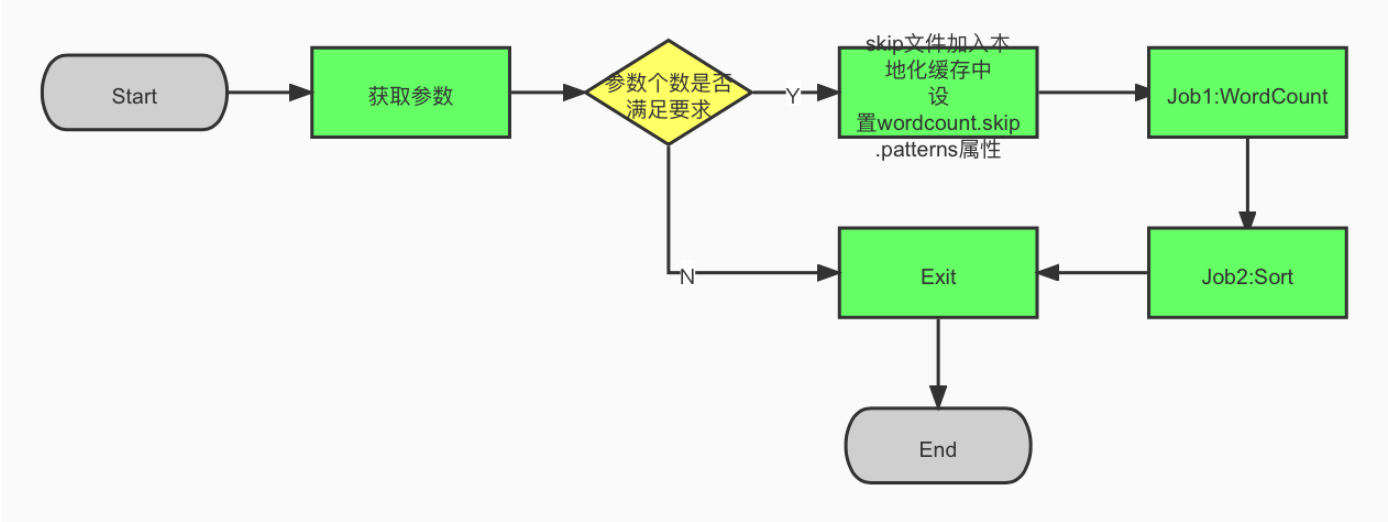
作业5 mapreduce 之 WordCount + Sort

一、设计思路

1. 总思路流程图及总设计思路

1.1总流程图：

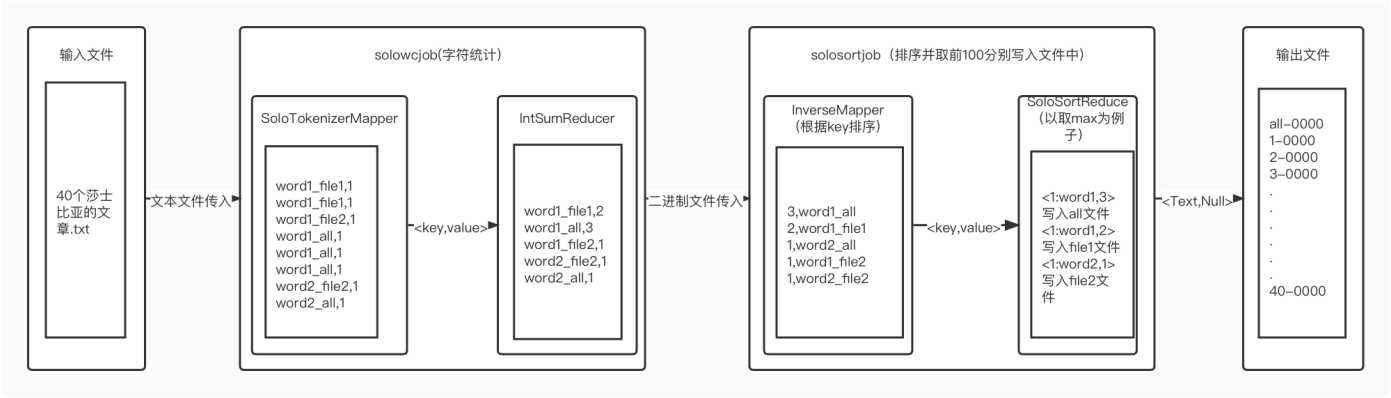


1.2设计思路：

本题主要解决两个问题，第一个是多文件的字符统计，第二个是根据value排序并按照格式要求输出。

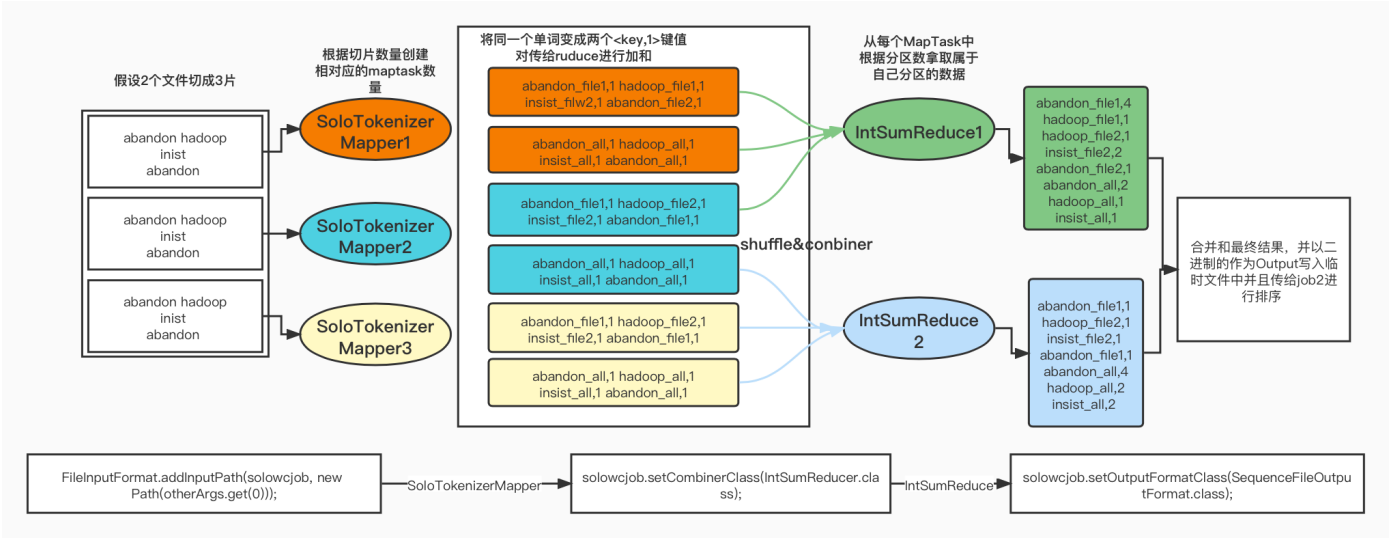
- 问题1:由于本题需要在多文件以及单文件分别进行排序输出，所以在字符统计时的map函数中，我们可以在Hadoop官网的的WordCount2.0基础上进行改写，即在map阶段，将文件名加入<key，value>中的key，key写为word------filename，以此识别不同文件中的单词，这样在使用reduce函数进行词频统计时就可以统计在某个文件中单词出现的次数，考虑到本题还有一个需求，就是统计所有文件的前100个高频词，我们可以将其看作一个特殊的单独文件，即“all”文件，即对于每一个符合统计要求的单词，我们将其写为两个<key,value>，一个key时 word------filename，另一个是word------all，这样进行词频统计时就可以同时完成作业要求中的两个任务。
- 问题2: 在解决问题1的基础上，即我们已经有一个临时文件存储所有<word+filename, num>的二进制统计文件，在此基础上，我们要完成第二个任务：排序并且输出前100个高频词汇到不同文件夹中。
    - 排序方法：利用已有的class：InverseMapper.class交换<word+filename, num>和shuffle阶段的自动根据key进行排序的特性，这样传给reduce节点的就是格式为<num,word+filename>且根据num排好序的键值对。
    - 将reduce节点处理完的数据输出不同文件方法：利用value值，即word+filename中的filename以及MultipleOutputs，将传入reduce节点的键值对，根据filename写到不同文件中去。
    - 取每个文件的前100个方法：利用hashmap的key不可重复特性，统计已经写入文件的键值对属于哪个文件，如果filename已经在hashmap中作为key，那么 value+1 ； 如果该filename的value已经100了，就不再写入。如果filename不在hashmap中，那么把这个filename写进去，并且rank设置为0，如果一共有filecount个键并且每个value都是100，break。

