Q0 [10 marks]: Secure Virtual Machines Setup on the Cloud

Initial setting:

	Name	Type	Targets	Filters	Protocols / ports	Action	
☐	allow-ssh	Ingress	Apply to all	IP ranges: 137.189.0.0/16	tcp:22	Allow	^
☐	default-allow-internal	Ingress	Apply to all	IP ranges: 10.128.0.0/9	tcp:0-65535 udp:0-65535 icmp	Allow	^

Allowing SSH connection from CUHK IP address and keep the default internal connection as stated in the question.

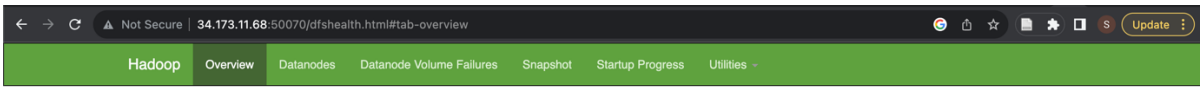
I later added some other connection to allow me to connect ssh from google cloud platform page and I also added some other ports, like 50070, to allow me to access the Hadoop WebUI on CUHK network.

Q1 [90 marks + 20 bonus marks]: Hadoop Cluster Setup

a. [20 marks] Single-node Hadoop Setup

i)

Screenshot of successfully setup single node cluster on http://34.173.11.68:50070/



Overview 'localhost:9000' (active)

Started:	Thu Sep 14 16:52:19 +0800 2023
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 20:42:00 +0800 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-62bf6f85-5460-41f5-8f82-e11024e6bc42
Block Pool ID:	BP-1364017626-10.128.0.2-1694681531799

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 65.58 MB of 293 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 47.66 MB of 48.65 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	96.73 GB
DFS Used:	24 KB (0%)
Non DFS Used:	3.73 GB
DFS Remaining:	92.99 GB (96.13%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 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	0
Block Deletion Start Time	Thu Sep 14 16:52:19 +0800 2023
Last Checkpoint Time	Thu Sep 14 16:52:11 +0800 2023

NameNode Journal Status

Current transaction ID: 1

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-cscmarkchan/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-cscmarkchan/dfs/name/current/edits_inprogress_0000000000000000001)

NameNode Storage

NameNode Storage

Storage Directory	Type	State
/tmp/hadoop-cscmarkchan/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

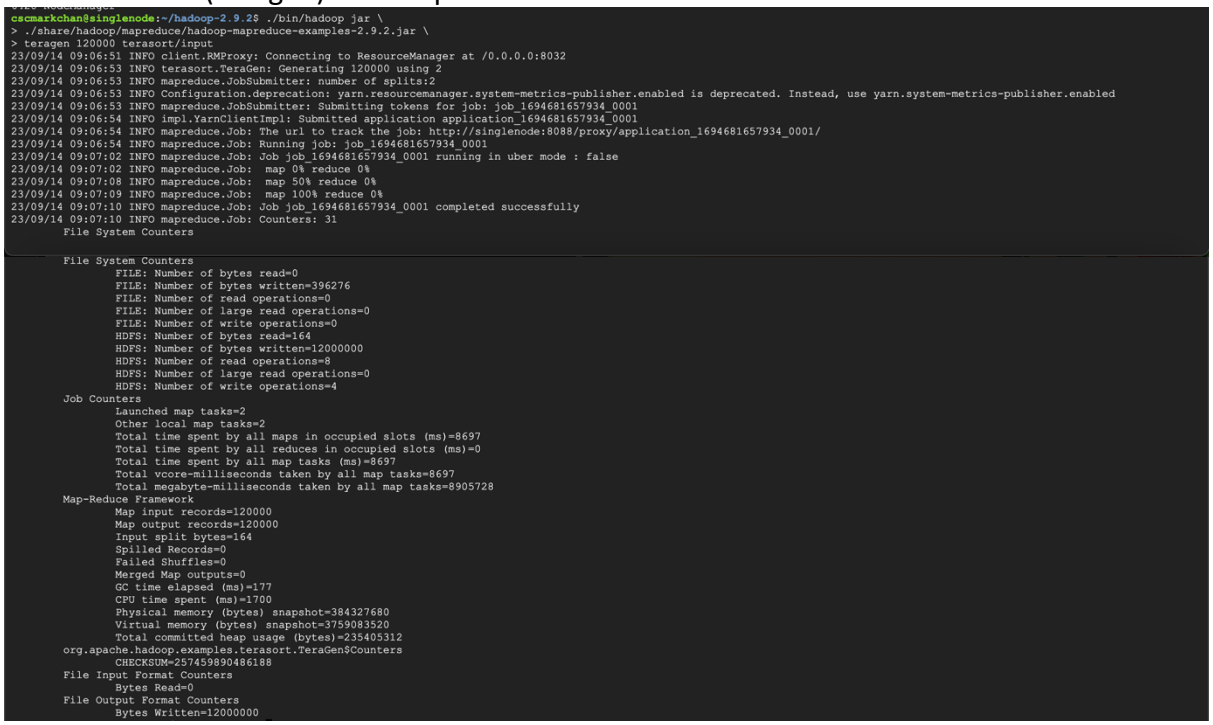
Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	96.73 GB	24 KB (0%)	92.99 GB (96.13%)	24 KB	1

Hadoop, 2018.

By following the guideline on Hadoop website, I successfully setup single node cluster on VM External IP. In this case my external IP is 34.173.11.68.

