

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：2021 上机实践成绩：
指导教师：徐辰 姓名：杨茜雅 学号：10215501435
上机实践名称：准备工作 上机实践日期：2024.3.6

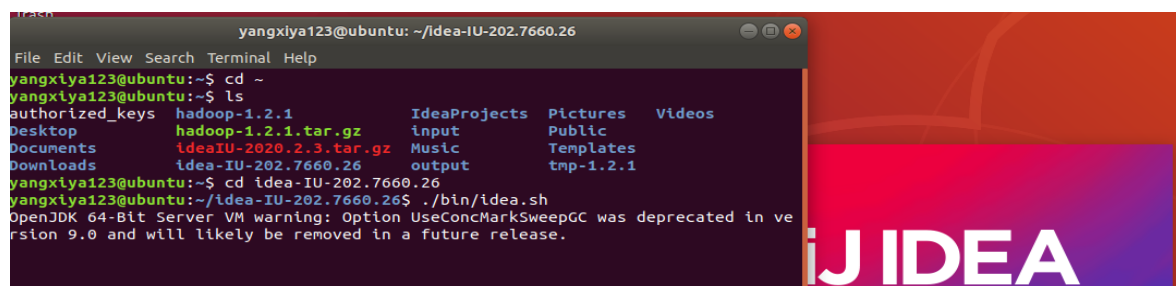
实验一：准备工作

- 1、在准备过程 lab1 中，在本地的 Ubuntu 虚拟机中装 idea 时，运行 ./bin/idea.sh 后并没有进入 IDEA 安装指导程序

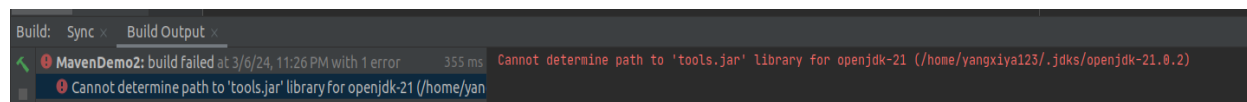
```
hadoop-1.2.1 input templates
yangxiya123@ubuntu:~$ cd idea-IU-202.7660.26/
yangxiya123@ubuntu:~/idea-IU-202.7660.26$ ./bin/idea.sh
OpenJDK 64-Bit Server VM warning: Option UseConcMarkSweepGC was deprecated in version 9.0 and will likely be removed in a future release.

Startup Error: Unable to detect graphics environment
yangxiya123@ubuntu:~/idea-IU-202.7660.26$
```

最后在 <http://t.csdnimg.cn/PGdyQ> 找到了解决方法。问题出现的原因是上一步我在设置免密登录时，在结尾运行了 ssh localhost 指令来看有没有配置成功。此时我的电脑通过 SSH 登录到了另一台机器，而 xshell 不支持图形显示。所以，**不忘妄图通过 SSH 去显示图形界面。如果要使用软件的图形界面，最直接的就是直接在 linux 上去操作，而不是通过 SSH（如 Xshell 等）。解决方法：retart 虚拟机，不要 ssh localhost。**



- 2、运行 run HelloWorld 时，如果出现 build 错误，可能是 language level 和 jdk 版本不适配的问题。对于使用 openjdk-21 的同学，可以考虑修改 project structure 的 language level，注意 Modules 和 Project 中的 language level 版本都要修改并且保持一致。但是这个方法在我的电脑上不 work，我的解决办法是把 jdk 换成 1.8 java version 即可，同时我还注意到 project structure 中 Modules 和 Project 中的 language level 版本并不一样，但是这也不影响运行。



