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5 Spark Streaming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](#_Toc494984240)

[二、实验平台 1](#_Toc494984241)

[三、实验内容和要求 1](#_Toc494984242)

[1.安装Flume 1](#_Toc494984243)

[2. 使用Avro数据源测试Flume 2](#_Toc494984244)

[3. 使用netcat数据源测试Flume 2](#_Toc494984245)

[4.使用Flume作为Spark Streaming数据源 2](#_Toc494984246)

[四、实验报告 2](#_Toc494984247)

[**附录1:任课教师介绍** 3](#_Toc494984248)

[**附录2：课程教材介绍** 3](#_Toc494984249)

[**附录3：高校大数据课程公共服务平台介绍** 4](#_Toc494984250)

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

第6章 Spark Streaming

教材配套机房上机实验指南

实验5 Spark Streaming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](http://www.cs.xmu.edu.cn/linziyu)

# 一、实验目的

（1）通过实验学习日志采集工具Flume的安装和使用方法；

（2）掌握采用Flume作为Spark Streaming数据源的编程方法。

# 二、实验平台

操作系统： Ubuntu16.04

Spark版本：2.4.0

Flume版本：1.7.0

Python版本：3.4.3

# 三、实验内容和要求

## 1.安装Flume

Flume是Cloudera提供的一个分布式、可靠、可用的系统，它能够将不同数据源的海量日志数据进行高效收集、聚合、移动，最后存储到一个中心化数据存储系统中。Flume 的核心是把数据从数据源收集过来，再送到目的地。请到Flume官网下载Flume1.7.0安装文件，下载地址如下：

http://www.apache.org/dyn/closer.lua/flume/1.7.0/apache-flume-1.7.0-bin.tar.gz

或者也可以直接到教材官网的“下载专区”中的“软件”目录中下载apache-flume-1.7.0-bin.tar.gz。

下载后，把Flume1.7.0安装到Linux系统的“/usr/local/flume”目录下，具体安装和使用方法可以参考教材官网的“实验指南”栏目中的“日志采集工具Flume的安装与使用方法”。

## 2. 使用Avro数据源测试Flume

Avro可以发送一个给定的文件给Flume，Avro 源使用AVRO RPC机制。请对Flume的相关配置文件进行设置，从而可以实现如下功能：在一个终端中新建一个文件helloworld.txt（里面包含一行文本“Hello World”），在另外一个终端中启动Flume以后，可以把helloworld.txt中的文本内容显示出来。

## 3. 使用netcat数据源测试Flume

请对Flume的相关配置文件进行设置，从而可以实现如下功能：在一个Linux终端（这里称为“Flume终端”）中，启动Flume，在另一个终端（这里称为“Telnet终端”）中，输入命令“telnet localhost 44444”，然后，在Telnet终端中输入任何字符，让这些字符可以顺利地在Flume终端中显示出来。

## 4.使用Flume作为Spark Streaming数据源

Flume是非常流行的日志采集系统，可以作为Spark Streaming的高级数据源。请把Flume Source设置为netcat类型，从终端上不断给Flume Source发送各种消息，Flume把消息汇集到Sink，这里把Sink类型设置为avro，由Sink把消息推送给Spark Streaming，由自己编写的Spark Streaming应用程序对消息进行处理。

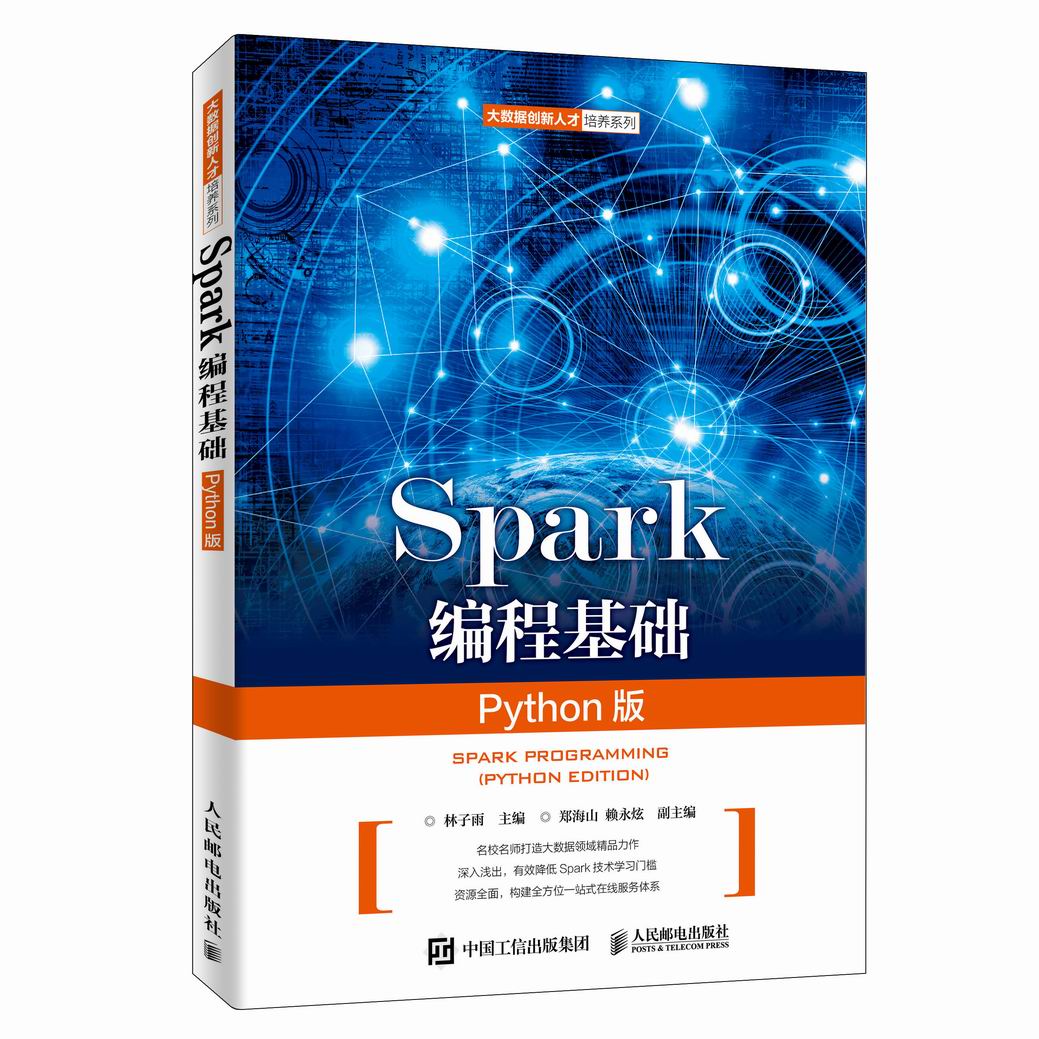
# 四、实验报告

|  |  |  |
| --- | --- | --- |
| 《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