作业5

191840373 朱家萱

对设计思路，实验结果等给出说明，并给出提交作业运行成功的WEB页面截图。可以进一步对性能、扩展性等方面存在的不足和可能的改进之处进行分析。

一、设计思路

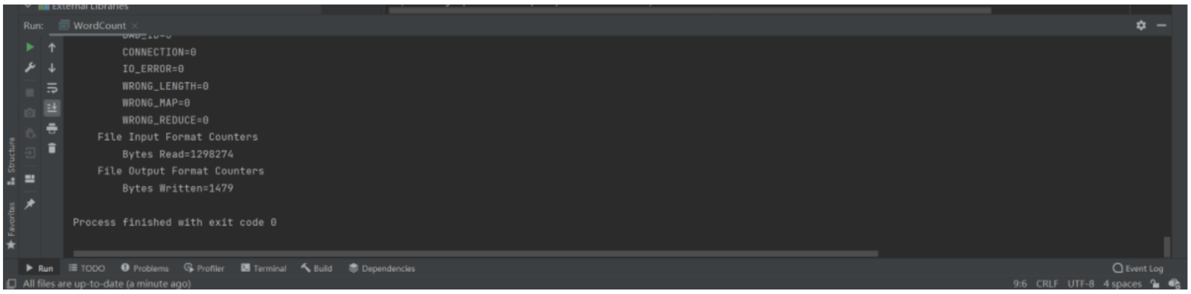
首先单词计数的程序很容易实现，与hadoop官网的wordcount2.0采取相同的设计思路，难点在于按照单词出现的次数从大到小排列、停词的处理和筛选前100个高频词输出。wordcount中Map函数的输入键值对是行和文本，输出是单词和次数1，经过Combiner合并，Reduce函数输入是单词和频数，输出单词和累加和。由于原本的排序是MapReduce的默认排序，就需要一个函数使得全局有序，并且降序输出，就要写一个继承类指定为逆序，并在主函数里调用。Map读取分布式缓存中的停词表，延用wordcount2中的parseSkipFile，读取filename路径下的文件，将文件中的需要词加入停用词列表wordsToSkip中，一开始的每个单词在所有文件中词频和每个单词在单个文件中词频的结果都存放在临时文件中，最后将临时文件删去。

二、实验结果

1、按照CSDN教程，在本地搭建平台

[Hadoop: Intellij结合Maven本地运行和调试MapReduce程序 (无需搭载Hadoop和HDFS环境)]

安装Intellij IDEA、JDK 1.8、hadoop 3.2.1、cygwin进行配置，方便对代码进行debug，具体的配置、调错由于与实验本身关系不大，故不再赘述，只呈现idea中运行成功的截图



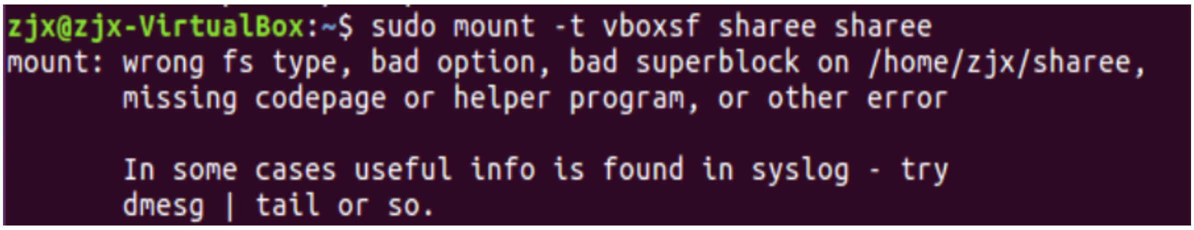
```
            CONNECTION=0
            IO_ERROR=0
            WRONG_LENGTH=0
            WRONG_MAP=0
            WRONG_REDUCE=0
        File Input Format Counters
            Bytes Read=1298274
        File Output Format Counters
            Bytes Written=1479

Process finished with exit code 0
```

2、Linux下的运行

（1）把要共享的文件夹挂载到虚拟机某一个文件

为了方便文件传输，由于winscp总是连接失败，故决定采用挂载本机文件夹的方式进行本机与虚拟机的文件共享。

- 出现错误：



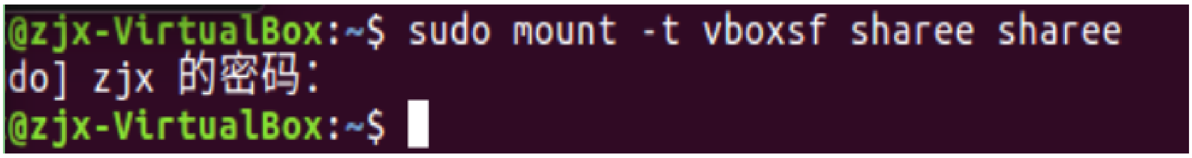
```
zjx@zjx-VirtualBox:~$ sudo mount -t vboxsf sharee sharee
mount: wrong fs type, bad option, bad superblock on /home/zjx/sharee,
       missing codepage or helper program, or other error

       In some cases useful info is found in syslog - try
       dmesg | tail or so.
```

- 解决方法：

在终端输入，挂载成功

```
zjx@zjx-VirtualBox:~$ sudo apt-get install nfs-common
```

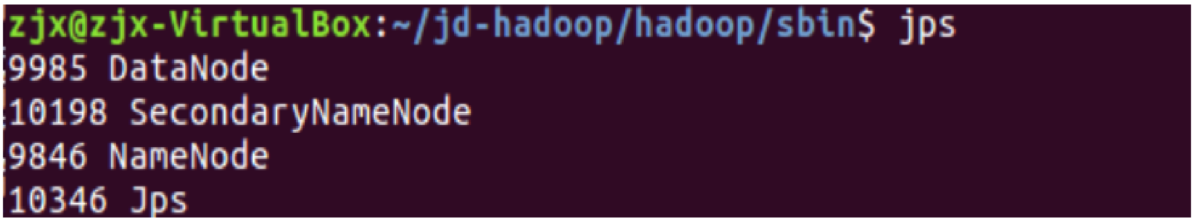
- 重启后挂载失效问题，需要在终端输入，重新挂载成功



```
@zjx-VirtualBox:~$ sudo mount -t vboxsf sharee sharee
do] zjx 的密码：
@zjx-VirtualBox:~$ 
```

