华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统 **年级**：2019 **上机实践成绩**：

指导教师：徐辰 **姓名**：赵煜硕 **学号**：10195501415

上机实践名称：Spark 部署 **上机实践日期**：

很顺利无坑

华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　　**年级**：2019　　**上机实践成绩**：

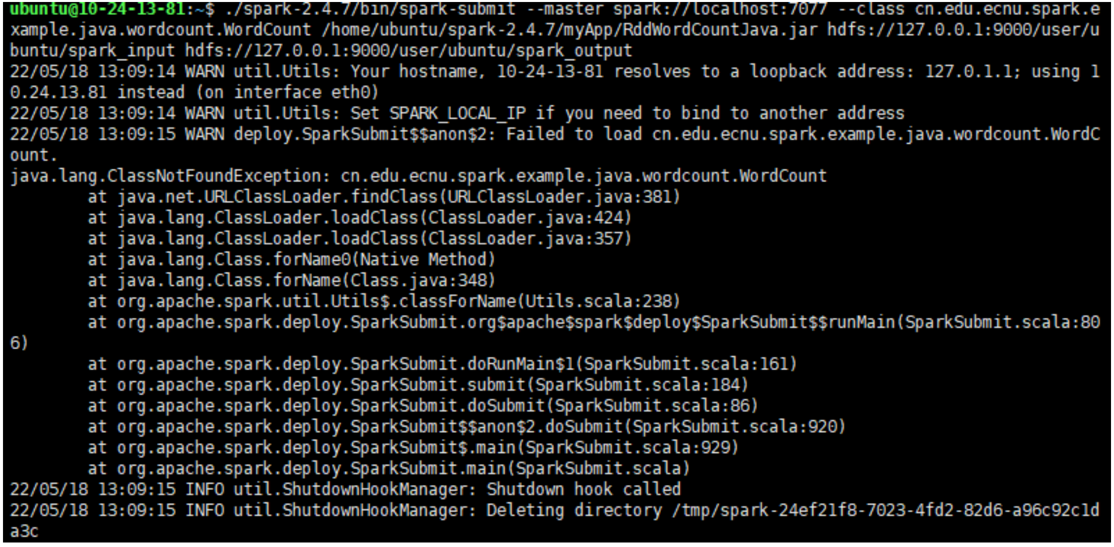
指导教师：徐辰　　**姓名**：赵煜硕　　**学号**：10195501415

上机实践名称：　Spark 部署　　**上机实践日期**：

ScalaSDK 死活装不上，解决方案：不用 scala

执行

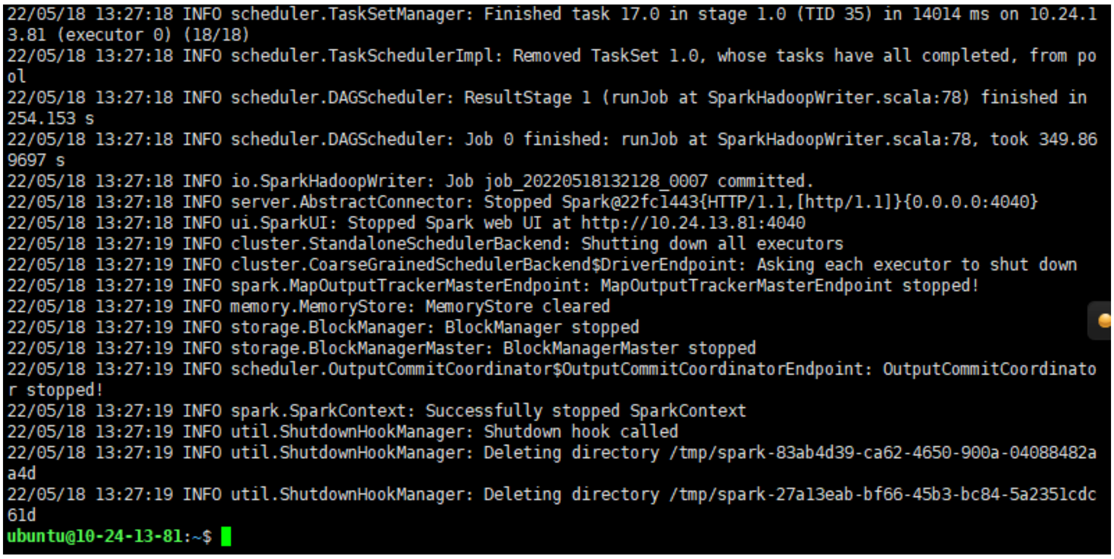
~/spark-2.4.7/bin/spark-submit
--master spark://localhost:7077 --class
cn.edu.ecnu.spark.example.java.wordcount.WordCount
/home/ubuntu/spark-2.4.7/myApp/RddWordCountJava.jar
hdfs://localhost:9000/user/ubuntu/spark_input
hdfs://localhost:9000/user/ubuntu/spark_output



```
ubuntu@10-24-13-81:~$ ./spark-2.4.7/bin/spark-submit --master spark://localhost:7077 --class cn.edu.ecnu.spark.e
xample.java.wordcount.WordCount /home/ubuntu/spark-2.4.7/myApp/RddWordCountJava.jar hdfs://127.0.0.1:9000/user/u
buntu/spark_input hdfs://127.0.0.1:9000/user/ubuntu/spark_output
22/05/18 13:09:14 WARN util.Utils: Your hostname, 10-24-13-81 resolves to a loopback address: 127.0.1.1; using 1
0.24.13.81 instead (on interface eth0)
22/05/18 13:09:14 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
22/05/18 13:09:15 WARN deploy.SparkSubmit$$anon$2: Failed to load cn.edu.ecnu.spark.example.java.wordcount.WordC
ount.
java.lang.ClassNotFoundException: cn.edu.ecnu.spark.example.java.wordcount.WordCount
        at java.net.URLClassLoader.findClass(URLClassLoader.java:381)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:424)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:357)
        at java.lang.Class.forName0(Native Method)
        at java.lang.Class.forName(Class.java:348)
        at org.apache.spark.util.Utils$.classForName(Utils.scala:238)
        at org.apache.spark.deploy.SparkSubmit.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:80
6)
        at org.apache.spark.deploy.SparkSubmit.doRunMain$1(SparkSubmit.scala:161)
        at org.apache.spark.deploy.SparkSubmit.submit(SparkSubmit.scala:184)
