中国地质大学

本科生课程论文

课程名称 大数据技术基础

教师姓名 卢超

学生姓名 常文瀚

学生学号 20181001095

学生班级 191181

所在学院 计算机学院

完成日期 2021 年 6 月 20 日

目录

第一章 熟悉常用的HDFS操作

1.1 实验目的

1. 理解HDFS在Hadoop体系结构中的角色；
2. 熟练使用HDFS操作常用的Shell命令；

1.2 实验平台

操作系统：Linux

操作平台：Docker

Hadoop版本：2.7.1或以上版本

JDK版本：1.8或以上版本

1.3 实验内容和要求

1. 编程实现以下指定功能，并利用Hadoop提供的Shell命令完成相同任务：

（1） 向HDFS中上传任意文本文件，如果指定的文件在HDFS中已经存在，由用户指定是追加到原有文件末尾还是覆盖原有的文件；

（2） 从HDFS中下载指定文件，如果本地文件与要下载的文件名称相同，则自动对下载的文件重命名；

（3） 将HDFS中指定文件的内容输出到终端中；

（4） 显示HDFS中指定的文件的读写权限、大小、创建时间、路径等信息；

（5） 给定HDFS中某一个目录，输出该目录下的所有文件的读写权限、大小、创建时间、路径等信息，如果该文件是目录，则递归输出该目录下所有文件相关信息；

（6） 提供一个HDFS内的文件的路径，对该文件进行创建和删除操作。如果文件所在目录不存在，则自动创建目录；

（7） 提供一个HDFS的目录的路径，对该目录进行创建和删除操作。创建目录时，如果目录文件所在目录不存在则自动创建相应目录；删除目录时，由用户指定当该目录不为空时是否还删除该目录；

（8） 向HDFS中指定的文件追加内容，由用户指定内容追加到原有文件的开头或结尾；

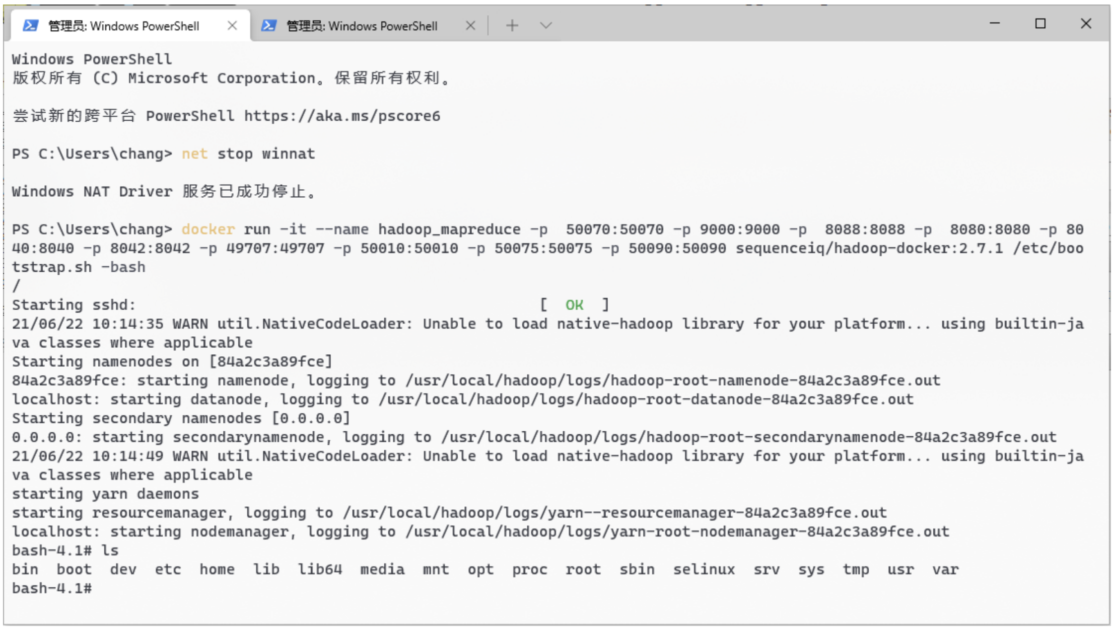
（9） 删除 HDFS 中指定的文件；

（10）删除 HDFS 中指定的目录，由用户指定目录中如果存在文件时是否删除目录；

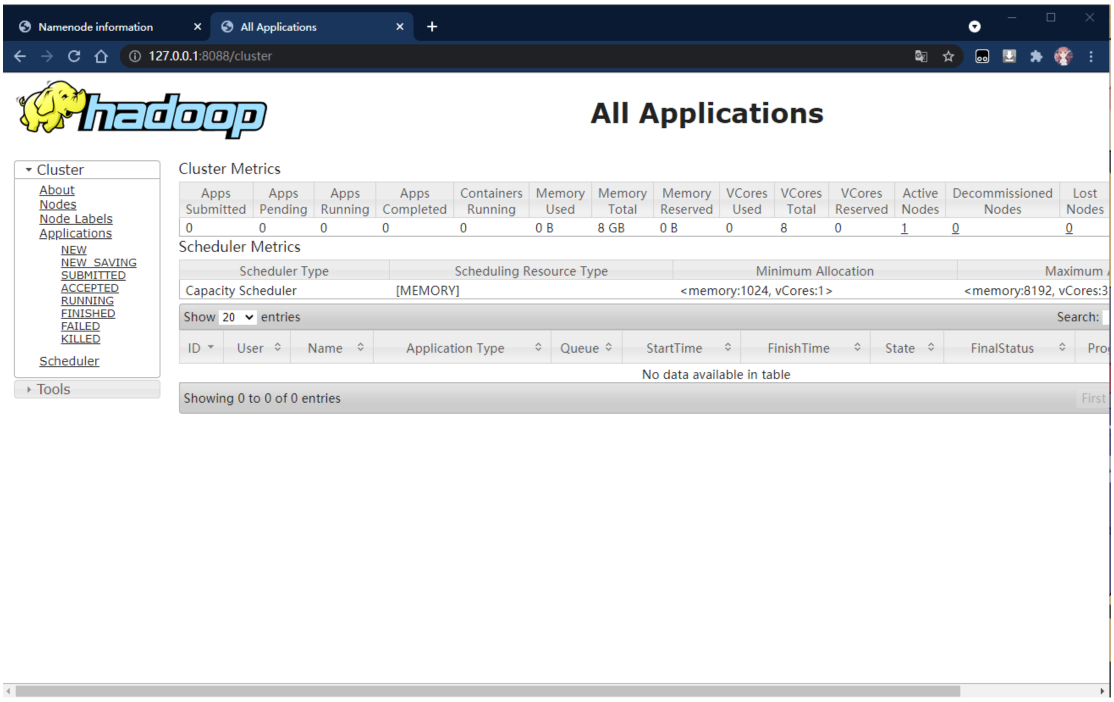
（11）在 HDFS 中，将文件从源路径移动到目的路径。

1. 4 实验过程

（1）安装 Docker，拉取在线镜像，并且运行。



```
Windows PowerShell
版权所有 (C) Microsoft Corporation。保留所有权利。

尝试新的跨平台 PowerShell https://aka.ms/pscore6

PS C:\Users\chang> net stop winnat

Windows NAT Driver 服务已成功停止。

PS C:\Users\chang> docker run -it --name hadoop_mapreduce -p  50070:50070 -p 9000:9000 -p  8088:8088 -p  8080:8080 -p 80
40:8040 -p 8042:8042 -p 49707:49707 -p 50010:50010 -p 50075:50075 -p 50090:50090 sequenceiq/hadoop-docker:2.7.1 /etc/boo
tstrap.sh -bash
/
Starting sshd:                                             [  OK  ]
21/06/22 10:14:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
Starting namenodes on [84a2c3a89fce]
84a2c3a89fce: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-84a2c3a89fce.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-84a2c3a89fce.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-84a2c3a89fce.out
21/06/22 10:14:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-84a2c3a89fce.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-84a2c3a89fce.out
bash-4.1# ls
bin  boot  dev  etc  home  lib  lib64  media  mnt  opt  proc  root  sbin  selinux  srv  sys  tmp  usr  var
bash-4.1#
```



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	8	0	1	0	0

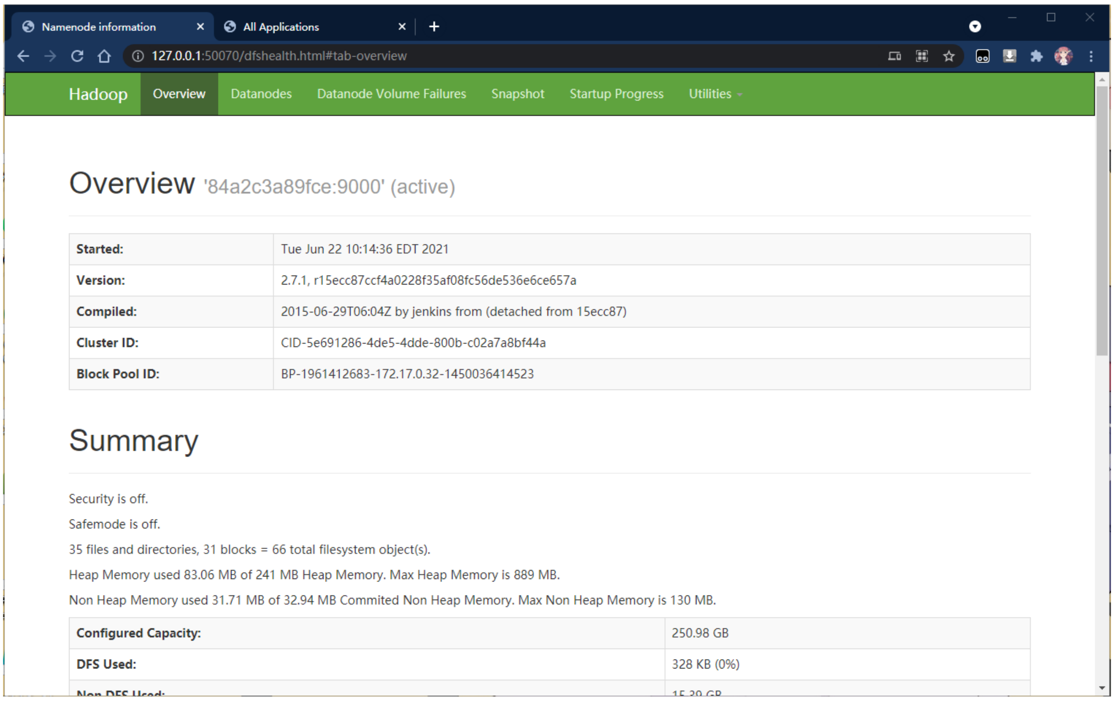
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:3

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Pro
No data available in table									

Showing 0 to 0 of 0 entries First



Overview '84a2c3a89fce:9000' (active)

Started:	Tue Jun 22 10:14:36 EDT 2021
Version:	2.7.1, r15ecc87ccf4a0228f35af08fc56de536e6ce657a
Compiled:	2015-06-29T06:04Z by jenkins from (detached from 15ecc87)
Cluster ID:	CID-5e691286-4de5-4dde-800b-c02a7a8bf44a
Block Pool ID:	BP-1961412683-172.17.0.32-1450036414523

Summary

Security is off.

Safemode is off.

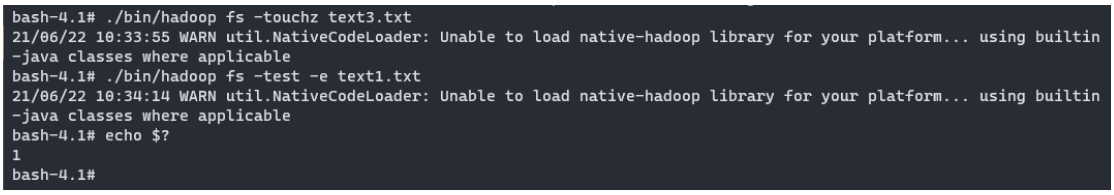
35 files and directories, 31 blocks = 66 total filesystem object(s).

Heap Memory used 83.06 MB of 241 MB Heap Memory. Max Heap Memory is 889 MB.

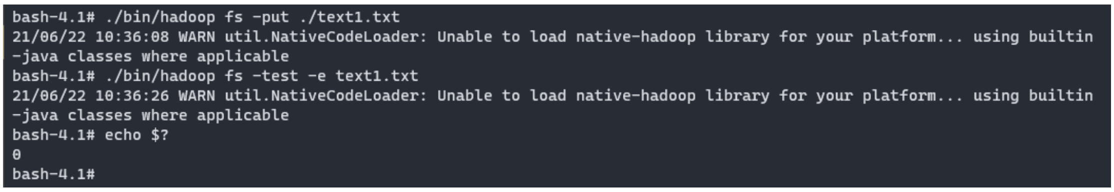
Non Heap Memory used 31.71 MB of 32.94 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.

Configured Capacity:	250.98 GB
DFS Used:	328 KB (0%)
Non DFS Used:	15.39 GB

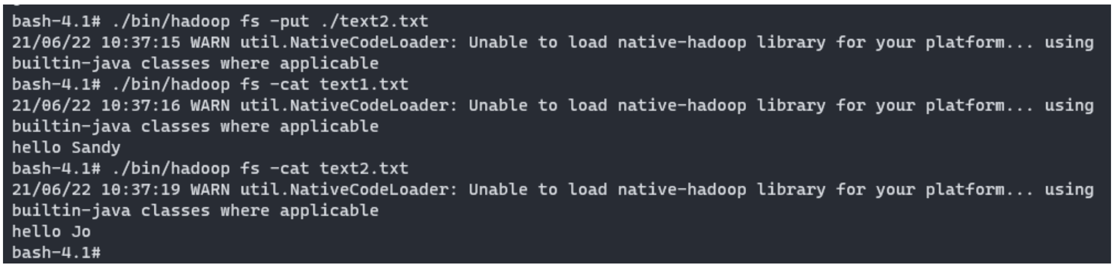
（2）查看文件是否存在于 HDFS，并编辑新的文本文件，将其上传，读取文件后将内容添加在末尾，并显示文本内容。



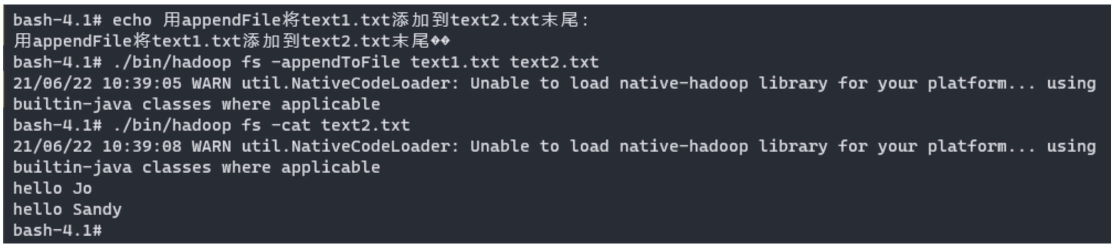
```
bash-4.1# ./bin/hadoop fs -touchz text3.txt
21/06/22 10:33:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
bash-4.1# ./bin/hadoop fs -test -e text1.txt
21/06/22 10:34:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
bash-4.1# echo $?
1
bash-4.1#
```



```
bash-4.1# ./bin/hadoop fs -put ./text1.txt
21/06/22 10:36:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
bash-4.1# ./bin/hadoop fs -test -e text1.txt
21/06/22 10:36:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
bash-4.1# echo $?
0
bash-4.1#
```

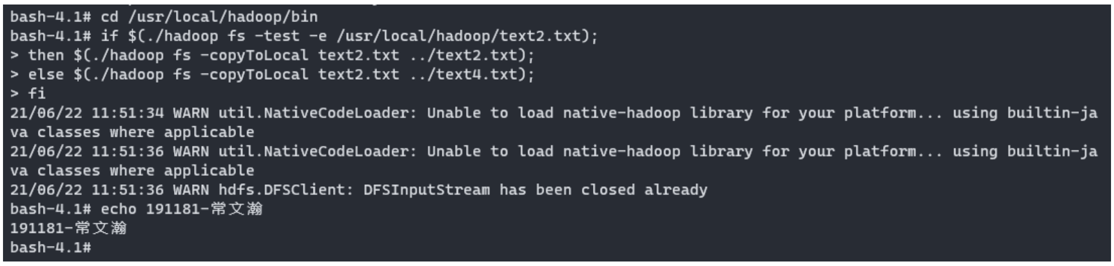


```
bash-4.1# ./bin/hadoop fs -put ./text2.txt
21/06/22 10:37:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
bash-4.1# ./bin/hadoop fs -cat text1.txt
21/06/22 10:37:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
hello Sandy
bash-4.1# ./bin/hadoop fs -cat text2.txt
21/06/22 10:37:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
hello Jo
bash-4.1#
```



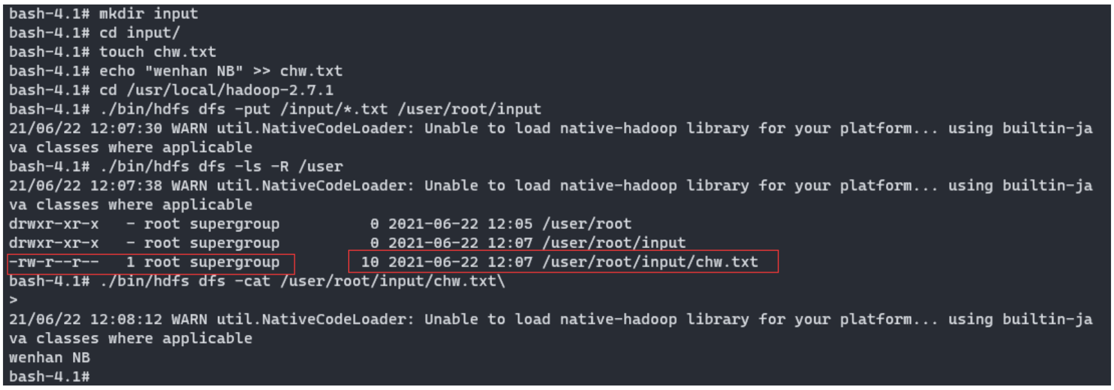
```
bash-4.1# echo 用appendFile将text1.txt添加到text2.txt末尾：
用appendFile将text1.txt添加到text2.txt末尾��
bash-4.1# ./bin/hadoop fs -appendToFile text1.txt text2.txt
21/06/22 10:39:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
bash-4.1# ./bin/hadoop fs -cat text2.txt
21/06/22 10:39:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
hello Jo
hello Sandy
bash-4.1#
```

（3）从 HDFS 中下载指定文件，如果本地文件与要下载的文件名称相同，则自动对下载的文件重命名。



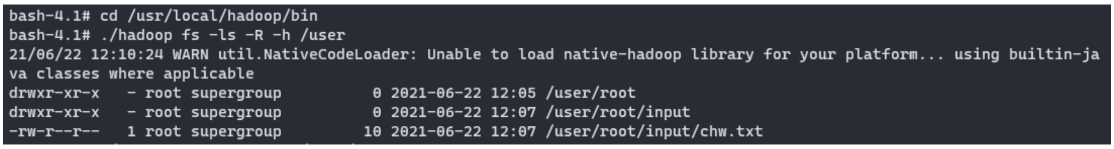
```
bash-4.1# cd /usr/local/hadoop/bin
bash-4.1# if $(./hadoop fs -test -e /usr/local/hadoop/text2.txt);
> then $(./hadoop fs -copyToLocal text2.txt ../text2.txt);
> else $(./hadoop fs -copyToLocal text2.txt ../text4.txt);
> fi
21/06/22 11:51:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 11:51:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 11:51:36 WARN hdfs.DFSClient: DFSInputStream has been closed already
bash-4.1# echo 191181-常文瀚
191181-常文瀚
bash-4.1#
```