（2）开启dfs节点

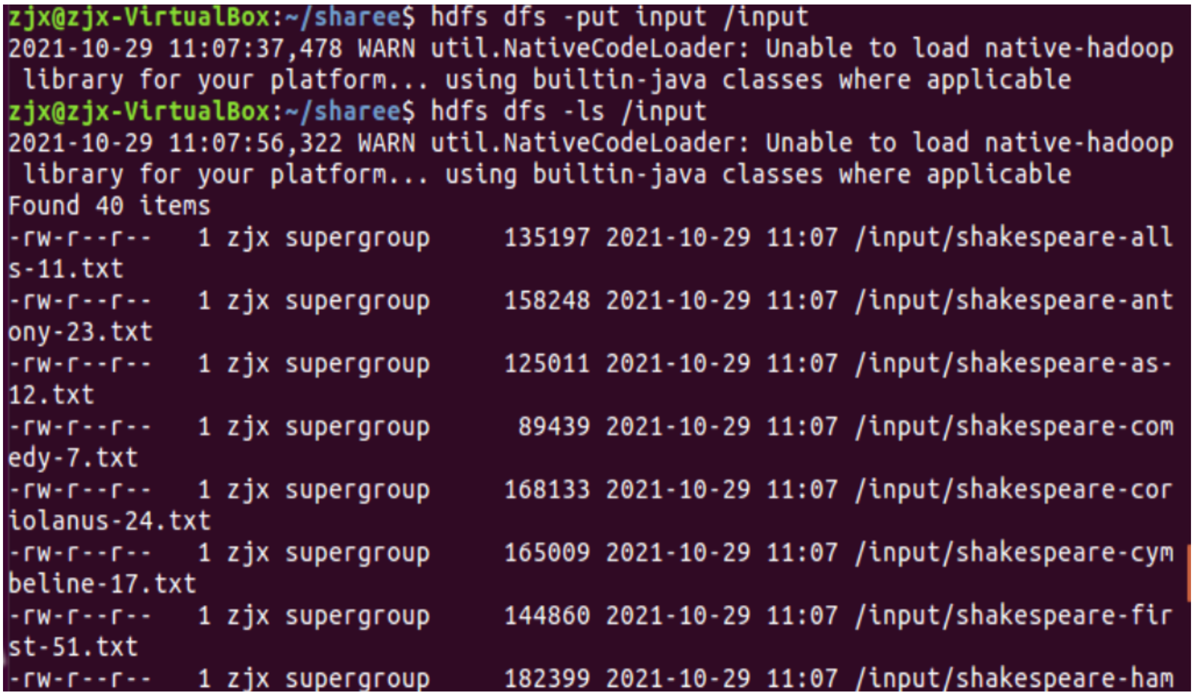
- 查看状态，正常



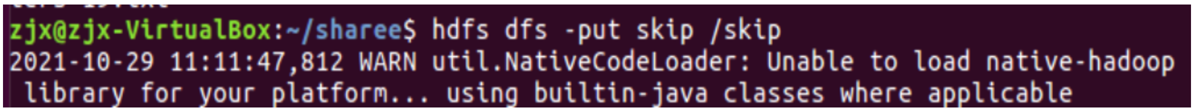
```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/sbin$ jps
9985 DataNode
10198 SecondaryNameNode
9846 NameNode
10346 Jps
```

- 若不幸没有DataNode节点，只需将tmp文件里的dfs全部删除，再格式化、重启，即可重新出现DataNode

（3）正式运行

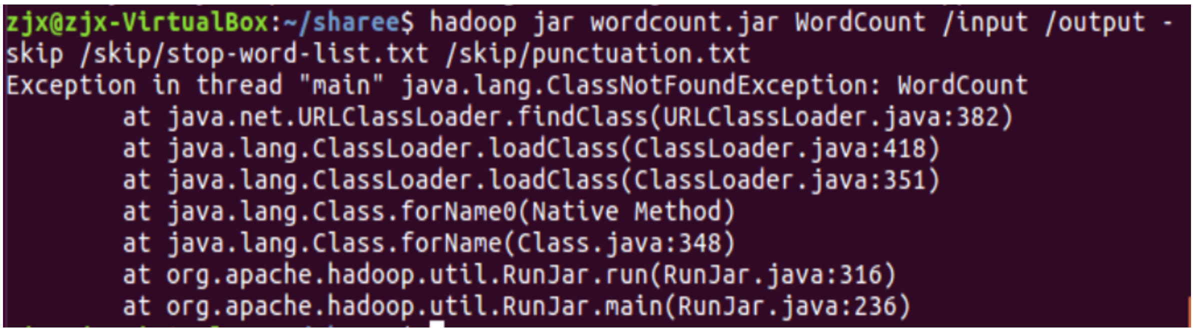


```
zjx@zjx-VirtualBox:~/sharee$ hdfs dfs -put input /input
2021-10-29 11:07:37,478 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
zjx@zjx-VirtualBox:~/sharee$ hdfs dfs -ls /input
2021-10-29 11:07:56,322 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Found 40 items
-rw-r--r--   1 zjx supergroup     135197 2021-10-29 11:07 /input/shakespeare-all
s-11.txt
-rw-r--r--   1 zjx supergroup     158248 2021-10-29 11:07 /input/shakespeare-ant
ony-23.txt
-rw-r--r--   1 zjx supergroup     125011 2021-10-29 11:07 /input/shakespeare-as-
12.txt
-rw-r--r--   1 zjx supergroup      89439 2021-10-29 11:07 /input/shakespeare-com
edy-7.txt
-rw-r--r--   1 zjx supergroup     168133 2021-10-29 11:07 /input/shakespeare-cor
iolanus-24.txt
-rw-r--r--   1 zjx supergroup     165009 2021-10-29 11:07 /input/shakespeare-cym
beline-17.txt
-rw-r--r--   1 zjx supergroup     144860 2021-10-29 11:07 /input/shakespeare-fir
st-51.txt
-rw-r--r--   1 zjx supergroup     182399 2021-10-29 11:07 /input/shakespeare-ham
```



```
zjx@zjx-VirtualBox:~/sharee$ hdfs dfs -put skip /skip
2021-10-29 11:11:47,812 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
```

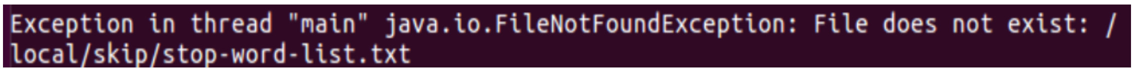
- 不幸报错：



```
zjx@zjx-VirtualBox:~/sharee$ hadoop jar wordcount.jar WordCount /input /output -
skip /skip/stop-word-list.txt /skip/punctuation.txt
Exception in thread "main" java.lang.ClassNotFoundException: WordCount
        at java.net.URLClassLoader.findClass(URLClassLoader.java:382)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:418)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:351)
        at java.lang.Class.forName0(Native Method)
        at java.lang.Class.forName(Class.java:348)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:316)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
```

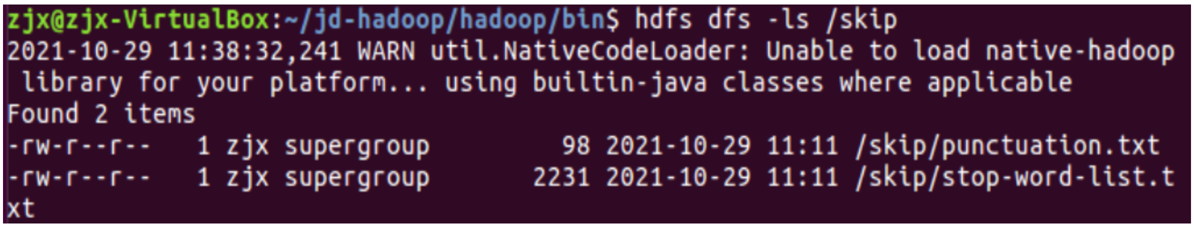
原因是打包的jar不对，在本机重新打包jar文件并共享后，报错解决

- 再报错：



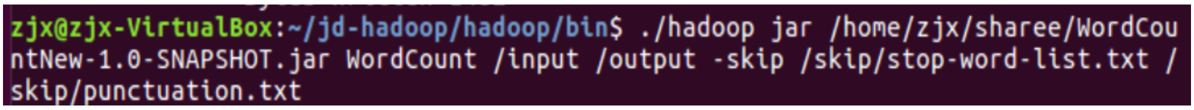
```
Exception in thread "main" java.io.FileNotFoundException: File does not exist: /
local/skip/stop-word-list.txt
```