2. Job1+Job2流程图：



3. Job1流程图及两个Class设计思路：

3.1 流程图



3.2SoloTokenizerMapper

- setup：读取配置文件，将停词文件中的单词分别读出到patternsToSkip、patternsToStop，供map函数处理value值。
- parseStopFile：将停词文件stop-word-list.txt中的单词写入patternsToStop
- parseSkipFile：将符号文件punctuation.txt中的字符写入patternsToSkip
- map主要：
  - 忽略大小写：`String line = value.toString().toLowerCase()`
  - 满足patternsToSkip的pattern都过滤掉：

```
for (String pattern : patternsToSkip)
   line = line.replaceAll(pattern, " ");
```

  - 在停词范围的都过滤掉：

```
if(patternsToStop.contains(nextword)){
  continue;}
```

- 获取当前文件名：

```
FileSplit fileSplit = (FileSplit)context.getInputSplit();
String textName = fileSplit.getPath().getName();
```

- 转为<key,value>

```
String allword = nextword + "------all";
 if(nextword.length() >= 3) {
    word.set(word + "------" + textName);
    context.write(word, one);
    context.write(new Text(allword),one);
```

3.3IntSumReducer.Class

- 和原WordCountv2.0没有变化

3.4main函数设置：

```
        solowcjob.setMapperClass(SoloTokenizerMapper.class);
        solowcjob.setCombinerClass(IntSumReducer.class);
        solowcjob.setReducerClass(IntSumReducer.class);
        solowcjob.setOutputKeyClass(Text.class);
        solowcjob.setOutputValueClass(IntWritable.class);
        solowcjob.setOutputFormatClass(SequenceFileOutputFormat.class);
```

4. Job2的Class设计思路：

- InverseMapper.class
  - 已有class
- SoloSortReducer.Class
  - 设置多文件输出：

```
        private MultipleOutputs<Text, NullWritable> mos=null;
        public void setup(Context context) throws IOException
        {
            mos=new MultipleOutputs<Text,NullWritable>(context);
        }
        public void cleanup(Context context)
        {
            try {
                mos.close();
            } catch (IOException | InterruptedException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
            }
        }
```

```
Configuration conf = context.getConfiguration();
int filenum = Integer.parseInt(conf.get("fileCount")); //读取配置文件中的需要输出的文件的个数。
mos.write(filename,new Text(sum +": "+word+","+key),NullWritable.get());
```

- 设置write条件：`Map<String, Integer> filemap = new HashMap<String,Integer>();`

  // 如果该文件名已经在filemap中作为key，那么 原来的value+1 ；

  // 如果该文件名的value已经100了，就不再写入，跳过；

  // 如果该文件名不在filemap的key中，那么把这个filename写进去，并且value设置为0；

  // 如果一共有filecount（从配置文件中读入）个键并且每个键的value都是100，那么break出去。

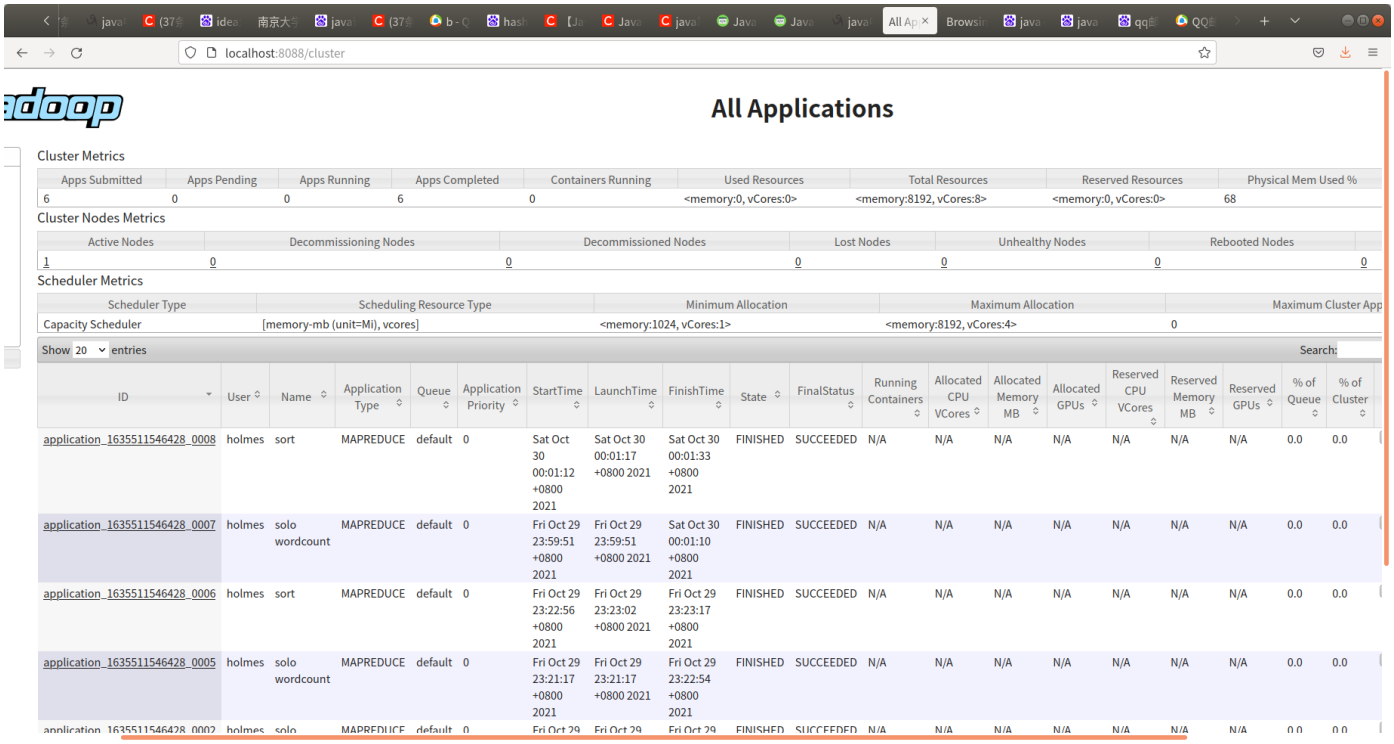
- main函数设置：

```
        FileInputFormat.addInputPath(solosortjob, tempDir2);
        solosortjob.setInputFormatClass(SequenceFileInputFormat.class);
        solosortjob.setMapperClass(InverseMapper.class);
        solosortjob.setReducerClass(SoloSortReducer.class);
        FileOutputFormat.setOutputPath(solosortjob, new Path(otherArgs.get(1)));
        for (String filename : allfile) {
            MultipleOutputs.addNamedOutput(solosortjob, filename,
TextOutputFormat.class,Text.class, NullWritable.class);
        }
```

