IEMS 5730 Spring 2021 Homework #0

Q1:Hadoop Cluster Setup

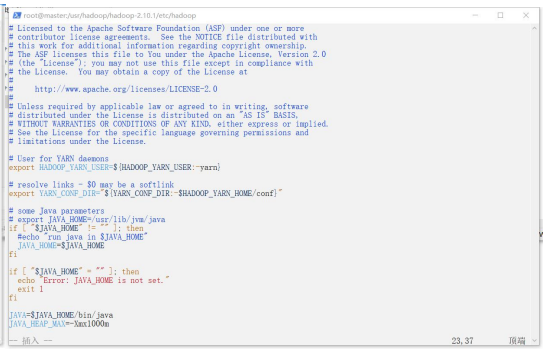
I used the VM Workstation to set up four virtual machines and build a multi-node Hadoop cluster. (1 Master and 3Slaves)

Step1: setup the Hadoop environment for the master

P1: hdfs configuration

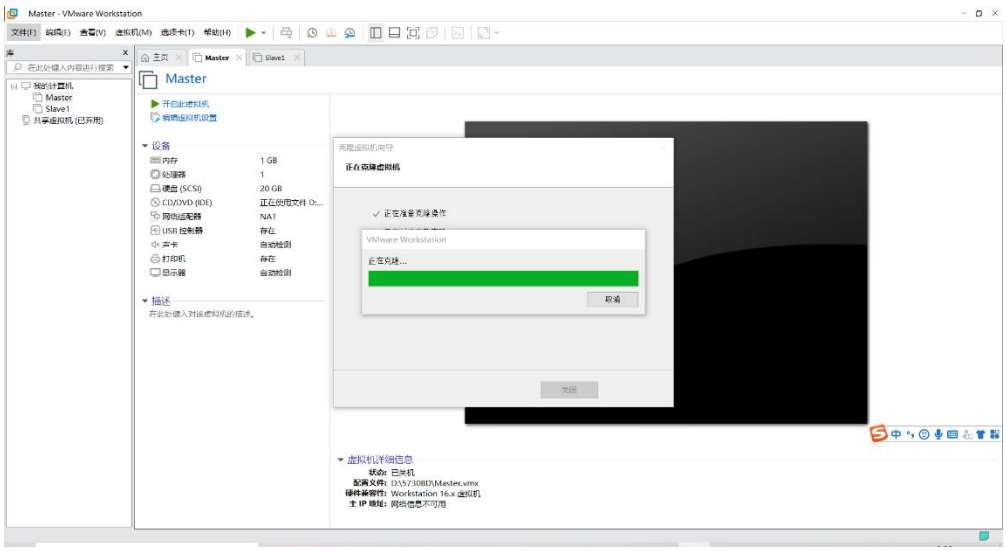
P2: mapred-site.xml configuration

P3: yarn-site.xml configuration

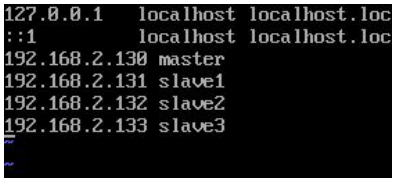


P4: Set the Hadoop path

Step2: clone the master virtual machine

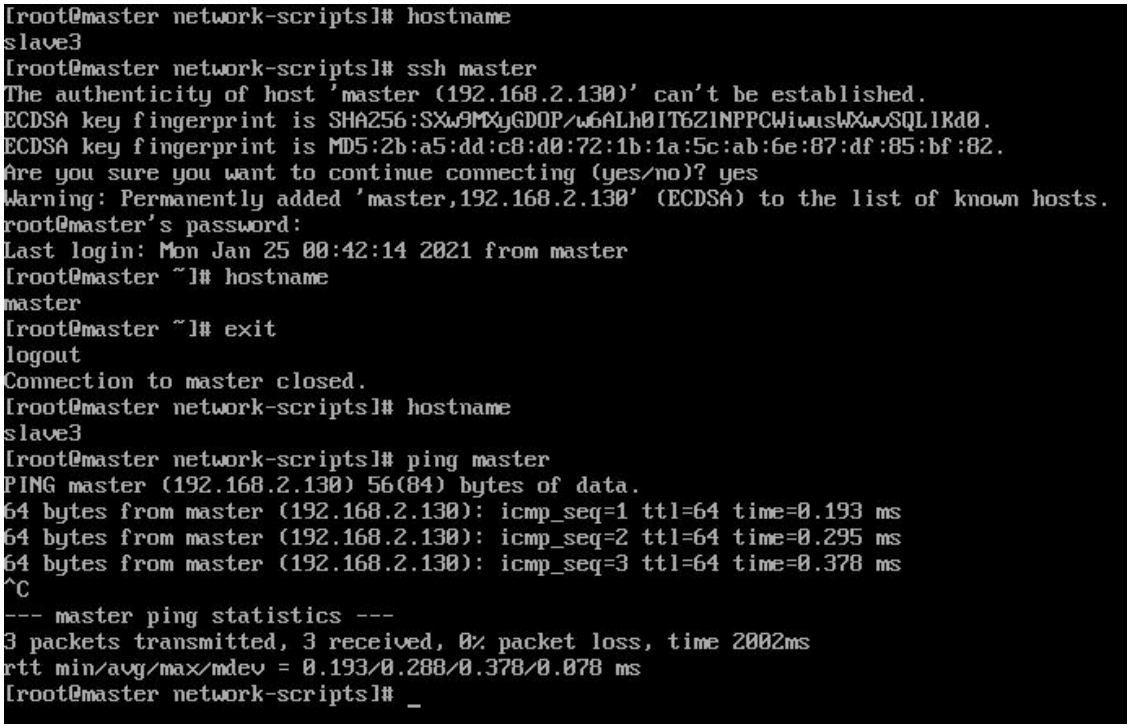


P5: Clone the master(three times)