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：2021

上机实践成绩：

指导教师：徐辰

姓名：杨茜雅

学号：10215501435

上机实践名称：Hadoop 1.x 部署

上机实践日期：

2024.3.13

实验二：Hadoop 1.x 部署

非常顺利的第一次实验，全程没遇到任何问题，实验手册上该有的输出都有，按流程一步一步来即可。下面记录一下同学们遇到的一些问题及解决措施，以免以后遇到同样的问题。

1、没有出现 FsShell 并不是实验步骤出问题了，而是文件上传速度太快，该过程已经结束，另起终端 jps 运行晚了。

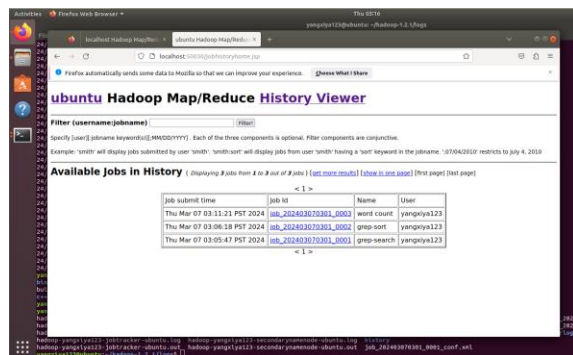
```
21076 Jps
14437 SecondaryNameNode
14007 NameNode
14221 DataNode
20878 FsShell
```

图 2.5 文件上传过程中的进程

```
yangxiya123@ubuntu:~$ jps
6372 SecondaryNameNode
7031 Jps
6009 NameNode
6205 DataNode
```

2、只有 namenode，据有类似问题的同学反馈，该情况是因为 ssh 挂了，重新配一遍 ssh 即可（可以理解为虚拟机没有 ssh 这个功能了）

3、查看类似网页需要在虚拟机里面的火狐浏览器看！！不是在外部的火狐浏览器



连接失败

Firefox 无法建立到 localhost:27017 服务器的连接。

- 此站点暂时无法使用或者太过忙碌。请过几分钟后重试。
- 如果您无法加载任何网页，请检查您计算机的网络连接状态。
- 如果您的计算机或网络受到防火墙或者代理服务器的保护，请确认 Firefox 已被授权访问网络。

重试

4、实验结束给助教老师发截图的时候，运行 `./bin/hadoop fs -cat output/grep/p*` 指令前，记得先启动 MapReduce 服务和 Hadoop 服务

```
yangxiya123@ubuntu:~/hadoop-1.2.1$ ./bin/start-mapred.sh
starting jobtracker, logging to /home/yangxiya123/hadoop-1.2.1/libexec/./logs/hadoop-yangxiya123-jobtracker-ubuntu.out
localhost: starting tasktracker, logging to /home/yangxiya123/hadoop-1.2.1/libexec/./logs/hadoop-yangxiya123-tasktracker-ubuntu.out
yangxiya123@ubuntu:~/hadoop-1.2.1$ ./bin/start-dfs.sh
bash: ./bin/start-dfs.sh: No such file or directory
yangxiya123@ubuntu:~/hadoop-1.2.1$ ./bin/start-dfs.sh
starting namenode, logging to /home/yangxiya123/hadoop-1.2.1/libexec/./logs/hadoop-yangxiya123-namenode-ubuntu.out
localhost: starting datanode, logging to /home/yangxiya123/hadoop-1.2.1/libexec/./logs/hadoop-yangxiya123-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /home/yangxiya123/hadoop-1.2.1/libexec/./logs/hadoop-yangxiya123-secondarynamenode-ubuntu.out
yangxiya123@ubuntu:~/hadoop-1.2.1$ ./bin/hadoop fs -cat output/grep/p*
1      dfs.datanode.name.dir
1      dfs.namenode.name.dir
1      dfs.replication
1      dfs.server.namenode.
1      dfsadmin
yangxiya123@ubuntu:~/hadoop-1.2.1$
```

如果没启动的话会出现这样的问题

```
yangxly123@ubuntu:~/hadoop-1.2.1$ ./bin/hadoop fs -cat output/grep/p
24/03/07 21:44:25 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:26 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:27 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:28 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:29 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:30 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:31 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:32 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:33 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
24/03/07 21:44:34 INFO ipc.Client: Retrying connect to server: localhost/127.0.0.1:9000. Already tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10,
sleepTime=1 SECONDS)
cat: Call to localhost/127.0.0.1:9000 failed on connection exception: java.net.ConnectException: Connection refused
yangxly123@ubuntu:~/hadoop-1.2.1$ ./bin/start-mapred.sh
```

