实验四 bug 报告

陈朴炎 2021211138

目录

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：

```
ie tasks for this job
24/05/25 20:35:55 INFO YarnScheduler: Killing all running tasks in stage 2: Stage finished
24/05/25 20:35:55 INFO DAGScheduler: Job 1 finished: collect at ScalaWordCount.scala:37, took 1.057801 s
(hello,1),(Iam,1),(ChenPuYan,1),(hello,1),(spark,,1),(my,1),(student,1),(id,1),(is,1),(2021211138,1),(hello,1),(teacher,,1),(this,1),(is,1),(experiment,1),(4,1),(today,1),(is,1),(2024/5/25,,1),(love,1),(world,1)24/05/25 20:35:55 WARN FileSystem: Failed to initialize fileystem hdfs://cpy-2021211138-0001:8020/spark-test: java.lang.IllegalArgumentException: java.net.UnknownHostException: cpy-2021211138-0001
Exception in thread "main" java.lang.IllegalArgumentException: java.net.UnknownHostException: cpy-2021211138-0001
        at org.apache.hadoop.security.SecurityUtil.buildTokenService(SecurityUtil.java:466)
        at org.apache.hadoop.hdfs.NameNodeProxiesClient.createProxyWithClientProtocol(NameNodeProxiesClient.java:134)
        at org.apache.hadoop.hdfs.DFSClient.<init>(DFSClient.java:374)
        at org.apache.hadoop.hdfs.DFSClient.<init>(DFSClient.java:308)
        at org.apache.hadoop.hdfs.DistributedFileSystem.initDFSClient(DistributedFileSystem.java:202)
        at org.apache.hadoop.hdfs.DistributedFileSystem.initialize(DistributedFileSystem.java:187)
        at org.apache.hadoop.fs.FileSystem.createFileSystem(FileSystem.java:3469)
        at org.apache.hadoop.fs.FileSystem.access$300(FileSystem.java:174)
        at org.apache.hadoop.fs.FileSystem$Cache.getInternal(FileSystem.java:3574)
        at org.apache.hadoop.fs.FileSystem$Cache.get(FileSystem.java:3521)
        at org.apache.hadoop.fs.FileSystem.get(FileSystem.java:540)
        at org.apache.hadoop.fs.Path.getFileSystem(Path.java:365)
        at org.apache.spark.internal.io.SparkHadoopWriterUtils$.createPathFromString(SparkHadoopWriterUtils.scala:87)
```

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：

```
::1     localhost       localhost.localdomain   localhost6      localhost6.localdomain6


# 127.0.0.1     localhost       localhost.localdomain   localhost4      localhost4.localdomain4
# 127.0.0.1     localhost       localhost
# 127.0.0.1     cpy-2021211138  cpy-2021211138

# 192.168.0.30   node1
# 120.46.149.118 node2
# 1.92.86.3      node3
# 120.46.87.42   node4

192.168.0.30    node1
192.168.0.213   node2
192.168.0.161   node3
192.168.0.135   node4

192.168.0.30    node1   node1
```

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

/home/modules/hadoop-2.7.7/etc/hadoop/core-site.xml – root@1.92.114.12 – 编辑器 – WinSCP

```
<property>
  <name>fs.obs.connection.maximum</name>
  <value>1000</value>
</property>
<property>
  <name>fs.defaultFS</name>
  <value>hdfs://node1:8020</value>
</property>
```

hdfs-site.xml 用的也是 node 的编号

/home/modules/hadoop-2.7.7/etc/hadoop/hdfs-site.xml – root@1.92.114.12 – 编辑器 – WinSCP

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
 <property>
   <name>dfs.replication</name>
   <value>3</value>
 </property>
 <property>
   <name>dfs.namenode.secondary.http-address</name>
   <value>node1:50090</value>
 </property>
 <property>
   <name>dfs.namenode.secondary.https-address</name>
   <value>node4:50091</value>
 </property>
```

yarn-site.xml 中也没有

找了半天，发现原来是代码错了：

```
val ret=wordAndOne.sortBy(kv=>kv._2, ascending = false)
print(ret.collect().mkString(","))
ret.saveAsTextFile( path = "hdfs://cpy-2021211138-0001:8020/spark-test")
sc.stop()
```

我应该在标蓝的这里写 node1 的

修改如下：