（4）显示 HDFS 中指定的文件的读写权限、大小、创建时间、路径等信息。

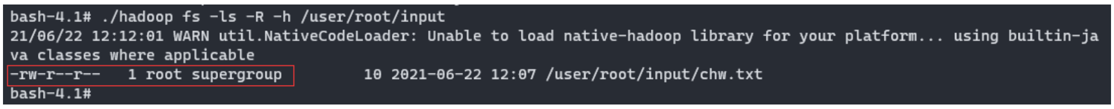


```
bash-4.1# mkdir input
bash-4.1# cd input/
bash-4.1# touch chw.txt
bash-4.1# echo "wenhan NB" >> chw.txt
bash-4.1# cd /usr/local/hadoop-2.7.1
bash-4.1# ./bin/hdfs dfs -put /input/*.txt /user/root/input
21/06/22 12:07:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hdfs dfs -ls -R /user
21/06/22 12:07:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
drwxr-xr-x   - root supergroup          0 2021-06-22 12:05 /user/root
drwxr-xr-x   - root supergroup          0 2021-06-22 12:07 /user/root/input
-rw-r--r--   1 root supergroup         10 2021-06-22 12:07 /user/root/input/chw.txt
bash-4.1# ./bin/hdfs dfs -cat /user/root/input/chw.txt\
>
21/06/22 12:08:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
wenhan NB
bash-4.1#
```

（5）给定 HDFS 中某一个目录，输出该目录下的所有文件的读写权限、大小、创建时间、路径等信息，如果该文件是目录，则递归输出该目录下所有文件相关信息。

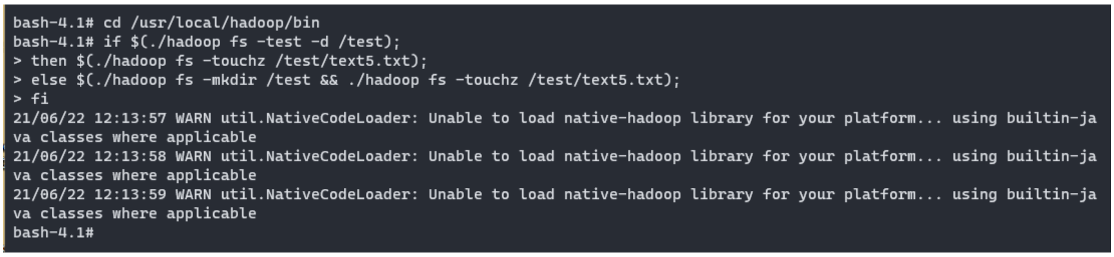


```
bash-4.1# cd /usr/local/hadoop/bin
bash-4.1# ./hadoop fs -ls -R -h /user
21/06/22 12:10:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
drwxr-xr-x   - root supergroup          0 2021-06-22 12:05 /user/root
drwxr-xr-x   - root supergroup          0 2021-06-22 12:07 /user/root/input
-rw-r--r--   1 root supergroup         10 2021-06-22 12:07 /user/root/input/chw.txt
```

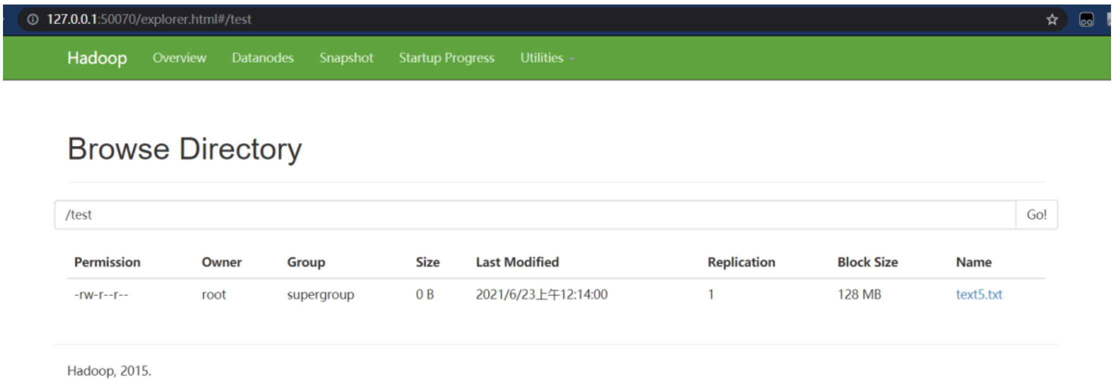


```
bash-4.1# ./hadoop fs -ls -R -h /user/root/input
21/06/22 12:12:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
-rw-r--r--   1 root supergroup         10 2021-06-22 12:07 /user/root/input/chw.txt
bash-4.1#
```

（6）提供一个 HDFS 内的文件的路径，对该文件进行创建和删除操作。如果文件所在目录不存在，则自动创建目录。



```
bash-4.1# cd /usr/local/hadoop/bin
bash-4.1# if $(./hadoop fs -test -d /test);
> then $(./hadoop fs -touchz /test/text5.txt);
> else $(./hadoop fs -mkdir /test && ./hadoop fs -touchz /test/text5.txt);
> fi
21/06/22 12:13:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 12:13:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 12:13:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1#
```

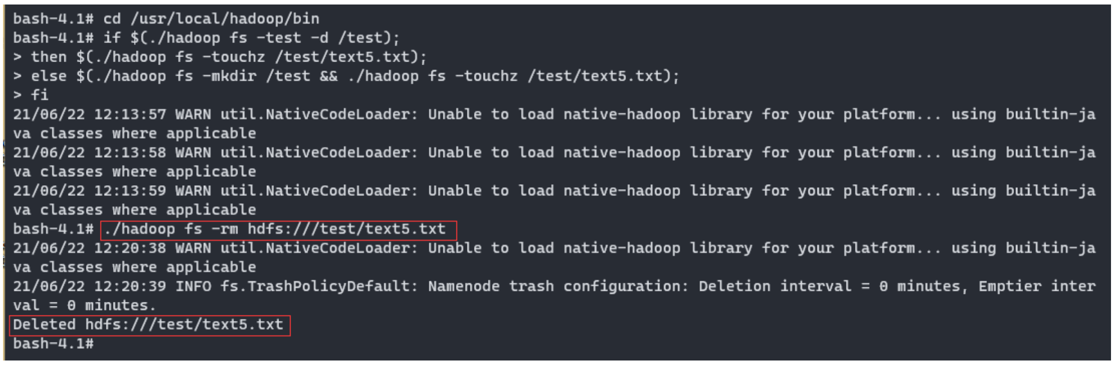


Browse Directory

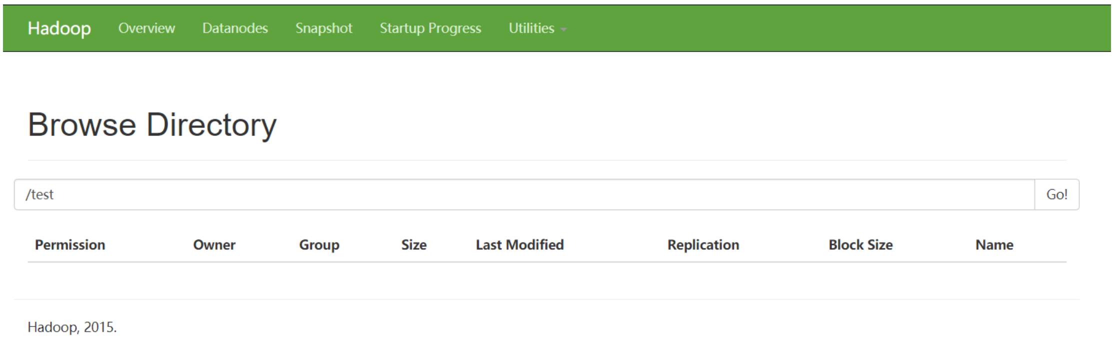
/test

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	0 B	2021/6/23上午12:14:00	1	128 MB	text5.txt

Hadoop, 2015.



```
bash-4.1# cd /usr/local/hadoop/bin
bash-4.1# if $(./hadoop fs -test -d /test);
> then $(./hadoop fs -touchz /test/text5.txt);
> else $(./hadoop fs -mkdir /test && ./hadoop fs -touchz /test/text5.txt);
> fi
21/06/22 12:13:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 12:13:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 12:13:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./hadoop fs -rm hdfs:///test/text5.txt
21/06/22 12:20:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 12:20:39 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted hdfs:///test/text5.txt
bash-4.1#
```



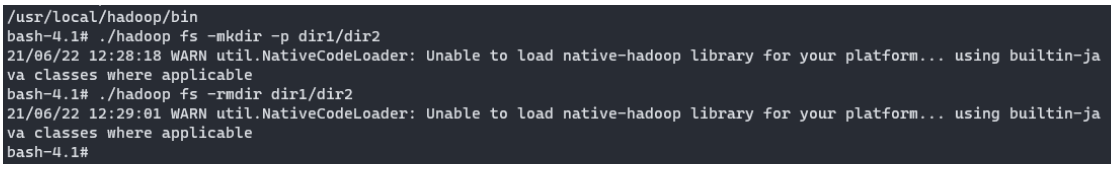
Browse Directory

/test

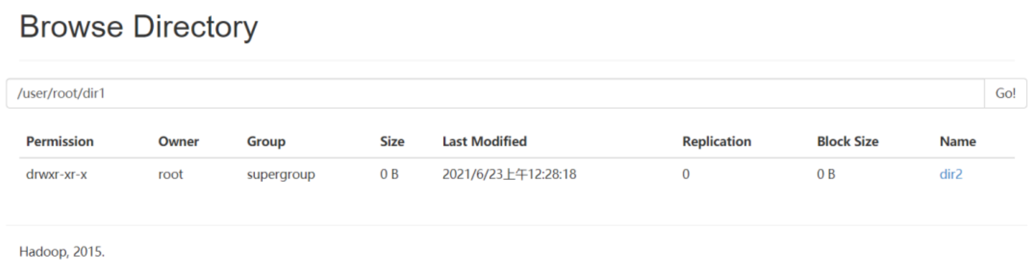
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name

Hadoop, 2015.

（7）提供一个 HDFS 的目录的路径，对该目录进行创建和删除操作，创建目录时，如果目录文件所在目录不存在则自动创建相应目录，删除目录时，由用户指定当该目录不为空时是否还删除该目录。



```
/usr/local/hadoop/bin
bash-4.1# ./hadoop fs -mkdir -p dir1/dir2
21/06/22 12:28:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./hadoop fs -rmdir dir1/dir2
21/06/22 12:29:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1#
```

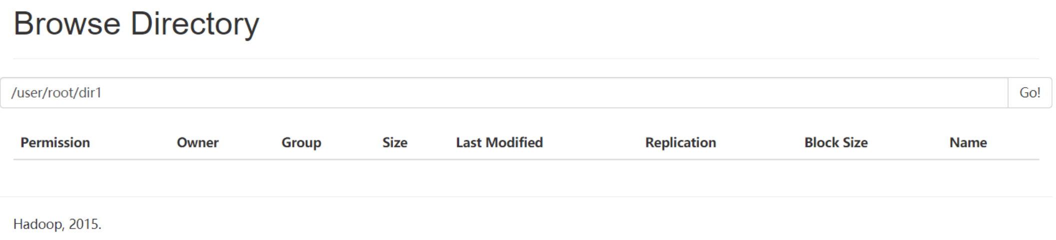


Browse Directory

/user/root/dir1

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	2021/6/23上午12:28:18	0	0 B	dir2

Hadoop, 2015.



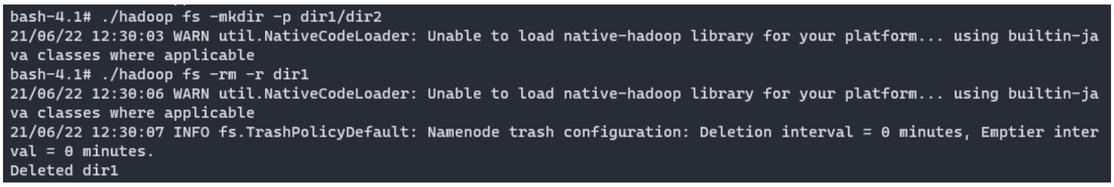
Browse Directory

/user/root/dir1 Go!

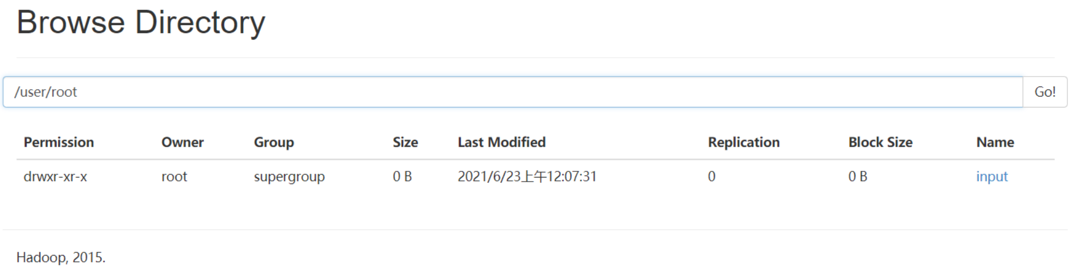
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name

Hadoop, 2015.

#rmdir 只能删除空目录，不能删除非空目录。



```
bash-4.1# ./hadoop fs -mkdir -p dir1/dir2
21/06/22 12:30:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./hadoop fs -rm -r dir1
21/06/22 12:30:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 12:30:07 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted dir1
```



Browse Directory

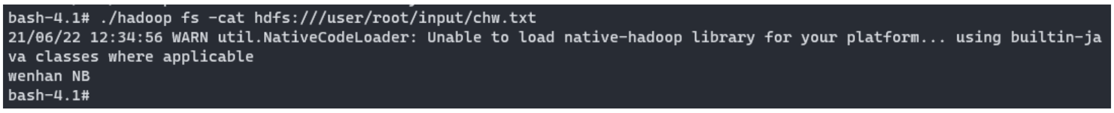
/user/root Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	2021/6/23上午12:07:31	0	0 B	input

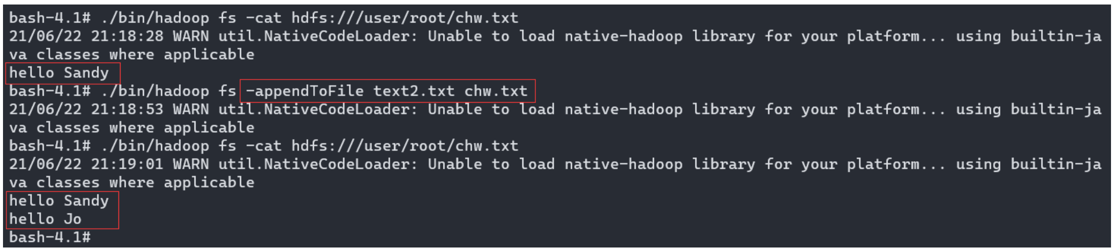
Hadoop, 2015.

dir1/dir2 全部被删除

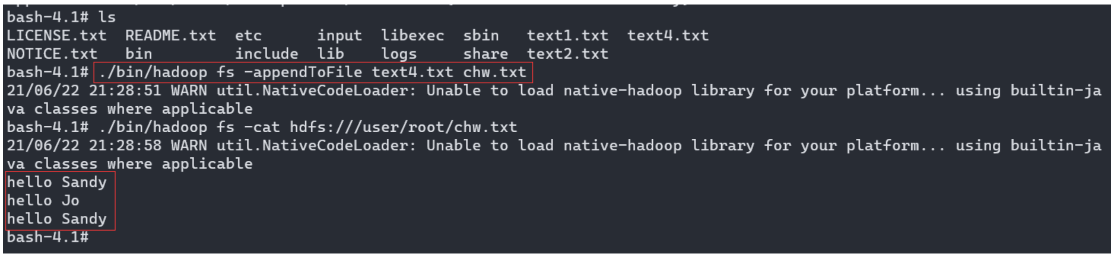
（8）向 HDFS 中指定的文件追加内容，由用户指定内容追加到原有文件的开头或结尾。



```
bash-4.1# ./hadoop fs -cat hdfs:///user/root/input/chw.txt
21/06/22 12:34:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
wenhan NB
bash-4.1#
```

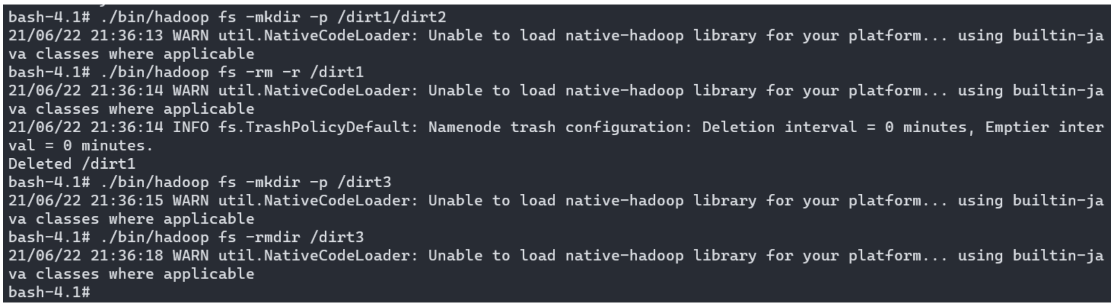


```
bash-4.1# ./bin/hadoop fs -cat hdfs:///user/root/chw.txt
21/06/22 21:18:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
hello Sandy
bash-4.1# ./bin/hadoop fs -appendToFile text2.txt chw.txt
21/06/22 21:18:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -cat hdfs:///user/root/chw.txt
21/06/22 21:19:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
hello Sandy
hello Jo
bash-4.1#
```



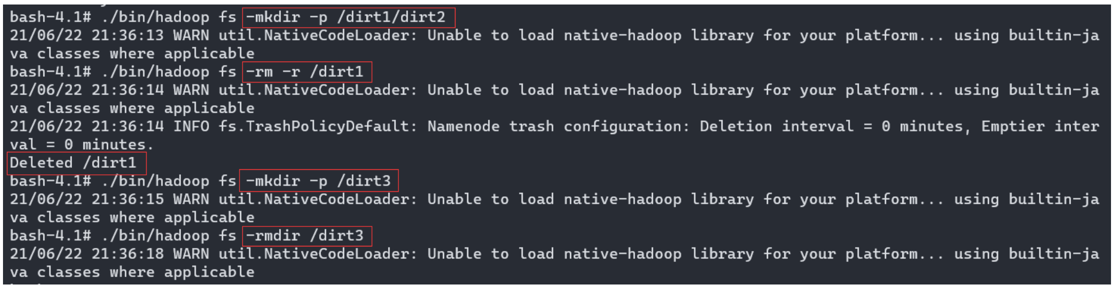
```
bash-4.1# ls
LICENSE.txt  README.txt  etc      input  libexec  sbin   text1.txt  text4.txt
NOTICE.txt   bin         include  lib    logs     share  text2.txt
bash-4.1# ./bin/hadoop fs -appendToFile text4.txt chw.txt
21/06/22 21:28:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -cat hdfs:///user/root/chw.txt
21/06/22 21:28:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
hello Sandy
hello Jo
hello Sandy
bash-4.1#
```

（9）删除 HDFS 中指定的文件



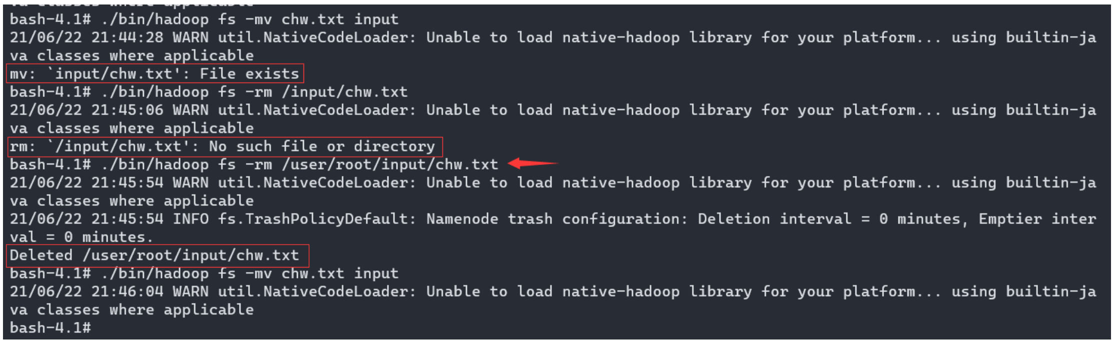
```
bash-4.1# ./bin/hadoop fs -mkdir -p /dirt1/dirt2
21/06/22 21:36:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -rm -r /dirt1
21/06/22 21:36:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 21:36:14 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted /dirt1
bash-4.1# ./bin/hadoop fs -mkdir -p /dirt3
21/06/22 21:36:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -rmdir /dirt3
21/06/22 21:36:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1#
```