  显示找不到停词文件，而传输是成功的



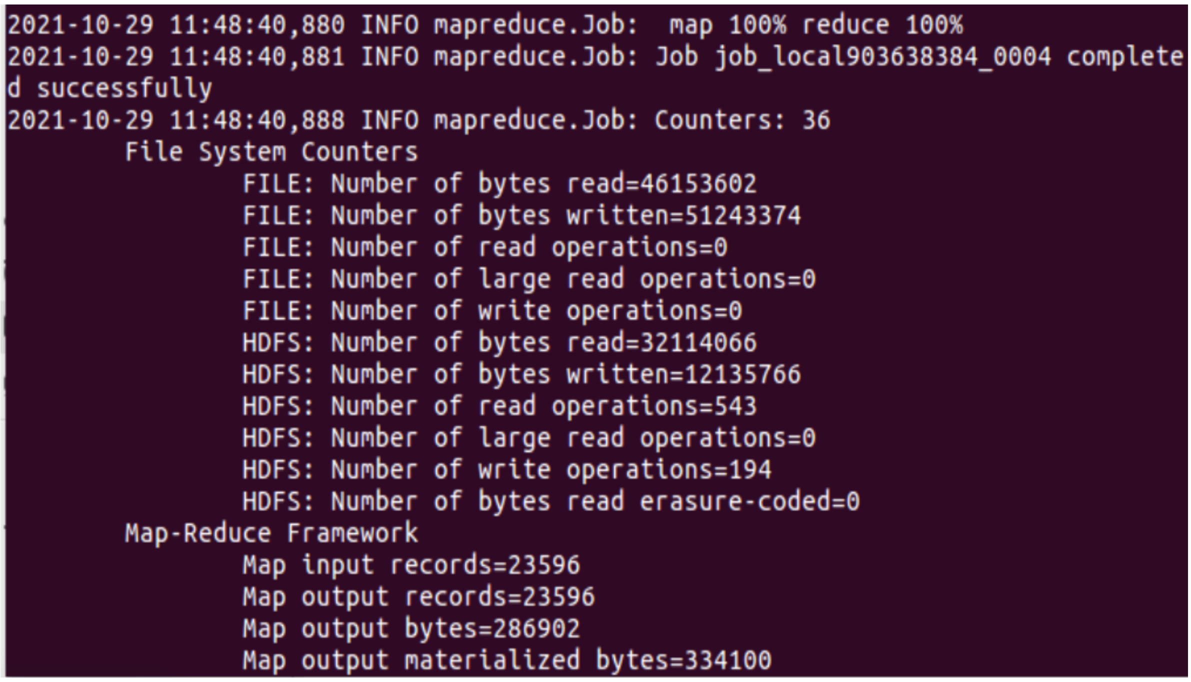
```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ hdfs dfs -ls /skip
2021-10-29 11:38:32,241 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 zjx supergroup         98 2021-10-29 11:11 /skip/punctuation.txt
-rw-r--r--   1 zjx supergroup       2231 2021-10-29 11:11 /skip/stop-word-list.t
xt
```

再观察命令，发现文件地址输入错了，更改命令后：

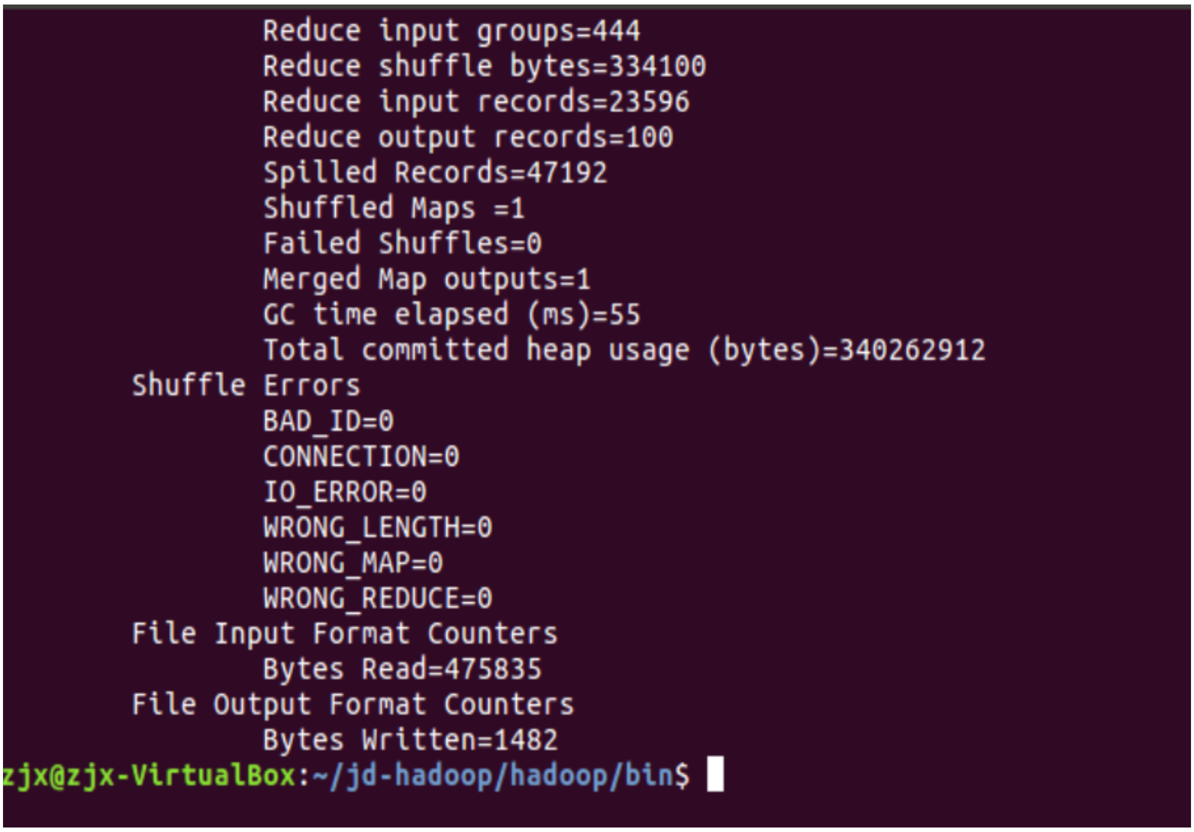


```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hadoop jar /home/zjx/sharee/WordCou
ntNew-1.0-SNAPSHOT.jar WordCount /input /output -skip /skip/stop-word-list.txt /
skip/punctuation.txt
```