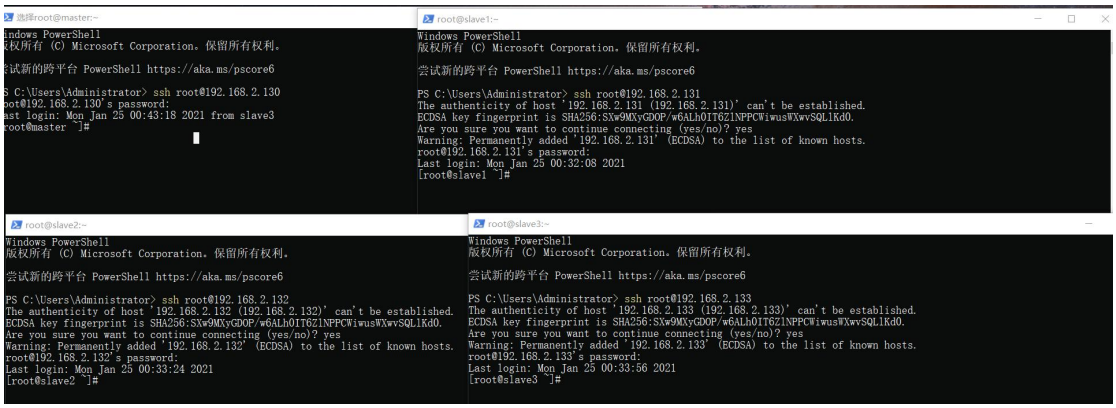
```
127.0.0.1   localhost localhost.loc
::1         localhost localhost.loc
192.168.2.130 master
192.168.2.131 slave1
192.168.2.132 slave2
192.168.2.133 slave3
```

P6: Modify the hosts

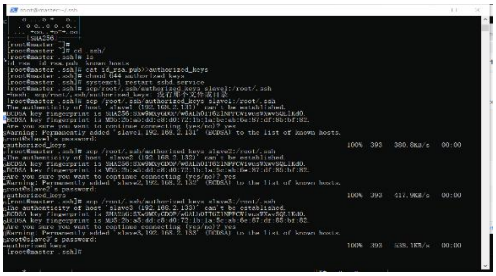


```
[root@master network-scripts]# hostname
slave3
[root@master network-scripts]# ssh master
The authenticity of host 'master (192.168.2.130)' can't be established.
ECDSA key fingerprint is SHA256:SXw9MXyGDOP/w6ALh0IT6ZINPPCWiwusWXwvSQLlKd0.
ECDSA key fingerprint is MD5:2b:a5:dd:c8:d0:72:1b:1a:5c:ab:6e:87:df:85:bf:82.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'master,192.168.2.130' (ECDSA) to the list of known hosts.
root@master's password:
Last login: Mon Jan 25 00:42:14 2021 from master
[root@master ~]# hostname
master
[root@master ~]# exit
logout
Connection to master closed.
[root@master network-scripts]# hostname
slave3
[root@master network-scripts]# ping master
PING master (192.168.2.130) 56(84) bytes of data.
64 bytes from master (192.168.2.130): icmp_seq=1 ttl=64 time=0.193 ms
64 bytes from master (192.168.2.130): icmp_seq=2 ttl=64 time=0.295 ms
64 bytes from master (192.168.2.130): icmp_seq=3 ttl=64 time=0.378 ms
^C
--- master ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 0.193/0.288/0.378/0.078 ms
[root@master network-scripts]# _
```

