

华东师范大学数据科学与工程学院实验报告

课程名称：分布式计算系统

年级：2022

上机实践成绩：

指导教师：徐辰

姓名：郭夏辉

学号：10211900416

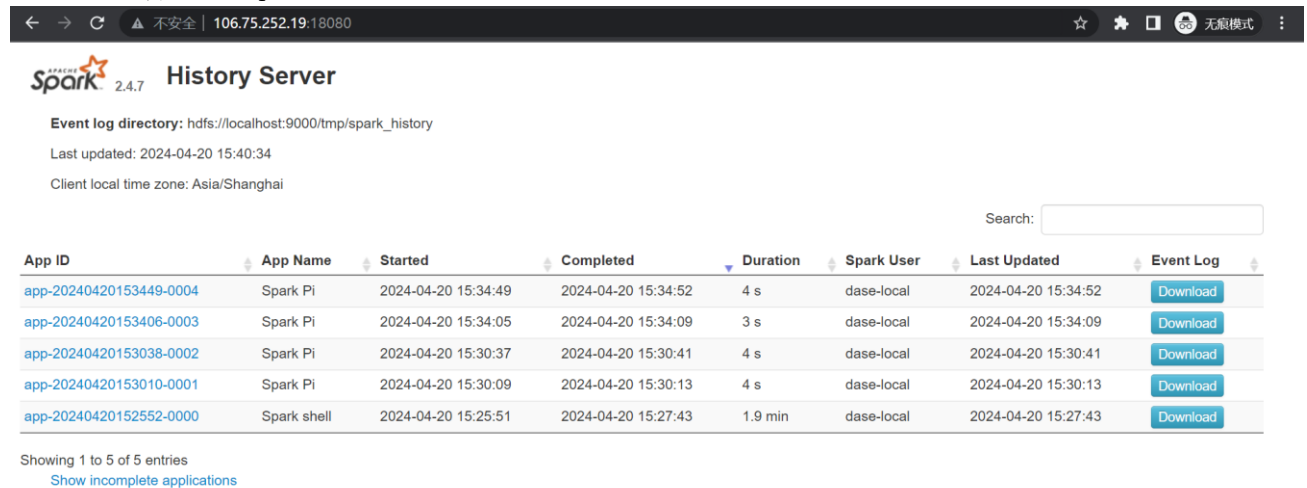
上机实践名称：Spark 部署

上机实践日期：

2024.5.7

问题 1：Spark 单机伪分布式实验时，无法在浏览器中看到 Spark 内部结点的运行情况。

我发现自己输入 `http://localhost:8080` 就看不到 Master 和 Worker。经过一段时间的排查，我发现自己的云主机压根没有开启 8080 这个端口。在 ucloud 的防火墙设置中我顺便开启了 4040 和 18080 这两个端口再次进行试验，也方便了之后的实验，最后我终于能在浏览器中看到了 Spark 的运行情况了。



The screenshot shows the Spark History Server interface at `http://106.75.252.19:18080`. The page title is "History Server" for Spark 2.4.7. It displays the event log directory as `hdfs://localhost:9000/tmp/spark_history` and the last update time as 2024-04-20 15:40:34. Below this is a table with columns: App ID, App Name, Started, Completed, Duration, Spark User, Last Updated, and Event Log. The table contains 5 entries for Spark Pi and Spark shell applications. Each entry has a "Download" button for the event log. At the bottom, it shows "Showing 1 to 5 of 5 entries" and a link to "Show incomplete applications".

