【大数据挖掘分析与应用】

【Big data mining analysis and application】

一、基本信息

课程代码：【2059337】
课程学分：【3】
面向专业：【计算机科学与技术（云计算）】
课程性质：【系级必修课】
开课院系：【信息技术学院计算机科学与技术系】
使用教材：

教材：【Spark 编程基础 林子雨 人民邮电出版社 2018-08-01】

参考书目：【Hadoop 权威指南（第四版）（美）怀特 清华大学出版社 2017-07-01】、【Spark 大数据技术与应用（第 2 版）肖芳 ，张良均 2022-09-01】、【大数据开发项目实战 祝锡永 ，张良均 2022-09-01】

课程网站网址：https://elearning.gench.edu.cn:8443/
先修课程：【云计算导论】、【Linux 系统应用】、【云网络组建与管理】、【深入浅出统计学】

二、课程简介

该课程是计算机类相关专业的核心课程，是信息技术学院的院级平台课程和重点课程之一。通过课程学习，学生通过识记、理解和应用三个层次来掌握相关知识点，具备掌握和应用 Spark 系统相关理论知识并掌握 Spark 核心组件的安装、部署、研发等能力。本课程以 Scala 作为开发 Spark 应用程序的编程语言，系统介绍了 Spark 编程的基础知识。课程共 8 单元，内容包括大数据技术概述、Scala 语言基础、Spark 的设计与运行原理、Spark 环境搭建和使用方法、RDD 编程、Spark SQL、Spark Streaming、Spark MLlib 等。本课程每个单元都安排了入门级的编程实践操作，以便学生更好地学习和掌握 Spark 编程方法，通过对 Spark 编程体系的系统学习，具备了大数据开发的基本要求，也具备了企业的用人要求。本课程特别注重学生的动手能力，望学生多多操练。

三、选课建议

大数据挖掘分析及应用课程适合计算机类专业云计算方向的学生必修，除了学过计算机相关课程外，学生还必须掌握基础统计知识，从而具备了学好该课程必要的知识。

四、课程与专业毕业要求的关联性

专业毕业要求	关联

LO1：工程知识：能够将数学、自然科学、工程基础和专业知识用于解决复杂工程问题	
LO2：问题分析：能够应用数学、自然科学和工程科学的基本原理，识别、表达、并通过文献研究分析复杂工程问题，以获得有效结论	●
LO3：设计/开发解决方案：能够设计针对复杂工程问题的解决方案，设计满足特定需求的系统、单元（部件）或工艺流程，并能够在设计环节中体现创新意识	
LO4：研究：能够基于科学原理并采用科学方法对复杂工程问题进行研究，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论	●
LO5：使用现代工具：能够针对复杂工程问题，开发、选择与使用恰当的技术、资源、现代工程工具和信息技术工具，包括对复杂工程问题的预测与模拟，并能够理解其局限性	
LO6：工程与社会：能够基于工程相关背景知识进行合理分析，评价专业工程实践和复杂工程问题解决方案对社会、健康、安全、法律以及文化的影响，并理 解应承担的责任	
LO7：环境和可持续发展：能够理解和评价针对复杂工程问题的专业工程实践对环境、社会可持续发展的影响	
LO8：职业规范：具有人文社会科学素养、社会责任感，能够在工程实践中理解并遵守工程职业道德和规范，履行责任	
LO9：个人和团队：能够在多学科背景下的团队中承担个体、团队成员以及负责人的角色	●
LO10：沟通：能够就复杂工程问题与业界同行及社会公众进行有效沟通和交流，包括撰写报告和设计文稿、陈述发言、清晰表达或回应指令。并具备一定的国际视野，能够在跨文化背景下进行沟通和交流	
LO11：项目管理：理解并掌握工程管理原理与经济决策方法，并能在多学科环境中应用	
LO12：终身学习：具有自主学习和终身学习的意识，有不断学习和适应发展的能力	

五、课程目标/课程预期学习成果

序号	课程预期学习成果	课程目标 （细化的预期学习成果）	教与学方式	评价方式
1	L023 具备对复杂工程问题进行分析和求解的能力	1. 了解大数据定义。 2. 熟悉大数据特点。 3. 熟悉大数据处理的挑战。 4. 掌握 Linux 操作系统安装、网络环境配置的技能。 5. 掌握常用 Linux 命令实操技能。 6. 掌握简单 Linux Shell 编程技能。	讲授、练习、实践	大作业
2	L041 能够基于科学原理，结合智能制造行业，具有将智能制造中关于应用系统开发各方面知识集成的能力，并根据实际对系统设计进行优化	1. 了解 Spark 应用场景、Spark 生态圈、核心组件。 2. 理解 Spark 原理和体系架构。 3. 掌握 Spark 安装、部署、配置。	讲授、练习、实践	大作业
3	L091 能够理解团队合作的意义，能与团队成员有效沟通，用人单位评价好	1. 与同学们合作完成大数据项目。 2. 把握好自己在项目中的职责。	实践	大作业