成功运行：

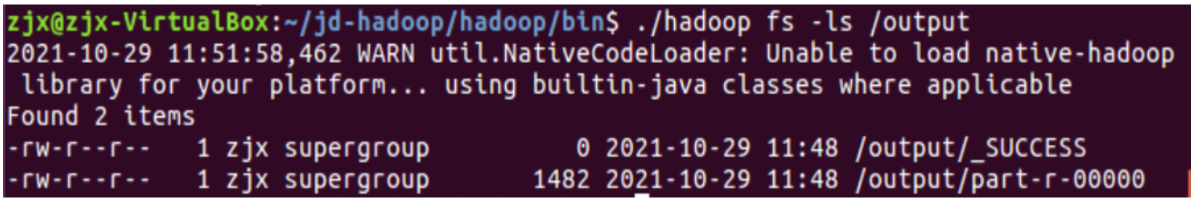


```
2021-10-29 11:48:40,880 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-29 11:48:40,881 INFO mapreduce.Job: Job job_local903638384_0004 complete
d successfully
2021-10-29 11:48:40,888 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=46153602
                FILE: Number of bytes written=51243374
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=32114066
                HDFS: Number of bytes written=12135766
                HDFS: Number of read operations=543
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=194
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=23596
                Map output records=23596
                Map output bytes=286902
                Map output materialized bytes=334100
```

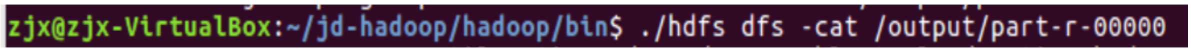


```
                Reduce input groups=444
                Reduce shuffle bytes=334100
                Reduce input records=23596
                Reduce output records=100
                Spilled Records=47192
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=55
                Total committed heap usage (bytes)=340262912
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=475835
        File Output Format Counters
                Bytes Written=1482
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ 
```

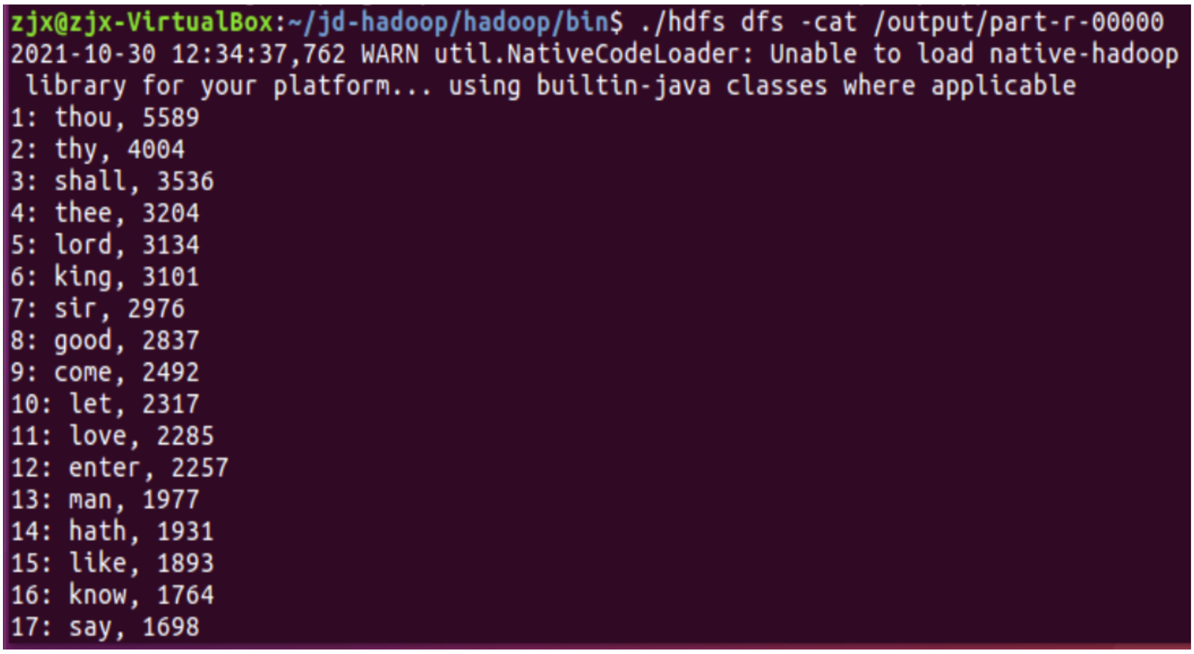
- 查看运行结果：



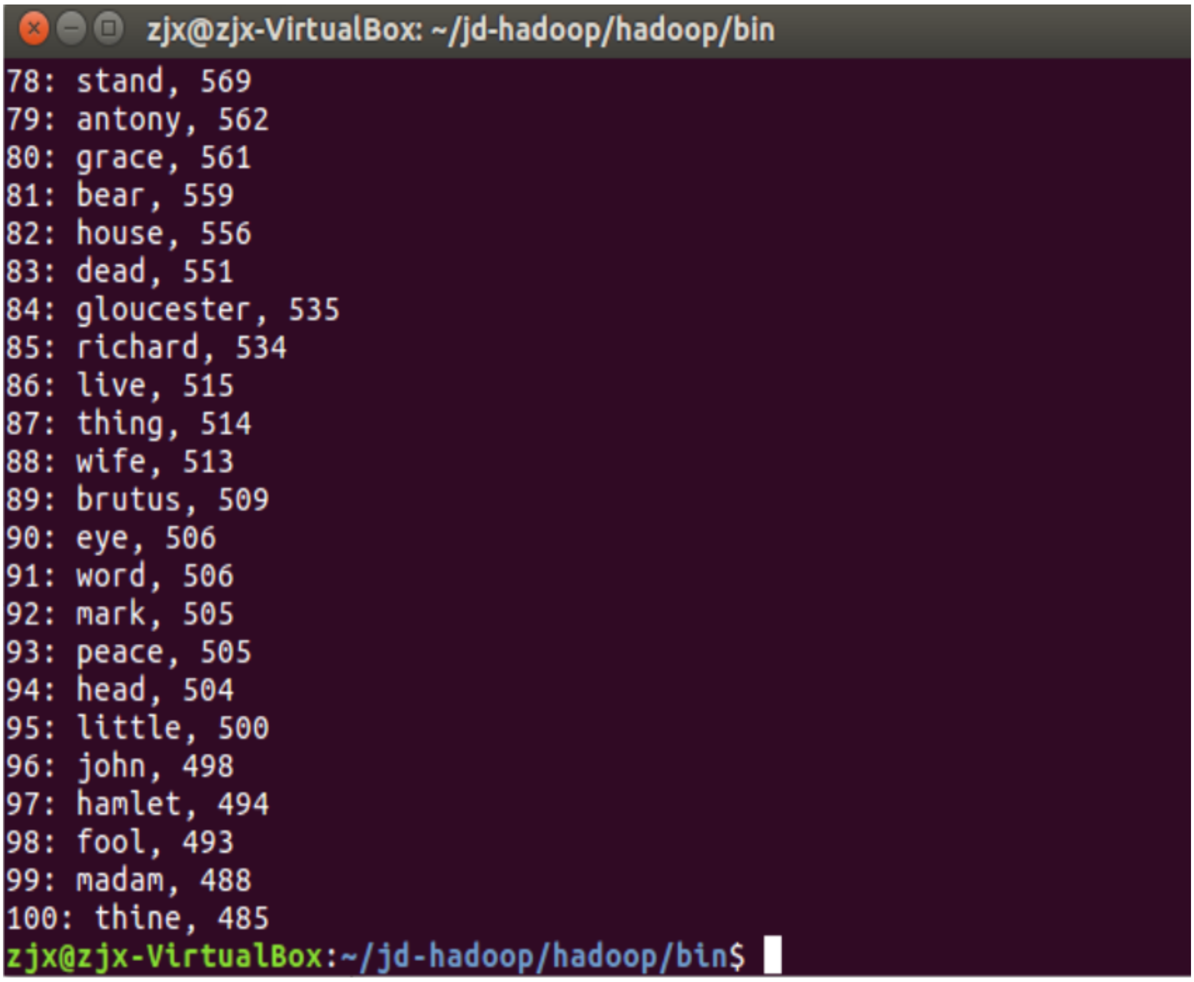
```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hadoop fs -ls /output
2021-10-29 11:51:58,462 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 zjx supergroup          0 2021-10-29 11:48 /output/_SUCCESS
-rw-r--r--   1 zjx supergroup       1482 2021-10-29 11:48 /output/part-r-00000
```



```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hdfs dfs -cat /output/part-r-00000
```

