第2届全国高校大数据教学研讨会特邀报告
大会官网：http://dblab.xmu.edu.cn/post/bdts2018/

大数据处理技术Spark

课程资源和教学经验分享

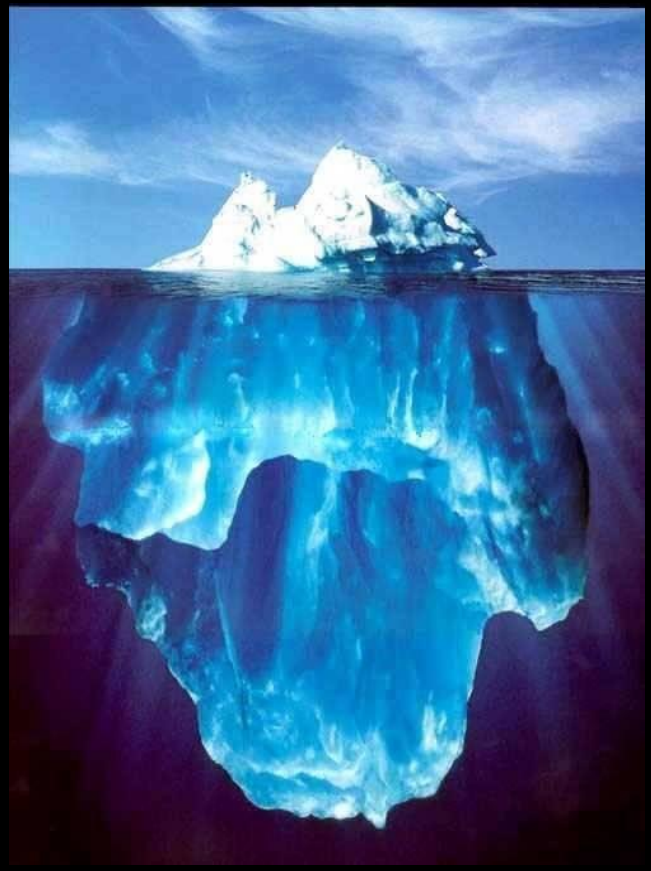
厦门大学 林子雨 博士/助理教授
ziyulin@xmu.edu.cn

2018年5月12日 厦门

报告全文阅读地址：http://dblab.xmu.edu.cn/post/10715/

内容提要

- 高校大数据课程公共服务平台
- 厦门大学建设的大数据课程体系
- 先修课程《大数据技术原理与应用》
- 《Spark编程基础》课程资源与教学经验



高校大数据课程公共服务平台

为高校提供大数据教学一站式服务

- 大数据专业建设方案
- 系列课程教材
- 讲义PPT、习题、实验、案例
- 教师备课指南
- 学生学习指南
- 授课视频
- 教师培训交流
- 大数据教学研讨会

