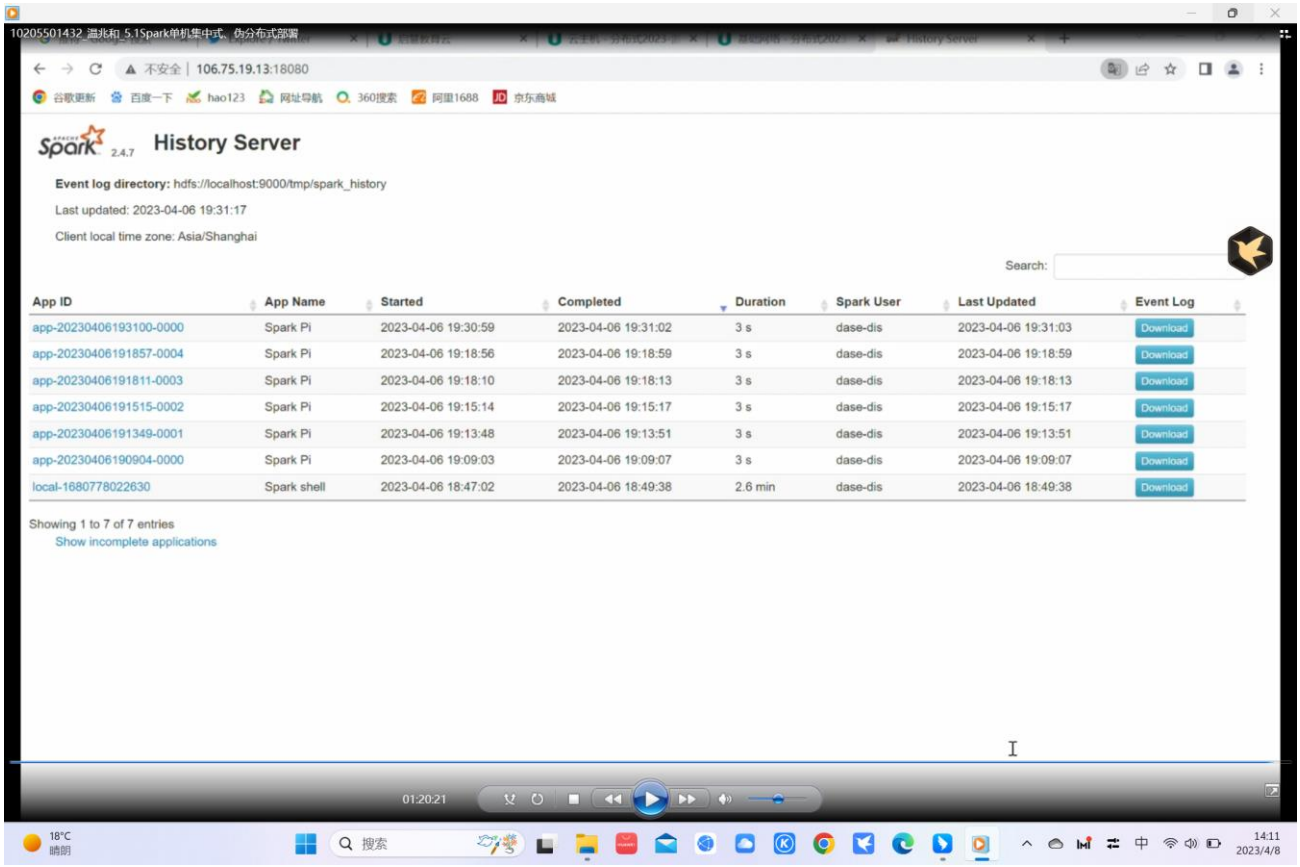
华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　　**年级：**2020　　**上机实践成绩：**

