《大数据基础编程、实验和案例教程（第2版）》

教材官网：
http://dblab.xmu.edu.cn/post/bigdatapractice2/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第10章 Flink的安装和基础编程

（PPT版本号：2020年12月版本）

扫一扫访问教材官网

林子雨

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

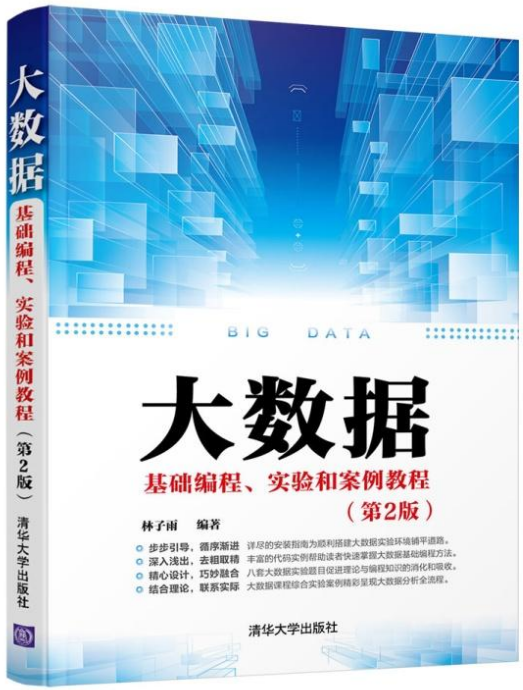
主页：**http://dblab.xmu.edu.cn/linziyu**

教材简介

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元，2020年10月第2版
教材官网： http://dblab.xmu.edu.cn/post/bigdatapractice2/

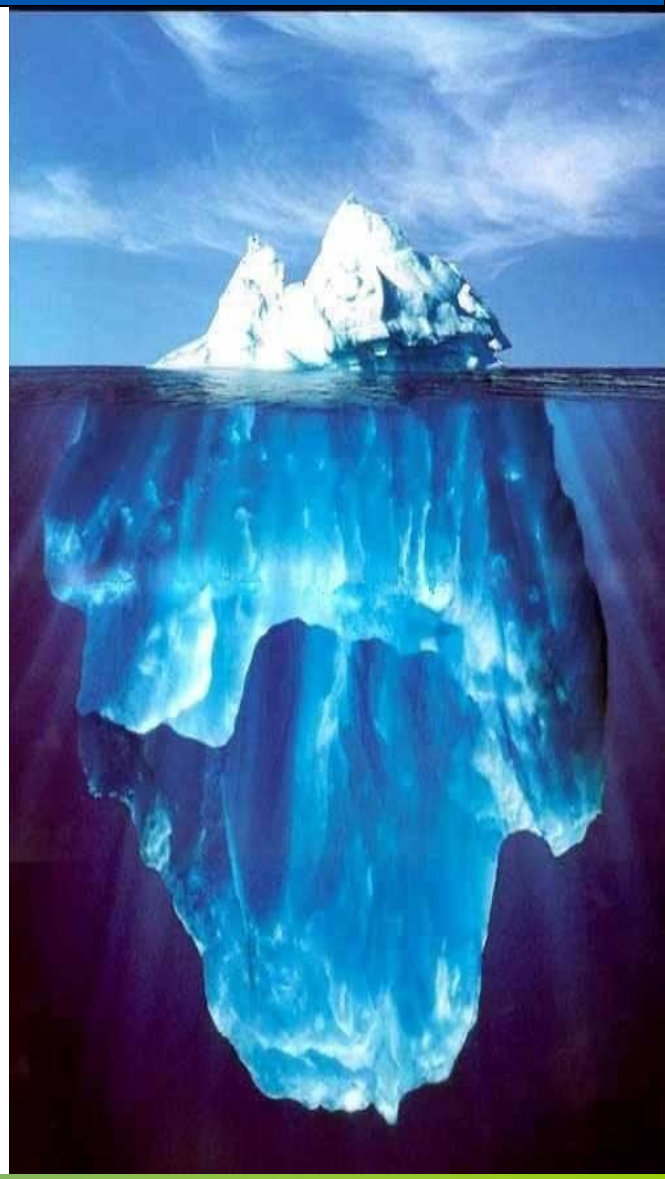
扫一扫访问
教材官网



- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

提纲

10.1 安装Flink
10.2 编程实现WordCount程序



高校大数据课程
公 共 服 务 平 台

百度搜索厦门大学数据库实验室网站访问平台

10.1 安装Flink

Flink的运行需要Java环境的支持，因此，在安装Flink之前，请先参照相关资料安装Java环境（比如Java8）。然后，到Flink官网（https://flink.apache.org/downloads.html）下载安装包flink-1.9.1-bin-scala_2.11.tgz。