全国高校大数据教学知名品牌

平台构建在厦门大学数据库实验室官网上，在线资源全部免费开放

平台访问地址：http://dblab.xmu.edu.cn/post/8197/

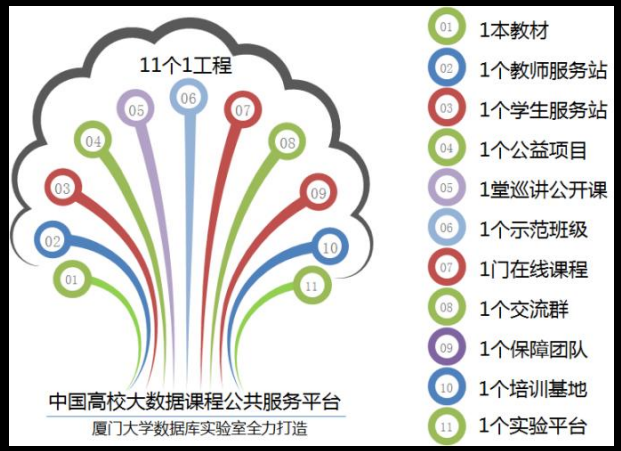
高校大数据课程公共服务平台

建设周期
五年（2013-2018）

投入资金
100万+

高校大数据课程公共服务平台

打造11大工程



平台每年访问量超过100万次

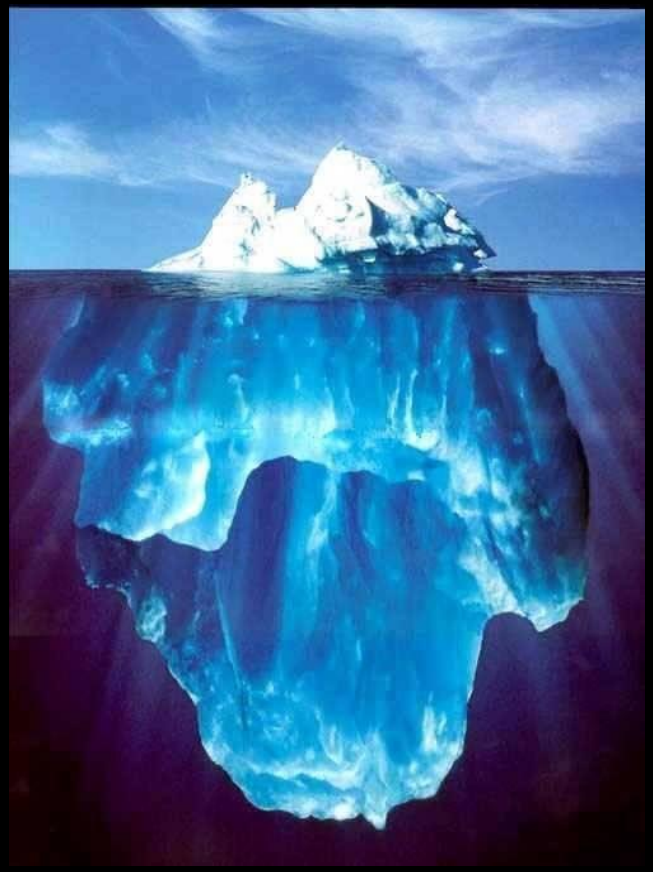
高校大数据课程公共服务平台

学习路线图涵盖了高校大数据课程公共服务平台的大量免费大数据学习资源

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

内容提要

- 高校大数据课程公共服务平台
- 厦门大学建设的大数据课程体系
- 先修课程《大数据技术原理与应用》
- 《Spark编程基础》课程资源与教学经验



厦门大学建设的大数据课程体系

全面训练学生
大数据分析全流程的能力

让学生掌握
一种分布式并行编程框架

引导学生进入大数据世界
由单机环境到分布式环境

学生具备单机编程
开展数据分析的能力

实训　大数据实习实训案例

进阶　Spark编程基础

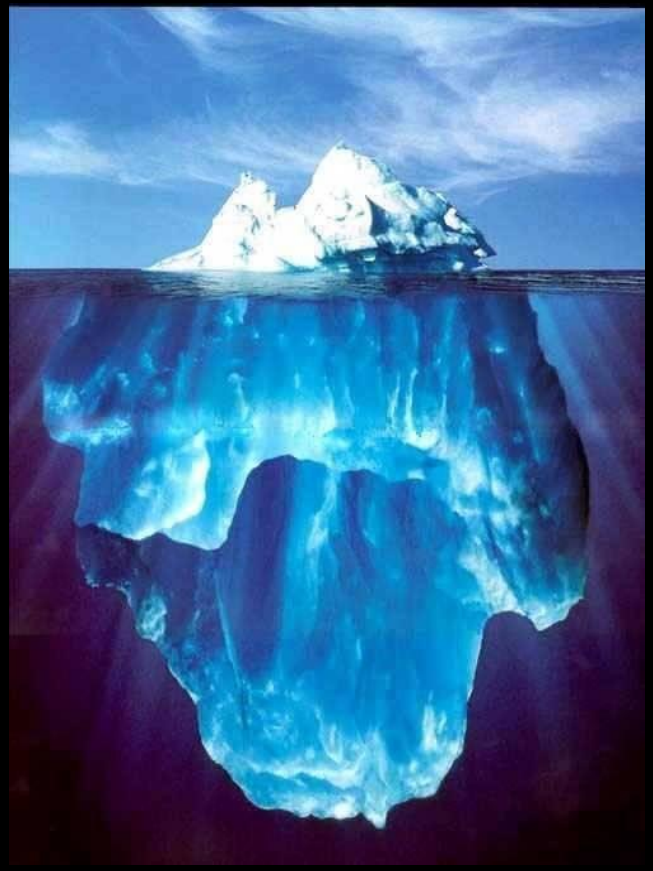