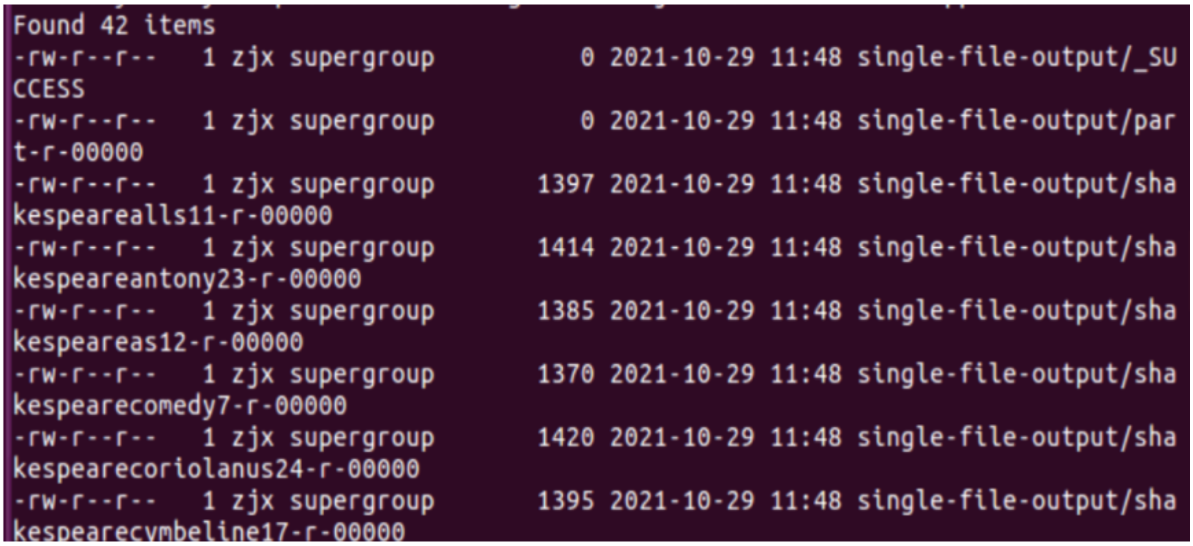


```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hdfs dfs -cat /output/part-r-00000
2021-10-30 12:34:37,762 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
1: thou, 5589
2: thy, 4004
3: shall, 3536
4: thee, 3204
5: lord, 3134
6: king, 3101
7: sir, 2976
8: good, 2837
9: come, 2492
10: let, 2317
11: love, 2285
12: enter, 2257
13: man, 1977
14: hath, 1931
15: like, 1893
16: know, 1764
17: say, 1698
```



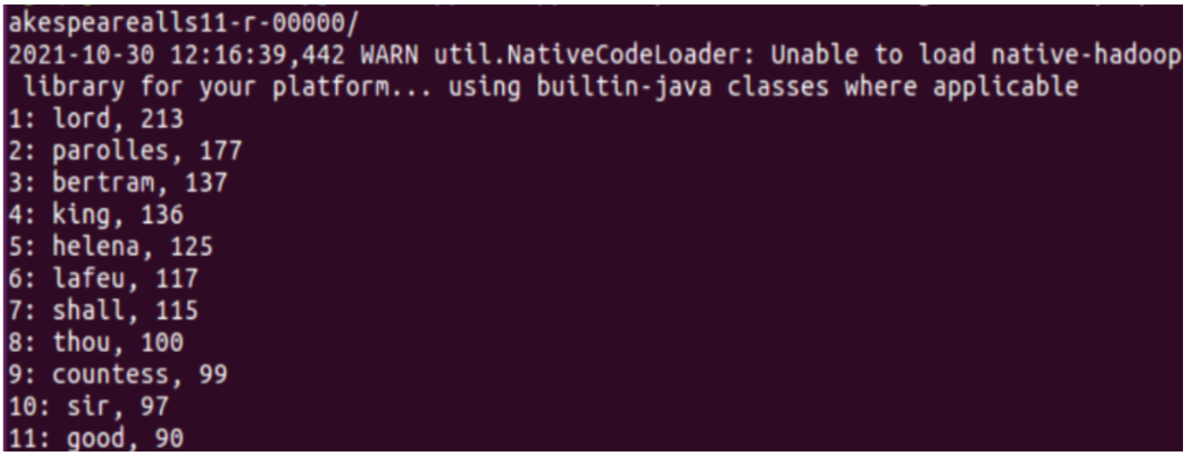
```
78: stand, 569
79: antony, 562
80: grace, 561
81: bear, 559
82: house, 556
83: dead, 551
84: gloucester, 535
85: richard, 534
86: live, 515
87: thing, 514
88: wife, 513
89: brutus, 509
90: eye, 506
91: word, 506
92: mark, 505
93: peace, 505
94: head, 504
95: little, 500
96: john, 498
97: hamlet, 494
98: fool, 493
99: madam, 488
100: thine, 485
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ 
```