六、课程内容

模块 1　大数据技术概述

通过本章学习，学生能够熟悉大数据概念，了解大数据框架最核心的设计，熟悉大数据的发展历史，熟悉大数据的发展现状，掌握大数据的特点，掌握大数据核心。

重点：大数据的特点，大数据核心。

难点：大数据框架最核心的设计。

理论课时：2

实践课时：0

模块 2　Scala 语言基础

通过本章学习，学生需要理解 Scala 语言概述，掌握 Scala 基础知识，了解面向对象编程基础，了解函数式编程基础，熟悉 Scala 应用场景。

重点：面向对象编程基础，函数式编程基础。

难点：面向对象编程基础，函数式编程基础。

理论课时：4

实践课时：4

模块 3 Spark 的设计与运行原理

通过本章学习，了解Spark设计概念，了解Spark生态系统，掌握Spark运行架构，理解Spark的运行原理。

重点：掌握Spark运行架构，Spark的运行原理。

难点：Spark的运行原理。

理论课时：4

实践课时：4

模块 4 Spark 环境搭建和使用方法

通过本章学习，学生能够掌握 Spark 的单节点搭建，集群搭建，掌握在 Spark 环境运行程序。

重点：Spark 的单节点搭建，集群搭建。

难点：Spark 程序的运行。

理论课时：2

实践课时：4

模块 5 RDD 编程

通过本章学习，学生能够掌握 RDD 编程基础，能够掌握键值对 RDD，掌握数据的读写方式。

重点：RDD 编程基础，数据读写。

难点：RDD 编程。

理论课时：2

实践课时：4

模块 6 Spark SQL

通过本章学习，学生能够了解 Spark 简介信息，了解 DataFrame 数据框架的概述，掌握 DataFrame 的创建、保存等，掌握 DataFrame 的常用操作，掌握使用 Spark SQL 读写数据库。

重点：DataFrame 的常用操作，使用 Spark SQL 读写数据库。

难点：DataFrame 的常用操作，使用 Spark SQL 读写数据库。

理论课时：2

实践课时：4

模块 7 Spark Streaming

通过本章学习，学生能够了解流计算概述，理解 Spark Streaming 的设计，掌握 Spark Streaming 的工作机制，掌握 Spark Streaming 的基础操作。

重点：Spark Streaming 的工作机制，Spark Streaming 的基础操作。

难点：Spark Streaming 的基础操作。

理论课时：2

实践课时：4

模块 8 Spark MLlib

通过本章学习，学生能够了解 Spark 机器学习的库，能够使用 Spark 完成机器学习中的分类、聚类算法等。

重点：Spark 机器学习中的分类、聚类算法。

难点：Spark 机器学习中的分类、聚类算法。

理论课时：2

实践课时：4

七、课内实验名称及基本要求

序号	实验名称	主要内容	实验时数	实验类型	备注

1	Spark 安装配置方法	Spark 集群的安装配置大致为如下流程:选定一台机器作为 Master，在 Master 节点上配置 hadoop 用户、安装 SSH server、安装 Java 环境，在 Master 节点上安装 Hadoop，并完成配置 在其他 Slave 节点上配置 hadoop 用户、安装 SSH server、安装 Java 环境将 Master 节点上的 /usr/local/hadoop 目录复制到其他 Slave 节点上在 Master 节点上开启 Hadoop。	12	验证型	
2	RDD 编程初级实践	（1）spark-shell 交互式编程； （2）编写独立应用程序实现数据去重； （3）编写独立应用程序实现求平均值问题；	12	验证型	

3	Spark SQL 编程初级实践	1. Spark SQL 基本操作； 2. 编程实现将 RDD 转换为 DataFrame； 3. 编程实现利用 DataFrame 读写 MySQL 的数据。	12	验证型	
4	Spark Streaming 编程初级实践	1. 安装 Flume； 2. 使用 Avro 数据源测试 Flume； 3. 使用 netcat 数据源测试 Flume； 4. 使用 Flume 作为 Spark Streaming 数据源。	12	验证型	

八、评价方式与成绩

总评构成（X）	评价方式	占比

X1	期末大作业	40%
X2	实验报告	30%
X3	课后作业	20%
X4	签到与课堂表现	10%

撰写人：胡敏彦　系主任审核签名：戴智明 审核时间：2023 年 2 月