导论　大数据技术原理与应用

数学、编程、算法、数据结构、操作系统、
数据采集与预处理、数据库、数据挖掘

厦门大学建设的课程

内容提要

- 高校大数据课程公共服务平台
- 厦门大学建设的大数据课程体系
- 先修课程《大数据技术原理与应用》
- 《Spark编程基础》课程资源与教学经验



先修课程《大数据技术原理与应用》

课程定位

01

《大数据技术原理与应用》是《Spark编程基础》的先修课程

入门级课程
构建知识体系、阐明基本原理
引导初级实践、了解相关应用

授课对象：

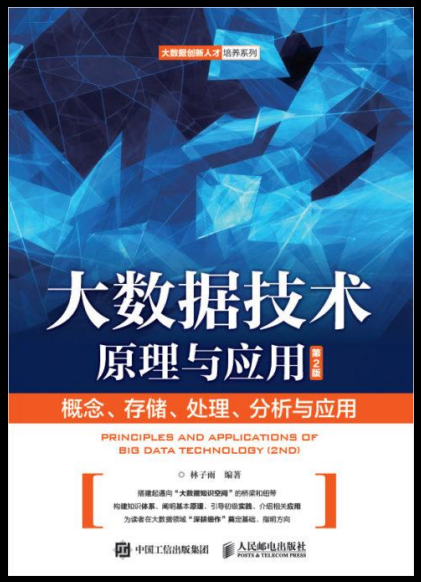
本科生（计算机、软件工程、数据科学与大数据技术）

知识储备：编程、操作系统、数据库

先修课程《大数据技术原理与应用》

教材选择

02



人民邮电出版社
2017年度好书

教材定位为大数据技术入门教材
为读者搭建起通向“大数据知识空间”的桥梁和纽带

- 构建知识体系
- 阐明基本原理
- 引导初级实践
- 了解相关应用

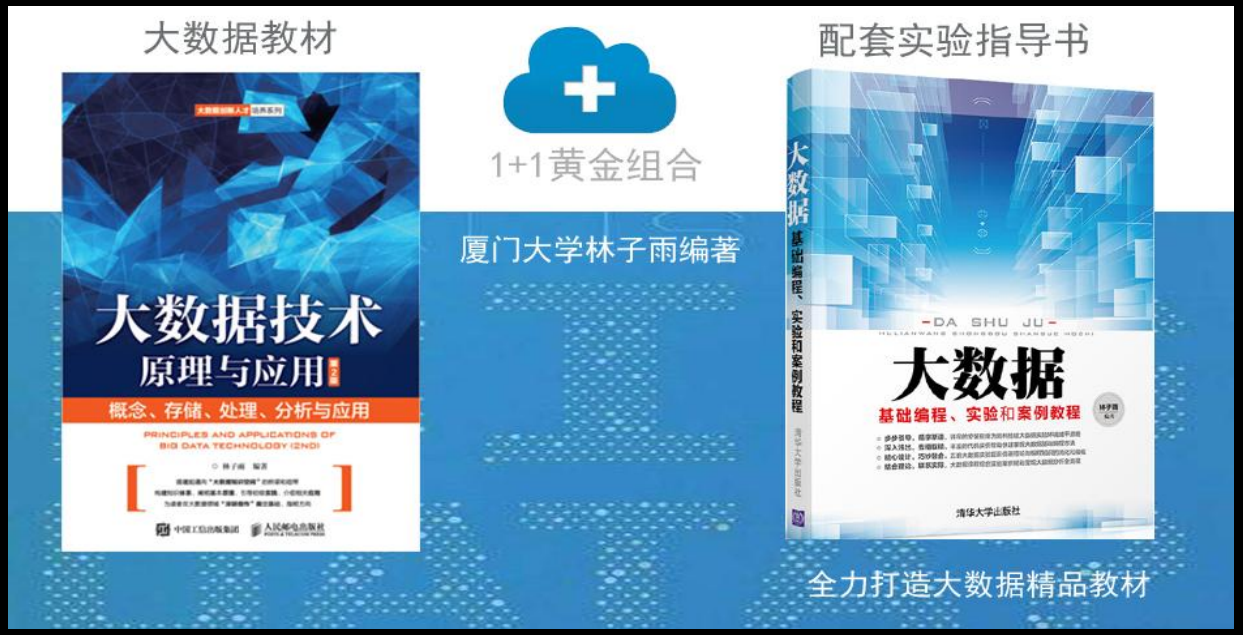
•为读者在大数据领域“深耕细作”奠定基础、指明方向
•Hadoop、HDFS、HBase、NoSQL、云数据库、MapReduce、流计算、图计算、数据可视化、Spark

