金融大数据处理技术 作业5（任务一）

221275010 屈航

1、设计思路

1.1、 Mapper的设计思路

`Mapper` **重构了一个** `map` **函数。**

对于 `analyst_ratings.csv` 里面每一行数据的处理要抽出来**最后一个字段** `stock`。这可以采 `string.split(",")` 把最后一个字段 `stock` 提取出来，然后加入 `context.write(stock, 1)`。

以下是 `map` **函数**的主要功能语句：

```
tring line = value.toString();
String[] columns = line.split(","); // Assuming CSV format with ',' as delimiter
if (columns.length >= 4) { // Ensure there are enough columns
    stockCode.set(columns[columns.length - 1].trim()); // Assuming stock code is
in the last column
    context.write(stockCode, one);
}
```

注意：实验开始前要对 `analyst_ratings.csv` **中的第一行删去，第一行并不是需要统计的内容。**

1.2、 Reducer的设计思路

`Reducer` **包含有** `reduce` **函数和** `cleanup` **函数。**

1. `reduce` **函数**就是实现对于同一个key的 `Iterable<IntWritable> values` 中的每一个元素也就是统计数求和，最后再将求和结果存入一个叫做 `Map<String, Integer> stockCountMap = new HashMap<>()` 的类里面以待 `cleanup` 函数实现对输出格式的处理。
   以下是 `reduce` 函数的主要功能语句：

   ```
   int sum = 0;
   for (IntWritable val : values) {
       sum += val.get();
   }
   stockCountMap.put(key.toString(), sum);
   ```

2. `cleanup` **函数**就是实现对输出结果的格式控制。首先建立一个列表 `List<Map.Entry<String, Integer>> sortedList = new ArrayList<>(stockCountMap.entrySet())`，然后通过**重构** `sort` **方法**的比较方法实现按照**词频**进行**倒序排列**；最后按照排序好的 `sortedList` 依次按照格式进行输出。
   以下是 `cleanup` **函数**的主要功能语句：

```
List<Map.Entry<String, Integer>> sortedList = new ArrayList<>
(stockCountMap.entrySet());
sortedList.sort((a, b) -> b.getValue().compareTo(a.getValue())); // Sort by
count descending

// Output the results in the required format
int rank = 1;
for (Map.Entry<String, Integer> entry : sortedList) {
    context.write(new Text(rank + ": " + entry.getKey() + ", " +
entry.getValue()), null);
    rank++;
}
```

1.3、项目运行的配置设计

- 此次项目主要使用`Maven`进行项目管理，通过编辑`pom.xml`文件对该项目进行配置。`pom.xml`文件的配置信息包含有该项目需要哪些库文件需要下载，该项目的项目文件有哪些。
- 依次使用`mvn clean install`进行配置，同时还可以使用`mvn compile`对`.class`文件进行生成，`mvn package`实现对项目文件的`.class`文件打包成`jar`文件。
- 将`analyst_ratings.csv`上传至**HDFS**的`/input`文件夹里面，最后运行该项目的`jar`文件，运行命令为：