使用如下命令对安装文件进行解压缩：

```
$ cd ~/Downloads
$ sudo tar -zxvf flink-1.9.1-bin-scala_2.11.tgz -C /usr/local
```

修改目录名称，并设置权限，命令如下：

```
$ cd /usr/local
$ sudo mv ./ flink-1.9.1 ./flink
$ sudo chown -R hadoop:hadoop ./flink
```

10.1 安装Flink

Flink对于本地模式是开箱即用的，如果要修改Java运行环境，可以修改“/usr/local/flink/conf/flink-conf.yaml”文件中的env.java.home参数，设置为本地Java的绝对路径。
使用如下命令添加环境变量：

```
$ vim ~/.bashrc
```

在.bashrc文件中添加如下内容：

```
export FLNK_HOME=/usr/local/flink
export PATH=$FLINK_HOME/bin:$PATH
```

保存并退出.bashrc文件，然后执行如下命令让配置文件生效：

```
$ source ~/.bashrc
```

10.1 安装Flink

使用如下命令启动Flink：

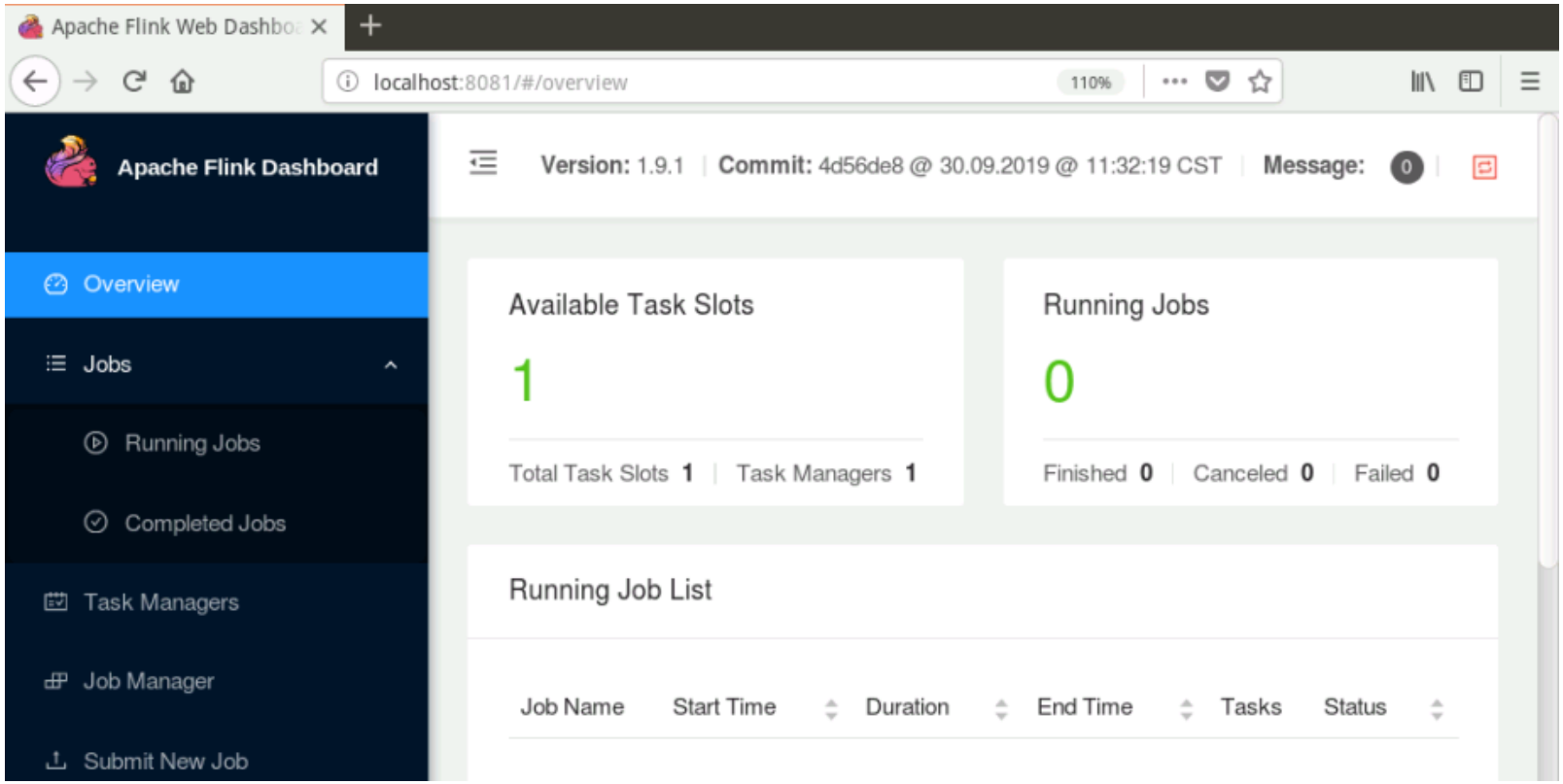
```
$ cd /usr/local/flink
$ ./bin/start-cluster.sh
```