（10）删除 HDFS 中指定的目录，由用户指定目录中如果存在文件时是否删除目录；



```
bash-4.1# ./bin/hadoop fs -mkdir -p /dirt1/dirt2
21/06/22 21:36:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -rm -r /dirt1
21/06/22 21:36:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 21:36:14 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted /dirt1
bash-4.1# ./bin/hadoop fs -mkdir -p /dirt3
21/06/22 21:36:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -rmdir /dirt3
21/06/22 21:36:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
```

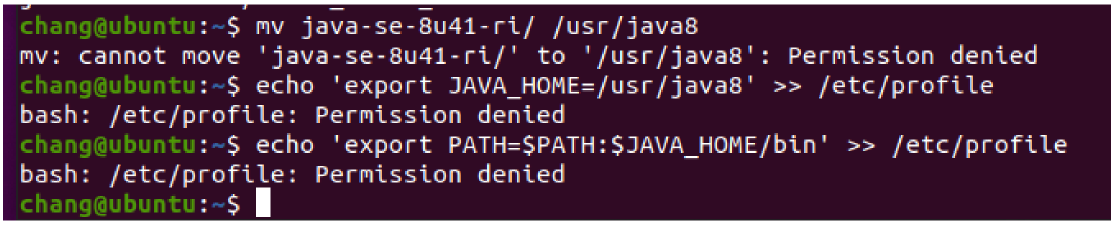
（11）在 HDFS 中将文件从源路径移动到目的路径。



```
bash-4.1# ./bin/hadoop fs -mv chw.txt input
21/06/22 21:44:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
mv: `input/chw.txt': File exists
bash-4.1# ./bin/hadoop fs -rm /input/chw.txt
21/06/22 21:45:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
rm: `/input/chw.txt': No such file or directory
bash-4.1# ./bin/hadoop fs -rm /user/root/input/chw.txt
21/06/22 21:45:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/22 21:45:54 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted /user/root/input/chw.txt
bash-4.1# ./bin/hadoop fs -mv chw.txt input
21/06/22 21:46:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1#
```

1.5 实验中的问题与解决方法

（1）配置 java 环境时，配置环境路径失败。

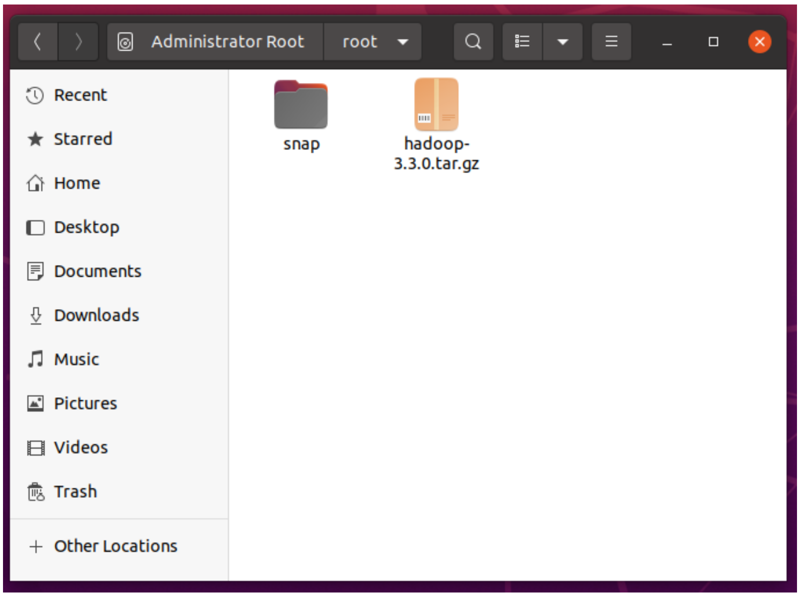


```
chang@ubuntu:~$ mv java-se-8u41-ri/ /usr/java8
mv: cannot move 'java-se-8u41-ri/' to '/usr/java8': Permission denied
chang@ubuntu:~$ echo 'export JAVA_HOME=/usr/java8' >> /etc/profile
bash: /etc/profile: Permission denied
chang@ubuntu:~$ echo 'export PATH=$PATH:$JAVA_HOME/bin' >> /etc/profile
bash: /etc/profile: Permission denied
chang@ubuntu:~$
```

原因：修改路径等信息时，需要 Ubuntu 管理员权限。

解决方法：使用 sudo su，输入系统密码，获得管理员权限进行修改。

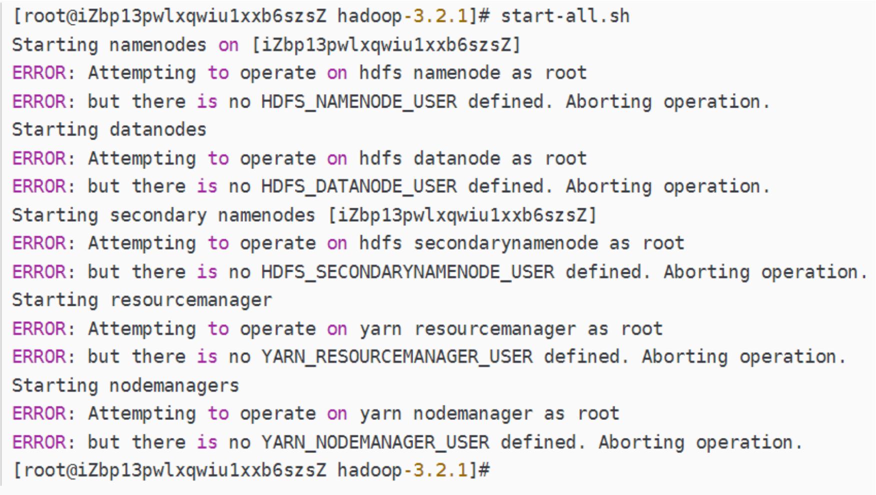
（2）下载 Hadoop 时，下载速度过慢。



原因：软件源设置到了 Ubuntu 默认软件源。

解决方法：可以使用国内镜像或者下载到 Windows，再上传到 Ubuntu 上。

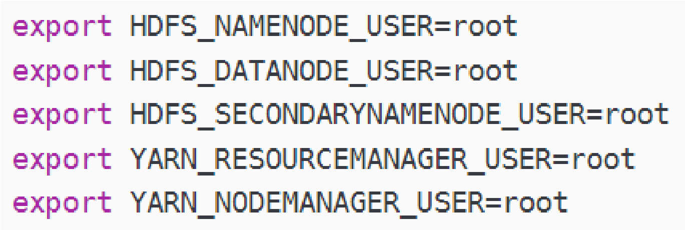
（3）执行 start-dfs.sh start-yarn.sh 两个文件会报错，例如：



```
[root@iZbp13pwlxqwiu1xxb6szsZ hadoop-3.2.1]# start-all.sh
Starting namenodes on [iZbp13pwlxqwiu1xxb6szsZ]
ERROR: Attempting to operate on hdfs namenode as root
ERROR: but there is no HDFS_NAMENODE_USER defined. Aborting operation.
Starting datanodes
ERROR: Attempting to operate on hdfs datanode as root
ERROR: but there is no HDFS_DATANODE_USER defined. Aborting operation.
Starting secondary namenodes [iZbp13pwlxqwiu1xxb6szsZ]
ERROR: Attempting to operate on hdfs secondarynamenode as root
ERROR: but there is no HDFS_SECONDARYNAMENODE_USER defined. Aborting operation.
Starting resourcemanager
ERROR: Attempting to operate on yarn resourcemanager as root
ERROR: but there is no YARN_RESOURCEMANAGER_USER defined. Aborting operation.
Starting nodemanagers
ERROR: Attempting to operate on yarn nodemanager as root
ERROR: but there is no YARN_NODEMANAGER_USER defined. Aborting operation.
[root@iZbp13pwlxqwiu1xxb6szsZ hadoop-3.2.1]#
```

原因：Hadoop 为不同的用户安装，你为不同的用户启动 yarn 服务，或者是在 Hadoop 配置的 hadoop-env.sh 中指定了 HDFS_NAMENODE_USER 但是 HDFS_DATANODE_USER 用户是别的东西。

解决方法：输入下列代码，配置用户为 root

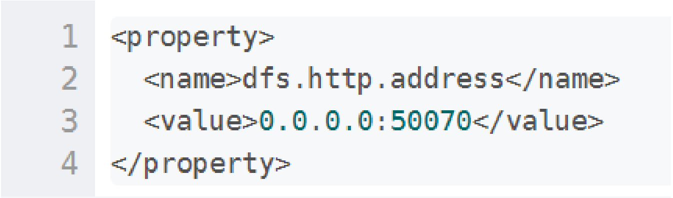


```
export HDFS_NAMENODE_USER=root
export HDFS_DATANODE_USER=root
export HDFS_SECONDARYNAMENODE_USER=root
export YARN_RESOURCEMANAGER_USER=root
export YARN_NODEMANAGER_USER=root
```

（4）安装后 8088 端口可以访问 50070 无法访问，防火墙开放后仍然无法访。

原因：Namenode 初始化默认端口失效，需要修改配置文件。

解决方法：手动修改配置文件设置默认端口，hdfs-site.xml 添加如下代码。



```
<property>
  <name>dfs.http.address</name>
  <value>0.0.0.0:50070</value>
</property>
```

第二章 熟悉常用的 HBase 操作

2.1 实验目的

1. 理解 HBase 在 Hadoop 体系结构中的角色；
2. 熟练使用 HBase 操作常用的 Shell 命令；
3. 熟悉 HBase 操作常用的 Java API（选做）。

2.2 实验平台

操作系统：Linux

Hadoop 版本：2.7.1 或以上版本

HBase 版本：1.1.2 或以上版本

JDK 版本：1.8 或以上版本

Java IDE：未安装桌面系统，使用 vim 编辑代码

2.3 实验内容和要求

1. 编程实现以下指定功能，并用Hadoop提供的HBase Shell命令完成相同任务：

   （1） 列出 HBase 所有的表的相关信息，例如表名；

   （2） 在终端打印出指定的表的所有记录数据；

   （3） 向已经创建好的表添加和删除指定的列族或列；

   （4） 清空指定的表的所有记录数据；

   （5） 统计表的行数。

2. 现有以下关系型数据库中的表和数据，要求将其转换为适合于 HBase 存储的表并插入数据：

学生表（Student）

学号 （S_No）	姓名 （S_Name）	性别 （S_Sex）	年龄 （S_Age）
2015001	Zhangsan	male	23
2015003	Mary	female	22
2015003	Lisi	male	24

课程表（Course）

课程号（C_No）	课程名（C_Name）	学分（C_Credit）
123001	Math	2.0
123002	Computer Science	5.0
123003	English	3.0

选课表（SC）

学号（SC_Sno）	课程号（SC_Cno）	成绩（SC_Score）
2015001	123001	86
2015001	123003	69
2015002	123002	77
2015002	123003	99
2015003	123001	98

2015003	123002	95

3. 利用 HBase 和 MapReduce 完成如下任务：

假设 HBase 有 2 张表，表的逻辑视图及部分数据如下所示：

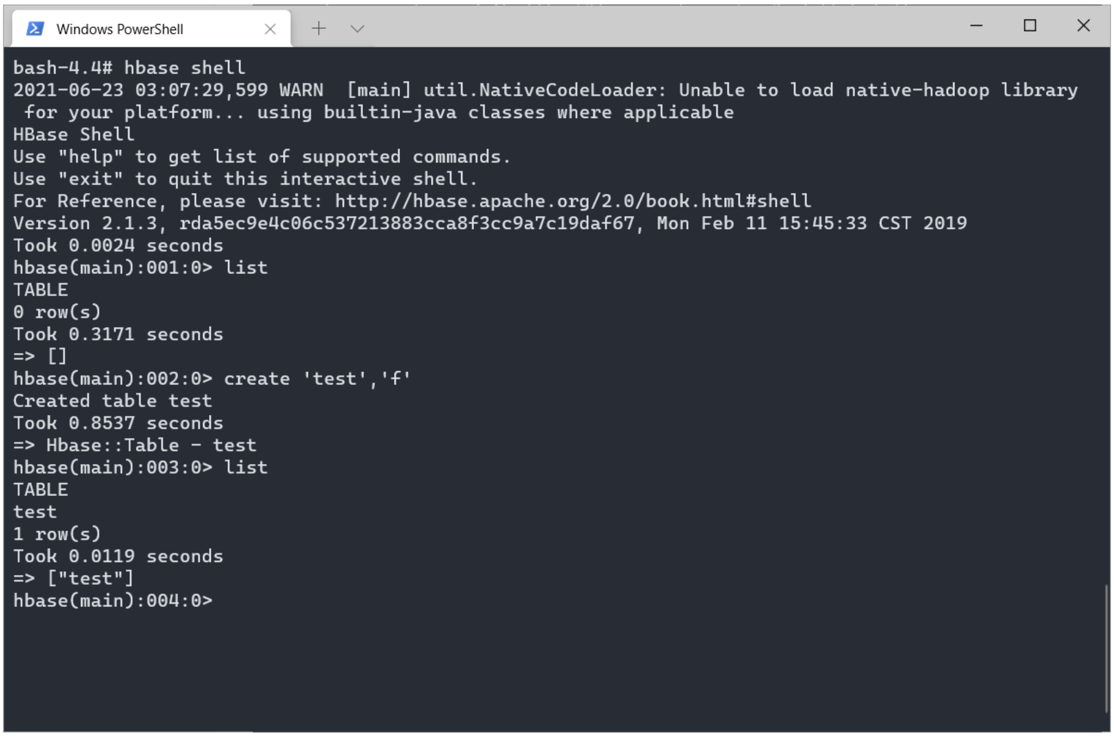
表 逻辑视图及部分数据

书名（bookName）	价格（price）
Database System Concept	30$
Thinking in Java	60$
Data Mining	25$

要求：从 HBase 读出上述两张表的数据，对“price”的排序，并将结果存储到 HBase 中。

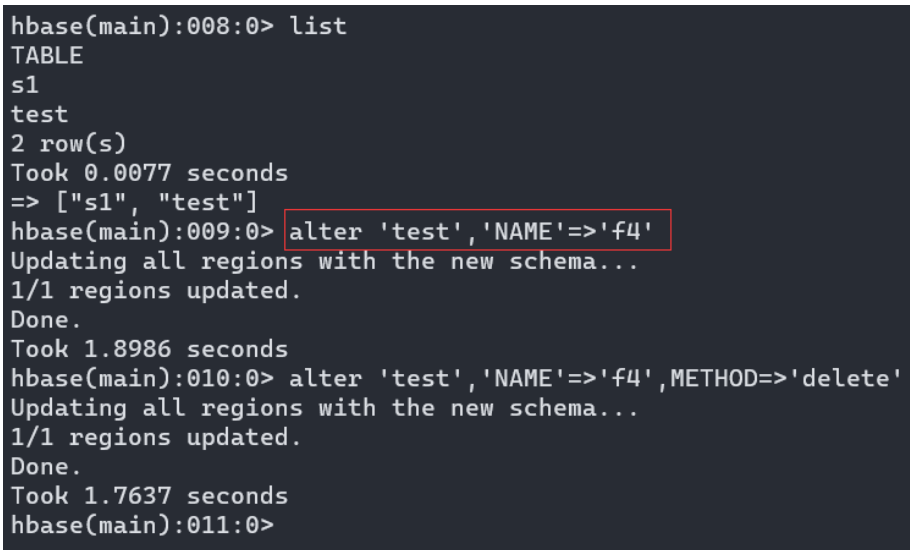
2. 4 实验过程

（1）创建一个表，并查看 Linux 上的表名称



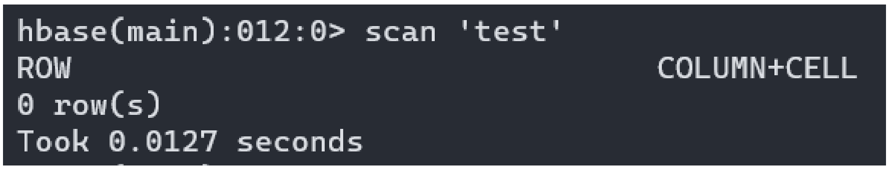
```
bash-4.4# hbase shell
2021-06-23 03:07:29,599 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library
 for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.1.3, rda5ec9e4c06c537213883cca8f3cc9a7c19daf67, Mon Feb 11 15:45:33 CST 2019
Took 0.0024 seconds
hbase(main):001:0> list
TABLE
0 row(s)
Took 0.3171 seconds
=> []
hbase(main):002:0> create 'test','f'
Created table test
Took 0.8537 seconds
=> Hbase::Table - test
hbase(main):003:0> list
TABLE
test
1 row(s)
Took 0.0119 seconds
=> ["test"]
hbase(main):004:0>
```

（2）添加信息，删除信息

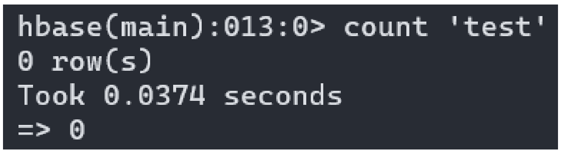


```
hbase(main):008:0> list
TABLE
s1
test
2 row(s)
Took 0.0077 seconds
=> ["s1", "test"]
hbase(main):009:0> alter 'test','NAME'=>'f4'
Updating all regions with the new schema...
1/1 regions updated.
Done.
Took 1.8986 seconds
hbase(main):010:0> alter 'test','NAME'=>'f4',METHOD=>'delete'
Updating all regions with the new schema...
1/1 regions updated.
Done.
Took 1.7637 seconds
hbase(main):011:0>
```

（3）扫描表，统计行数

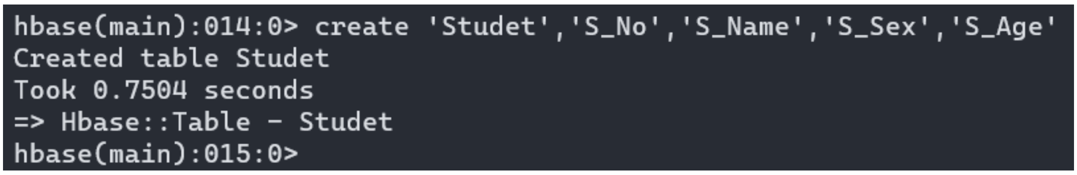


```
hbase(main):012:0> scan 'test'
ROW                                  COLUMN+CELL
0 row(s)
Took 0.0127 seconds
```

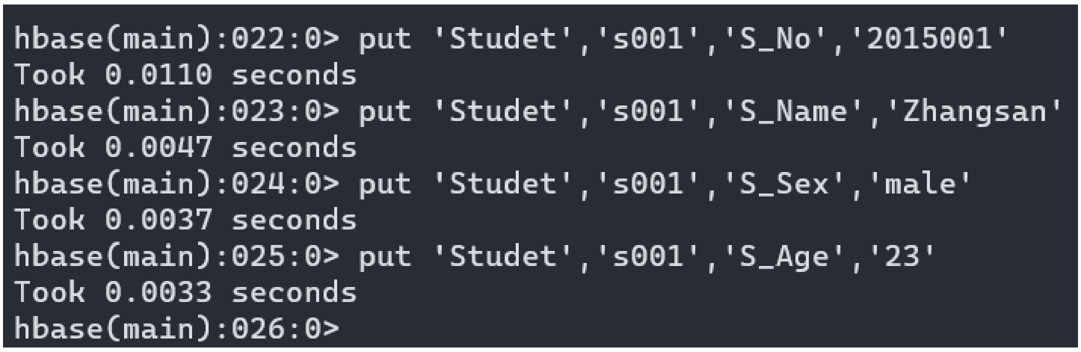


```
hbase(main):013:0> count 'test'
0 row(s)
Took 0.0374 seconds
=> 0
```

（4）创建 Student 表，并添加数据

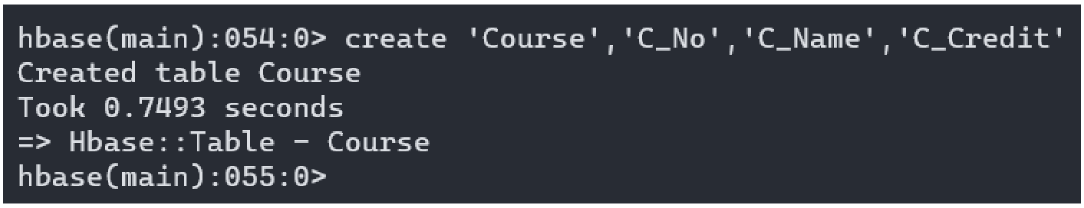


```
hbase(main):014:0> create 'Studet','S_No','S_Name','S_Sex','S_Age'
Created table Studet
Took 0.7504 seconds
=> Hbase::Table - Studet
hbase(main):015:0>
```