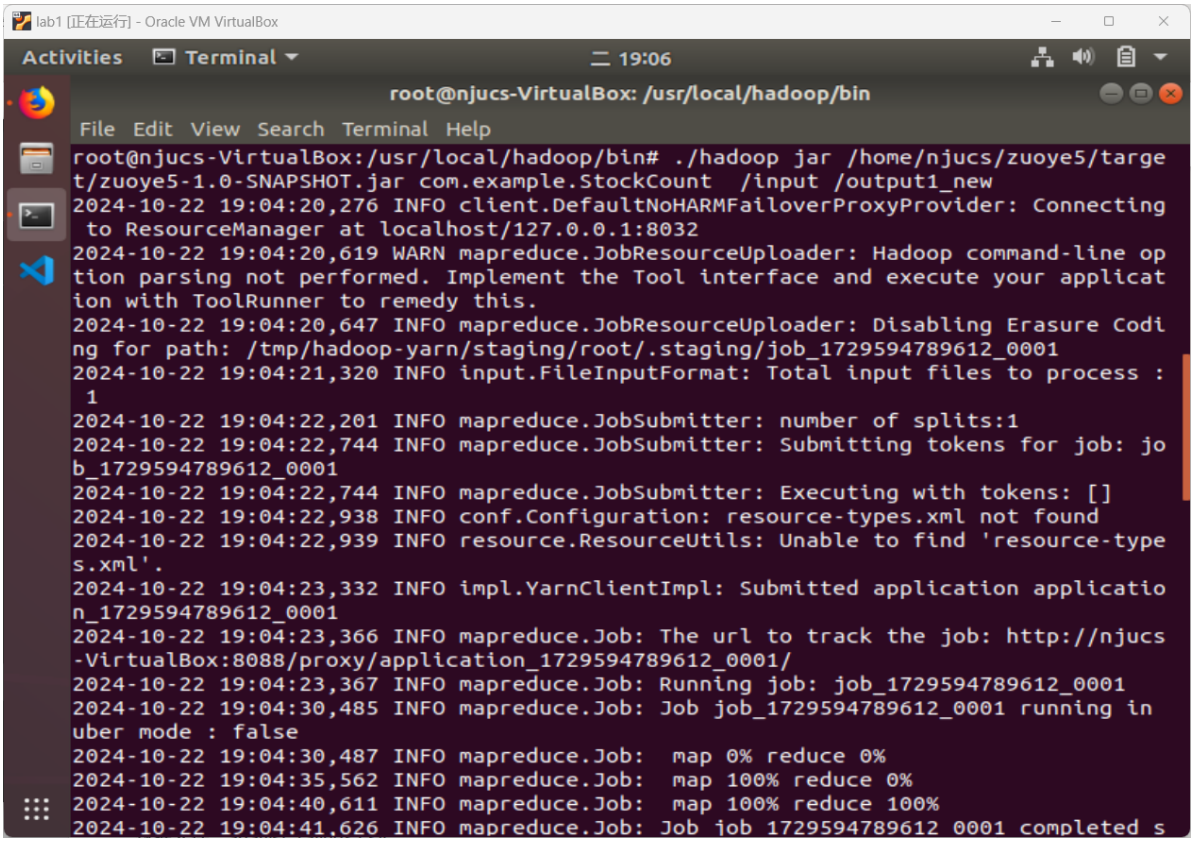
```
./hadoop jar /home/njucs/zuoye5/target/zuoye5-1.0-SNAPSHOT.jar
com.example.StockCount /input /output
```

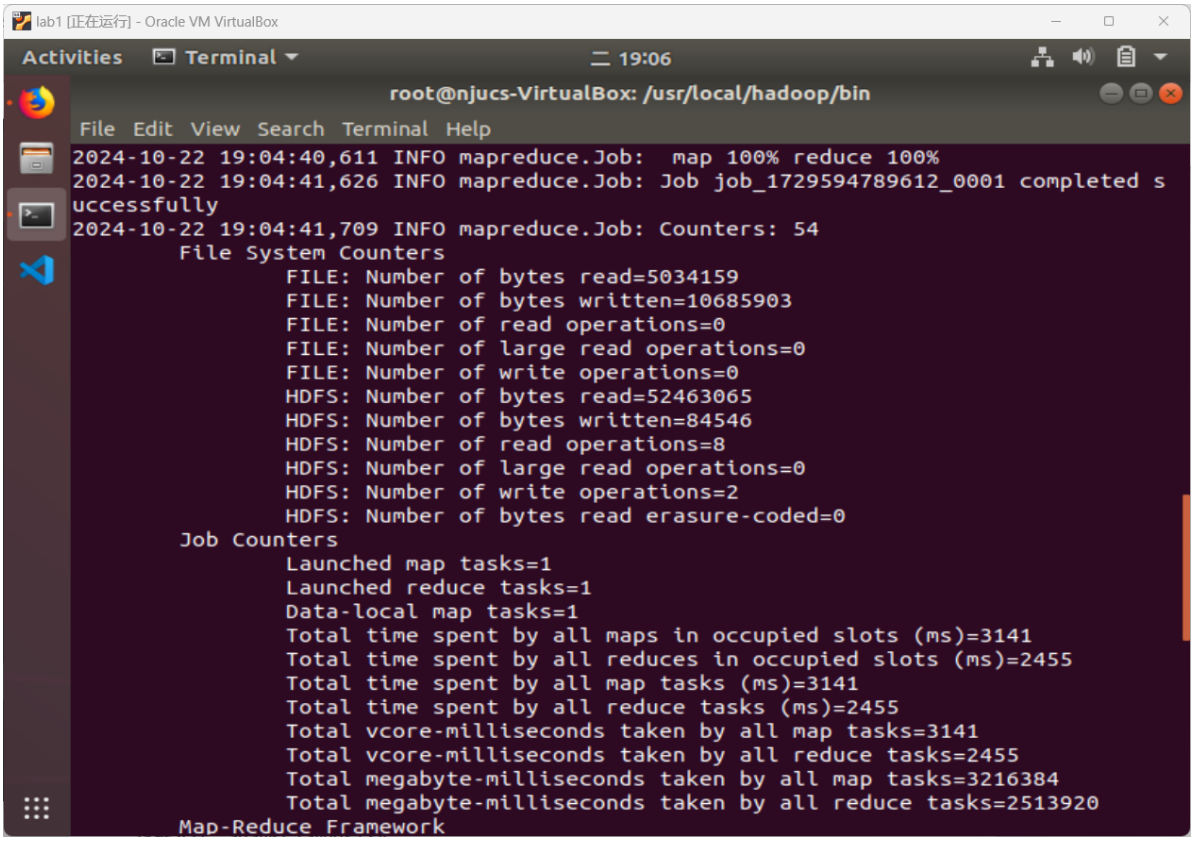
注意要把导出来的 *part-r-00000*解锁，以实现普通用户可以打开，命令如下：