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
app-20240420153449-0004	Spark Pi	2024-04-20 15:34:49	2024-04-20 15:34:52	4 s	dase-local	2024-04-20 15:34:52	Download
app-20240420153406-0003	Spark Pi	2024-04-20 15:34:05	2024-04-20 15:34:09	3 s	dase-local	2024-04-20 15:34:09	Download
app-20240420153038-0002	Spark Pi	2024-04-20 15:30:37	2024-04-20 15:30:41	4 s	dase-local	2024-04-20 15:30:41	Download
app-20240420153010-0001	Spark Pi	2024-04-20 15:30:09	2024-04-20 15:30:13	4 s	dase-local	2024-04-20 15:30:13	Download
app-20240420152552-0000	Spark shell	2024-04-20 15:25:51	2024-04-20 15:27:43	1.9 min	dase-local	2024-04-20 15:27:43	Download

问题 2：下载 spark 的源代码时候一直卡在 connecting 环节

经过一段时间的检查，我发现出现这个问题的原因是连接外网存在障碍。我现在本地下载好了之后再通过 scp 上传问题就解决了。

```
dase-local@10-23-114-123:~$ sudo wget https://archive.apache.org/dist/spark/spark-2.4.7/spark-2.4.7-bin-without-hadoop.tgz
--2024-04-20 14:50:25-- https://archive.apache.org/dist/spark/spark-2.4.7/spark-2.4.7-bin-without-hadoop.tgz
Resolving archive.apache.org (archive.apache.org) ... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443 ... ^C
```

问题 3：分布式实验中 HDFS 和 Spark 服务开启之后还是无法启动 Spark Shell, 提示 Connection Refused

经过一段时间的检查，配合着实验手册，我发现自己在进行分布式实验之前没有修改 Hadoop 和 Spark 配置，配置文件中的主机名还是伪分布式实验中的 `localhost` 而非 `ecnu01`。在我把里面的 `localhost` 全部改成 `ecnu01` 并重新将配置发送到另外三台主机之后，我再次开始实验，就成功地打开了 Spark Shell 并运行了 WordCount 程序。

问题 4: Spark 启动日志服务器时候报错 File does not exist

```
dase-dis@ecnu01:~/spark-2.4.7$ ./sbin/start-history-server.sh
starting org.apache.spark.deploy.history.HistoryServer, logging to /home/dase-dis/spark-2.4.7/logs/spark-das
e-dis-org.apache.spark.deploy.history.HistoryServer-1-ecnu01.out
failed to launch: nice -n 0 /home/dase-dis/spark-2.4.7/bin/spark-class org.apache.spark.deploy.history.Histo
ryServer
    at org.apache.spark.deploy.history.FsHistoryProvider.<init>(FsHistoryProvider.scala:207)
    at org.apache.spark.deploy.history.FsHistoryProvider.<init>(FsHistoryProvider.scala:86)
    ... 6 more
Caused by: java.io.FileNotFoundException: File does not exist: hdfs://ecnu01:9000/tmp/spark_history
    at org.apache.hadoop.hdfs.DistributedFileSystem$29.doCall(DistributedFileSystem.java:1528)
    at org.apache.hadoop.hdfs.DistributedFileSystem$29.doCall(DistributedFileSystem.java:1521)
    at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
    at org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(DistributedFileSystem.java:1521)
    at org.apache.spark.deploy.history.FsHistoryProvider.org$apache$spark$deploy$history$FsHistoryProvid
er$$startPolling(FsHistoryProvider.scala:257)
    ... 9 more
full log in /home/dase-dis/spark-2.4.7/logs/spark-dase-dis-org.apache.spark.deploy.history.HistoryServer-1-e
cnu01.out
dase-dis@ecnu01:~/spark-2.4.7$
```

解决这个问题很简单，我只需要在 HDFS 中新建一个对应的的目录即可（~/hadoop-2.10.1/bin/hdfs dfs -mkdir -p /tmp/spark_history）

