**实验四 bug报告**

陈朴炎 2021211138

目录

[1 bug报告 3](#_Toc167816915)

[bug1 3](#_Toc167816916)

[出错部分 3](#_Toc167816917)

[修改方式 3](#_Toc167816918)

[bug2 5](#_Toc167816919)

[出错部分 5](#_Toc167816920)

[修改方式 6](#_Toc167816921)

[bug3 8](#_Toc167816922)

[出错部分 8](#_Toc167816923)

[修改方式 8](#_Toc167816924)

[bug4 9](#_Toc167816925)

[出错部分 9](#_Toc167816926)

[修改方式 9](#_Toc167816927)

[bug5 11](#_Toc167816928)

[出错部分 11](#_Toc167816929)

[修改方式 11](#_Toc167816930)

[bug6 12](#_Toc167816931)

[出错部分 12](#_Toc167816932)

[修改方式 12](#_Toc167816933)

[bug7 14](#_Toc167816934)

[出错部分 14](#_Toc167816935)

[修改方式 14](#_Toc167816936)

[bug8 14](#_Toc167816937)

[出错部分 14](#_Toc167816938)

[修改方式 15](#_Toc167816939)

[bug9 15](#_Toc167816940)

[出错部分 15](#_Toc167816941)

[修改方式 16](#_Toc167816942)

[bug10 16](#_Toc167816943)

[出错部分 16](#_Toc167816944)

[修改方式 17](#_Toc167816945)

[2 实验结果截图 17](#_Toc167816946)

[2.1 Hadoop集群测试结果 17](#_Toc167816947)

[2.2 Spark集群搭建完成的测试结果 18](#_Toc167816948)

[2.3 Scala单词计数实验结果 18](#_Toc167816949)

[2.4 RDD编程结果 19](#_Toc167816950)

[2.5 Spark sql读写数据库结果 19](#_Toc167816951)

# 1 bug报告

## bug1

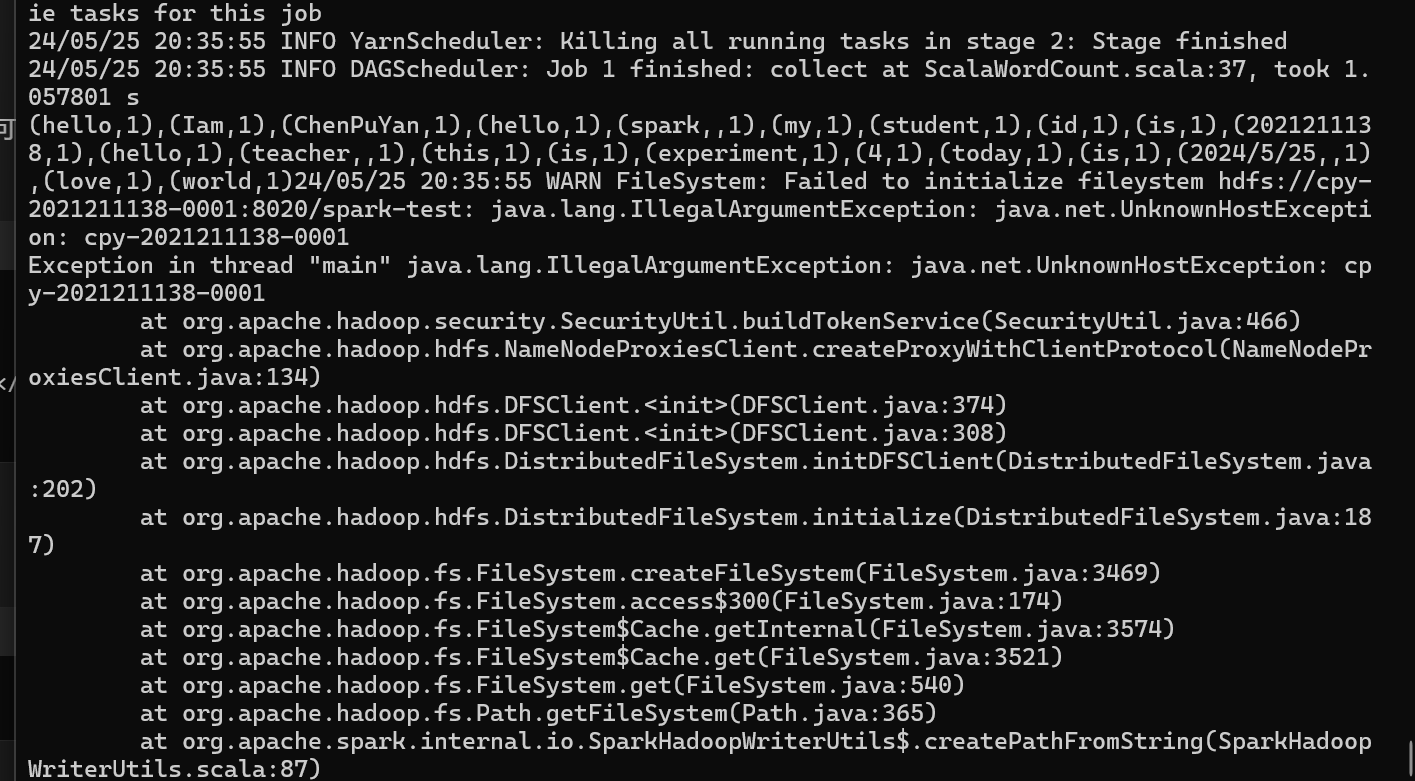
### 出错部分

在第一部分中，配置完spark环境，上传完程序，输入spark-submmit命令后，

spark-submit --class org.example.ScalaWordCount --master yarn --num-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test.jar

执行结果如下：

报错：



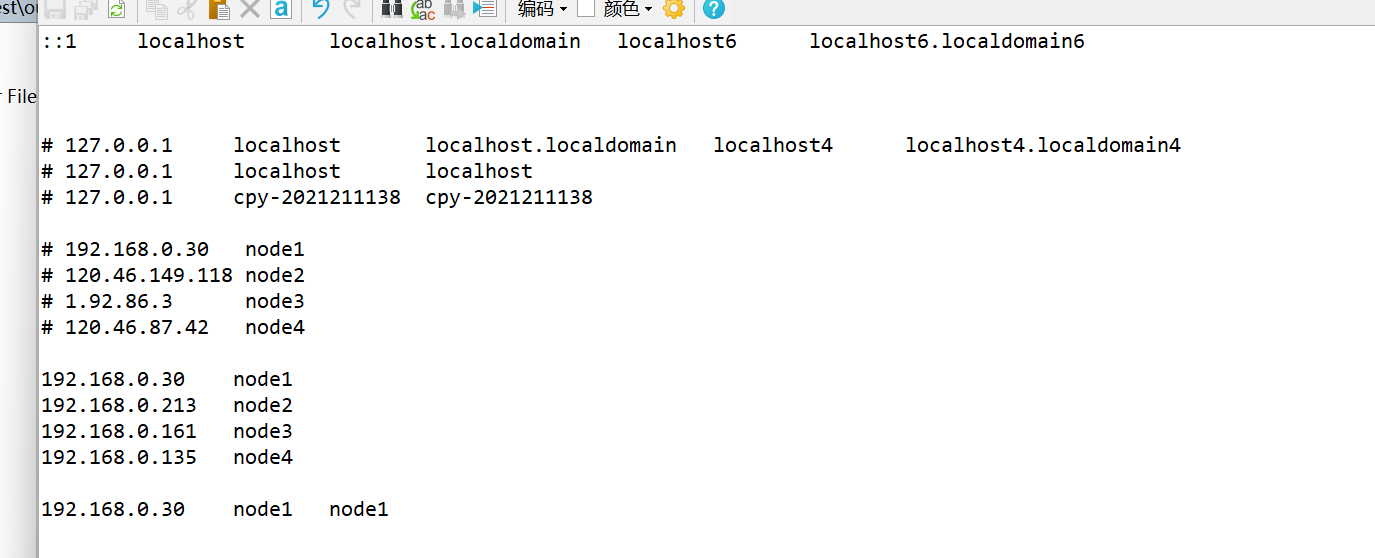
查看异常信息，主要错在：

java.lang.IllegalArgumentException: java.net.UnknownHostException: cpy-2021211138-0001

