基于“1+1+N”体系的“大数据采集与集成”课程教学改革与实践*

刘宁

北京林业大学信息学院，北京 100091

摘 要 本文根据工程教育认证理念，基于大数据专业培养应用型、创新型人才的定位，针对当前“大数据采集与集成”课程存在的理论体系不完备、实验体系不合理、考核指标不完善的现状，提出课程实践教学改革思路。通过构建一套“1+1+N”的教学体系，在完善课程理论体系、增加单元实验及综合实验案例以及完善过程化考核等方面进行改革与实践，以解决目前课程存在的重理论轻实践、重标准化考核轻创新能力培养的问题。课程教学改革效果显著，有效增加了学生对大数据专业知识体系的构建，提高了学生对课程学习的主观能动性，学生的创新能力和综合解决实际问题的能力得到了显著提升。

关键字 大数据采集，理论体系构建，综合实验案例，过程化考核

Teaching Reform and Practice of the "Big Data Collection and Integration" Course Based on the "1+1+N" System

Liu Ning

School of Information Science and Technology of Beijing Forestry University
Beijing 100091, China
liuning0928@bjfu.edu.cn

Abstract—This paper proposes a reform idea for the practical teaching of the current "Big Data Collection and Integration" course based on the concept of engineering education certification and the positioning of cultivating applied and innovative talents through big data majors. It addresses the current situation of incomplete theoretical systems, unreasonable experimental systems, and imperfect assessment indicators in the course. By constructing a "1+1+N" teaching system, we have carried out reforms and practices in improving the theoretical system of courses, adding unit experiments and comprehensive experimental cases, and improving process-based assessment, in order to address the current problems of emphasizing theory over practice and standardized assessment over innovation ability cultivation in the curriculum. The curriculum teaching reform has achieved remarkable results, effectively increasing students' understanding of the construction of big data professional knowledge system, enhancing students' subjective initiative in learning the curriculum, and significantly improving students' innovative ability and comprehensive ability to solve practical problems.

Keywords—Big data collection, construction of theoretical system, comprehensive experimental cases, process-based assessment

1 引 言

“大数据采集与集成”是数据科学与大数据技术专业的一门选修课程，是大数据理论体系中重要的环节，主要研究大数据处理过程中的数据采集、数据传输、数据集成和数据预处理，这是大数据处理中的核心内容，也是开展数据可视化、数据建模、数据分析等工作的重要先驱步骤。该课程具有章节内容相对独立、实践性强的特点。因此，在授课过程中，注重课程理论知识体系的构建，将通过单元实验及综合案例实验将课程实践与理论讲述有机结合[1]，不仅可以让学生加深对课程理论知识的理解，还可以锻炼学生解决数据采集、传输、集成、预处理各环节问题的能力，提升学生专业实践及综合解决问题的能力，具有重要的研究价值。

2 “大数据采集与集成”课程问题

2．1 各章理论相对独立，体系化程度不足

“大数据采集与集成”理论部分主要涉及数据采集、数据传输、数据集成和数据预处理四大核心模

*基金资助：本文得到北京林业大学教育教学改革与研究项目“‘大数据采集与集成’课程资源建设与实践教学研究”（BJFU2022JY083）资助。

块，具有知识点多、实践性强以及各模块相对独立的特点。这使教师在授课过程中倾向于对各概念、理论、实践工具的详解，学生被动灌输大量概念、理论、函数及其参数含义等多且杂的知识点。因缺乏对课程知识体系的理解和对综合案例的实践，会导致学生不能深刻理解每部分理论知识在整个大数据处理中的环节及作用，对于各知识点细节记忆不深，且不同核心模块内容相对独立，可能会导致学生“学新忘旧”的现象，很难对课程建立体系化认知。

2．2 实验设计不合理，缺乏综合实验案例

目前课程的实验不合理的地方主要有三点。首先，实验均为各章节内的简单模拟实验，缺乏真实应用背景，使学生很难通过实验了解各部分内容在大数据处理流程中的真实作用，不能充分锻炼学生的动手实践能力；其次，实验设计均为设计好的固定流程，缺少让学生主动思考及探索的元素，使得实验流于形式，不能充分调动学生自主思考、主动学习的能力；最后，缺乏将各章节内容贯穿到一起的综合实验，因而不能让学生通过实验对课程知识体系有更深入的理解，也不能有效锻炼学生利用所学知识综合解决问题的能力。

2．3 考核体系不完备，缺乏过程性考核

“大数据采集与集成”当前的考核方式，以考勤、日常作业、实验考查及考试为主，考核指标体系并不完善，缺乏对学生主动学习、综合解决问题能力等方面的考核，不能有效激发学生对课程的学习兴趣，使得学习效果大打折扣。

