课程名称：大数据处理与分析：Spark编程实践

课程编码：7337001

课程学分：2学分

课程学时：32学时

适用专业：计算机科学与技术

先修课程：Python或C程序设计、数据库原理、操作系统、高等数学、线性代数、概率论与数理统计

课程类别：专业选修课

《大数据处理与分析：Spark 编程实践》

课程教学大纲

一、课程简介与目标

本课程为计算机专业选修课程，授课对象为计算机科学与技术、数据科学与大数据技术、信息安全等相关专业的本科生。当今时代是大数据时代，大数据已经渗透到当今每一个行业和领域，相较于传统数据，大数据具有体量巨大、模态繁多、关联复杂等特点，对大数据的处理和分析是从大数据中获取价值的决定性因素。本课程将介绍大数据处理和分析的基本概念、大数据处理和分析框架 Spark 等设计与运行的基本原理、大数据处理和分析环境的搭建和使用方法、基本编程方法、流式数据处理的基本方法、大数据分析的基本方法，为学生将来进一步在相关领域深入学习或工作打下良好的理论和实践基础。

1. 课程拟达到的教学目标：该课程的目标是从大数据处理与分析的实践出发、让学生掌握基于 Spark 平台进行大数据处理与分析的基本技能，能够实现对大规模结构化和非结构化数据的基本处理与分析，进而建构大数据处理与分析的基础知识体系。

2. 课程思政目标：本门课程在培养学生专业素质和思维能力的同时，通过将大数据技术和社会需求紧密结合，加深学生对所学知识与社会发展、改善民生、提升国家竞争力之间关系的认识，引导学生勤于学习、善于学习、乐于钻研、勇于探索、立志学好科学技术回馈国家、回馈社会。

二、教学基本内容及基本要求

大数据处理与分析：Spark 编程实践课程共 32 学时，其中理论授课 22 学时、上机教学 10 学时。学时分配如表 1 所示。

表 1 学时分配表

总学时	讲授学时	课内上机
32	22	10

1．课程重点

掌握大数据的基本概念、大数据处理与分析框架基本概念、Spark 的部署和安装、RDD 编程基础、Spark SQL 的基本概念、DataFrame 的基本概念、Spark Streaming 的基本概念、DStream 操作、Spark MLlib 基本使用方法；理解大数据计算模式、流计算概念；能编写基于 Spark 的大数据处理与分析程序；能够上机操作和调试程序。

2．课程难点

RDD 设计与运行原理、大数据处理作业阶段划分的原理和运用、RDD 编程模型、Spark Streaming 编程。

3．课堂教学（22 学时）

表 2 各知识单元教学内容、考核要求和学时分配

第一知识单元　大数据处理与分析技术概述					
学时分配	2 学时	教学方式	课堂讲授，ppt 电子课件，板书		
教学内容				重点	难点
1	大数据的技术背景：了解计算机软硬件的发展历程、IT 基础设施发展的历程、互联网、物联网等数据的激增、互联网应用形态的演化，了解大数据产生的必然性。				
2	大数据的基本概念：什么是大数据、大数据的 5V 特征。			√	
3	大数据处理与分析关键技术的基础知识：大数据技术的不同层面及其功能、大数据的几种经典计算模式。			√	
4	大数据处理和分析的典型应用。				
考核要点	什么是大数据、大数据的 5V 特征、大数据技术的不同层面及其功能、大数据的几种经典计算模式				
第二知识单元　Python 编程基础					
学时分配	4 学时	教学方式	课堂讲授，ppt 电子课件，板书		
教学内容				重点	难点
1	Python 语言概述：Python 变量、数字以及运算符				
2	Python 列表、元组、字符串、字典、集合等基础概念和用法			√	√
3	Python 字符串基本操作：索引、切片、操作符、内置函数			√	√

4	Python 选择与循环：分支结构、迭代器、Rang 函数				
5	Python 函数：参数、return 语句、变量作用域				
考核要点	Python 语言基础，变量类型、运算符、序列类型、字符串操作				
第三知识单元　Spark 基础					
学时分配	2 学时	教学方式	课堂讲授，ppt 电子课件，板书		
教学内容				重点	难点
1	Spark 基本概念、历史、Hadoop 对比				
2	Spark 生态系统、Spark 运行架构、Spark 的部署方式			√	
3	安装 Spark 的几种方式、在 spark-shell 中运行代码、编写 Spark 独立应用程序			√	
4	Spark 集群环境搭建、单机搭建、代码运行方式				
考核要点	Spark 基本运行方式、安装、独立程序编写				
第四知识单元　Spark 编程模型					
学时分配	6 学时	教学方式	课堂讲授，ppt 电子课件，板书		
教学内容				重点	难点
1	MapReduce：基本模型概念、编程常用函数				
2	RDD 编程基础、键值对 RDD、数据读写			√	√
3	Spark SQL 简介、DataFrame 概述、DataFrame 的创建			√	√
4	DataFrame 存储、DataFrame 的常用操作、从 RDD 转换得到 DataFrame			√	√
5	使用 Spark SQL 读写数据库、Spark SQL 基本使用举例				
考核要点	RDD 基本模型概念、数据读写、Spark SQL 读写操作				
第五知识单元　Spark Streaming 和 Spark MLib					
学时分配	8 学时	教学方式	课堂讲授，ppt 电子课件，板书		
教学内容				重点	难点
1	流计算简介、Spark Streaming、DStream 操作概述、基本输入源			√	
2	高级数据源、转换操作、输出操作、Structured Streaming			√	√
3	Spark MLlib 简介、机器学习工作流、特征抽取、转化和选择			√	√
4	分类算法在 Spark MLib 上的使用				
5	聚类算法、机器学习参数调优				
考核要点	Spark Streaming 的概念、数据源、操作、Spark MLlib 的使用、常见功能				