P7: Slave tries to connect to the host

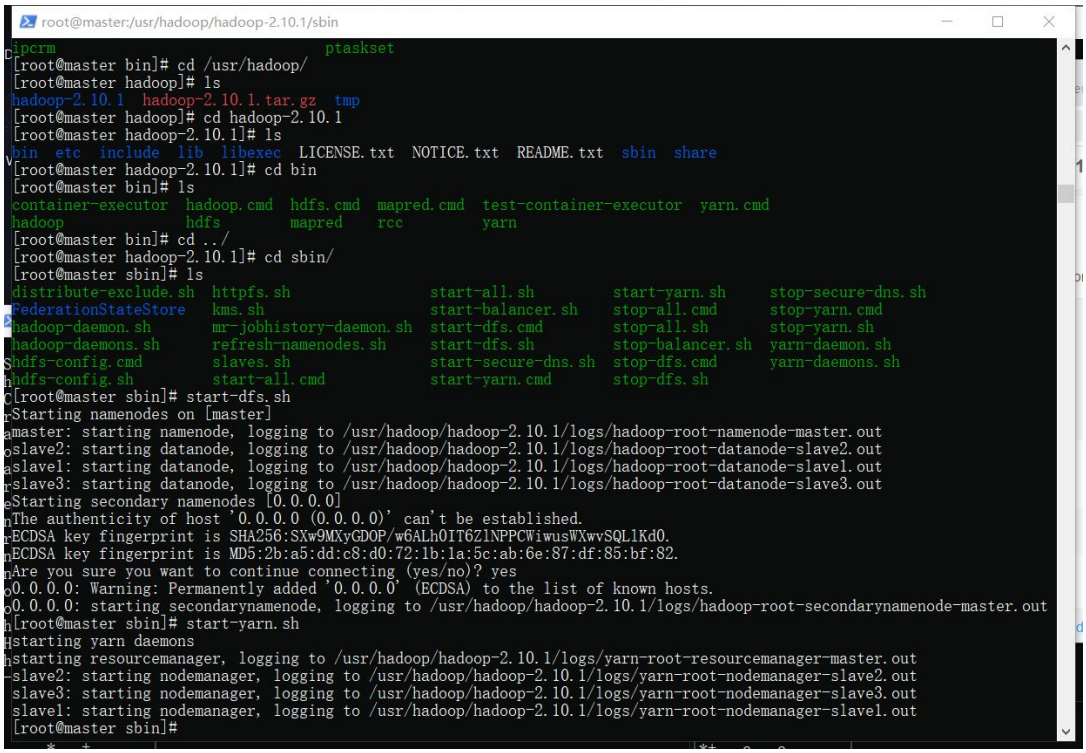


P8: All four hosts are connected



P9: Generate a public key and send it to each slave for easy connection

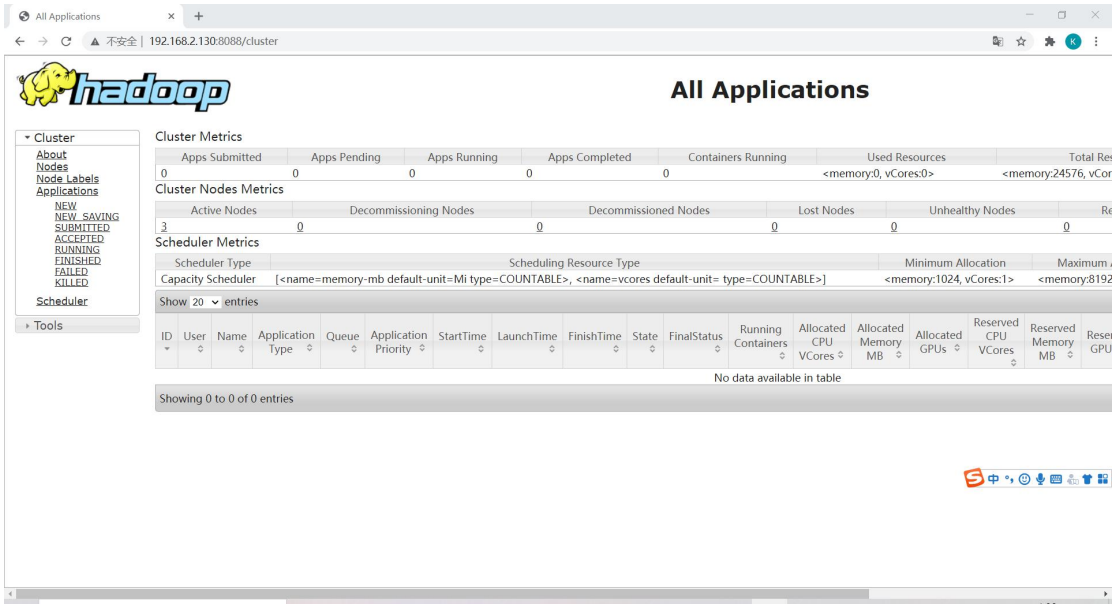
Step3: Hadoop Test



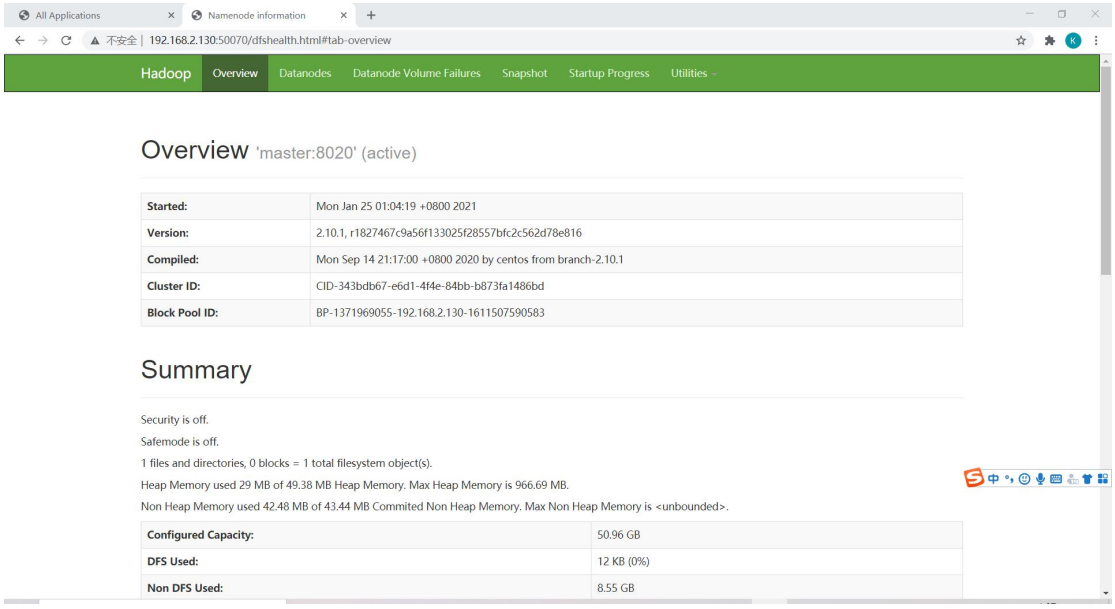
P10: Start the Hadoop (using start-dfs.sh and start-yarn.sh)