```
val ret=wordAndOne.sortBy(kv=>kv._2, ascending = false)
print(ret.collect().mkString(","))
ret.saveAsTextFile( path = "hdfs://node1:8020/spark-test")
sc.stop()
```

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

```
24/05/25 21:24:20 INFO DAGScheduler: Job 1 finished: collect at ScalaWordCount.scala:37, took 0.
983221 s
(hello,1),(spark,,1),(my,1),(student,1),(id,1),(is,1),(2021211138,1),(hello,1),(Iam,1),(ChenPuYa
n,1),(hello,1),(teacher,,1),(this,1),(is,1),(experiment,1),(4,1),(today,1),(is,1),(2024/5/25,,1)
,(love,1),(world,1)24/05/25 21:24:20 INFO deprecation: mapred.output.dir is deprecated. Instead,
 use mapreduce.output.fileoutputformat.outputdir
24/05/25 21:24:20 INFO HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoo
p.mapred.FileOutputCommitter
24/05/25 21:24:20 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
24/05/25 21:24:20 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders
under output directory:false, ignore cleanup failures: false
24/05/25 21:24:20 INFO BlockManagerInfo: Removed broadcast_2_piece0 on node1:45601 in memory (si
ze: 3.1 KiB, free: 434.4 MiB)
24/05/25 21:24:20 INFO BlockManagerInfo: Removed broadcast_2_piece0 on node3:37675 in memory (si
ze: 3.1 KiB, free: 434.4 MiB)
24/05/25 21:24:20 INFO BlockManagerInfo: Removed broadcast_2_piece0 on node2:39127 in memory (si
ze: 3.1 KiB, free: 434.4 MiB)
24/05/25 21:24:20 INFO SparkContext: Starting job: runJob at SparkHadoopWriter.scala:83
```

bug2

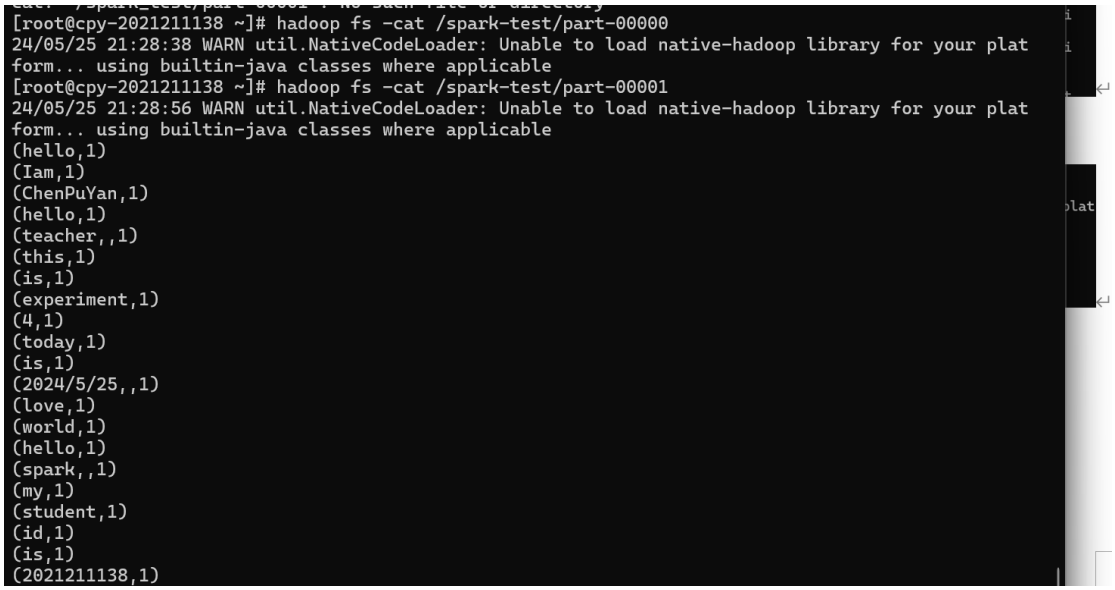
出错部分

在 hdfs 上查看程序输出：