- 5、如果运行过程出现 Java Heap Space 异常，那么说明进程的堆内存不足。请按照实验说明进行修改。如果是 grep 过程中卡住了，可以尝试调大虚拟机的内存设置。整个实验过程中，使用云主机的同学遇到的问题似乎比使用本地 Ubuntu 的同学要多。如果使用云主机，不要改它的 hostname。

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：2021

上机实践成绩：

指导教师：徐辰

姓名：杨茜雅

学号：10215501435

上机实践名称：Hadoop2.x 部署

上机实践日期：

2024.3.27

此次实验分成两个部分：伪分布式（必做）和分布式部署（选做）。我两个实验都完成了，并且都是**独自完成**，没有进行小组合作，分布式部署开设了**四个本地的虚拟机**。

- 1、在分布式部署中我没有使用云主机，我在本地开设了四个虚拟机。在这一步需要知道四个主机的 IP 地址，如果你使用云主机的话，需要使用内网 ip，如果使用虚拟机的话可以采用几种指令来查询 ip: `hostname -I`、`ip addr`、`ip a`、`ifconfig`。我的四个虚拟机的 ip 地址是连在一起的。

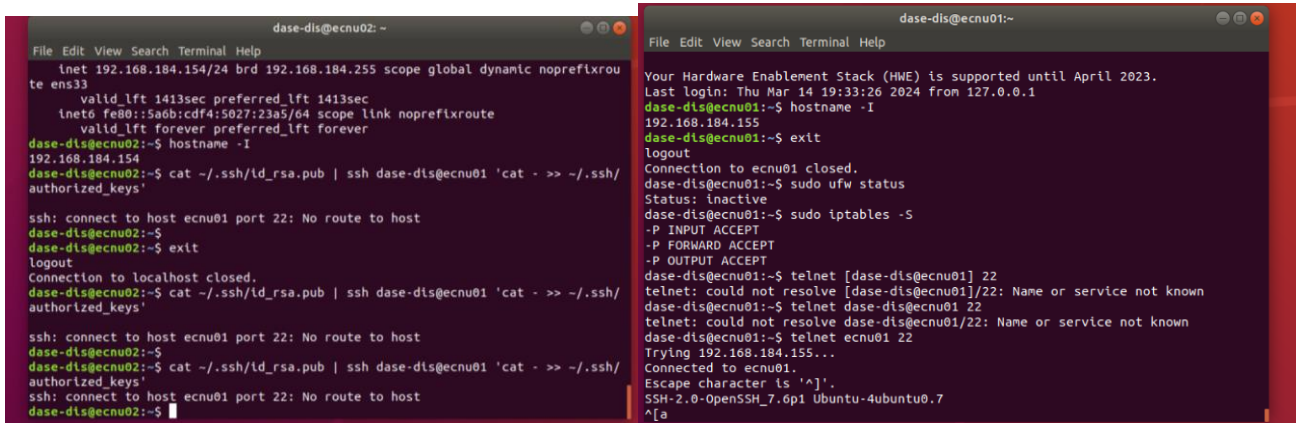
```
1 #IP地址 主机名
2 10.24.15.11 ecnu01
3 10.24.15.12 ecnu02
4 10.24.15.13 ecnu03
5 10.24.15.14 ecnu04
6 # 注意，以上10.24.15.11、10.24.15.12、10.24.15.13、10.24.15.14
   四个IP地址仅作示范，请分别替换为四台机器实际的IP地址
```