3 课程教学改革思路与实践方法

3．1 “1+1+N”教学体系

根据工程教育认证的理念[8,9]，本文针对当前“大数据采集与集成”课程存在的理论体系不完备、实验体系不合理、考核指标不完善的现状，提出一套基于“1+1+N”的教学体系：

在理论知识方面，构建一个以大数据处理流程为核心的课程知识体系。让学生理解课程内容在大数据处理中的环节及重要意义[2]，帮助学生掌握大数据处理任务中的数据采集、传输、集成及预处理的基本概念、方法及常用工具。

在实验设计方面，完成一个真实业务需求下的单元实验及综合实验案例。培养和锻炼学生利用课程知识和工具综合解决实际问题的能力[3,4]，让学生增加对理论知识体系及大数据处理流程中每个步骤意义和作用的理解。

在考核体系方面，完善从知识掌握、实验水平、创新思维[7]、学习态度这四个维度以及细分维度构成对学生全方位的考核。

3．2 理论知识体系

按照数据采集与集成的处理过程，将课程内容分为：数据采集、数据传输、数据集成、数据预处理四个部分，如图 1 所示。图 1 中 “Scrapy 网络爬虫”、“Kafka”、“Flume”、“ETL”、“Pandas”为不同理论模块对应的不同实验工具，每个工具都配备了相应的实验，以巩固对理论知识的理解，以及提升学生动手能力。

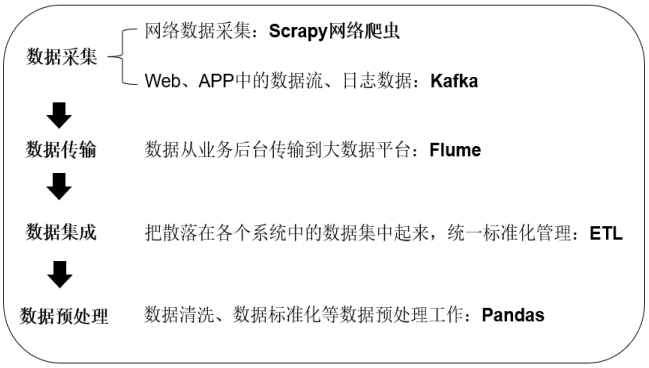

图 1 课程理论知识体系

（1） 数据采集

在数据采集部分，分网络数据采集和 Web、APP 中的数据流日志数据采集两个部分依次介绍。在网络数据采集部分，主要介绍网页基础知识，网络爬虫概述，用 Python 实现 HTTP 请求、定制 requests、解析网页，最后以使用 Scrapy 为案例来介绍网络爬虫的综合使用案例。在 Web、APP 中的数据流日志数据采集部分，主要介绍分布式消息系统 Kafka，包括 Kafka 简介、Kafka 在大数据生态系统中的作用、Kafka 与 Flume 的区别与联系、Kafka 相关概念、Kafka 安装和使用、使用 Python 操作 Kafka、Kafka 与 MySQL 的组合使用。

通过这部分的讲解，让学生了解在数据获取中，对网络数据源及业务系统数据源的不同采集方式，了解网页基础知识，了解在各类数据采集中遇到的常见问题，以及如何通过使用 Scrapy 和 Kafka 等工具去解决这些问题。