教材官网：http://dblab.xmu.edu.cn/post/bigdata/

先修课程《大数据技术原理与应用》

教材选择

02



实验指导书官网：http://dblab.xmu.edu.cn/post/bigdatapractice/

先修课程《大数据技术原理与应用》

03

实验内容

•全套机房上机实验指南，包含题目和答案
•用于入门级大数据课程的上机实验课
•每个实验都需要连续4节上机课来完成
•每个实验的设计，都充分考虑了学生的基础和能力，力求学生能够在连续4节课的上机时间内，顺利完成课程实验，提交实验报告

实验一：熟悉常用的Linux操作和Hadoop操作
实验二：熟悉常用的HDFS操作
实验三：熟悉常用的HBase操作
实验四：NoSQL和关系数据库的操作比较
实验五：MapReduce初级编程实践

免费在线访问地址：http://dblab.xmu.edu.cn/post/6131/

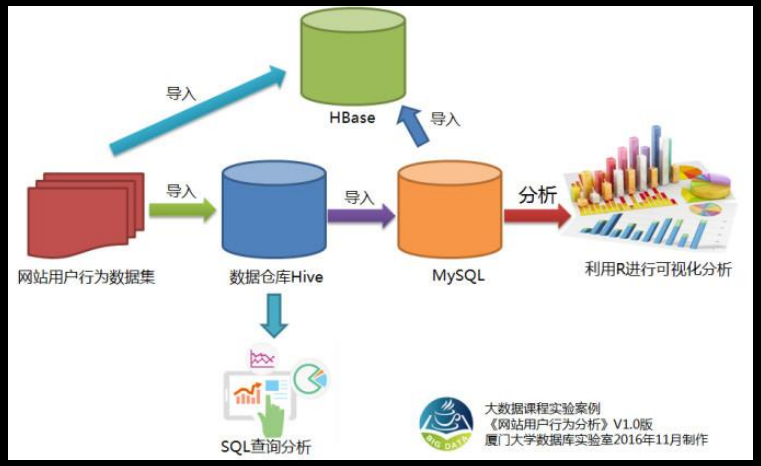
先修课程《大数据技术原理与应用》

03

实验内容

大数据课程实验案例《网站用户购物行为分析》

- 采用2000万条用户购物数据集
- 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、HBase、Hive、Sqoop、R、Eclipse等系统和软件的安装和使用方法
- 案例适合高校（高职）大数据教学，可以作为学生学习大数据课程后的综合实践案例