华东师范大学数据科学与工程学院实验报告

课程名称：分布式计算系统

年级：2022

上机实践成绩：

指导教师：徐辰

姓名：郭夏辉

学号：10211900416

上机实践名称：Spark 编程

上机实践日期：

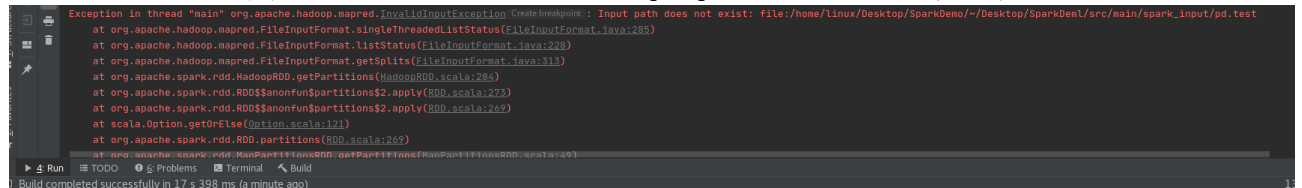
2024.5.14

问题 1: Error: JAVA_HOME is not set and could not be found

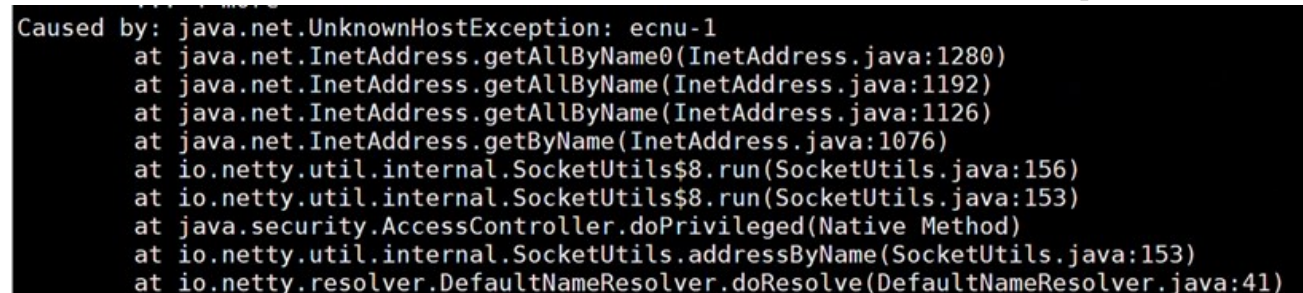
本来我以为自己出现了 lab2 时候的问题，即在/etc/hadoop/hadoop-env.sh 中没有将 JAVA_HOME 设置为绝对路径（export JAVA_HOME=/usr/local/jdk1.8）。但我后来发现自己早就已经设置了，那为什么会出现这样的问题呢？经过一段时间的排查，我发现自己不知道什么时候在 export 前面多加了一个空格，而 sh 文件又对空格敏感，所以就无法被系统识别。修改了这个之后问题就解决了。

问题 2: WordCount 本地调试时候程序并没有正常运行

在虚拟机中进行 Spark WordCount 程序的本地调试时，我发现程序没有正常运行。在查看了报错之后，我发现输入的路径和预期并不一致，就点击右上角的“WordCountJava”调整 configuration。我看到下面一栏已经设置了工作路径，就把输入路径设置为/src/main/spark_input/pd.test，然而程序还是没有正常运行。于是，我又在输入路径的前面加上了工作路径，即/home/linux/Desktop/SparkDemo，程序才在本地正常运行。