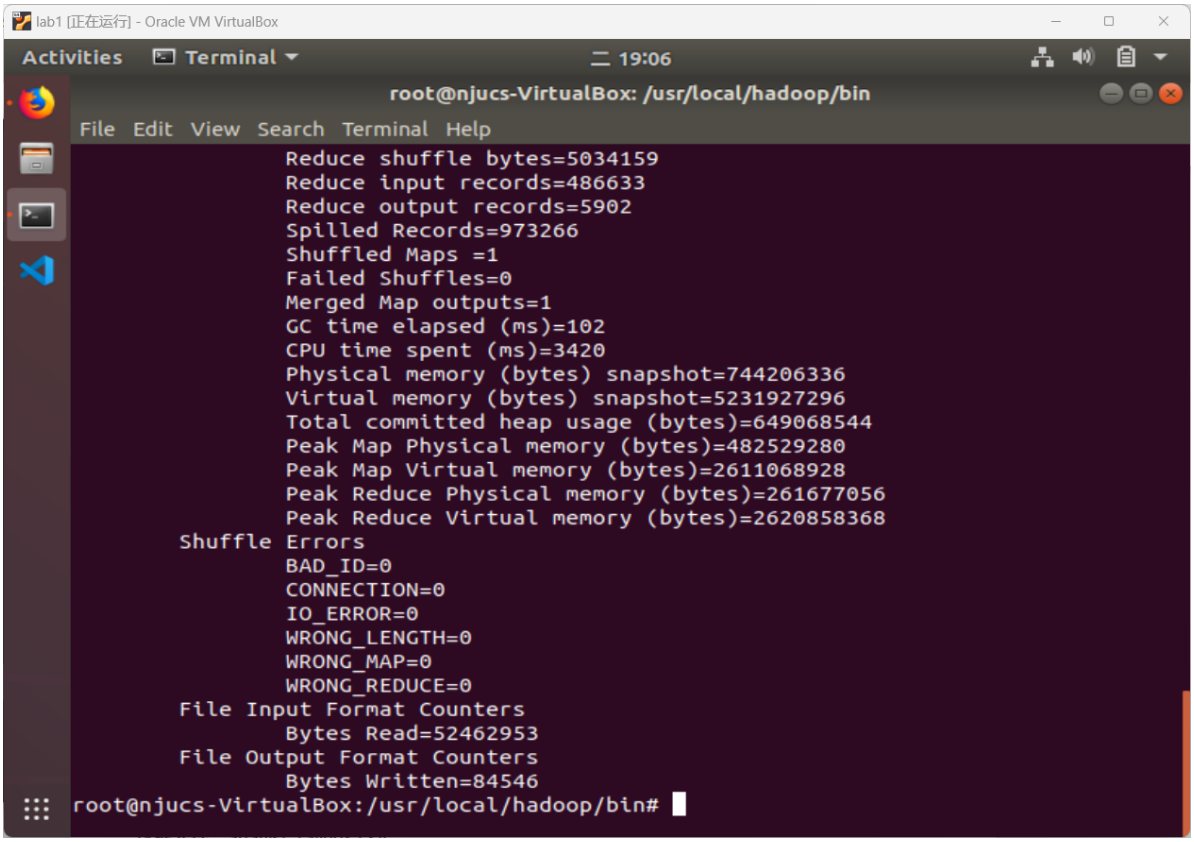
```
sudo chown $USER part-r-00000
```

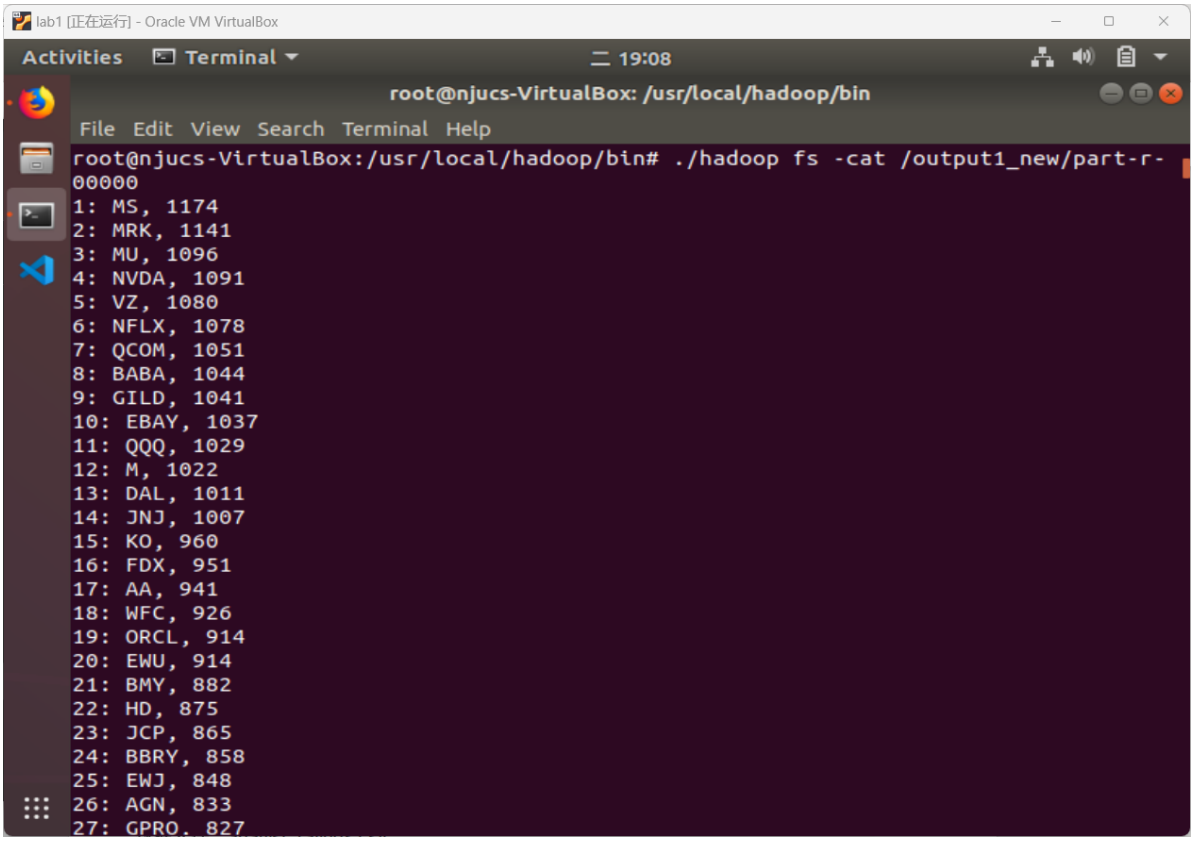
2、程序运行结果

以下即为`StockCount.java`程序执行的任务一（统计数据集上市公司股票代码（“stock”列）的出现次数，按出现次数从大到小输出，输出格式为"<排名>：<股票代码>，<次数>“）的运行结果：