```
drwxr-xr-x   - root supergroup          0 2024-05-25 20:23 /user
[root@cpy-2021211138 ~]# hadoop fs -ls /spark-test
24/05/25 21:26:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
Found 3 items
-rw-r--r--   3 root supergroup          0 2024-05-25 21:24 /spark-test/_SUCCESS
-rw-r--r--   3 root supergroup          0 2024-05-25 21:24 /spark-test/part-00000
-rw-r--r--   3 root supergroup        212 2024-05-25 21:24 /spark-test/part-00001
```

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：

```
val words:RDD[String]=lines.flatMap((line:String)=>{line.split( regex = " ")})

// 将RDD中的每一个单词转换为 kv 对，key是String 类型的单词，value 是 Int 类型的 1 并赋值给新的RDD对象wordAndOne
val wordAndOne:RDD[(String,Int)]=words.map((word:String)=>{(word,1)})
val wordAndNum:RDD[(String,Int)]=wordAndOne.reduceByKey((count1:Int,count2:Int)=>{count1+count2})
val ret=wordAndOne.sortBy(kv=>kv._2, ascending = false)
print(ret.collect().mkString(","))
ret.saveAsTextFile( path = "hdfs://node1:8020/spark-test")
sc.stop()
```

```
// 将RDD中的每一个单词转换为 kv 对，key是String 类型的单词，value 是 Int 类型的 1 并赋值给新的RDD对象wordAndOne
val wordAndOne:RDD[(String,Int)]=words.map((word:String)=>{(word,1)})
val wordAndNum:RDD[(String,Int)]=wordAndOne.reduceByKey((count1:Int,count2:Int)=>{count1+count2})
val ret=wordAndOne.sortBy(kv=>kv._2, ascending = false)
print(ret.collect().mkString(","))
ret.saveAsTextFile( path = "hdfs://node1:8020/spark-test")
sc.stop()
}
```

标蓝的地方写错了，应该是 wordAndNum 而不是 wordAndOne。

修改后，如下：

```
// 将RDD中的每一个单词转换为 kv 对，key是String 类型的单词，value 是 Int 类型的 1 并赋值给新的RDD对象wordAndOne
val wordAndOne:RDD[(String,Int)]=words.map((word:String)=>{(word,1)})
val wordAndNum:RDD[(String,Int)]=wordAndOne.reduceByKey((count1:Int,count2:Int)=>{count1+count2})
val ret=wordAndNum.sortBy(kv=>kv._2, ascending = false)
print(ret.collect().mkString(","))
ret.saveAsTextFile( path = "hdfs://node1:8020/spark-test")
sc.stop()
```

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：

```
[root@cpy-2021211138 ~]# hadoop fs -ls /spark-test
24/05/25 21:42:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   3 root supergroup          0 2024-05-25 21:42 /spark-test/_SUCCESS
-rw-r--r--   3 root supergroup          0 2024-05-25 21:42 /spark-test/part-00000
-rw-r--r--   3 root supergroup         17 2024-05-25 21:42 /spark-test/part-00001
-rw-r--r--   3 root supergroup        161 2024-05-25 21:42 /spark-test/part-00002
[root@cpy-2021211138 ~]#
```

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：

```
[root@cpy-2021211138 ~]# hadoop fs -cat /spark-test/part-00000
24/05/25 21:43:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
[root@cpy-2021211138 ~]# hadoop fs -cat /spark-test/part-00001
24/05/25 21:43:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
(is,3)
(hello,3)
[root@cpy-2021211138 ~]# hadoop fs -cat /spark-test/part-00002
24/05/25 21:43:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
(spark,,1)
(love,1)
(student,1)
(experiment,1)
(Iam,1)
(world,1)
(this,1)
(teacher,,1)
(today,1)
(my,1)
(2024/5/25,,1)
(4,1)
(ChenPuYan,1)
(id,1)
(2021211138,1)
```

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：