也就是说我的主机名不对。

### 修改方式

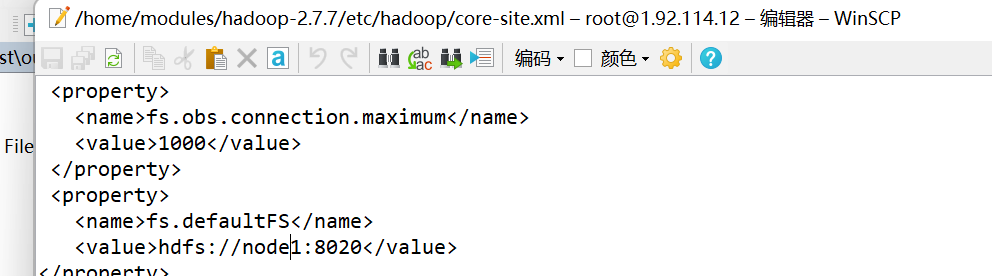
查看/etc/hosts文件：



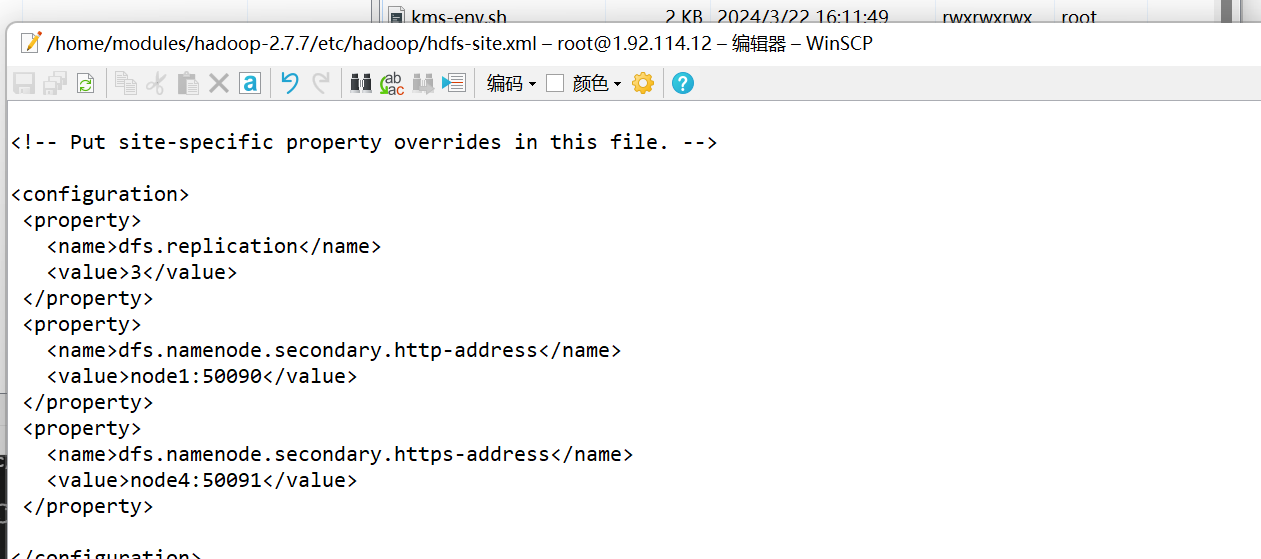
这里都是用node1、node2、node3、node4，并没有其他的主机名

但是我自己回想一下，好像有些地方确实改了主机名，于是一个一个排查：

hadoop下的core-site.xml中用的是node1，不是这个文件出错

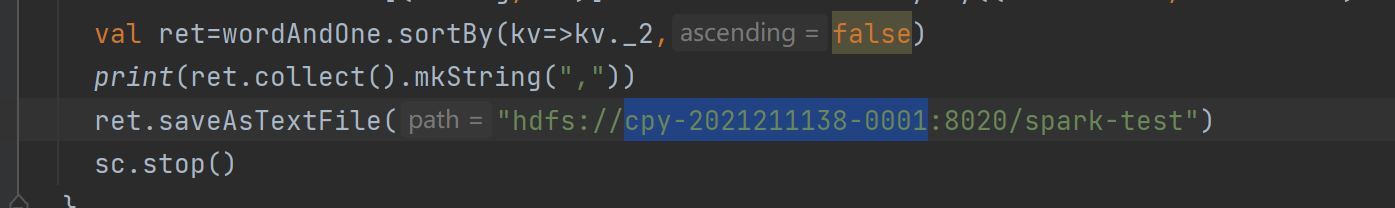


hdfs-site.xml用的也是node的编号



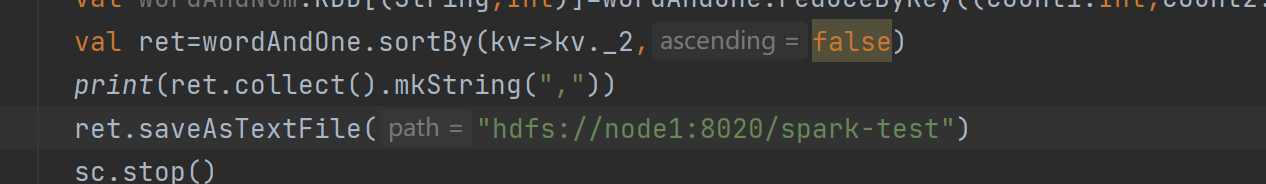
yarn-site.xml中也没有

找了半天，发现原来是代码错了：



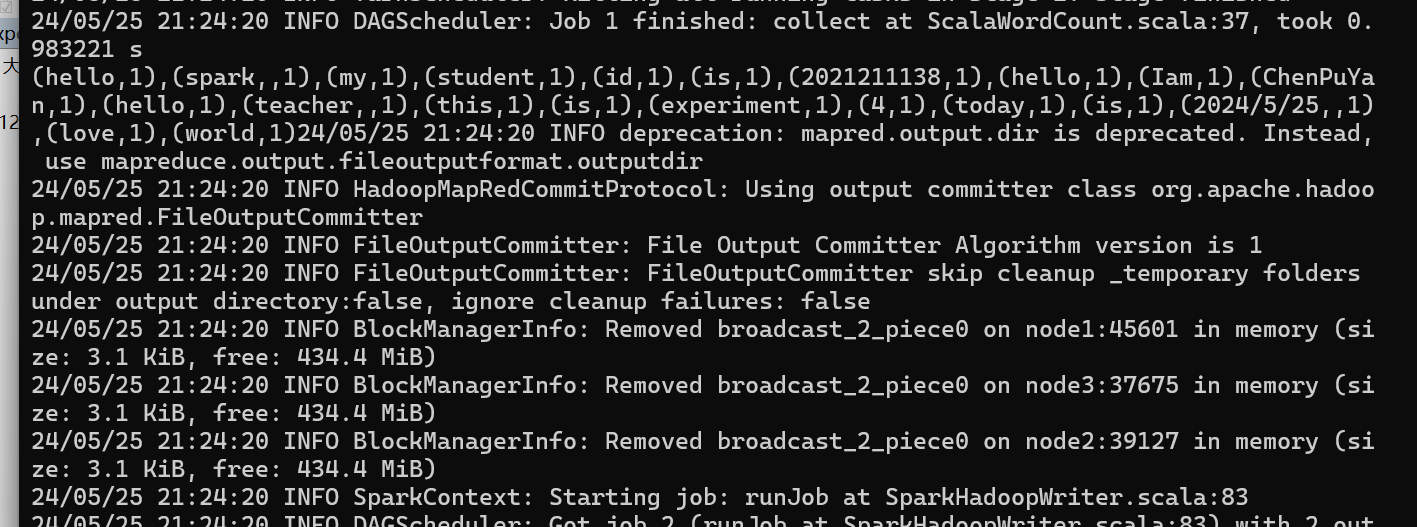
我应该在标蓝的这里写node1的

修改如下：



之后重新打包，重新上传

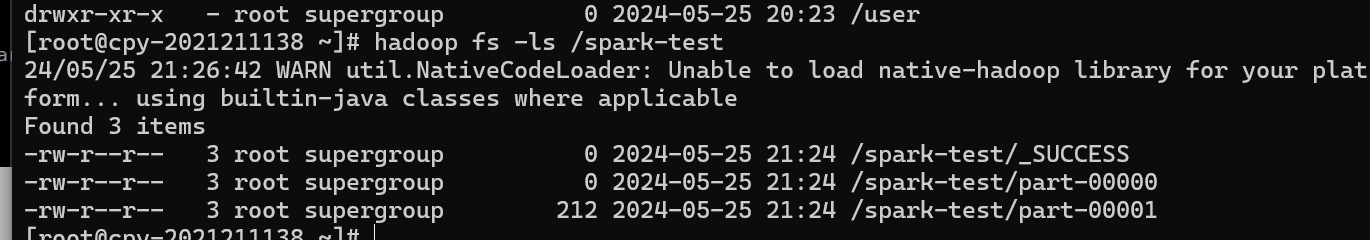
之后再运行一次就没有明显的报错了



## bug2

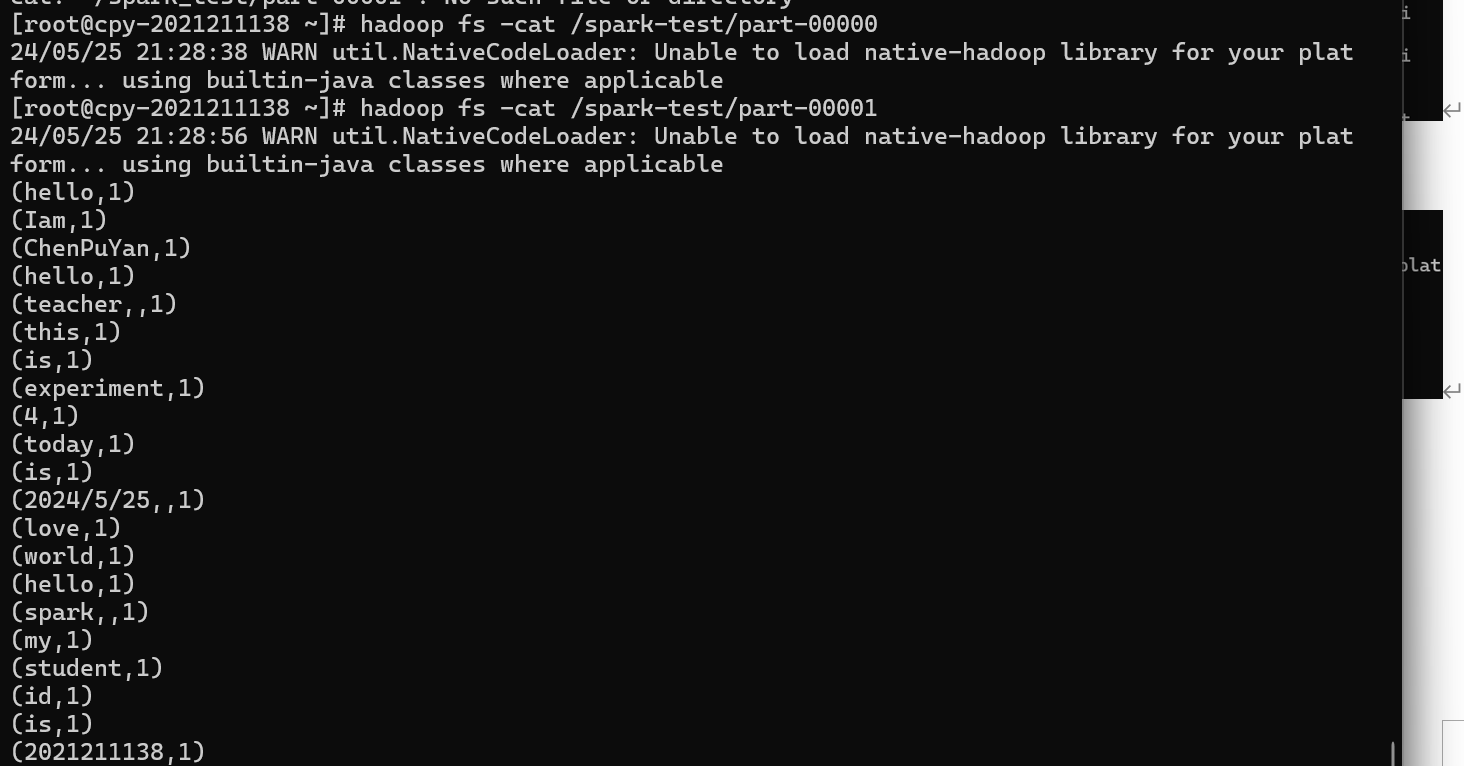
### 出错部分

在hdfs上查看程序输出：



出现两个文件

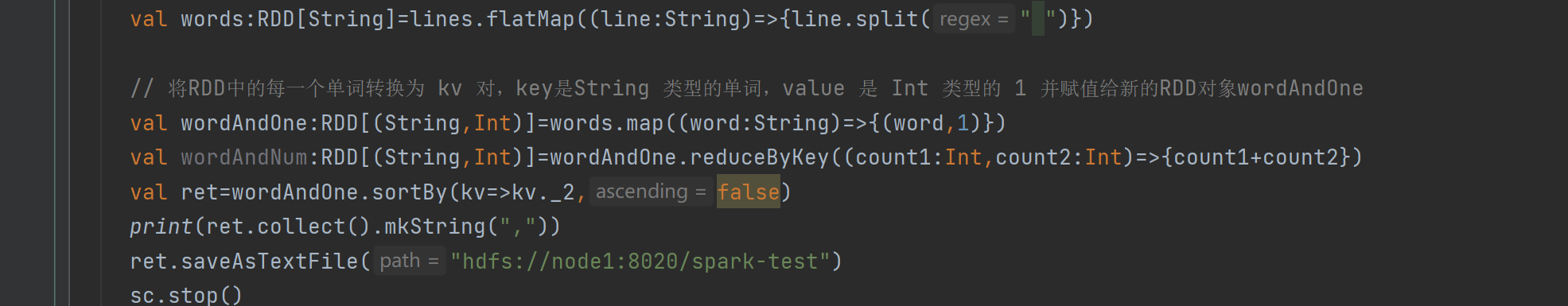
使用 hadoop fs -cat /spark-test/part-0000x来查看输出信息，得到如下内容：