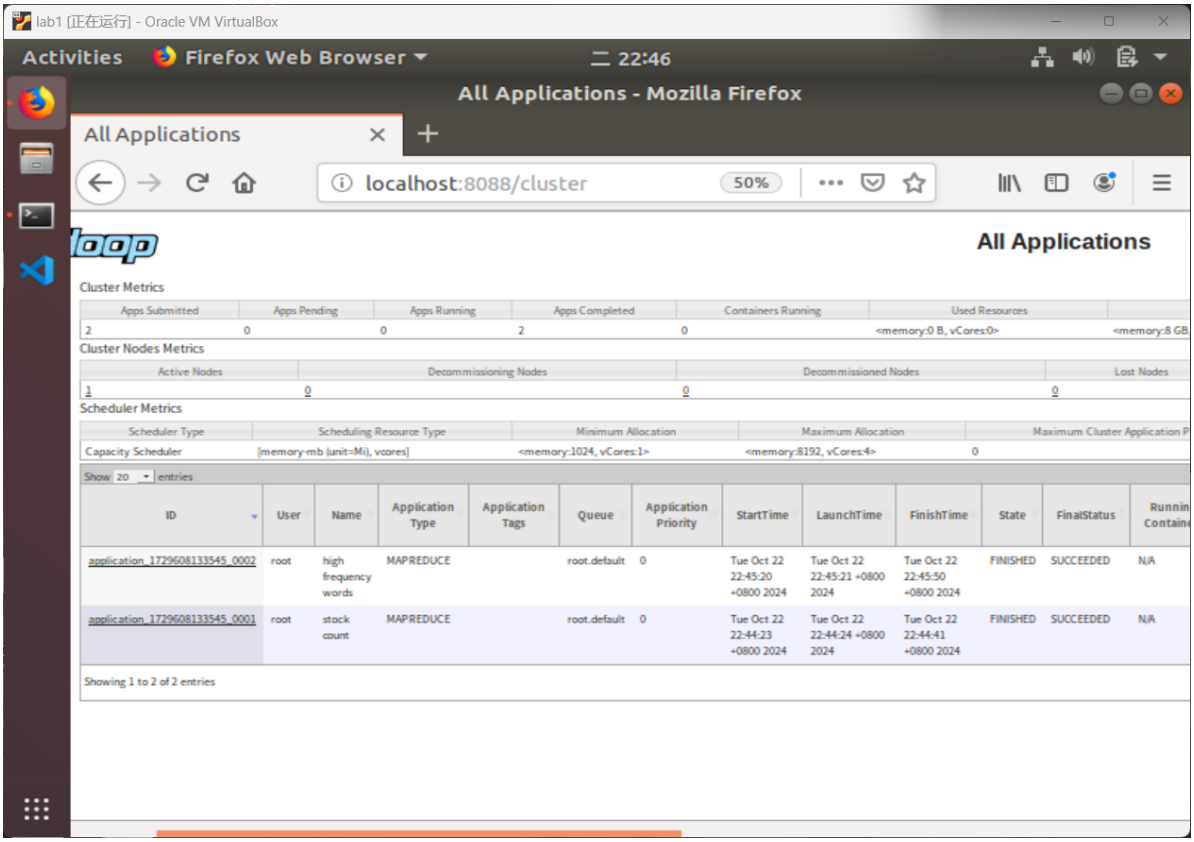

lab1 [正在运行] - Oracle VM VirtualBox
Activities
Terminal
19:06
root@njucs-VirtualBox: /usr/local/hadoop/bin
File Edit View Search Terminal Help
root@njucs-VirtualBox:/usr/local/hadoop/bin# ./hadoop jar /home/njucs/zuoye5/target/zuoye5-1.0-SNAPSHOT.jar com.example.StockCount /input /output1_new
2024-10-22 19:04:20,276 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at localhost/127.0.0.1:8032
2024-10-22 19:04:20,619 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-10-22 19:04:20,647 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1729594789612_0001
2024-10-22 19:04:21,320 INFO input.FileInputFormat: Total input files to process : 1
2024-10-22 19:04:22,201 INFO mapreduce.JobSubmitter: number of splits:1
2024-10-22 19:04:22,744 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729594789612_0001
2024-10-22 19:04:22,744 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-22 19:04:22,938 INFO conf.Configuration: resource-types.xml not found
2024-10-22 19:04:22,939 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-22 19:04:23,332 INFO impl.YarnClientImpl: Submitted application application_1729594789612_0001
2024-10-22 19:04:23,366 INFO mapreduce.Job: The url to track the job: http://njucs-VirtualBox:8088/proxy/application_1729594789612_0001/
2024-10-22 19:04:23,367 INFO mapreduce.Job: Running job: job_1729594789612_0001
2024-10-22 19:04:30,485 INFO mapreduce.Job: Job job_1729594789612_0001 running in uber mode : false
2024-10-22 19:04:30,487 INFO mapreduce.Job: map 0% reduce 0%
2024-10-22 19:04:35,562 INFO mapreduce.Job: map 100% reduce 0%
2024-10-22 19:04:40,611 INFO mapreduce.Job: map 100% reduce 100%
2024-10-22 19:04:41,626 INFO mapreduce.Job: Job job_1729594789612_0001 completed s