- 2、实验过程中我有遇到 ip 名和主机名不匹配的问题。我在初始设置的时候就将主机命名为了 ecnu01，但是实际打开一看发现还是 ubuntu。于是我选择用指令修改，但是使用指令修改完以后，虽然显示修改成功但是主机名仍然是 ubuntu。于是我选择用 `hostname` 或者 `hostnamectl status` 指令在当前终端会话中查询目前的主机名，它显示是 ecnu01，最后 `restart` 一下虚拟机，主机名就变了。可以得出结论：当使用 `hostnamectl set-hostname` 命令后，如果立即在终端查看主机名可能不会显示变化，这是因为当前会话和终端仍然保持着旧的主机名。如果要想看到主机名的变化，需要重启虚拟机。

```
File Edit View Search Terminal Help
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

dase-dis@ubuntu:~$ sudo hostnamectl set-hostname ecnu01
[sudo] password for dase-dis:
dase-dis@ubuntu:~$ sudo hostnamectl set-hostname ecnu01
dase-dis@ubuntu:~$
```

- 3、在进行四台主机之间的免密钥登陆时，出现如图的问题，我的系统无法找到目标地址的路径，这可能有几个原因：1、主机地址错误 2、端口 22 被阻塞 3、目标主机没有运行 SSH 服务 4、路由配置问题 5、防火墙问题 6、远程主机没有运行



The image shows two terminal windows. The left window, titled 'dase-dis@ecnu02: ~', displays network configuration commands and the output of 'cat ~/.ssh/ld_rsa.pub'. It shows the configuration of the network interface 'ens33' with IP 192.168.184.154 and the creation of a public key. The right window, titled 'dase-dis@ecnu01: ~', shows the output of 'hostname -I' (192.168.184.155), 'exit', 'logout', and 'sudo ufw status' (Status: inactive). It also shows the output of 'sudo iptables -F' and 'sudo iptables -P INPUT ACCEPT'. The bottom of the right window shows the output of 'telnet [dase-dis@ecnu01] 22', which fails with the message 'telnet: could not resolve [dase-dis@ecnu01]/22: Name or service not known'.

根据实验和我本身设备的特殊性，我觉得原因会在 1 和 2 中，因为前一天晚上还能运行免密钥登陆，第二天早上就不可以了，不太可能是虚拟机本身的问题。于是我打算检查 22 端口和 ip 地址是否正确。检查 22 端口可以使用 telnet ecnu02（主机名） 22 来检查，命令成功连接到 ecnu01 的 22 端口，并显示了 SSH 的版本信息，这表明该端口是开放的，并且 SSH 服务正在运行。此外，截图还显示了 UFW 防火墙的状态是非活动（inactive），并且 iptables 显示所有传入、传出和转发的流量都被接受（ACCEPT），这表明没有防火墙规则阻止 22 端口。所以只可能是 ip 问题，后来检查 ip 地址才发现第二天的虚拟机 ip 地址相比于第一天的有所改变了？我也不知道是为什么，总而言之问题排查成功。

4、在 dase-dis@ecnu01 上运行 ssh dase-dis@ecnu02 可以成功但是在 dase-dis@ecnu02 上运行 dase-dis@ecnu01 就会发生如图问题，ecnu02 在尝试建立网络连接到 ecnu01 和 ecnu03 时，无法找到一条可用的路径到达目标主机。如果在 ecnu01 上运行 ssh dase-dis@ecnu02 可以成功，这说明 ecnu02 主机是可以被访问的。问题出现在从 ecnu02 访问其他主机，后来我发现是在进行 lab1 那一步的时候，我只修改了 ecnu01 的文件 /etc/hosts，后面把 ecnu020304 的都同步修改即可。

```
dase-dis@ecnu02: ~
File Edit View Search Terminal Help
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage

51 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

New release '20.04.6 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Your Hardware Enablement Stack (HWE) is supported until April 2023.
*** System restart required ***
Last login: Thu Mar 14 19:58:06 2024 from 192.168.184.155
dase-dis@ecnu02:~$ exit
logout
Connection to localhost closed.
dase-dis@ecnu02:~$ ssh dase-dis@ecnu01
ssh: connect to host ecnu01 port 22: No route to host
dase-dis@ecnu02:~$ ssh dase-dis@ecnu03
ssh: connect to host ecnu03 port 22: No route to host
dase-dis@ecnu02:~$
```

c) 打开文件后，按“i”键进入编辑模式，添加映射信息，映射信息如以下列表所示。映射信息添加完成后，按“Esc”键退出编辑模式，并输入“:wq!”保存并退出

| #IP地址 | 主机名 |
|--|--------|
| 10.24.15.11 | ecnu01 |
| 10.24.15.12 | ecnu02 |
| 10.24.15.13 | ecnu03 |
| 10.24.15.14 | ecnu04 |
| # 注意，以上10.24.15.11、10.24.15.12、10.24.15.13、10.24.15.14 四个IP地址仅作示范，请分别替换为四台机器实际的IP地址 | |