（2） 数据传输

数据传输的主要功能是把大数据从业务后台传输到大数据平台。这部分主要介绍日志采集系统 Flume，包括 Flume 简介、Flume 的安装和使用、Flume

和 Kafka 的组合使用、采集日志文件到 HDFS、采集 MySQL 数据到 HDFS。

通过这部分的讲解，让学生理解为何要把数据从业务后台传输到大数据后台，以及为何要实时进行数据传输，如何实时进行传输。

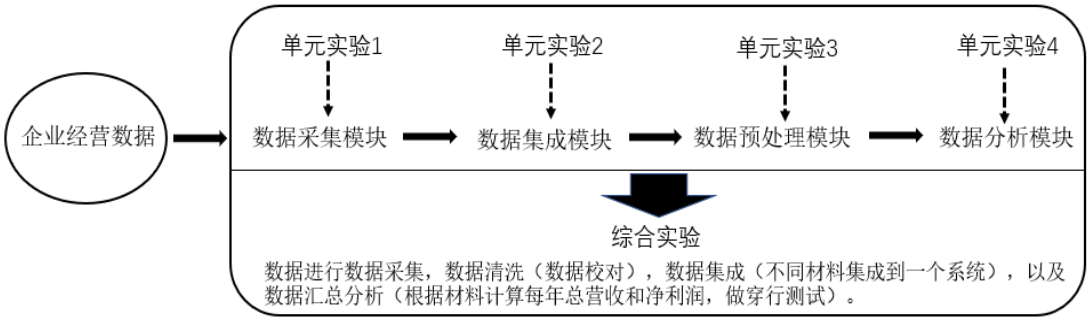

图 2 一贯式单元实验和综合实验设计架构

（3） 数据集成

数据集成部分主要介绍数据仓库中数据集成，即如何把不同数据源中的数据集成到数据仓库。重点介绍数据仓库概念、数据集成、ETL、CDC；第 7 章 ETL 工具 Kettle，介绍 Kettle 的基本概念、Kettle 的基本功能、安装 Kettle、数据抽取、数据清洗与转换、数据加载等内容。

通过这部分的讲解，让学生理解数据集成的必要性，在实际数据集成中的常见问题，如何用 ETL 工具去解决数据集成中的问题，如数据抽取、数据清洗、数据转换等。

（4） 数据预处理

数据预处理部分主要介绍基于 Pandas 的数据清洗过程，重点介绍 NumPy 的基本使用方法、pandas 数据结构、基本功能、汇总和描述统计、处理缺失数据、综合实例等内容。

通过这部分的讲解，主要让学生理解数据清洗的必要性及重要意义，以及基于 Pandas，NumPy 如何对数据进行汇总、统计、处理缺失数据和数据清洗等工作。

以上四个部分的内容构成了本课程的核心理论知识体系。在课程理论讲述的过程中，始终让学生了解每部分知识在整个大数据处理中的环节、重要意义、常见问题[5,6]，让学生对知识体系有清晰的了解。

3．3 实验设计与实践教学

（1） 实践教学工具

结合理论教学知识体系和实践教学任务设计，在课程四大核心模块，均涉及实践教学工具的讲述及实战[15]。数据采集模块的实践教学工具为 Scrapy 网络爬虫及 Kafka；数据传输的实践教学工具为 Flume；数据集成模块的实践教学工具为 ETL 工具；数据预处理模块的实践教学工具为 Pandas。

（2） 实践教学综合案例

如本文上一章节介绍，课程的理论部分由数据采集、数据传输、数据集成和数据预处理四个主要部分组成。

本课程在数据科学专业体系中，具有实践性强的特点，为了巩固学生对课程理论知识的理解，提高学生综合运用所学知识解决实际问题的能力，本文探索以一个大数据应用的真实案例贯穿整个教学过程[11,12]，如图 2 所示。综合考察学生对数据采集、数据传输、数据集成、数据预处理及数据分析这些知识点的掌握情况，增进学生对课程相关知识实际应用的理解[13,14]，提升学生专业实践能力。

（3） 课程综合案例背景介绍

在企业实际经营过程中，需要对过往经营情况做汇总分析，以便核查其经营状况。其中，企业的营收和支出为最为核心的数据。企业的每一笔营收或支出，都会有相应的合同、入库/出库单、发票、银行流水、台账等核心材料构成完成的证据链，在审计企业经营状况中，对每一条营收或支出，都要审计这些材料，是否齐全，以及信息是否一致。只有材料齐全以及各种材料信息一致，才可以认为该条营收/支出为真实有效的，这一过程被称为穿行测试。

（4） 数据及需求分析

现提供某家公司的历史经营数据，包括合同、发票、银行流水三种材料。在处理过程中，先对这些材料进行扫描或拍照，形成 PDF 或者 JPG 等图片文件。需要将上述三种材料的数据进行数据采集，数据清洗（数据校对），数据集成（不同材料集成到一个系统），以及数据汇总分析（根据材料计算每年总营收和净利润，做穿行测试）。

表 1 “大数据采集与集成”多维评价模型

知识掌握	期末测试（15%）
	平时作业完成度（10 %）
	阶段性测试（5 %）
实验水平	实验完成程度（15 %）
	实验报告（10%）
	代码质量（5 %）
综合能力	创新思维（15 %）
	扩展学习（5 %）
	团队合作（5%）
学习态度	出勤率（5 %）
	预习与复习情况（5 %）
	上课表现（5 %）

