厦门大学林子雨，郑海山，赖永炫 编著

《Spark编程基础（Python）》

教材配套

机房上机实验指南

实验3 Spark和Hadoop的安装

（版本号：2020年4月版本）

（答案）



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](#_Toc495048680)

[二、实验平台 1](#_Toc495048681)

[三、实验内容和要求 1](#_Toc495048682)

[1．安装Hadoop和Spark 1](#_Toc495048683)

[2．HDFS常用操作 2](#_Toc495048684)

[3. Spark读取文件系统的数据 3](#_Toc495048685)

[四、实验报告 5](#_Toc495048686)

[**附录1:任课教师介绍** 5](#_Toc495048687)

[**附录2：课程教材介绍** 5](#_Toc495048688)

[**附录3：高校大数据课程公共服务平台介绍** 6](#_Toc495048689)

厦门大学林子雨,郑海山,赖永炫 编著《Spark编程基础（Python版）》

第4章 Spark环境搭建和使用方法

教材配套机房上机实验指南

实验3 Spark和Hadoop的安装

（答案）

**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）掌握在Linux虚拟机中安装Hadoop和Spark的方法；

（2）熟悉HDFS的基本使用方法；

（3）掌握使用Spark访问本地文件和HDFS文件的方法。

# 二、实验平台

操作系统：Ubuntu16.04；

Spark版本：2.1.0；

Hadoop版本：2.7.1。

Python版本：3.4.3。

# 三、实验内容和要求

## 1．安装Hadoop和Spark

进入Linux系统，参照本教程官网“实验指南”栏目的“Hadoop的安装和使用”，完成Hadoop伪分布式模式的安装。完成Hadoop的安装以后，再安装Spark（Local模式）。

【参考答案】

请参照下面网络教程完成Hadoop的安装：

http://dblab.xmu.edu.cn/blog/install-hadoop/

请参照下面网络教程完成Spark（Local模式）的安装：

http://dblab.xmu.edu.cn/blog/1307-2/

## 2．HDFS常用操作

使用hadoop用户名登录进入Linux系统，启动Hadoop，参照相关Hadoop书籍或网络资料，或者也可以参考本教程官网的“实验指南”栏目的“HDFS操作常用Shell命令”，使用Hadoop提供的Shell命令完成如下操作：

1. 启动Hadoop，在HDFS中创建用户目录“/user/hadoop”；

【参考答案】

使用hadoop用户名登录进入Linux系统，假设Hadoop系统被安装在“/usr/local/hadoop”目录。使用如下命令启动Hadoop：

|  |
| --- |
| $cd /usr/local/hadoop  $./sbin/start-dfs.sh #启动HDFS  $./bin/hdfs dfs -mkdir -p /user/hadoop #在HDFS中创建用户目录/user/hadoop |

1. 在Linux系统的本地文件系统的“/home/hadoop”目录下新建一个文本文件test.txt，并在该文件中随便输入一些内容，然后上传到HDFS的“/user/hadoop”目录下；

【参考答案】

|  |
| --- |
| $cd /home/hadoop  $vim test.txt  #在test.txt中随便输入一些内容，并保存退出vim编辑器  $cd /usr/local/hadoop  $./bin/hdfs dfs -put /home/hadoop/test.txt /user/hadoop |

1. 把HDFS中“/user/hadoop”目录下的test.txt文件，下载到Linux系统的本地文件系统中的“/home/hadoop/下载”目录下；

【参考答案】

|  |
| --- |
| $ cd /usr/local/hadoop  $./bin/hdfs dfs -get /user/hadoop/test.txt /home/hadoop/下载 |

1. 将HDFS中“/user/hadoop”目录下的test.txt文件的内容输出到终端中进行显示；

【参考答案】

|  |
| --- |
| $ cd /usr/local/hadoop  $./bin/hdfs dfs -cat /user/hadoop/test.txt |

1. 在HDFS中的“/user/hadoop”目录下，创建子目录input，把HDFS中“/user/hadoop”目录下的test.txt文件，复制到“/user/hadoop/input”目录下；

【参考答案】

|  |
| --- |
| $ cd /usr/local/hadoop  $./bin/hdfs dfs -mkdir /user/hadoop/input  $./bin/hdfs dfs -cp /user/hadoop/test.txt /user/hadoop/input |

1. 删除HDFS中“/user/hadoop”目录下的test.txt文件，删除HDFS中“/user/hadoop”目录下的input子目录及其子目录下的所有内容。

【参考答案】

|  |
| --- |
| $ cd /usr/local/hadoop  $./bin/hdfs dfs -rm /user/hadoop/test.txt  $./bin/hdfs dfs -rm -r /user/hadoop/input |

## 3. Spark读取文件系统的数据

（1）在pyspark中读取Linux系统本地文件“/home/hadoop/test.txt”，然后统计出文件的行数；

【参考答案】假设Spark安装在“/usr/local/spark”目录。

|  |
| --- |
| $ cd /usr/local/spark  $./bin/pyspark  >>> textFile=sc.textFile("file:///home/hadoop/test.txt")  >>> linecount=textFile.count()  >>> print(linecount) |

（2）在pyspark中读取HDFS系统文件“/user/hadoop/test.txt”（如果该文件不存在，请先创建），然后，统计出文件的行数；

【参考答案】

|  |
| --- |
| >>> textFile=sc.textFile("hdfs://localhost:9000/user/hadoop/test.txt")  >>> linecount=textFile.count()  >>> print(linecount) |

（3）编写独立应用程序，读取HDFS系统文件“/user/hadoop/test.txt”（如果该文件不存在，请先创建），然后，统计出文件的行数；通过 spark-submit 提交到 Spark 中运行程序。

