《Spark 大数据分析与实践》课程教学大纲

（课程英文名称）

课程编号：201800522062

学　　分：5 学分

学　　时：50 学时　（其中：讲课学时 40　上机学时：10）

先修课程：

后续课程：数据清洗

适用专业：大数据应用技术

开课部门：

一、课程的性质与目标

《Spark 大数据分析与实践》作为高等院校本、专科计算机相关专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考，是一本适合广大计算机编程爱好者的优秀读物。通过学习课程使得学生掌握 Spark 对大规模数据的交互式分析、编写 Spark 应用以及使用 SparkStreaming 处理高速数据流的主要思想和基本步骤；同时对 Spark 平台的应用与开发的理论知识有所了解，如分布式数据收集、分布式数据存储、分布式数据计算、分布式数据展示。

开设本学科的目的是让学生掌握如何使用大数据分析技术解决特定业务领域的问题。完成本课程学习后能够熟练的应用大数据技术解决企业中的实际生产问题。

二、教学条件要求

操作系统：Center OS

Spark 版本：Spark 2.3.2

开发工具：IDEA

三、课程的主要内容及基本要求

第 1 章 Scala 语言基础

章名	Scala 语言基础	学时	6

学习目标	1. 了解 Scala 的特点 2. 掌握 Scala 和 IDEA 的下载安装 3. 掌握 Scala 的基础语法 4. 掌握 Scala 的数据结构 5. 熟悉 Scala 面向对象的特性 6. 掌握 Scala 的模式匹配与样例类			
知识点	了解	掌握	重点	难点
Scala 的概述	√			
Scala 的下载安装		√		
在 IDEA 开发工具中下载安装 Scala 插件		√		
开发第一个 Scala 程序		√		
Scala 的基础语法		√		
Scala 的数据结构			√	
Scala 面向对象的特性		√		√
Scala 的模式匹配与样例类		√		

第 2 章 Spark 基础

章名	Spark 基础	学时	7
学习目标	1. 掌握 Spark 集群的搭建和配置 2. 掌握 Spark HA 集群的搭建和配置 3. 掌握 Spark 集群架构 4. 理解 Spark 作业提交的工作原理		

知识点	了解	掌握	重点	难点
Spark 概述	√			
Spark 的特点	√			
Spark 应用场景		√		
Spark 与 Hadoop 对比		√		
搭建 Spark 开发环境			√	
Spark 运行架构与原理		√		
体验第一个 Spark 程序		√		
启动 Spark-Shell		√		
IDEA 开发 WordCount 程序		√		

第 3 章 Spark RDD 弹性分布式数据集

章名	Spark RDD 弹性分布式数据集	学时	6
学习目标	1. 了解 HDFS 演变 2. 掌握 HDFS 特点 3. 掌握 HDFS 的架构和原理		

	4. 掌握 HDFS 的 Shell 和 Java Api 操作			
知识点	了解	掌握	重点	难点
RDD 简介	√			
RDD 的创建方式		√		
转换算子		√	√	
行动算子		√	√	
RDD 的分区	√			
RDD 的依赖关系	√			
RDD 机制	√			
DAG 概念	√			
RDD 在 Spark 中的运行流程		√		

第 4 章 Spark SQL 结构化数据文件处理

章名	Spark SQL 结构化数据文件处理	学时	6	
学习目标	1. 理解 Spark SQL 基本概念及其架构 2. 掌握 DataFrame/Dataset 的常用操作 3. 掌握 RDD 转换 DataFrame 的方式 4. 掌握 Spark SQL 操作数据源			
知识点	了解	掌握	重点	难点
Spark SQL 的简介	√			
Spark SQL 的架构		√		√
DataFrame 简介	√			
DataFrame 的创建		√		
DataFrame 的常用操作		√	√	
Dataset 简介	√		√	
Dataset 对象的创建		√		
RDD 转换 DataFrame		√	√	
Spark SQL 操作 MySQL		√		
操作 Hive 数据集		√		

第 5 章 HBase 分布式数据库

章名	HBase 分布式数据库	学时	10	
学习目标	1. 理解 HBase 的数据模型 2. 掌握 HBase 的集群部署 3. 理解 HBase 的架构 4. 理解 HBase 读写数据流程 5. 掌握 HBase 与 Hive 的整合			
知识点	了解	掌握	重点	难点
HBase 的简介	√			
HBase 的数据模型		√		√

HBase 的集群部署		√	√	
HBase 的 Shell 操作		√		
HBase 的 Java API 操作		√		
HBase 的架构		√		√
物理存储	√			
寻址机制	√			
HBase 读写数据流程		√		
HBase 和 Hive 的整合		√	√	