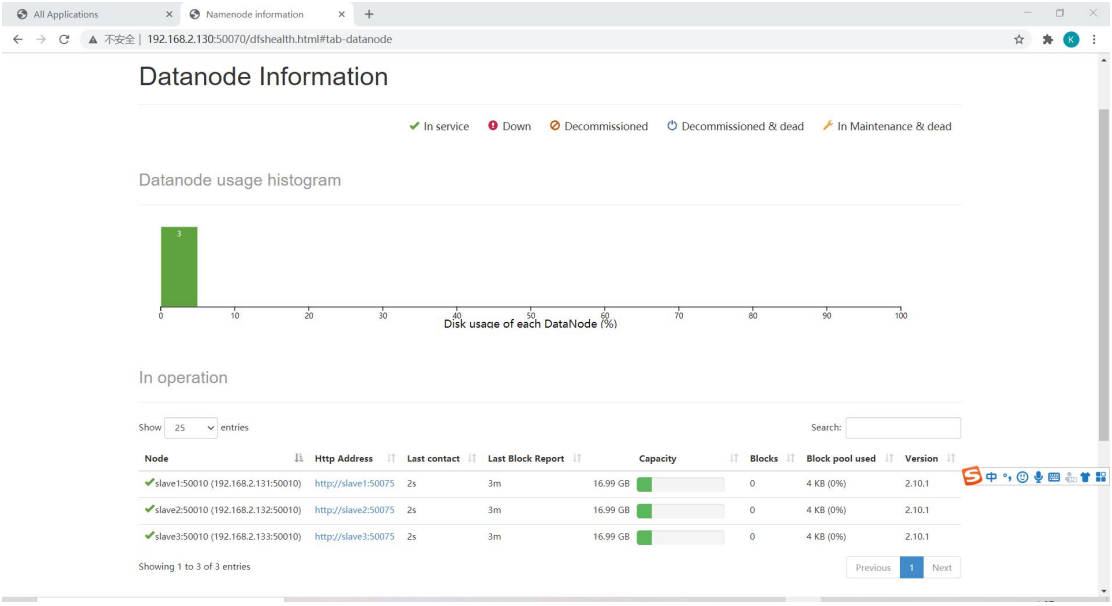
```
[root@master sbin]# jps
1634 SecondaryNameNode
1444 NameNode
1781 ResourceManager
2047 Jps
```

P11: Checking the Hadoop process (using jps)



P12: Yarn interface display (port 8088)





P12 & P13: Namenode information (port 50070)

Step4: Terasort Test

```
root@master:/usr/hadoop/hadoop-2.10.1/sbin
[root@master sbin]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.jar ter
agen  10737418  /terasort/1G-input
21/01/25 01:18:45 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.2.130:8032
21/01/25 01:18:46 INFO terasort.TeraGen: Generating 10737418 using 2
21/01/25 01:18:46 INFO mapreduce.JobSubmitter: number of splits:2
21/01/25 01:18:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1611507897104_0001
21/01/25 01:18:47 INFO conf.Configuration: resource-types.xml not found
21/01/25 01:18:47 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/01/25 01:18:47 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/01/25 01:18:47 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/01/25 01:18:47 INFO impl.YarnClientImpl: Submitted application application_1611507897104_0001
21/01/25 01:18:47 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1611507897104_0001/
21/01/25 01:18:47 INFO mapreduce.Job: Running job: job_1611507897104_0001
21/01/25 01:18:57 INFO mapreduce.Job: Job job_1611507897104_0001 running in uber mode : false
21/01/25 01:18:57 INFO mapreduce.Job:  map 0% reduce 0%
21/01/25 01:19:17 INFO mapreduce.Job:  map 19% reduce 0%
21/01/25 01:19:18 INFO mapreduce.Job:  map 38% reduce 0%
21/01/25 01:19:23 INFO mapreduce.Job:  map 47% reduce 0%
21/01/25 01:19:24 INFO mapreduce.Job:  map 56% reduce 0%
21/01/25 01:19:30 INFO mapreduce.Job:  map 65% reduce 0%
21/01/25 01:19:31 INFO mapreduce.Job:  map 76% reduce 0%
21/01/25 01:19:36 INFO mapreduce.Job:  map 86% reduce 0%
21/01/25 01:19:37 INFO mapreduce.Job:  map 97% reduce 0%
21/01/25 01:19:38 INFO mapreduce.Job:  map 100% reduce 0%
21/01/25 01:19:39 INFO mapreduce.Job: Job job_1611507897104_0001 completed successfully
21/01/25 01:19:39 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=416530
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=167
                HDFS: Number of bytes written=1073741800
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
```

P14: map/reduce process of teragen

```
[root@master ~]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.jar terage
n  10737418  /terasort/2G-input
```

P15: Generate 2G input

```
[root@master ~]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.jar teraso
rt /terasort/2G-input/terasort/2G-output
```

P16: Terasort process

When I am dealing with the terasort process, an error happened(shown in P17) and I can't find a solution for a long time. After checking, I found that I forgot to type a space before “hadoop” so the process can’t find the correct path.

```
[root@master hadoop-2.10.1]# bin/hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-
2.10.1.jar terasort  terasort/2G-input terasort/2G-output
21/01/25 03:10:45 INFO terasort.TeraSort: starting
21/01/25 03:10:46 ERROR terasort.TeraSort: Input path does not exist: hdfs://master:8020/user/root/terasort/2G-input
```

P17: Error

```
[root@master mapreduce]# hadoop jar /usr/hadoop/hadoop-2.10.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.1.ja
r terasort  /terasort/2G-input /terasort/2G-output

21/01/25 03:36:27 INFO terasort.TeraSort: starting
21/01/25 03:36:28 INFO input.FileInputFormat: Total input files to process : 2
Spent 162ms computing base-splits.
Spent 2ms computing TeraScheduler splits.
Computing input splits took 164ms
Sampling 8 splits of 8
Making 1 from 100000 sampled records
Computing paritition took 865ms
Spent 1031ms computing partitions.
21/01/25 03:36:29 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.2.130:8032
21/01/25 03:36:30 INFO mapreduce.JobSubmitter: number of splits:8
21/01/25 03:36:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1611507897104_0003
21/01/25 03:36:31 INFO conf.Configuration: resource-types.xml not found
21/01/25 03:36:31 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/01/25 03:36:31 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
21/01/25 03:36:31 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
21/01/25 03:36:31 INFO impl.YarnClientImpl: Submitted application application_1611507897104_0003
21/01/25 03:36:31 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1611507897104_0003/
21/01/25 03:36:31 INFO mapreduce.Job: Running job: job_1611507897104_0003
21/01/25 03:36:41 INFO mapreduce.Job: Job job_1611507897104_0003 running in uber mode : false
21/01/25 03:36:41 INFO mapreduce.Job:  map 0% reduce 0%
```

