厦门大学林子雨，郑海山，赖永炫 编著

《Spark编程基础（Python版）》

教材配套

机房上机实验指南

实验4 Spark SQL编程初级实践

（版本号：2020年4月版本）

（题目）



主讲教师：林子雨

厦门大学数据库实验室

二零二零年四月

目录

[一、实验目的 1](#_Toc519775182)

[二、实验平台 1](#_Toc519775183)

[三、实验内容和要求 1](#_Toc519775184)

[1．Spark SQL基本操作 1](#_Toc519775185)

[2．编程实现将RDD转换为DataFrame 2](#_Toc519775186)

[3. 编程实现利用DataFrame读写MySQL的数据 2](#_Toc519775187)

[四、实验报告 2](#_Toc519775188)

[**附录1:任课教师介绍** 3](#_Toc519775189)

[**附录2：课程教材介绍** 4](#_Toc519775190)

[**附录3：高校大数据课程公共服务平台介绍** 5](#_Toc519775191)

厦门大学林子雨,赖永炫,陶继平 编著

《Spark编程基础（Scala版）》

第5章 Spark SQL

教材配套机房上机实验指南

实验4 Spark SQL编程初级实践

（题目）

**主讲教师：林子雨**

E-mail: [ziyulin@xmu.edu.cn](mailto:ziyulin@xmu.edu.cn) 个人主页：<http://dblab.xmu.edu.cn/post/linziyu>

# 一、实验目的

（1）通过实验掌握Spark SQL的基本编程方法；

（2）熟悉RDD到DataFrame的转化方法；

（3）熟悉利用Spark SQL管理来自不同数据源的数据。

# 二、实验平台

操作系统： Ubuntu16.04

Spark版本：2.4.0

数据库：MySQL

Python版本：3.4.3

# 三、实验内容和要求

## 1．Spark SQL基本操作

将下列JSON格式数据复制到Linux系统中，并保存命名为employee.json。

|  |
| --- |
| { "id":1 , "name":" Ella" , "age":36 }  { "id":2, "name":"Bob","age":29 }  { "id":3 , "name":"Jack","age":29 }  { "id":4 , "name":"Jim","age":28 }  { "id":4 , "name":"Jim","age":28 }  { "id":5 , "name":"Damon" }  { "id":5 , "name":"Damon" } |

为employee.json创建DataFrame，并写出Python语句完成下列操作：

1. 查询所有数据；
2. 查询所有数据，并去除重复的数据；
3. 查询所有数据，打印时去除id字段；
4. 筛选出age>30的记录；
5. 将数据按age分组；
6. 将数据按name升序排列；
7. 取出前3行数据；
8. 查询所有记录的name列，并为其取别名为username；
9. 查询年龄age的平均值；
10. 查询年龄age的最小值。

## 2．编程实现将RDD转换为DataFrame

源文件内容如下（包含id,name,age）：

|  |
| --- |
| 1,Ella,36  2,Bob,29  3,Jack,29 |

请将数据复制保存到Linux系统中，命名为employee.txt，实现从RDD转换得到DataFrame，并按“id:1,name:Ella,age:36”的格式打印出DataFrame的所有数据。请写出程序代码。

## 3. 编程实现利用DataFrame读写MySQL的数据

（1）在MySQL数据库中新建数据库sparktest，再创建表employee，包含如表5-2所示的两行数据。

**表5-2 employee表原有数据**

|  |  |  |  |
| --- | --- | --- | --- |
| id | name | gender | Age |
| 1 | Alice | F | 22 |
| 2 | John | M | 25 |

（2）配置Spark通过JDBC连接数据库MySQL，编程实现利用DataFrame插入如表5-3所示的两行数据到MySQL中，最后打印出age的最大值和age的总和。

**表5-3 employee表新增数据**

|  |  |  |  |
| --- | --- | --- | --- |
| id | name | gender | age |
| 3 | Mary | F | 26 |
| 4 | Tom | M | 23 |