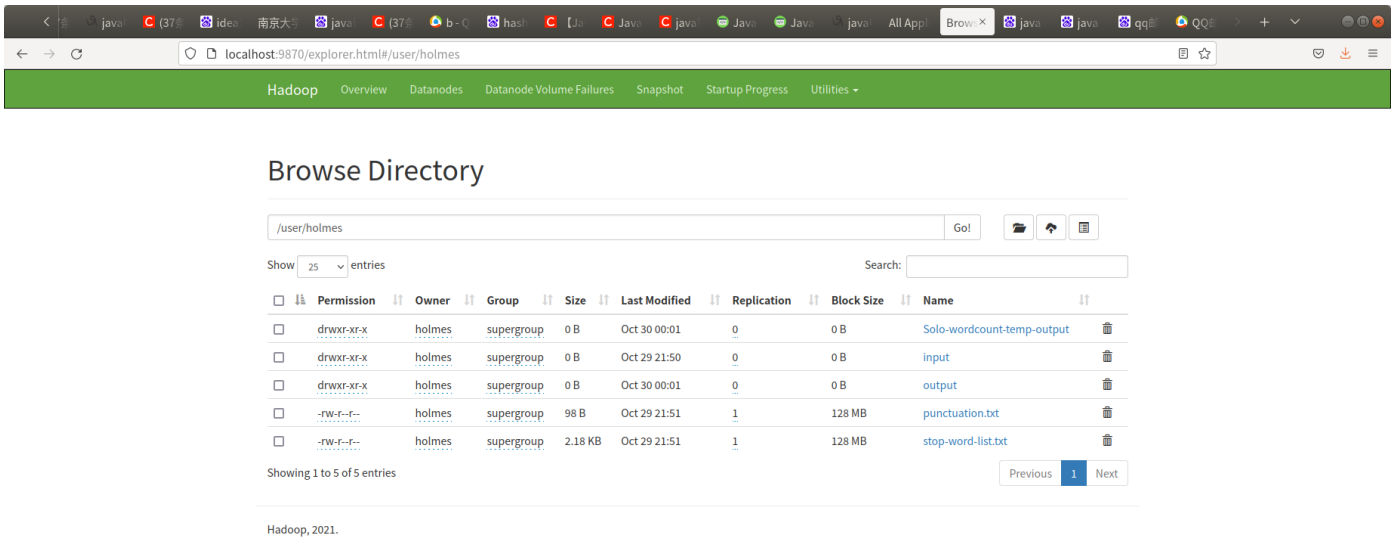
二、运行截图

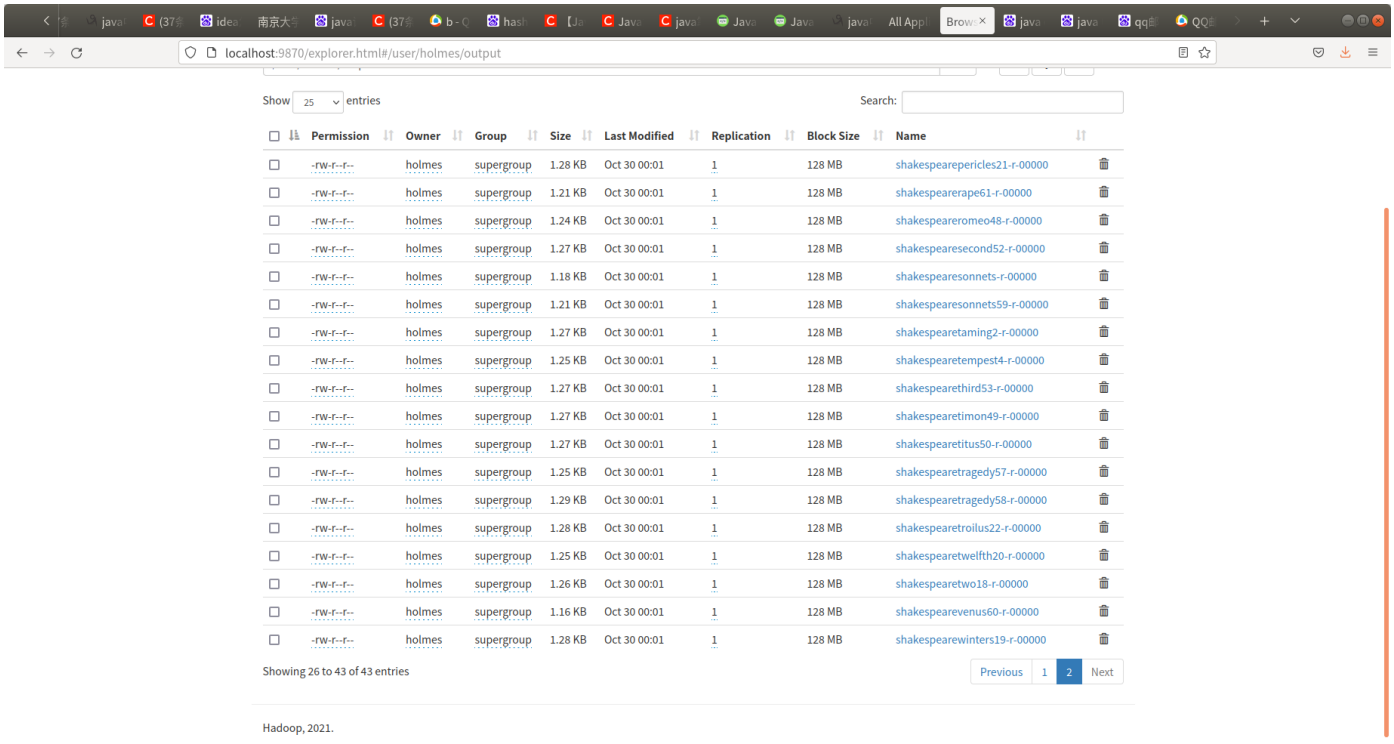
1. 伪分布式运行截图

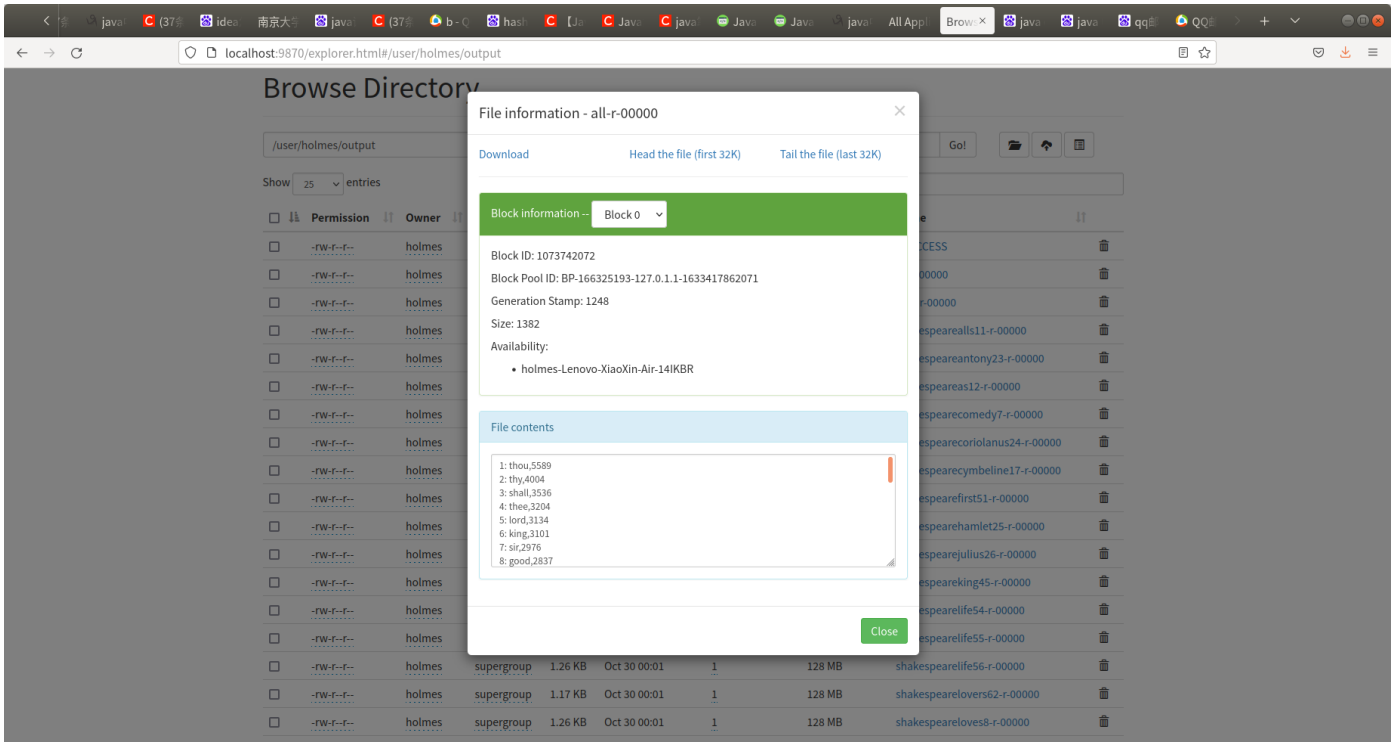
1.1 8088端口WEB截图



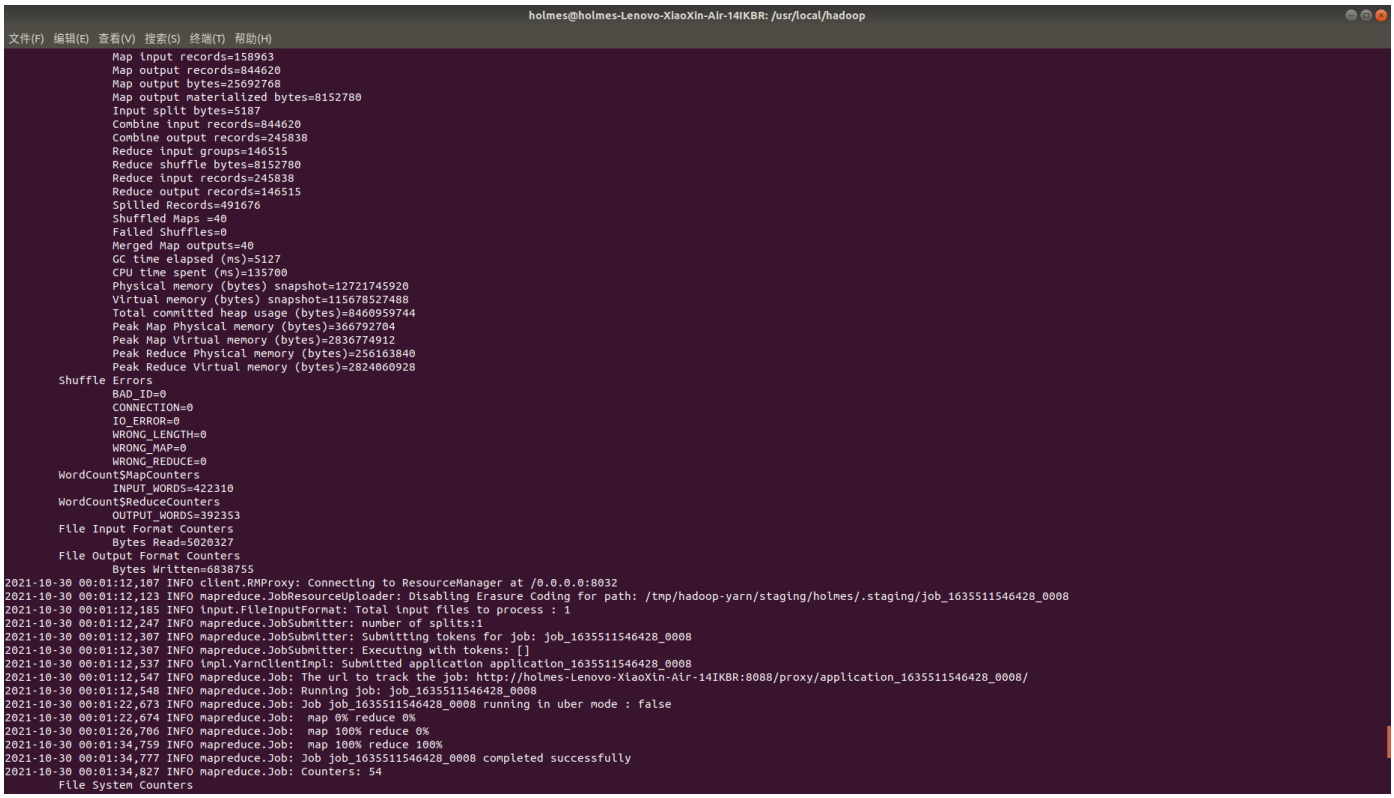
1.2 9870端口运行截图





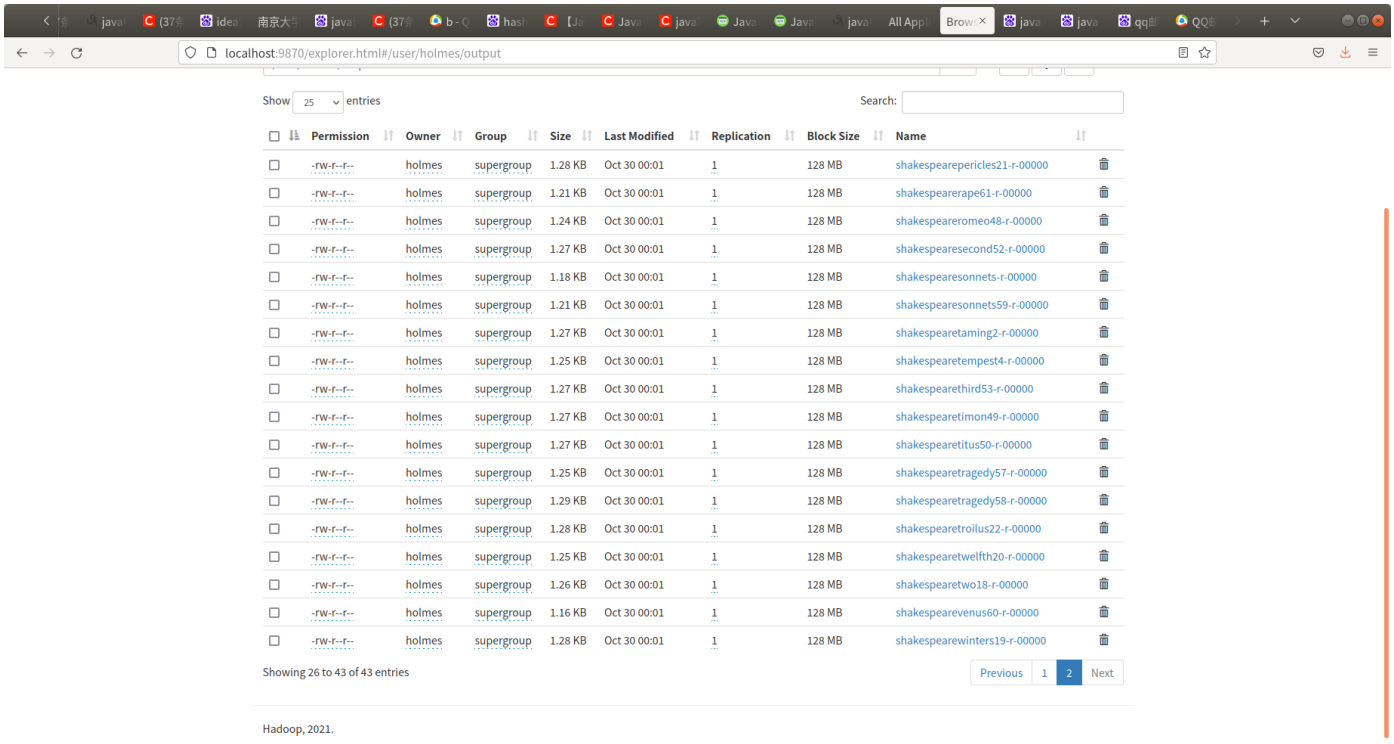


1.3 终端运行截图

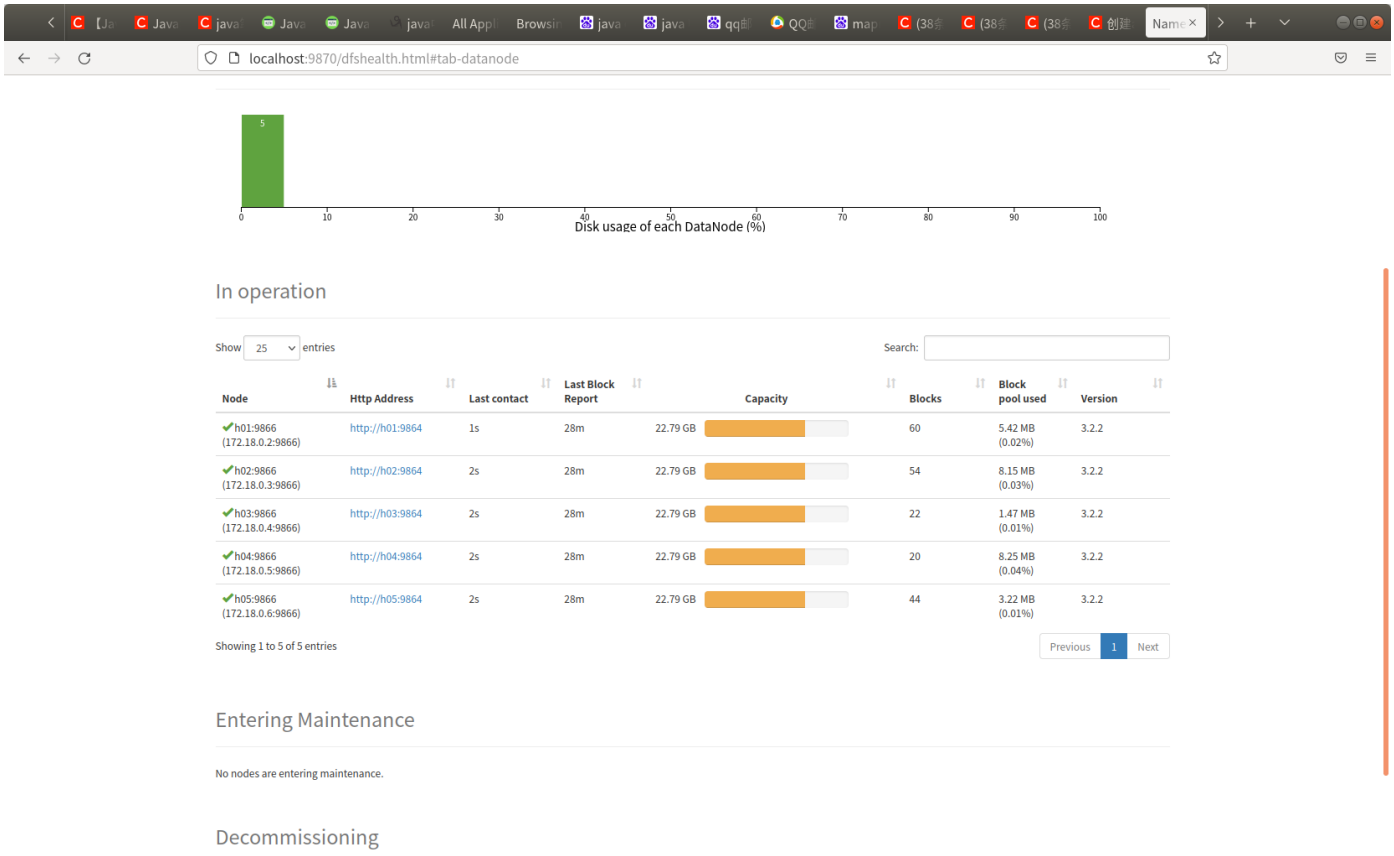