免费访问地址：http://dblab.xmu.edu.cn/post/7499/

先修课程《大数据技术原理与应用》

04

课程资源

林子雨主讲《大数据技术原理与应用》授课视频

授课视频观看地址：http://www.icourse163.org/course/XMU-1002335004

先修课程《大数据技术原理与应用》

05

课程资源

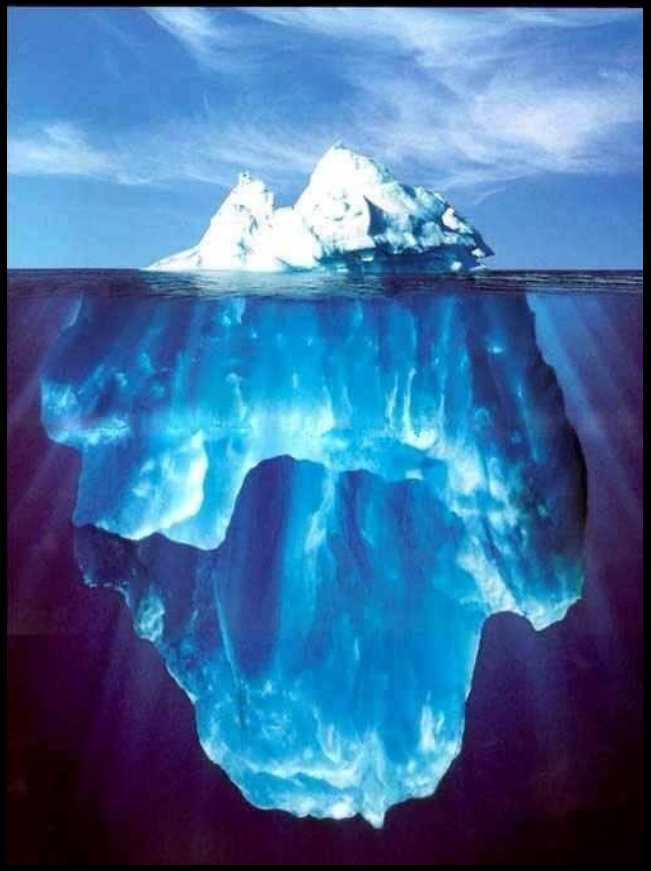
大数据课程教师交流群（QQ群号：461510122）
促进大数据课程教师之间的沟通和交流

截至目前，已经有来自全国500多所高校的700多名教师加入交流群

厦门大学、福建师范大学、厦门理工学院、同济大学、浙江财经大学、安徽大学、大连海洋大学、中北大学、河海大学、中山大学、浙江大学、中国农业大学、重庆邮电大学、华中师范大学、武汉理工大学、贵州师范大学、江西财经大学、山西大学、河北经贸大学、东北大学、山东农业大学、海南大学、中国地质大学、武汉大学、中国传媒大学、湖南大学、中国科技大学.....

内容提要

- 高校大数据课程公共服务平台
- 厦门大学建设的大数据课程体系
- 先修课程《大数据技术原理与应用》
- 《Spark编程基础》课程资源与教学经验



《Spark编程基础》课程资源与教学经验

01 课程定位

02 教材选择

03 课时安排

04 交叉知识

05 实验内容

06 考核方法

07 教学资源

08 师资培训

《Spark编程基础》课程资源与教学经验

课程定位

01

大数据技术进阶学习课程
授课对象：本科生、研究生（计算机相关专业）
知识储备：Java编程、数据库、操作系统、Hadoop
先修课程：入门级大数据课程，比如：大数据技术原理与应用

《Spark编程基础》课程资源与教学经验

教材选择

02

选择教材时，必须首先确定编程语言

Spark支持多种编程语言：Scala、Java、Python、R

首选语言是Scala，可以把Python作为课程拓展学习

《Spark编程基础》课程资源与教学经验

教材选择

02

在线免费《Spark入门教程》

Spark是当前最热门的大数据处理框架，林子雨编著《Spark入门教程》，让初学者零基础零障碍学习Spark。教程采用Scala语言编写Spark应用程序，因此，教程包括Scala入门和Spark入门两个部分的内容

扫一扫手机访问在线教程

免费在线教程：http://dblab.xmu.edu.cn/blog/spark/

《Spark编程基础》课程资源与教学经验

教材选择

02

第一部分：快学Scala
第一章 Scala简介
第二章 Scala安装
第三章 Scala基础
声明值和变量、基本数据类型和操作、Range、打印语句、读写文件
第四章 控制结构
if条件表达式、while循环、for循环、数据结构、数组、列表、元组、集、映射、迭代器
第六章 类
第七章 对象
第八章 继承
第九章 特质
第十章 模式匹配
第十一章 函数式编程
函数定义和高阶函数、针对集合的操作、遍历操作、map操作和flatMap操作、filter操作、reduce操作、fold操作、函数式编程实例WordCount