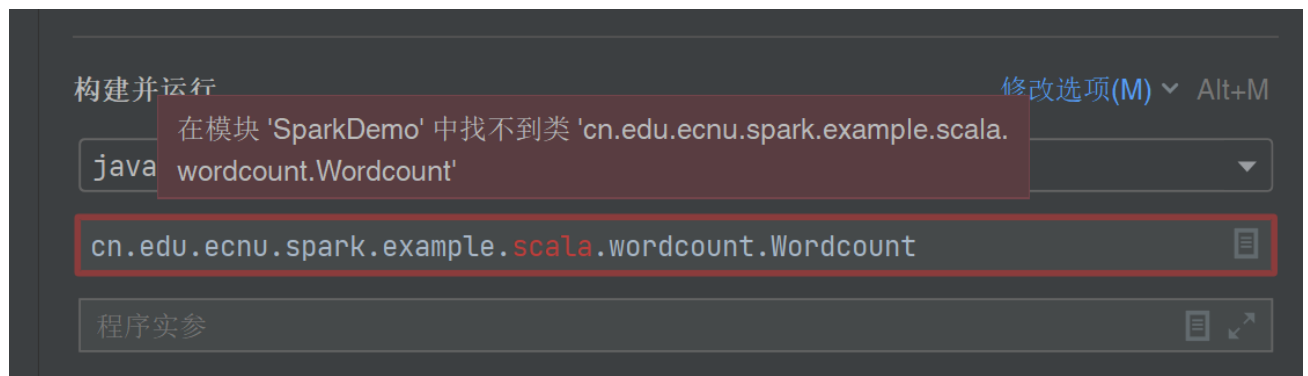
问题 3: 云主机上程序也并没有正常运行，报错 UnknownHostException



经过一段时间的检查，我发现出现这个问题的原因是我在输入命令时把一处 ecnu01 写成了 ecnu-1。在改正了拼写问题后，故障便解决了。

问题 4: IDEA 中如果没有标记 scala 目录为代码目录，则会报错

出现这个问题的原因还是 SparkDemo 的 idea 设置只把 java 给标记为了代码目录，而没有把 scala 标记，标记了之后问题就立刻解决了。



华东师范大学数据科学与工程学院实验报告

课程名称：分布式计算系统

年级：2022

上机实践成绩：

指导教师：徐辰

姓名：郭夏辉

学号：10211900416

上机实践名称：Spark+Yarn

上机实践日期：

2024.5.21

问题 1：单机伪分布式启动 Yarn 时候竟然出现了 ecnu02 和 ecnu03，提示 Connection timeout

出现这个问题的主要原因是自己登录错帐号了，竟然用 dase-dis 来做单机伪分布式的工作，这样的话因为是从镜像创建的云实例，默认还是分布式部署的，但是当时另外几台云主机并未打开，所以不可能运行成功的。我在检查之后重新用 dase-local 来登录，然后沿着实验手册的指导去做，Yarn 就成功启动了。

```
dase-dis@ecnu01:~$ vim ~/spark-2.4.7/conf/spark-env.sh
dase-dis@ecnu01:~$ vim ~/hadoop-2.10.1/etc/hadoop/yarn-site.xml
dase-dis@ecnu01:~$ ~/hadoop-2.10.1/sbin/start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/dase-dis/hadoop-2.10.1/logs/yarn-dase-dis-resourcemanager-ecnu01.out
ecnu03: ssh: connect to host ecnu03 port 22: Connection timed out
```

问题 2：分布式实验时 Spark 启动 historyserver 过程报错：File does not exist

我本来以为这个就像 lab5 里面类似问题那么好解决，但是经过一番排查发现延续之前实验的配置是错误的。

```
dase-dis@ecnu01:~$ spark-2.4.7/sbin/start-history-server.sh
starting org.apache.spark.deploy.history.HistoryServer, logging to /home/dase-dis/spark-2.4.7/logs/spark-dase-dis-org.apache.spark.deploy.history.HistoryServer-1-ecnu01.out
failed to launch: nice -n 0 /home/dase-dis/spark-2.4.7/bin/spark-class org.apache.spark.deploy.history.HistoryServer
    at org.apache.spark.deploy.history.FsHistoryProvider.<init>(FsHistoryProvider.scala:207)
    at org.apache.spark.deploy.history.FsHistoryProvider.<init>(FsHistoryProvider.scala:86)
    ... 6 more
Caused by: java.io.FileNotFoundException: File does not exist: hdfs://ecnu01:9000/tmp/spark_history
    at org.apache.hadoop.hdfs.DistributedFileSystem$29.doCall(DistributedFileSystem.java:1528)
    at org.apache.hadoop.hdfs.DistributedFileSystem$29.doCall(DistributedFileSystem.java:1521)
    at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
    at org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(DistributedFileSystem.java:1521)
    at org.apache.spark.deploy.history.FsHistoryProvider.org$apache$spark$deploy$history$FsHistoryProvider$$startPolling(FsHistoryProvider.scala:257)
    ... 9 more
full log in /home/dase-dis/spark-2.4.7/logs/spark-dase-dis-org.apache.spark.deploy.history.HistoryServer-1-ecnu01.out
```