指导教师：徐辰　　**姓名：**温兆和　　**学号：**10205501432

上机实践名称：Spark 部署　　**上机实践日期：**2022.04.06

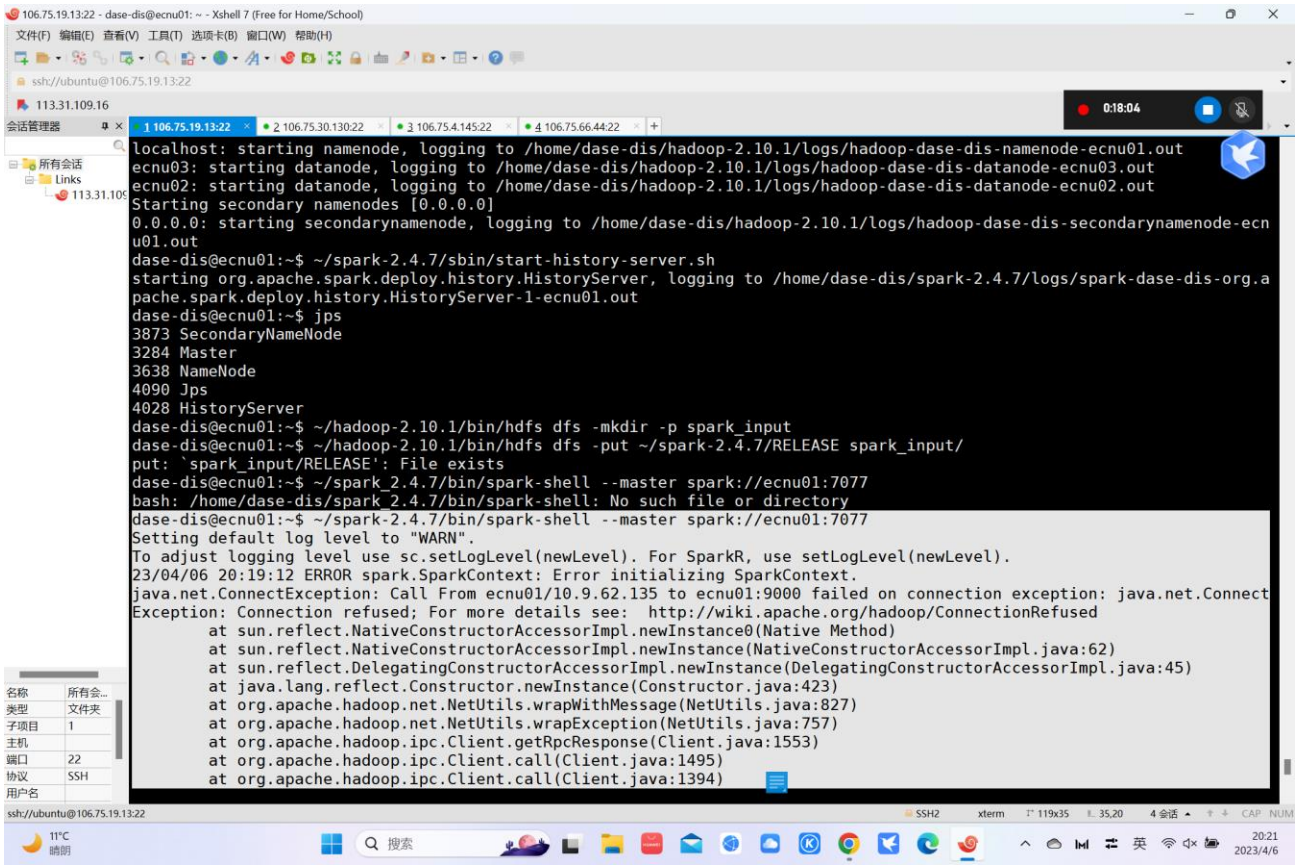
问题一



在进行 Spark 单机伪分布式部署实验的时候，我发现自己无法在网页上看到 Spark 内部节点的运行情况，比如输入 http://localhost:8080 就看不到 Master 和 Worker。经过助教的指导，我知道我必须自己在云主机上添加 8080 这个端口，并且确保主机的名称是 localhost。于是，我修改了云主机的安全模式->防火墙，在里面添加了本次实验要用到的 8080、4040 和 18080 端口并再次进行实验，终于在浏览器看到了 Spark 的运行情况，如上图所示。助教还告诉我，以后在进行试验前就要把所有需要用到的端口号添加到云主机中。

问题二

在进行 Spark 的多主机分布式部署实验时，我发现自己虽然实现了多台主机之间的免密钥登录并打开了 HDFS 和 Spark 服务，但是无法启动 Spark Shell。我为此询问了两位助教，但一直没有解决。后来，在仔细阅读实验手册后，我发现我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 而不是 ecnu01。于是，我仔细检查了所有的 Hadoop 和 Spark 配置，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并再次开始实验，就成功地打开了 Spark Shell 并运行了 WordCount 程序。这让我加深了对 Hadoop 和 Spark 环境配置的中的内容的理解。