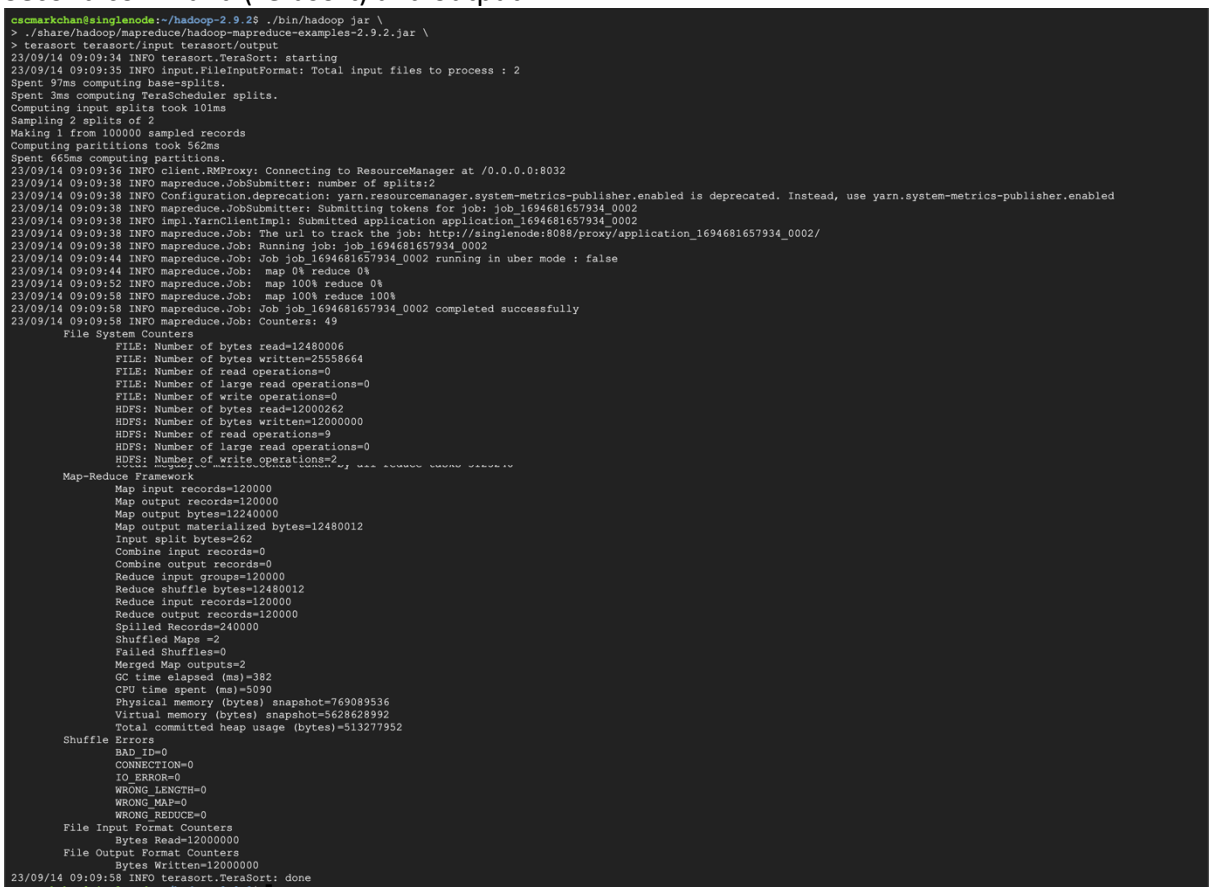
ii) Run the Terasort example

First command (Teragen) and output:



```
cscmarkchan@singlenode:~/hadoop-2.9.2$ ./bin/hadoop jar \
> ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar \
> teragen 120000 terasort/input
23/09/14 09:06:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/09/14 09:06:53 INFO terasort.TeraGen: Generating 120000 using 2
23/09/14 09:06:53 INFO mapreduce.JobSubmitter: number of splits:2
23/09/14 09:06:53 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/14 09:06:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694681657934_0001
23/09/14 09:06:54 INFO impl.YarnClientImpl: Submitted application application_1694681657934_0001
23/09/14 09:06:54 INFO mapreduce.Job: The url to track the job: http://singlenode:8088/proxy/application_1694681657934_0001/
23/09/14 09:06:54 INFO mapreduce.Job: Running job: job_1694681657934_0001
23/09/14 09:07:02 INFO mapreduce.Job: Job job_1694681657934_0001 running in uber mode : false
23/09/14 09:07:02 INFO mapreduce.Job:  map 0% reduce 0%
23/09/14 09:07:08 INFO mapreduce.Job:  map 50% reduce 0%
23/09/14 09:07:09 INFO mapreduce.Job:  map 100% reduce 0%
23/09/14 09:07:10 INFO mapreduce.Job: Job job_1694681657934_0001 completed successfully
23/09/14 09:07:10 INFO mapreduce.Job: Counters: 31
        File System Counters
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396276
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=164
                HDFS: Number of bytes written=12000000
                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
                Total time spent by all maps in occupied slots (ms)=8697
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=8697
                Total vcore-milliseconds taken by all map tasks=8697
                Total megabyte-milliseconds taken by all map tasks=8905728
        Map-Reduce Framework
                Map input records=120000
                Map output records=120000
                Input split bytes=164
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=177
                CPU time spent (ms)=1700
                Physical memory (bytes) snapshot=384327680
                Virtual memory (bytes) snapshot=3759083520
                Total committed heap usage (bytes)=235405312
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=257459890486188
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=12000000
```

Second command (Terasort) and output



```
cscmarkchan@singlenode:~/hadoop-2.9.2$ ./bin/hadoop jar \
> ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar \
> terasort terasort/input terasort/output
23/09/14 09:09:34 INFO terasort.TeraSort: starting
23/09/14 09:09:35 INFO input.FileInputFormat: Total input files to process : 2
Spent 97ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
Computing input splits took 101ms
Sampling 2 splits of 2
Making 1 from 100000 sampled records
Computing parititions took 562ms
Spent 665ms computing partitions.
23/09/14 09:09:36 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/09/14 09:09:38 INFO mapreduce.JobSubmitter: number of splits:2
23/09/14 09:09:38 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/14 09:09:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694681657934_0002
23/09/14 09:09:38 INFO impl.YarnClientImpl: Submitted application application_1694681657934_0002
23/09/14 09:09:38 INFO mapreduce.Job: The url to track the job: http://singlenode:8088/proxy/application_1694681657934_0002/
23/09/14 09:09:38 INFO mapreduce.Job: Running job: job_1694681657934_0002
23/09/14 09:09:44 INFO mapreduce.Job: Job job_1694681657934_0002 running in uber mode : false
23/09/14 09:09:44 INFO mapreduce.Job:  map 0% reduce 0%
23/09/14 09:09:52 INFO mapreduce.Job:  map 100% reduce 0%
23/09/14 09:09:58 INFO mapreduce.Job:  map 100% reduce 100%
23/09/14 09:09:58 INFO mapreduce.Job: Job job_1694681657934_0002 completed successfully
23/09/14 09:09:58 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=12480006
                FILE: Number of bytes written=25558664
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=12000262
                HDFS: Number of bytes written=12000000
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Map-Reduce Framework
                Map input records=120000
                Map output records=120000
                Map output bytes=12240000
                Map output materialized bytes=12480012
                Input split bytes=262
                Combine input records=0
                Combine output records=0
                Reduce input groups=120000
                Reduce shuffle bytes=12480012
                Reduce input records=120000
                Reduce output records=120000
                Spilled Records=240000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=382
                CPU time spent (ms)=5090
                Physical memory (bytes) snapshot=769089536
                Virtual memory (bytes) snapshot=5628628992
                Total committed heap usage (bytes)=513277952
        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=12000000
        File Output Format Counters
                Bytes Written=12000000
23/09/14 09:09:58 INFO terasort.TeraSort: done
```

Third command (Teravalidate) and output

```
cscmarkchan@singlenode:~/hadoop-2.9.2$ ./bin/hadoop jar \
> ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar \
> teravalidate terasort/output terasort/check
23/09/14 09:12:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/09/14 09:12:31 INFO input.FileInputFormat: Total input files to process : 1
Spent 13ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
23/09/14 09:12:31 INFO mapreduce.JobSubmitter: number of splits:1
23/09/14 09:12:31 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/14 09:12:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694681657934_0003
23/09/14 09:12:31 INFO impl.YarnClientImpl: Submitted application application_1694681657934_0003
23/09/14 09:12:31 INFO mapreduce.Job: The url to track the job: http://singlenode:8088/proxy/application_1694681657934_0003/
23/09/14 09:12:31 INFO mapreduce.Job: Running job: job_1694681657934_0003
23/09/14 09:12:38 INFO mapreduce.Job: Job job_1694681657934_0003 running in uber mode : false
23/09/14 09:12:38 INFO mapreduce.Job:  map 0% reduce 0%
23/09/14 09:12:43 INFO mapreduce.Job:  map 100% reduce 0%
23/09/14 09:12:47 INFO mapreduce.Job:  map 100% reduce 100%
23/09/14 09:12:48 INFO mapreduce.Job: Job job_1694681657934_0003 completed successfully
23/09/14 09:12:48 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=92
                FILE: Number of bytes written=397293
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=12000132
                HDFS: Number of bytes written=22
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=2748
                Total time spent by all reduces in occupied slots (ms)=2409
                Total time spent by all map tasks (ms)=2748
                Total time spent by all reduce tasks (ms)=2409
                Total vcore-milliseconds taken by all map tasks=2748
                Total vcore-milliseconds taken by all reduce tasks=2409
                Total megabyte-milliseconds taken by all map tasks=2813952
                Total megabyte-milliseconds taken by all reduce tasks=2466816
        Map-Reduce Framework
                Map input records=120000
                Map output records=3
                Map output bytes=80
                Map output materialized bytes=92
                Input split bytes=132
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=92
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=190
                CPU time spent (ms)=1540
                Physical memory (bytes) snapshot=488738816
                Virtual memory (bytes) snapshot=3764727808
                Total committed heap usage (bytes)=310902784
        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=12000000
        File Output Format Counters
                Bytes Written=22
```

b. [40 marks] Multi-node Hadoop Cluster Setup

bi)