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:86)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:920)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:929)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
22/05/18 13:09:15 INFO util.ShutdownHookManager: Shutdown hook called
22/05/18 13:09:15 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-24ef21f8-7023-4fd2-82d6-a96c92c1d
a3c
```

原因是在本地编写目录 cn.edu.ecnu.spark.example.java.wordcount 时出错

重新执行



```
22/05/18 13:27:18 INFO scheduler.TaskSetManager: Finished task 17.0 in stage 1.0 (TID 35) in 14014 ms on 10.24.1
3.81 (executor 0) (18/18)
22/05/18 13:27:18 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from po
ol
22/05/18 13:27:18 INFO scheduler.DAGScheduler: ResultStage 1 (runJob at SparkHadoopWriter.scala:78) finished in
254.153 s
22/05/18 13:27:18 INFO scheduler.DAGScheduler: Job 0 finished: runJob at SparkHadoopWriter.scala:78, took 349.86
9697 s
22/05/18 13:27:18 INFO io.SparkHadoopWriter: Job job_20220518132128_0007 committed.
22/05/18 13:27:18 INFO server.AbstractConnector: Stopped Spark@22fc1443{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
22/05/18 13:27:18 INFO ui.SparkUI: Stopped Spark web UI at http://10.24.13.81:4040
22/05/18 13:27:19 INFO cluster.StandaloneSchedulerBackend: Shutting down all executors
22/05/18 13:27:19 INFO cluster.CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
22/05/18 13:27:19 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/05/18 13:27:19 INFO memory.MemoryStore: MemoryStore cleared
22/05/18 13:27:19 INFO storage.BlockManager: BlockManager stopped
22/05/18 13:27:19 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
22/05/18 13:27:19 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinato
r stopped!
22/05/18 13:27:19 INFO spark.SparkContext: Successfully stopped SparkContext
22/05/18 13:27:19 INFO util.ShutdownHookManager: Shutdown hook called
22/05/18 13:27:19 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-83ab4d39-ca62-4650-900a-04088482a
a4d
22/05/18 13:27:19 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-27a13eab-bf66-45b3-bc84-5a2351cdc
61d
ubuntu@10-24-13-81:~$
```