第 6 章 Kafka 分布式发布订阅消息系统

章名	Kafka 分布式发布订阅消息系统	学时	5	
学习目标	1. 掌握基本的消息传递模式 2. 掌握 Kafka 集群部署 3. 掌握 Kafka 基本操作 4. 了解 Kafka Streams API 的使用			
知识点	了解	掌握	重点	难点
消息传递模式简介		√		
Kafka 简介		√		
Kafka 核心组件介绍		√		
Kafka 工作流程分析		√		√
安装 Kafka		√		√
启动 Kafka 服务		√		
基于命令行方式使用 Kafka		√		
基于 Java API 方式使用 Kafka		√		
Kafka Streams 概述		√		
Kafka Streams 开发单词计数		√		

第 7 章 Spark Streaming 实时计算框架

章名	Spark Streaming 实时计算框架	学时	4	
学习目标	1. 了解什么是实时计算 2. 理解 Spark Streaming 工作原理 3. 掌握 DStream 的转换操作 4. 掌握 DStream 的窗口操作 5. 掌握 DStream 的输出操作 6. 掌握 Spark Streaming 和 Kafka 整合			
知识点	了解	掌握	重点	难点
什么是实时计算	√			

常用的实时计算框架	√			
Spark Streaming 简介		√	√	√
Spark Streaming 工作原理		√	√	√
DStream 简介		√		
DStream 编程模型		√	√	√
DStream 转换操作		√	√	
DStream 窗口操作		√	√	
DStream 输出操作		√	√	
DStream 实例—实现网站热词排序		√		
KafkaUtils.createDstream 方式		√		
KafkaUtils.createDirectStream 方式		√		

第 8 章 Spark MLlib 机器学习算法库

章名	Spark MLlib 机器学习算法库	学时	4
学习目标	1. 了解什么是机器学习 2. 掌握机器学习的工作流程 3. 了解 Spark MLlib 的基本使用方式 4. 了解电影推荐系统的构建流程		

知识点	了解	掌握	重点	难点
什么是机器学习	√		√	
机器学习的应用	√			√
MLlib 的简介	√			
Spark 机器学习工作流程	√		√	
本地向量	√			√
标注点	√			√
本地矩阵	√			√
摘要统计	√			√
相关统计	√			√
分层抽样	√			√
线性支持向量机	√			√
逻辑回归	√			√
推荐模型分类	√			
利用 MLlib 实现电影推荐		√	√	

第 9 章 综合案例——Spark 实时交易数据统计

章名	综合案例——Spark 实时交易数据统计	学时	5

学习目标	1. 熟悉 Spark 实时计算系统架构 2. 掌握看板平台开发业务流程 3. 熟悉系统环境搭建步骤 4. 掌握 Redis 和 WebSocket 基本使用方式			
知识点	了解	掌握	重点	难点
系统背景介绍	√			
系统架构设计		√	√	
Redis 介绍	√			
Redis 部署与启动		√		
Redis 操作及命令		√		
模块开发-构建工程结构		√		
模块开发-构建订单系统		√		
模块开发-分析订单数据		√		√
模块开发-数据展示		√		√

四、学时分配

章目	讲课	上机	合计
第 1 章　Scala 语言基础	5 学时	1 学时	6 学时
第 2 章　Spark 基础	6 学时	1 学时	7 学时
第 3 章　Spark RDD 弹性分布式数据集	5 学时	1 学时	6 学时
第 4 章　Spark SQL 结构化数据文件处理	5 学时	1 学时	6 学时
第 5 章　HBase 分布式数据库	5 学时	1 学时	6 学时
第 6 章　Kafka 分布式发布订阅消息系统	4 学时	1 学时	5 学时
第 7 章　Spark Streaming 实时计算框架	3 学时	1 学时	4 学时
第 8 章　Spark MLlib 机器学习算法库	3 学时	1 学时	4 学时
第 9 章　Spark 实时计算案例	4 学时	2 学时	6 学时
合计	40 学时	10 学时	50 学时

五、考核模式与成绩评定办法

本课程为考试课程，期末考试采用百分制的闭卷考试模式。学生的考试成绩由平时成绩（30%）和期末考试（70%）组成，其中，平时成绩包括出勤（5%）、作业（5%）、上机成绩（20%）。

六、选用教材和主要参考书

本大纲是根据教材《Spark 大数据分析与实践》所设计的。

七、大纲说明

本课程的授课模式为：课堂授课+上机，其中，课堂主要采用多媒体的方式进行授课，并且会通过测试题阶段测试学生的掌握程度；上机主要是编写程序，要求学生动手完成指定的程序设计或验证。

撰写人： 审定人：

批准人： 执行时间：