```
ze: 434.4 MiB)
24/05/25 22:17:05 INFO BlockManagerInfo: Removed broadcast_3_piece0 on node2:45635 in memory (size: 3.1 KiB, free: 434.4 MiB)
24/05/25 22:17:05 INFO BlockManagerInfo: Removed broadcast_3_piece0 on node3:41043 in memory (size: 3.1 KiB, free: 434.4 MiB)
24/05/25 22:17:05 INFO BlockManagerInfo: Removed broadcast_3_piece0 on node4:34631 in memory (size: 3.1 KiB, free: 434.4 MiB)
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://node1:8020/spark-test already exists
        at org.apache.hadoop.mapred.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:131)
        at org.apache.spark.internal.io.HadoopMapRedWriteConfigUtil.assertConf(SparkHadoopWriter.scala:299)
        at org.apache.spark.internal.io.SparkHadoopWriter$.write(SparkHadoopWriter.scala:71)
        at org.apache.spark.rdd.PairRDDFunctions.$anonfun$saveAsHadoopDataset$1(PairRDDFunctions.scala:1091)
        at org.apache.spark.rdd.PairRDDFunctions$$Lambda$1654/0x0000000000000000.apply$mcV$sp(Unknown Source)
        at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.scala:18)
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:112)
        at org.apache.spark.rdd.RDD.withScope(RDD.scala:410)
```

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

```
[root@cpy-2021211138 ~]# hadoop fs -cp /spark-test /spark-test1
24/05/25 22:20:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[root@cpy-2021211138 ~]# hadoop fs -ls /spark-test1
24/05/25 22:20:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   3 root supergroup          0 2024-05-25 22:20 /spark-test1/_SUCCESS
-rw-r--r--   3 root supergroup          0 2024-05-25 22:20 /spark-test1/part-00000
-rw-r--r--   3 root supergroup         17 2024-05-25 22:20 /spark-test1/part-00001
-rw-r--r--   3 root supergroup        161 2024-05-25 22:20 /spark-test1/part-00002
```

删除

```
[root@cpy-2021211138 ~]# hadoop fs -rm -r /spark-test
24/05/25 22:21:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
24/05/25 22:21:16 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval =
0 minutes, Emptier interval = 0 minutes.
Deleted /spark-test
```

bug4

出错部分

第二个编程实验中，发现结果错误了：

```
Error: Could not find or load main class fs
[root@cpy-2021211138 ~]# hadoop fs -ls /spark-test
24/05/25 22:23:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   3 root supergroup          0 2024-05-25 22:23 /spark-test/_SUCCESS
-rw-r--r--   3 root supergroup          0 2024-05-25 22:23 /spark-test/part-00000
-rw-r--r--   3 root supergroup         17 2024-05-25 22:23 /spark-test/part-00001
-rw-r--r--   3 root supergroup        161 2024-05-25 22:23 /spark-test/part-00002
[root@cpy-2021211138 ~]#
```

修改方式

问题在于，运行的命令输错了

```
Deleted /spark-test
[root@cpy-2021211138 ~]# spark-submit --class org.example.ScalaWordCount --master yarn --num-exe
cutors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test2.jar
24/05/25 22:22:42 INFO SparkContext: Running Spark version 3.5.1
24/05/25 22:22:42 INFO SparkContext: OS info Linux, 4.18.0-80.7.2.el7.aarch64, aarch64
24/05/25 22:22:42 INFO SparkContext: Java version 1.8.0_292-ea
24/05/25 22:22:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform.
... using builtin-java classes where applicable
```

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
```

```
[root@cpy-2021211138 ~]# spark-submit --class org.example.ScalaDuplicateRemove -master yarn --nu
m-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test2.jar
24/05/25 22:26:56 WARN DependencyUtils: Local jar /root/-master does not exist, skipping.
Error: Failed to load class org.example.ScalaDuplicateRemove.
24/05/25 22:26:56 INFO ShutdownHookManager: Shutdown hook called
24/05/25 22:26:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-21568385-be1e-4a9a-972
a-ee97a5292158
[root@cpy-2021211138 ~]# spark-submit --class org.example.ScalaDuplicateRemove --master yarn --n
um-executors 3 --driver-memory 1g --executor-memory 1g --executor-cores 1 spark-test2.jar
24/05/25 22:27:22 INFO SparkContext: Running Spark version 3.5.1
24/05/25 22:27:22 INFO SparkContext: OS info Linux, 4.18.0-80.7.2.el7.aarch64, aarch64
24/05/25 22:27:22 INFO SparkContext: Java version 1.8.0_292-ea
24/05/25 22:27:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform.
.. using builtin-java classes where applicable
24/05/25 22:27:22 INFO ResourceUtils: ==============================================================
====
24/05/25 22:27:22 INFO ResourceUtils: No custom resources configured for spark.driver.
24/05/25 22:27:22 INFO ResourceUtils: ==============================================================
====
24/05/25 22:27:22 INFO SparkContext: Submitted application: Scala Duplicate Remover
24/05/25 22:27:22 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map
(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, scri
pt: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cp
us -> name: cpus, amount: 1.0)
24/05/25 22:27:22 INFO ResourceProfile: Limiting resource is cpus at 1 tasks per executor
24/05/25 22:27:22 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/05/25 22:27:22 INFO SecurityManager: Changing view acls to: root
24/05/25 22:27:22 INFO SecurityManager: Changing modify acls to: root
24/05/25 22:27:22 INFO SecurityManager: Changing view acls groups to:
24/05/25 22:27:22 INFO SecurityManager: Changing modify acls groups to:
24/05/25 22:27:22 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabl
ed; users with view permissions: root; groups with view permissions: EMPTY; users with modify pe
rmissions: root; groups with modify permissions: EMPTY
24/05/25 22:27:23 INFO Utils: Successfully started service 'sparkDriver' on port 38743.
```