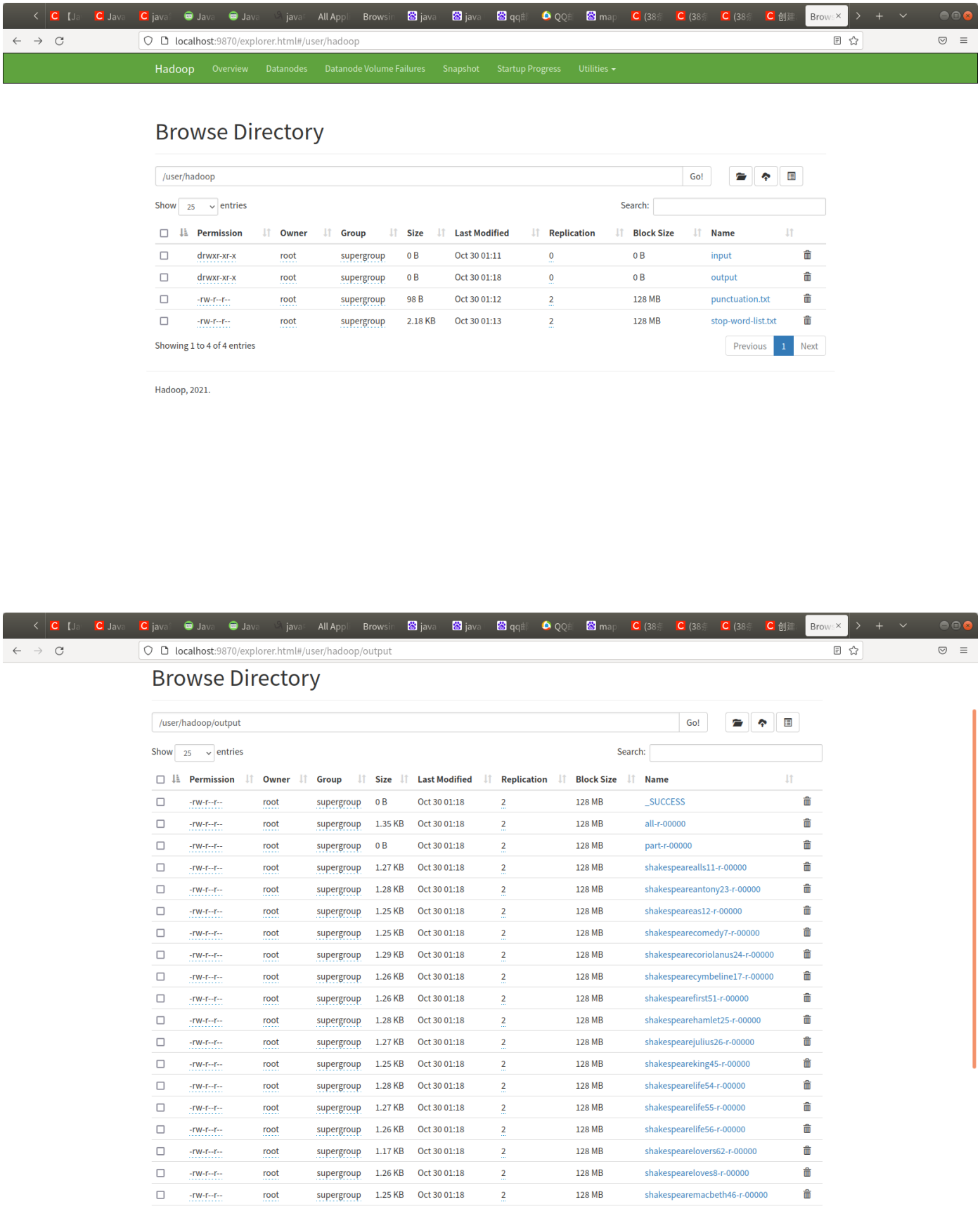
2. docker集群运行截图

2.1 8088端口WEB截图



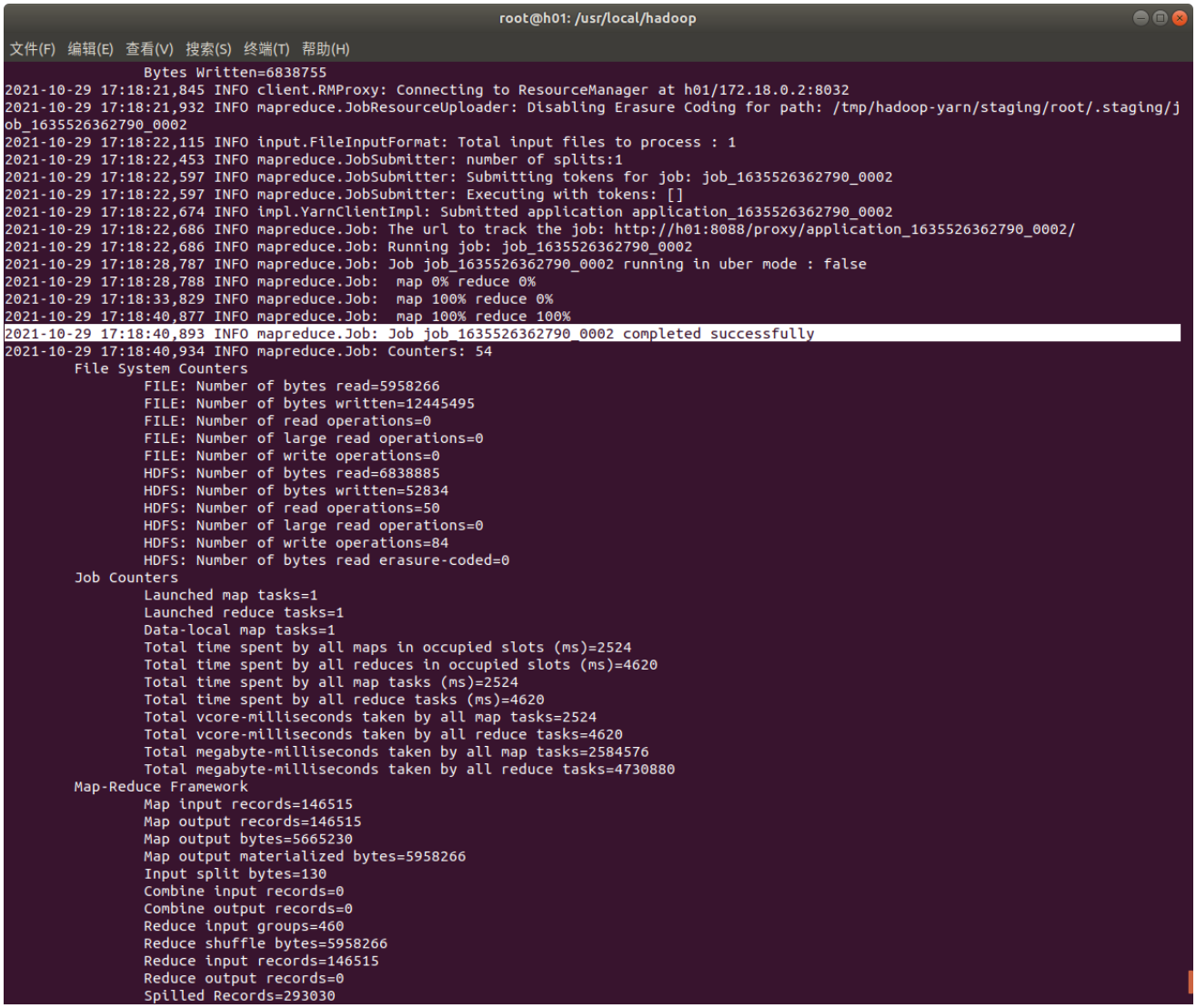
2.2 9870端口节点运行状态




localhost:9870/explorer.html#/user/hadoop
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/user/hadoop
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
drwxr-xr-x root supergroup 0 B Oct 30 01:11 0 0 B input
drwxr-xr-x root supergroup 0 B Oct 30 01:18 0 0 B output
-rw-r--r-- root supergroup 98 B Oct 30 01:12 2 128 MB punctuation.txt
-rw-r--r-- root supergroup 2.18 KB Oct 30 01:13 2 128 MB stop-word-list.txt
Showing 1 to 4 of 4 entries
Previous
1
Next
Hadoop, 2021.

