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）

（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](#_Toc495046512)

[二、实验平台 1](#_Toc495046513)

[三、实验内容和要求 1](#_Toc495046514)

[1.安装Flume 1](#_Toc495046515)

[2. 使用Avro数据源测试Flume 2](#_Toc495046516)

[3. 使用netcat数据源测试Flume 2](#_Toc495046517)

[4.使用Flume作为Spark Streaming数据源 2](#_Toc495046518)

[四、实验报告 2](#_Toc495046519)

[**附录1:任课教师介绍** 3](#_Toc495046520)

[**附录2：课程教材介绍** 3](#_Toc495046521)

[**附录3：高校大数据课程公共服务平台介绍** 4](#_Toc495046522)

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

（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）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

或者也可以直接到本教程官网的“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程官网的“实验指南”栏目中的“日志采集工具Flume的安装与使用方法”。

【参考答案】

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

a.解压安装包：

|  |
| --- |
| 1.sudo tar -zxvf apache-flume-1.7.0-bin.tar.gz -C /usr/local  2.sudo mv ./apache-flume-1.7.0-bin ./flume  3.sudo chown -R hadoop:hadoop ./flume |

b.配置环境变量

|  |
| --- |
| 1.sudo vim ~/.bashrc  #在bashrc中加入以下内容  export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64;  export FLUME\_HOME=/usr/local/flume  export FLUME\_CONF\_DIR=$FLUME\_HOME/conf  export PATH=$PATH:$FLUME\_HOME/bin  #Java\_Home如果已经有了就不用添加了  #使环境变量生效  2.source ~/.bashrc  #修改/usr/local/flume/conf/flume-env.sh  3.sudo cp ./flume-env.sh.template ./flume-env.sh  4.sudo vim flume-env.sh  #在文件开头增加一行，设置JAVA\_HOME  export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64; |

c.查看flume版本信息

|  |
| --- |
| 1./usr/local/flume/bin/flume-ng version  ##显示如下信息  Flume 1.7.0  Source code repository: https://git-wip-us.apache.org/repos/asf/flume.git  Revision: 511d868555dd4d16e6ce4fedc72c2d1454546707  Compiled by bessbd on Wed Oct 12 20:51:10 CEST 2016  From source with checksum 0d21b3ffdc55a07e1d08875872c00523 |

## 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。

【参考答案】

a.创建agent配置文件

|  |  |
| --- | --- |
| 1.cd /usr/local/flume  2.sudo vim ./conf/avro.conf  #复制以下内容到avro.conf   |  | | --- | | a1.sources = r1  a1.sinks = k1  a1.channels = c1    # Describe/configure the source  a1.sources.r1.type = avro  a1.sources.r1.channels = c1  a1.sources.r1.bind = 0.0.0.0  a1.sources.r1.port = 4141  #注意这个端口名，在后面的教程中会用得到    # Describe the sink  a1.sinks.k1.type = logger    # Use a channel which buffers events in memory  a1.channels.c1.type = memory  a1.channels.c1.capacity = 1000  a1.channels.c1.transactionCapacity = 100    # Bind the source and sink to the channel  a1.sources.r1.channels = c1  a1.sinks.k1.channel = c1 |   上面Avro Source参数说明如下：  Avro Source的别名是avro,也可以使用完整类别名称org.apache.flume.source.AvroSource，因此，上面有一行设置是a1.sources.r1.type = avro，表示数据源的类型是avro。  bind绑定的ip地址或主机名，使用0.0.0.0表示绑定机器所有的接口。 a1.sources.r1.bind = 0.0.0.0，就表示绑定机器所有的接口。  port表示绑定的端口。a1.sources.r1.port = 4141，表示绑定的端口是4141。  a1.sinks.k1.type = logger，表示sinks的类型是logger。 |

b.启动flume agent a1

|  |
| --- |
| /usr/local/flume/bin/flume-ng agent -c . -f /usr/local/flume/conf/avro.conf -n a1 -Dflume.root.logger=INFO,console  #这个终端不要关闭 |

c.创建指定文件

|  |
| --- |
| 1.cd /usr/local/flume  2.echo “Hello World”>> ./helloworld.txt  #然后在另一个终端执行：  3./usr/local/flume/bin/flume-ng avro-client --conf conf -H localhost -p 4141 -F /usr/local/flume/helloworld.txt  #执行之后，我们就可以在前面不让关闭的那个终端看到Hello World了 |