5、 哪一个真实的 ip 地址？

```
dase-dis@ecnu02:~$ telnet ecnu02 22
Trying 192.168.184.151...
telnet: Unable to connect to remote host: No route to host
dase-dis@ecnu02:~$ hostname -I
192.168.184.154
dase-dis@ecnu02:~$
```

hostname -I 命令返回了 192.168.184.154，这应该是 ecnu02 的真实内网 IP 地址。而在尝试使用 telnet ecnu02 22 连接时，尝试连接的 IP 地址是 192.168.184.151，由于出现了“No route to host”的错误，这说明 192.168.184.151 可能不是 ecnu02 的正确 IP 地址，或者这个地址当前不可达（原因是 ecnu02 主机的/etc/hosts 文件中，我设置的是 192.168.184.151，这是个错误的地址）

6、显示没有 jps，按要求安装即可。出现 vim 功能没有的情况也是按要求手动装一下就行。

指令：sudo apt update、sudo apt install vim

```
File Edit View Search Terminal Help
Connection to ecnu04 closed.
dase-dis@ecnu03:~$ jps
Command 'jps' not found, but can be installed with:
sudo apt install openjdk-11-jdk-headless
sudo apt install openjdk-8-jdk-headless
dase-dis@ecnu03:~$ sudo apt install openjdk-8-jdk-headless
[sudo] password for dase-dis:
Reading package lists... Done
Building dependency tree
Reading state information... Done
```

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：2021

上机实践成绩：

指导教师：徐辰

姓名：杨茜雅

学号：10215501435

上机实践名称：MapReduce 2 编程

上机实践日期：

2024.4.17

实验四：MapReduce 2.x 编程

此次实验分成三个部分：在单机集中式、单机伪分布式、分布式部署方式下运行应用程序。三个部分已全完成，并且都是独自完成，没有进行小组合作，分布式部署开设了两个本地的虚拟机。（一个主节点一个客户端节点）

1、在调试 MapReduce 应用程序环节出现如下问题

```
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException Create breakpoint : Output directory
at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:146)
at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:272)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:145)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1578)
at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1567) <1 internal call>
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1926)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1588)
at cn.edu.ecnu.mapreduce.example.java.wordcount.WordCount.run(WordCount.java:37)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:76)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:98)
at cn.edu.ecnu.mapreduce.example.java.wordcount.WordCount.main(WordCount.java:42)
```

发现是配置 a 环节实验手册有误，应该为/home/dase-local/input/pd.train /home/dase-local/IdeaProjects/MapReduceDemo/output

- 配置 Main Class 为 `cn.edu.ecnu.mapreduce.example.java.wordcount.WordCount`
- 配置 Program arguments 为 `/home/dase-local/input/ /home/dase-local/IdeaProjects/MapReduceDemo/output`

2、打包 jar 后，运行 ./bin/hadoop jar ./myApp/WordCount.jar input/pd.train output 时报错

```
t
yangxiya123@ubuntu:~$ cd hadoop-2.10.1/
yangxiya123@ubuntu:~/hadoop-2.10.1$ ./bin/hadoop jar ./myApp/WordCount.jar input/pd.train output
Exception in thread "main" java.lang.ClassNotFoundException: cn.edu.ecnu.mapreduce.example.java.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.hadoop.util.RunJar.run(RunJar.java:237)
at org.apache.hadoop.util.RunJar.main(RunJar.java:158)
yangxiya123@ubuntu:~/hadoop-2.10.1$
```

