基于云平台的海量数据管理课程实验教学研究

◆李　贺　袁　航
（西安电子科技大学）

摘要：近年来随着各行业对处理海量数据专业人才的迫切需求，很多高校都陆续开设了海量数据管理课程。然而大部分的教学研究多数侧重教学内容探讨和教学方法及手段的改进，对于实践教学涉及的不是很多。该课程有很多知识点比较抽象、难以直观的进行理解，通过相应的实验教学可以帮助学生通过分析和编写实验代码来加深对相关理论方法的理解。本文针对海量数据管理课程提出了一种基于云平台的实验教学方案，使学生们可以在真实的大数据环境中进行实验学习。

1.目前国内海量数据管理课程实验现状与存在的问题

我国非常重视对大数据技术的研究和学习，在很多国内顶尖高校海量数据管理都已经成为计算机科学与技术专业的核心专业课程，但是目前针对海量数据管理的实验设计方面还缺乏有效可行的成熟方案。由于该课程的特殊性，理论课程的学习只能帮助学生对海量数据管理的整体架构和算法方面的了解，海量数据管理的实验教学不但可以帮助学生对架构和理论的理解还可以促进学生掌握利用海量数据管理技术来解决实际问题的能力。因此，海量数据管理实验是整个教学过程中不可缺少的重要组成部分。

由于海量数据管理课程所涉及的内容比较广泛，有一定的理论基础，该课程有很多知识点比较抽象、难以直观的进行理解，通过相应的实验教学可以帮助学生通过分析和编写实验代码来加深对相关理论方法的理解，从而进一步加深理解一些特殊的框架结构在计算机解决实际问题中的重要作用，提高学生利用计算机解决问题的能力和软件开发的能力。通过海量数据管理的实验教学，教学主体会发生改变。教师由传统的课堂教学中的知识讲授者变成信息组织者，成为学生学习的引导者；学生的学习由以老师讲授为主的被动学习；学生变为主动的学习；学生成为学习的主体，有助于提高学习积极性和培养创新能力。实验教学增强课程之间的衔接，有助于学生对课程体系的整体了解与认识，通过相关实验设计再在计算机上加以实现，进行反复调试，不仅能巩固学生的理论知识，而且有助于提高学生的动手动脑能力，促进学生对后续课程的学习。通过面对面的实验辅导，增进了师生之间的认识和理解，缩短了师生间的距离。实验教学作为一种有效的辅助教学手段，能产生较好的学习效果。将海量数据管理实验教学引人数学当中，可以激发学生的学习兴趣，调动学生的学习积极性，更加有利于学生学习这门课程。因此，在海量数据管理课程的教学过程中增加实验环节是十分有必要的。

2.基于云平台的海量数据管理课程实验教学

由于目前拥有的实验设备不满足海量数据管理课程的实验教学需求，开发一个用于实验教学的云平台是有必要的。云平台将在高性能设备集群中实现，学生通过目前实验室的设备连接云平台进行实验。本文提出的海量数据管理课程的实验教学云平台，主要包含一个海量数据的数据源和一个 Hadoop 集群。整个实验过程包括海量数据采集，海量数据存储，海量数据计算，海量数据存储。实验的主体在 Hadoop 集群中完成。

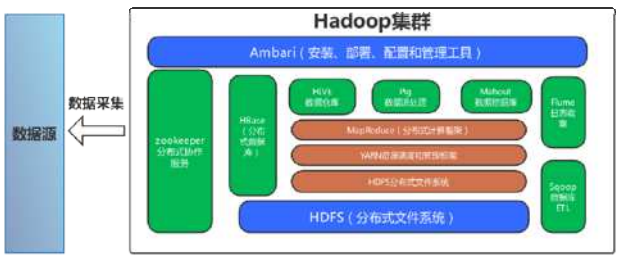

图 1　海量数据管理课程的实验教学云平台架构图

1)海量数据采集：海量数据的采集工作是海量数据管理技术的第一步，通过学习海量数据采集的目标和数据预处理(ETL)技术，掌握数据采集的熟悉常见数据采集系统的组成，掌握海量数据管理中数据的来源分类及特征，熟悉各类数据源的常用数据采集方法。在这个基础上了解互联网数据的抓取方法，熟悉常用网络爬虫的实现方法和工具，了解文档(结构化)数据的采集方法。在进行海量数据抓去的过程中运用数据预处理技术，这需要学习常用的数据预处理方法和工具。

2)海量数据存储：海量数据存储是海量数据管理技术的核心内容，也是整个实验内容中最重要的部分。通过学习数据存储技术的发展历史及各阶段的代表性技术(SAN,NAS 等)，了解存储架构的演进的趋势，掌握云数据中心的特征、架构，了解云数据中心的部署、管理、容灾备份及运维。掌握 nosql 的基本概念、核心思想，通过分析常用 nosql 系统的特点，总结 nosql 与关系型数据库的区别，熟悉数据存储中的 CAP 理论。通过学习 Hadoop 平台的架构及组成，掌握 HDFS 的原理及基本指令，能够熟练配置伪分布式 Hadoop 和完全分布式的 Hadoop。在 Hadoop 的基础之上学习 Hbase 的原理及基本操作指令，能够在 Hadoop 的基础上配置 Habase，并根据 Hbase 的原理结合实际问题设计 Hbase 数据库。

3)海量数据计算：通过学习分布式计算的概念，了解分布式计算和集中式计算的区别，重点是学习 MapReduce 编程框架及其工作过程，能够利用 MapReduce 的编程方法解决实际应用中的海量数据计算问题。根据给定的特定环境，在 Hadoop 系统的基础上设计并编写 MapReduce 程序。

4)海量数据检索：海量数据检索是整个海量数据管理系统中提升数据查询效率的关键步骤，数据量越大，索引的重要性越能体现出来。通过重点学习 Hash 法，外存排序方法、B 树、B+树、R 树、KD 树和倒排索引等索引技术，分析各种索引对于不同的数据检索环境的效果，能够利用 Hive 根据实际应用设计数据仓库来对海量数据进行高效的管理。

结语

海量数据管理课程具有技术性强、涉及内容广泛和贴近编程等特性，增加实验教学可以激发学生的学习兴趣，增进对相应知识点的理解，能有效提高学生的编程能力，提高课程教学质量。本文提出了海量数据管理课程的实验教学云平台，脱离了普通的实验环境，使学生能够真正体会大数据的意义。学生在做实验的过程中，会从数据采集、数据存储、编写 MapReduce 程序、调试和验证等各个环节进行训练。这样能够更深刻地理解和牢固地掌握海量数据管理的整个体系结构，培养学生用编程技术解决实际问题的能力。由于该课程在全国各个高校都处于刚起步阶段，在对实验教学的研究方面还有大量工作要做，需要制定切实可行的课程实验教学大纲，研究有效的实验教学计划，进一步完善实验教学的各个环节，加强整个课程体系的建设，不断地改进教学模式改革。

作者简介：李贺（1983 年 1 月–），男，汉族，籍贯：河南，职称：副教授，学历：博士，研究方向：数据挖掘，单位：西安电子科技大学计算机科学与技术学院。