After starting the multi-node Hadoop cluster, we can use command jps to confirm the running is correct. For 1 namenode and 4 slaves

```
hduser@namenode-1:~$ jps
16960 DataNode
17808 Jps
17298 ResourceManager
17170 SecondaryNameNode
17571 NodeManager
16777 NameNode
```

```
hduser@datanode-1:~$ jps
1156 NodeManager
2570 Jps
987 DataNode
```

```
hduser@datanode-2:~$ jps
2563 Jps
1143 NodeManager
974 DataNode
```

```
hduser@datanode-3:~$ jps
994 DataNode
2487 Jps
1164 NodeManager
```

bii)

My SID information:

1155157657 % 3 + 1 = 2

1155157657 % 20 + 10 = 27

Teragen for smaller size dataset (2GB):

```
hduser@namenode-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teragen 20000000 smaller/input
23/09/15 21:05:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/15 21:05:22 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/15 21:05:23 INFO terasort.TeraGen: Generating 20000000 using 2
23/09/15 21:05:23 INFO mapreduce.JobSubmitter: number of splits:2
23/09/15 21:05:23 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/15 21:05:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694811589268_0001
23/09/15 21:05:24 INFO impl.YarnClientImpl: Submitted application application_1694811589268_0001
23/09/15 21:05:24 INFO mapreduce.Job: The url to track the job: http://namenode-1:8088/proxy/application_1694811589268_0001/
23/09/15 21:05:24 INFO mapreduce.Job: Running job: job_1694811589268_0001
23/09/15 21:05:32 INFO mapreduce.Job: Job job_1694811589268_0001 running in uber mode : false
23/09/15 21:05:32 INFO mapreduce.Job:  map 0% reduce 0%
23/09/15 21:05:50 INFO mapreduce.Job:  map 75% reduce 0%
23/09/15 21:05:52 INFO mapreduce.Job:  map 85% reduce 0%
23/09/15 21:05:54 INFO mapreduce.Job:  map 100% reduce 0%
23/09/15 21:05:54 INFO mapreduce.Job: Job job_1694811589268_0001 completed successfully
23/09/15 21:05:54 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396714
                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=2000000000
                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
                Total time spent by all maps in occupied slots (ms)=37152
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=37152
                Total vcore-milliseconds taken by all map tasks=37152
                Total megabyte-milliseconds taken by all map tasks=38043648
        Map-Reduce Framework
                Map input records=20000000
                Map output records=20000000
                Input split bytes=167
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=393
                CPU time spent (ms)=33140
                Physical memory (bytes) snapshot=393490432
                Virtual memory (bytes) snapshot=3771244544
                Total committed heap usage (bytes)=237502464
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=42957274697559007
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=2000000000
```

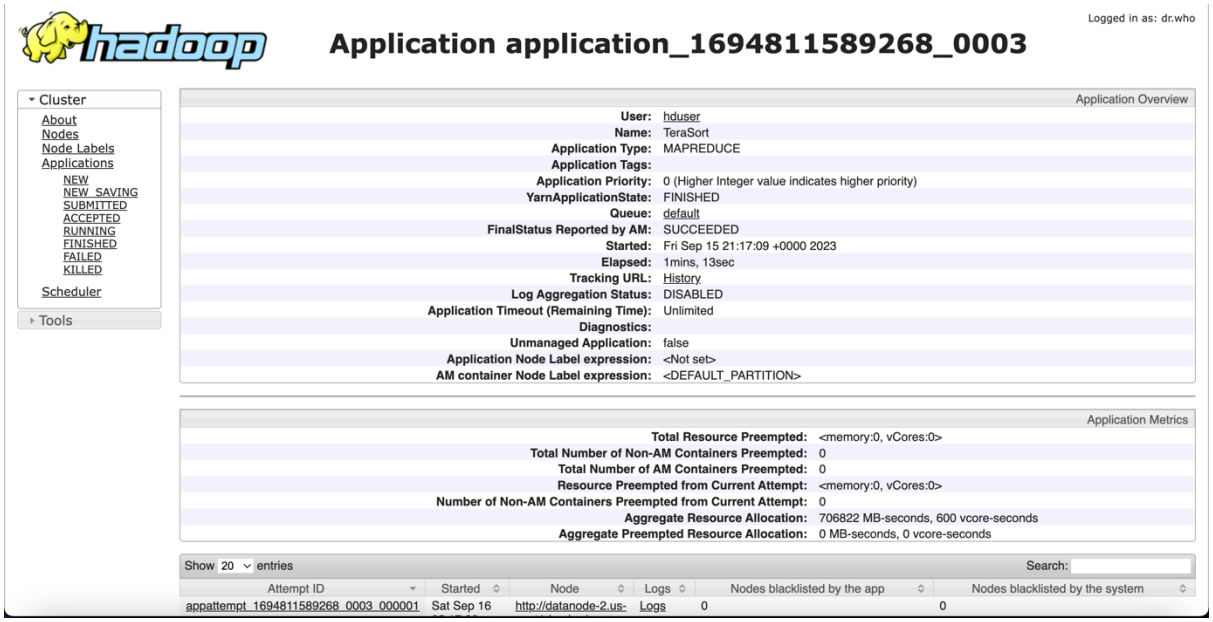
Teragen for larger size dataset (27GB):

```
hduser@namenode-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teragen 270000000 bigger/input
23/09/15 21:10:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/15 21:10:04 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/15 21:10:04 INFO terasort.TeraGen: Generating 270000000 using 2
23/09/15 21:10:05 INFO mapreduce.JobSubmitter: number of splits:2
23/09/15 21:10:05 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/15 21:10:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694811589268_0002
23/09/15 21:10:05 INFO impl.YarnClientImpl: Submitted application application_1694811589268_0002
23/09/15 21:10:05 INFO mapreduce.Job: The url to track the job: http://namenode-1:8088/proxy/application_1694811589268_0002/
23/09/15 21:10:05 INFO mapreduce.Job: Running job: job_1694811589268_0002
23/09/15 21:10:11 INFO mapreduce.Job: Job job_1694811589268_0002 running in uber mode : false
23/09/15 21:10:11 INFO mapreduce.Job:  map 0% reduce 0%
23/09/15 21:10:28 INFO mapreduce.Job:  map 7% reduce 0%
23/09/15 21:10:33 INFO mapreduce.Job:  map 9% reduce 0%
23/09/15 21:10:34 INFO mapreduce.Job:  map 11% reduce 0%
23/09/15 21:10:40 INFO mapreduce.Job:  map 12% reduce 0%
23/09/15 21:10:41 INFO mapreduce.Job:  map 14% reduce 0%
23/09/15 21:10:46 INFO mapreduce.Job:  map 16% reduce 0%
23/09/15 21:10:47 INFO mapreduce.Job:  map 18% reduce 0%
23/09/15 21:10:52 INFO mapreduce.Job:  map 20% reduce 0%
23/09/15 21:10:53 INFO mapreduce.Job:  map 22% reduce 0%
23/09/15 21:10:58 INFO mapreduce.Job:  map 24% reduce 0%
23/09/15 21:10:59 INFO mapreduce.Job:  map 25% reduce 0%
23/09/15 21:11:04 INFO mapreduce.Job:  map 27% reduce 0%
23/09/15 21:11:05 INFO mapreduce.Job:  map 29% reduce 0%
23/09/15 21:11:10 INFO mapreduce.Job:  map 33% reduce 0%
23/09/15 21:11:16 INFO mapreduce.Job:  map 37% reduce 0%
23/09/15 21:11:22 INFO mapreduce.Job:  map 41% reduce 0%
23/09/15 21:11:28 INFO mapreduce.Job:  map 44% reduce 0%
23/09/15 21:11:34 INFO mapreduce.Job:  map 48% reduce 0%
23/09/15 21:11:40 INFO mapreduce.Job:  map 52% reduce 0%
23/09/15 21:11:46 INFO mapreduce.Job:  map 55% reduce 0%
23/09/15 21:11:52 INFO mapreduce.Job:  map 59% reduce 0%
23/09/15 21:11:58 INFO mapreduce.Job:  map 63% reduce 0%
23/09/15 21:12:04 INFO mapreduce.Job:  map 66% reduce 0%
23/09/15 21:12:10 INFO mapreduce.Job:  map 70% reduce 0%
23/09/15 21:12:16 INFO mapreduce.Job:  map 74% reduce 0%
23/09/15 21:12:22 INFO mapreduce.Job:  map 78% reduce 0%
23/09/15 21:12:28 INFO mapreduce.Job:  map 81% reduce 0%
23/09/15 21:12:34 INFO mapreduce.Job:  map 86% reduce 0%
23/09/15 21:12:40 INFO mapreduce.Job:  map 90% reduce 0%
23/09/15 21:12:46 INFO mapreduce.Job:  map 94% reduce 0%
23/09/15 21:12:53 INFO mapreduce.Job:  map 98% reduce 0%
23/09/15 21:12:59 INFO mapreduce.Job:  map 100% reduce 0%
23/09/15 21:13:01 INFO mapreduce.Job: Job job_1694811589268_0002 completed successfully
23/09/15 21:13:01 INFO mapreduce.Job: Counters: 31
```

```
23/09/15 21:13:01 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396714
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=170
                HDFS: Number of bytes written=27000000000
                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
                Total time spent by all maps in occupied slots (ms)=326717
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=326717
                Total vcore-milliseconds taken by all map tasks=326717
                Total megabyte-milliseconds taken by all map tasks=334558208
        Map-Reduce Framework
                Map input records=270000000
                Map output records=270000000
                Input split bytes=170
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2388
                CPU time spent (ms)=318430
                Physical memory (bytes) snapshot=480358400
                Virtual memory (bytes) snapshot=3763679232
                Total committed heap usage (bytes)=275775488
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=57985185607766483
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=27000000000
```