- 单个文集的运行结果：

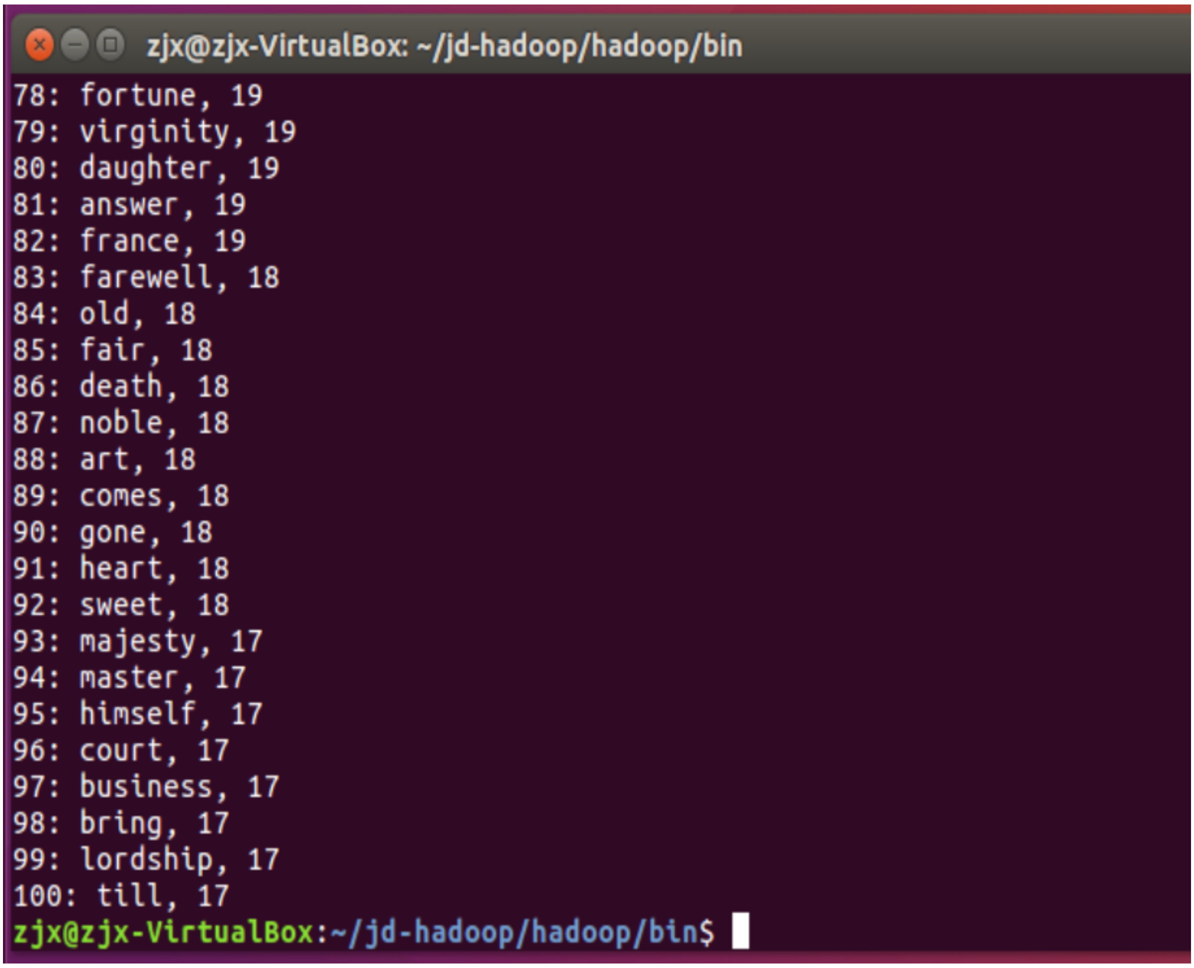


```
Found 42 items
-rw-r--r--   1 zjx supergroup          0 2021-10-29 11:48 single-file-output/_SU
CCESS
-rw-r--r--   1 zjx supergroup          0 2021-10-29 11:48 single-file-output/par
t-r-00000
-rw-r--r--   1 zjx supergroup       1397 2021-10-29 11:48 single-file-output/sha
kespearealls11-r-00000
-rw-r--r--   1 zjx supergroup       1414 2021-10-29 11:48 single-file-output/sha
kespeareantony23-r-00000
-rw-r--r--   1 zjx supergroup       1385 2021-10-29 11:48 single-file-output/sha
kespeareas12-r-00000
-rw-r--r--   1 zjx supergroup       1370 2021-10-29 11:48 single-file-output/sha
kespearecomedy7-r-00000
-rw-r--r--   1 zjx supergroup       1420 2021-10-29 11:48 single-file-output/sha
kespearecoriolanus24-r-00000
-rw-r--r--   1 zjx supergroup       1395 2021-10-29 11:48 single-file-output/sha
kespearecymbeline17-r-00000
```

以其中一个莎士比亚文档为例：

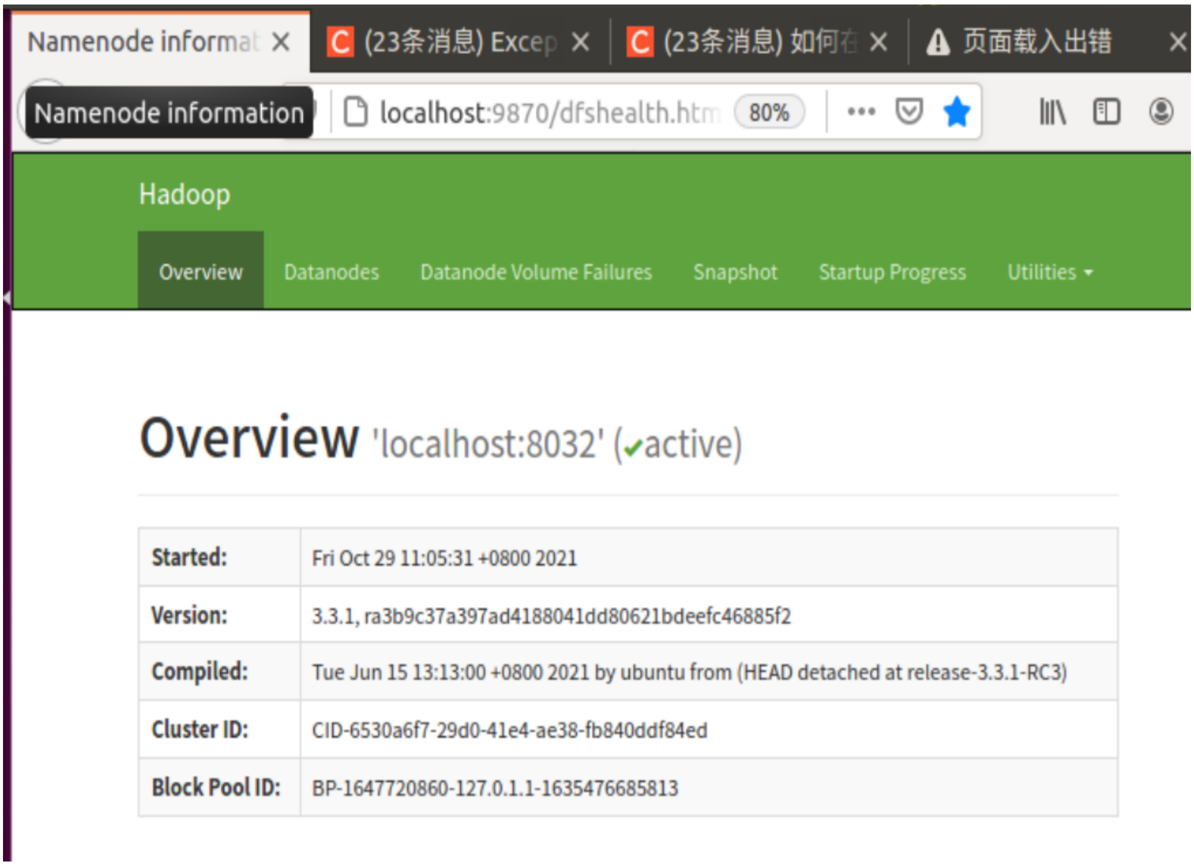


```
akespearealls11-r-00000/
2021-10-30 12:16:39,442 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
1: lord, 213
2: parolles, 177
3: bertram, 137
4: king, 136
5: helena, 125
6: lafeu, 117
7: shall, 115
8: thou, 100
9: countess, 99
10: sir, 97
11: good, 90
```