但是这里只有一个文件有内容，并且我也只有两个文件，跟实验报告中的4个文件显然不同。

### 修改方式

原因如下：





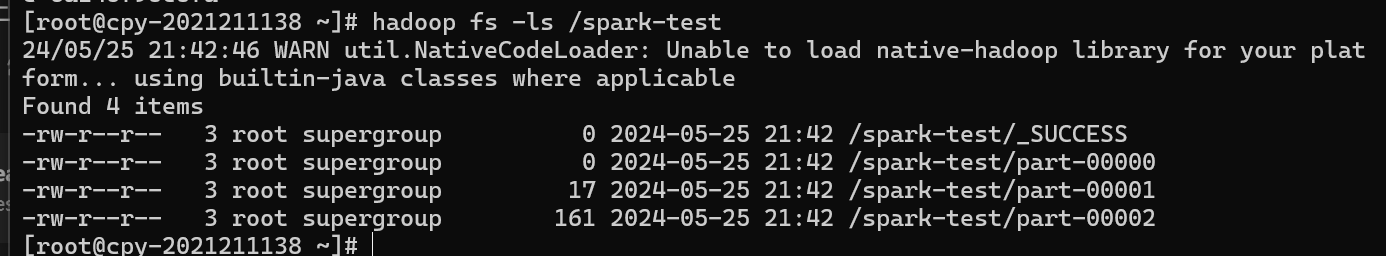
标蓝的地方写错了，应该是wordAndNum而不是wordAndOne。

修改后，如下：



重新打包，删除MANIFEST.FT文件，重新上传，并删除hadoop fs里面的/spark-test文件夹，

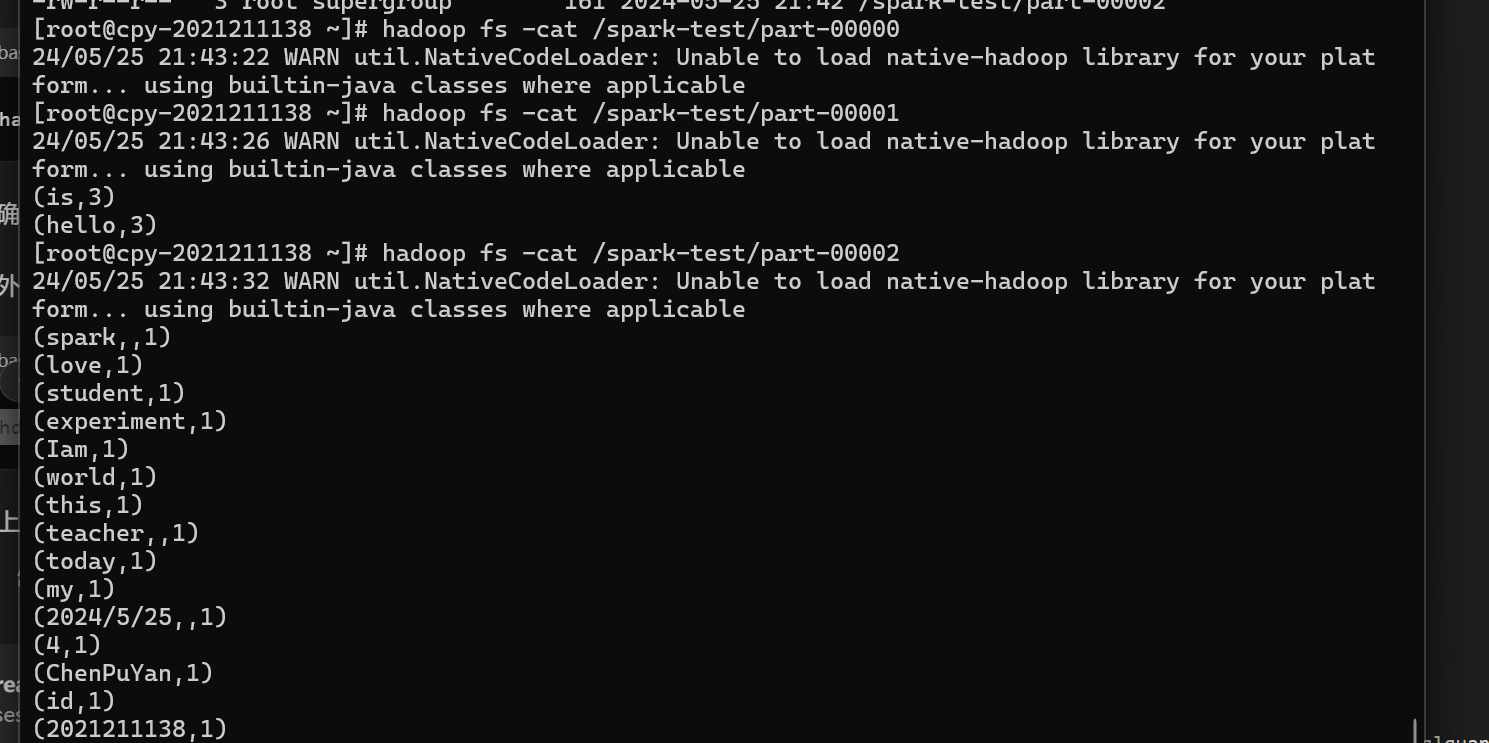
重新运行：



可以看到现在有了三个输出文件

检擦三个输出文件

结果如下：



得到正确的结果。

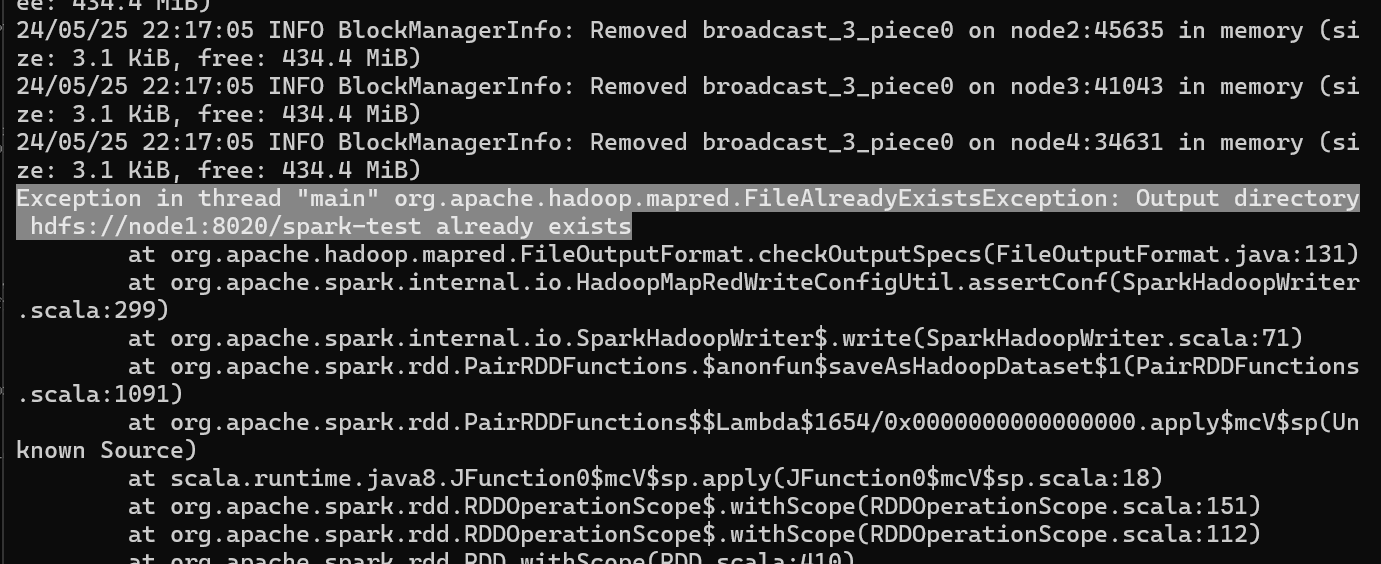
## bug3

### 出错部分

第二个编程实验中，上传完程序，输入如下submmit指令：

spark-submit --class org.example.ScalaWordCount --master yarn --num-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test2.jar