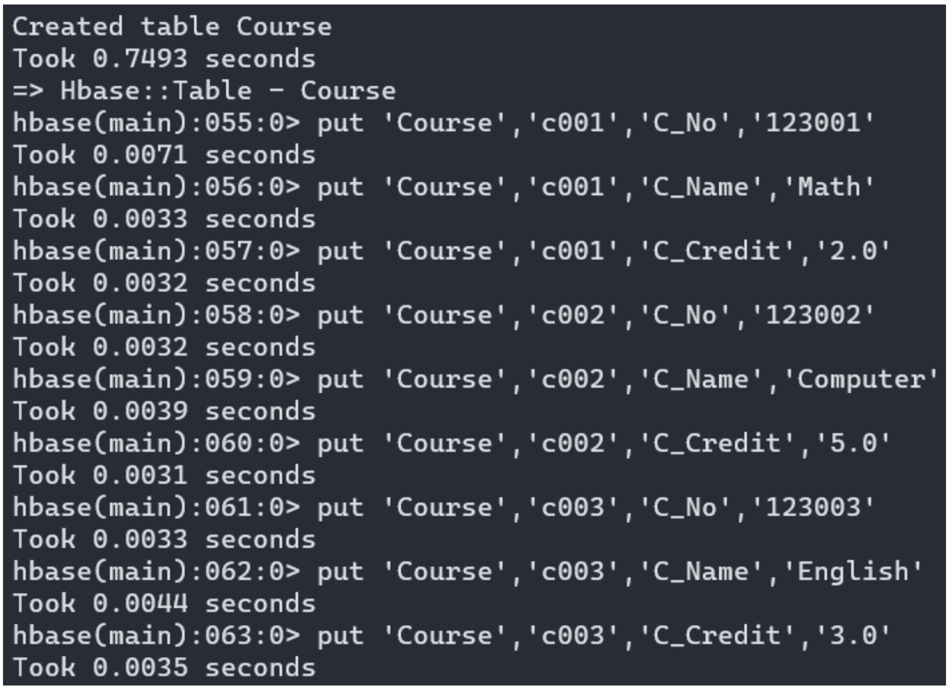


```
hbase(main):022:0> put 'Studet','s001','S_No','2015001'
Took 0.0110 seconds
hbase(main):023:0> put 'Studet','s001','S_Name','Zhangsan'
Took 0.0047 seconds
hbase(main):024:0> put 'Studet','s001','S_Sex','male'
Took 0.0037 seconds
hbase(main):025:0> put 'Studet','s001','S_Age','23'
Took 0.0033 seconds
hbase(main):026:0>
```

（5）创建 Course 表，添加课程

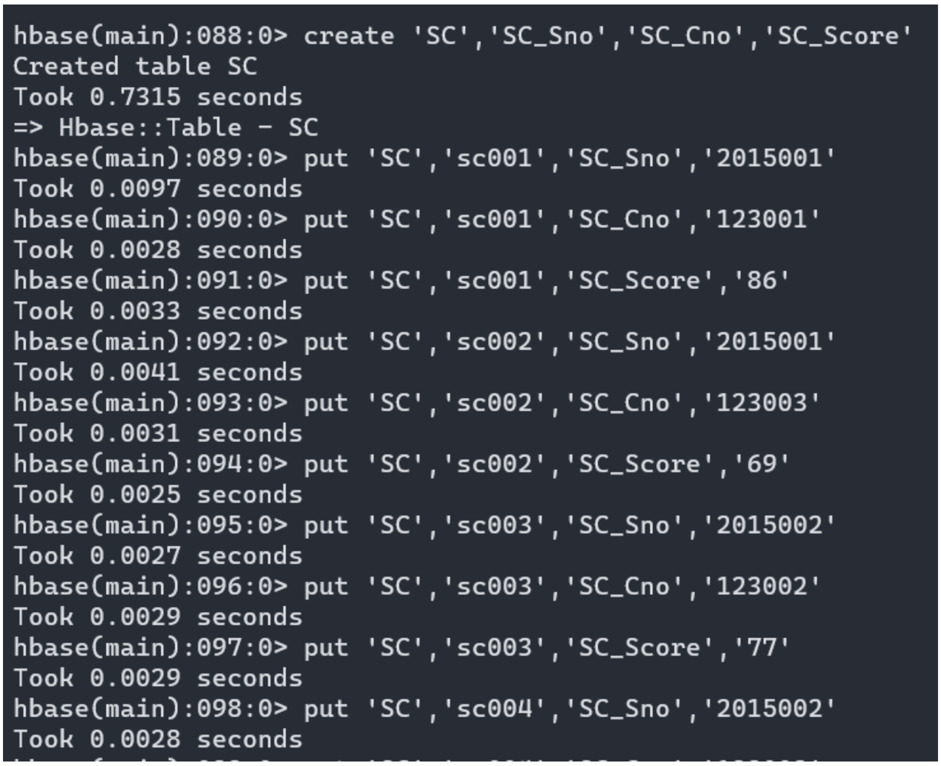


```
hbase(main):054:0> create 'Course','C_No','C_Name','C_Credit'
Created table Course
Took 0.7493 seconds
=> Hbase::Table - Course
hbase(main):055:0>
```



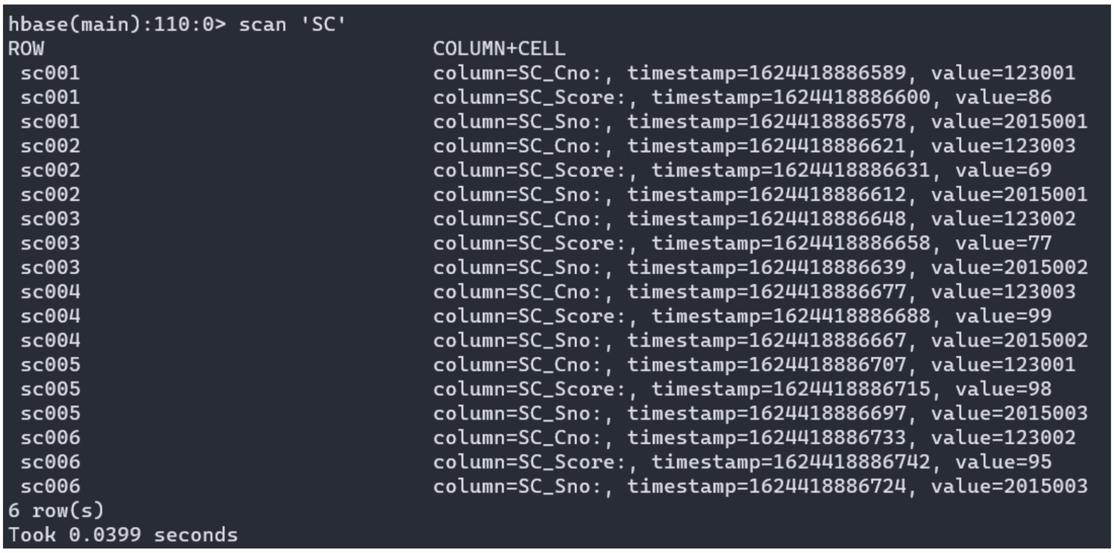
```
Created table Course
Took 0.7493 seconds
=> Hbase::Table - Course
hbase(main):055:0> put 'Course','c001','C_No','123001'
Took 0.0071 seconds
hbase(main):056:0> put 'Course','c001','C_Name','Math'
Took 0.0033 seconds
hbase(main):057:0> put 'Course','c001','C_Credit','2.0'
Took 0.0032 seconds
hbase(main):058:0> put 'Course','c002','C_No','123002'
Took 0.0032 seconds
hbase(main):059:0> put 'Course','c002','C_Name','Computer'
Took 0.0039 seconds
hbase(main):060:0> put 'Course','c002','C_Credit','5.0'
Took 0.0031 seconds
hbase(main):061:0> put 'Course','c003','C_No','123003'
Took 0.0033 seconds
hbase(main):062:0> put 'Course','c003','C_Name','English'
Took 0.0044 seconds
hbase(main):063:0> put 'Course','c003','C_Credit','3.0'
Took 0.0035 seconds
```

（6）创建 SC 表，添加信息



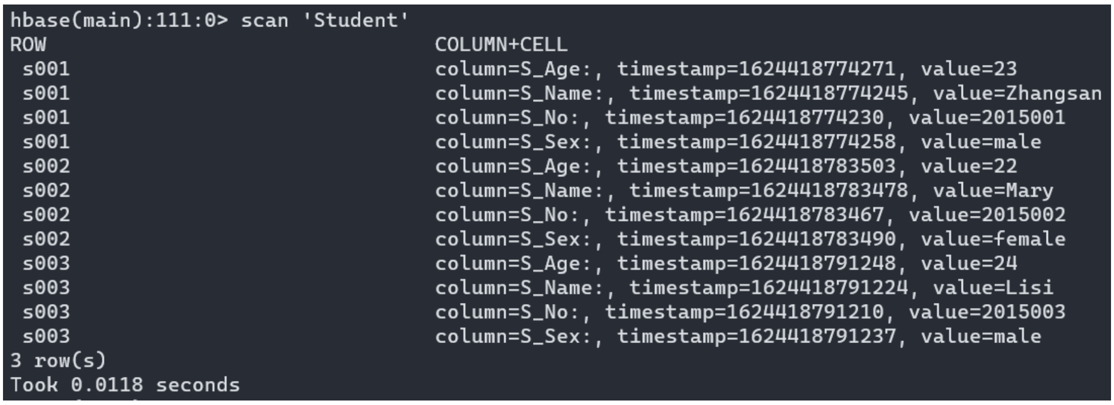
```
hbase(main):088:0> create 'SC','SC_Sno','SC_Cno','SC_Score'
Created table SC
Took 0.7315 seconds
=> Hbase::Table - SC
hbase(main):089:0> put 'SC','sc001','SC_Sno','2015001'
Took 0.0097 seconds
hbase(main):090:0> put 'SC','sc001','SC_Cno','123001'
Took 0.0028 seconds
hbase(main):091:0> put 'SC','sc001','SC_Score','86'
Took 0.0033 seconds
hbase(main):092:0> put 'SC','sc002','SC_Sno','2015001'
Took 0.0041 seconds
hbase(main):093:0> put 'SC','sc002','SC_Cno','123003'
Took 0.0031 seconds
hbase(main):094:0> put 'SC','sc002','SC_Score','69'
Took 0.0025 seconds
hbase(main):095:0> put 'SC','sc003','SC_Sno','2015002'
Took 0.0027 seconds
hbase(main):096:0> put 'SC','sc003','SC_Cno','123002'
Took 0.0029 seconds
hbase(main):097:0> put 'SC','sc003','SC_Score','77'
Took 0.0029 seconds
hbase(main):098:0> put 'SC','sc004','SC_Sno','2015002'
Took 0.0028 seconds
```

（7）打印 SC 表



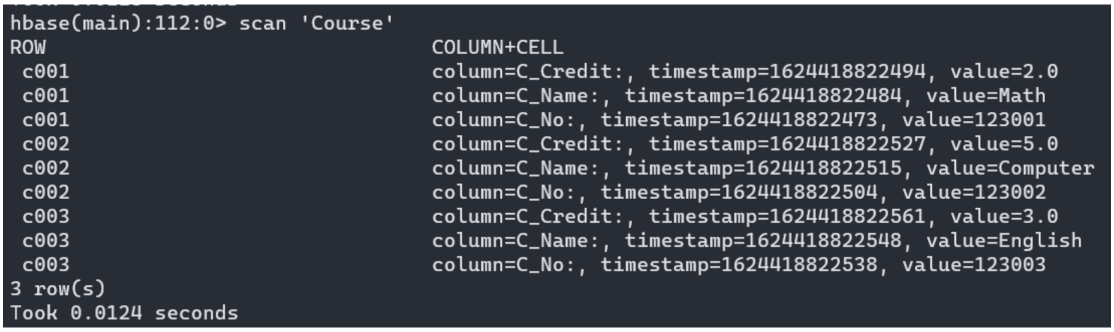
```
hbase(main):110:0> scan 'SC'
ROW                                COLUMN+CELL
 sc001                             column=SC_Cno:, timestamp=1624418886589, value=123001
 sc001                             column=SC_Score:, timestamp=1624418886600, value=86
 sc001                             column=SC_Sno:, timestamp=1624418886578, value=2015001
 sc002                             column=SC_Cno:, timestamp=1624418886621, value=123003
 sc002                             column=SC_Score:, timestamp=1624418886631, value=69
 sc002                             column=SC_Sno:, timestamp=1624418886612, value=2015001
 sc003                             column=SC_Cno:, timestamp=1624418886648, value=123002
 sc003                             column=SC_Score:, timestamp=1624418886658, value=77
 sc003                             column=SC_Sno:, timestamp=1624418886639, value=2015002
 sc004                             column=SC_Cno:, timestamp=1624418886677, value=123003
 sc004                             column=SC_Score:, timestamp=1624418886688, value=99
 sc004                             column=SC_Sno:, timestamp=1624418886667, value=2015002
 sc005                             column=SC_Cno:, timestamp=1624418886707, value=123001
 sc005                             column=SC_Score:, timestamp=1624418886715, value=98
 sc005                             column=SC_Sno:, timestamp=1624418886697, value=2015003
 sc006                             column=SC_Cno:, timestamp=1624418886733, value=123002
 sc006                             column=SC_Score:, timestamp=1624418886742, value=95
 sc006                             column=SC_Sno:, timestamp=1624418886724, value=2015003
6 row(s)
Took 0.0399 seconds
```

（8）打印 Student 表



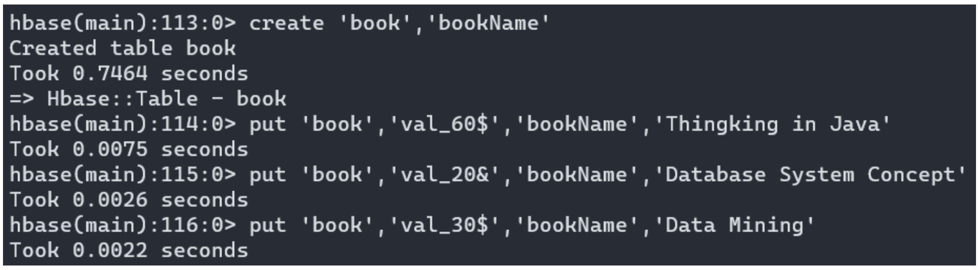
```
hbase(main):111:0> scan 'Student'
ROW                                COLUMN+CELL
 s001                              column=S_Age:, timestamp=1624418774271, value=23
 s001                              column=S_Name:, timestamp=1624418774245, value=Zhangsan
 s001                              column=S_No:, timestamp=1624418774230, value=2015001
 s001                              column=S_Sex:, timestamp=1624418774258, value=male
 s002                              column=S_Age:, timestamp=1624418783503, value=22
 s002                              column=S_Name:, timestamp=1624418783478, value=Mary
 s002                              column=S_No:, timestamp=1624418783467, value=2015002
 s002                              column=S_Sex:, timestamp=1624418783490, value=female
 s003                              column=S_Age:, timestamp=1624418791248, value=24
 s003                              column=S_Name:, timestamp=1624418791224, value=Lisi
 s003                              column=S_No:, timestamp=1624418791210, value=2015003
 s003                              column=S_Sex:, timestamp=1624418791237, value=male
3 row(s)
Took 0.0118 seconds
```

（9）打印 Course 表

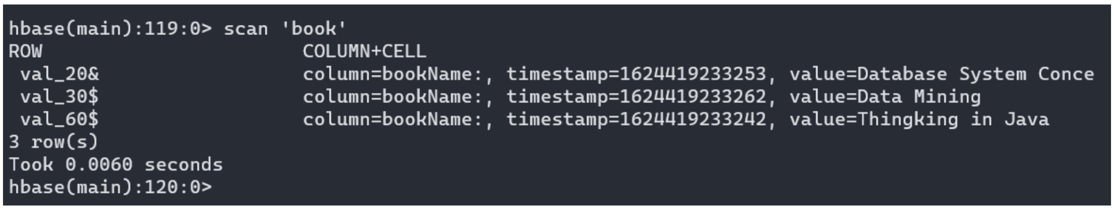


```
hbase(main):112:0> scan 'Course'
ROW                                COLUMN+CELL
 c001                              column=C_Credit:, timestamp=1624418822494, value=2.0
 c001                              column=C_Name:, timestamp=1624418822484, value=Math
 c001                              column=C_No:, timestamp=1624418822473, value=123001
 c002                              column=C_Credit:, timestamp=1624418822527, value=5.0
 c002                              column=C_Name:, timestamp=1624418822515, value=Computer
 c002                              column=C_No:, timestamp=1624418822504, value=123002
 c003                              column=C_Credit:, timestamp=1624418822561, value=3.0
 c003                              column=C_Name:, timestamp=1624418822548, value=English
 c003                              column=C_No:, timestamp=1624418822538, value=123003
3 row(s)
Took 0.0124 seconds
```

（10）创建书籍表，直接打印，可以根据设置的 value 自动排序



```
hbase(main):113:0> create 'book','bookName'
Created table book
Took 0.7464 seconds
=> Hbase::Table - book
hbase(main):114:0> put 'book','val_60$','bookName','Thingking in Java'
Took 0.0075 seconds
hbase(main):115:0> put 'book','val_20&','bookName','Database System Concept'
Took 0.0026 seconds
hbase(main):116:0> put 'book','val_30$','bookName','Data Mining'
Took 0.0022 seconds
```



```
hbase(main):119:0> scan 'book'
ROW                         COLUMN+CELL
 val_20&                    column=bookName:, timestamp=1624419233253, value=Database System Conce
 val_30$                    column=bookName:, timestamp=1624419233262, value=Data Mining
 val_60$                    column=bookName:, timestamp=1624419233242, value=Thingking in Java
3 row(s)
Took 0.0060 seconds
hbase(main):120:0>
```

2.5 实验中的问题与解决方法

（1）启动 Docker，运行 start-hbase 失败

原因：在启动 Docker 时，封装好的 Hbase 直接运行了起来，所以不需要手动启动。

解决方法：Hbase 已运行，可以直接操作。

第三章 MapReduce 编程初级实践

3.1 实验目的

1.通过实验掌握基本的 MapReduce 编程方法；

2.掌握用 MapReduce 解决一些常见的数据处理问题，包括数据去重、数据排序和数据挖掘等。

3.2 实验平台

已经配置完成的 Hadoop 伪分布式环境。

3.3 实验内容和要求

1.编程实现文件合并和去重操作

对于两个输入文件，即文件 A 和文件 B，请编写 MapReduce 程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新的输出文件 C。下面是输入文件和输出文件的一个样例供参考。

输入文件 A 的样例如下：

```
20150101     x

20150102     y
```

```
20150103     x

20150104     y

20150105     z

 20150106     x
```

输入文件 B 的样例如下：

```
20150101       y

20150102       y

20150103       x

20150104       z

20150105       y
```

根据输入文件 A 和 B 合并得到的输出文件 C 的样例如下：

```
 20150101       x

 20150101       y

 20150102       y

 20150103       x

 20150104       y

 20150104       z

 20150105       y

20150105       z

 20150106       x
```

2. 编写程序实现对输入文件的排序

现在有多个输入文件，每个文件中的每行内容均为一个整数。要求读取所有文件中的整数，进行升序排序后，输出到一个新的文件中，输出的数据格式为每行两个整数，第一个数字为第二个整数的排序位次，第二个整数为原待排列的整数。下面是输入文件和输出文件的一个样例供参考。

输入文件 1 的样例如下：

```
33
37
12
40
```

输入文件 2 的样例如下：

```
4
16
39
5
```

输入文件 3 的样例如下：

```
1
45
25
```

根据输入文件 1、2 和 3 得到的输出文件如下：

```
1 1
2 4
3 5
```

```
4 12

5 16

6 25

7 33

8 37

9 39

10 40

11 45
```