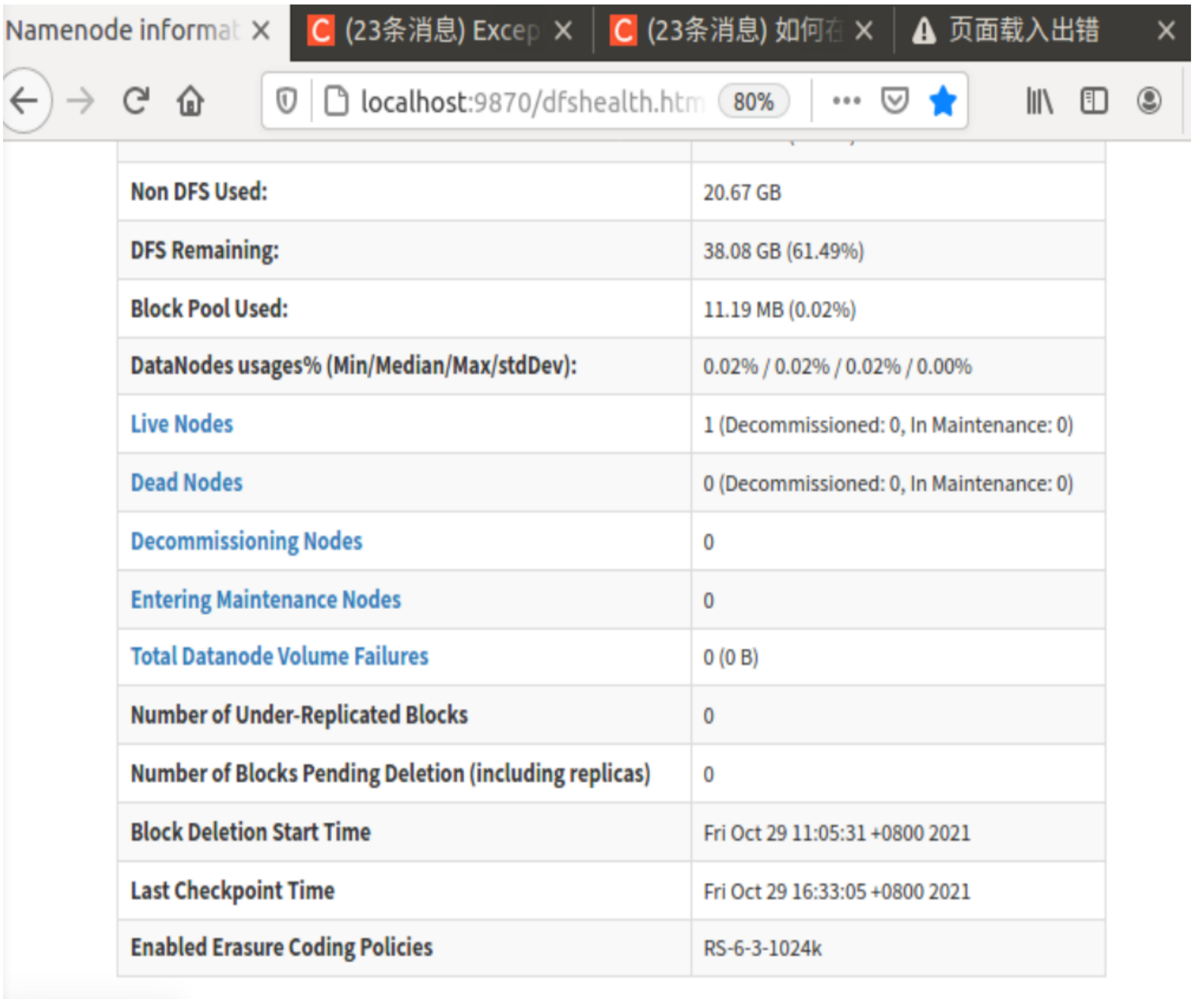
```
78: fortune, 19
79: virginity, 19
80: daughter, 19
81: answer, 19
82: france, 19
83: farewell, 18
84: old, 18
85: fair, 18
86: death, 18
87: noble, 18
88: art, 18
89: comes, 18
90: gone, 18
91: heart, 18
92: sweet, 18
93: majesty, 17
94: master, 17
95: himself, 17
96: court, 17
97: business, 17
98: bring, 17
99: lordship, 17
100: till, 17
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$
```

- Web端的显示：

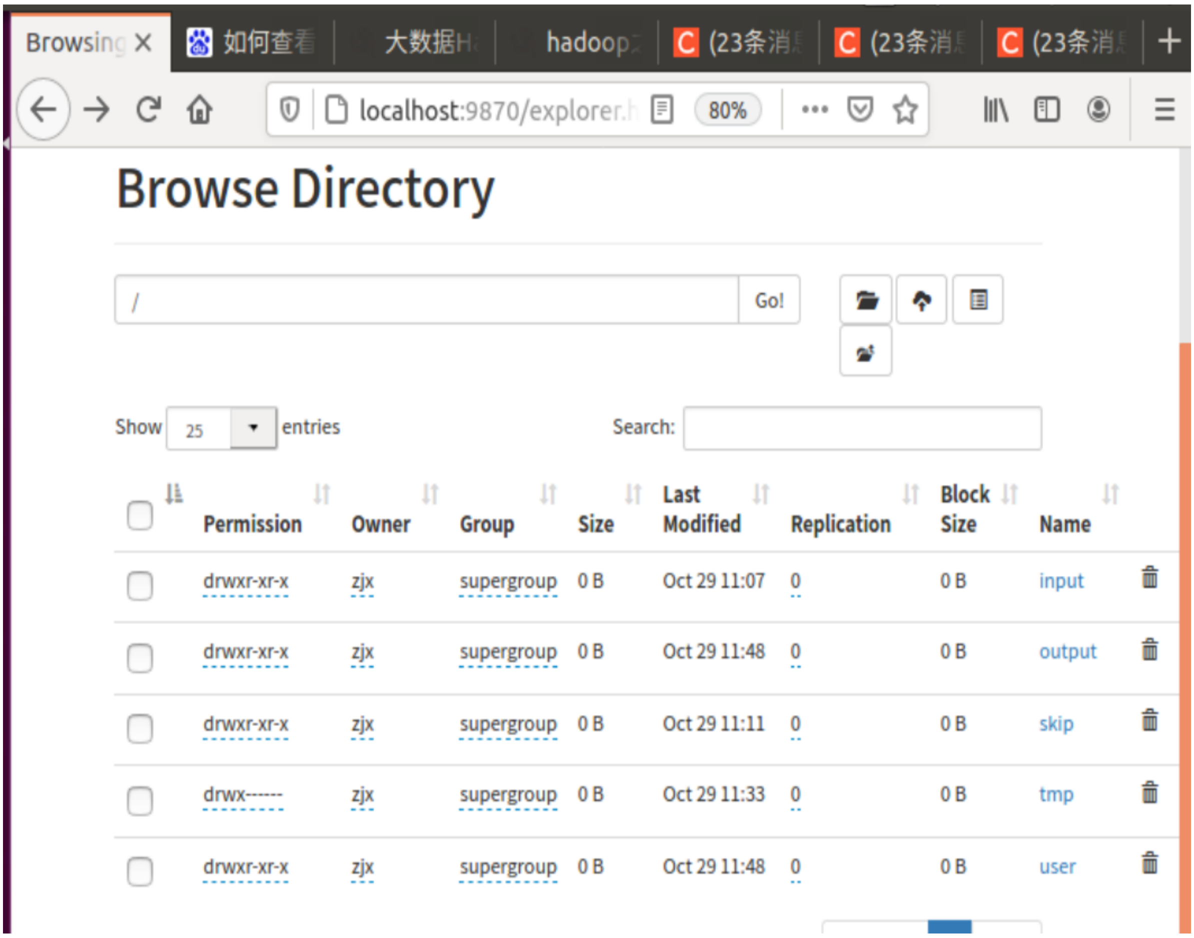


Overview 'localhost:8032' (✓active)

