生还需要对大数据的计算特点和存储特点做到了解,对其中出现的信息技术手段能够熟练应用,对 HDFS、MapReduce、YARN、HBase 熟练掌握^[1]。这对后期学生学习其他分布式计算框架具有十分重要的意义。

二、基于 Hadoop 的《大数据开发技术》实验平台建设

1. 转变教学理念,充分利用现有实验平台

我国虽然在互联网教学中已经取得了一定的成果,然而大数据专业开设的时间比较短,很多学校在建立实验平台时,还处于摸索的阶段,这给教师在教学期间带来了不小的难度。所以教师对学生开展实验活动教学时,应针对当前已存在的实验平台进行综合分析,分析其优缺点,选择合适的平台,按照授课内容对照国内外开放的 Hadoop 实验课程进行分析,借鉴相应的经验,这样既可以弥补现有教学中存在的不足,还可以创新教学方式,提高学生的学习质量。当然也要确保实验教学费用在可承受范围之内^[5]。

2. 科学搭建实验教学平台

在课内实践环节,开展实验教学活动时,可以根据教材内容,利用学校实验设备搭建一些可行的实验教学平台环境。比较常见的有两种方式:普通 PC 机搭建环境和新建大数据实验室。普通 PC 机构建环境主要是在实验活动开展的过程中,利用普通 PC 机安装 Linux 虚拟机,通过对这些虚拟机完成相应的 Hadoop 伪分布式环境搭建或者利用完全分布式搭建。从实验过程来看,这种普通 PC 机在构建环境的过程中,由于对硬件环境的要求比较低,可以应用到较多的范围下,提高实验平台教学效率。这种实验方式,一方面可以锻炼学生对基础内容的理解能力,另外一方面还可以开展简单的实验活动。但这种实验环境和平台在搭建过程中具有一定的缺陷,不论是单机的配置方式、还是单机性能,学生都难以感知到分布式环境的优势,不利于后期综合实践活动的开展。而且

教学时间具有一定的限制, 学生前期需要对集群进行连接, 才可以进行实践, 以 MapReduce 编程为例, 环境搭建会花费较多的时间, 无法保证学生在实验过程中掌握到较多的知识内容。

另外一部分是新建大数据实验室的方式, 进行实验平台搭建, 学校可以选择同其他企业进行合作, 建设相应的大数据实验室, 由于企业的加入, 高校在建设实验平台时, 可以选择 Linux 服务器集群与 Docker 容器技术进行搭建^[6]。这种模式可以给学生提供较多的教学资源, 另外一方面也可为多门课程的开展提供实验平台资源。而且教师在实际上课过程中, 可以根据学生的学习情况, 定制适合学生实验的教学活动。例如 Hadoop 完全分布式实验, 可以根据实验需求进行多方面调整, 从而达到较为理想的教学效果。但是这种情况学生利用平台开展实践活动时, 分配到的虚拟化资源和教学资源相对来说比较有限, 当学生人数较多, 同时占用实验资源时, 很容易出现卡顿的现象, 所以在安排课程时, 需要对学生的学习能力与学习课程内容进行分层, 尽可能避免出现高峰现象。同企业建设实验平台时, 要结合高校的人才培养方案进行分析, 在完善系统的过程中, 能够将教学、实验、监控等这些流程做到全方位规划, 从不同的角度提高学生实践能力。

3. 注重教学环境和教学平台的科学搭配

由于实验环境的限制, 在实践教学中, 教师设计相应的实验教学方案, 满足学生的基本实验需求。例如在综合实践环节, 可以让学生在 PC 机中的伪分布式环境中主动完成编码, 如果数据集中的编码任务数量比较大, 可以让学生在弹性大数据实践环境下进行实验^[7], 可以将学生分成不同的小组, 按照学生的学习能力, 对其中的弹性大数据实践环境资源做到综合性分配, 这种方式下, 既可以提高学生的学习效率, 还可以让学生在实践活动中, 相互协调, 实现共同合作, 搭建可行的实践环境, 从而提高学生的大数据编程能力, 深化学生对 Hadoop 的理解。