106.75.19.13:22 - dase-dis@ecnu01: ~ - Xshell 7 (Free for Home/School)
localhost: starting namenode, logging to /home/dase-dis/hadoop-2.10.1/logs/hadoop-dase-dis-namenode-ecnu01.out
ecnu03: starting datanode, logging to /home/dase-dis/hadoop-2.10.1/logs/hadoop-dase-dis-datanode-ecnu03.out
ecnu02: starting datanode, logging to /home/dase-dis/hadoop-2.10.1/logs/hadoop-dase-dis-datanode-ecnu02.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/dase-dis/hadoop-2.10.1/logs/hadoop-dase-dis-secondarynamenode-ecnu01.out
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
dase-dis@ecnu01:~$ jps
3873 SecondaryNameNode
3284 Master
3638 NameNode
4090 Jps
4028 HistoryServer
dase-dis@ecnu01:~$ ~/hadoop-2.10.1/bin/hdfs dfs -mkdir -p spark_input
dase-dis@ecnu01:~$ ~/hadoop-2.10.1/bin/hdfs dfs -put ~/spark-2.4.7/RELEASE spark_input/
put: `spark_input/RELEASE': File exists
dase-dis@ecnu01:~$ ~/spark_2.4.7/bin/spark-shell --master spark://ecnu01:7077
bash: /home/dase-dis/spark_2.4.7/bin/spark-shell: No such file or directory
dase-dis@ecnu01:~$ ~/spark-2.4.7/bin/spark-shell --master spark://ecnu01:7077
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/04/06 20:19:12 ERROR spark.SparkContext: Error initializing SparkContext.
java.net.ConnectException: Call From ecnu01/10.9.62.135 to ecnu01:9000 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
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 org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:827)
at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:757)
at org.apache.hadoop.ipc.Client.getRpcResponse(Client.java:1553)
at org.apache.hadoop.ipc.Client.call(Client.java:1495)
at org.apache.hadoop.ipc.Client.call(Client.java:1394)

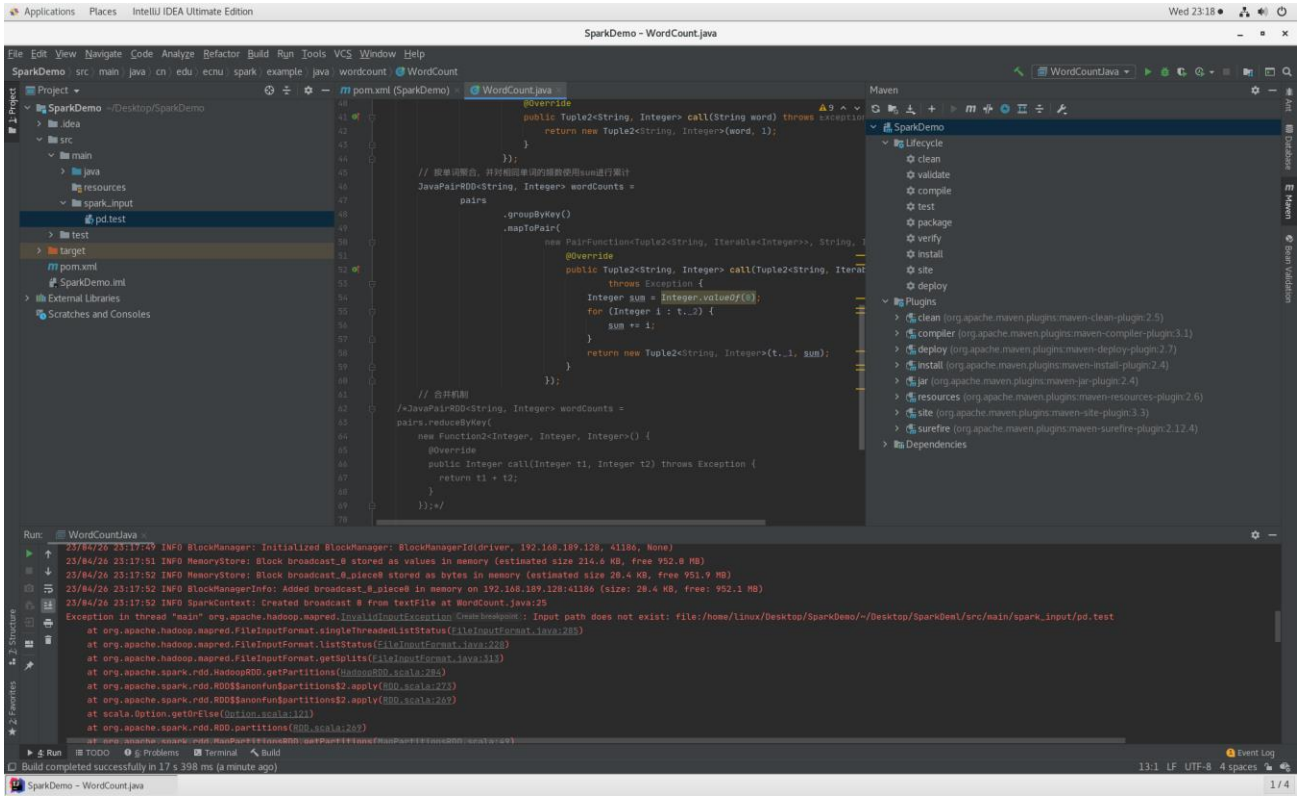
华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　　**年级：**2020　　**上机实践成绩：**