## 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。

【参考答案】

a.创建netcat的agent配置

|  |  |
| --- | --- |
| 1.cd /usr/local/flume  2.sudo vim ./conf/netcat.conf  #复制以下内容   |  | | --- | | # Name the components on this agent  a1.sources = r1  a1.sinks = k1  a1.channels = c1    # Describe/configure the source  a1.sources.r1.type = netcat  a1.sources.r1.bind = localhost  a1.sources.r1.port = 44444  #同上，记住该端口名    # Describe the sink  a1.sinks.k1.type = logger    # Use a channel which buffers events in memory  a1.channels.c1.type = memory  a1.channels.c1.capacity = 1000  a1.channels.c1.transactionCapacity = 100    # Bind the source and sink to the channel  a1.sources.r1.channels = c1  a1.sinks.k1.channel = c1 | |

b.启动flume agent

|  |
| --- |
| /usr/local/flume/bin/flume-ng agent --conf ./conf --conf-file ./conf/netcat.conf --name a1 -Dflume.root.logger=INFO,console  #不要关闭这个终端 |

c.新打开一个终端输入：

|  |
| --- |
| telnet localhost 44444  #在这个终端输入字符串就可以显示在前面那个终端里了，但是中文是不支持的，显示长度也有限 |

## 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。

【参考答案】

a.创建flume-to-spark.conf

|  |  |
| --- | --- |
| 1.cd /usr/local/flume/conf  2.sudo vim ./flume-to-spark.conf  #复制以下内容   |  | | --- | | #flume-to-spark.conf: A single-node Flume configuration  # Name the components on this agent  a1.sources = r1  a1.sinks = k1  a1.channels = c1  # Describe/configure the source  a1.sources.r1.type = netcat  a1.sources.r1.bind = localhost  a1.sources.r1.port = 33333  # Describe the sink  a1.sinks.k1.type = avro  a1.sinks.k1.hostname = localhost  a1.sinks.k1.port =44444  # Use a channel which buffers events in memory  a1.channels.c1.type = memory  a1.channels.c1.capacity = 1000000  a1.channels.c1.transactionCapacity = 1000000  # Bind the source and sink to the channel  a1.sources.r1.channels = c1  a1.sinks.k1.channel = c1 |   #说明：  1.Flume suorce类为netcat，绑定到localhost的33333端口，消息可以通过telnet localhost 33333 发送到flume suorce  2.Flume Sink类为avro，绑定44444端口，flume sink通过localhost 44444端口把消息发送出来。而spark streaming程序一直监听44444端口。  #注意！！先不要启动Flume agent，因为44444端口还没打开，sink的消息无处可去，44444端口由spark streaming程序打开。 |

b.spark准备工作

|  |  |
| --- | --- |
| 1.下载spark-streaming-flume\_2.11-2.1.0.jar，2.11对应scala，2.10对应spark版本  下载地址：<http://central.maven.org/maven2/org/apache/spark/spark-streaming-flume_2.11/2.1.0/spark-streaming-flume_2.11-2.1.0.jar>  #把这个jar文件放到/usr/local/spark/jars/flume目录下  2.sudo cp ./spark-streaming-flume\_2.11-2.1.0.jar /usr/local/spark/jars/flume/  3.修改spark目录下conf/spark-env.sh文件中的SPARK\_DIST\_CLASSPATH变量.把flume的相关jar包添加到此文件中。   |  | | --- | | export SPARK\_DIST\_CLASSPATH=$(/usr/local/hadoop/bin/hadoop classpath):$(/usr/local/hbase/bin/hbase classpath):/usr/local/spark/examples/jars/\*:/usr/local/spark/jars/kafka/\*:/usr/local/kafka/libs/\*:/usr/local/spark/jars/flume/\*:/usr/local/flume/lib/\* |   这样我们就准备好spark了。 |

c.编写spark程序使用Flume数据源

|  |  |  |  |
| --- | --- | --- | --- |
| 1.创建python文件   |  |  | | --- | --- | | cd /usr/local/spark/mycode  mkdir flume  cd flume  sudo vim FlumeEventCount.py  #复制以下代码   |  | | --- | | from \_\_future\_\_ import print\_function  import sys  from pyspark import SparkContext  from pyspark.streaming import StreamingContext  from pyspark.streaming.flume import FlumeUtils  import pyspark  if \_\_name\_\_ == "\_\_main\_\_":  if len(sys.argv) != 3:  print("Usage: flume\_wordcount.py <hostname> <port>", file=sys.stderr)  exit(-1)  sc = SparkContext(appName="FlumeEventCount")  ssc = StreamingContext(sc, 2)  hostname= sys.argv[1]  port = int(sys.argv[2])  stream = FlumeUtils.createStream(ssc, hostname, port,pyspark.StorageLevel.MEMORY\_AND\_DISK\_SER\_2)  stream.count().map(lambda cnt : "Recieve " + str(cnt) +" Flume events!!!!").pprint()  ssc.start()  ssc.awaitTermination() |   # 注意可能需要安装pyspark，命令  sudo pip install pyspark |   2.测试实际效果   |  | | --- | | 首先启动Spark streaming程序  ./bin/spark-submit --driver-class-path /usr/local/spark/jars/\*:/usr/local/spark/jars/flume/\* ./mycode/flume/FlumeEventCount.py localhost 44444  然后启动一个新的终端，启动Flume Agent  cd /usr/local/flume  bin/flume-ng agent --conf ./conf --conf-file ./conf/flume-to-spark.conf --name a1 -Dflume.root.logger=INFO,console  最后再启动一个新的终端连接33333端口  telnet localhost 33333  现在你可以在最后这个终端里输入一些字符了。在你输入字符后可以看到第一个终端会显示如下的信息  -------------------------------------------  Time: 1488029430000 ms  -------------------------------------------  Received 1 flume events！！！ | |