使用jps命令查看进程：

```
$ jps
17942 TaskManagerRunner
18022 Jps
17503 StandaloneSessionClusterEntrypoint
```

10.1 安装Flink

Flink的JobManager同时会在8081端口上启动一个Web前端，可以在浏览器中输入“http://localhost:8081”来访问



10.1 安装Flink

Flink安装包中自带了测试样例，这里可以运行WordCount样例程序来测试Flink的运行效果，具体命令如下：

```
$ cd /usr/local/flink/bin
$ ./flink run /usr/local/flink/examples/batch/WordCount.jar
```

执行上述命令以后，如果执行成功，应该可以看到类似如下的屏幕信息：

```
Starting execution of program
Executing WordCount example with default input data set.
Use --input to specify file input.
Printing result to stdout. Use --output to specify output path.
(a,5)
(action,1)
(after,1)
(against,1)
(all,2)
……
```

10.2 编程实现WordCount程序

10.2.1安装Maven
10.2.2 编写代码
10.2.3使用Maven打包Java程序
10.2.4 通过flink run命令运行程序

10.2.1安装Maven

Ubuntu中没有自带安装Maven，需要手动安装Maven。可以访问Maven官网下载安装文件，下载地址如下：

https://downloads.apache.org/maven/maven-3/3.6.3/binaries/apache-maven-3.6.3-bin.zip

可以选择安装在“/usr/local/maven”目录中，命令如下：

```
$ sudo unzip ~/Downloads/apache-maven-3.6.3-bin.zip -d /usr/local
$ cd /usr/local
$ sudo mv apache-maven-3.6.3/ ./maven
$ sudo chown -R hadoop ./maven
```

10.2.2 编写代码

在Linux终端中执行如下命令，在用户主文件夹下创建一个文件夹flinkapp作为应用程序根目录：

```
$ cd ~ #进入用户主文件夹
$ mkdir -p ./flinkapp/src/main/java
```

然后，使用vim编辑器在“./flinkapp/src/main/java”目录下建立三个代码文件，即WordCountData.java、WordCountTokenizer.java和WordCount.java。

10.2.2 编写代码

WordCountData.java用于提供原始数据，其内容如下：

```
package cn.edu.xmu;
import org.apache.flink.api.java.DataSet;
import org.apache.flink.api.java.ExecutionEnvironment;

public class WordCountData {
    public static final String[] WORDS=new String[]{"To be, or not to be,“,…….};
    public WordCountData() {
    }
    public static DataSet<String> getDefaultTextLineDataset(ExecutionEnvironment env){
        return env.fromElements(WORDS);
    }
}
```

10.2.2 编写代码

WordCountTokenizer.java用于切分句子，其内容如下：

```
package cn.edu.xmu;
import org.apache.flink.api.common.functions.FlatMapFunction;
import org.apache.flink.api.java.tuple.Tuple2;
import org.apache.flink.util.Collector;
public class WordCountTokenizer implements FlatMapFunction<String, Tuple2<String,Integer>>{
   public WordCountTokenizer(){}
   public void flatMap(String value, Collector<Tuple2<String, Integer>> out) throws Exception {
      String[] tokens = value.toLowerCase().split("\\W+");
      int len = tokens.length;
      for(int i = 0; i<len;i++){
         String tmp = tokens[i];
         if(tmp.length()>0){
            out.collect(new Tuple2<String, Integer>(tmp,Integer.valueOf(1)));
         }
      }
   }
}
```