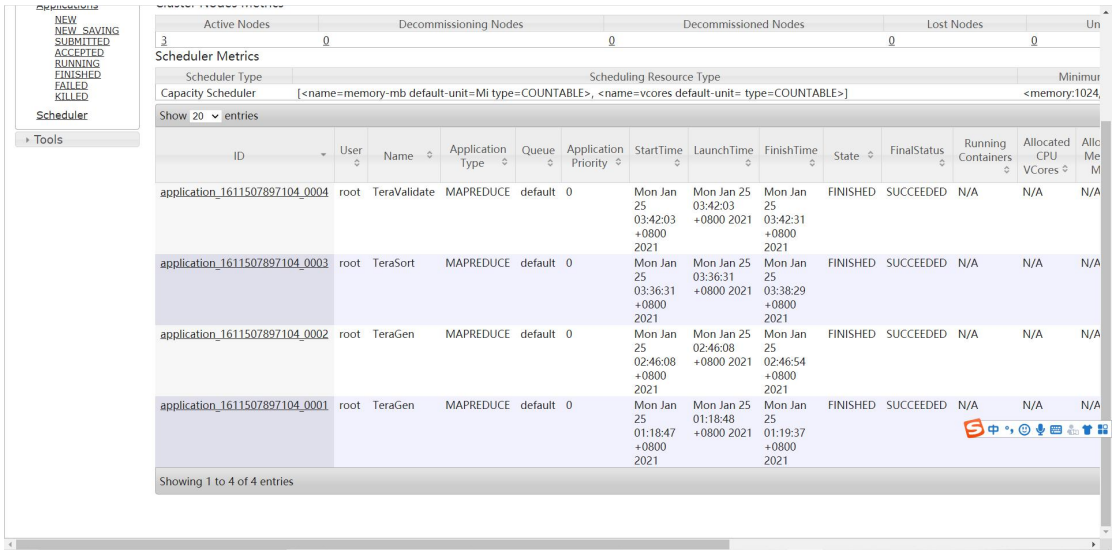
P18: Problem solved

```
    GC time elapsed (ms)=14024
    CPU time spent (ms)=78590
```

P19: Terasort running time

```
[root@master mapreduce]# hdfs dfs -ls /terasort/
Found 4 items
drwxr-xr-x   - root supergroup          0 2021-01-25 01:19 /terasort/1G-input
drwxr-xr-x   - root supergroup          0 2021-01-25 02:46 /terasort/2G-input
drwxr-xr-x   - root supergroup          0 2021-01-25 03:38 /terasort/2G-output
drwxr-xr-x   - root supergroup          0 2021-01-25 03:42 /terasort/2G-validate
[root@master mapreduce]#
```

P20: Teravalidate succeeded



P21: Three test process in yarn

Step5: Running python on Hadoop

In this part, I first create a folder named “WordCount” in hdfs, and then use the command “Hadoop fs – put” to put the two Shakespeare-basket into the folder, making it easier for me to count the word.

```
[root@master hadoop]# hdfs fs -put /usr/hadoop/shakespeare-basket1 /WordCount
错误: 找不到或无法加载主类 fs
[root@master hadoop]# hadoop fs -put /usr/hadoop/shakespeare-basket1 /WordCount
[root@master hadoop]# hadoop fs -ls /WordCount
Found 1 items
-rw-r--r--   3 root supergroup  124620254 2021-01-25 04:26 /WordCount/shakespeare-basket1
[root@master hadoop]#
```

P22: Put Shakespeare-basket1 and Shakespeare-basket2 into WordCount

```
[root@master hduser]# echo "foo foo quux labs foo bar quux" | /home/hduser/mapper.py
/home/hduser/mapper.py:行4: import: 未找到命令
/home/hduser/mapper.py:行9: 未预期的符号 `line' 附近有语法错误
/home/hduser/mapper.py:行9:     line = line.strip()'
[root@master hduser]# python --version
Python 2.7.5
[root@master hduser]# vim mapper.py
[root@master hduser]# echo "foo foo quux labs foo bar quux" | /home/hduser/mapper.py
foo     1
foo     1
quux    1
labs    1
foo     1
bar     1
quux    1
```