指导教师：徐辰　　**姓名：**温兆和　　**学号：**10205501432

上机实践名称：Spark 编程　　**上机实践日期：**2022.04.27

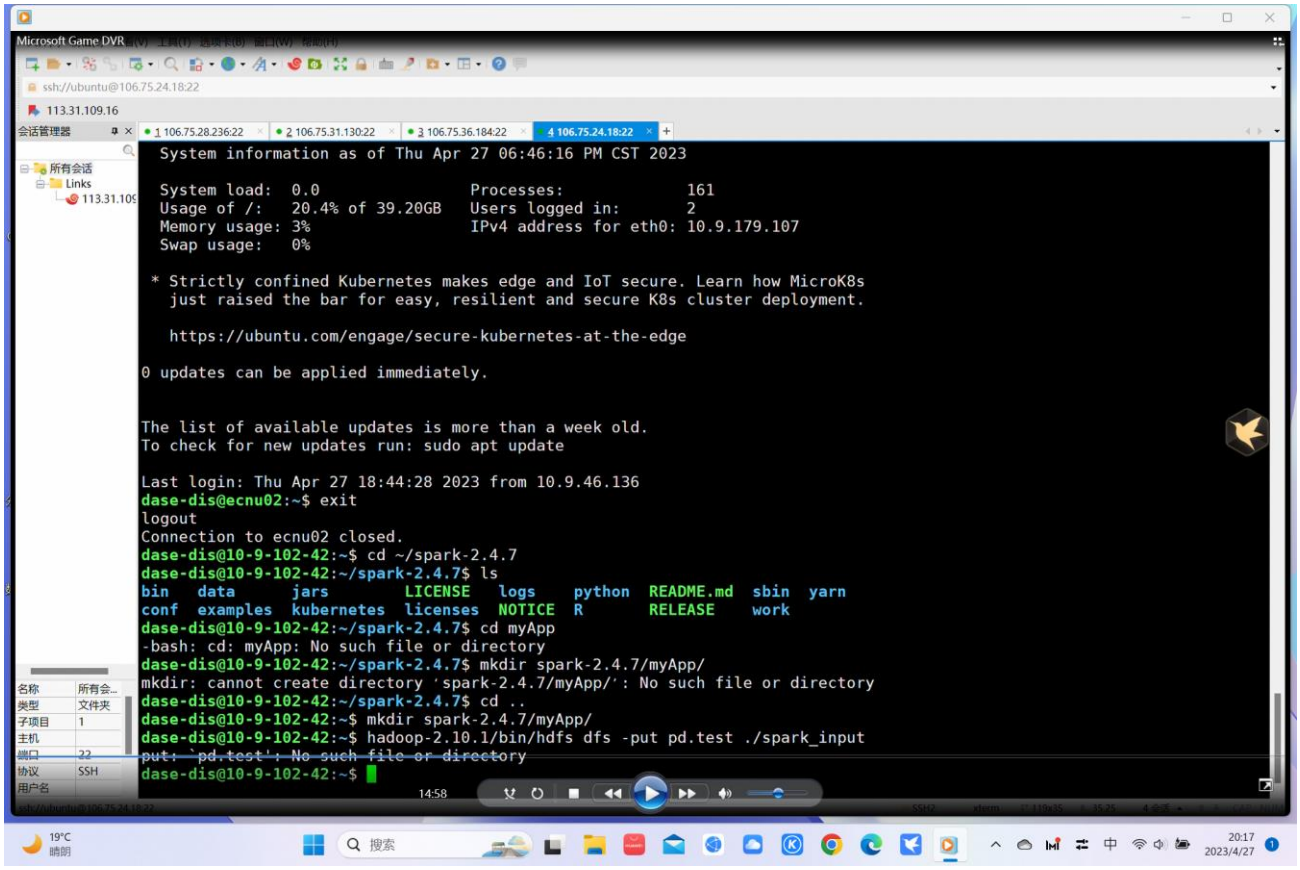
问题一



在 Linux 虚拟机上进行 Spark WordCount 程序的本地调试时，我发现程序没有正常运行。在仔细查看了报错之后，我发现输入的路径和自己预想的并不一样，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，但是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。