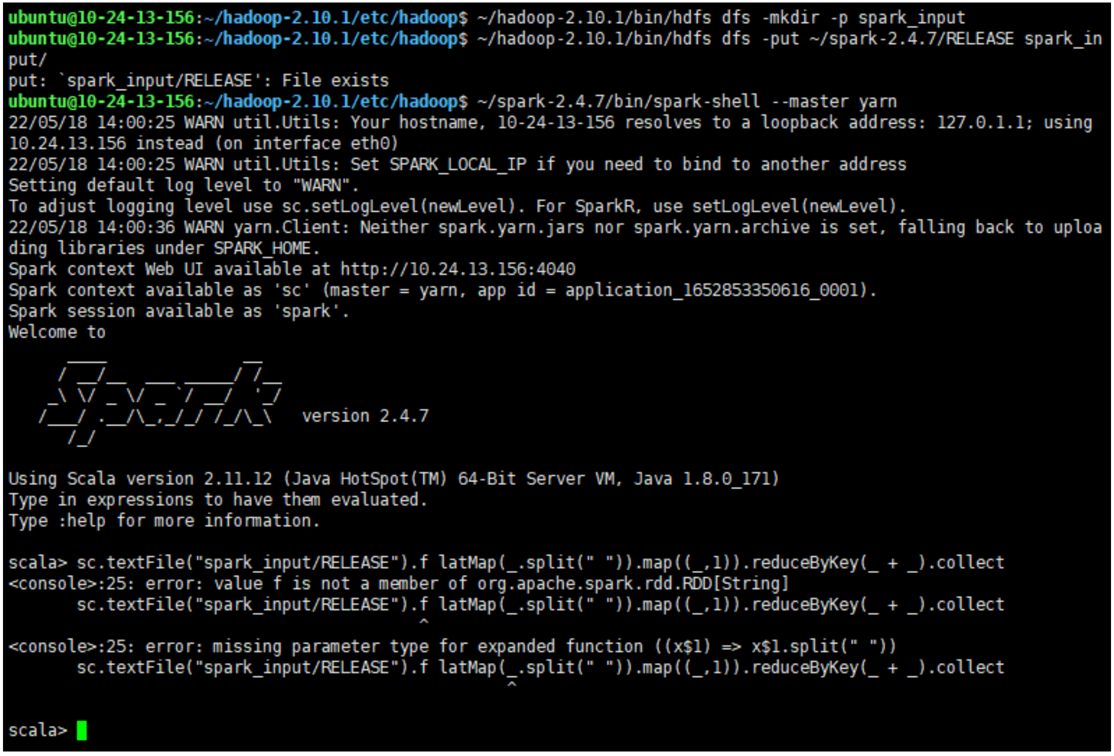
华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　　**年级：**2019　　**上机实践成绩：**

指导教师：徐辰　　**姓名：**赵煜硕　　**学号：**10195501415

上机实践名称：yarn 部署　　**上机实践日期：**

运行 Spark 程序
sc.textFile("spark_input/RELEASE").flatMap(_.split(" ")).map((_,1)).reduceByKey(_ + _).collect