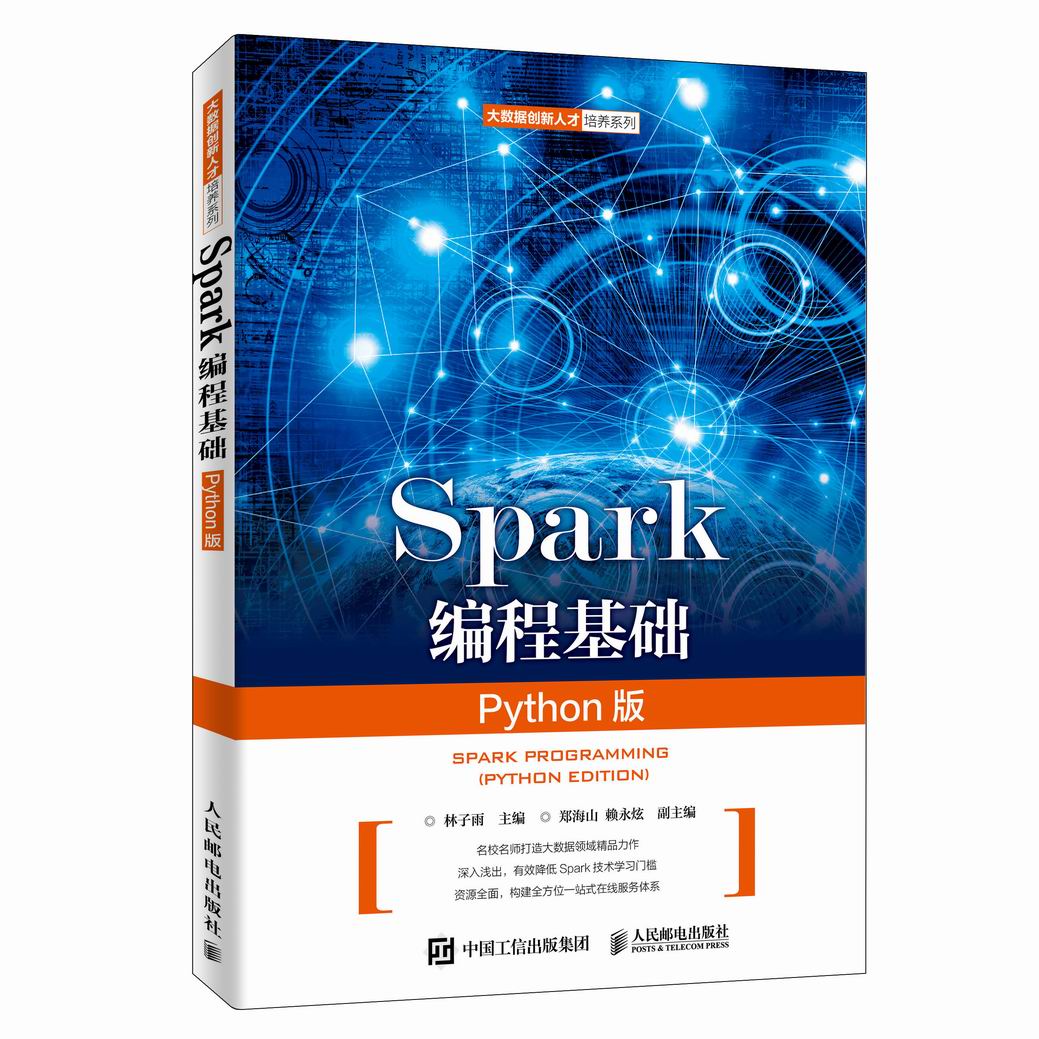
# 四、实验报告

|  |  |  |
| --- | --- | --- |
| 《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