00

第0章 Spark编程基础

授课教师：[李阳源]

15138617705@163.com

主页：https://sites.google.com/view/dalguri



SL-01:先修课程

在学习本课程之前，需要学生已经学习过如下课程

- 大数据技术导论
  - 比如林子雨编著《大数据技术原理与应用（第4版）》
  - 教材官网： https://dblab.xmu.edu.cn/post/bigdata4/
- 数据库系统原理
  - 了解关系数据库与SQL语句
- 编程
  - 比如，Python或者Java，了解面向对象编程基础知识

课程特色





开启Spark学习大门

扫除学习障碍，铺平学习道路

课程内容提要



- Spark是当前热门的大数据处理技术
- 本课程重点介绍Spark的技术原理与编程方法
- Spark支持采用Scala、Java、Python和R语言进行编程，本课程采用Python语言编写Spark应用程序
- 课程内容包括Spark简介、设计与运行原理、大数据实验环境搭建、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等
- 通过本课程的学习，将帮助学生形成基础的Spark应用程序编程能力，为学生使用Spark技术解决实际科研问题和未来从事大数据相关工作奠定基础

教学大纲



章（或节）	主要内容	学时安排
第1章 大数据技术概述	大数据的基本概念、关键技术和代表性软件	2
第2章 Spark的设计与运行原理	Spark简介、运行架构、RDD的设计与运行原理、部署模式	18
第3章 大数据实验环境搭建	Linux系统的安装、Hadoop的安装、MySQL的安装、Kafka的安装、Anaconda的安装	12
第4章Spark环境搭建和使用方法	Spark的基本安装方法，如何在spark-shell中运行代码以及如何开发Spark独立应用程序	8
第5章 RDD编程	RDD编程、键值对RDD、数据读写	20
第6章 Spark SQL	Spark SQL简介、DataFrame、读写Parquet、通过JDBC连接数据库、PySpark和pandas的整合、综合实例	20
~~第7章 Spark Streaming~~	~~流计算简介、Spark Streaming简介、DStream操作~~	~~0~~
第8章 Structured Streaming	Structured Streaming简介、编程方法、输入源、输出操作、容错处理、迟到数据处理、查询的管理和监控	10
第9章 Spark MLlib	Spark MLlib简介、spark.mllib库、spark.ml库	6
合计		

课程配套教学资源



课程实验

实验1　Linux、Hadoop和MySQL的安装和使用

实验2　Spark的安装和使用

实验3-RDD编程初级实践

实验4-Spark SQL编程初级实践

实验5-Spark Streaming编程初级实践

实验6-Spark Structured Streaming编程实践

实验7-Spark MLlib编程初级实践

课程配套教学资源



Spark课程综合实验案例1：淘宝双11数据分析与预测

本案例涉及数据预处理、存储、查询和可视化分析等数据处理全流程所涉及的各种典型操作，涵盖Linux、MySQL、Hadoop、Hive、Sqoop、Eclipse、ECharts、Spark等系统和软件的安装和使用方法