3. 对给定的表格进行信息挖掘

下面给出一个 child-parent 的表格，要求挖掘其中的父子辈关系，给出祖孙辈关系的表格。

输入文件内容如下：

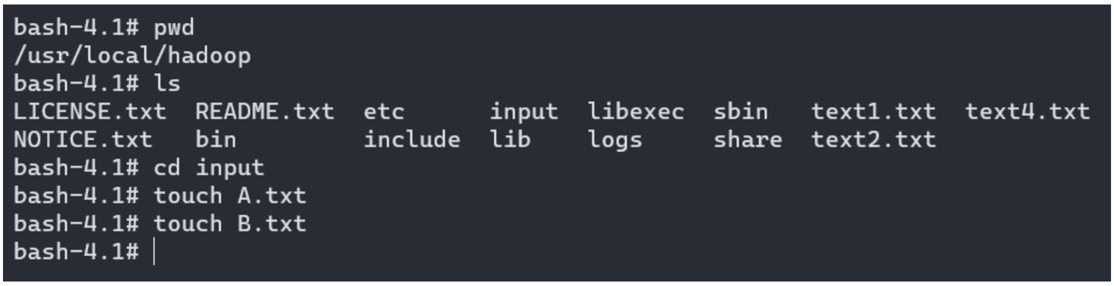
```
child           parent

Steven          Lucy

Steven          Jack

Jone            Lucy

Jone            Jack

Lucy            Mary

Lucy            Frank

Jack            Alice

Jack            Jesse

David           Alice

David           Jesse
```

```
 Philip      David

 Philip      Alma

 Mark       David

Mark       Alma
```

输出文件内容如下：

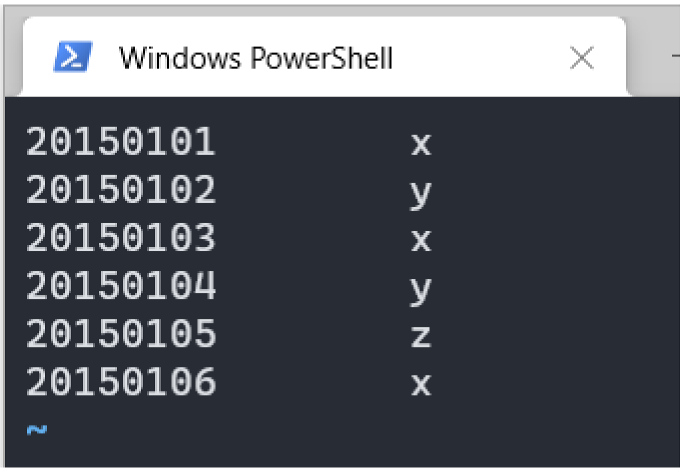
```
 grandchild      grandparent

 Steven          Alice

 Steven          Jesse

 Jone            Alice

 Jone            Jesse

 Steven          Mary

 Steven          Frank

 Jone            Mary

 Jone            Frank

 Philip          Alice

 Philip          Jesse

 Mark            Alice

Mark            Jesse
```

3.4 实验过程

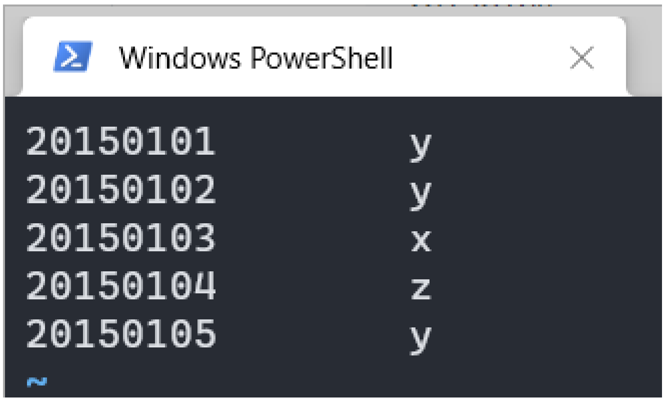


```
bash-4.1# pwd
/usr/local/hadoop
bash-4.1# ls
LICENSE.txt  README.txt  etc      input  libexec  sbin   text1.txt  text4.txt
NOTICE.txt   bin         include  lib    logs     share  text2.txt
bash-4.1# cd input
bash-4.1# touch A.txt
bash-4.1# touch B.txt
bash-4.1# 
```

（2）创建文本并输入信息

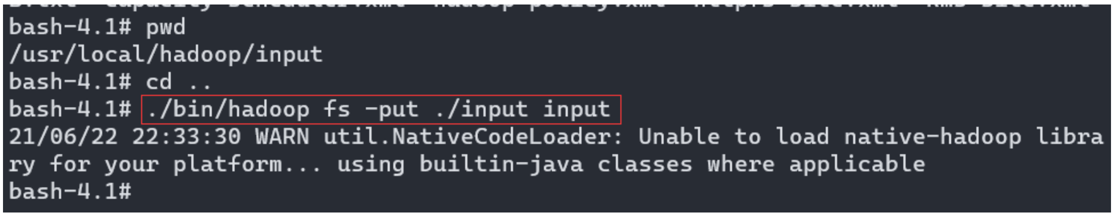


```
20150101        x
20150102        y
20150103        x
20150104        y
20150105        z
20150106        x
~
```

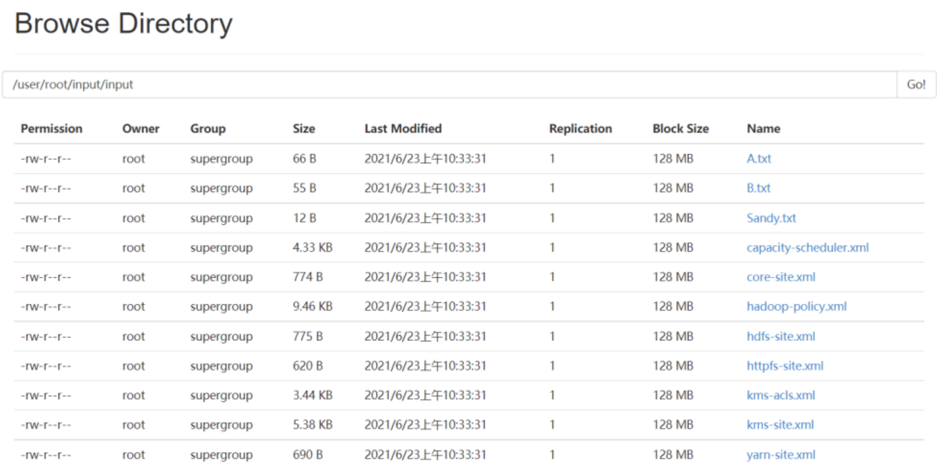


```
20150101        y
20150102        y
20150103        x
20150104        z
20150105        y
~
```

（3）上传文件



```
bash-4.1# pwd
/usr/local/hadoop/input
bash-4.1# cd ..
bash-4.1# ./bin/hadoop fs -put ./input input
21/06/22 22:33:30 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
bash-4.1#
```

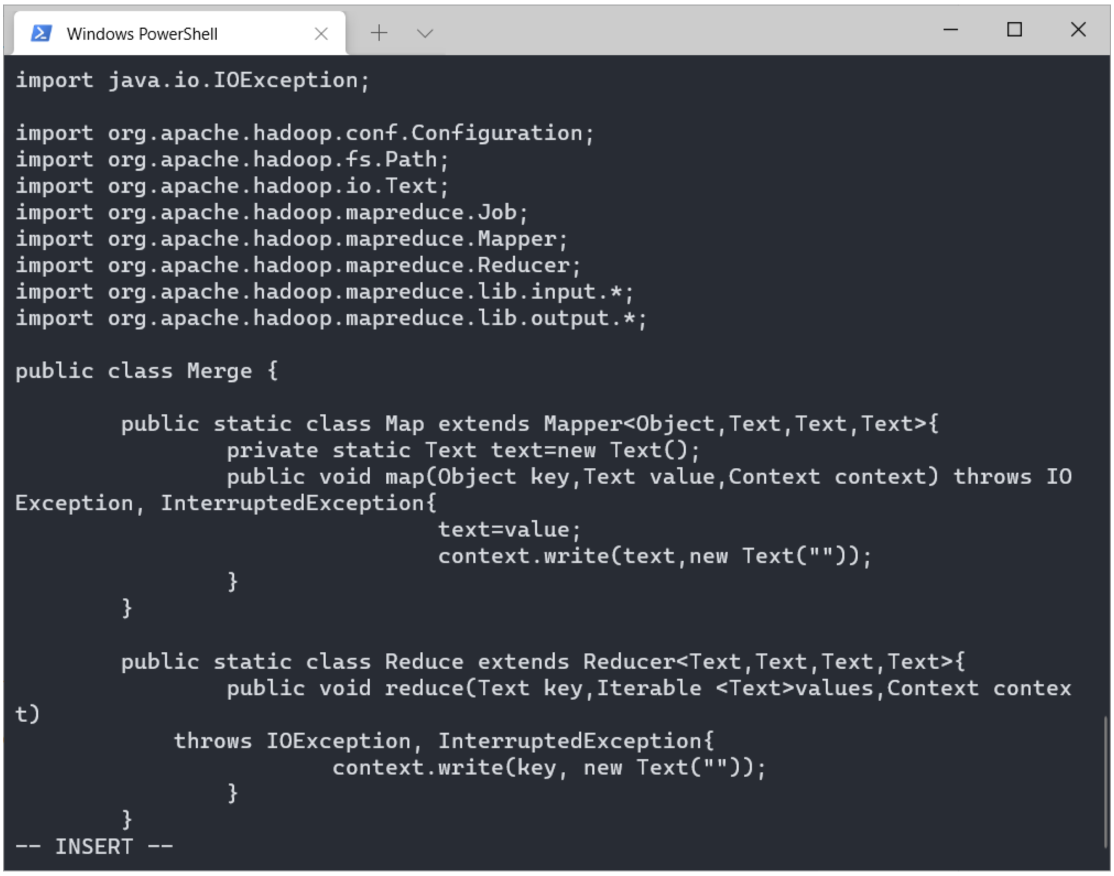


Browse Directory

/user/root/input/input

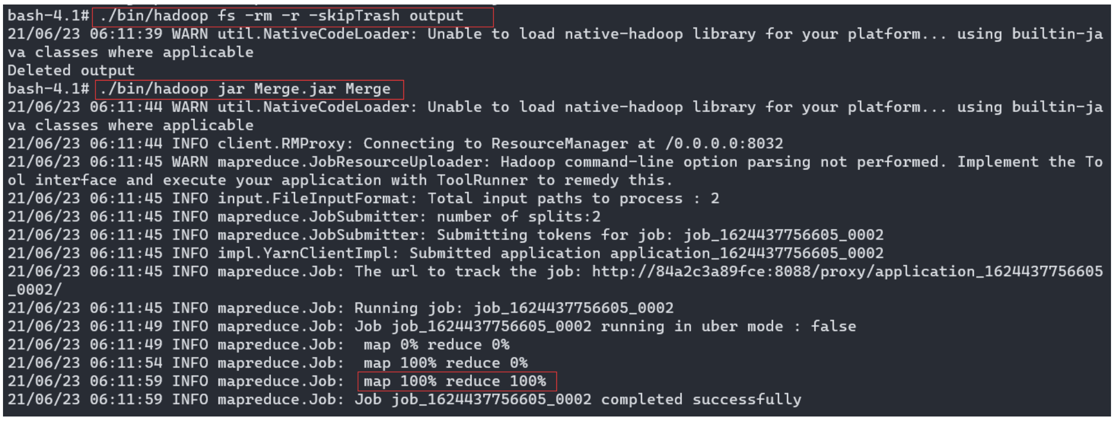
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	66 B	2021/6/23上午10:33:31	1	128 MB	A.txt
-rw-r--r--	root	supergroup	55 B	2021/6/23上午10:33:31	1	128 MB	B.txt
-rw-r--r--	root	supergroup	12 B	2021/6/23上午10:33:31	1	128 MB	Sandy.txt
-rw-r--r--	root	supergroup	4.33 KB	2021/6/23上午10:33:31	1	128 MB	capacity-scheduler.xml
-rw-r--r--	root	supergroup	774 B	2021/6/23上午10:33:31	1	128 MB	core-site.xml
-rw-r--r--	root	supergroup	9.46 KB	2021/6/23上午10:33:31	1	128 MB	hadoop-policy.xml
-rw-r--r--	root	supergroup	775 B	2021/6/23上午10:33:31	1	128 MB	hdfs-site.xml
-rw-r--r--	root	supergroup	620 B	2021/6/23上午10:33:31	1	128 MB	httpfs-site.xml
-rw-r--r--	root	supergroup	3.44 KB	2021/6/23上午10:33:31	1	128 MB	kms-acls.xml
-rw-r--r--	root	supergroup	5.38 KB	2021/6/23上午10:33:31	1	128 MB	kms-site.xml
-rw-r--r--	root	supergroup	690 B	2021/6/23上午10:33:31	1	128 MB	yarn-site.xml

（4）编辑代码



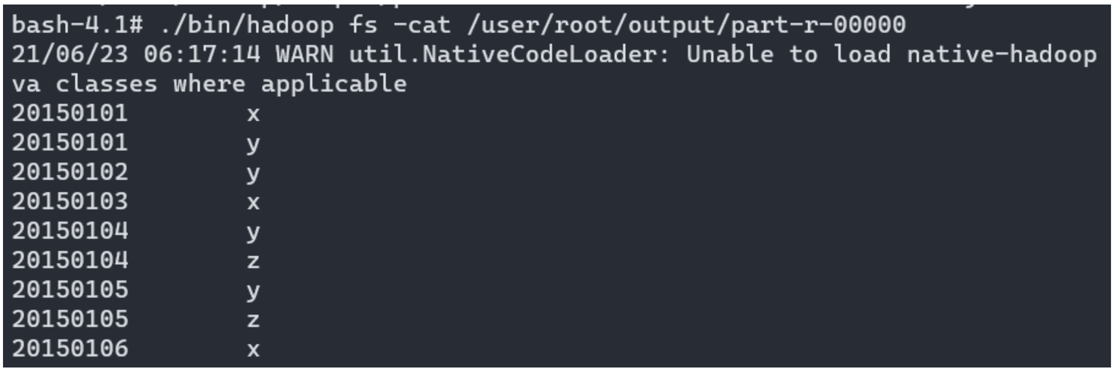
```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.*;
import org.apache.hadoop.mapreduce.lib.output.*;

public class Merge {

        public static class Map extends Mapper<Object,Text,Text,Text>{
                private static Text text=new Text();
                public void map(Object key,Text value,Context context) throws IO
Exception, InterruptedException{
                                text=value;
                                context.write(text,new Text(""));
                }
        }

        public static class Reduce extends Reducer<Text,Text,Text,Text>{
                public void reduce(Text key,Iterable <Text>values,Context contex
t)
            throws IOException, InterruptedException{
                        context.write(key, new Text(""));
                }
        }
-- INSERT --
```

（5）编译 java 代码，运行 jar 包



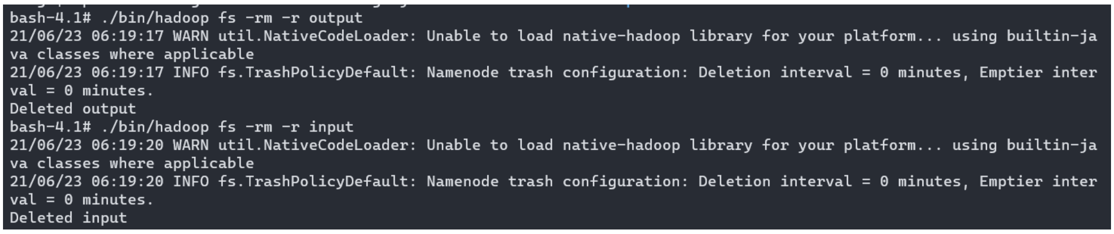
```
bash-4.1# ./bin/hadoop fs -rm -r -skipTrash output
21/06/23 06:11:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
Deleted output
bash-4.1# ./bin/hadoop jar Merge.jar Merge
21/06/23 06:11:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:11:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/06/23 06:11:45 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the To
ol interface and execute your application with ToolRunner to remedy this.
21/06/23 06:11:45 INFO input.FileInputFormat: Total input paths to process : 2
21/06/23 06:11:45 INFO mapreduce.JobSubmitter: number of splits:2
21/06/23 06:11:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624437756605_0002
21/06/23 06:11:45 INFO impl.YarnClientImpl: Submitted application application_1624437756605_0002
21/06/23 06:11:45 INFO mapreduce.Job: The url to track the job: http://84a2c3a89fce:8088/proxy/application_1624437756605
_0002/
21/06/23 06:11:45 INFO mapreduce.Job: Running job: job_1624437756605_0002
21/06/23 06:11:49 INFO mapreduce.Job: Job job_1624437756605_0002 running in uber mode : false
21/06/23 06:11:49 INFO mapreduce.Job:  map 0% reduce 0%
21/06/23 06:11:54 INFO mapreduce.Job:  map 100% reduce 0%
21/06/23 06:11:59 INFO mapreduce.Job:  map 100% reduce 100%
21/06/23 06:11:59 INFO mapreduce.Job: Job job_1624437756605_0002 completed successfully
```

（6）得到结果



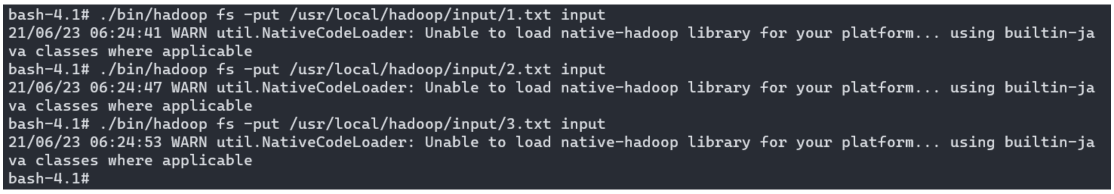
```
bash-4.1# ./bin/hadoop fs -cat /user/root/output/part-r-00000
21/06/23 06:17:14 WARN util.NativeCodeLoader: Unable to load native-hadoop
va classes where applicable
20150101        x
20150101        y
20150102        y
20150103        x
20150104        y
20150104        z
20150105        y
20150105        z
20150106        x
```

（7）做第二个实验，先删除 input output



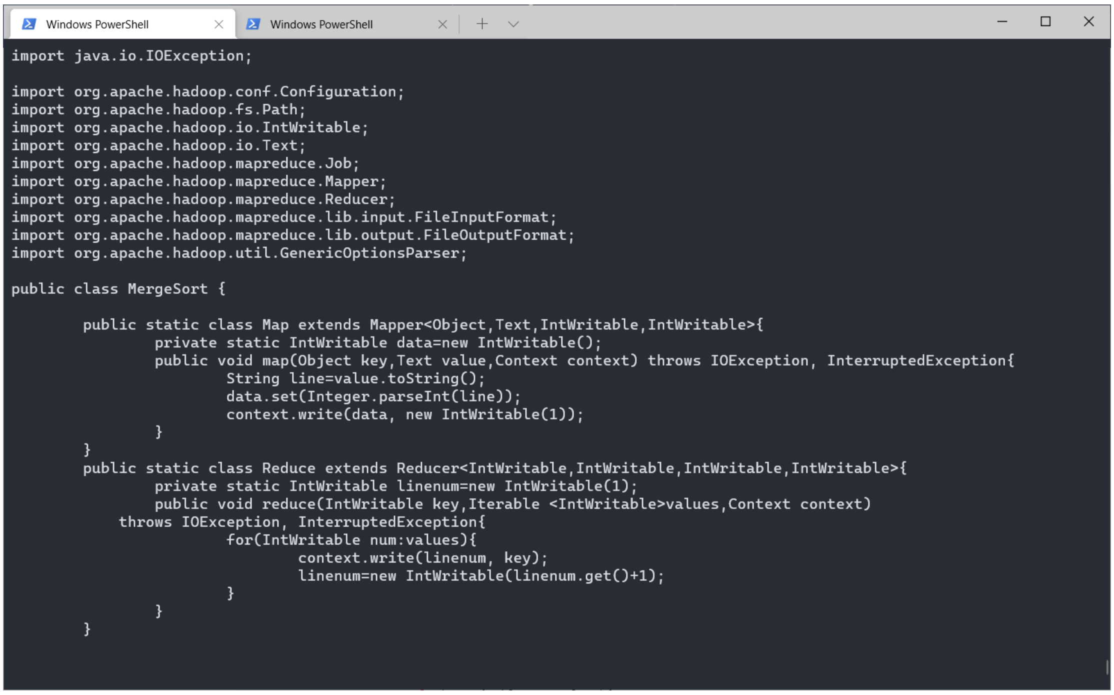
```
bash-4.1# ./bin/hadoop fs -rm -r output
21/06/23 06:19:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:19:17 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted output
bash-4.1# ./bin/hadoop fs -rm -r input
21/06/23 06:19:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:19:20 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted input
```