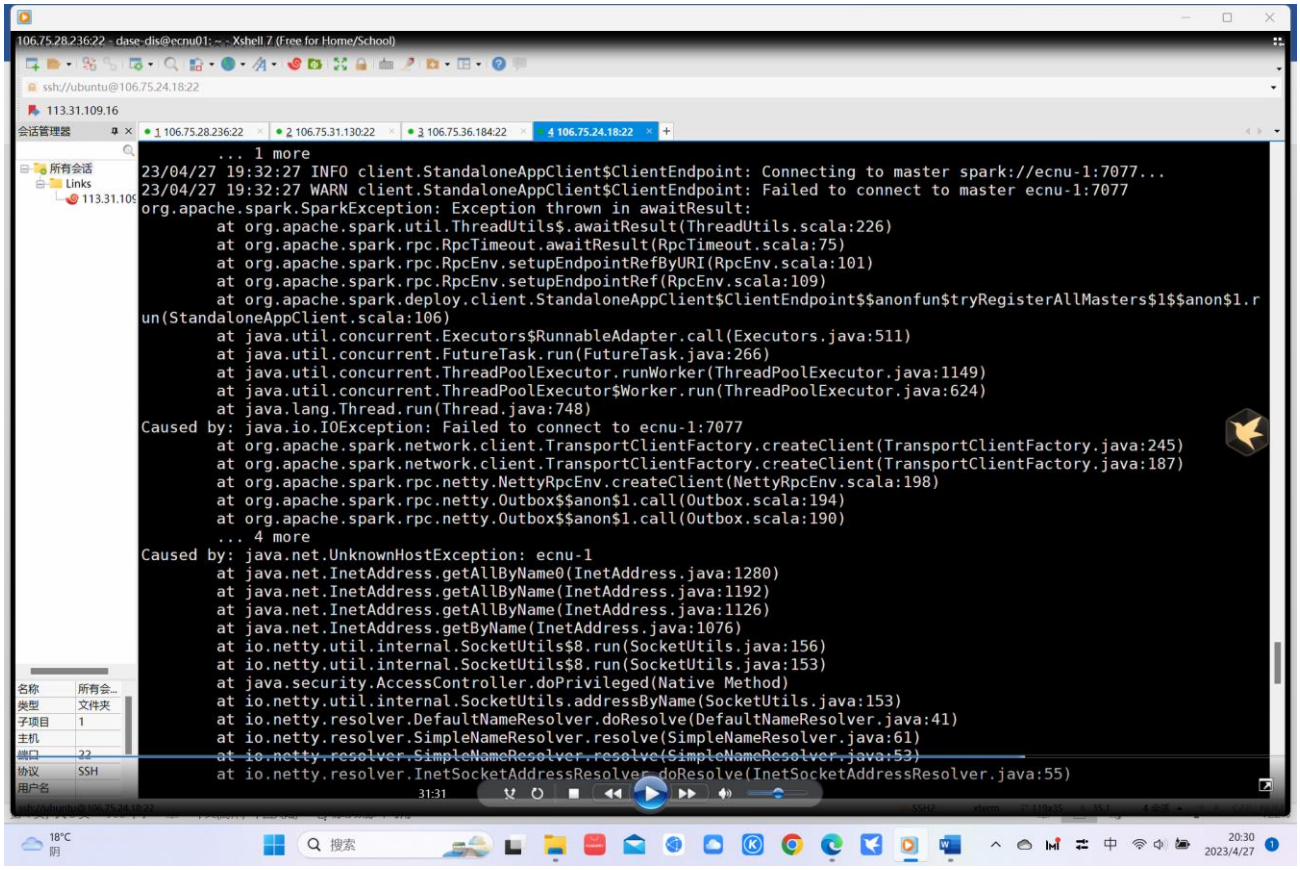
问题二

在将本地代码打包成 jar 包提交到云主机上进行分布式环境下的运行时，我想要将输入文件 pd.test 移动到./spark_input 下面，但是云主机上并没有这个文件。实验手册上说这个文件已经提交到 hdfs:///user/dase-dis/input，但我并没有找到这个路径。最后，在助教的建议下，我在本地找了一个文件，命名为 pd.test 并提交到云主机上。



问题三

在云主机上，我的代码还是没有正常运行。在仔细查看了报错后，我发现这是因为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，故障排除。