《Spark编程基础》课程资源与教学经验

第二部分：Spark速成（Spark2.1.0版本）

第1章 Spark的设计与运行原理
Spark简介、Spark运行架构、RDD的设计与运行原理、Spark的部署模式
第2章 Spark的安装与使用
Spark的安装与使用、第一个Spark应用程序：WordCount、使用开发工具IntelliJ IDEA和Eclipse编写Spark应用程序、Spark集群环境搭建、在集群上运行Spark应用程序
第3章 Spark编程基础
RDD编程、键值对RDD、数据读写（文件数据读写、读写HBase数据）
第4章 Spark SQL
Spark SQL简介、DataFrame与RDD的区别、DataFrame的创建、从RDD转换得到DataFrame、读取和保存数据（读写Parquet、通过JDBC连接数据库、连接Hive读写数据）
第5章 Spark Streaming
流计算简介、Spark Streaming简介、DStream操作（DStream操作概述、 输入源[文件流、套接字流、RDD队列流、Apache Kafka、Apache Flume]、转换操作、输出操作）
第6章 Spark MLlib
Spark MLlib简介、机器学习工作流（机器学习工作流、构建一个机器学习工作流、特征抽取、转化和选择[TF-IDF、Word2Vec、CountVectorizer、标签和索引的转化、卡方选择器]）、分类与回归（逻辑斯蒂回归分类器、决策树分类器）、聚类算法（KMeans聚类算法、高斯混合模型(GMM)聚类算法）、推荐算法（协同过滤算法）

《Spark编程基础》课程资源与教学经验

教材选择

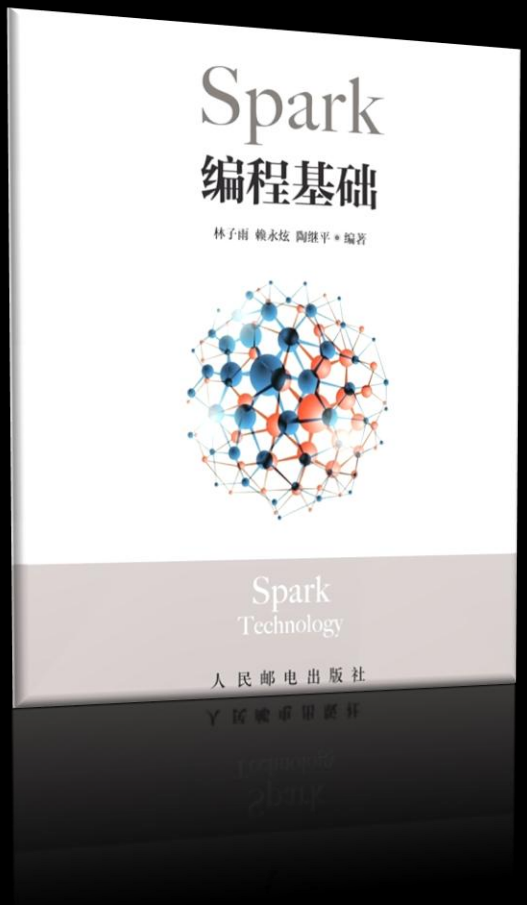
02

《Spark编程基础》

厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-47598-5
教材官网：http://dblab.xmu.edu.cn/post/spark/