P23: Test mapper.py and reducer.py

```
root@master:/home/hduser
21/01/25 07:46:24 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/01/25 07:46:24 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units
 = Mi, type = COUNTABLE
21/01/25 07:46:24 INFO resource.ResourceUtils: Adding resource type - name = vcores, units =
, type = COUNTABLE
21/01/25 07:46:24 INFO impl.YarnClientImpl: Submitted application application_1611507897104_0
010
21/01/25 07:46:24 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/appl
ication_1611507897104_0010/
21/01/25 07:46:24 INFO mapreduce.Job: Running job: job_1611507897104_0010
21/01/25 07:46:34 INFO mapreduce.Job: Job job_1611507897104_0010 running in uber mode : false
21/01/25 07:46:34 INFO mapreduce.Job:  map 0% reduce 0%
21/01/25 07:46:59 INFO mapreduce.Job:  map 28% reduce 0%
21/01/25 07:47:06 INFO mapreduce.Job:  map 33% reduce 0%
21/01/25 07:47:12 INFO mapreduce.Job:  map 44% reduce 0%
21/01/25 07:47:18 INFO mapreduce.Job:  map 46% reduce 0%
21/01/25 07:47:19 INFO mapreduce.Job:  map 50% reduce 0%
21/01/25 07:47:24 INFO mapreduce.Job:  map 52% reduce 0%
21/01/25 07:47:25 INFO mapreduce.Job:  map 53% reduce 0%
21/01/25 07:47:43 INFO mapreduce.Job:  map 54% reduce 0%
21/01/25 07:47:47 INFO mapreduce.Job:  map 54% reduce 11%
21/01/25 07:47:49 INFO mapreduce.Job:  map 59% reduce 11%
21/01/25 07:47:50 INFO mapreduce.Job:  map 60% reduce 11%
21/01/25 07:47:55 INFO mapreduce.Job:  map 62% reduce 11%
21/01/25 07:47:56 INFO mapreduce.Job:  map 66% reduce 11%
21/01/25 07:48:02 INFO mapreduce.Job:  map 70% reduce 11%
21/01/25 07:48:08 INFO mapreduce.Job:  map 75% reduce 11%
21/01/25 07:48:14 INFO mapreduce.Job:  map 85% reduce 11%
21/01/25 07:48:17 INFO mapreduce.Job:  map 85% reduce 22%
21/01/25 07:48:26 INFO mapreduce.Job:  map 89% reduce 22%
21/01/25 07:48:40 INFO mapreduce.Job:  map 92% reduce 22%
21/01/25 07:48:46 INFO mapreduce.Job:  map 95% reduce 22%
21/01/25 07:48:52 INFO mapreduce.Job:  map 100% reduce 22%
21/01/25 07:48:59 INFO mapreduce.Job:  map 100% reduce 68%
21/01/25 07:49:05 INFO mapreduce.Job:  map 100% reduce 72%
21/01/25 07:49:11 INFO mapreduce.Job:  map 100% reduce 76%
21/01/25 07:49:17 INFO mapreduce.Job:  map 100% reduce 80%
21/01/25 07:49:23 INFO mapreduce.Job:  map 100% reduce 84%
21/01/25 07:49:30 INFO mapreduce.Job:  map 100% reduce 87%
21/01/25 07:49:36 INFO mapreduce.Job:  map 100% reduce 92%
21/01/25 07:49:42 INFO mapreduce.Job:  map 100% reduce 95%
21/01/25 07:49:48 INFO mapreduce.Job:  map 100% reduce 99%
21/01/25 07:49:49 INFO mapreduce.Job:  map 100% reduce 100%
21/01/25 07:49:49 INFO mapreduce.Job: Job job_1611507897104_0010 completed successfully
21/01/25 07:49:49 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=715074630
                FILE: Number of bytes written=1083568264
                FILE: Number of read operations=0
```

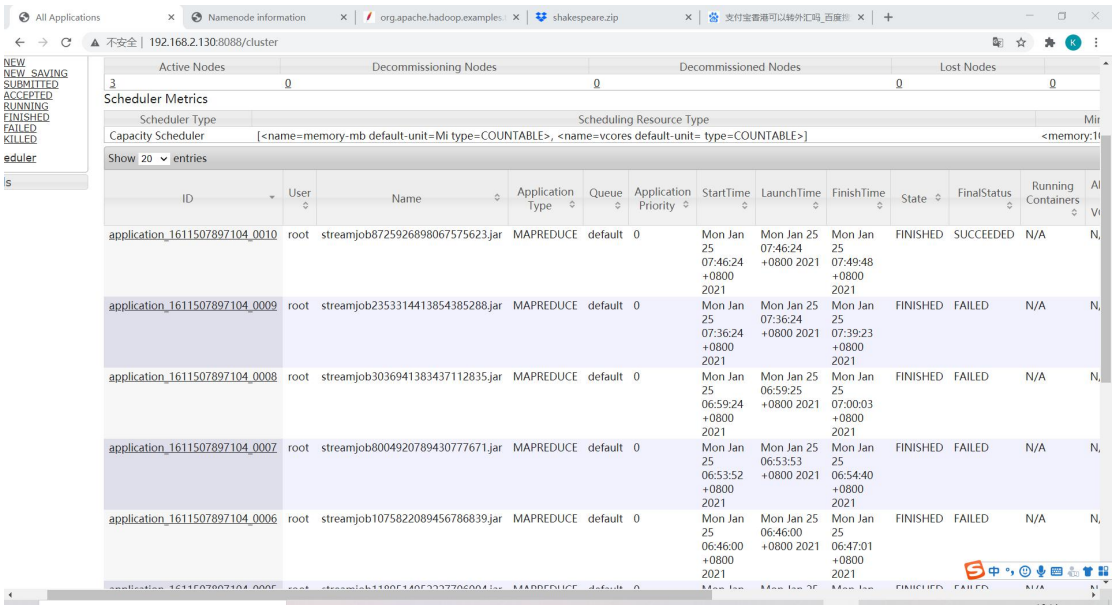
P24: Map/reduce process

```
Map-Reduce Framework
        Map input records=4340061
        Map output records=37015545
        Map output bytes=293614077
        Map output materialized bytes=367645185
        Input split bytes=300
        Combine input records=0
        Combine output records=0
        Reduce input groups=93358
        Reduce shuffle bytes=367645185
        Reduce input records=37015545
        Reduce output records=93358
        Spilled Records=109134525
        Shuffled Maps =3
        Failed Shuffles=0
        Merged Map outputs=3
        GC time elapsed (ms)=4744
        CPU time spent (ms)=131100
        Physical memory (bytes) snapshot=666931200
        Virtual memory (bytes) snapshot=8372662272
        Total committed heap usage (bytes)=447262720
```