（5） 综合案例与课程理论的结合

结合综合案例的真实需求及课程理论知识体系，将需求按理论体系拆解为数据采集模块、数据传输及集成模块、数据预处理模块及数据分析模块。

① 数据采集模块

将非结构化的数据进行结构化提取，并保存到数据库中，便于后续的分析。需将银行流水、合同、发票分别做采集和标准化存储。数据库和字段可根据需求自行设计。

② 数据传输及集成模块

需要根据银行流水情况，做企业经营状况汇总。每个企业会在多家银行有账户，我们会获取该企业所有银行账户的银行流水，把所有银行流水下的数据做集成，以便做后续的分析。有明确的需求为：需要存储后的数据支持计算企业每一年的总营收、总净利润；支持各类个性化需求，比如某段时间的营收、净利润，与某企业的总支出和总收入等。

③ 数据预处理模块

因为 OCR 识别有错误，对于识别后的各类数据，需要先进行清洗和校对，再入库。有部分样本为必做样本，有部分样本为选做样本。

④ 数据分析模块

对于每一个合同，需要根据发票、银行流水信息来确认该合同是否是正确的（交易金额及交易方名称对应，即判断为正确）。

（6） 案例重点难点分析

源文件为拍照或扫描文件，有些源文件质量较低，存在对比度低、倾斜、手写或印章噪声污染等问题。OCR 识别会存在一定的错误率，在获取 OCR 识别的结果后，需要对数据进行校对和清洗，然后才能入库；合同等文件为非制式文件，格式不统一，关键词命名也不统一。比如合同中“甲方乙方”，有些合同会称之为“买方卖方”、“需求方、供货方”等，这些都需要统一为标准字段。此外，有些命名实体没有关键词，因此我们找不到对应的标签来进行提取，只能利用基于 NLP 的实体识别（Named Entity Recognition,NER）技术来进行识别。我们需要把非标准的数据进行标准化，即把合同中设计的一些标准字段提取出来，比如合同编号、甲乙方名称、货物名称、总金额、签订日期等。

不同的银行流水版式各异，需要想办法将这些不同版式的银行流水数据进行数据集成。

4 构建“多维度-重实践-考能力”课程考核评价体系

有效的考核方式可以激发学生积极性和创造力，因此建立有效的考核评价体系是对教学质量评估的重要环节[10]。如表 1 所示，本文构建一套“多维度-重实践-考能力”的课程考核评价体系，将考核分为“知识掌握”、“实验水平”、“综合能力”、“学习态度”这四个维度。“知识掌握”细分为“期末测试”、“平时作业完成度”、“阶段性测试”这三个子维度；“实验水平”细分为“实验完成度”、“实验报告”、“代码质量”这三个子维度；“综合能力”细分为“创新思维”、“扩展学习”、“团队合作”三个子维度；“学习态度”细分为“出勤率”、“预习与复习情况”、“上课表现”三个子维度。该考核评价体系强调在教学过程中的全程化考核，加大对学生学习过程的考核，增强对实验完成情况的全面考察，增加对学生创新思维、拓展学习等综合能力的考核。引导学生在学习过程中，重视知识理论体系的构建，通过综合实验提升学生分析问题及解决问题，在学生学习全流程都增加了对学生创新思维、扩展学习等方面的考核，有助于提升学生对课程内容的理解，运用所学知识综合解决问题的能力，锻炼他们的创新思维能力。

5 课程教学改革成效

本文提出的教学实践改革方法强调学生通过对知识的学习来构建对大数据处理流程体系的认知，通过设立一个真实案例，将单元实验和综合实验有机结合，让同学们更加清楚课程涉及的数据采集、数据传输、数据集成和数据预处理内容在整个大数据处理链路上的作用和意义。此外，在理论教学和实验设计过程中，设计多个开放式问题，增强同学们主动学习的能动性，培养他们综合解决问题的能力，培养他们的创新思维能力。后续授课教师也将持续改进，持续深化课程体系构建，优化实验设计，形成“教学-实验-评价”一体化的课程教学方法。

表 2 课程评价统计表

题目	评价人数	很高	高	一般	低	很低
对课程的期望	22	11	10	1	0	0
题目	评价学生人数	非常投入	投入	一般	不太投入	完全不投入
学习课程投入程度	22	12	9	1	0	0
题目	评价学生人数	非常清楚	清楚	基本清楚	不清楚	非常不清楚
知道该课程的用途和意义	22	12	9	1	0	0

从学生反馈来看，和以往的教学效果相比，“大数据采集与集成”课程开展教学改革以来，学生对课程学习的积极性显著增强。表 2 展示了学生对课程的调查问卷统计结果，在填写问卷的 22 个学生中，有 21 个学生表示对该门课程的投入程度为“投入”或“非常投入”，并且通过课程的学习，了解该门课程的用途和意义，占比超过了 95%。学生在评教留言中普遍反映对课程理论体系有了更全面和深入的了解，通过完成课程实验案例，调用了他们主动学习的热情。课后学生在对课程的综合评价中给出了 97.75 的高分，在同类课程中名列前茅。这都充分说明实施教学改革以来，学生对课程有了更清晰的认知，并且学生学习该课程的积极性有了显著的提高。