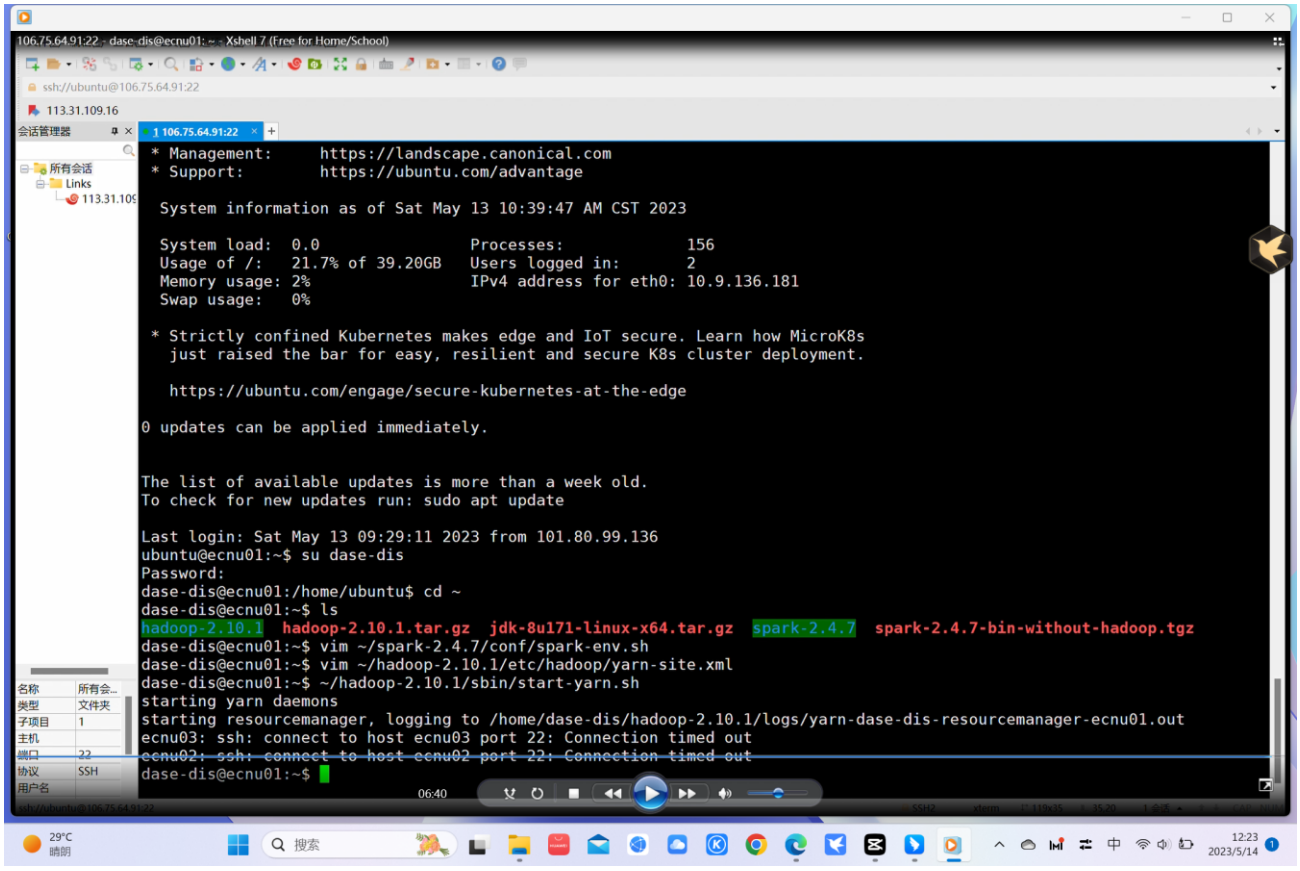
华东师范大学数据科学与工程学院实验报告

课程名称：分布式编程模型与系统　　**年级：**2020　　**上机实践成绩：**

指导教师：徐辰　　**姓名：**温兆和　　**学号：**10205501432

上机实践名称：基于 Yarn 部署 Spark　　**上机实践日期：**2022.05.11

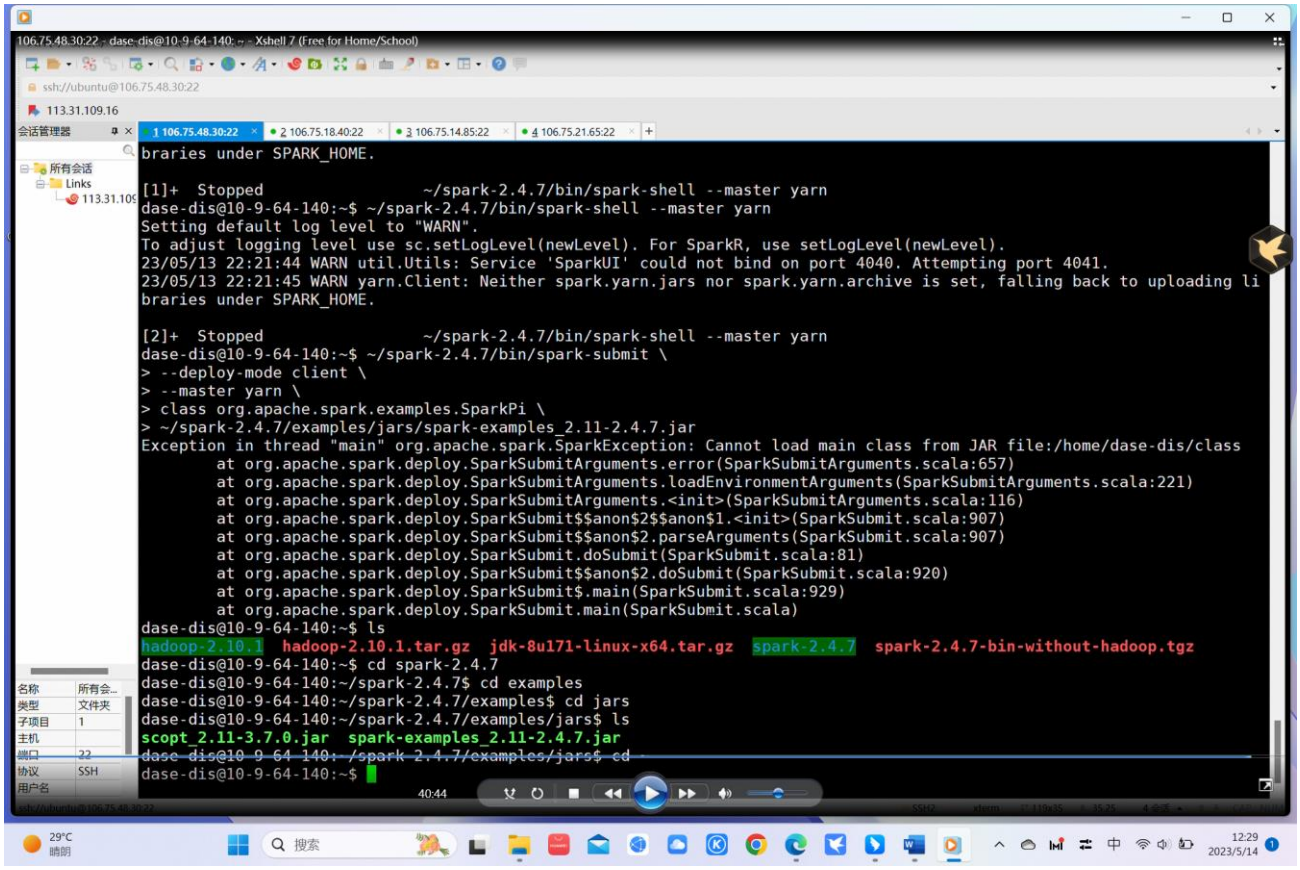
问题一



在进行 Spark+Yarn 单机伪分布式部署实验时，我发现无法启动 Yarn 服务。在仔细查看报错后，我发现我使用的是之前分布式部署的镜像，所以这里也默认我是分布式部署的，但是当时另外几台云主机并未打开。所以，我重新打开了一台虚拟机，从头开始配置 Java 环境、Spark 和 Hadoop，完成了单机伪分布式部署实验。

问题二

在进行 Spark+Yarn 分布式部署实验时，我用此前分布式实验后生成的镜像生成了四台云主机并进行试验，但是无法提交 jar 包。最终，我重新打开了四台虚拟机，从头开始设置单机和多机免密钥登录，配置 Java 环境、Spark 和 Hadoop，完成了单机分布式部署实验。


106.75.48.30:22 - dase-dis@10-9-64-140: ~ - Xshell 7 (Free for Home/School)
ssh://ubuntu@106.75.48.30:22
113.31.109.16
会话管理器
1 106.75.48.30:22
2 106.75.18.40:22
3 106.75.14.85:22
4 106.75.21.65:22
braries under SPARK_HOME.
[1]+ Stopped ~/spark-2.4.7/bin/spark-shell --master yarn