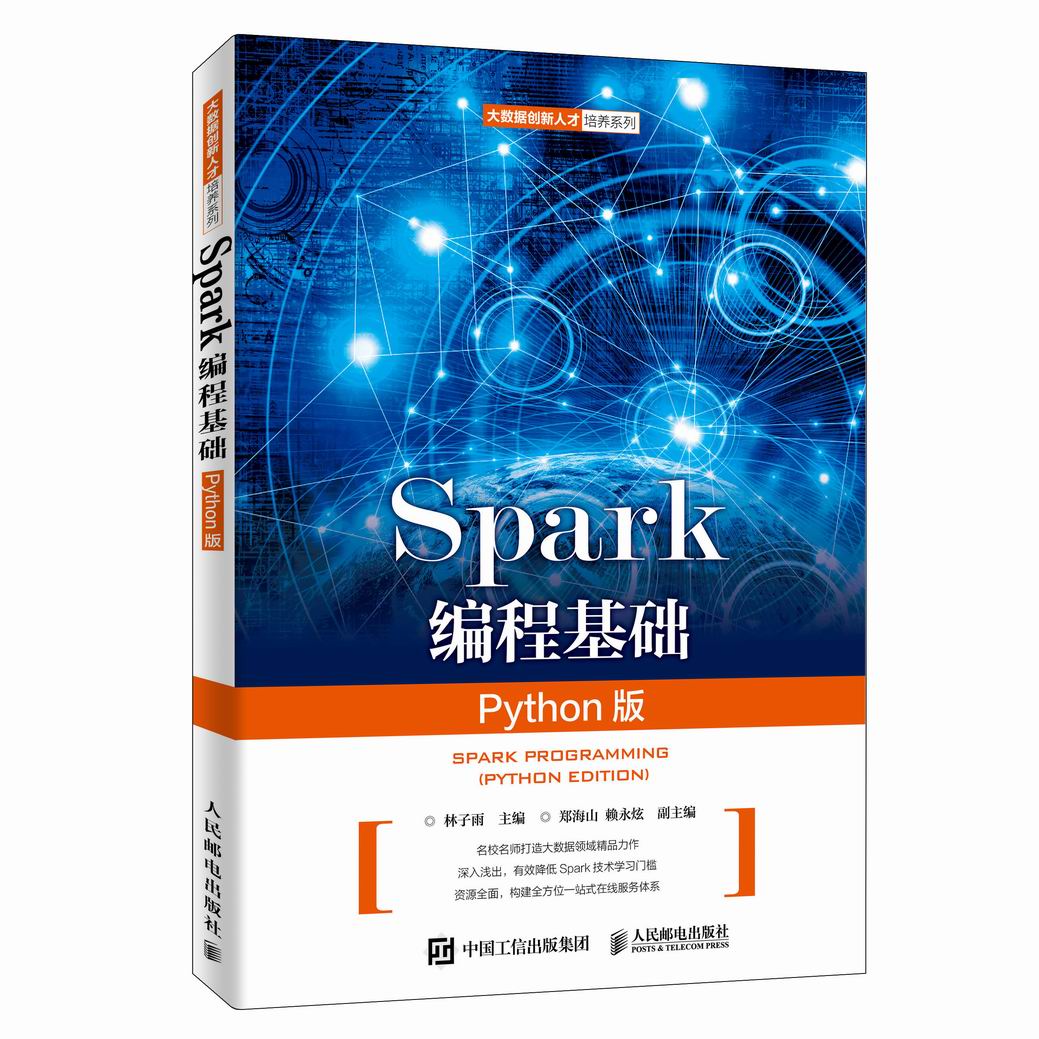
# 四、实验报告

|  |  |  |
| --- | --- | --- |
| 《Spark编程基础（Python版）》实验报告 | | |
| 题目： | 姓名： | 日期： |
| 实验环境： | | |
| 实验内容与完成情况： | | |
| 出现的问题： | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）： | | |

**附录1:任课教师介绍**

|  |  |
| --- | --- |
| [http://www.cs.xmu.edu.cn/database/linziyu/images/linziyu2016.jpg](http://dblab.xmu.edu.cn/post/linziyu/) | 林子雨（1978－），男，博士，国内高校知名大数据教师，厦门大学计算机科学系副教授，厦门大学云计算与大数据研究中心创始成员，厦门大学数据库实验室负责人，中国计算机学会数据库专委会委员，中国计算机学会信息系统专委会委员，中国高校首个“数字教师”提出者和建设者。2013年开始在厦门大学开设大数据课程，建设了国内高校首个大数据课程公共服务平台，平台累计网络访问量超过1000万次，成为全国高校大数据教学知名品牌，并荣获“2018年福建省教学成果二等奖”，主持的课程《大数据技术原理与应用》获评“2018年国家精品在线开放课程”。  E-mail: ziyulin@xmu.edu.cn  个人主页：http://dblab.xmu.edu.cn/post/linziyu  数据库实验室网站：http://dblab.xmu.edu.cn |

**附录2：课程教材介绍**



林子雨、郑海山、赖永炫编著《Spark编程基础（Python版）》

人民邮电出版社 ISBN:978-7-115-52439-3 定价：49.80元

厦门大学林子雨、郑海山、赖永炫老师编著《Spark编程基础（Python版）》，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、授课视频、上机实验指南等。

本书可以作为高等院校计算机、软件工程、数据科学与大数据技术等专业的进阶级大数据课程教材，用于指导Spark编程实践，也可供相关技术人员参考。

欢迎访问《Spark编程基础(Python版)》教材官方网站：http://dblab.xmu.edu.cn/post/spark-python/



扫一扫访问教材官网

**附录3：高校大数据课程公共服务平台介绍**



高校大数据课程公共服务平台，由中国高校首个“数字教师”的提出者和建设者——林子雨老师发起，由厦门大学数据库实验室全力打造，由厦门大学云计算与大数据研究中心、海峡云计算与大数据应用研究中心携手共建。这是国内第一个服务于高校大数据课程建设的公共服务平台，旨在促进国内高校大数据课程体系建设，提高大数据课程教学水平，降低大数据课程学习门槛，提升学生课程学习效果。平台服务对象涵盖高校、教师和学生。平台为高校开设大数据课程提供全流程辅助，为教师开展教学工作提供一站式服务，为学生学习大数据课程提供全方位辅导。平台重点打造“11个1工程”，即1本教材（含官网）、1个教师服务站、1个学生服务站、1个公益项目、1堂巡讲公开课、1个示范班级、1门在线课程、1个交流群（QQ群、微信群）、1个保障团队、1个培训基地和1个实验平台。目前平台每年访问量已经超过200万次，累计访问量超过1000万次，成为国内高校大数据教学知名品牌。

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



扫一扫访问平台主页