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： 准备工作	**上机实践日期**：	2024.3.6

问题 1：主机名修改后无论是什么命令都报错 Name or service not known。

在配置多台主机之间的免密登录时，理论上来说，根据实验指导文档，修改完 /etc/sudoers 文件之后就可以通过 sudo hostnamectl set-hostname ecnu01 将主机名修改为 ecnu01，这一步也没有问题，但是之后输入所有的命令都出现 Name or service not known 而无法进行了。这主要是因为我没有在修改好主机名之前便添加映射，只要先在 /etc/hosts 中添加映射再去利用 hostnamectl 修改主机名，这个问题便不会出现了。

问题 2：不在 sudoers 文件中，此事将被报告。

我最开始创建 dase-dis 用户是在 ubuntu 用户下通过 sudo useradd dase-dis 创建的，但是转到 dase-dis 用户之后使用 sudo 赋权时就出现了这样的报错。结合水杉上实验的演示视频，我发现正常的创建用户步骤是这样的（假定当前的用户是 ubuntu）：

1. sudo useradd -m dase-dis -d /home/dase-dis -s /bin/bash
2. sudo passwd dase-dis(修改 dase-dis 密码)
3. sudo passwd root(修改 root 密码)接下来 su 登录 root
4. vim /etc/sudoers 在#User privilege specification 中加入：dase-dis ALL=(ALL) ALL 保存文件
5. 之后 su dase-dis 之后 sudo 赋权就正常了。

问题 3：wget 报错：解析代理服务器 URL https://<proxy_name>:<port_number> 时发生错误：端口号错误。

在检查了一下自己 wget 的链接发现没有什么问题之后，我觉得出现这个问题的主要原因是没有给 wget 赋予相应的权限。在 wget 之前加入 sudo 之后下载任务便正常开始了。

问题 4：在第二个终端中 jps not found

出现这个问题时我考虑到了环境变量可能还没来得及重新加载，输入 source /etc/profile 之后 jps 命令便可以正常使用了。

问题 5：IDEA 中 Build 程序时出现错误

经过一段时间的排查，我才发现自己的 language level 和 jdk 版本不适配。由于我使用的是 openjdk-21，所以需要修改 project structure 的 language level，使得 Modules 和 Project 中 language level 版本一致。具体来说，我把 SDK 中 jdk 版本切换到了 1.8，然后再改一下 Modules 和 Project 中的 language level 使之为 8 就行了。

问题 6：su dase-dis 时 Password: su: Authentication failure

查阅网上的资料后，我发现这是因为 useradd 创建一个新用户但没有设置密码时，该用户并无有效的密码，因此我不能用 su 切换到该用户。我在 sudo passwd dase-dis 修改 dase-dis 的密码之后，再 su 就不报错了。

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： Hadoop 1.x 部署	**上机实践日期**：	2024.3.13

问题 1：在云主机上新建用户后终端只显示$

经过排查，我发现自己新建用户时输入的命令是 sudo useradd -m dase-dis -d /home/dase-dis 没有添加-s /bin/bash，加上这个选项之后问题便解决了。

问题 2：Hadoop 启动时 Error: JAVA_HOME is not set and could not be found

为了解决这个问题，我觉得 JAVA_HOME 环境变量可能还没有生效，就去 source /etc/profile,但是依然没有解决。然后我修改/etc/hadoop/hadoop-env.sh 中设 JAVA_HOME 为绝对路径，即 export JAVA_HOME=/usr/local/jdk1.8，然后再运行 Hadoop 发现能正常运行了。

问题 3：我在 dase-dis 用户运行 wordcount 时，另外一个终端如果是 dase-dis 用户 jps 什么都没有，但是如果是 root 用户就能看到 RunJar 进程。

经过漫长的检查，我终于发现自己的 hadoop 命令之前错误地添加了一个 sudo，就比如原本应该是./bin/hadoop jar hadoop-examples-1.2.1.jar wordcount ~/input/pd.train ~/output/wordcount,这如果是 dase-dis 运行,另外一个终端如果再用 dase-dis 登录当然 jps 是能看到 wordcount 运行着的。但如果在此之前加一个 sudo，运行命令的便是 root 了，所以另外一个终端用 dase-dis 再 jps 就什么也看不到了。

为什么我还要再加一个 sudo 呢？这主要是因为权限问题，不加 sudo 的话直接用 dase-dis 其实 hadoop 是跑不起来的。要想解决这个问题，我先在 dase-dis 的工作目录下 sudo chmod -R 777 ~/hadoop-1.2.1 修改权限，然后不带 sudo 运行 hadoop 即可。