dase-dis@10-9-64-140:~$ ~/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).
23/05/13 22:21:44 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
23/05/13 22:21:45 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading li
braries under SPARK_HOME.
[2]+ Stopped ~/spark-2.4.7/bin/spark-shell --master yarn
dase-dis@10-9-64-140:~$ ~/spark-2.4.7/bin/spark-submit \
> --deploy-mode client \
> --master yarn \
> class org.apache.spark.examples.SparkPi \
> ~/spark-2.4.7/examples/jars/spark-examples_2.11-2.4.7.jar
Exception in thread "main" org.apache.spark.SparkException: Cannot load main class from JAR file:/home/dase-dis/class
at org.apache.spark.deploy.SparkSubmitArguments.error(SparkSubmitArguments.scala:657)
at org.apache.spark.deploy.SparkSubmitArguments.loadEnvironmentArguments(SparkSubmitArguments.scala:221)
at org.apache.spark.deploy.SparkSubmitArguments.<init>(SparkSubmitArguments.scala:116)
at org.apache.spark.deploy.SparkSubmit$$anon$2$$anon$1.<init>(SparkSubmit.scala:907)
at org.apache.spark.deploy.SparkSubmit$$anon$2.parseArguments(SparkSubmit.scala:907)
at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:81)
at org.apache.spark.deploy.SparkSubmit$$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)
dase-dis@10-9-64-140:~$ ls
hadoop-2.10.1 hadoop-2.10.1.tar.gz jdk-8u171-linux-x64.tar.gz spark-2.4.7 spark-2.4.7-bin-without-hadoop.tgz
dase-dis@10-9-64-140:~$ cd spark-2.4.7
dase-dis@10-9-64-140:~/spark-2.4.7$ cd examples
dase-dis@10-9-64-140:~/spark-2.4.7/examples$ cd jars
dase-dis@10-9-64-140:~/spark-2.4.7/examples/jars$ ls
scopt_2.11-3.7.0.jar spark-examples_2.11-2.4.7.jar
dase-dis@10-9-64-140:~/spark-2.4.7/examples/jars$ cd ~
dase-dis@10-9-64-140:~$
名称 所有会...
类型 文件夹
子项目 1
主机
端口 22
协议 SSH
用户名
40:44
29°C 晴朗
搜索
12:29 2023/5/14