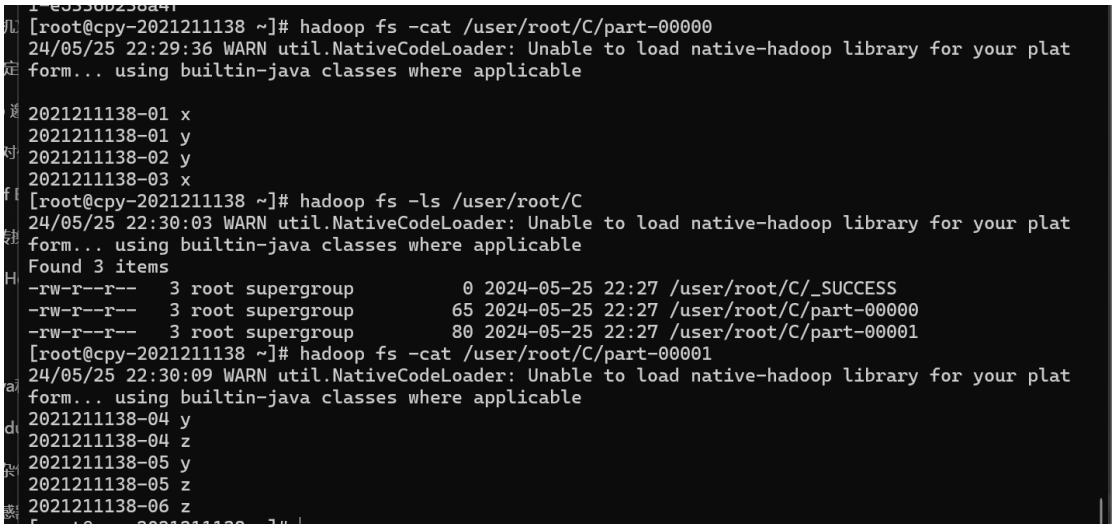
```
bleURLClassLoader@f3238d2b for default.
24/05/25 22:27:24 INFO Executor: Fetching spark://node1:38743/jars/spark-test2.jar with timestam
p 1716647242640
24/05/25 22:27:24 INFO TransportClientFactory: Successfully created connection to node1/192.168.
0.30:38743 after 46 ms (0 ms spent in bootstraps)
24/05/25 22:27:24 INFO Utils: Fetching spark://node1:38743/jars/spark-test2.jar to /tmp/spark-c0
e59554-096f-4e7b-b59e-3b530151ee54/userFiles-7f90388c-c745-4c56-92fd-e8962a323f2c/fetchFileTemp1
176633986625060616.tmp
24/05/25 22:27:25 INFO Executor: Adding file:/tmp/spark-c0e59554-096f-4e7b-b59e-3b530151ee54/user
rFiles-7f90388c-c745-4c56-92fd-e8962a323f2c/spark-test2.jar to class loader default
24/05/25 22:27:25 INFO Utils: Successfully started service 'org.apache.spark.network.netty.Netty
BlockTransferService' on port 38297.
24/05/25 22:27:25 INFO NettyBlockTransferService: Server created on node1:38297
24/05/25 22:27:25 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy
 for block replication policy
24/05/25 22:27:25 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, node1
, 38297, None)
```

```
24/05/25 22:27:28 INFO Executor: Finished task 1.0 in stage 3.0 (TID 5). 2030 bytes result sent
to driver
24/05/25 22:27:28 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 5) in 51 ms on node1
(executor driver) (2/2)
24/05/25 22:27:28 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, f
rom pool
24/05/25 22:27:28 INFO DAGScheduler: ShuffleMapStage 3 (sortBy at ScalaDuplicateRemove.scala:29)
 finished in 0.136 s
24/05/25 22:27:28 INFO DAGScheduler: looking for newly runnable stages
24/05/25 22:27:28 INFO DAGScheduler: running: HashSet()
24/05/25 22:27:28 INFO DAGScheduler: waiting: HashSet(ResultStage 4)
```

```
4/05/25 22:27:29 INFO SparkContext: SparkContext is stopping with exitCode 0.
4/05/25 22:27:29 INFO SparkUI: Stopped Spark web UI at http://node1:4040
4/05/25 22:27:29 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
4/05/25 22:27:29 INFO MemoryStore: MemoryStore cleared
4/05/25 22:27:29 INFO BlockManager: BlockManager stopped
4/05/25 22:27:29 INFO BlockManagerMaster: BlockManagerMaster stopped
4/05/25 22:27:29 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoor
inator stopped!
4/05/25 22:27:29 INFO SparkContext: Successfully stopped SparkContext
4/05/25 22:27:29 INFO ShutdownHookManager: Shutdown hook called
4/05/25 22:27:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-c0e59554-096f-4e7b-b59
-3b530151ee54
4/05/25 22:27:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-44036e60-5f31-4fbe-a96
-e5336b238a4f
root@cpy-2021211138 ~]#
```