localhost:9870/explorer.html#/user/hadoop/output
Browse Directory
/user/hadoop/output
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r-- root supergroup 0 B Oct 30 01:18 2 128 MB _SUCCESS
-rw-r--r-- root supergroup 1.35 KB Oct 30 01:18 2 128 MB all-r-00000
-rw-r--r-- root supergroup 0 B Oct 30 01:18 2 128 MB part-r-00000
-rw-r--r-- root supergroup 1.27 KB Oct 30 01:18 2 128 MB shakespearealls11-r-00000
-rw-r--r-- root supergroup 1.28 KB Oct 30 01:18 2 128 MB shakespeareantony23-r-00000
-rw-r--r-- root supergroup 1.25 KB Oct 30 01:18 2 128 MB shakespeareas12-r-00000
-rw-r--r-- root supergroup 1.25 KB Oct 30 01:18 2 128 MB shakespearecomedy7-r-00000
-rw-r--r-- root supergroup 1.29 KB Oct 30 01:18 2 128 MB shakespearecoriolanus24-r-00000
-rw-r--r-- root supergroup 1.26 KB Oct 30 01:18 2 128 MB shakespearecymbeline17-r-00000
-rw-r--r-- root supergroup 1.26 KB Oct 30 01:18 2 128 MB shakespearefirst51-r-00000
-rw-r--r-- root supergroup 1.28 KB Oct 30 01:18 2 128 MB shakespearehamlet25-r-00000
-rw-r--r-- root supergroup 1.27 KB Oct 30 01:18 2 128 MB shakespearejulius26-r-00000
-rw-r--r-- root supergroup 1.25 KB Oct 30 01:18 2 128 MB shakespeareking45-r-00000
-rw-r--r-- root supergroup 1.28 KB Oct 30 01:18 2 128 MB shakespearelife54-r-00000
-rw-r--r-- root supergroup 1.27 KB Oct 30 01:18 2 128 MB shakespearelife55-r-00000
-rw-r--r-- root supergroup 1.26 KB Oct 30 01:18 2 128 MB shakespearelife56-r-00000
-rw-r--r-- root supergroup 1.17 KB Oct 30 01:18 2 128 MB shakespearelovers62-r-00000
-rw-r--r-- root supergroup 1.26 KB Oct 30 01:18 2 128 MB shakespeareloves8-r-00000
-rw-r--r-- root supergroup 1.25 KB Oct 30 01:18 2 128 MB shakespearemacbeth46-r-00000

2.3 终端运行成功截图

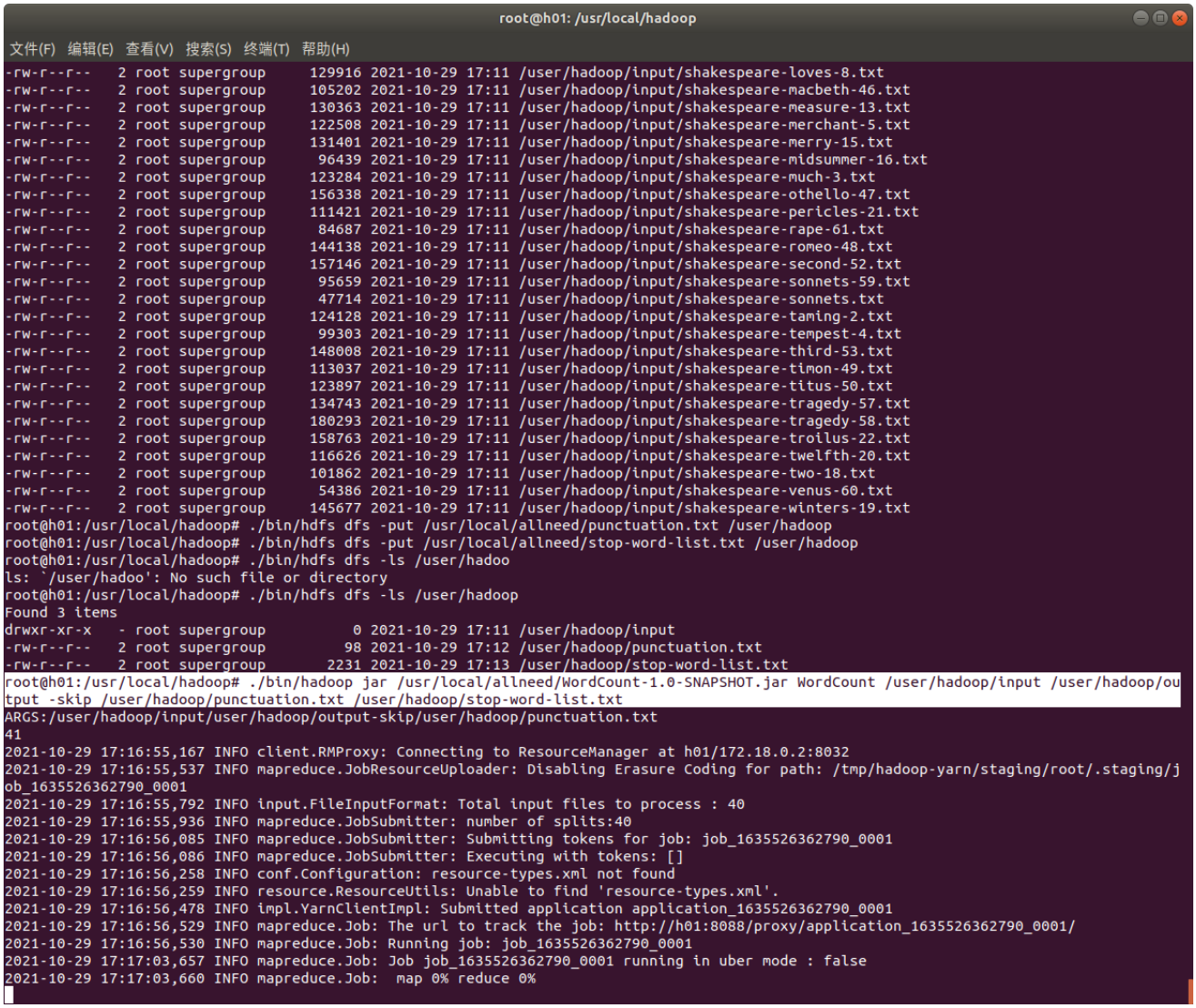

root@h01: /usr/local/hadoop
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
Bytes Written=6838755
2021-10-29 17:18:21,845 INFO client.RMProxy: Connecting to ResourceManager at h01/172.18.0.2:8032
2021-10-29 17:18:21,932 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1635526362790_0002
2021-10-29 17:18:22,115 INFO input.FileInputFormat: Total input files to process : 1
2021-10-29 17:18:22,453 INFO mapreduce.JobSubmitter: number of splits:1
2021-10-29 17:18:22,597 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635526362790_0002
2021-10-29 17:18:22,597 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-29 17:18:22,674 INFO impl.YarnClientImpl: Submitted application application_1635526362790_0002
2021-10-29 17:18:22,686 INFO mapreduce.Job: The url to track the job: http://h01:8088/proxy/application_1635526362790_0002/
2021-10-29 17:18:22,686 INFO mapreduce.Job: Running job: job_1635526362790_0002
2021-10-29 17:18:28,787 INFO mapreduce.Job: Job job_1635526362790_0002 running in uber mode : false
2021-10-29 17:18:28,788 INFO mapreduce.Job: map 0% reduce 0%
2021-10-29 17:18:33,829 INFO mapreduce.Job: map 100% reduce 0%
2021-10-29 17:18:40,877 INFO mapreduce.Job: map 100% reduce 100%
2021-10-29 17:18:40,893 INFO mapreduce.Job: Job job_1635526362790_0002 completed successfully
2021-10-29 17:18:40,934 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=5958266
FILE: Number of bytes written=12445495
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=6838885
HDFS: Number of bytes written=52834
HDFS: Number of read operations=50
HDFS: Number of large read operations=0
HDFS: Number of write operations=84
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=1
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=2524
Total time spent by all reduces in occupied slots (ms)=4620
Total time spent by all map tasks (ms)=2524
Total time spent by all reduce tasks (ms)=4620
Total vcore-milliseconds taken by all map tasks=2524
Total vcore-milliseconds taken by all reduce tasks=4620
Total megabyte-milliseconds taken by all map tasks=2584576
Total megabyte-milliseconds taken by all reduce tasks=4730880
Map-Reduce Framework
Map input records=146515
Map output records=146515
Map output bytes=5665230
Map output materialized bytes=5958266
Input split bytes=130
Combine input records=0
Combine output records=0
Reduce input groups=460
Reduce shuffle bytes=5958266
Reduce input records=146515
Reduce output records=0
Spilled Records=293030

```
root@h01: /usr/local/hadoop
文件(F)  编辑(E)  查看(V)  搜索(S)  终端(T)  帮助(H)
2021-10-29 17:17:03,660 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-29 17:17:46,184 INFO mapreduce.Job:  map 13% reduce 0%
2021-10-29 17:17:48,389 INFO mapreduce.Job:  map 15% reduce 0%
2021-10-29 17:17:58,743 INFO mapreduce.Job:  map 17% reduce 0%
2021-10-29 17:18:00,452 INFO mapreduce.Job:  map 22% reduce 0%
2021-10-29 17:18:01,495 INFO mapreduce.Job:  map 25% reduce 0%
2021-10-29 17:18:02,525 INFO mapreduce.Job:  map 28% reduce 0%
2021-10-29 17:18:03,548 INFO mapreduce.Job:  map 29% reduce 0%
2021-10-29 17:18:05,611 INFO mapreduce.Job:  map 31% reduce 0%
2021-10-29 17:18:06,796 INFO mapreduce.Job:  map 34% reduce 0%
2021-10-29 17:18:07,830 INFO mapreduce.Job:  map 39% reduce 0%
2021-10-29 17:18:08,868 INFO mapreduce.Job:  map 50% reduce 0%
2021-10-29 17:18:09,916 INFO mapreduce.Job:  map 60% reduce 0%
2021-10-29 17:18:10,980 INFO mapreduce.Job:  map 64% reduce 0%
2021-10-29 17:18:12,001 INFO mapreduce.Job:  map 67% reduce 0%
2021-10-29 17:18:13,036 INFO mapreduce.Job:  map 69% reduce 0%
2021-10-29 17:18:14,069 INFO mapreduce.Job:  map 74% reduce 0%
2021-10-29 17:18:15,090 INFO mapreduce.Job:  map 78% reduce 0%
2021-10-29 17:18:16,104 INFO mapreduce.Job:  map 82% reduce 0%
2021-10-29 17:18:17,114 INFO mapreduce.Job:  map 85% reduce 0%
2021-10-29 17:18:18,122 INFO mapreduce.Job:  map 93% reduce 0%
2021-10-29 17:18:19,126 INFO mapreduce.Job:  map 100% reduce 0%
2021-10-29 17:18:21,143 INFO mapreduce.Job:  map 100% reduce 100%
2021-10-29 17:18:21,164 INFO mapreduce.Job: Job job_1635526362790_0001 completed successfully
2021-10-29 17:18:21,703 INFO mapreduce.Job: Counters: 58
        File System Counters
                FILE: Number of bytes read=8152546
                FILE: Number of bytes written=26012543
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5025274
                HDFS: Number of bytes written=6838755
                HDFS: Number of read operations=125
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=2
                Launched map tasks=41
                Launched reduce tasks=1
                Data-local map tasks=34
                Rack-local map tasks=7
                Total time spent by all maps in occupied slots (ms)=2360045
                Total time spent by all reduces in occupied slots (ms)=31716
                Total time spent by all map tasks (ms)=2360045
                Total time spent by all reduce tasks (ms)=31716
                Total vcore-milliseconds taken by all map tasks=2360045
                Total vcore-milliseconds taken by all reduce tasks=31716
                Total megabyte-milliseconds taken by all map tasks=2416686080
                Total megabyte-milliseconds taken by all reduce tasks=32477184
        Map-Reduce Framework
                Map input records=158963
                Map output records=844620
```