三、基于 Hadoop 的《大数据开发技术》实践课程的数据集选择

由于《大数据开发技术》实践任务具有容量较大的特点, 而且对时效性要求比较高, 所以选择合适的数据集是学生实践课程教学质量提升的重要因素^[2]。之前主要采取传统的方式进行工作, 借助大数据平台实现综合处理, 会受到实验设备的限制, 教师需要对部分实验内容进行改革。

1. 控制数据集的数量

教师在选择实验数据集时, 需要注意到学生的学习能力和课时安排, 尽量将数据集大小控制到一千万到两千万条左右。如果教师在数据集选择的过程中, 数量比较小, 学生在实践活动中可以借助 Mysql 等其他的技术手段进行处理, 不利于凸显出分布式计算与分布式存储的优势, 让学生在实践中很容易形成一种 Hadoop 无用的心理。然而教师在设计期间将数据集设计过大时, 受到实

验设备的限制, 导致 HDFS 中的一些配置存储空间不足, 而且会花费学生较多的时间进行运算, 很难在课堂中完成实验活动。

2. 科学选择数据集

基于 Hadoop 的《大数据开发技术》综合实践, 在选择数据的过程中, 要注意数据集的内容具有真实性, 可以供学生做到深层次学习和分析。在选择前, 还需要对数据集的多样性展开分析, 注意结构化与非结构化的数据分析, 而且其中也含有一定的噪声数据, 这些数据可以避免学生在实践过程中, 得到千篇一律的结果, 调动学生的学习积极性, 激发学生的学习欲望, 保证学生可以在实践活动中掌握有关 Hadoop 的计算架构方式, 进一步提高学生的学习效率。

3. 学生自行选择数据集

对于有能力的同学想进行深层次学习, 对数据有时间和精力进行探索。教师应该对学生讲解如何借助相关的资源平台下载数据, 例如: 全球气象数据集、一些开放政府平台数据作为数据集等。学生也可以根据自身的喜好, 利用课后时间对这些数据集做到全面整理和清洗^[4]。

四、结束语

总而言之, 为提高基于 Hadoop 的《大数据开发技术》实践质量, 需要对当前的实验平台进行分析, 认识到实践活动开展中遇到的阻力, 根据学生的学习特点和学校的资源优势, 制定出可行的实践方案。此外, 教师还需要关注数据集在实践活动中的重要价值, 帮助学生在实践过程中, 融会贯通知识内容, 锻炼自身的实践能力。

参考文献:

- [1] 于建. 大数据时代软件开发与维护技术要点及运用实践 [J]. 数字通信世界, 2022(04):188-190.
- [2] 兰娅勋. 大数据技术在计算机软件系统开发中的应用 [J]. 软件, 2021,42(08):98-100.
- [3] 张晓明, 李海生, 陈明, 王芳. 基于 OBE 理念的大大数据技术开发课程教学设计与实践 [J]. 计算机教育, 2021(08):166-170.
- [4] 张敏. 大数据技术在大港油田勘探开发中的应用研究 [J]. 中国管理信息化, 2021,24(14):76-77.
- [5] 徐宏. 基于大数据技术的农业电商大数据开发系统的设计与实现 [D]. 武汉轻工大学, 2021.
- [6] 黄志超. 基于大数据技术的《数据库开发技术》课程项目化教学改革 [J]. 电脑知识与技术, 2021,17(12):129-130+155.
- [7] 米永宁, 耿志杰. 应用大数据技术开发数字档案信息资源的现实困境与策略研究 [J]. 北京档案, 2016(11):16-19.

作者简介: 王英瑛 (1978.10—), 女, 汉族, 黑龙江伊春, 硕士, 山东工商学院, 讲师, 大数据技术。

课题项目: 烟台市校地融合项目: 以信息化特色助力烟台新兴产业发展的信息与计算科学学科体系建设 (03117806)