问题 4：jps 能看到 secondarynamenode 但是竟然看不到 namenode

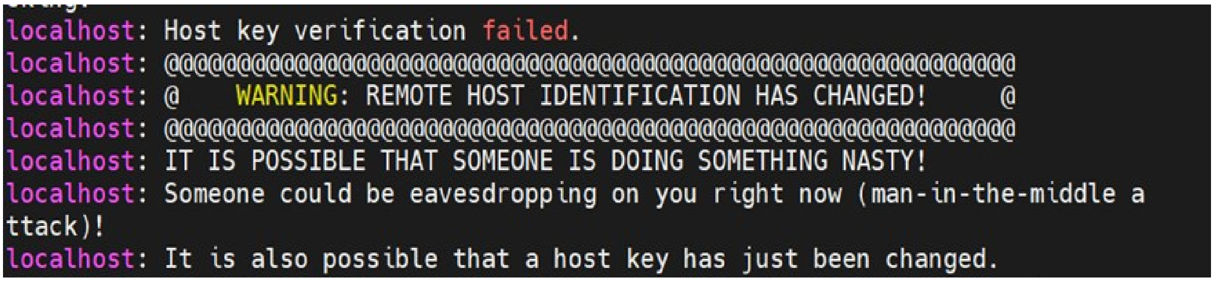
我不小心多次格式化了 namenode，然后我在启动 hdfs 后，进程列表中只有 secondarynamenode 但却没 namenode 进程，之后的各项操作也都是报错。经过查阅 namenode.log 日志，我发现自己的 tmp 文件夹已经损坏，需要杀死所有进程，把 tmp 文件全部删除并重新格式化 namenode。我这样做并重启了 hdfs 服务之后，进程列表里又可以看见 namenode 了

问题 5：运行时出现 Java Heap Space 异常

经过助教学长的提醒，我知道了这个问题是因为进程的堆内存不足所导致的。我调大了虚拟机的内存，从原来的 1G 调整到了 16G，这个问题还是没有解决。后来经过沟通，我才知道如果在云主机上做这个实验不能修改 hostname，保持它不变的情况下我再以一个较大的内存来做，就正常了。

问题 6：即便镜像中的云主机已经配置了免密登录，但是从之新建的云主机验证时还是出错了。

具体来说，我觉得每次从镜像中创建的云主机存在一些差异，彼此之间是不太一样的，所以镜像中那个.ssh 文件中记录的密钥信息并不匹配。因此，我就把.ssh 文件重置了一下，然后按照免密登录的配置流程操作了一下，问题就解决了。

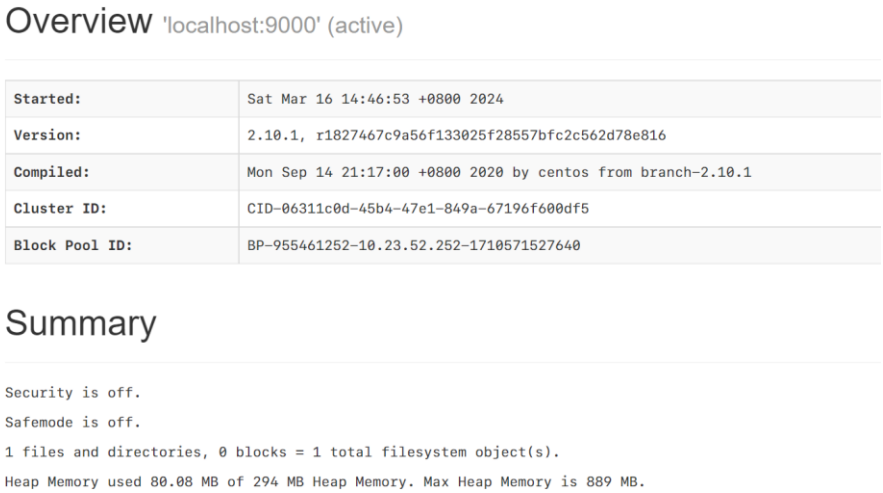


华东师范大学数据科学与工程学院实验报告

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：Hadoop 2 部署	**上机实践日期：**	2024.3.27

问题 1：在浏览器中输入 localhost:50070 查看结点内部情况时发现无法访问

这主要是因为自己的浏览器是本地浏览器，而不是运行 HDFS 的云主机。输入 xxx:50070 问题便解决了（xxx 为云主机的外网 ip，要现在防火墙中把所有的端口打开）。