《Spark编程基础》课程资源与教学经验

课时安排

03

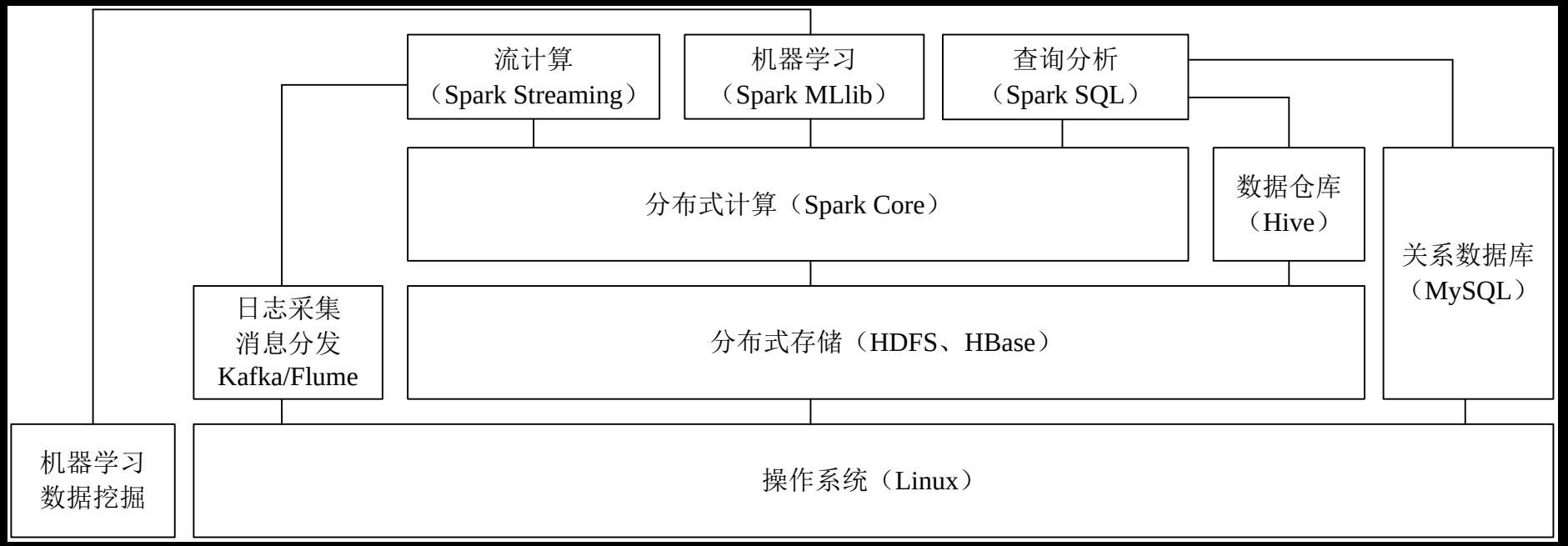
理论32学时，可另外增加实验上机学时

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 Scala语言基础	介绍Scala语言基础语法	6
第3章 Spark设计与运行原理	Spark概述、Spark生态系统、Spark运行架构、Spark的署和应用方式	3
第4章 Spark安装和使用方法	安装Spark、在Spark Shell中运行代码、编写Spark独立应用程序、第一个Spark应用程序：WordCount、使用开发工具编写Spark应用程序、Spark集群环境搭建、在集群上运行Spark应用程序	3
第5章 Spark编程基础	RDD编程、键值对RDD、数据读写（文件数据读写、读写HBase数据）	4
第6章 Spark SQL	Spark SQL简介、DataFrame、读写Parquet、通过JDBC连接数据库、连接Hive读写数据	2
第7章 Spark Streaming	流计算简介、Spark Streaming简介、DStream操作	4
第8章 Spark MLlib	Spark MLlib简介、机器学习工作流、特征抽取、转化和选择、分类与回归、聚类算法、推荐算法	4
综合案例	Spark综合案例	2
合计		32

《Spark编程基础》课程资源与教学经验

交叉知识

04



《Spark编程基础》课程资源与教学经验

04

实验内容

（1）实验环境搭建

- 单机构建实验环境（虚拟机、8GB内存）
  参考指南： http://dblab.xmu.edu.cn/post/5663/
- 实验室多机构建分布式环境
- 统一大数据实验机房
  - 一台服务器推送云桌面到多台终端机器
  - 多台物理机器构建分布式环境
- 在云端构建大数据实验环境
  参考指南：http://dblab.xmu.edu.cn/blog/1952-2/
- 采用Docker容器搭建大数据实验环境

《Spark编程基础》课程资源与教学经验

04

实验内容

（2）教材配套实验

实验1-Linux系统的安装和常用命令
实验2-Scala编程初级实践
实验3-Spark和Hadoop的安装
实验4-RDD编程初级实践
实验5-Spark SQL编程初级实践
实验6-Spark Streaming编程初级实践
实验7-Spark机器学习库MLlib编程实践