【参考答案】

使用hadoop用户名登录Linux系统，打开一个终端，在Linux终端中，执行如下命令创建一个Linux本地文件：

|  |
| --- |
| $ cd ~  $ vim test.txt |

上面命令使用vim编辑器创建一个test.txt文件，请在该文件中随便输入一些英文语句，然后保存退出。接下来，执行如下命令，把该文件上传到Hadoop的HDFS中：

|  |
| --- |
| $ cd /usr/local/hadoop #进入Hadoop的安装目录  $ ./sbin/start-dfs.sh #启动Hadoop  $ ./bin/hdfs dfs -put ~/test.txt /user/hadoop #把本地文件test.txt上传到HDFS中  $ ./bin/hdfs dfs -ls /user/hadoop/ #验证文件是否上传到HDFS中  $ ./bin/hdfs dfs -cat /user/hadoop/test.txt #查看HDFS中的test.txt文件内容 |

在终端中执行如下命令创建一个SimpleApp.py应用程序：

|  |
| --- |
| $ cd ~ # 进入用户主文件夹  $ mkdir ./sparkapp # 创建应用程序根目录  $ cd sparkapp  $ vim SimpleApp.py |

在SimpleApp.py中输入如下代码：

|  |
| --- |
| #!/usr/bin/env python3  from pyspark import SparkConf, SparkContext  conf = SparkConf().setMaster("local").setAppName("My App")  sc = SparkContext(conf = conf)  logFile = "hdfs://localhost:9000/user/hadoop/test.txt "  logData = sc.textFile(logFile)  linecount = logData.count()  print(linecount) |

然后，在终端中执行如下命令运行程序：

|  |
| --- |
| $ cd ~/sparkapp  $ /usr/local/spark/bin/spark-submit SimpleApp.py |

这时就可以看到实验结果。

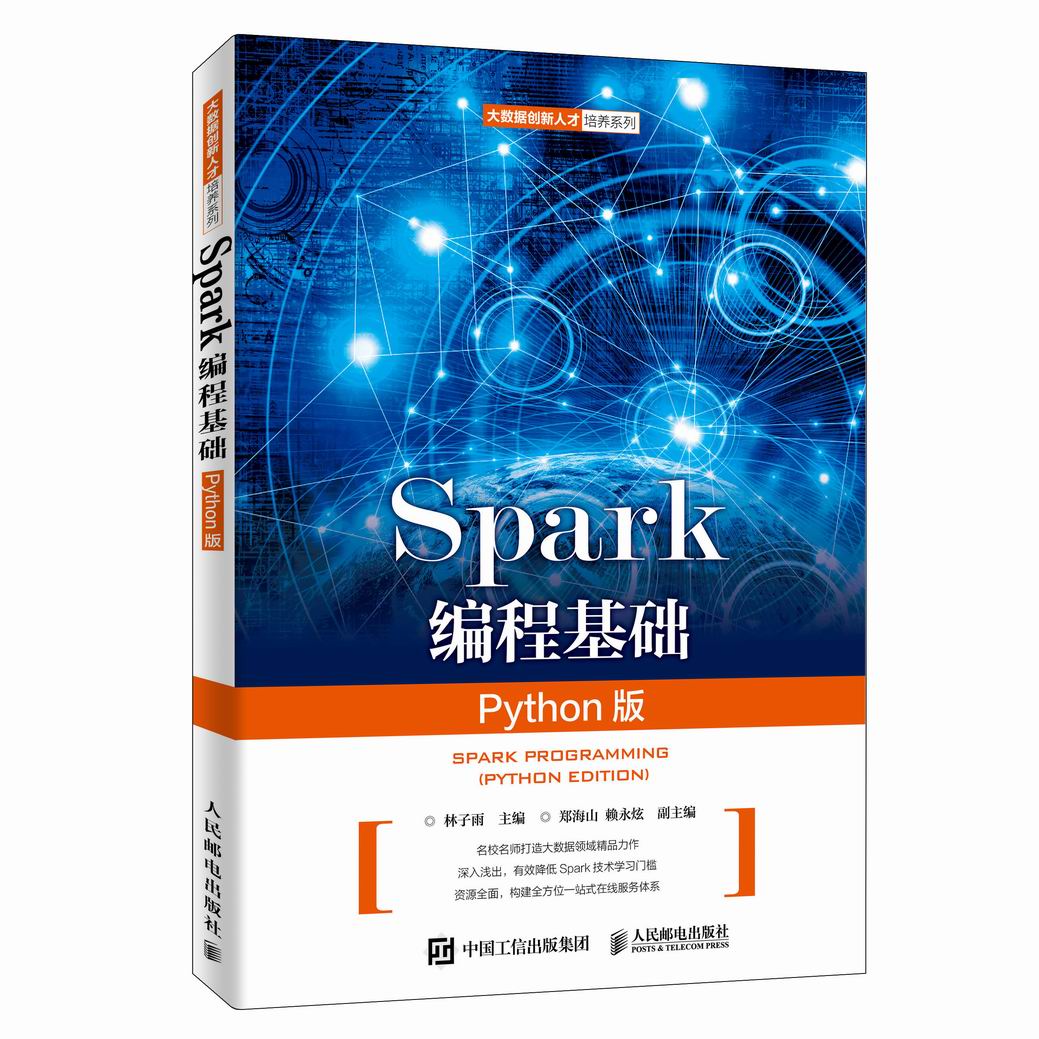
# 四、实验报告

|  |  |  |
| --- | --- | --- |
| 《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