问题 2：分布式实验时，四台主机单机和它们互相之间的免密钥登录设置后还是需要输入密码

经过一段时间的排查，我发现自己错误的原因是 authorized 被错打成 authorised 了，然后就一错再错导致免密登录并没有真的设置成功。

问题 3：多次格式化 namenode 导致启动 HDFS 后 ecnu01 中没有 namenode 进程

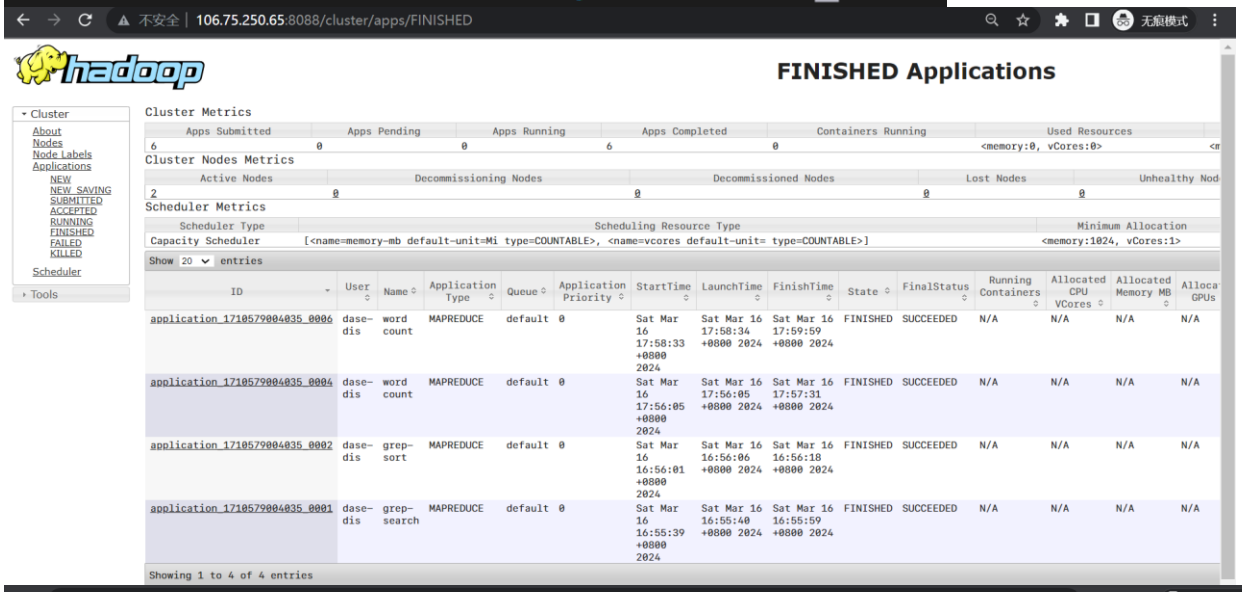
在详细查看了 namenode.log 日志后，我发现 tmp 文件夹已经损坏。我应该杀死所有进程之后，把四台主机上的 tmp 文件全部删除并重新格式化。我在这样做并重启了 HDFS 服务，发现 ecnu01 的进程列表里又有 namenode 了。