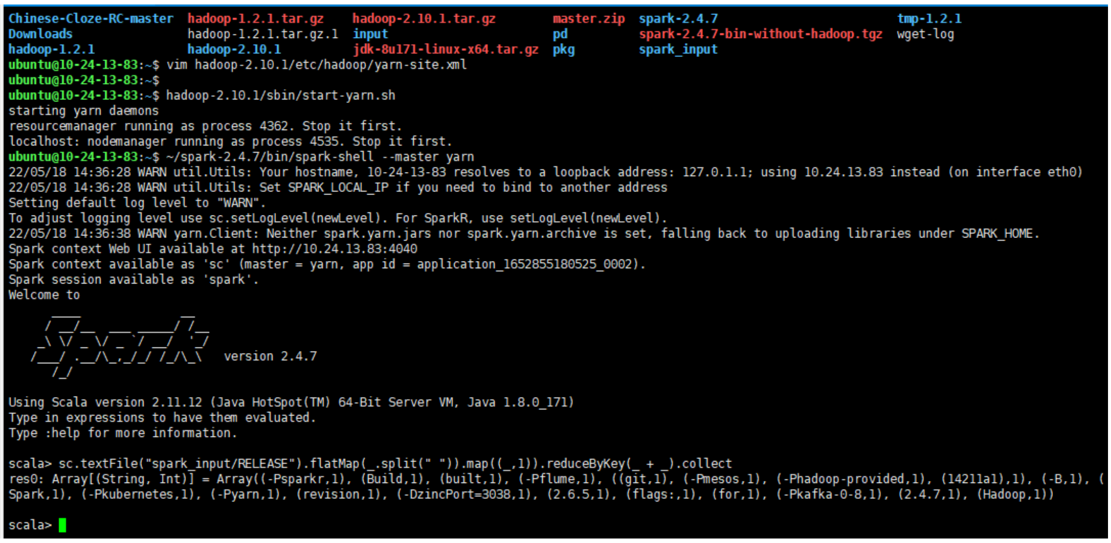
```
ubuntu@10-24-13-156:~/hadoop-2.10.1/etc/hadoop$ ~/hadoop-2.10.1/bin/hdfs dfs -mkdir -p spark_input
ubuntu@10-24-13-156:~/hadoop-2.10.1/etc/hadoop$ ~/hadoop-2.10.1/bin/hdfs dfs -put ~/spark-2.4.7/RELEASE spark_in
put/
put: `spark_input/RELEASE': File exists
ubuntu@10-24-13-156:~/hadoop-2.10.1/etc/hadoop$ ~/spark-2.4.7/bin/spark-shell --master yarn
22/05/18 14:00:25 WARN util.Utils: Your hostname, 10-24-13-156 resolves to a loopback address: 127.0.1.1; using
10.24.13.156 instead (on interface eth0)
22/05/18 14:00:25 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/18 14:00:36 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploa
ding libraries under SPARK_HOME.
Spark context Web UI available at http://10.24.13.156:4040
Spark context available as 'sc' (master = yarn, app id = application_1652853350616_0001).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.7
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_171)
Type in expressions to have them evaluated.
Type :help for more information.

scala> sc.textFile("spark_input/RELEASE").f latMap(_.split(" ")).map((_,1)).reduceByKey(_ + _).collect
<console>:25: error: value f is not a member of org.apache.spark.rdd.RDD[String]
       sc.textFile("spark_input/RELEASE").f latMap(_.split(" ")).map((_,1)).reduceByKey(_ + _).collect
                                          ^
<console>:25: error: missing parameter type for expanded function ((x$1) => x$1.split(" "))
       sc.textFile("spark_input/RELEASE").f latMap(_.split(" ")).map((_,1)).reduceByKey(_ + _).collect
                                                   ^

scala> 
```

发现是书本印刷错误，faltMap 连
重新执行



```
Chinese-Cloze-RC-master  hadoop-1.2.1.tar.gz    hadoop-2.10.1.tar.gz        master.zip  spark-2.4.7                            tmp-1.2.1
Downloads                hadoop-1.2.1.tar.gz.1  input                       pd          spark-2.4.7-bin-without-hadoop.tgz  wget-log
hadoop-1.2.1             hadoop-2.10.1          jdk-8u171-linux-x64.tar.gz  pkg         spark_input
ubuntu@10-24-13-83:~$ vim hadoop-2.10.1/etc/hadoop/yarn-site.xml
ubuntu@10-24-13-83:~$
ubuntu@10-24-13-83:~$ hadoop-2.10.1/sbin/start-yarn.sh
starting yarn daemons
resourcemanager running as process 4362. Stop it first.
localhost: nodemanager running as process 4535. Stop it first.
ubuntu@10-24-13-83:~$ ~/spark-2.4.7/bin/spark-shell --master yarn
22/05/18 14:36:28 WARN util.Utils: Your hostname, 10-24-13-83 resolves to a loopback address: 127.0.1.1; using 10.24.13.83 instead (on interface eth0)
22/05/18 14:36:28 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/18 14:36:38 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Spark context Web UI available at http://10.24.13.83:4040
Spark context available as 'sc' (master = yarn, app id = application_1652855180525_0002).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.7
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_171)
Type in expressions to have them evaluated.
Type :help for more information.

scala> sc.textFile("spark_input/RELEASE").flatMap(_.split(" ")).map((_,1)).reduceByKey(_ + _).collect
res0: Array[(String, Int)] = Array((-Psparkr,1), (Build,1), (built,1), (-Pflume,1), ((git,1), (-Pmesos,1), (-Phadoop-provided,1), (14211a1),1), (-B,1), (
Spark,1), (-Pkubernetes,1), (-Pyarn,1), (revision,1), (-DzincPort=3038,1), (2.6.5,1), (flags:,1), (for,1), (-Pkafka-0-8,1), (2.4.7,1), (Hadoop,1))

scala> 
```