（8）创建文件，并将其上传到 HDFS



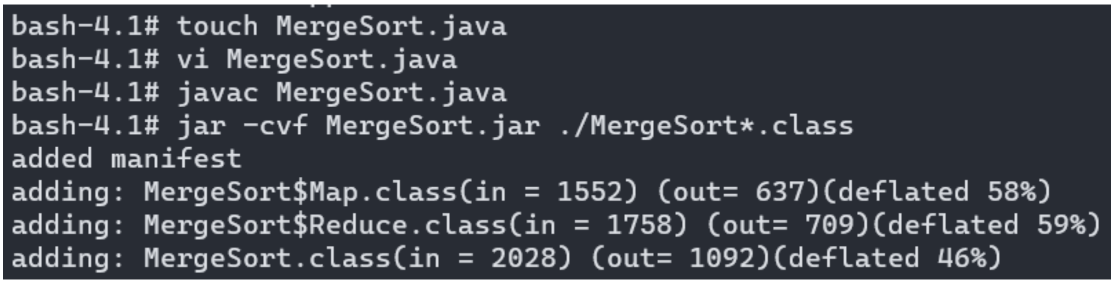
```
bash-4.1# ./bin/hadoop fs -put /usr/local/hadoop/input/1.txt input
21/06/23 06:24:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -put /usr/local/hadoop/input/2.txt input
21/06/23 06:24:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1# ./bin/hadoop fs -put /usr/local/hadoop/input/3.txt input
21/06/23 06:24:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
bash-4.1#
```

（9）编写代码



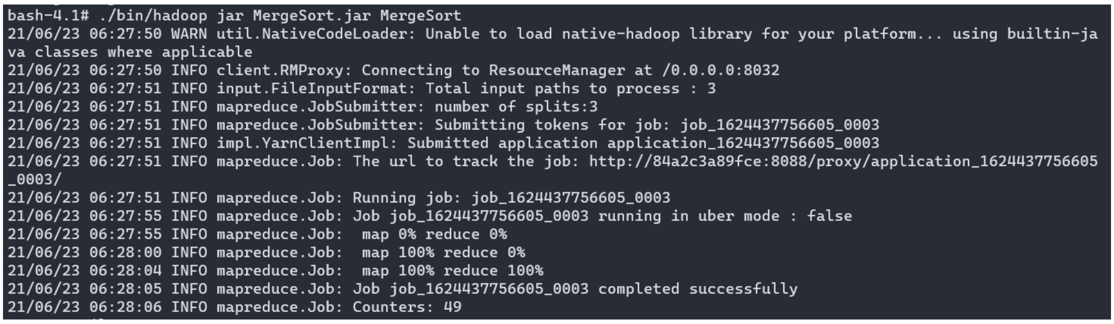
```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class MergeSort {

        public static class Map extends Mapper<Object,Text,IntWritable,IntWritable>{
                private static IntWritable data=new IntWritable();
                public void map(Object key,Text value,Context context) throws IOException, InterruptedException{
                        String line=value.toString();
                        data.set(Integer.parseInt(line));
                        context.write(data, new IntWritable(1));
                }
        }
        public static class Reduce extends Reducer<IntWritable,IntWritable,IntWritable,IntWritable>{
                private static IntWritable linenum=new IntWritable(1);
                public void reduce(IntWritable key,Iterable <IntWritable>values,Context context)
            throws IOException, InterruptedException{
                        for(IntWritable num:values){
                                context.write(linenum, key);
                                linenum=new IntWritable(linenum.get()+1);
                        }
                }
        }
```

（10）编译代码



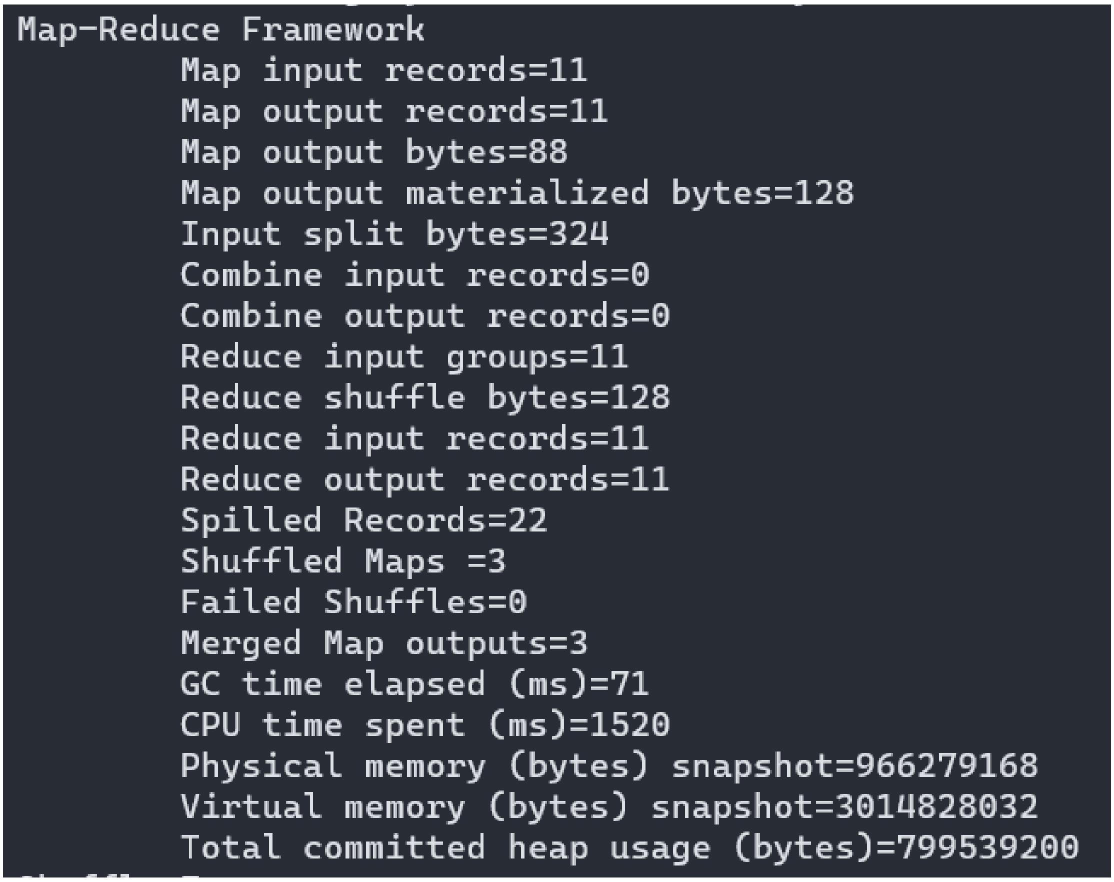
```
bash-4.1# touch MergeSort.java
bash-4.1# vi MergeSort.java
bash-4.1# javac MergeSort.java
bash-4.1# jar -cvf MergeSort.jar ./MergeSort*.class
added manifest
adding: MergeSort$Map.class(in = 1552) (out= 637)(deflated 58%)
adding: MergeSort$Reduce.class(in = 1758) (out= 709)(deflated 59%)
adding: MergeSort.class(in = 2028) (out= 1092)(deflated 46%)
```

（11）运行 jar 包

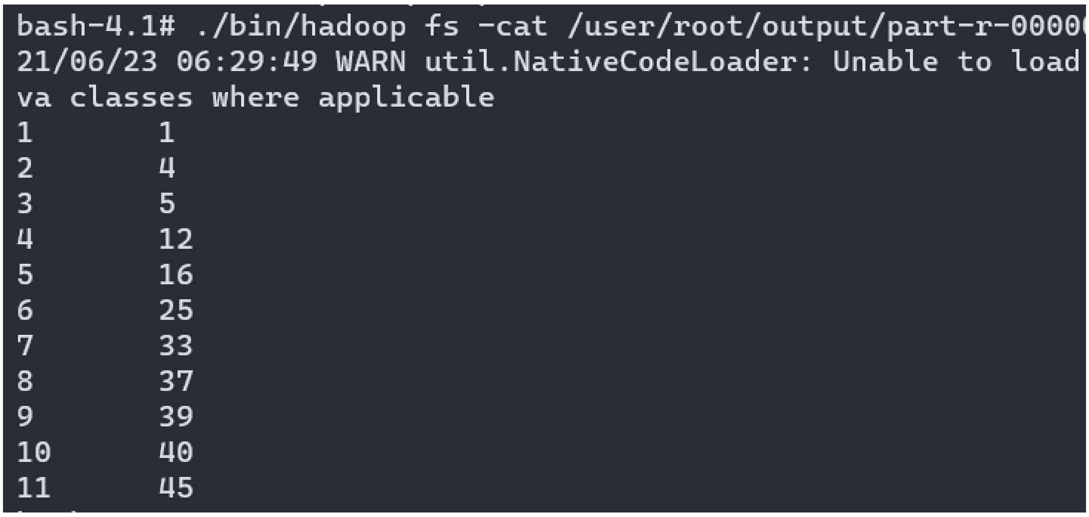


```
bash-4.1# ./bin/hadoop jar MergeSort.jar MergeSort
21/06/23 06:27:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:27:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/06/23 06:27:51 INFO input.FileInputFormat: Total input paths to process : 3
21/06/23 06:27:51 INFO mapreduce.JobSubmitter: number of splits:3
21/06/23 06:27:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624437756605_0003
21/06/23 06:27:51 INFO impl.YarnClientImpl: Submitted application application_1624437756605_0003
21/06/23 06:27:51 INFO mapreduce.Job: The url to track the job: http://84a2c3a89fce:8088/proxy/application_1624437756605
_0003/
21/06/23 06:27:51 INFO mapreduce.Job: Running job: job_1624437756605_0003
21/06/23 06:27:55 INFO mapreduce.Job: Job job_1624437756605_0003 running in uber mode : false
21/06/23 06:27:55 INFO mapreduce.Job:  map 0% reduce 0%
21/06/23 06:28:00 INFO mapreduce.Job:  map 100% reduce 0%
21/06/23 06:28:04 INFO mapreduce.Job:  map 100% reduce 100%
21/06/23 06:28:05 INFO mapreduce.Job: Job job_1624437756605_0003 completed successfully
21/06/23 06:28:06 INFO mapreduce.Job: Counters: 49
```

运行成功，结果如下：

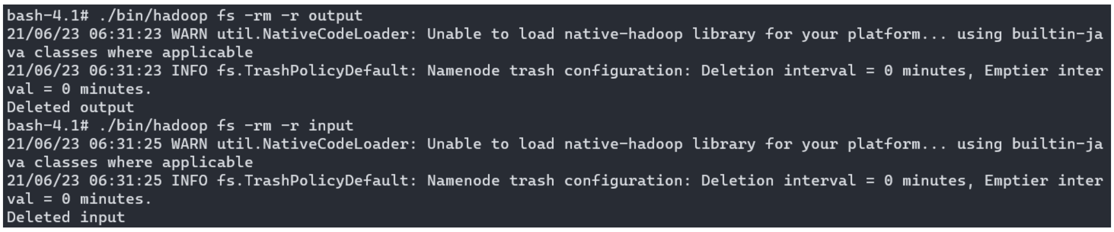


```
Map-Reduce Framework
        Map input records=11
        Map output records=11
        Map output bytes=88
        Map output materialized bytes=128
        Input split bytes=324
        Combine input records=0
        Combine output records=0
        Reduce input groups=11
        Reduce shuffle bytes=128
        Reduce input records=11
        Reduce output records=11
        Spilled Records=22
        Shuffled Maps =3
        Failed Shuffles=0
        Merged Map outputs=3
        GC time elapsed (ms)=71
        CPU time spent (ms)=1520
        Physical memory (bytes) snapshot=966279168
        Virtual memory (bytes) snapshot=3014828032
        Total committed heap usage (bytes)=799539200
```



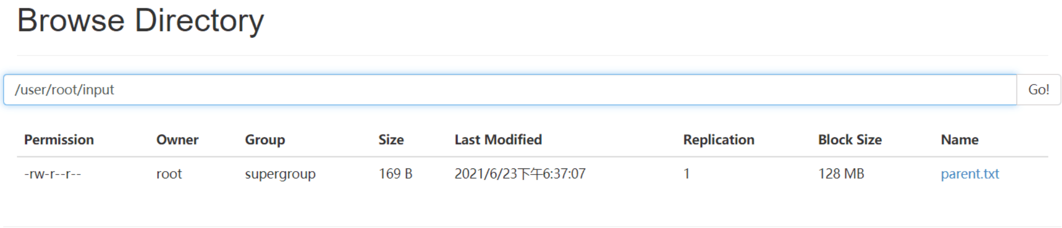
```
bash-4.1# ./bin/hadoop fs -cat /user/root/output/part-r-0000
21/06/23 06:29:49 WARN util.NativeCodeLoader: Unable to load
va classes where applicable
1       1
2       4
3       5
4       12
5       16
6       25
7       33
8       37
9       39
10      40
11      45
```

（12）删除 input 和 output 文件



```
bash-4.1# ./bin/hadoop fs -rm -r output
21/06/23 06:31:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:31:23 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted output
bash-4.1# ./bin/hadoop fs -rm -r input
21/06/23 06:31:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:31:25 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier inter
val = 0 minutes.
Deleted input
```

（13）创建新的文本文档并编辑，parent.txt，上传

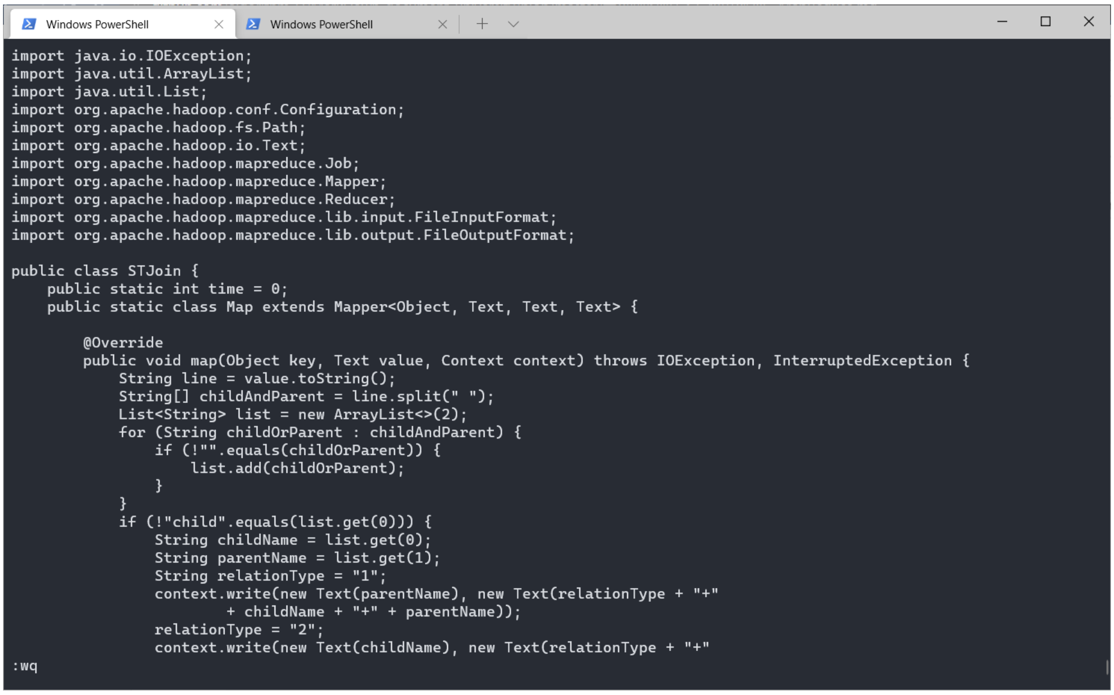


Browse Directory

/user/root/input

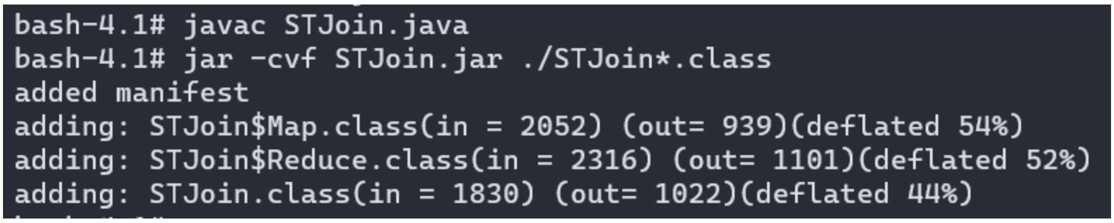
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	169 B	2021/6/23下午6:37:07	1	128 MB	parent.txt

（14）编写代码



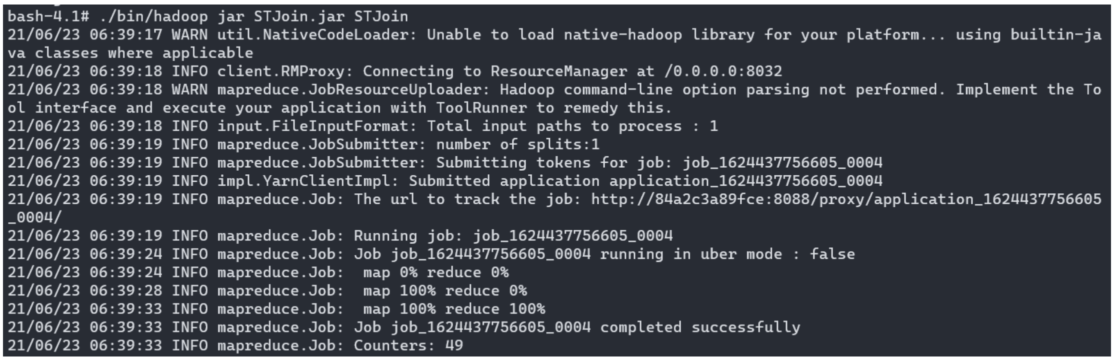
```
import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class STJoin {
    public static int time = 0;
    public static class Map extends Mapper<Object, Text, Text, Text> {

        @Override
        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            String line = value.toString();
            String[] childAndParent = line.split(" ");
            List<String> list = new ArrayList<>(2);
            for (String childOrParent : childAndParent) {
                if (!"".equals(childOrParent)) {
                    list.add(childOrParent);
                }
            }
            if (!"child".equals(list.get(0))) {
                String childName = list.get(0);
                String parentName = list.get(1);
                String relationType = "1";
                context.write(new Text(parentName), new Text(relationType + "+"
                        + childName + "+" + parentName));
                relationType = "2";
                context.write(new Text(childName), new Text(relationType + "+"
:wq
```

（15）编译代码，生成 jar 包

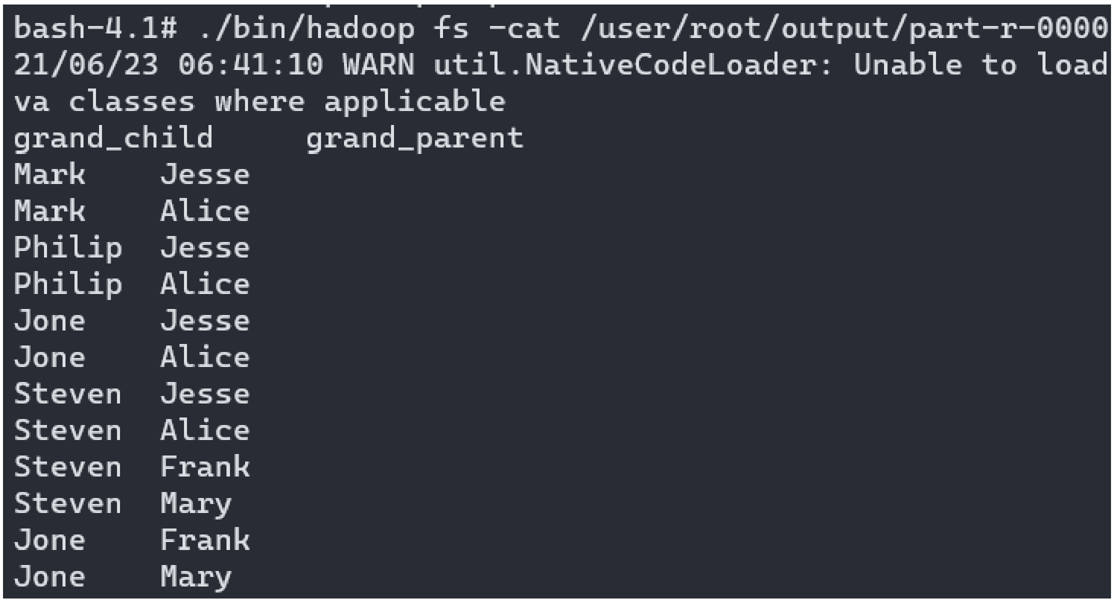


```
bash-4.1# javac STJoin.java
bash-4.1# jar -cvf STJoin.jar ./STJoin*.class
added manifest
adding: STJoin$Map.class(in = 2052) (out= 939)(deflated 54%)
adding: STJoin$Reduce.class(in = 2316) (out= 1101)(deflated 52%)
adding: STJoin.class(in = 1830) (out= 1022)(deflated 44%)
```