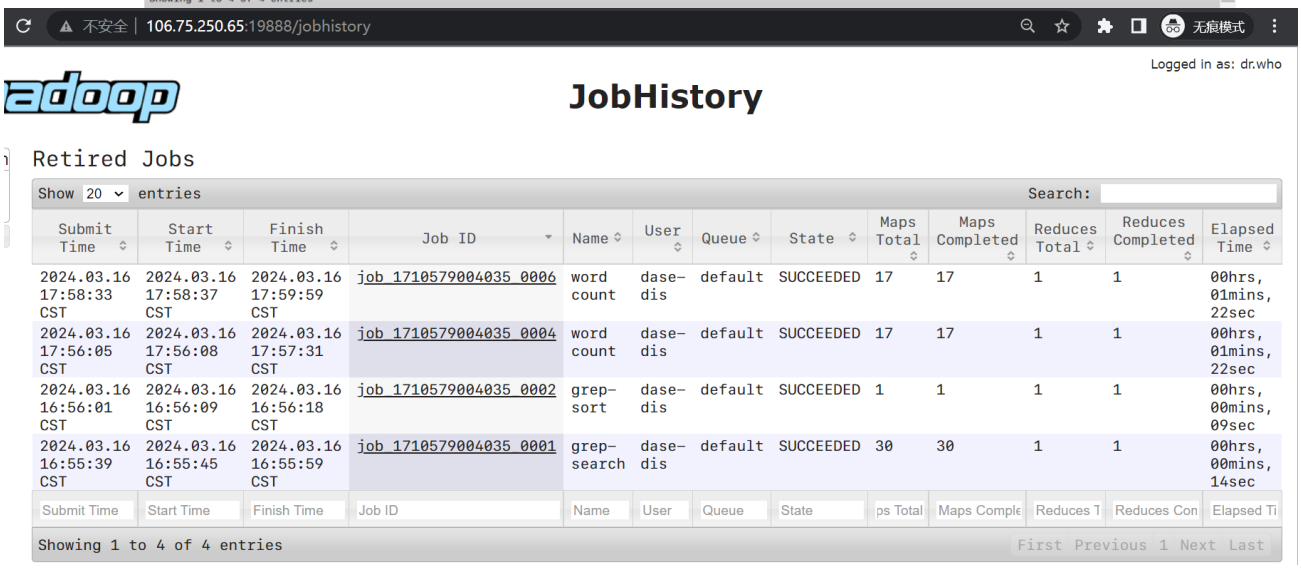
其实这个实验我严格按照实验操作文档一步一步去做，并没有出现很多不好解决的问题。在分布式实验时，grep 从客户端提交之后运行：

```
dase-dis@ecnu04:~/hadoop-2.10.1$ ./bin/hdfs dfs -cat output/grep/p*
6       dfs.audit.logger
4       dfs.class
3       dfs.logger
3       dfs.server.namenode.
2       dfs.audit.log.maxbackupindex
2       dfs.period
2       dfs.audit.log.maxfilesize
1       dfs.replication
1       dfs.log
1       dfs.file
1       dfs.datanode.data.dir
1       dfs.servers
1       dfsadmin
1       dfsmetrics.log
1       dfs.namenode.name.dir
dase-dis@ecnu04:~/hadoop-2.10.1$
```

在运行 wordcount 时从节点的进程信息：

```
dase-dis@ecnu02:~/hadoop-2.10.1$ jps
4960 YarnChild
4739 MRAppMaster
4995 YarnChild
2101 NodeManager
4999 YarnChild
2329 DataNode
4922 YarnChild
5180 Jps
4956 YarnChild
4927 YarnChild
dase-dis@ecnu02:~/hadoop-2.10.1$
```





Submit Time	Start Time	Finish Time	Job ID	Name	User	Queue	State	Maps Total	Maps Completed	Reduces Total	Reduces Completed	Elapsed Time
2024.03.16 17:58:33 CST	2024.03.16 17:58:37 CST	2024.03.16 17:59:59 CST	job_1710579004035_0006	word count	dase-dis	default	SUCCEEDED	17	17	1	1	00hrs, 01mins, 22sec
2024.03.16 17:56:05 CST	2024.03.16 17:56:08 CST	2024.03.16 17:57:31 CST	job_1710579004035_0004	word count	dase-dis	default	SUCCEEDED	17	17	1	1	00hrs, 01mins, 22sec
2024.03.16 16:56:01 CST	2024.03.16 16:56:09 CST	2024.03.16 16:56:18 CST	job_1710579004035_0002	grep-sort	dase-dis	default	SUCCEEDED	1	1	1	1	00hrs, 00mins, 09sec
2024.03.16 16:55:39 CST	2024.03.16 16:55:45 CST	2024.03.16 16:55:59 CST	job_1710579004035_0001	grep-search	dase-dis	default	SUCCEEDED	30	30	1	1	00hrs, 00mins, 14sec

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：MapReduce 2 编程	**上机实践日期**：	2024.4.17

问题 1：org.apache.hadoop.io.nativeio.NativeIO$Windows.access0

由于 Windows 不支持 HDFS，因此我们在 Windows 本地运行 MapReduce 程序时需要使用 winutils.exe 文件（winutils.exe 提供了一个包装器）

1.如果本地没有安装 winutils.exe，那首先就需要我们下载 winutils.exe 和 hadoop.dll 这两个文件（https://github.com/steveloughran/winutils，选择对应的 hadoop 版本）。

2. 然后，将上面提到的两个文件拷贝到本地 $HADOOP_HOME/bin 目录下（需要先安装 Hadoop 并配置好 HADOOP_HOME 环境变量），通常到这里就可以解决上述问题，如果仍然报错，可以尝试重启系统，让环境变量生效。