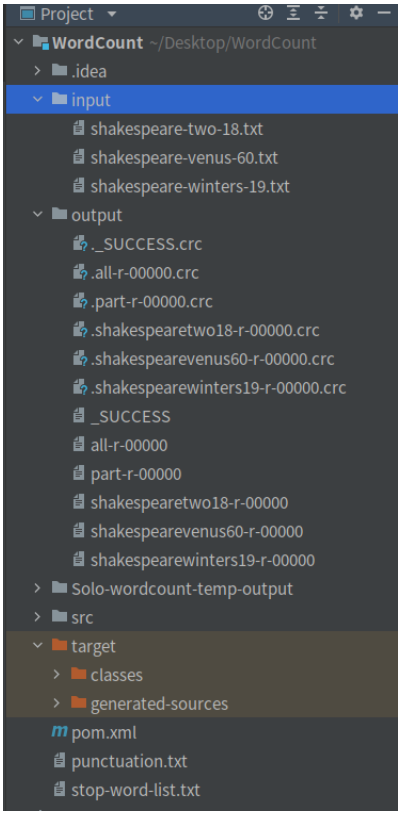
2.4 传入文件及运行截图

```
root@h01: /usr/local/hadoop
文件(F)  编辑(E)  查看(V)  搜索(S)  终端(T)  帮助(H)

SUBCOMMAND may print help when invoked w/o parameters or with -h.
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -put /usr/local/allneed/shakespeare-txt/*.txt /user/hadoop/input
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -ls /user/hadoop/input
Found 40 items
-rw-r--r--   2 root supergroup     135197 2021-10-29 17:11 /user/hadoop/input/shakespeare-alls-11.txt
-rw-r--r--   2 root supergroup     158248 2021-10-29 17:11 /user/hadoop/input/shakespeare-antony-23.txt
-rw-r--r--   2 root supergroup     125011 2021-10-29 17:11 /user/hadoop/input/shakespeare-as-12.txt
-rw-r--r--   2 root supergroup      89439 2021-10-29 17:11 /user/hadoop/input/shakespeare-comedy-7.txt
-rw-r--r--   2 root supergroup     168133 2021-10-29 17:11 /user/hadoop/input/shakespeare-coriolanus-24.txt
-rw-r--r--   2 root supergroup     165009 2021-10-29 17:11 /user/hadoop/input/shakespeare-cymbeline-17.txt
-rw-r--r--   2 root supergroup     144860 2021-10-29 17:11 /user/hadoop/input/shakespeare-first-51.txt
-rw-r--r--   2 root supergroup     182399 2021-10-29 17:11 /user/hadoop/input/shakespeare-hamlet-25.txt
-rw-r--r--   2 root supergroup     117902 2021-10-29 17:11 /user/hadoop/input/shakespeare-julius-26.txt
-rw-r--r--   2 root supergroup     157894 2021-10-29 17:11 /user/hadoop/input/shakespeare-king-45.txt
-rw-r--r--   2 root supergroup     154933 2021-10-29 17:11 /user/hadoop/input/shakespeare-life-54.txt
-rw-r--r--   2 root supergroup     148351 2021-10-29 17:11 /user/hadoop/input/shakespeare-life-55.txt
-rw-r--r--   2 root supergroup     122448 2021-10-29 17:11 /user/hadoop/input/shakespeare-life-56.txt
-rw-r--r--   2 root supergroup      14364 2021-10-29 17:11 /user/hadoop/input/shakespeare-lovers-62.txt
-rw-r--r--   2 root supergroup     129916 2021-10-29 17:11 /user/hadoop/input/shakespeare-loves-8.txt
-rw-r--r--   2 root supergroup     105202 2021-10-29 17:11 /user/hadoop/input/shakespeare-macbeth-46.txt
-rw-r--r--   2 root supergroup     130363 2021-10-29 17:11 /user/hadoop/input/shakespeare-measure-13.txt
-rw-r--r--   2 root supergroup     122508 2021-10-29 17:11 /user/hadoop/input/shakespeare-merchant-5.txt
-rw-r--r--   2 root supergroup     131401 2021-10-29 17:11 /user/hadoop/input/shakespeare-merry-15.txt
-rw-r--r--   2 root supergroup      96439 2021-10-29 17:11 /user/hadoop/input/shakespeare-midsummer-16.txt
-rw-r--r--   2 root supergroup     123284 2021-10-29 17:11 /user/hadoop/input/shakespeare-much-3.txt
-rw-r--r--   2 root supergroup     156338 2021-10-29 17:11 /user/hadoop/input/shakespeare-othello-47.txt
-rw-r--r--   2 root supergroup     111421 2021-10-29 17:11 /user/hadoop/input/shakespeare-pericles-21.txt
-rw-r--r--   2 root supergroup      84687 2021-10-29 17:11 /user/hadoop/input/shakespeare-rape-61.txt
-rw-r--r--   2 root supergroup     144138 2021-10-29 17:11 /user/hadoop/input/shakespeare-romeo-48.txt
-rw-r--r--   2 root supergroup     157146 2021-10-29 17:11 /user/hadoop/input/shakespeare-second-52.txt
-rw-r--r--   2 root supergroup      95659 2021-10-29 17:11 /user/hadoop/input/shakespeare-sonnets-59.txt
-rw-r--r--   2 root supergroup      47714 2021-10-29 17:11 /user/hadoop/input/shakespeare-sonnets.txt
-rw-r--r--   2 root supergroup     124128 2021-10-29 17:11 /user/hadoop/input/shakespeare-taming-2.txt
-rw-r--r--   2 root supergroup      99303 2021-10-29 17:11 /user/hadoop/input/shakespeare-tempest-4.txt
-rw-r--r--   2 root supergroup     148008 2021-10-29 17:11 /user/hadoop/input/shakespeare-third-53.txt
-rw-r--r--   2 root supergroup     113037 2021-10-29 17:11 /user/hadoop/input/shakespeare-timon-49.txt
-rw-r--r--   2 root supergroup     123897 2021-10-29 17:11 /user/hadoop/input/shakespeare-titus-50.txt
-rw-r--r--   2 root supergroup     134743 2021-10-29 17:11 /user/hadoop/input/shakespeare-tragedy-57.txt
-rw-r--r--   2 root supergroup     180293 2021-10-29 17:11 /user/hadoop/input/shakespeare-tragedy-58.txt
-rw-r--r--   2 root supergroup     158763 2021-10-29 17:11 /user/hadoop/input/shakespeare-troilus-22.txt
-rw-r--r--   2 root supergroup     116626 2021-10-29 17:11 /user/hadoop/input/shakespeare-twelfth-20.txt
-rw-r--r--   2 root supergroup     101862 2021-10-29 17:11 /user/hadoop/input/shakespeare-two-18.txt
-rw-r--r--   2 root supergroup      54386 2021-10-29 17:11 /user/hadoop/input/shakespeare-venus-60.txt
-rw-r--r--   2 root supergroup     145677 2021-10-29 17:11 /user/hadoop/input/shakespeare-winters-19.txt
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -put /usr/local/allneed/punctuation.txt /user/hadoop
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -put /usr/local/allneed/stop-word-list.txt /user/hadoop
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -ls /user/hadoo
ls: `/user/hadoo': No such file or directory
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -ls /user/hadoop
Found 3 items
drwxr-xr-x   - root supergroup          0 2021-10-29 17:11 /user/hadoop/input
-rw-r--r--   2 root supergroup         98 2021-10-29 17:12 /user/hadoop/punctuation.txt
-rw-r--r--   2 root supergroup       2231 2021-10-29 17:13 /user/hadoop/stop-word-list.txt
root@h01:/usr/local/hadoop#
```