10.2.2 编写代码

WordCount.java提供主函数，其内容如下：

```
package cn.edu.xmu;
import org.apache.flink.api.java.DataSet;
import org.apache.flink.api.java.ExecutionEnvironment;
import org.apache.flink.api.java.operators.AggregateOperator;
import org.apache.flink.api.java.utils.ParameterTool;
public class WordCount {
   public WordCount(){}
   public static void main(String[] args) throws Exception {
      ParameterTool params = ParameterTool.fromArgs(args);
      ExecutionEnvironment env = ExecutionEnvironment.getExecutionEnvironment();
      env.getConfig().setGlobalJobParameters(params);
      Object text;
      //如果没有指定输入路径，则默认使用WordCountData中提供的数据
      if(params.has("input")){
         text = env.readTextFile(params.get("input"));
      }else{
         System.out.println("Executing WordCount example with default input data set.");
         System.out.println("Use -- input to specify file input.");
         text = WordCountData.getDefaultTextLineDataset(env);
      }
      AggregateOperator counts = ((DataSet)text).flatMap(new WordCountTokenizer()).groupBy(new int[]{0}).sum(1);
      //如果没有指定输出，则默认打印到控制台
      if(params.has("output")){
         counts.writeAsCsv(params.get("output"),"\n", " ");
         env.execute();
      }else{
         System.out.println("Printing result to stdout. Use --output to specify output path.");
         counts.print();
      }
   }
}
```

10.2.2 编写代码

在pom.xml文件中添加如下内容，用来声明该独立应用程序的信息以及与Flink的依赖关系：

```
<project>
  <groupId>cn.edu.xmu</groupId>
  <artifactId>simple-project</artifactId>
  <modelVersion>4.0.0</modelVersion>
  <name>Simple Project</name>
  <packaging>jar</packaging>
  <version>1.0</version>
  <repositories>
    <repository>
      <id>jboss</id>
      <name>JBoss Repository</name>
      <url>http://repository.jboss.com/maven2/</url>
    </repository>
  </repositories>
  <dependencies>
    <dependency>
      <groupId>org.apache.flink</groupId>
      <artifactId>flink-java</artifactId>
      <version>1.9.1</version>
    </dependency>
    <dependency>
      <groupId>org.apache.flink</groupId>
      <artifactId>flink-streaming-java_2.11</artifactId>
            <version>1.9.1</version>
    </dependency>
    <dependency>
      <groupId>org.apache.flink</groupId>
      <artifactId>flink-clients_2.11</artifactId>
      <version>1.9.1</version>
    </dependency>
  </dependencies>
</project>
```

10.2.3使用Maven打包Java程序

为了保证Maven能够正常运行，先执行如下命令检查整个应用程序的文件结构：

```
$ cd ~/flinkapp
$ find .
```

文件结构应该是类似如下的内容：

```
.
./src
./src/main
./src/main/java
./src/main/java/WordCountData.java
./src/main/java/WordCount.java
./src/main/java/WordCountTokenizer.java
./pom.xml
```

10.2.3使用Maven打包Java程序

接下来，我们可以通过如下代码将整个应用程序打包成JAR包

```
$ cd ~/flinkapp    #一定把这个目录设置为当前目录
$ /usr/local/maven/bin/mvn package
```

如果屏幕返回的信息中包含“BUILD SUCCESS”，则说明生成JAR包成功。

10.2.4 通过flink run命令运行程序

最后，可以将生成的JAR包通过flink run命令提交到Flink中运行（请确认已经启动Flink），命令如下：

```
$ /usr/local/flink/bin/flink run --class cn.edu.xmu.WordCount
~/flinkapp/target/simple-project-1.0.jar
```

执行成功后，可以在屏幕上看到词频统计结果。

10.3本章小结

Apache Flink是一个分布式处理引擎，用于对无界和有界数据流进行有状态计算。Flink以数据并行和流水线方式执行任意流数据程序，Flink的流水线运行时系统可以执行批处理和流处理程序。此外，Flink的运行时本身也支持迭代算法的执行。本章简要介绍了Flink的安装以及基本编程方法。