Java 运行时找不到 `cn.edu.ecnu.mapreduce.example.java.wordcount` 这个类。猜测原因有：

1. JAR 文件结构不对，可能是打包的时候漏了什么类，使得 `WordCount` 类没有位于 `cn/edu/ecnu/mapreduce/example/java/wordcount/` 这个路径下。

2. 主类在 MANIFEST.MF 中未正确指定：运行 JAR 时，如果发现 `MANIFEST.MF` 文件中没有指定主类，是需要我在命令行中指定的。

对于问题 1：我采用 `jar tf WordCount.jar` 查看 JAR 文件中的内容列表，内容无误。

```

yangxia123@ubuntu:~/hadoop-2.10.1/myApp$ ls
WordCount.jar
yangxia123@ubuntu:~/hadoop-2.10.1/myApp$ jar tf WordCount.jar
META-INF/MANIFEST.MF
cn/
cn/edu/
cn/edu/ecnu/
cn/edu/ecnu/mapreduce/
cn/edu/ecnu/mapreduce/example/
cn/edu/ecnu/mapreduce/example/java/
cn/edu/ecnu/mapreduce/example/java/wordcount/
cn/edu/ecnu/mapreduce/example/java/wordcount/WordCountMapper.class
cn/edu/ecnu/mapreduce/example/java/wordcount/WordCount.class
cn/edu/ecnu/mapreduce/example/java/wordcount/WordCountReducer.class
META-INF/

```

对于问题 2: 发现是打包时, Main Class 设置错了, 少了红框内容

```

'WordCount.jar' manifest properties:
Manifest File: /IdeaProjects/MapReduceDemo/META-INF/MANIFEST.MF
Main Class:   ecnu.mapreduce.example.java.wordcount.WordCount

```

3、紧跟上一步, 报了不同的错, 在 Hadoop MapReduce 中, 如果输出目录已经存在, 作业默认不会覆盖该目录, 以防止不小心删除重要数据, 而这里显示我的输出目录已经存在。

删除它就可以: `./bin/hdfs dfs -rm -r /user/yangxia123/output`

再运行: `./bin/hadoop jar ./myApp/WordCount.jar input/pd.train output`

```

./bin/hadoop jar ./myApp/WordCount.jar input/pd.train output
24/03/25 08:11:39 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory
hdfs://localhost:9000/user/yangxia123/output already exists
at
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java
at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:279)

```

4、实验手册已经写明了打包路径是这样的

jar 包名称为 WordCount.jar, 打包路径为 `/home/dase-local/IdeaProjects/HDFSFileOP/out/artifacts/WordCount/`, 配置界面如图4.3所示。

所以在下一步应该自己看着办改掉红框内容

```

cp -r /IdeaProjects/MapReduceDemo/out/artifacts/WordCount/WordCount.jar
~/hadoop-2.10.1/myApp/
#将打包好的可执行jar包拷贝到hadoop安装路径下的myApp/目录下

```

5、分布式部署需要两台能互通的虚拟机, 因为我的环境问题所以我又配置了一遍免密钥登陆。在执行 `su dase-dis` 时, 我遇到 `Password: su: Authentication failure` 的问题, 这是因为当使用 `useradd` 创建一个新用户但没有设置密码时, 该用户将没有有效的密码, 因此我无法直接用 `su` 切换到该用户。对此我需要 `sudo passwd dase-dis` 才能创建新用户。

6、

- 注意在配置免密钥登陆运行 `sudo vim /etc/hosts` 时, 最好需要互通的设备的使用一

样的 hosts 文件，你别管里面是不是有用不上的信息。

- `cat ~/.ssh/id_rsa.pub | ssh dase-dis@ecnu01 'cat - >> ~/.ssh/authorized_keys'` 要注意是谁发给谁，虽然不清楚原理，但我确实在 A 发给 B 时遇到了错误，并且在 B 发给 A 时错误消失，推测是要分辨好主节点和客户端节点。

- 后面在 `scp ~/.ssh/authorized_keys dase-dis@ecnu02:/home/dase-dis/.ssh/authorized_keys` 时提示你输密码，但总是 `permission denied`，八成是主机名和用户名写错了，仔细看看。