P25: Running time of the python wordcount

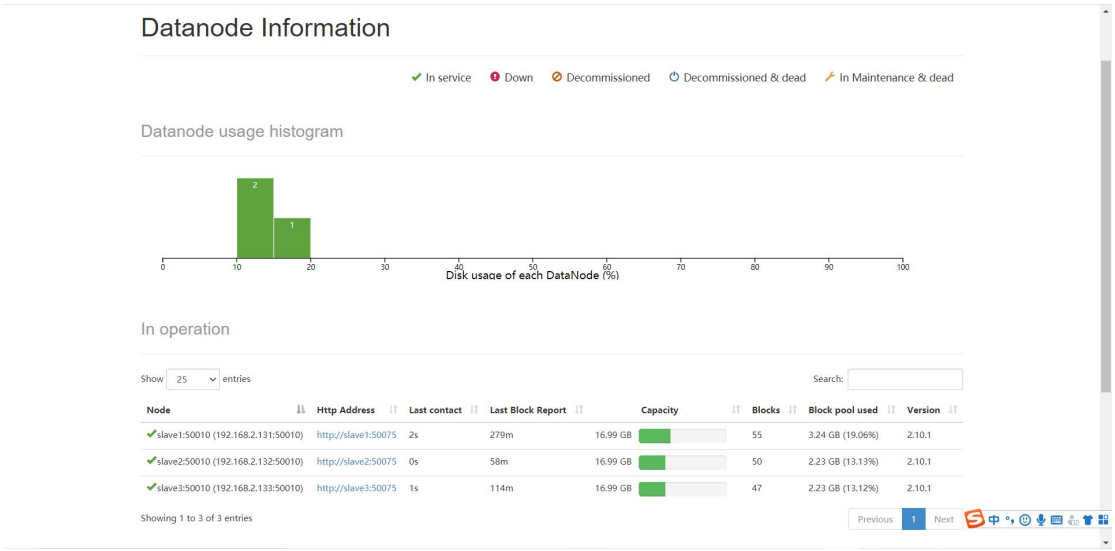
```
warranteth      42
warranties      70
warrantise      6
warrantize      80
warrantor       1
warrants        248
warranty        331
warre   3
warred  16
warren  111
warren's        82
warrener        40
warrenton       4
warring 46
warrington      24
warrior 792
warrior''       4
warrior's       26
warriors        1046
warriorship     1
wars    6274
wars'   43
warsaw  20
```

P26: Part of the count results



P27: Steamjob shown on yarn (status:succeeded)

We can see from P27, the time spent on steamjob is 204 seconds. We will later compare this to the time spent on the java wordcount program.



P28: The status of datanodes (after wordcount)

Step6: Running java on Hadoop

(bonus)

```
root@master:/usr/hadoop/hadoop-2.10.1
package org.myorg;

    import java.io.IOException;
    import java.util.*;

    import org.apache.hadoop.fs.Path;
    import org.apache.hadoop.conf.*;
    import org.apache.hadoop.io.*;
    import org.apache.hadoop.mapred.*;
    import org.apache.hadoop.util.*;

    public class WordCount {

      public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) thro
ws IOException {
          String line = value.toString();
          StringTokenizer tokenizer = new StringTokenizer(line);
          while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            output.collect(word, one);
          }
        }
      }

      public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
        public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter r
eporter) throws IOException {
          int sum = 0;
          while (values.hasNext()) {
            sum += values.next().get();
          }
          output.collect(key, new IntWritable(sum));
-- 插入 --                                                                    1,1          顶端
```

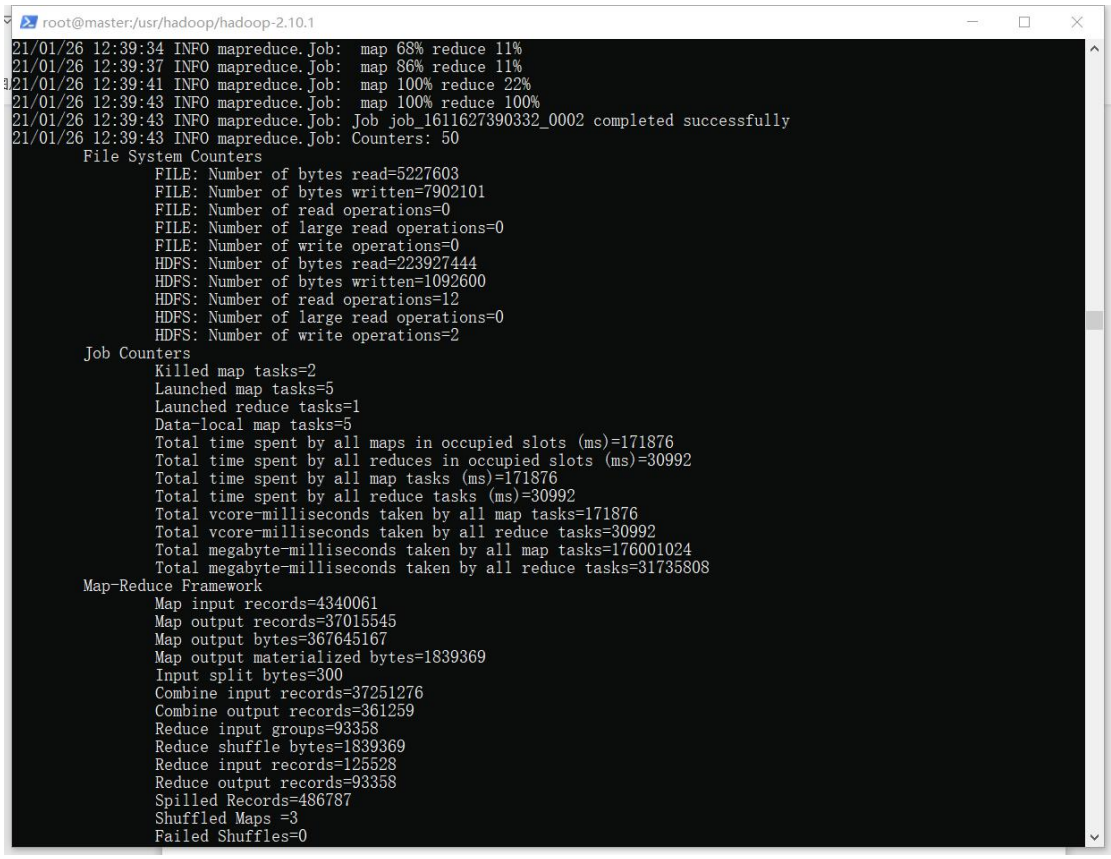
P29: Write java wordcount program

```
[root@master hadoop-2.10.1]# javac -classpath /usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-common-2.10.1.jar:/us
r/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-hdfs-2.10.1.jar:/usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-c
lient-2.10.1.jar:/usr/hadoop/hadoop-2.10.1/share/hadoop/common/hadoop-mapreduce-client-core-2.10.1.jar -d wordcount_clas
ses/ WordCount.java
```