Started:	Fri Oct 29 11:05:31 +0800 2021
Version:	3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled:	Tue Jun 15 13:13:00 +0800 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-6530a6f7-29d0-41e4-ae38-fb840ddf84ed
Block Pool ID:	BP-1647720860-127.0.1.1-1635476685813



Non DFS Used:	20.67 GB
DFS Remaining:	38.08 GB (61.49%)
Block Pool Used:	11.19 MB (0.02%)
DataNodes usages% (Min/Median/Max/stdDev):	0.02% / 0.02% / 0.02% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Fri Oct 29 11:05:31 +0800 2021
Last Checkpoint Time	Fri Oct 29 16:33:05 +0800 2021
Enabled Erasure Coding Policies	RS-6-3-1024k



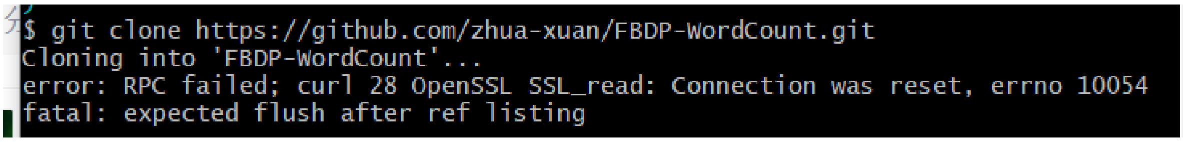
Browse Directory

/

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	zjx	supergroup	0 B	Oct 29 11:07	0	0 B	input
drwxr-xr-x	zjx	supergroup	0 B	Oct 29 11:48	0	0 B	output
drwxr-xr-x	zjx	supergroup	0 B	Oct 29 11:11	0	0 B	skip
drwx------	zjx	supergroup	0 B	Oct 29 11:33	0	0 B	tmp
drwxr-xr-x	zjx	supergroup	0 B	Oct 29 11:48	0	0 B	user

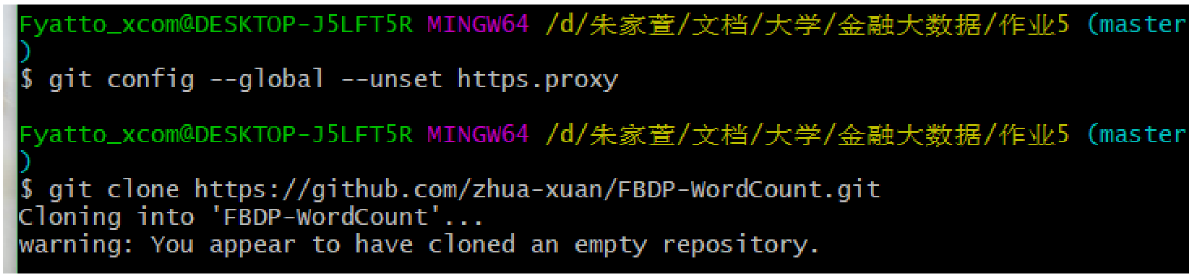
三、上传至github

从前都是上传至gitee，这次上传至github克隆时出现错误，

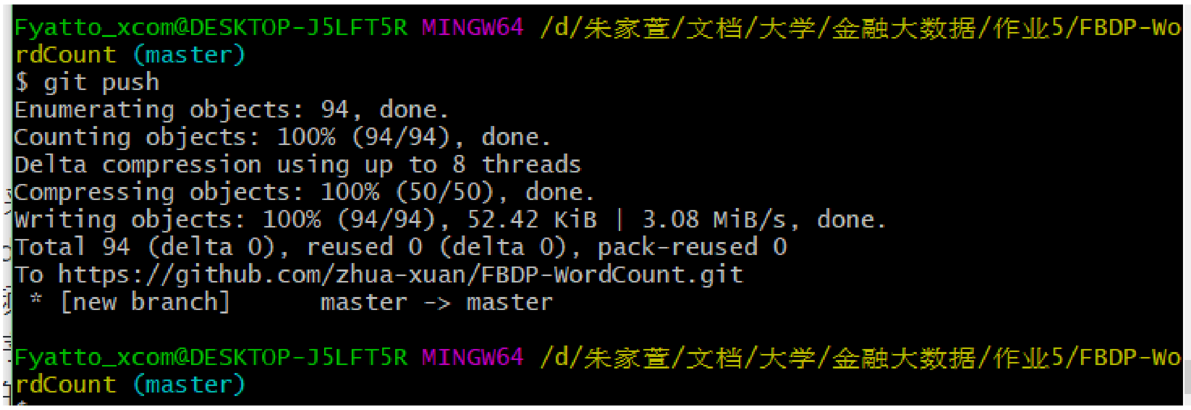


```
$ git clone https://github.com/zhua-xuan/FBDP-WordCount.git
Cloning into 'FBDP-WordCount'...
error: RPC failed; curl 28 OpenSSL SSL_read: Connection was reset, errno 10054
fatal: expected flush after ref listing
```

解决：



```
Fyatto_xcom@DESKTOP-J5LFT5R MINGW64 /d/朱家萱/文档/大学/金融大数据/作业5 (master
)
$ git config --global --unset https.proxy

Fyatto_xcom@DESKTOP-J5LFT5R MINGW64 /d/朱家萱/文档/大学/金融大数据/作业5 (master
)
$ git clone https://github.com/zhua-xuan/FBDP-WordCount.git
Cloning into 'FBDP-WordCount'...
warning: You appear to have cloned an empty repository.
```



```
Fyatto_xcom@DESKTOP-J5LFT5R MINGW64 /d/朱家萱/文档/大学/金融大数据/作业5/FBDP-Wo
rdCount (master)
$ git push
Enumerating objects: 94, done.
Counting objects: 100% (94/94), done.
Delta compression using up to 8 threads
Compressing objects: 100% (50/50), done.
Writing objects: 100% (94/94), 52.42 KiB | 3.08 MiB/s, done.
Total 94 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/zhua-xuan/FBDP-WordCount.git
 * [new branch]      master -> master

Fyatto_xcom@DESKTOP-J5LFT5R MINGW64 /d/朱家萱/文档/大学/金融大数据/作业5/FBDP-Wo
rdCount (master)
```