Terasort for smaller dataset (2GB):

```
hduser@namenode-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort smaller/input smaller/output
23/09/15 21:17:06 INFO terasort.TeraSort: starting
23/09/15 21:17:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/15 21:17:07 INFO input.FileInputFormat: Total input files to process : 2
Spent 111ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
Computing input splits took 115ms
Sampling 10 splits of 16
Making 1 from 100000 sampled records
Computing parititions took 792ms
Spent 910ms computing partitions.
23/09/15 21:17:08 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/15 21:17:09 INFO mapreduce.JobSubmitter: number of splits:16
23/09/15 21:17:09 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/15 21:17:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694811589268_0003
23/09/15 21:17:09 INFO impl.YarnClientImpl: Submitted application application_1694811589268_0003
23/09/15 21:17:09 INFO mapreduce.Job: The url to track the job: http://namenode-1:8088/proxy/application_1694811589268_0003/
23/09/15 21:17:09 INFO mapreduce.Job: Running job: job_1694811589268_0003
23/09/15 21:17:15 INFO mapreduce.Job: Job job_1694811589268_0003 running in uber mode : false
23/09/15 21:17:15 INFO mapreduce.Job:  map 0% reduce 0%
23/09/15 21:17:37 INFO mapreduce.Job:  map 6% reduce 0%
23/09/15 21:17:38 INFO mapreduce.Job:  map 13% reduce 0%
23/09/15 21:17:39 INFO mapreduce.Job:  map 39% reduce 0%
23/09/15 21:17:40 INFO mapreduce.Job:  map 58% reduce 0%
23/09/15 21:17:41 INFO mapreduce.Job:  map 67% reduce 0%
23/09/15 21:17:45 INFO mapreduce.Job:  map 78% reduce 0%
23/09/15 21:17:46 INFO mapreduce.Job:  map 85% reduce 0%
23/09/15 21:17:48 INFO mapreduce.Job:  map 88% reduce 0%
23/09/15 21:17:49 INFO mapreduce.Job:  map 92% reduce 0%
23/09/15 21:17:51 INFO mapreduce.Job:  map 95% reduce 0%
23/09/15 21:17:52 INFO mapreduce.Job:  map 99% reduce 0%
23/09/15 21:17:53 INFO mapreduce.Job:  map 100% reduce 0%
23/09/15 21:18:00 INFO mapreduce.Job:  map 100% reduce 51%
23/09/15 21:18:06 INFO mapreduce.Job:  map 100% reduce 70%
23/09/15 21:18:12 INFO mapreduce.Job:  map 100% reduce 80%
23/09/15 21:18:18 INFO mapreduce.Job:  map 100% reduce 90%
23/09/15 21:18:24 INFO mapreduce.Job:  map 100% reduce 100%
23/09/15 21:18:24 INFO mapreduce.Job: Job job_1694811589268_0003 completed successfully
23/09/15 21:18:24 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=4857932310
                FILE: Number of bytes written=6941328600
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2000002032
                HDFS: Number of bytes written=2000000000
                HDFS: Number of read operations=51
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=16
                Launched reduce tasks=1
                Data-local map tasks=16
                Total time spent by all maps in occupied slots (ms)=456727
                Total time spent by all reduces in occupied slots (ms)=44046
                Total time spent by all map tasks (ms)=456727
                Total time spent by all reduce tasks (ms)=44046
                Total vcore-milliseconds taken by all map tasks=456727
                Total vcore-milliseconds taken by all reduce tasks=44046
                Total megabyte-milliseconds taken by all map tasks=467688448
                Total megabyte-milliseconds taken by all reduce tasks=45103104
        Map-Reduce Framework
                Map input records=20000000
                Map output records=20000000
                Map output bytes=2040000000
                Map output materialized bytes=2080000096
                Input split bytes=2032
                Combine input records=0
                Combine output records=0
                Reduce input groups=20000000
                Reduce shuffle bytes=2080000096
                Reduce input records=20000000
                Reduce output records=20000000
                Spilled Records=66710885
                Shuffled Maps =16
                Failed Shuffles=0
                Merged Map outputs=16
                GC time elapsed (ms)=5220
                CPU time spent (ms)=179930
                Physical memory (bytes) snapshot=4768710656
                Virtual memory (bytes) snapshot=31860948992
                Total committed heap usage (bytes)=3303014400
        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=2000000000
        File Output Format Counters
                Bytes Written=2000000000
23/09/15 21:18:24 INFO terasort.TeraSort: done
```