出现报错：

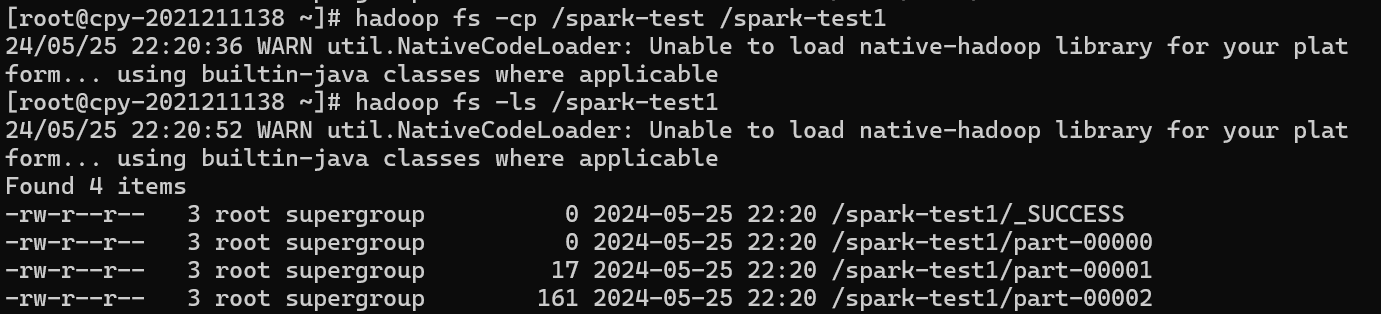


它说已经存在该路径了

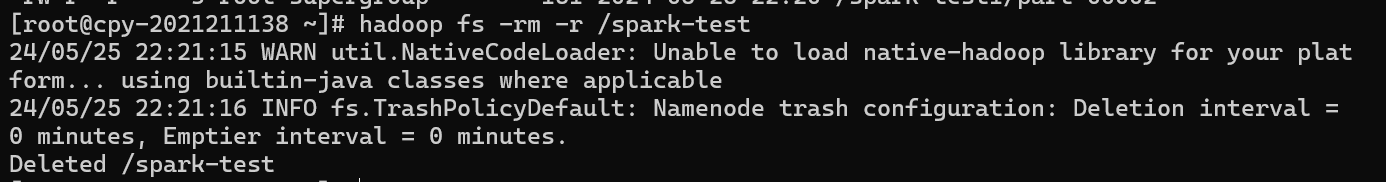
### 修改方式

那么就将原来test1的输出文件夹改为spark-test1，然后再将spark-test文件夹删掉，步骤如下：

复制，查看



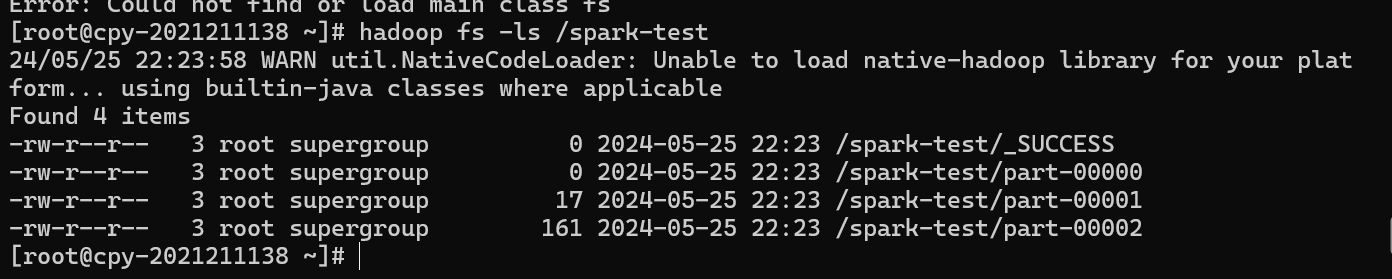
删除



## bug4

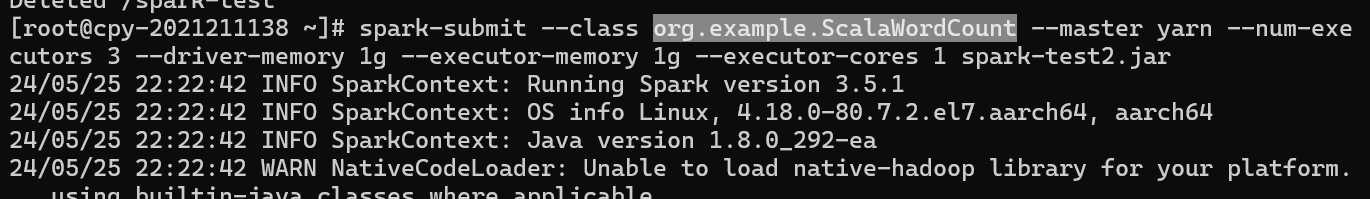
### 出错部分

第二个编程实验中，发现结果错误了：



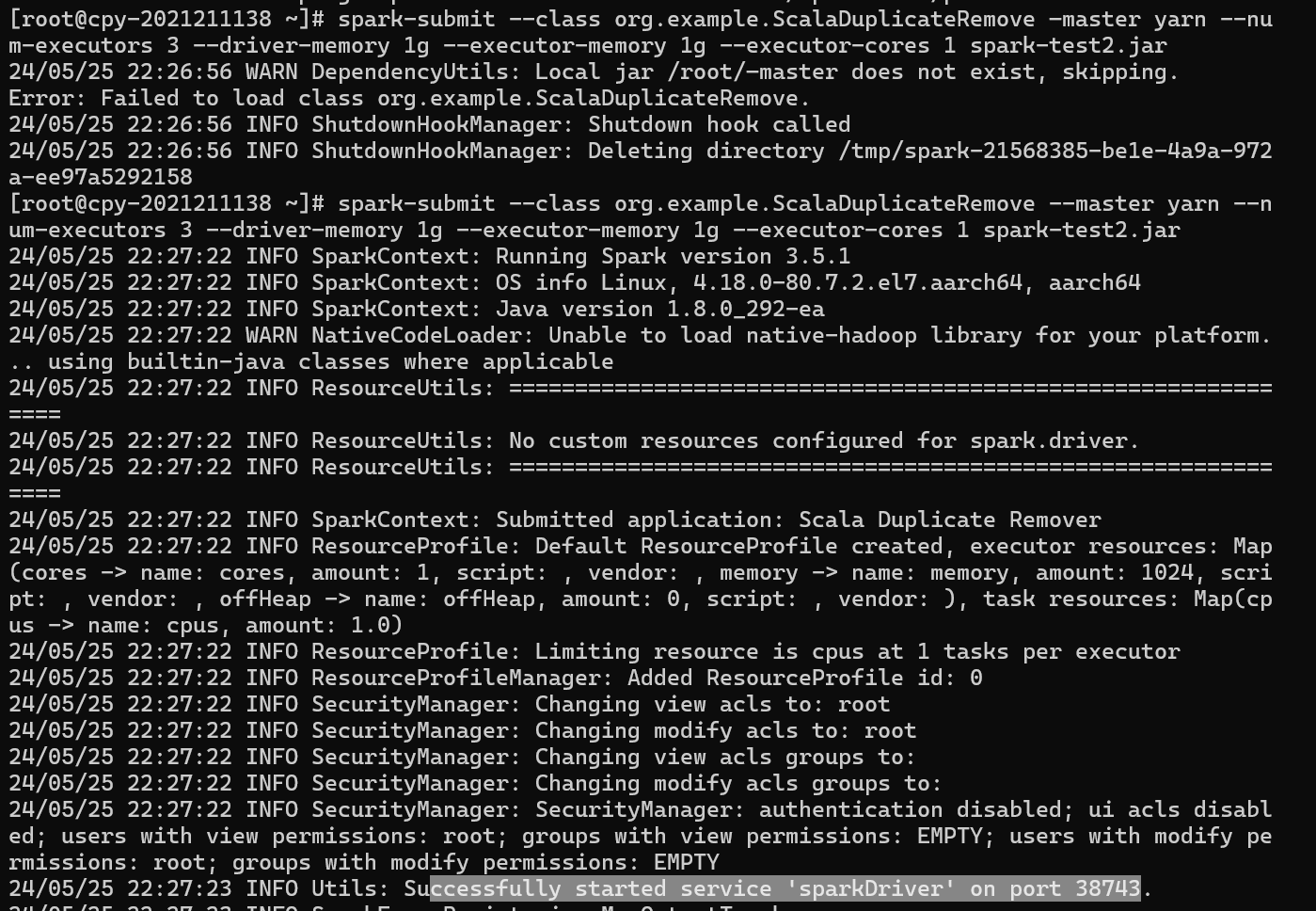
### 修改方式

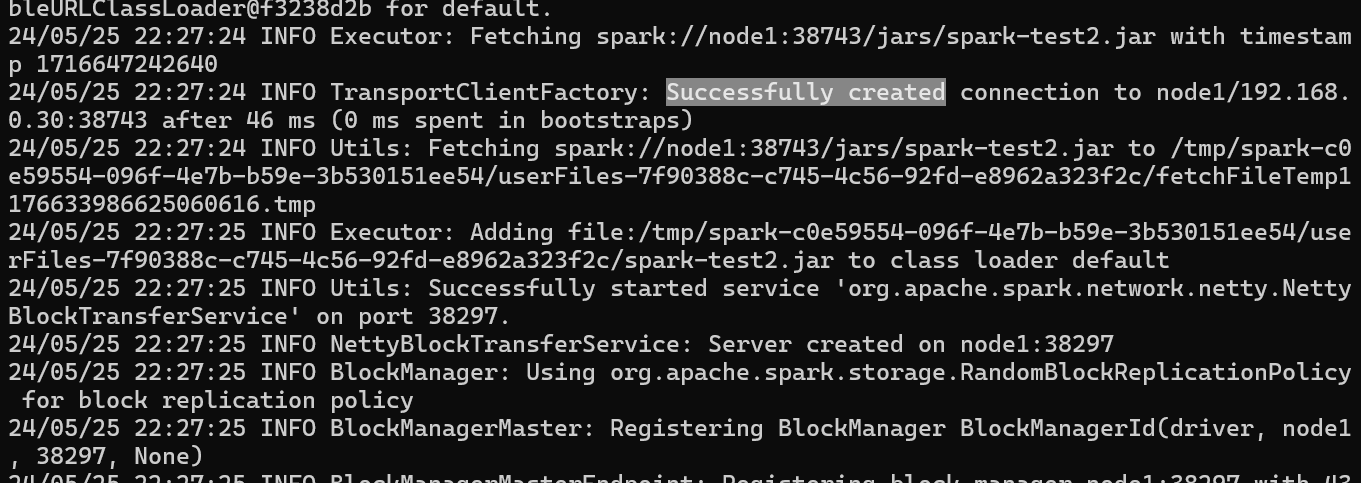
问题在于，运行的命令输错了

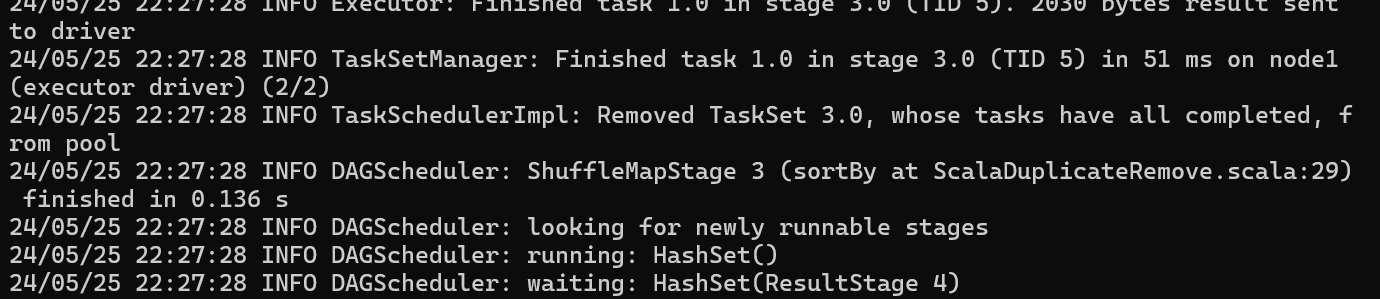


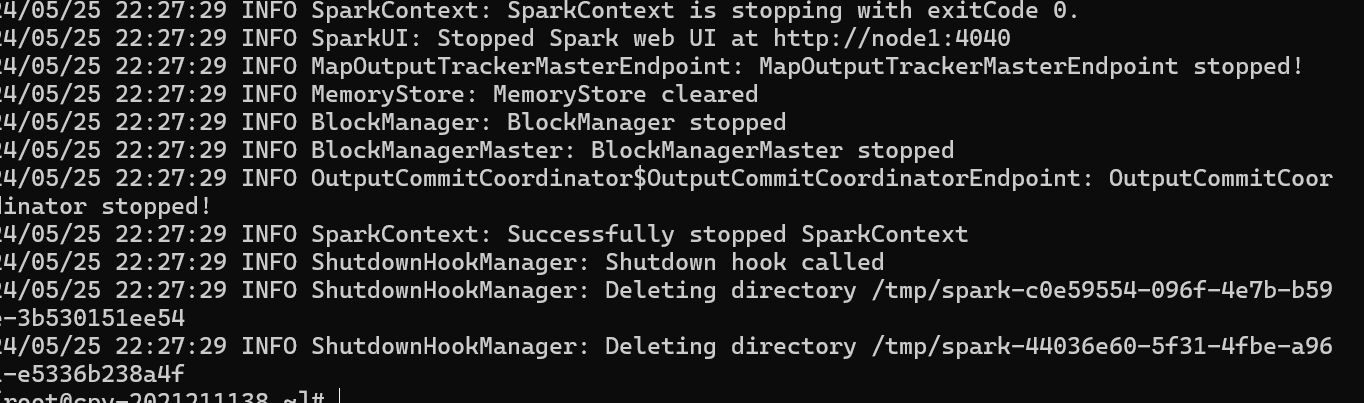
不是这个类，而是另一个，重新输入指令，如下：