（16）运行 jar 包，得到结果



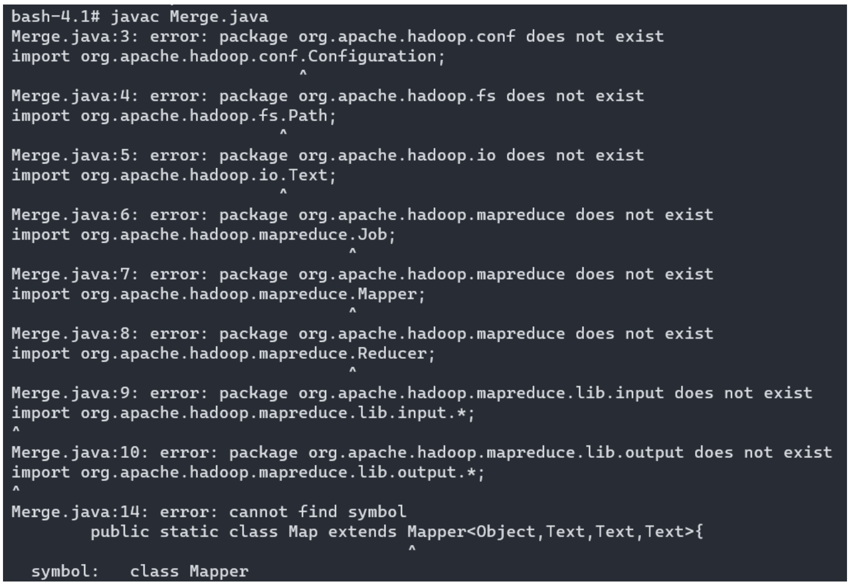
```
bash-4.1# ./bin/hadoop jar STJoin.jar STJoin
21/06/23 06:39:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:39:18 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/06/23 06:39:18 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the To
ol interface and execute your application with ToolRunner to remedy this.
21/06/23 06:39:18 INFO input.FileInputFormat: Total input paths to process : 1
21/06/23 06:39:19 INFO mapreduce.JobSubmitter: number of splits:1
21/06/23 06:39:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624437756605_0004
21/06/23 06:39:19 INFO impl.YarnClientImpl: Submitted application application_1624437756605_0004
21/06/23 06:39:19 INFO mapreduce.Job: The url to track the job: http://84a2c3a89fce:8088/proxy/application_1624437756605
_0004/
21/06/23 06:39:19 INFO mapreduce.Job: Running job: job_1624437756605_0004
21/06/23 06:39:24 INFO mapreduce.Job: Job job_1624437756605_0004 running in uber mode : false
21/06/23 06:39:24 INFO mapreduce.Job:  map 0% reduce 0%
21/06/23 06:39:28 INFO mapreduce.Job:  map 100% reduce 0%
21/06/23 06:39:33 INFO mapreduce.Job:  map 100% reduce 100%
21/06/23 06:39:33 INFO mapreduce.Job: Job job_1624437756605_0004 completed successfully
21/06/23 06:39:33 INFO mapreduce.Job: Counters: 49
```



```
bash-4.1# ./bin/hadoop fs -cat /user/root/output/part-r-0000
21/06/23 06:41:10 WARN util.NativeCodeLoader: Unable to load
va classes where applicable
grand_child     grand_parent
Mark    Jesse
Mark    Alice
Philip  Jesse
Philip  Alice
Jone    Jesse
Jone    Alice
Steven  Jesse
Steven  Alice
Steven  Frank
Steven  Mary
Jone    Frank
Jone    Mary
```

3.5 实验中的问题与解决方法

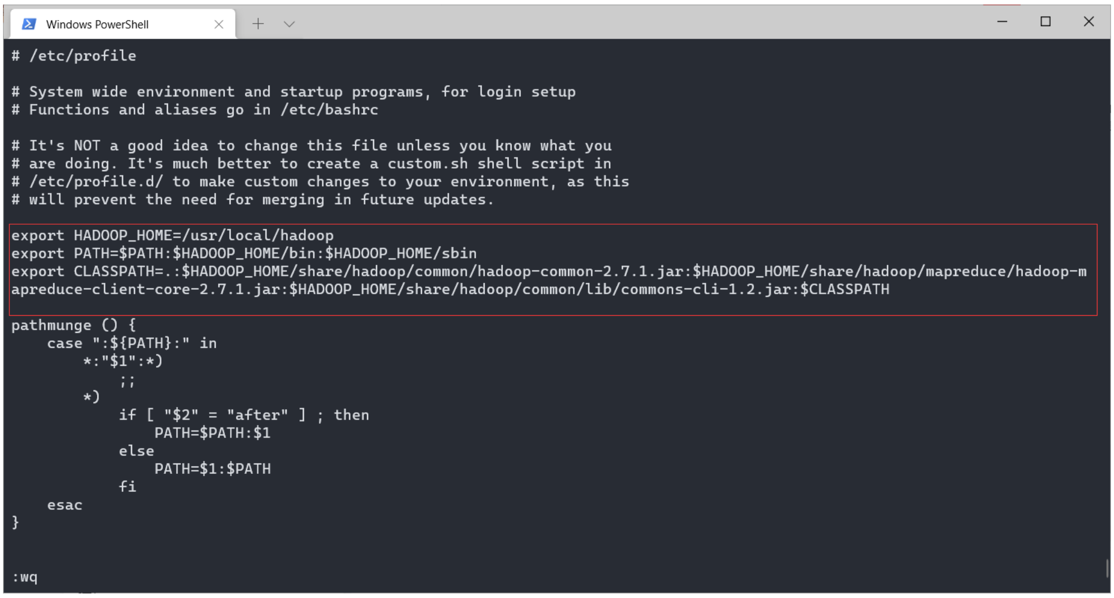
（1）编译 Java 代码时，找不到 Hadoop 中的 Java 依赖



```
bash-4.1# javac Merge.java
Merge.java:3: error: package org.apache.hadoop.conf does not exist
import org.apache.hadoop.conf.Configuration;
                             ^
Merge.java:4: error: package org.apache.hadoop.fs does not exist
import org.apache.hadoop.fs.Path;
                           ^
Merge.java:5: error: package org.apache.hadoop.io does not exist
import org.apache.hadoop.io.Text;
                           ^
Merge.java:6: error: package org.apache.hadoop.mapreduce does not exist
import org.apache.hadoop.mapreduce.Job;
                                  ^
Merge.java:7: error: package org.apache.hadoop.mapreduce does not exist
import org.apache.hadoop.mapreduce.Mapper;
                                  ^
Merge.java:8: error: package org.apache.hadoop.mapreduce does not exist
import org.apache.hadoop.mapreduce.Reducer;
                                  ^
Merge.java:9: error: package org.apache.hadoop.mapreduce.lib.input does not exist
import org.apache.hadoop.mapreduce.lib.input.*;
^
Merge.java:10: error: package org.apache.hadoop.mapreduce.lib.output does not exist
import org.apache.hadoop.mapreduce.lib.output.*;
^
Merge.java:14: error: cannot find symbol
        public static class Map extends Mapper<Object,Text,Text,Text>{
                                        ^
  symbol:   class Mapper
```

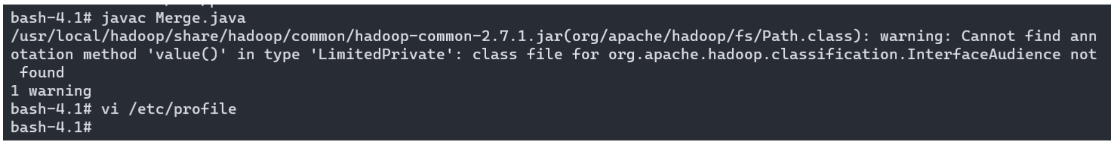
原因：安装 Java 和 Hadoop 时没有配置好 Hadoop 中依赖的路径。

解决方法：编辑/etc/profile 文件，添加 Hadoop 依赖的路径到环境变量。



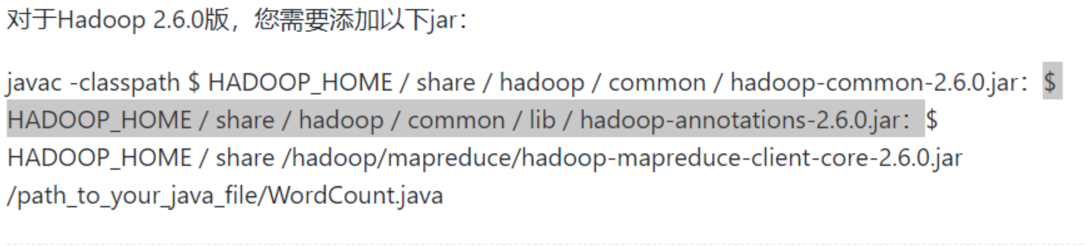
```
# /etc/profile

# System wide environment and startup programs, for login setup
# Functions and aliases go in /etc/bashrc

# It's NOT a good idea to change this file unless you know what you
# are doing. It's much better to create a custom.sh shell script in
# /etc/profile.d/ to make custom changes to your environment, as this
# will prevent the need for merging in future updates.

export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export CLASSPATH=.:$HADOOP_HOME/share/hadoop/common/hadoop-common-2.7.1.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-m
apreduce-client-core-2.7.1.jar:$HADOOP_HOME/share/hadoop/common/lib/commons-cli-1.2.jar:$CLASSPATH

pathmunge () {
    case ":${PATH}:" in
        *:"$1":*)
            ;;
        *)
            if [ "$2" = "after" ] ; then
                PATH=$PATH:$1
            else
                PATH=$1:$PATH
            fi
    esac
}

:wq
```

（2）编译时报错，缺少函数“value()”



```
bash-4.1# javac Merge.java
/usr/local/hadoop/share/hadoop/common/hadoop-common-2.7.1.jar(org/apache/hadoop/fs/Path.class): warning: Cannot find ann
otation method 'value()' in type 'LimitedPrivate': class file for org.apache.hadoop.classification.InterfaceAudience not
 found
1 warning
bash-4.1# vi /etc/profile
bash-4.1#
```

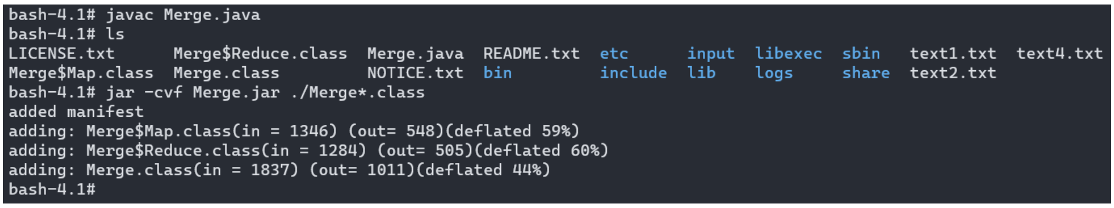
原因：在某些版本的 Hadoop 中，需要特别添加 hadoop-annotations-2.x.x.jar 到环境变量。格式如下图：



对于Hadoop 2.6.0版，您需要添加以下jar：

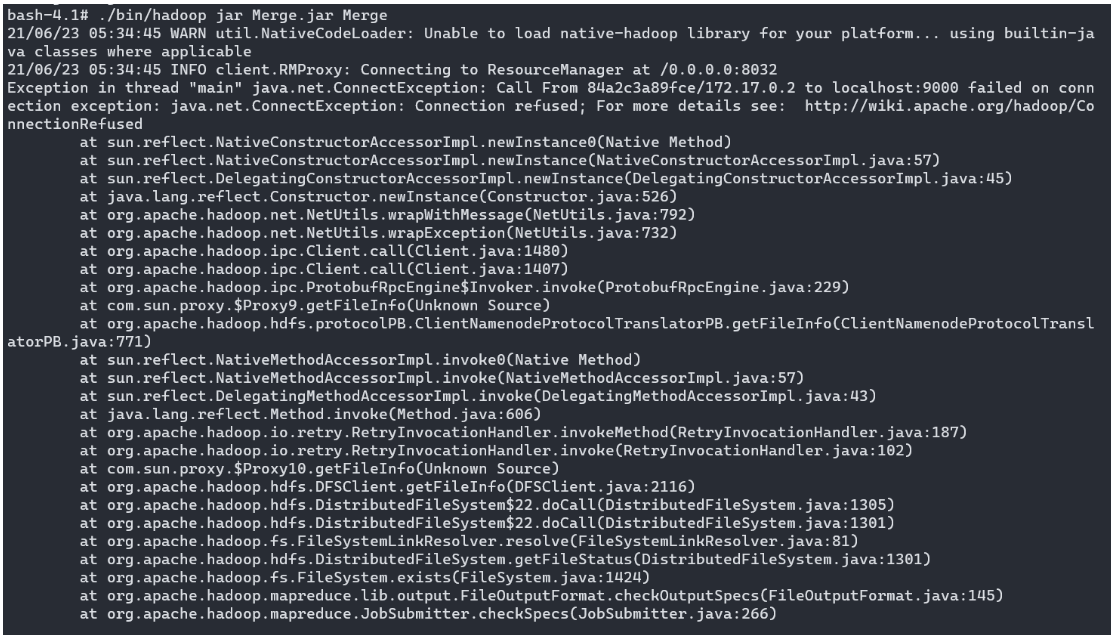
javac -classpath $ HADOOP_HOME / share / hadoop / common / hadoop-common-2.6.0.jar： $ HADOOP_HOME / share / hadoop / common / lib / hadoop-annotations-2.6.0.jar： $ HADOOP_HOME / share /hadoop/mapreduce/hadoop-mapreduce-client-core-2.6.0.jar /path_to_your_java_file/WordCount.java

解决方法：添加 hadoop-annotations-2.x.x.jar 到环境变量，此后编译成功



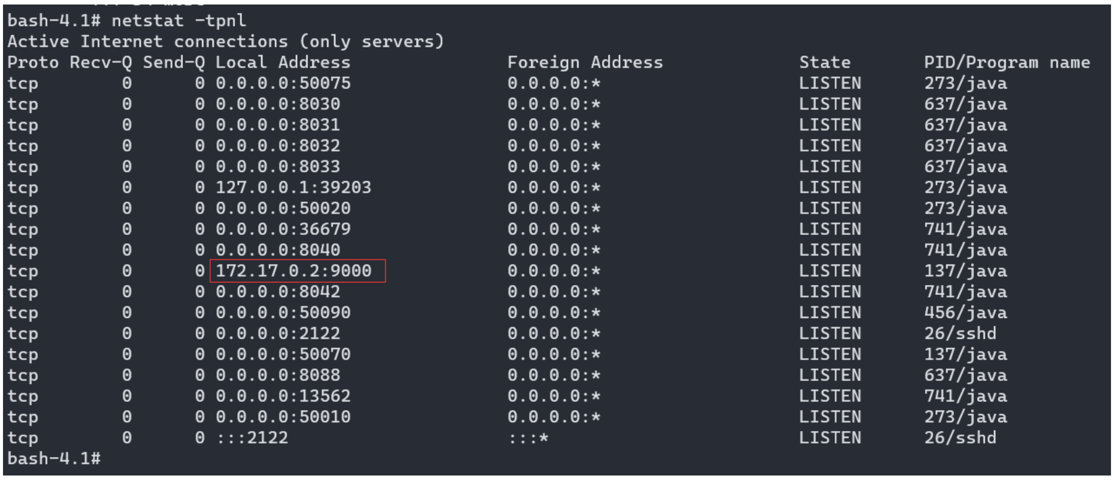
```
bash-4.1# javac Merge.java
bash-4.1# ls
LICENSE.txt      Merge$Reduce.class  Merge.java  README.txt  etc      input  libexec  sbin   text1.txt  text4.txt
Merge$Map.class  Merge.class         NOTICE.txt  bin         include  lib    logs     share  text2.txt
bash-4.1# jar -cvf Merge.jar ./Merge*.class
added manifest
adding: Merge$Map.class(in = 1346) (out= 548)(deflated 59%)
adding: Merge$Reduce.class(in = 1284) (out= 505)(deflated 60%)
adding: Merge.class(in = 1837) (out= 1011)(deflated 44%)
bash-4.1#
```

（3）9000 端口拒绝了访问



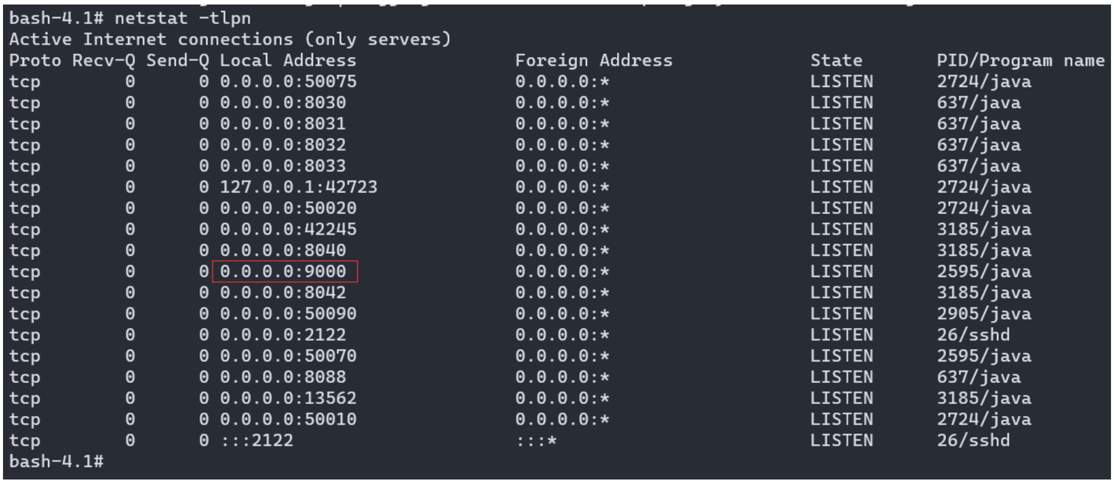
```
bash-4.1# ./bin/hadoop jar Merge.jar Merge
21/06/23 05:34:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 05:34:45 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
Exception in thread "main" java.net.ConnectException: Call From 84a2c3a89fce/172.17.0.2 to localhost:9000 failed on conn
ection exception: java.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/hadoop/Co
nnectionRefused
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:57)
        at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:526)
        at org.apache.hadoop.net.NetUtils.wrapWithMessage(NetUtils.java:792)
        at org.apache.hadoop.net.NetUtils.wrapException(NetUtils.java:732)
        at org.apache.hadoop.ipc.Client.call(Client.java:1480)
        at org.apache.hadoop.ipc.Client.call(Client.java:1407)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:229)
        at com.sun.proxy.$Proxy9.getFileInfo(Unknown Source)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.getFileInfo(ClientNamenodeProtocolTransl
atorPB.java:771)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:187)
        at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:102)
        at com.sun.proxy.$Proxy10.getFileInfo(Unknown Source)
        at org.apache.hadoop.hdfs.DFSClient.getFileInfo(DFSClient.java:2116)
        at org.apache.hadoop.hdfs.DistributedFileSystem$22.doCall(DistributedFileSystem.java:1305)
        at org.apache.hadoop.hdfs.DistributedFileSystem$22.doCall(DistributedFileSystem.java:1301)
        at org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkResolver.java:81)
        at org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(DistributedFileSystem.java:1301)
        at org.apache.hadoop.fs.FileSystem.exists(FileSystem.java:1424)
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:145)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:266)
```