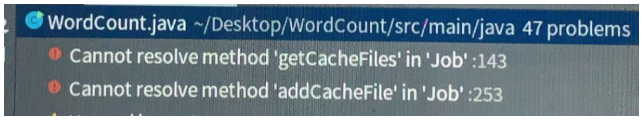

root@h01: /usr/local/hadoop
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
-rw-r--r-- 2 root supergroup 129916 2021-10-29 17:11 /user/hadoop/input/shakespeare-loves-8.txt
-rw-r--r-- 2 root supergroup 105202 2021-10-29 17:11 /user/hadoop/input/shakespeare-macbeth-46.txt
-rw-r--r-- 2 root supergroup 130363 2021-10-29 17:11 /user/hadoop/input/shakespeare-measure-13.txt
-rw-r--r-- 2 root supergroup 122508 2021-10-29 17:11 /user/hadoop/input/shakespeare-merchant-5.txt
-rw-r--r-- 2 root supergroup 131401 2021-10-29 17:11 /user/hadoop/input/shakespeare-merry-15.txt
-rw-r--r-- 2 root supergroup 96439 2021-10-29 17:11 /user/hadoop/input/shakespeare-midsummer-16.txt
-rw-r--r-- 2 root supergroup 123284 2021-10-29 17:11 /user/hadoop/input/shakespeare-much-3.txt
-rw-r--r-- 2 root supergroup 156338 2021-10-29 17:11 /user/hadoop/input/shakespeare-othello-47.txt
-rw-r--r-- 2 root supergroup 111421 2021-10-29 17:11 /user/hadoop/input/shakespeare-pericles-21.txt
-rw-r--r-- 2 root supergroup 84687 2021-10-29 17:11 /user/hadoop/input/shakespeare-rape-61.txt
-rw-r--r-- 2 root supergroup 144138 2021-10-29 17:11 /user/hadoop/input/shakespeare-romeo-48.txt
-rw-r--r-- 2 root supergroup 157146 2021-10-29 17:11 /user/hadoop/input/shakespeare-second-52.txt
-rw-r--r-- 2 root supergroup 95659 2021-10-29 17:11 /user/hadoop/input/shakespeare-sonnets-59.txt
-rw-r--r-- 2 root supergroup 47714 2021-10-29 17:11 /user/hadoop/input/shakespeare-sonnets.txt
-rw-r--r-- 2 root supergroup 124128 2021-10-29 17:11 /user/hadoop/input/shakespeare-taming-2.txt
-rw-r--r-- 2 root supergroup 99303 2021-10-29 17:11 /user/hadoop/input/shakespeare-tempest-4.txt
-rw-r--r-- 2 root supergroup 148008 2021-10-29 17:11 /user/hadoop/input/shakespeare-third-53.txt
-rw-r--r-- 2 root supergroup 113037 2021-10-29 17:11 /user/hadoop/input/shakespeare-timon-49.txt
-rw-r--r-- 2 root supergroup 123897 2021-10-29 17:11 /user/hadoop/input/shakespeare-titus-50.txt
-rw-r--r-- 2 root supergroup 134743 2021-10-29 17:11 /user/hadoop/input/shakespeare-tragedy-57.txt
-rw-r--r-- 2 root supergroup 180293 2021-10-29 17:11 /user/hadoop/input/shakespeare-tragedy-58.txt
-rw-r--r-- 2 root supergroup 158763 2021-10-29 17:11 /user/hadoop/input/shakespeare-troilus-22.txt
-rw-r--r-- 2 root supergroup 116626 2021-10-29 17:11 /user/hadoop/input/shakespeare-twelfth-20.txt
-rw-r--r-- 2 root supergroup 101862 2021-10-29 17:11 /user/hadoop/input/shakespeare-two-18.txt
-rw-r--r-- 2 root supergroup 54386 2021-10-29 17:11 /user/hadoop/input/shakespeare-venus-60.txt
-rw-r--r-- 2 root supergroup 145677 2021-10-29 17:11 /user/hadoop/input/shakespeare-winters-19.txt
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -put /usr/local/allneed/punctuation.txt /user/hadoop
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -put /usr/local/allneed/stop-word-list.txt /user/hadoop
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -ls /user/hadoo
ls: `/user/hadoo': No such file or directory
root@h01:/usr/local/hadoop# ./bin/hdfs dfs -ls /user/hadoop
Found 3 items
drwxr-xr-x - root supergroup 0 2021-10-29 17:11 /user/hadoop/input
-rw-r--r-- 2 root supergroup 98 2021-10-29 17:12 /user/hadoop/punctuation.txt
-rw-r--r-- 2 root supergroup 2231 2021-10-29 17:13 /user/hadoop/stop-word-list.txt
root@h01:/usr/local/hadoop# ./bin/hadoop jar /usr/local/allneed/WordCount-1.0-SNAPSHOT.jar WordCount /user/hadoop/input /user/hadoop/output -skip /user/hadoop/punctuation.txt /user/hadoop/stop-word-list.txt
ARGS:/user/hadoop/input/user/hadoop/output-skip/user/hadoop/punctuation.txt
41
2021-10-29 17:16:55,167 INFO client.RMProxy: Connecting to ResourceManager at h01/172.18.0.2:8032
2021-10-29 17:16:55,537 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1635526362790_0001
2021-10-29 17:16:55,792 INFO input.FileInputFormat: Total input files to process : 40
2021-10-29 17:16:55,936 INFO mapreduce.JobSubmitter: number of splits:40
2021-10-29 17:16:56,085 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635526362790_0001
2021-10-29 17:16:56,086 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-29 17:16:56,258 INFO conf.Configuration: resource-types.xml not found
2021-10-29 17:16:56,259 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-29 17:16:56,478 INFO impl.YarnClientImpl: Submitted application application_1635526362790_0001
2021-10-29 17:16:56,529 INFO mapreduce.Job: The url to track the job: http://h01:8088/proxy/application_1635526362790_0001/
2021-10-29 17:16:56,530 INFO mapreduce.Job: Running job: job_1635526362790_0001
2021-10-29 17:17:03,657 INFO mapreduce.Job: Job job_1635526362790_0001 running in uber mode : false
2021-10-29 17:17:03,660 INFO mapreduce.Job: map 0% reduce 0%

3. 本地文件结构



三、所遇问题

1. 本地运行存在以下问题



   原因：类库版本过低

   解决方法：修改pom.xml文件

2. 原想使用File方法读取目录下所有文件名于allfile文件中，并且计算文件个数写入配置文件中供reduce函数使用，以及可以实现自动MultipleOutputs操作，但是File方法只能在本地运行，无法读取HDFS文件系统中文件，在伪分布式中运行错误，于是舍弃该方法，变为通过该方法读取所有文件名，写入allfile（list）中。

   for (String filename : allfile) {
   MultipleOutputs.addNamedOutput(solosortjob, filename, TextOutputFormat.class,Text.class, NullWritable.class);}

四、总结分析不足之处

1. 性能：
   - 题目仅要求获得前100高频词汇，但是由于我使用InverseMapper.class类，得到所有单词的排序，有很多排序并不需要用到。其次对于后续写入文件时的操作是，跳过已满100的单词，即许多单词是不需要的。
2. 扩展性：
   - 由于无法在main函数中一次性获得所有文件名，所以不得不将所有文件名写入list中给MultipleOutputs.addNamedOutput中使用，因此传入别的文件，本代码需要修改才能运行。

- 只能传入两个停词文件，且按照一定顺序传入，因为stop文件和punc文件处理方式不同。
- 只能输出前一百的高频词汇，无法由用户自行设定。
- 没有 -skip参数，也无法正常运行。