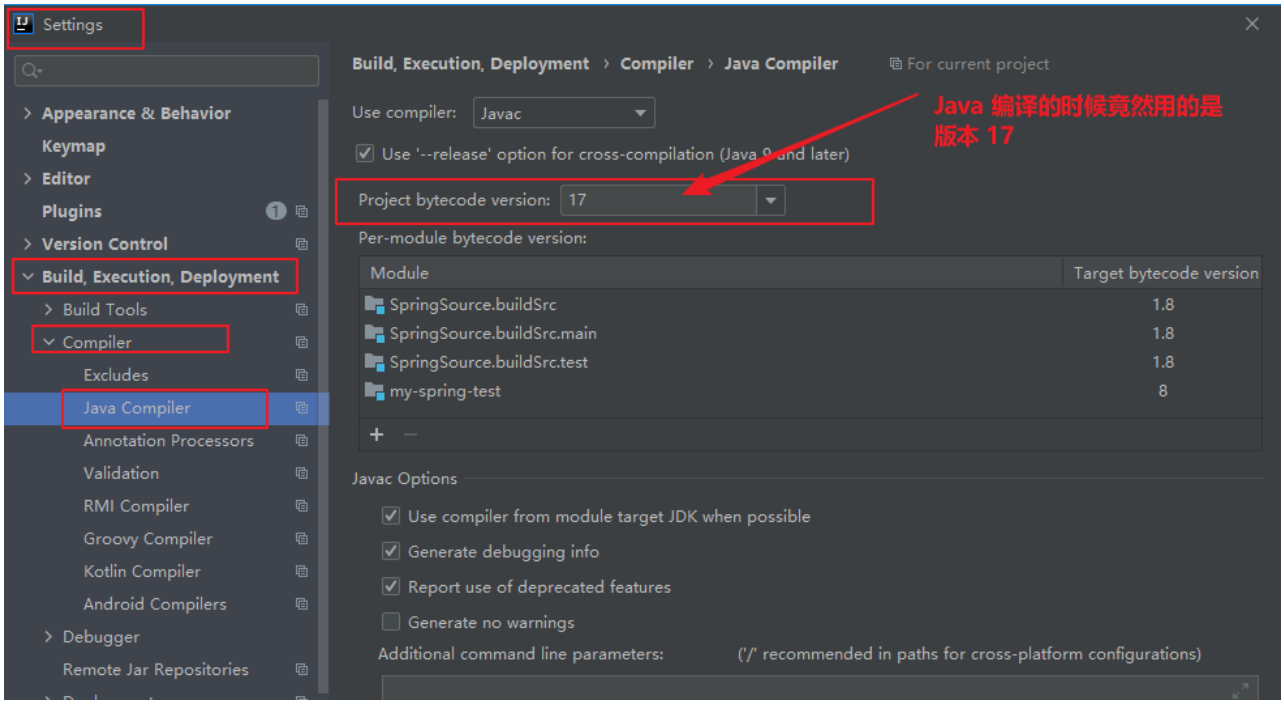
3. 如果通过步骤 2 仍然解决不了问题，那么可能是系统问题，可以将 hadoop.dll 拷贝到 C:/Windows/System32 目录下。

问题 2：Exception in thread "main" java.lang.UnsupportedClassVersionError: cn/edu/ecnu/mapreduce/example/java/wordcount/WordCount has been compiled by a more recent version of the Java Runtime (class file version 61.0), this version of the Java Runtime only recognizes class file versions up to 52.0

我查了一下 Java 的版本对应情况，是这样的——Java 8 (52)、Java 17 (61)。看来这个问题的原因是因为自己的 JDK 版本无法匹配到程序的版本。所以我在 IDEA 中下载了版本为 jdk1.8 的 SDK，尝试了几次之后发现只有 Azul Zulu version1.8.0_402 是可以下载的。安装之后再配置好对应的 SDK，该问题就解决了。

问题 3：idea 报错：无效的目标发行版：17 的解决办法

原因就是 JDK 版本不对。从 IDEA 编辑器中可以找到问题的原因所在，如下图是编辑器里的配置. 将 Settings --> Build, Execution, Deployment --> Complier --> Java Complier 配置下的 Project bytecode version: 17 改为 与 项目使用的 JDK 版本一样即可。



问题 4：jar 包里没有 WordCount 类

这是因为自己没有把类的名称写全，我应该把 package 的名称放在类的前面。于是，我把类的名称从 WordCount 改成了 cn.edu.ecnu.mapreduce.example.java.wordcount.WordCount，本地调试好的程序就能顺利地在云主机上运行了。

感觉这个实验和实验 3 还是挺相近的，我在实验过程中也是严格按照文档去做，并没有遇到什么太大的问题。单机伪分布式运行结束之后的结果和分布式运行结束之后的结果是一样的：