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

```
  Verifying  : mysql-community-client-8.0.37-1.el7.aarch64                  4/
  Verifying  : mysql-community-libs-compat-8.0.37-1.el7.aarch64             5/
  Verifying  : 1:mariadb-libs-5.5.68-1.el7.aarch64                          6/

Installed:
  mysql-community-client.aarch64 0:8.0.37-1.el7
  mysql-community-libs.aarch64 0:8.0.37-1.el7
  mysql-community-libs-compat.aarch64 0:8.0.37-1.el7

Dependency Installed:
  mysql-community-client-plugins.aarch64 0:8.0.37-1.el7
  mysql-community-common.aarch64 0:8.0.37-1.el7

Replaced:
  mariadb-libs.aarch64 1:5.5.68-1.el7

Complete!
```

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

```
scala> 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","root").
     | load()
```

出现：

```
scala> val jdbcDF=spark.read.format("jdbc").
     | option("url","jdbc:mysql://127.0.0.1:3306/spark").
     | option("driver","com.mysql.cj.jdbc.Driver").
     | option("dbtable","student").
     | option("user","root").
     | option("password","root").
     | load()
java.sql.SQLException: Access denied for user 'root'@'localhost' (using password: YES)
  at com.mysql.cj.jdbc.exceptions.SQLError.createSQLException(SQLError.java:129)
  at com.mysql.cj.jdbc.exceptions.SQLExceptionsMapping.translateException(SQLExceptionsMapping.java:122)
  at com.mysql.cj.jdbc.ConnectionImpl.createNewIO(ConnectionImpl.java:833)
  ... 62 elided
```

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

```
 for the right syntax to use near 'IDENTIFIED BY 'root' WITH GRANT OPTION' at lin
mysql> GRANT ALL PRIVILEGES ON *.* TO 'root'@'localhost' WITH GRANT OPTION;
Query OK, 0 rows affected (0.00 sec)

mysql> FLUSH PRIVILEGES;
Query OK, 0 rows affected (0.00 sec)
```

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()
```

```
scala> 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()
     |
val jdbcDF: org.apache.spark.sql.DataFrame = [id: int, name: string ... 2 more fields]
```

查看表单：