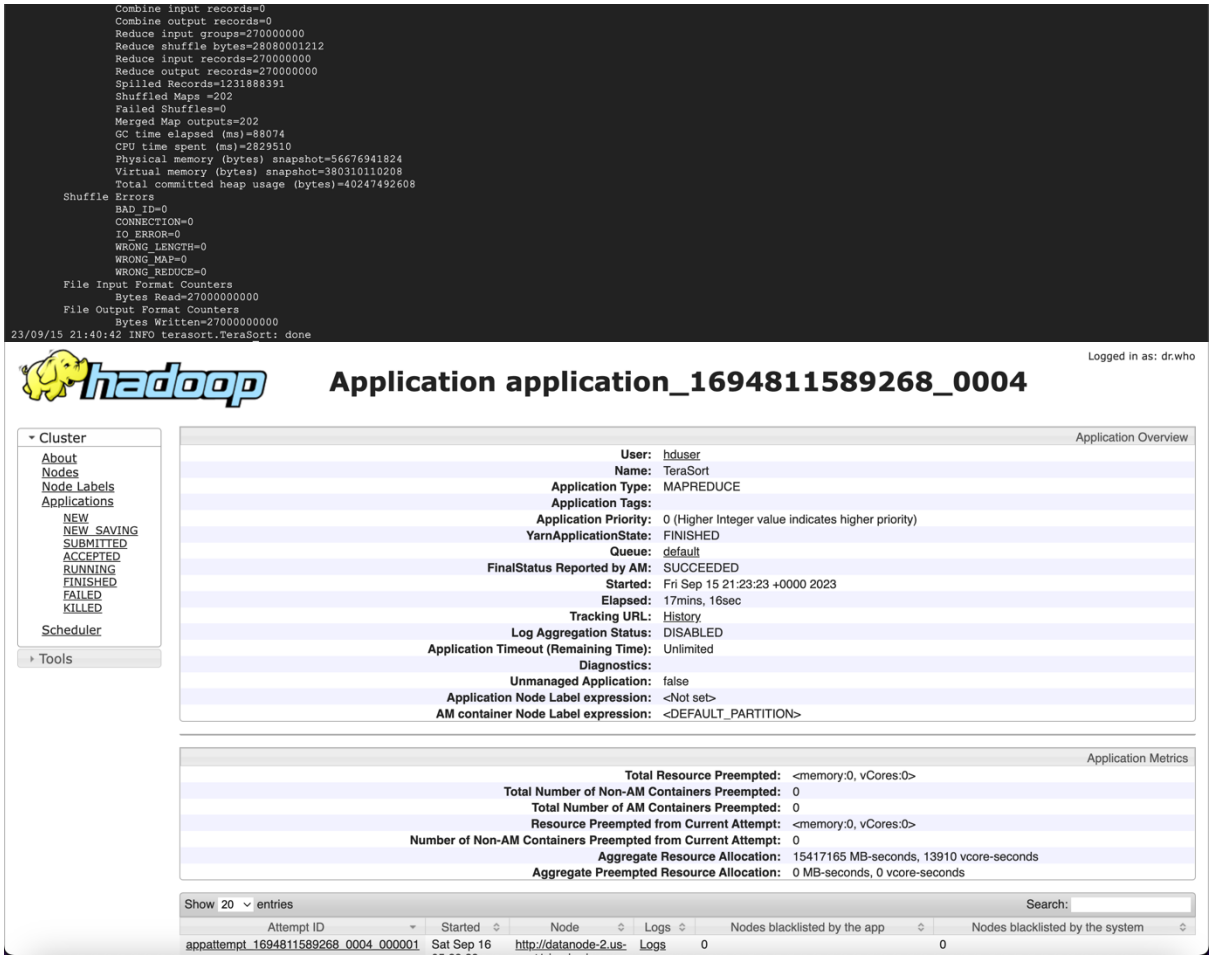


It uses 1m13s to map-reduce the smaller dataset

For the bigger dataset (27GB):

```
hduser@namenode-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort bigger/input bigger/output
23/09/15 21:23:20 INFO terasort.TeraSort: starting
23/09/15 21:23:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/15 21:23:21 INFO input.FileInputFormat: Total input files to process : 2
Spent 164ms computing base-splits.
Spent 7ms computing TeraScheduler splits.
Computing input splits took 172ms
Sampling 10 splits of 202
Making 1 from 100000 sampled records
Computing parititions took 735ms
Spent 910ms computing partitions.
23/09/15 21:23:22 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/15 21:23:23 INFO mapreduce.JobSubmitter: number of splits:202
23/09/15 21:23:23 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/15 21:23:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694811589268_0004
23/09/15 21:23:23 INFO impl.YarnClientImpl: Submitted application application_1694811589268_0004
23/09/15 21:23:23 INFO mapreduce.Job: The url to track the job: http://namenode-1:8088/proxy/application_1694811589268_0004/
23/09/15 21:23:23 INFO mapreduce.Job: Running job: job_1694811589268_0004
23/09/15 21:23:30 INFO mapreduce.Job: Job job_1694811589268_0004 running in uber mode : false
23/09/15 21:23:30 INFO mapreduce.Job:  map 0% reduce 0%
23/09/15 21:23:58 INFO mapreduce.Job:  map 1% reduce 0%
23/09/15 21:24:02 INFO mapreduce.Job:  map 5% reduce 0%
23/09/15 21:24:03 INFO mapreduce.Job:  map 6% reduce 0%
23/09/15 21:24:04 INFO mapreduce.Job:  map 7% reduce 0%
23/09/15 21:24:14 INFO mapreduce.Job:  map 8% reduce 0%
23/09/15 21:24:16 INFO mapreduce.Job:  map 9% reduce 0%
23/09/15 21:24:20 INFO mapreduce.Job:  map 10% reduce 0%
23/09/15 21:24:21 INFO mapreduce.Job:  map 11% reduce 0%
23/09/15 21:24:32 INFO mapreduce.Job:  map 13% reduce 0%
23/09/15 21:24:33 INFO mapreduce.Job:  map 14% reduce 0%
23/09/15 21:24:37 INFO mapreduce.Job:  map 15% reduce 0%
23/09/15 21:24:44 INFO mapreduce.Job:  map 16% reduce 0%
23/09/15 21:24:57 INFO mapreduce.Job:  map 17% reduce 0%
23/09/15 21:25:01 INFO mapreduce.Job:  map 18% reduce 0%
23/09/15 21:25:03 INFO mapreduce.Job:  map 19% reduce 0%
23/09/15 21:25:04 INFO mapreduce.Job:  map 20% reduce 0%
23/09/15 21:25:05 INFO mapreduce.Job:  map 20% reduce 3%
23/09/15 21:25:06 INFO mapreduce.Job:  map 21% reduce 3%
23/09/15 21:25:09 INFO mapreduce.Job:  map 23% reduce 3%
23/09/15 21:25:11 INFO mapreduce.Job:  map 23% reduce 4%
23/09/15 21:25:17 INFO mapreduce.Job:  map 23% reduce 5%
23/09/15 21:25:21 INFO mapreduce.Job:  map 24% reduce 5%
23/09/15 21:25:23 INFO mapreduce.Job:  map 24% reduce 6%
23/09/15 21:25:25 INFO mapreduce.Job:  map 25% reduce 6%
23/09/15 21:25:28 INFO mapreduce.Job:  map 26% reduce 6%
23/09/15 21:25:33 INFO mapreduce.Job:  map 27% reduce 6%

23/09/15 21:25:36 INFO mapreduce.Job:  map 28% reduce 6%
23/09/15 21:25:40 INFO mapreduce.Job:  map 29% reduce 6%
23/09/15 21:25:41 INFO mapreduce.Job:  map 30% reduce 6%
23/09/15 21:25:42 INFO mapreduce.Job:  map 30% reduce 7%
23/09/15 21:25:43 INFO mapreduce.Job:  map 31% reduce 7%
23/09/15 21:25:48 INFO mapreduce.Job:  map 31% reduce 9%
23/09/15 21:25:52 INFO mapreduce.Job:  map 32% reduce 9%
23/09/15 21:26:00 INFO mapreduce.Job:  map 32% reduce 10%
23/09/15 21:26:06 INFO mapreduce.Job:  map 33% reduce 10%
23/09/15 21:26:09 INFO mapreduce.Job:  map 34% reduce 10%
23/09/15 21:26:12 INFO mapreduce.Job:  map 34% reduce 11%
23/09/15 21:26:13 INFO mapreduce.Job:  map 35% reduce 11%
23/09/15 21:26:14 INFO mapreduce.Job:  map 36% reduce 11%
23/09/15 21:26:15 INFO mapreduce.Job:  map 37% reduce 11%
23/09/15 21:26:18 INFO mapreduce.Job:  map 38% reduce 11%
23/09/15 21:26:23 INFO mapreduce.Job:  map 39% reduce 11%
23/09/15 21:26:28 INFO mapreduce.Job:  map 40% reduce 11%
23/09/15 21:26:32 INFO mapreduce.Job:  map 41% reduce 11%
23/09/15 21:26:34 INFO mapreduce.Job:  map 42% reduce 11%
23/09/15 21:26:38 INFO mapreduce.Job:  map 43% reduce 11%
23/09/15 21:26:41 INFO mapreduce.Job:  map 44% reduce 11%
23/09/15 21:26:42 INFO mapreduce.Job:  map 44% reduce 12%
23/09/15 21:26:44 INFO mapreduce.Job:  map 45% reduce 12%
23/09/15 21:26:47 INFO mapreduce.Job:  map 46% reduce 12%
23/09/15 21:26:54 INFO mapreduce.Job:  map 46% reduce 13%
23/09/15 21:27:00 INFO mapreduce.Job:  map 46% reduce 15%
23/09/15 21:27:02 INFO mapreduce.Job:  map 47% reduce 15%
23/09/15 21:27:05 INFO mapreduce.Job:  map 48% reduce 15%
23/09/15 21:27:10 INFO mapreduce.Job:  map 49% reduce 15%
23/09/15 21:27:15 INFO mapreduce.Job:  map 50% reduce 15%
23/09/15 21:27:17 INFO mapreduce.Job:  map 51% reduce 15%
23/09/15 21:27:19 INFO mapreduce.Job:  map 52% reduce 15%
23/09/15 21:27:21 INFO mapreduce.Job:  map 53% reduce 15%
23/09/15 21:27:22 INFO mapreduce.Job:  map 54% reduce 15%
23/09/15 21:27:24 INFO mapreduce.Job:  map 54% reduce 16%
23/09/15 21:27:25 INFO mapreduce.Job:  map 55% reduce 16%
23/09/15 21:27:34 INFO mapreduce.Job:  map 56% reduce 16%
23/09/15 21:27:36 INFO mapreduce.Job:  map 56% reduce 17%
23/09/15 21:27:37 INFO mapreduce.Job:  map 57% reduce 17%
23/09/15 21:27:40 INFO mapreduce.Job:  map 58% reduce 17%
23/09/15 21:27:48 INFO mapreduce.Job:  map 59% reduce 18%
23/09/15 21:27:50 INFO mapreduce.Job:  map 60% reduce 18%
23/09/15 21:27:51 INFO mapreduce.Job:  map 61% reduce 18%
23/09/15 21:27:57 INFO mapreduce.Job:  map 62% reduce 18%
23/09/15 21:28:00 INFO mapreduce.Job:  map 62% reduce 19%
23/09/15 21:28:05 INFO mapreduce.Job:  map 63% reduce 19%

23/09/15 21:28:06 INFO mapreduce.Job:  map 63% reduce 20%
23/09/15 21:28:09 INFO mapreduce.Job:  map 64% reduce 20%
23/09/15 21:28:11 INFO mapreduce.Job:  map 65% reduce 20%
23/09/15 21:28:18 INFO mapreduce.Job:  map 65% reduce 21%
23/09/15 21:28:21 INFO mapreduce.Job:  map 66% reduce 21%
23/09/15 21:28:23 INFO mapreduce.Job:  map 67% reduce 21%
23/09/15 21:28:25 INFO mapreduce.Job:  map 69% reduce 21%
23/09/15 21:28:36 INFO mapreduce.Job:  map 70% reduce 21%
23/09/15 21:28:39 INFO mapreduce.Job:  map 71% reduce 21%
23/09/15 21:28:41 INFO mapreduce.Job:  map 72% reduce 21%
23/09/15 21:28:43 INFO mapreduce.Job:  map 73% reduce 21%
23/09/15 21:28:48 INFO mapreduce.Job:  map 74% reduce 22%
23/09/15 21:28:52 INFO mapreduce.Job:  map 75% reduce 22%
23/09/15 21:28:55 INFO mapreduce.Job:  map 76% reduce 22%
23/09/15 21:28:59 INFO mapreduce.Job:  map 77% reduce 22%
23/09/15 21:29:01 INFO mapreduce.Job:  map 77% reduce 23%
23/09/15 21:29:07 INFO mapreduce.Job:  map 77% reduce 25%