此外，很多同学基于课程的实践案例，对于理论和实践深入挖掘，在多个技术点实现上提出了不同解决方案，让他们分析问题和解决问题的实践能力都得到了有效提升。“大数据采集与集成”课程结束后，有数名学生组成小组，申请继续开展大数据采集和集成相关的科学研究，并以此为方向，拟申请大学生创业创新项目，撰写学术论文。

6 结束语

“大数据采集与集成”是数据科学与大数据技术专业一门重要的专业课程，课程涉及的数据采集、数据传输、数据集成与数据预处理技术都是数据处理的核心环节，是大数据专业学生培养的重要组成部分。该课程具有较强的实践性，是大数据专业实践中重要的组成部分，对学生构建大数据专业认知具有重要的意义。

本文针对当前课程存在的理论体系化程度不足，实验设计不合理，考核方式不完善等方面的问题，提出了一套“1+1+N”教学体系，在理论知识方面，构建一个以大数据处理流程为核心的知识体系；在实验设计方面，完成一个真实需求下的单元实验及综合实验案例；在考核体系方面，完善从知识掌握、实验水平、创新思维、学习态度这四个维度以及细分维度构成对学生全方位的考核。这不仅让学生对解课程各个理论知识模块在大数据处理中的位置、意义以及需要解决的重难点问题有了更深刻的理解，而且可以有效锻炼学生运用所学知识综合解决大数据采集、集成和预处理问题的能力，从而提高学生的专业实践能力。

参 考 文 献

[1] 陆悠,傅启明,邹恩岑,奚雪峰,张战成. 多课程联动的大数据技术课程实践教学方法研究[J]. 计算机教育, 2019, 5(6).

[2] 徐艳艳,李冬梅,陈志泊. "以赛启教"的"算法设计与分析(双语)"课程教学的探索[J]. 中国林业教育, 2022, 40(4):67-70.

[3] 李维,李昀,陈钊. 促进深度学习的混合式案例教学的探讨——以"IT 项目管理"课程为例[J]. 中国林业教育, 2021,039(004):57-60.

[4] 崔晓晖,张戎,陈俊生,齐建东. 基于混合驱动的本科教学评价指标体系分析系统的构建与应用——以北京市某大学为例[J]. 中国林业教育, 2018, 36(1):5.

[5] 王海燕, 刘润萌. AI 时代背景下"信息服务与信息检索"课程教学改革的探索[J]. 中国林业教育, 2022, 40(4):64-67.

[6] 李昀, 李维, 王海燕. "信息资源管理"课程教学改革的探索[J]. 中国林业教育, 2020, 38(5):54-57.

[7] 程宝雷,樊建席,张广泉. 高质量创新型本科人才的培养实践研究[J]. 计算机技术与教育学报, 2023, 11(4):5-9.

[8] 李文涛,邹彩玲,詹弢,刘士虎. 大数据背景下精准分层教学实践研究[J]. 计算机技术与教育学报, 2023, 11(4):16-21.

[9] 张锦,史长琼,向凌云,黄园媛. 结合工程教育认证的地方高校计算机类专业创新人才培养模式研究[J]. 计算机技术与教育学报, 2023, 11(4):71-76.

[10] 姚怡,梁微,孙宇. 基于 CIPP+AHP 的在线开放课程质量评价体系构建[J]. 计算机技术与教育学报, 2023, 11(3):57-60.

[11] 齐琪,房琛琛,崔舒宁,薄钧戈. “一例到底”在构建计算机求解能力中的探索与实践[J]. 计算机教育, 2024, 7(35):72-77.

[12] 安艳霞,何云峰,张丽. 高校教师实践育人素养的构成及涵养[J]. 中国林业教育, 2024, 42(2):1-6.

[13] 谭貌,段斌,周彦,旷怡. 面向产出落实工程教育认证标准的院系机制与实践[J]. 计算机技术与教育学报, 2023, 11(5):16-20.

[14] 张正鹏,卜丽静,李鹏,谭貌. 基于实践教学云平台的《数字图像处理》课程教学改革探索[J]. 计算机技术与教育学报, 2023, 11(5):27-32.

[15] 许莹,钟雄虎,周旭,肖德贵. 人工智能导论多元融合“五维全覆盖”信息化教学模式探索与实践[J]. 计算机技术与教育学报, 2023, 11(5):41-44.