spark-submit --class org.example.ScalaDuplicateRemove --master yarn --num-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test2.jar



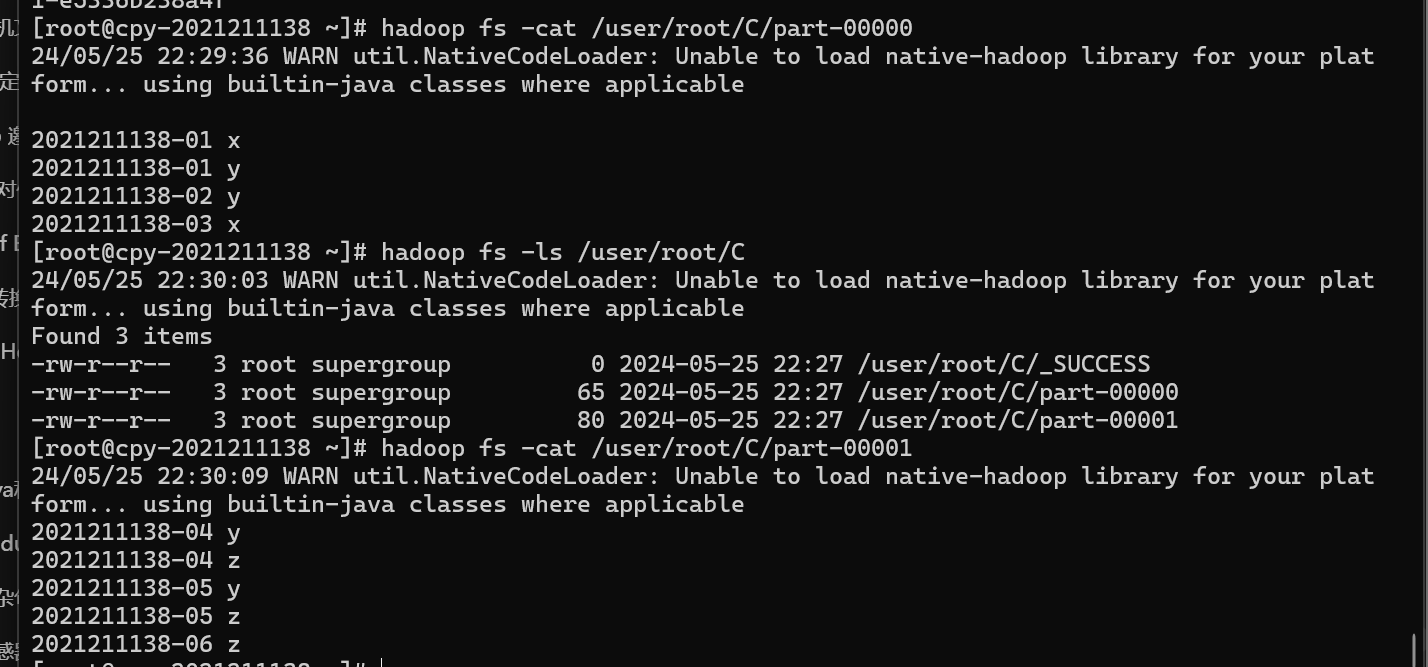






运行命令查看执行结果：

hadoop fs -cat /user/root/C/part-00000



## bug5

### 出错部分

安装mysql出现报错：

The GPG keys listed for the "MySQL 8.0 Community Server" repository are already installed but they are not correct for this package. Check that the correct key URLs are configured for this repository. Failing package is: mysql-community-common-8.0.37-1.el7.aarch64 GPG Keys are configured as: file:///etc/pki/rpm-gpg/RPM-GPG-KEY-mysql