23/09/15 21:29:11 INFO mapreduce.Job:  map 78% reduce 25%
23/09/15 21:29:14 INFO mapreduce.Job:  map 79% reduce 25%
23/09/15 21:29:24 INFO mapreduce.Job:  map 80% reduce 25%
23/09/15 21:29:26 INFO mapreduce.Job:  map 82% reduce 25%
23/09/15 21:29:30 INFO mapreduce.Job:  map 84% reduce 25%
23/09/15 21:29:31 INFO mapreduce.Job:  map 84% reduce 26%
23/09/15 21:29:33 INFO mapreduce.Job:  map 85% reduce 26%
23/09/15 21:29:42 INFO mapreduce.Job:  map 86% reduce 26%
23/09/15 21:29:43 INFO mapreduce.Job:  map 87% reduce 26%
23/09/15 21:29:44 INFO mapreduce.Job:  map 88% reduce 26%
23/09/15 21:29:49 INFO mapreduce.Job:  map 88% reduce 27%
23/09/15 21:29:52 INFO mapreduce.Job:  map 89% reduce 27%
23/09/15 21:29:55 INFO mapreduce.Job:  map 89% reduce 28%
23/09/15 21:30:00 INFO mapreduce.Job:  map 90% reduce 28%
23/09/15 21:30:03 INFO mapreduce.Job:  map 91% reduce 28%
23/09/15 21:30:07 INFO mapreduce.Job:  map 91% reduce 29%
23/09/15 21:30:09 INFO mapreduce.Job:  map 92% reduce 29%
23/09/15 21:30:11 INFO mapreduce.Job:  map 93% reduce 29%
23/09/15 21:30:13 INFO mapreduce.Job:  map 93% reduce 30%
23/09/15 21:30:15 INFO mapreduce.Job:  map 94% reduce 30%
23/09/15 21:30:16 INFO mapreduce.Job:  map 95% reduce 30%
23/09/15 21:30:18 INFO mapreduce.Job:  map 96% reduce 30%
23/09/15 21:30:19 INFO mapreduce.Job:  map 96% reduce 31%
23/09/15 21:30:22 INFO mapreduce.Job:  map 97% reduce 31%
23/09/15 21:30:27 INFO mapreduce.Job:  map 98% reduce 31%
23/09/15 21:30:29 INFO mapreduce.Job:  map 99% reduce 31%
23/09/15 21:30:31 INFO mapreduce.Job:  map 99% reduce 32%
23/09/15 21:30:39 INFO mapreduce.Job:  map 100% reduce 32%
```

```
23/09/15 21:30:43 INFO mapreduce.Job:  map 100% reduce 33%
23/09/15 21:33:25 INFO mapreduce.Job:  map 100% reduce 34%
23/09/15 21:33:31 INFO mapreduce.Job:  map 100% reduce 36%
23/09/15 21:33:37 INFO mapreduce.Job:  map 100% reduce 37%
23/09/15 21:33:43 INFO mapreduce.Job:  map 100% reduce 39%
23/09/15 21:33:49 INFO mapreduce.Job:  map 100% reduce 40%
23/09/15 21:33:55 INFO mapreduce.Job:  map 100% reduce 41%
23/09/15 21:34:01 INFO mapreduce.Job:  map 100% reduce 43%
23/09/15 21:34:07 INFO mapreduce.Job:  map 100% reduce 44%
23/09/15 21:34:13 INFO mapreduce.Job:  map 100% reduce 46%
23/09/15 21:34:19 INFO mapreduce.Job:  map 100% reduce 47%
23/09/15 21:34:26 INFO mapreduce.Job:  map 100% reduce 48%
23/09/15 21:34:32 INFO mapreduce.Job:  map 100% reduce 50%
23/09/15 21:34:38 INFO mapreduce.Job:  map 100% reduce 51%
23/09/15 21:34:44 INFO mapreduce.Job:  map 100% reduce 52%
23/09/15 21:34:50 INFO mapreduce.Job:  map 100% reduce 54%
23/09/15 21:34:56 INFO mapreduce.Job:  map 100% reduce 55%
23/09/15 21:35:02 INFO mapreduce.Job:  map 100% reduce 57%
23/09/15 21:35:08 INFO mapreduce.Job:  map 100% reduce 58%
23/09/15 21:35:14 INFO mapreduce.Job:  map 100% reduce 59%
23/09/15 21:35:20 INFO mapreduce.Job:  map 100% reduce 61%
23/09/15 21:35:26 INFO mapreduce.Job:  map 100% reduce 62%
23/09/15 21:35:32 INFO mapreduce.Job:  map 100% reduce 63%
23/09/15 21:35:38 INFO mapreduce.Job:  map 100% reduce 65%
23/09/15 21:35:44 INFO mapreduce.Job:  map 100% reduce 66%
23/09/15 21:35:50 INFO mapreduce.Job:  map 100% reduce 67%
23/09/15 21:35:56 INFO mapreduce.Job:  map 100% reduce 68%
23/09/15 21:36:08 INFO mapreduce.Job:  map 100% reduce 69%
23/09/15 21:36:14 INFO mapreduce.Job:  map 100% reduce 70%
23/09/15 21:36:26 INFO mapreduce.Job:  map 100% reduce 71%
23/09/15 21:36:32 INFO mapreduce.Job:  map 100% reduce 72%
23/09/15 21:36:44 INFO mapreduce.Job:  map 100% reduce 73%
23/09/15 21:36:50 INFO mapreduce.Job:  map 100% reduce 74%
23/09/15 21:37:02 INFO mapreduce.Job:  map 100% reduce 75%
23/09/15 21:37:08 INFO mapreduce.Job:  map 100% reduce 76%
23/09/15 21:37:20 INFO mapreduce.Job:  map 100% reduce 77%
23/09/15 21:37:26 INFO mapreduce.Job:  map 100% reduce 78%
23/09/15 21:37:32 INFO mapreduce.Job:  map 100% reduce 79%
23/09/15 21:37:44 INFO mapreduce.Job:  map 100% reduce 80%
23/09/15 21:37:50 INFO mapreduce.Job:  map 100% reduce 81%
23/09/15 21:38:02 INFO mapreduce.Job:  map 100% reduce 82%
23/09/15 21:38:08 INFO mapreduce.Job:  map 100% reduce 83%
23/09/15 21:38:20 INFO mapreduce.Job:  map 100% reduce 84%
23/09/15 21:38:26 INFO mapreduce.Job:  map 100% reduce 85%
23/09/15 21:38:38 INFO mapreduce.Job:  map 100% reduce 86%
23/09/15 21:38:44 INFO mapreduce.Job:  map 100% reduce 87%
23/09/15 21:38:56 INFO mapreduce.Job:  map 100% reduce 88%
23/09/15 21:39:02 INFO mapreduce.Job:  map 100% reduce 89%
23/09/15 21:39:08 INFO mapreduce.Job:  map 100% reduce 90%
23/09/15 21:39:20 INFO mapreduce.Job:  map 100% reduce 91%
23/09/15 21:39:26 INFO mapreduce.Job:  map 100% reduce 92%
23/09/15 21:39:38 INFO mapreduce.Job:  map 100% reduce 93%
23/09/15 21:39:44 INFO mapreduce.Job:  map 100% reduce 94%
23/09/15 21:39:56 INFO mapreduce.Job:  map 100% reduce 95%
23/09/15 21:40:02 INFO mapreduce.Job:  map 100% reduce 96%
23/09/15 21:40:14 INFO mapreduce.Job:  map 100% reduce 97%
23/09/15 21:40:20 INFO mapreduce.Job:  map 100% reduce 98%
23/09/15 21:40:32 INFO mapreduce.Job:  map 100% reduce 99%
23/09/15 21:40:38 INFO mapreduce.Job:  map 100% reduce 100%
23/09/15 21:40:41 INFO mapreduce.Job: Job job_1694811589268_0004 completed successfully
23/09/15 21:40:42 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=100036396432
                FILE: Number of bytes written=128156952614
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=27000025452
                HDFS: Number of bytes written=27000000000
                HDFS: Number of read operations=609
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=203
                Launched reduce tasks=1
                Data-local map tasks=201
                Rack-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=11723303
                Total time spent by all reduces in occupied slots (ms)=966120
                Total time spent by all map tasks (ms)=11723303
                Total time spent by all reduce tasks (ms)=966120
                Total vcore-milliseconds taken by all map tasks=11723303
                Total vcore-milliseconds taken by all reduce tasks=966120
                Total megabyte-milliseconds taken by all map tasks=12004662272
                Total megabyte-milliseconds taken by all reduce tasks=989306880
        Map-Reduce Framework
                Map input records=2700000000
                Map output records=2700000000
                Map output bytes=27540000000
                Map output materialized bytes=28080001212
                Input split bytes=25452
```