P30: Compile the java program we write

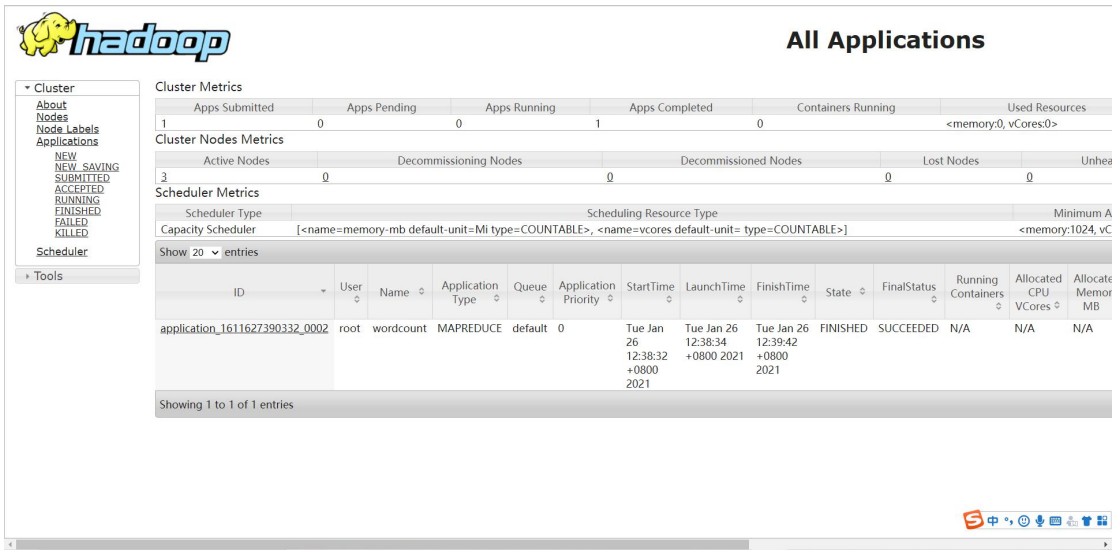
```
[root@master hadoop-2.10.1]# jar -cvf WordCount.jar -C wordcount_classes ./
已添加清单
正在添加: org/(输入 = 0) (输出 = 0)(存储了 0%)
正在添加: org/myorg/(输入 = 0) (输出 = 0)(存储了 0%)
正在添加: org/myorg/WordCount$Map.class(输入 = 1938) (输出 = 799)(压缩了 58%)
正在添加: org/myorg/WordCount$Reduce.class(输入 = 1611) (输出 = 649)(压缩了 59%)
正在添加: org/myorg/WordCount.class(输入 = 1534) (输出 = 752)(压缩了 50%)
正在添加: org/myorg/WordCount.jar(输入 = 338) (输出 = 183)(压缩了 45%)
[root@master hadoop-2.10.1]# ls
```

P31: Pack the jar package for wordcount



```
root@master:/usr/hadoop/hadoop-2.10.1
21/01/26 12:39:34 INFO mapreduce.Job:  map 68% reduce 11%
21/01/26 12:39:37 INFO mapreduce.Job:  map 86% reduce 11%
21/01/26 12:39:41 INFO mapreduce.Job:  map 100% reduce 22%
21/01/26 12:39:43 INFO mapreduce.Job:  map 100% reduce 100%
21/01/26 12:39:43 INFO mapreduce.Job: Job job_1611627390332_0002 completed successfully
21/01/26 12:39:43 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=5227603
                FILE: Number of bytes written=7902101
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=223927444
                HDFS: Number of bytes written=1092600
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=2
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=171876
                Total time spent by all reduces in occupied slots (ms)=30992
                Total time spent by all map tasks (ms)=171876
                Total time spent by all reduce tasks (ms)=30992
                Total vcore-milliseconds taken by all map tasks=171876
                Total vcore-milliseconds taken by all reduce tasks=30992
                Total megabyte-milliseconds taken by all map tasks=176001024
                Total megabyte-milliseconds taken by all reduce tasks=31735808
        Map-Reduce Framework
                Map input records=4340061
                Map output records=37015545
                Map output bytes=367645167
                Map output materialized bytes=1839369
                Input split bytes=300
                Combine input records=37251276
                Combine output records=361259
                Reduce input groups=93358
                Reduce shuffle bytes=1839369
                Reduce input records=125528
                Reduce output records=93358
                Spilled Records=486787
                Shuffled Maps =3
                Failed Shuffles=0
```

P32: Map/reduce process and running details



P33: Process shown on yarn

We can see the time spent on this wordcount program is 70 seconds (from start time to finish time), while the time spent on the python program is 204 seconds. So in my test, using java can take less time. But I don't know the principle, is it because the Hadoop framework is written in Java?