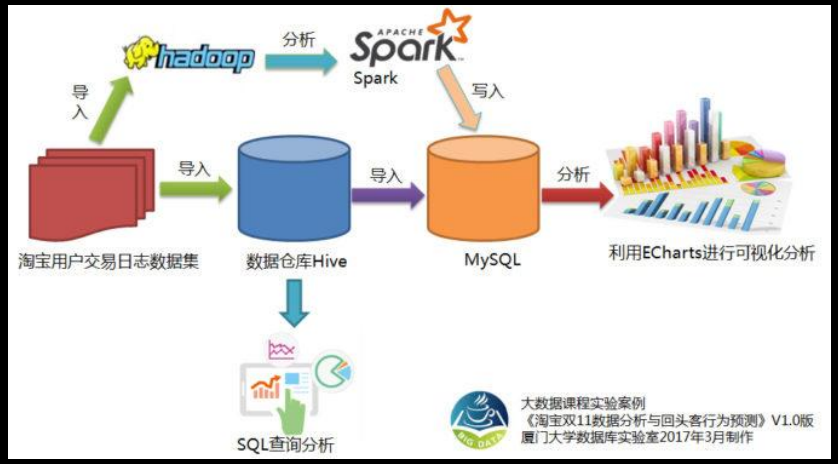
《Spark编程基础》课程资源与教学经验

04

实验内容

Spark课程综合实验案例：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



案例访问地址：http://dblab.xmu.edu.cn/post/8116/

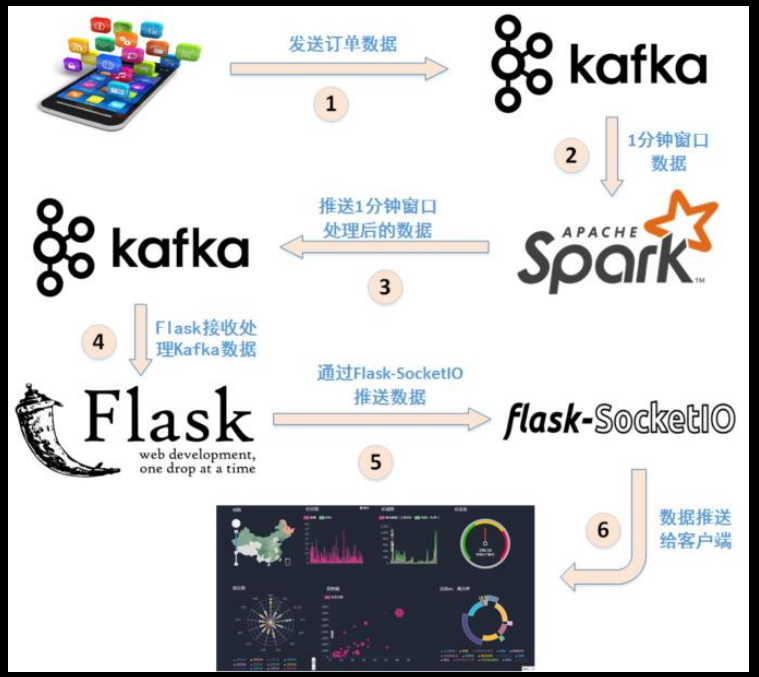
《Spark编程基础》课程资源与教学经验

04

实验内容

大数据课程实验案例：Spark+Kafka构建实时分析Dashboard案例

由厦门大学数据库实验室团队开发，旨在满足全国高校大数据教学对实验案例的迫切需求。本案例涉及数据预处理、消息队列发送和接收消息、数据实时处理、数据实时推送和实时展示等数据处理全流程所涉及的各种典型操作，涵盖Linux、Spark、Kafka、Flask、Flask-SocketIO、Highcharts.js、sockert.io.js、PyCharm等系统和软件的安装和使用方法。案例适合高校（高职）大数据教学，可以作为学生学习大数据课程后的综合实践案例。



免费在线实验案例主页：http://dblab.xmu.edu.cn/post/8274/

《Spark编程基础》课程资源与教学经验

05

考核方法

- 平时签到考勤10%
- 上机实验报告10%
- 期末大实验10%
- 期末笔试成绩70%

《Spark编程基础》课程资源与教学经验

06

教学视频

2018年3月在网易云课堂正式上线
http://study.163.com/course/introduction.htm?courseId=1005031005

《Spark编程基础》课程资源与教学经验

07

师资培训

第8期大数据师资培训班报名主页（Hadoop和Spark综合班，厦门，2018年7月24日-31日）报名主页（http://dblab.xmu.edu.cn/post/5899/）

总结：大数据课程建设模式

- 以大量教学实践推动课程和教材建设
- 以平台思维促进教学资源汇聚和共享
- 以迭代方法不断优化升级教学内容
- 自我造血为课程建设提供稳定资金保障

THANKS

敬请指正

@林子雨

附录：林子雨简介

林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn
中国高校首个“数字教师”的提出者和建设者
中国高校首个大数据课程公共服务平台建设者

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度和2017年度厦门大学教学类奖教金获得者，荣获2017年福建省精品在线开放课程和2018年厦门大学高等教育成果特等奖。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过500万字高价值的研究和教学资料，累计网络访问量超过500万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过100万次。