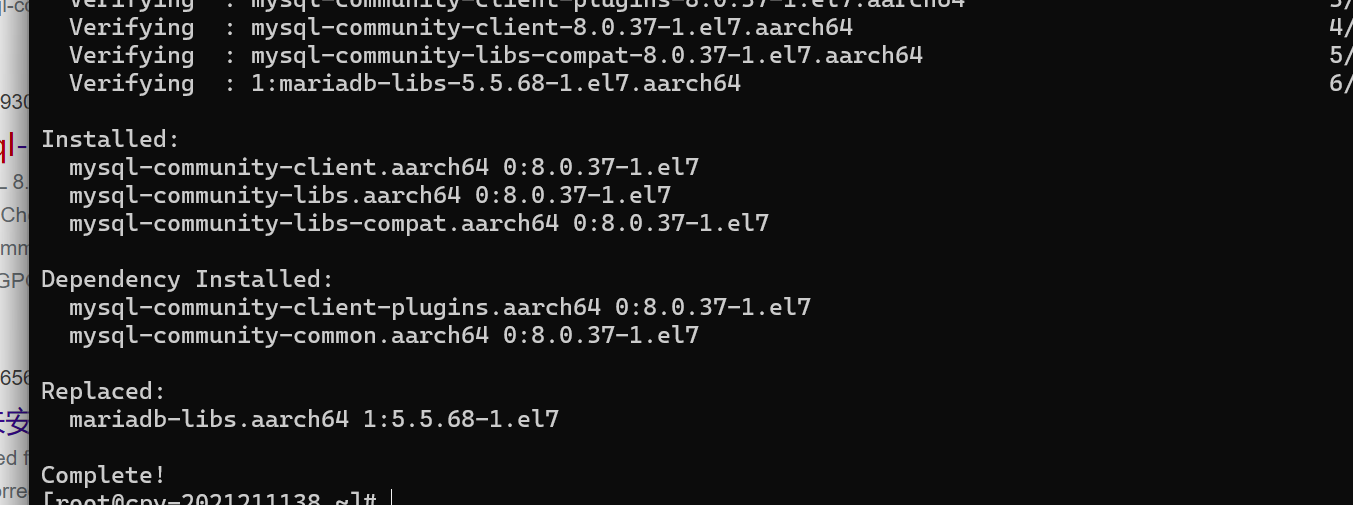
### 修改方式

这是因为GPG公钥验证问题

输入：

sudo rpm --import <https://repo.mysql.com/RPM-GPG-KEY-mysql-2023>

就能解决

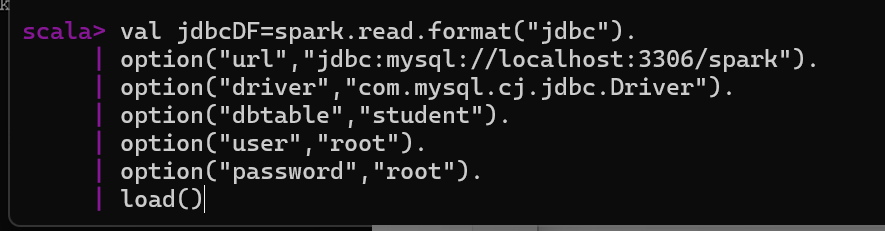


下载完成

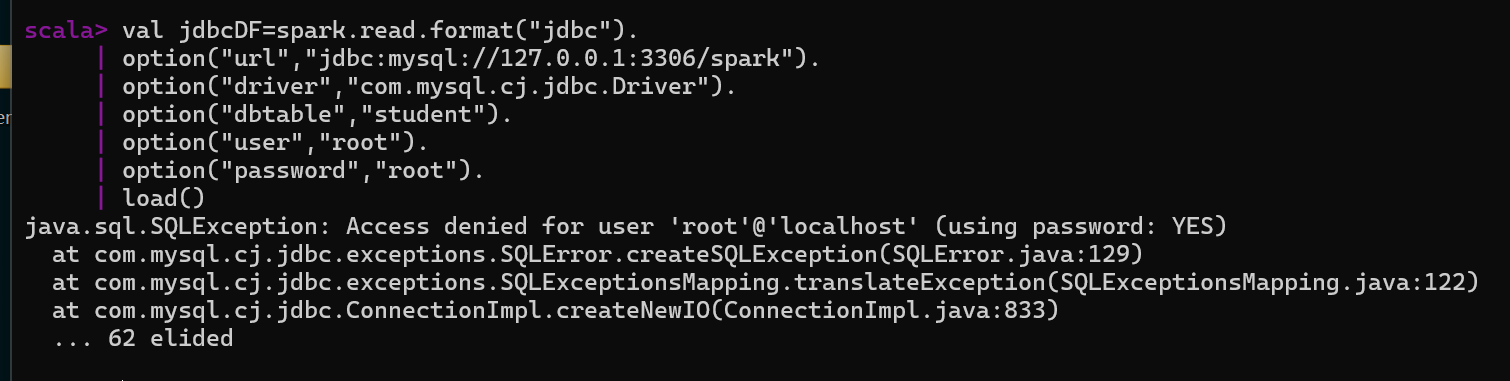
## bug6

### 出错部分

通过mysql连接驱动jar包之后，输入实验指导书中的代码



出现：



出现access denied错误

### 修改方式

原因可能是因为mysql还没启动

输入：

sudo systemctl start mysqld

启动mysql，并重新登录spark-shell

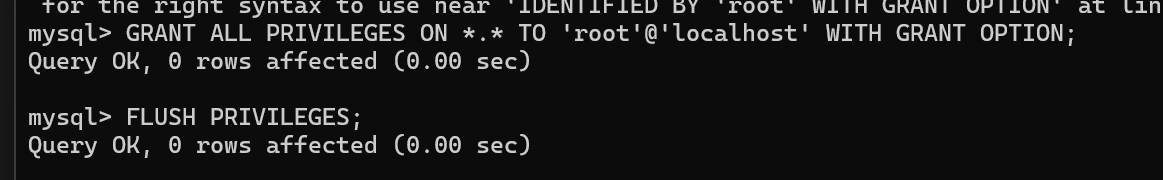
还是同样的报错。

仔细看报错原因，是access denied for user root

可能是因为root用户权限不够，要为root用户打开更高的权限

GRANT ALL PRIVILEGES ON \*.\* TO 'root'@'localhost' WITH GRANT OPTION;

输入之后刷新



后面发现是密码输错了

我的密码和助教的不一样

val jdbcDF=spark.read.format("jdbc").

option("url","jdbc:mysql://localhost:3306/spark").

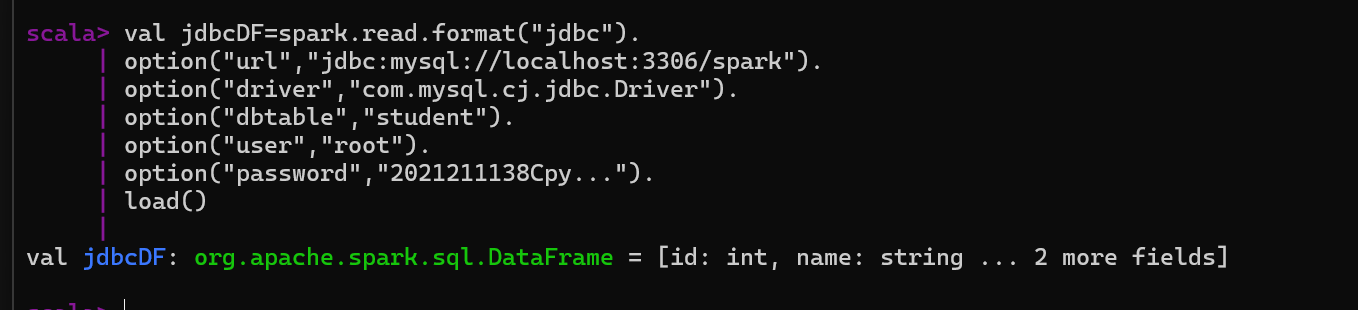
option("driver","com.mysql.cj.jdbc.Driver").

option("dbtable","student").

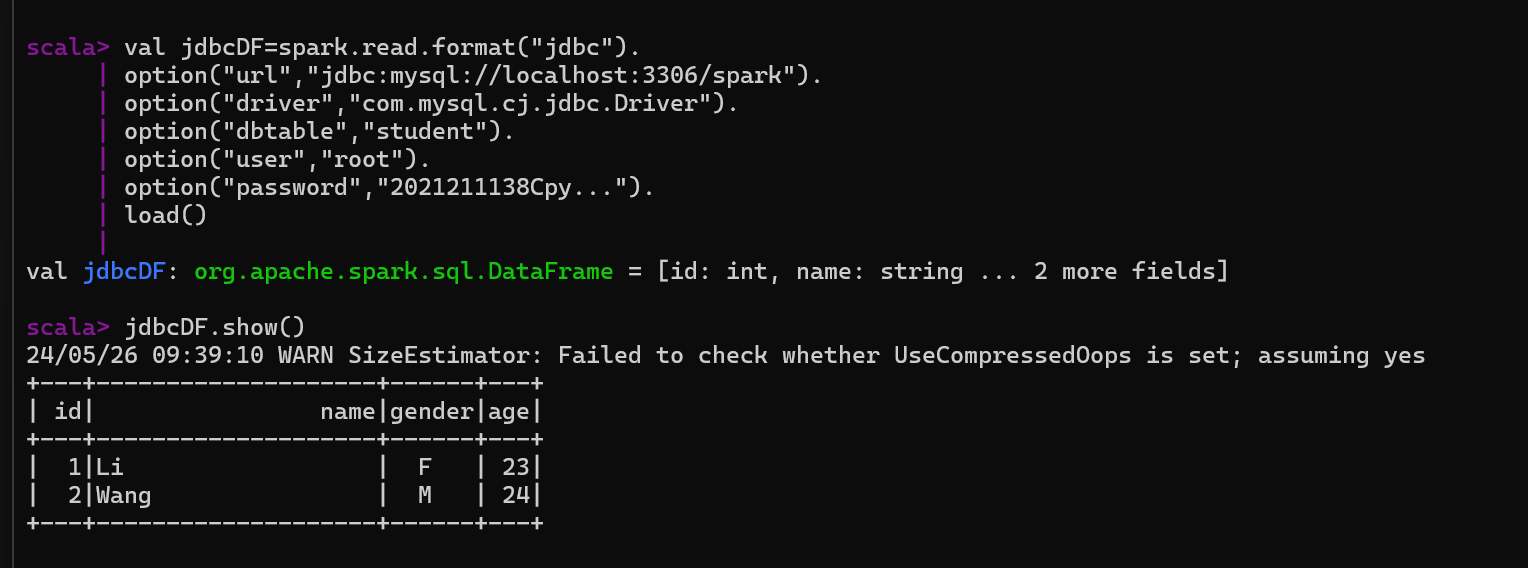
option("user","root").