```
scala> 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()
     |
val jdbcDF: org.apache.spark.sql.DataFrame = [id: int, name: string ... 2 more fields]

scala> jdbcDF.show()
24/05/26 09:39:10 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming yes
+---+--------------------+------+---+
| id|                name|gender|age|
+---+--------------------+------+---+
|  1|Li                  |  F   | 23|
|  2|Wang                |  M   | 24|
+---+--------------------+------+---+
```

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>(JDBC
Options.scala:83)
  at
org.apache.spark.sql.execution.datasources.jdbc.JDBCOptions.<init>(JDBC
Options.scala:34)
  at
org.apache.spark.sql.execution.datasources.jdbc.JdbcRelationProvider.cr
eateRelation(JdbcRelationProvider.scala:32)
  at
org.apache.spark.sql.execution.datasources.DataSource.resolveRelation(D
ataSource.scala:330)
  at
org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:152)
  at
org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:125)
```

```
  ... 53 elided
```

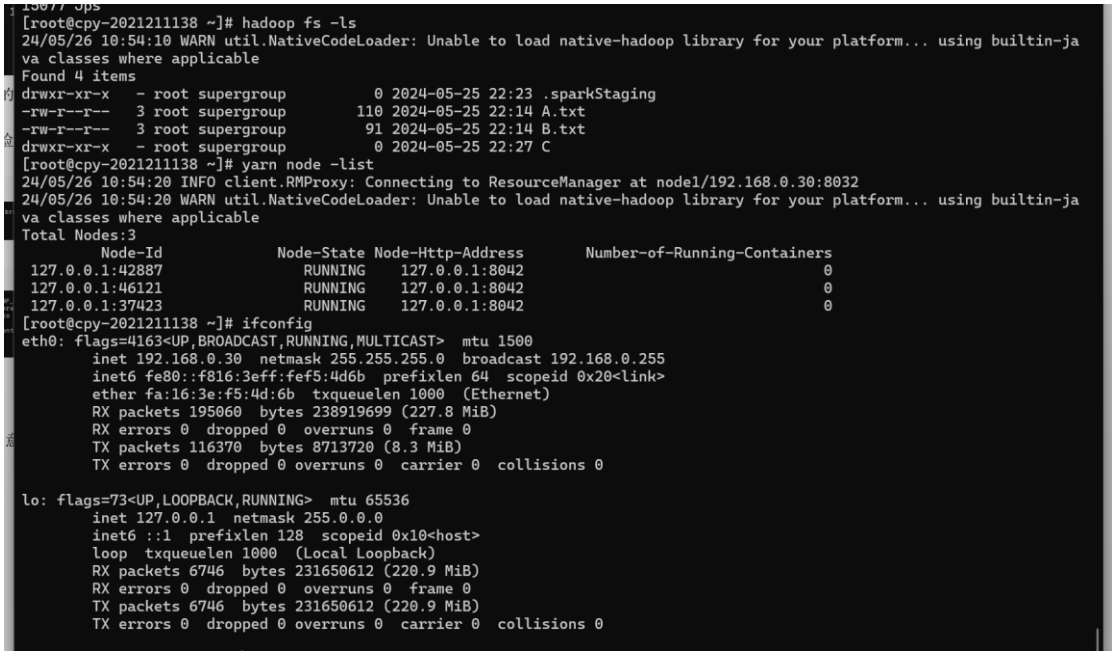
修改方式

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

```
scala> val nums = sc.parallelize(Seq(1, 2, 3, 4))
     | val sum = nums.reduce(_ + _)
     | println(sum)
24/05/26 10:56:12 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming yes
10
val nums: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:1
val sum: Int = 10

scala> :quit
[root@cpy-2021211138 ~]# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.0.30  netmask 255.255.255.0  broadcast 192.168.0.255
        inet6 fe80::f816:3eff:fef5:4d6b  prefixlen 64  scopeid 0x20<link>
        ether fa:16:3e:f5:4d:6b  txqueuelen 1000  (Ethernet)
        RX packets 195817  bytes 239048661 (227.9 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 116774  bytes 8758222 (8.3 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 6758  bytes 231651502 (220.9 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 6758  bytes 231651502 (220.9 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

2.3 Scala 单词计数实验结果

```
[root@cpy-2021211138 ~]# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.0.30  netmask 255.255.255.0  broadcast 192.168.0.255
        inet6 fe80::f816:3eff:fef5:4d6b  prefixlen 64  scopeid 0x20<link>
        ether fa:16:3e:f5:4d:6b  txqueuelen 1000  (Ethernet)
        RX packets 195817  bytes 239048661 (227.9 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 116774  bytes 8758222 (8.3 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 6758  bytes 231651502 (220.9 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 6758  bytes 231651502 (220.9 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