原因：查看网络状态，9000 端口已开启，但 ip 设置成了本机，没有向外界开放。



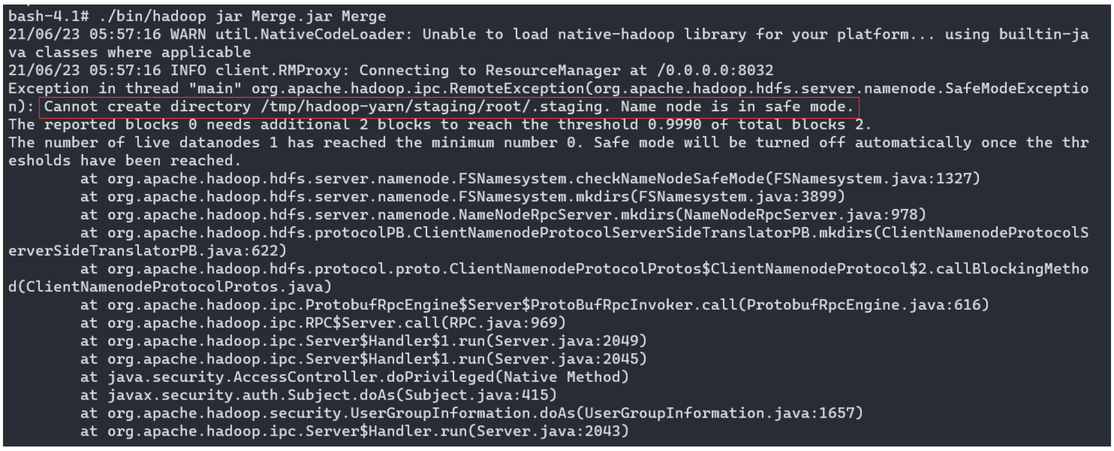
```
bash-4.1# netstat -tpnl
Active Internet connections (only servers)
Proto Recv-Q Send-Q Local Address               Foreign Address             State       PID/Program name
tcp        0      0 0.0.0.0:50075               0.0.0.0:*                   LISTEN      273/java
tcp        0      0 0.0.0.0:8030                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:8031                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:8032                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:8033                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 127.0.0.1:39203             0.0.0.0:*                   LISTEN      273/java
tcp        0      0 0.0.0.0:50020               0.0.0.0:*                   LISTEN      273/java
tcp        0      0 0.0.0.0:36679               0.0.0.0:*                   LISTEN      741/java
tcp        0      0 0.0.0.0:8040                0.0.0.0:*                   LISTEN      741/java
tcp        0      0 172.17.0.2:9000             0.0.0.0:*                   LISTEN      137/java
tcp        0      0 0.0.0.0:8042                0.0.0.0:*                   LISTEN      741/java
tcp        0      0 0.0.0.0:50090               0.0.0.0:*                   LISTEN      456/java
tcp        0      0 0.0.0.0:2122                0.0.0.0:*                   LISTEN      26/sshd
tcp        0      0 0.0.0.0:50070               0.0.0.0:*                   LISTEN      137/java
tcp        0      0 0.0.0.0:8088                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:13562               0.0.0.0:*                   LISTEN      741/java
tcp        0      0 0.0.0.0:50010               0.0.0.0:*                   LISTEN      273/java
tcp        0      0 :::2122                     :::*                        LISTEN      26/sshd
bash-4.1#
```

解决方法：把 9000 端口修改为向所有人开放，重启 Hadoop。



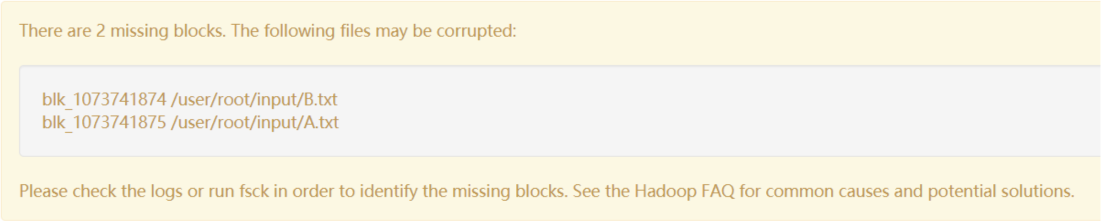
```
bash-4.1# netstat -tlpn
Active Internet connections (only servers)
Proto Recv-Q Send-Q Local Address               Foreign Address             State       PID/Program name
tcp        0      0 0.0.0.0:50075               0.0.0.0:*                   LISTEN      2724/java
tcp        0      0 0.0.0.0:8030                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:8031                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:8032                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:8033                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 127.0.0.1:42723             0.0.0.0:*                   LISTEN      2724/java
tcp        0      0 0.0.0.0:50020               0.0.0.0:*                   LISTEN      2724/java
tcp        0      0 0.0.0.0:42245               0.0.0.0:*                   LISTEN      3185/java
tcp        0      0 0.0.0.0:8040                0.0.0.0:*                   LISTEN      3185/java
tcp        0      0 0.0.0.0:9000                0.0.0.0:*                   LISTEN      2595/java
tcp        0      0 0.0.0.0:8042                0.0.0.0:*                   LISTEN      3185/java
tcp        0      0 0.0.0.0:50090               0.0.0.0:*                   LISTEN      2905/java
tcp        0      0 0.0.0.0:2122                0.0.0.0:*                   LISTEN      26/sshd
tcp        0      0 0.0.0.0:50070               0.0.0.0:*                   LISTEN      2595/java
tcp        0      0 0.0.0.0:8088                0.0.0.0:*                   LISTEN      637/java
tcp        0      0 0.0.0.0:13562               0.0.0.0:*                   LISTEN      3185/java
tcp        0      0 0.0.0.0:50010               0.0.0.0:*                   LISTEN      2724/java
tcp        0      0 :::2122                     :::*                        LISTEN      26/sshd
bash-4.1#
```

（4）Namenode 进入了安全模式，无法运行 jar 包



```
bash-4.1# ./bin/hadoop jar Merge.jar Merge
21/06/23 05:57:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 05:57:16 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
Exception in thread "main" org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.server.namenode.SafeModeExceptio
n): Cannot create directory /tmp/hadoop-yarn/staging/root/.staging. Name node is in safe mode.
The reported blocks 0 needs additional 2 blocks to reach the threshold 0.9990 of total blocks 2.
The number of live datanodes 1 has reached the minimum number 0. Safe mode will be turned off automatically once the thr
esholds have been reached.
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkNameNodeSafeMode(FSNamesystem.java:1327)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.mkdirs(FSNamesystem.java:3899)
        at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.mkdirs(NameNodeRpcServer.java:978)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTranslatorPB.mkdirs(ClientNamenodeProtocolS
erverSideTranslatorPB.java:622)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingMetho
d(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:616)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:969)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2049)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Server.java:2045)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1657)
        at org.apache.hadoop.ipc.Server$Handler.run(Server.java:2043)
```

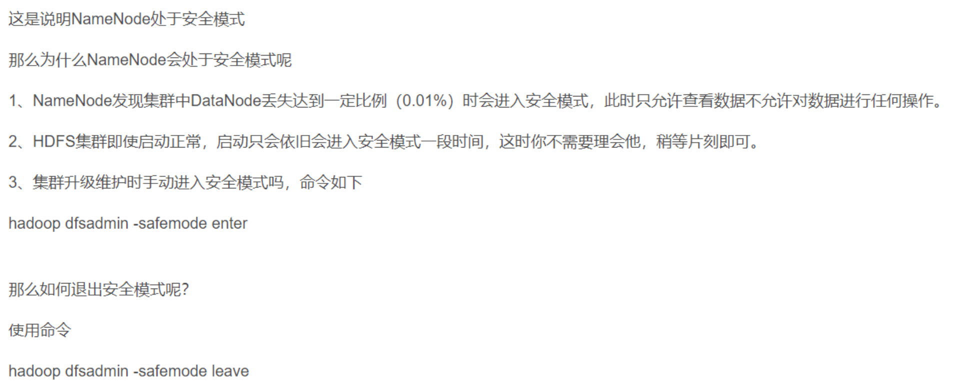
原因：在修改 9000 端口向所有人开启后，需要重启 Hadoop，但此时需要把上次运行时的 data 删除，Hadoop 重启后读取不到原来的文件，认为 File Block 被破坏，所以自动安全模式。



There are 2 missing blocks. The following files may be corrupted:

blk_1073741874 /user/root/input/B.txt
blk_1073741875 /user/root/input/A.txt

Please check the logs or run fsck in order to identify the missing blocks. See the Hadoop FAQ for common causes and potential solutions.



这是说明NameNode处于安全模式

那么为什么NameNode会处于安全模式呢

1、NameNode发现集群中DataNode丢失达到一定比例（0.01%）时会进入安全模式，此时只允许查看数据不允许对数据进行任何操作。

2、HDFS集群即使启动正常，启动只会依旧会进入安全模式一段时间，这时你不需要理会他，稍等片刻即可。

3、集群升级维护时手动进入安全模式吗，命令如下

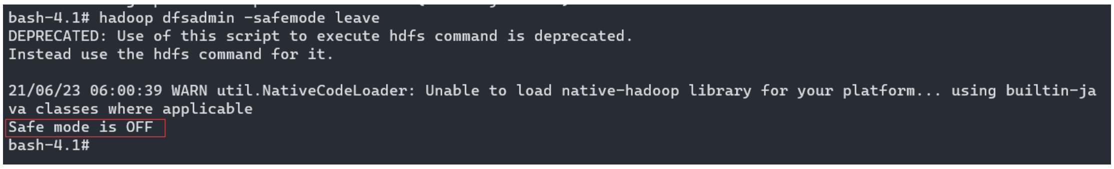
hadoop dfsadmin -safemode enter

那么如何退出安全模式呢?

使用命令

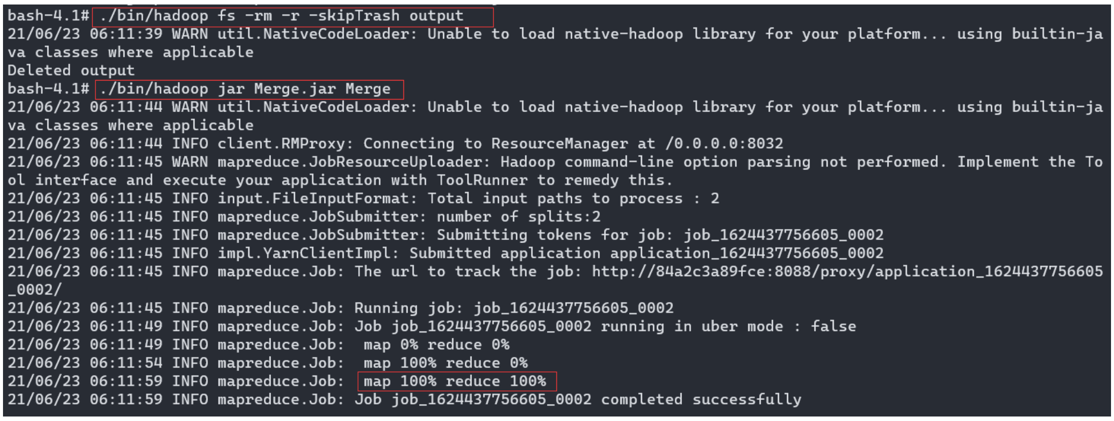
hadoop dfsadmin -safemode leave

解决方法：关闭安全模式，此时虽然没有了文件，但是 HDFS 还是对原始数据有记录，所以要把命令行删除原有文件，即使他已经不在了，之后重新上传需要用到的文件即可。



```
bash-4.1# hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

21/06/23 06:00:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
Safe mode is OFF
bash-4.1#
```

删除原有的 input 和 output 文件夹，再次运行，即可编译成功。



```
bash-4.1# ./bin/hadoop fs -rm -r -skipTrash output
21/06/23 06:11:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
Deleted output
bash-4.1# ./bin/hadoop jar Merge.jar Merge
21/06/23 06:11:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
21/06/23 06:11:44 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/06/23 06:11:45 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the To
ol interface and execute your application with ToolRunner to remedy this.
21/06/23 06:11:45 INFO input.FileInputFormat: Total input paths to process : 2
21/06/23 06:11:45 INFO mapreduce.JobSubmitter: number of splits:2
21/06/23 06:11:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1624437756605_0002
21/06/23 06:11:45 INFO impl.YarnClientImpl: Submitted application application_1624437756605_0002
21/06/23 06:11:45 INFO mapreduce.Job: The url to track the job: http://84a2c3a89fce:8088/proxy/application_1624437756605
_0002/
21/06/23 06:11:45 INFO mapreduce.Job: Running job: job_1624437756605_0002
21/06/23 06:11:49 INFO mapreduce.Job: Job job_1624437756605_0002 running in uber mode : false
21/06/23 06:11:49 INFO mapreduce.Job:  map 0% reduce 0%
21/06/23 06:11:54 INFO mapreduce.Job:  map 100% reduce 0%
21/06/23 06:11:59 INFO mapreduce.Job:  map 100% reduce 100%
21/06/23 06:11:59 INFO mapreduce.Job: Job job_1624437756605_0002 completed successfully
```

第四章 总结与体会

大数据技术是计算机研究领域的一个重要分支，它已经渗透到生活中的各个领域，大数据技术的高速发展为各行业的生命注入了新的血液，给我们的生活带来了极大的便利，这同时对各行业的发展也是一个考验，人们将更加离不开大数据技术，而计算机通过利用海量数据也将更好地服务于人类，使人们的生活更加丰富。未来大数据技术的应用将更加适应人们的生活。

当前，数据科学正在蓬勃发展，研究智能计算的领域十分活跃。虽然目前智能计算和大数据的研究水平暂时还很难使“智能机器”真正具备人类的智能，但大数据技术将在 21 世纪蓬勃发展，人工智能将不仅是模仿生物脑的功能，而且两者具有相同的特性，这两者的结合将使人工智能的研究向着更广和更深的方向发展，将开辟一个全新的领域，开辟很多新的研究方向。大数据技术将探索智能的新概念、新理论、新方法和新技术，而这些研究将在以后的发展中取得重大的成就。

经过课程设计，使我加深了对所学理论知识的理解与巩固，并能将课程中的纯理论应用到实践中，进一步加深了对知识的认识。同时，也有助于对其他知识的理解。我不但对分布式文件管理有了更深入的理解，还熟练的应用 Hadoop、HDFS、Hbase 对文件进行各种操作。

第五章 参考与引用

[1] https://blog.csdn.net/liu16659/article/details/80212233

[2] https://blog.csdn.net/ystyaoshengting/article/details/103026872

[3] https://zhuanlan.zhihu.com/p/269047002

[4] https://www.shuzhiduo.com/A/GBJrBBQRJ0/

[5] https://stackoverflow.com/questions/48107616/hadoop-blockmissingexception

[6] https://blog.csdn.net/liu16659/article/details/80212233

[7]https://blog.csdn.net/qq_52679708/article/details/115448087?utm_medium=distribute.pc_relevant.none-task-blog-baidujs_title-0&spm=1001.2101.3001.4242

附录一. MapReduce 编程初级实践实验代码

(1) Merge.java

```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.*;

import org.apache.hadoop.mapreduce.lib.output.*;


public class Merge {


    public static class Map extends Mapper<Object,Text,Text,Text>{

        private static Text text=new Text();

        public void map(Object key,Text value,Context context) throws IOException, InterruptedException{

                text=value;

                context.write(text,new Text(""));

        }

    }


    public static class Reduce extends Reducer<Text,Text,Text,Text>{

        public void reduce(Text key,Iterable <Text>values,Context context)

            throws IOException, InterruptedException{

              context.write(key, new Text(""));

        }

    }

    public static void main(String[] args) throws IOException,
    ClassNotFoundException, InterruptedException{
        Configuration conf=new Configuration();
        conf.set("fs.defaultFS","hdfs://localhost:9000");
        String[] otherArgs=new String[]{"input","output"};
        if(otherArgs.length!=2){
              System.err.println("Usage:Merge and duplicate removal<in><out>");
              System.exit(2);
        }
        Job job=Job.getInstance(conf,"Merge and duplicate removal");
        job.setJarByClass(Merge.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);
        FileInputFormat.addInputPath(job,new Path(otherArgs[0]));
        FileOutputFormat.setOutputPath(job,new Path(otherArgs[1]));
        System.exit(job.waitForCompletion(true)?0:1);
    }
}
```

(2) MergeSort.java

```
import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class MergeSort {
    public static class Map extends Mapper<Object,Text,IntWritable,IntWritable>{
        private static IntWritable data=new IntWritable();
        public void map(Object key,Text value,Context context) throws IOException, InterruptedException{
            String line=value.toString();
            data.set(Integer.parseInt(line));
            context.write(data, new IntWritable(1));
        }
    }
    public static class Reduce extends Reducer<IntWritable,IntWritable,IntWritable,IntWritable>{
        private static IntWritable linenum=new IntWritable(1);
        public void reduce(IntWritable key,Iterable <IntWritable>values,Context context)
          throws IOException, InterruptedException{
            for(IntWritable num:values){
                context.write(linenum, key);
                linenum=new IntWritable(linenum.get()+1);
            }
        }
    }

    /**
     * @param args
     * @throws IOException
     * @throws InterruptedException
     * @throws ClassNotFoundException
     */
    public static void main(String[] args) throws IOException,
    ClassNotFoundException, InterruptedException{
        Configuration conf=new Configuration();
        conf.set("fs.defaultFS","hdfs://localhost:9000");
        String[] str=new String[]{"input","output"};
        String[] otherArgs=new GenericOptionsParser(conf,str).getRemainingArgs();
        if(otherArgs.length!=2){
            System.err.println("Usage:mergesort<in><out>");
            System.exit(2);
        }
        Job job=Job.getInstance(conf,"mergesort");
        job.setJarByClass(MergeSort.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setOutputKeyClass(IntWritable.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job,new Path(otherArgs[0]));
        FileOutputFormat.setOutputPath(job,new Path(otherArgs[1]));
        System.exit(job.waitForCompletion(true)?0:1);
    }
}
```

(3) STJoin.java

```
import java.io.IOException;
import java.util.ArrayList;
import java.util.List;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class STJoin {
    public static int time = 0;
    public static class Map extends Mapper<Object, Text, Text, Text> {

        @Override
        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            String line = value.toString();
            String[] childAndParent = line.split(" ");
            List<String> list = new ArrayList<>(2);
            for (String childOrParent : childAndParent) {
                if (!"".equals(childOrParent)) {
                    list.add(childOrParent);
                }
            }
            if (!"child".equals(list.get(0))) {
                String childName = list.get(0);
                String parentName = list.get(1);
                String relationType = "1";
                context.write(new Text(parentName), new Text(relationType + "+"
                        + childName + "+" + parentName));
                relationType = "2";
                context.write(new Text(childName), new Text(relationType + "+"
                        + childName + "+" + parentName));
            }
        }
    }

    public static class Reduce extends Reducer<Text, Text, Text, Text> {
        @Override
        public void reduce(Text key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
            if (time == 0) {
                context.write(new Text("grand_child"), new Text("grand_parent"));
                time++;
            }
            List<String> grandChild = new ArrayList<>();
            List<String> grandParent = new ArrayList<>();
            for (Text text : values) {
                String s = text.toString();
                String[] relation = s.split("\\+");
                String relationType = relation[0];
                String childName = relation[1];
                String parentName = relation[2];
                if ("1".equals(relationType)) {
                    grandChild.add(childName);
                } else {
                    grandParent.add(parentName);
                }
            }
            int grandParentNum = grandParent.size();
            int grandChildNum = grandChild.size();
            if (grandParentNum != 0 && grandChildNum != 0) {
                for (int m = 0; m < grandChildNum; m++) {
                    for (int n = 0; n < grandParentNum; n++) {
                        context.write(new Text(grandChild.get(m)), new Text(
                                grandParent.get(n)));
                    }
                }
            }
        }
    }

    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        conf.set("fs.defaultFS", "hdfs://localhost:9000");
        String[] otherArgs = new String[]{"input", "output"};
        if (otherArgs.length != 2) {
            System.err.println("Usage: Single Table Join <in> <out>");
            System.exit(2);
        }
        Job job = Job.getInstance(conf, "Single table Join ");
        job.setJarByClass(STJoin.class);
        job.setMapperClass(Map.class);
        job.setReducerClass(Reduce.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);
        FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
        FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```