option("password","2021211138Cpy...").

load()



查看表单：



## bug7

### 出错部分

[root@lxl-2021211146-0002 ~]# hadoop jar ./hadoop-2.7.7/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar pi 10 10

Not a valid JAR: /root/hadoop-2.7.7/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar

### 修改方式

修改目录为之前的安装目录/home/modules/hadoop-2.7.7/bin/hadoop

hadoop jar /home/modules/hadoop-2.7.7/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar pi 10 10

## bug8

### 出错部分

[root@lxl-2021211146-0002 mapreduce]# hadoop jar /home/modules/hadoop-2.7.7/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.7.jar pi 10 10

Number of Maps  = 10

Samples per Map = 10

24/05/27 13:19:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.server.namenode.SafeModeException): Cannot create directory /user/root/QuasiMonteCarlo\_1716787141457\_275444856/in. Name node is in safe mode.

### 修改方式

使用命令离开安全模式

hadoop dfsadmin -safemode get  # 查看安全模式状态

hadoop dfsadmin -safemode enter    # 进入安全模式状态

hadoop dfsadmin -safemode leave   # 离开安全模式

## bug9

### 出错部分

[root@lxl-2021211146-0002 download]# spark-submit --class org.example.ScalaDuplicateRemove --master yarn --num-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test-2.jar

java.lang.ClassNotFoundException: org.example.ScalaDuplicateRemove

        at java.lang.Class.forNameImpl(Native Method)

        at java.lang.Class.forName(Class.java:402)

        at org.apache.spark.util.Utils$.classForName(Utils.scala:229)

        at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:700)

        at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:187)

        at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:212)

        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:126)

        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)

### 修改方式

忘记删除Jar包的MetaInfo文件Manifest删掉之后就可以了

## bug10

### 出错部分

scala> val jdbcDF = spark.read.format("jdbc").

     | | option("url","jdbc:mysql://localhost:3306/spark").

     | | option("dbtable","student").

     | | option("user","root").

     | | option("password","").

     | | load()

java.sql.SQLException: No suitable driver

  at java.sql.DriverManager.getDriver(DriverManager.java:315)

  at org.apache.spark.sql.execution.datasources.jdbc.JDBCOptions$$anonfun$7.apply(JDBCOptions.scala:84)

  at org.apache.spark.sql.execution.datasources.jdbc.JDBCOptions$$anonfun$7.apply(JDBCOptions.scala:84)

  at scala.Option.getOrElse(Option.scala:121)

  at org.apache.spark.sql.execution.datasources.jdbc.JDBCOptions.<init>(JDBCOptions.scala:83)

  at org.apache.spark.sql.execution.datasources.jdbc.JDBCOptions.<init>(JDBCOptions.scala:34)

  at org.apache.spark.sql.execution.datasources.jdbc.JdbcRelationProvider.createRelation(JdbcRelationProvider.scala:32)

  at org.apache.spark.sql.execution.datasources.DataSource.resolveRelation(DataSource.scala:330)

  at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:152)

  at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:125)