lab1 [正在运行] - Oracle VM VirtualBox
Activities
Terminal
19:06
root@njucs-VirtualBox: /usr/local/hadoop/bin
File Edit View Search Terminal Help
2024-10-22 19:04:40,611 INFO mapreduce.Job: map 100% reduce 100%
2024-10-22 19:04:41,626 INFO mapreduce.Job: Job job_1729594789612_0001 completed successfully
2024-10-22 19:04:41,709 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=5034159
FILE: Number of bytes written=10685903
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=52463065
HDFS: Number of bytes written=84546
HDFS: Number of read operations=8
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
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)=3141
Total time spent by all reduces in occupied slots (ms)=2455
Total time spent by all map tasks (ms)=3141
Total time spent by all reduce tasks (ms)=2455
Total vcore-milliseconds taken by all map tasks=3141
Total vcore-milliseconds taken by all reduce tasks=2455
Total megabyte-milliseconds taken by all map tasks=3216384
Total megabyte-milliseconds taken by all reduce tasks=2513920
Map-Reduce Framework


lab1 [正在运行] - Oracle VM VirtualBox
Activities
Terminal
19:06
root@njucs-VirtualBox: /usr/local/hadoop/bin
File Edit View Search Terminal Help
Reduce shuffle bytes=5034159
Reduce input records=486633
Reduce output records=5902
Spilled Records=973266
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=102
CPU time spent (ms)=3420
Physical memory (bytes) snapshot=744206336
Virtual memory (bytes) snapshot=5231927296
Total committed heap usage (bytes)=649068544
Peak Map Physical memory (bytes)=482529280
Peak Map Virtual memory (bytes)=2611068928
Peak Reduce Physical memory (bytes)=261677056
Peak Reduce Virtual memory (bytes)=2620858368
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=52462953
File Output Format Counters
Bytes Written=84546
root@njucs-VirtualBox:/usr/local/hadoop/bin#


lab1 [正在运行] - Oracle VM VirtualBox
Activities
Terminal
19:08
root@njucs-VirtualBox: /usr/local/hadoop/bin
File Edit View Search Terminal Help
root@njucs-VirtualBox:/usr/local/hadoop/bin# ./hadoop fs -cat /output1_new/part-r-00000
1: MS, 1174
2: MRK, 1141
3: MU, 1096
4: NVDA, 1091
5: VZ, 1080
6: NFLX, 1078
7: QCOM, 1051
8: BABA, 1044
9: GILD, 1041
10: EBAY, 1037
11: QQQ, 1029
12: M, 1022
13: DAL, 1011
14: JNJ, 1007
15: KO, 960
16: FDX, 951
17: AA, 941
18: WFC, 926
19: ORCL, 914
20: EWU, 914
21: BMY, 882
22: HD, 875
23: JCP, 865
24: BBRY, 858
25: EWJ, 848
26: AGN, 833
27: GPRO, 827

3、WEB页面截图


lab1 [正在运行] - Oracle VM VirtualBox
Activities
Firefox Web Browser
22:46
All Applications - Mozilla Firefox
All Applications
localhost:8088/cluster
50%
All Applications
Cluster Metrics
Apps Submitted
Apps Pending
Apps Running
Apps Completed
Containers Running
Used Resources
2
0
0
2
0
<memory:0 B, vCores:0>
<memory:8 GB
Cluster Nodes Metrics
Active Nodes
Decommissioning Nodes
Decommissioned Nodes
Lost Nodes
1
0
0
0
Scheduler Metrics
Scheduler Type
Scheduling Resource Type
Minimum Allocation
Maximum Allocation
Capacity Scheduler
[memory-mb (unit=Mi), vcores]
<memory:1024, vCores:1>
<memory:8192, vCores:4>
0
Show 20 entries
ID
User
Name
Application Type
Application Tags
Queue
Application Priority
StartTime
LaunchTime
FinishTime
State
FinalStatus
application_1729608133545_0002
root
high frequency words
MAPREDUCE
root.default
0
Tue Oct 22 22:45:20 +0800 2024
Tue Oct 22 22:45:21 +0800 2024
Tue Oct 22 22:45:50 +0800 2024
FINISHED
SUCCEEDED
N/A
application_1729608133545_0001
root
stock count
MAPREDUCE
root.default
0
Tue Oct 22 22:44:23 +0800 2024
Tue Oct 22 22:44:24 +0800 2024
Tue Oct 22 22:44:41 +0800 2024
FINISHED
SUCCEEDED
N/A
Showing 1 to 2 of 2 entries