采用 lab5 中的方法之后还是一样的报错：

```
dase-dis@ecnu01:~$ hadoop-2.10.1/bin/hdfs dfs -mkdir -p /user/dase-dis/tmp/spark_history
```

然后我细看了一下 hdfs://ecnu01:9000 这个地址，发现它的实际情况和预期有所差距：

```
dase-dis@ecnu01:~$ hadoop-2.10.1/bin/hdfs dfs -ls hdfs://ecnu01:9000
Found 1 items
drwxr-xr-x - dase-dis supergroup 0 2024-05-21 15:57 hdfs://ecnu01:9000/user/dase-dis/tmp
dase-dis@ecnu01:~$
```

可以看到 spark-defaults.conf 中配置的并不对，按如下方式修改：

```
# spark.executor.extraJavaOptions -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two"
spark.eventLog.enabled=true
spark.eventLog.dir=hdfs://ecnu01:9000/user/dase-dis/tmp/spark_history
spark.history.fs.logDirectory=hdfs://ecnu01:9000/user/dase-dis/tmp/spark_history
```


问题 3: [ERROR] Failed to construct terminal; falling back to unsupported java.lang.NumberFormatException: For input string: "0x100"

```
dase-dis@ecnu01:~$ spark-2.4.7/bin/spark-shell --master yarn
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
[ERROR] Failed to construct terminal; falling back to unsupported
java.lang.NumberFormatException: For input string: "0x100"
    at java.lang.NumberFormatException.forInputString(NumberFormatException.java:65)
    at java.lang.Integer.parseInt(Integer.java:580)
    at java.lang.Integer.valueOf(Integer.java:766)
    at scala.tools.jline_embedded.internal.InfoCmp.parseInfoCmp(InfoCmp.java:59)
    at scala.tools.jline_embedded.UnixTerminal.parseInfoCmp(UnixTerminal.java:242)
    at scala.tools.jline_embedded.UnixTerminal.<init>(UnixTerminal.java:65)
    at scala.tools.jline_embedded.UnixTerminal.<init>(UnixTerminal.java:50)
    at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
    at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
    at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:
45)
    at java.lang.reflect.Constructor.newInstance(Constructor.java:423)
    at java.lang.Class.newInstance(Class.java:442)
    at scala.tools.jline_embedded.TerminalFactory.getFlavor(TerminalFactory.java:211)
    at scala.tools.jline_embedded.TerminalFactory.create(TerminalFactory.java:102)
    at scala.tools.jline_embedded.TerminalFactory.get(TerminalFactory.java:186)
    at scala.tools.jline_embedded.TerminalFactory.get(TerminalFactory.java:192)
    at scala.tools.jline_embedded.console.ConsoleReader.<init>(ConsoleReader.java:243)
    at scala.tools.jline_embedded.console.ConsoleReader.<init>(ConsoleReader.java:235)
    at scala.tools.jline_embedded.console.ConsoleReader.<init>(ConsoleReader.java:223)
    at scala.tools.nsc.interpreter.jline_embedded.JLineConsoleReader.<init>(JLineReader.scala:64)
    at scala.tools.nsc.interpreter.jline_embedded.InteractiveReader.<init>(JLineReader.scala:33)
    at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
```

结合这个帖子 (<https://stackoverflow.com/questions/56457685/how-to-fix-the-sbt-crash-java-lang-numberformatexception-for-input-string-0x>) 中的解决方案, 我在 ~/.bashrc 中添加 export TERM=xterm-color, 然后 source ~/.bashrc 问题就解决了。