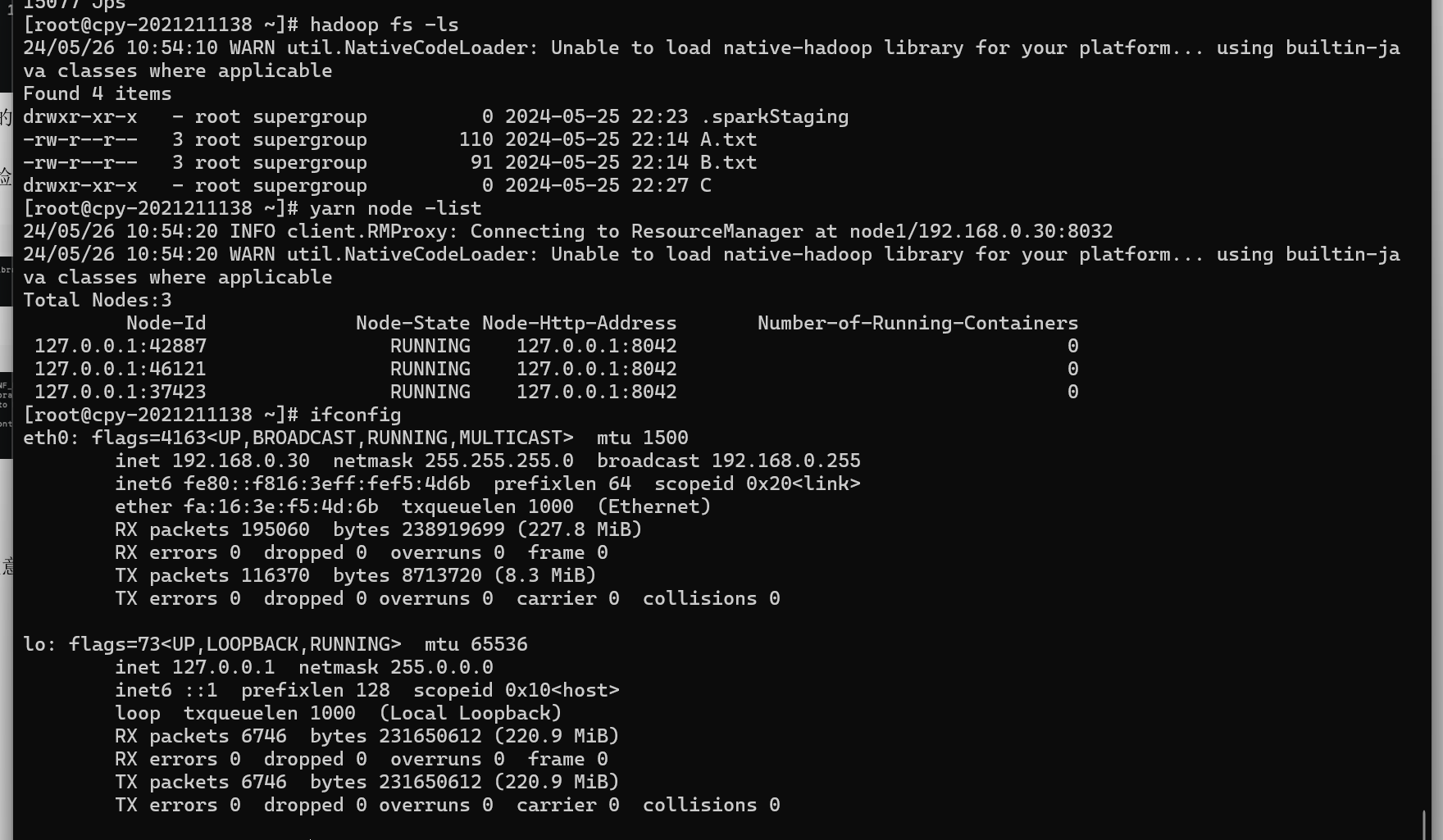
  ... 53 elided

### 修改方式

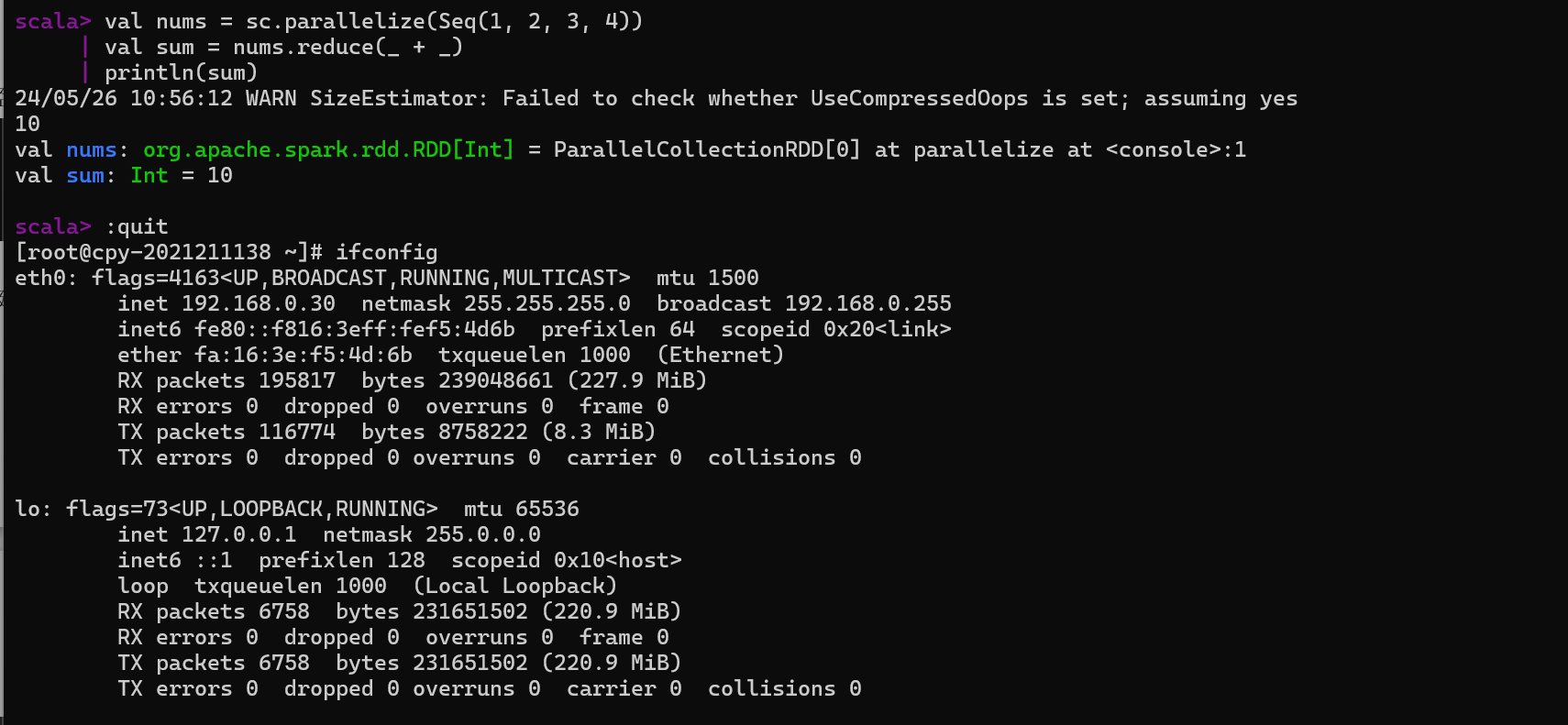
忘记配置相关数据库登录参数了。重新配置好数据库登录参数，就不再报错了。

# 2 实验结果截图

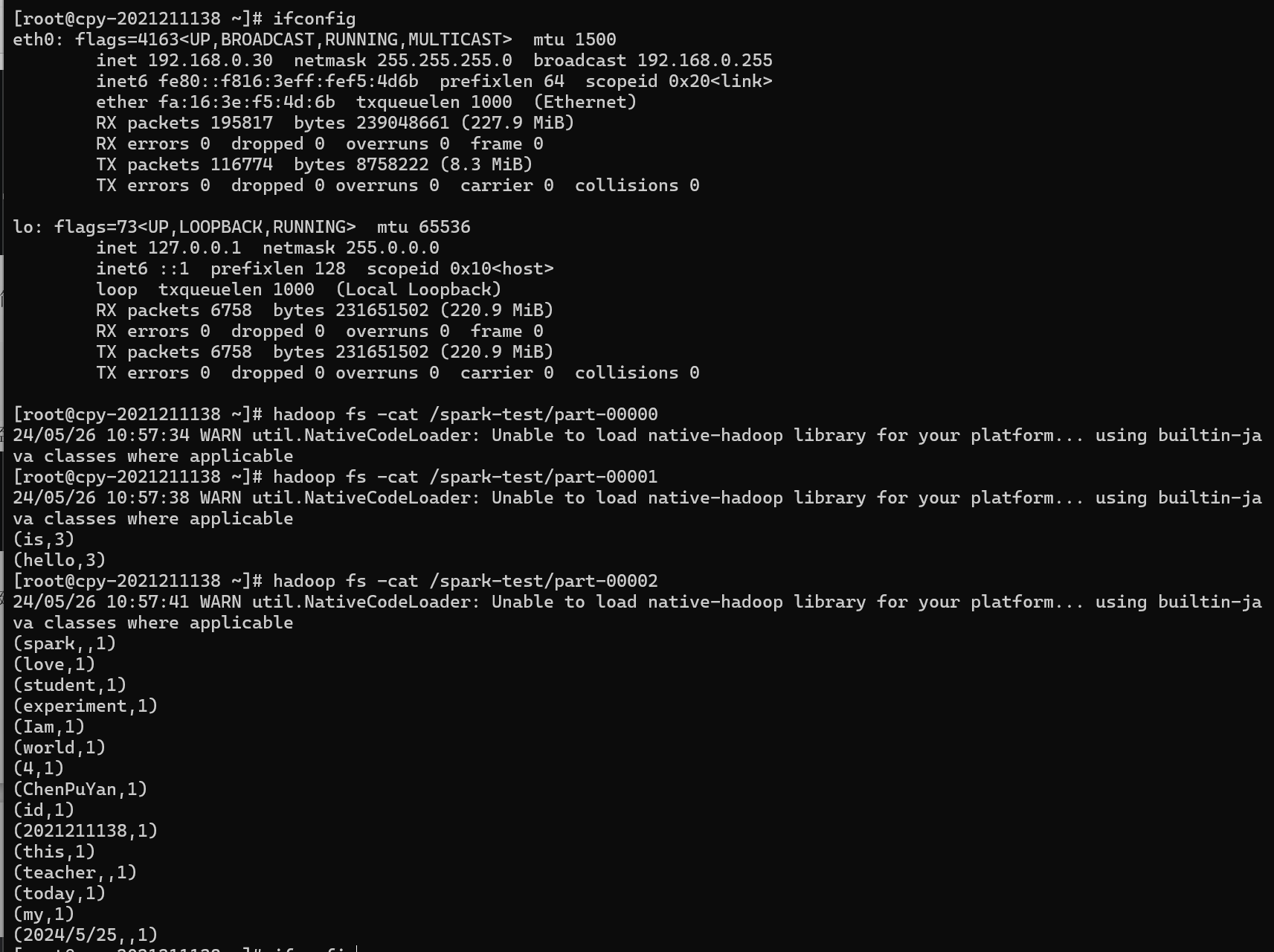
## 2.1 Hadoop集群测试结果



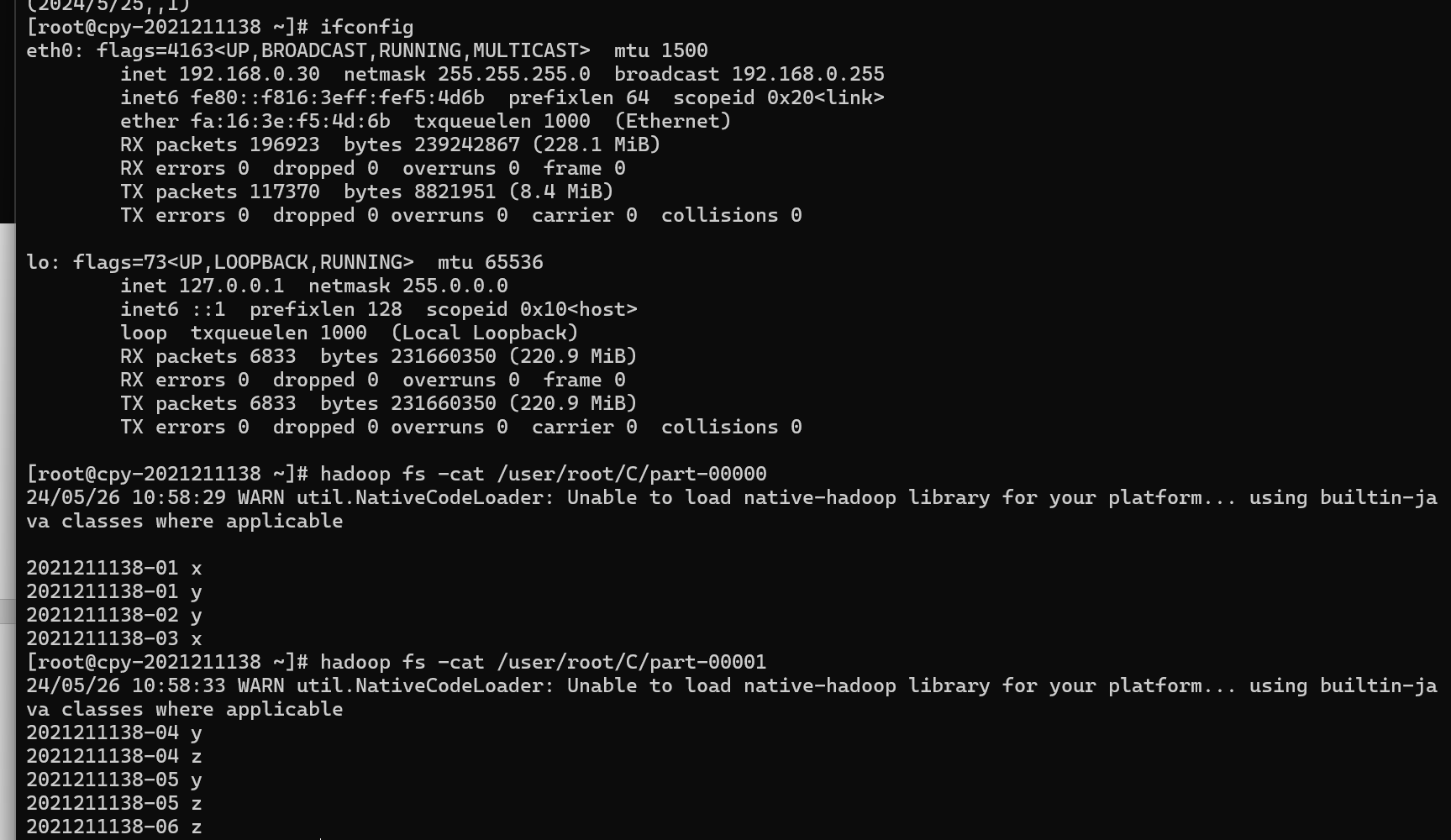
## 2.2 Spark集群搭建完成的测试结果



## 2.3 Scala单词计数实验结果



## 2.4 RDD编程结果



## 2.5 Spark sql读写数据库结果