[root@cpy-2021211138 ~]# hadoop fs -cat /spark-test/part-00000
24/05/26 10:57:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
[root@cpy-2021211138 ~]# hadoop fs -cat /spark-test/part-00001
24/05/26 10:57:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
(is,3)
(hello,3)
[root@cpy-2021211138 ~]# hadoop fs -cat /spark-test/part-00002
24/05/26 10:57:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
(spark,,1)
(love,1)
(student,1)
(experiment,1)
(Iam,1)
(world,1)
(4,1)
(ChenPuYan,1)
(id,1)
(2021211138,1)
(this,1)
(teacher,,1)
(today,1)
(my,1)
(2024/5/25,,1)
```

2.4 RDD 编程结果

```
(2024/5/25,,1)
[root@cpy-2021211138 ~]# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.0.30  netmask 255.255.255.0  broadcast 192.168.0.255
        inet6 fe80::f816:3eff:fef5:4d6b  prefixlen 64  scopeid 0x20<link>
        ether fa:16:3e:f5:4d:6b  txqueuelen 1000  (Ethernet)
        RX packets 196923  bytes 239242867 (228.1 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 117370  bytes 8821951 (8.4 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 6833  bytes 231660350 (220.9 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 6833  bytes 231660350 (220.9 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

[root@cpy-2021211138 ~]# hadoop fs -cat /user/root/C/part-00000
24/05/26 10:58:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable

2021211138-01 x
2021211138-01 y
2021211138-02 y
2021211138-03 x
[root@cpy-2021211138 ~]# hadoop fs -cat /user/root/C/part-00001
24/05/26 10:58:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
2021211138-04 y
2021211138-04 z
2021211138-05 y
2021211138-05 z
2021211138-06 z
```

2.5 Spark sql 读写数据库结果

```
mysql> create database spark
    -> ;
Query OK, 1 row affected (0.03 sec)

mysql> use spark;
Database changed
mysql> create table student(id int(4), name char(20), gender char(4), age int(4));
Query OK, 0 rows affected, 2 warnings (0.03 sec)

mysql> insert into student values(1, 'Li', 'F', 23);
Query OK, 1 row affected (0.01 sec)

mysql> insert into student values(2, 'Wang', 'M', 24);
Query OK, 1 row affected (0.01 sec)

mysql> select * from student;
+------+------+--------+------+
| id   | name | gender | age  |
+------+------+--------+------+
|    1 | Li   | F      |   23 |
|    2 | Wang | M      |   24 |
+------+------+--------+------+
2 rows in set (0.00 sec)
```

```
scala> 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()
     |
val jdbcDF: org.apache.spark.sql.DataFrame = [id: int, name: string ... 2 more fields]

scala> jdbcDF.show()
24/05/26 09:39:10 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming yes
+---+--------------------+------+---+
| id|                name|gender|age|
+---+--------------------+------+---+
|  1|Li                  |  F   | 23|
|  2|Wang                |  M   | 24|
+---+--------------------+------+---+
```

```
24/05/26 10:01:56 INFO BlockManager: BlockManager stopped
24/05/26 10:01:56 INFO BlockManagerMaster: BlockManagerMaster stopped
24/05/26 10:01:56 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/05/26 10:01:56 INFO SparkContext: Successfully stopped SparkContext
24/05/26 10:01:56 INFO ShutdownHookManager: Shutdown hook called
24/05/26 10:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-6674309c-4979-4ceb-a220-8209fd233c32
24/05/26 10:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-67a67b45-ad1e-4078-9020-90976828d970
[root@cpy-2021211138 ~]# mysql -uroot -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 19
Server version: 8.0.37 MySQL Community Server - GPL

Copyright (c) 2000, 2024, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> use spark
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from student;
+------+------+--------+------+
| id   | name | gender | age  |
+------+------+--------+------+
|    1 | Li   | F      |   23 |
|    2 | Wang | M      |   24 |
|    3 | Chen | M      |   21 |
|    4 | liu  | M      |   27 |
+------+------+--------+------+
4 rows in set (0.00 sec)
```