附录A：主讲教师林子雨简介

主讲教师：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://dblab.xmu.edu.cn/post/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

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），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，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、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。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1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。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200万次，累计访问量超过1000万次。

附录B：大数据学习路线图

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

附录C：林子雨大数据系列教材

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元

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

附录E：《大数据导论》教材

- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

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

开设大数据专业导论课的优质教材

扫一扫访问教材官网

附录F：《大数据技术原理与应用》教材

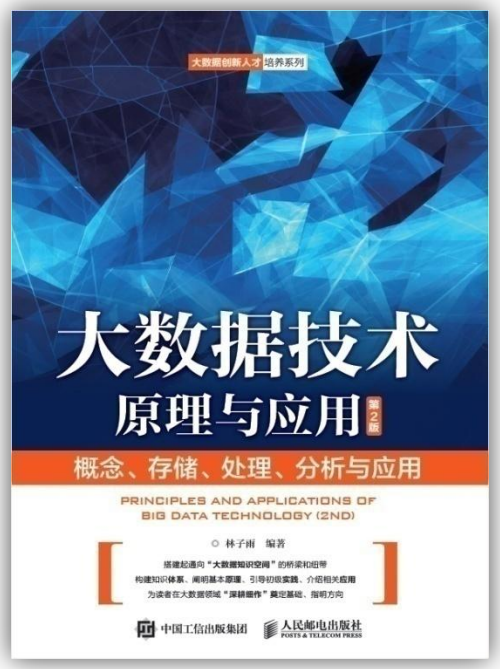
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、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、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录G：《大数据基础编程、实验和案例教程（第2版）》

本书是与《大数据技术原理与应用（第3版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

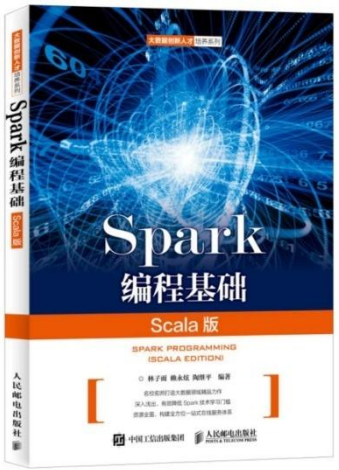
厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，八套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

林子雨编著《大数据基础编程、实验和案例教程（第2版）》
清华大学出版社 ISBN:978-7-302-55977-1 定价：69元 2020年10月第2版

附录H：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

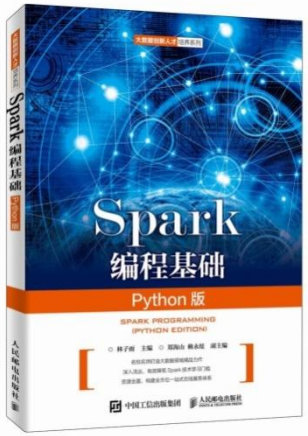
厦门大学 林子雨，赖永炫，陶继平 编著

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:978-7-115-48816-9
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以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。全书共8章，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。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录I：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 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:978-7-115-52439-3
教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python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。全书共8章，内容包括大数据技术概述、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、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、上机实验指南等。

附录J：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

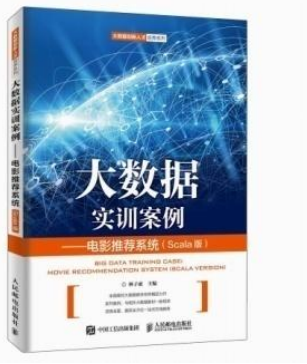
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

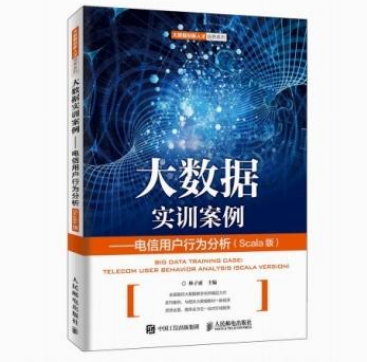
附录K：高校大数据实训课程系列案例教材

为了更好满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《实时日志流处理分析》
《微博用户情感分析》
《互联网广告预测分析》
《网站日志处理分析》



系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！
http://dblab.xmu.edu.cn/post/shixunkecheng/



扫一扫访问大数据实训课程系列案例教材主页


Thank You!