4. 上机教学（10 学时）

1）Linux 系统的安装和常用命令（2 学时）

掌握 Linux 虚拟机的安装方法，利用虚拟机软件加载镜像文件，熟悉 Linux 系统的基本用法，尤其是一些常用命令的使用方法。

2）Python 编程基础回顾与练习（2 学时）

掌握在 Jupyter Notebook 环境中编写 Python 脚本的基本方法、熟悉 Python 的基本语法。

3）Spark 环境搭建与使用方法（2 学时）

掌握 Vim 的使用，掌握 JDK 的安装，掌握 Spark 环境的搭建方法，熟悉 Spark 环境的启动、利用 PySpark 以及 Jupyter Notebook 编写、运行 Spark 的基本方法。

4）通过单词计数程序填空题熟悉 RDD 编程（2 学时）

掌握 RDD 及 pair RDD 的创建，并使用 pair RDD 进行计数，掌握利用 Spark RDD 编程对文件中的单词进行计数的程序编写。

5）使用 RDD 编程进行 Apache Web 服务器日志文件统计分析（2 学时）

进一步熟悉 RDD 及 pair RDD 的创建和使用；学习并掌握利用 Spark RDD 编程进行 Apache Web 服务器日志文件分析。

三、课程采用的教学方法

本课程是大数据处理和分析的入门学习，需要培养学生对大数据处理和分析的兴趣，在理解大数据处理和分析框架运行原理的基础上，掌握大数据处理和分析框架使用、RDD 编程的基本方法，进而逐步建构大数据处理与分析的基础知识体系，为将来进一步在相关领域深入学习或工作打下良好的理论和实践基础。因此，在抓好课堂教学效果的同时，应做好课前预习、课后复习、上机和完成实验报告等环节，通过增强师生间、同学间的多种形式的讨论，来提高课程的教学效果和教学质量。安排有课后答疑、课下讨论、网上讨论等环节。

课程教学方法及具体要求如下：

1. 课堂讲授

1）以能力培养为导向，注重理解大数据处理和分析中的各种概念、原理、方法和技能。为保证教学质量，课堂讲授中应重点突出、点面结合，既要保证完成使广大学生接受完整的程序设计知识体系结构的教学目标，又要针对关键问题、重点内容作较为详尽、多引入实例的透彻讲解，使学生真正领会和掌握本课程的知识要领及技术要点。

2）结合实例和上机教学。为使广大同学对大数据处理和分析思想和方法有更为直观、深刻的认识，应在重点和实践相关内容结合实例进行讲授，对于课程的教学重点或难点，通过编程实践增强感性认识和促进学生认知掌握，安排相应上机题。

3）多媒体课件、板书结合的教学手段与多种教学方法兼施并用。教学方法则采取在教师讲授基本教学内容的过程中适当穿插引入个体针对性提问、集体提问、答疑、讨论等教学形式。

2. 讨论与自学

鼓励同学之间或同学与教师之间针对大数据处理和分析重点和难点内容展开讨论，以使学生掌握知识要点、扩大知识面和培养独立思考能力及创新能力。对于有能力的同学，鼓励其广泛阅读相关书籍、资料，扩大知识结构。

3. 课前预习和课后复习

建议学生课前预习相应教学内容；课后复习以课堂讲授内容为主线。

四、建议教材及教学参考书

1. 教材

[1] 《Spark 快速大数据分析》，人民邮电出版社，Holden Karau 等著，王道远译，ISBN: 9787115403094，2018.7

2. 教学参考书

[1]《Spark 编程基础（Python 版）》，人民邮电出版社，林子雨、郑海山、赖永炫编著，ISBN:978-7-115-52439-3，2020.4

五、知识单元对课程目标的达成度设计

1. 知识单元支撑课程目标情况表

围绕每一个具体的课程目标，从相关支撑知识单元的角度设计不同的考核方式，如下表：

课程目标	知识单元	考核方式设计
目标 1	第一知识单元：大数据处理与分析技术概述 第二知识单元：Python 编程基础	以选择题和问答题方式考核。
目标 2	第三知识单元：Spark 概述 第四知识单元：Spark 编程模型	以编程填空题方式考核。
目标 3	第五知识单元 Spark Streaming 和 Spark MLib	以选择和编程填空方式考核

2. 课程的总体考核方法及量化评定标准

依照每部分知识单元对课程目标的支撑情况设计考核方法与成绩评定，本课程成绩由平时成绩和期末考试成绩两部分组成，以百分制计算，平时成绩占 30%，期末考试为开卷考试，成绩占 70%。平时成绩由考勤和上机成绩确定。

六、其它问题的说明

无。

大纲撰写人：王桂玲、杨中国

大纲审阅人：方英兰

系负责人：段建勇

学院负责人：马礼

制（修）订日期：2021 年 8 月