The bigger dataset took 17m16s to finish the map-reduce task

Compare:
Smaller dataset (2GB): 1m13s
Bigger dataset (27GB): 17m16s

Proving both job is running on multi-node cluster with 1 namenode and 4 slaves

41 files and directories, 447 blocks = 488 total filesystem object(s).

Heap Memory used 57.38 MB of 267 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 56.55 MB of 57.66 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	386.93 GB
DFS Used:	108.94 GB (28.16%)
Non DFS Used:	12.53 GB
DFS Remaining:	265.4 GB (68.59%)
Block Pool Used:	108.94 GB (28.16%)
DataNodes usages% (Min/Median/Max/stdDev):	22.13% / 22.54% / 45.67% / 10.11%
Live Nodes	4 (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	0
Block Deletion Start Time	Sat Sep 16 04:59:24 +0800 2023
Last Checkpoint Time	Sat Sep 16 05:10:43 +0800 2023

In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✔datanode-1.us-east1-b.c.hadoop-cluster-399115.internal:50010 (10.142.0.11:50010)	http://datanode-1.us-east1-b.c.hadoop-cluster-399115.internal:50075	2s	46m	96.73 GB	174	21.4 GB (22.13%)	2.9.2
✔datanode-2.us-east1-b.c.hadoop-cluster-399115.internal:50010 (10.142.0.12:50010)	http://datanode-2.us-east1-b.c.hadoop-cluster-399115.internal:50075	2s	46m	96.73 GB	184	21.55 GB (22.28%)	2.9.2
✔datanode-3.us-east1-b.c.hadoop-cluster-399115.internal:50010 (10.142.0.13:50010)	http://datanode-3.us-east1-b.c.hadoop-cluster-399115.internal:50075	2s	46m	96.73 GB	357	44.18 GB (45.67%)	2.9.2
✔namenode-1.us-east1-b.c.hadoop-cluster-399115.internal:50010 (10.142.0.10:50010)	http://namenode-1.us-east1-b.c.hadoop-cluster-399115.internal:50075	2s	5m	96.73 GB	180	21.8 GB (22.54%)	2.9.2

c. [30 marks] Running Python Code on Hadoop

i) Hadoop streaming

The mapper function copied from homework reference and I modify the syntax, making it running on python3:

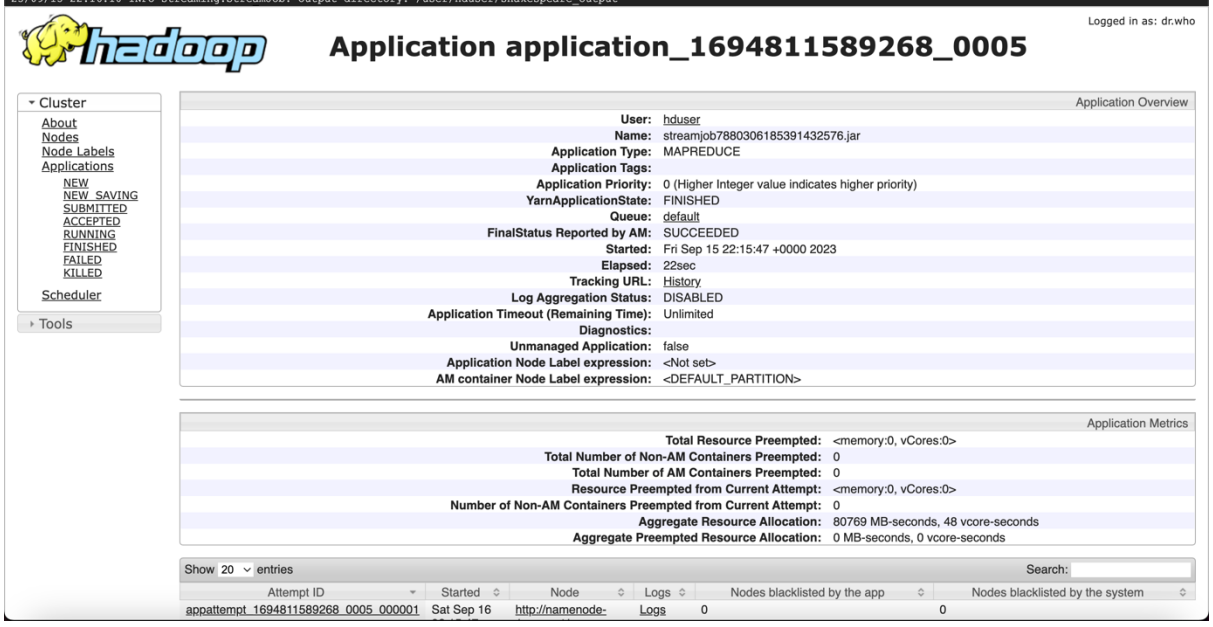
```
hduser@namenode-1:/usr/local/hadoop$ cat /home/hduser/mapper.py
#!/usr/bin/env python3
"""mapper.py"""

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print('%s\t%s' % (word, 1))
```

The reducer function copied from homework reference and I modify the syntax, making it running on python3:

```
hduser@namenode-1:/usr/local/hadoop$ cat /home/hduser/reducer.py
#!/usr/bin/env python3
"""reducer.py"""

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print('%s\t%s' % (current_word, current_count))
```

I copied the command and modified the location of the files in the command, and here is the output:

```
hduser@namenode-1:/usr/local/hadoop$ ./bin/hadoop jar ./share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar -file /home/hduser/mapper.py -mapper /home/hduser/mapper.py -file /home/hd
user/reducer.py -reducer /home/hduser/reducer.py -input /user/hduser/shakespeare -output /user/hduser/shakespeare_output
23/09/15 22:15:45 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
23/09/15 22:15:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/home/hduser/mapper.py, /home/hduser/reducer.py, /tmp/hadoop-unjar1765225785163078174/] [] /tmp/streamjob7880306185391432576.jar tmpDir=null
23/09/15 22:15:46 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/15 22:15:46 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/15 22:15:46 INFO mapred.FileInputFormat: Total input files to process : 1
23/09/15 22:15:46 INFO mapreduce.JobSubmitter: number of splits:2
23/09/15 22:15:47 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/15 22:15:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694811589268_0005
23/09/15 22:15:47 INFO impl.YarnClientImpl: Submitted application application_1694811589268_0005
23/09/15 22:15:47 INFO mapreduce.Job: The url to track the job: http://namenode-1:8088/proxy/application_1694811589268_0005/
23/09/15 22:15:47 INFO mapreduce.Job: Running job: job_1694811589268_0005
23/09/15 22:15:54 INFO mapreduce.Job: Job job_1694811589268_0005 running in uber mode : false
23/09/15 22:15:54 INFO mapreduce.Job:  map 0% reduce 0%
23/09/15 22:16:03 INFO mapreduce.Job:  map 100% reduce 0%
23/09/15 22:16:10 INFO mapreduce.Job:  map 100% reduce 100%
23/09/15 22:16:10 INFO mapreduce.Job: Job job_1694811589268_0005 completed successfully
23/09/15 22:16:10 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=7841073
                FILE: Number of bytes written=16288901
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4585857
                HDFS: Number of bytes written=878840
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=13881
                Total time spent by all reduces in occupied slots (ms)=4196
                Total time spent by all map tasks (ms)=13881
                Total time spent by all reduce tasks (ms)=4196
                Total vcore-milliseconds taken by all map tasks=13881
                Total vcore-milliseconds taken by all reduce tasks=4196
                Total megabyte-milliseconds taken by all map tasks=14214144
                Total megabyte-milliseconds taken by all reduce tasks=4296704
        Map-Reduce Framework
                Map input records=111396
                Map output records=814671
                Map output bytes=6211725
                Map output materialized bytes=7841079
                Input split bytes=198
                Combine input records=0
                Combine output records=0
                Reduce input groups=80364
                Reduce shuffle bytes=7841079
                Reduce input records=814671
                Reduce output records=80364
                Spilled Records=1629342
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=371
                CPU time spent (ms)=6840
                Physical memory (bytes) snapshot=753680384
                Virtual memory (bytes) snapshot=5622190080
                Total committed heap usage (bytes)=507510784
        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=4585659
        File Output Format Counters
                Bytes Written=878840
23/09/15 22:16:10 INFO streaming.StreamJob: Output directory: /user/hduser/shakespeare_output
```



Running time 22s

d. [Bonus 20 marks] Compiling the Java WordCount program

(This is a bonus question for me as I am NOT ESTR student)

added this line to ~/.bashrc and update the environment variable

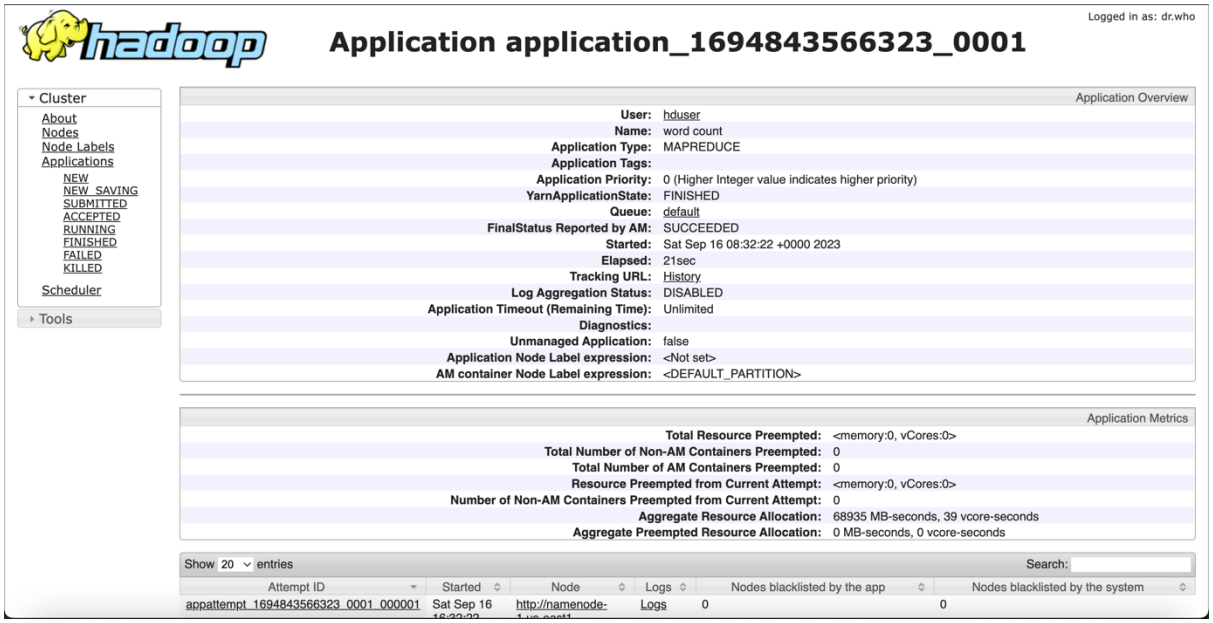
```
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
```

After copying the file from the template from Hadoop website, I save it to /usr/local/Hadoop and run the following two command according to the instruction

```
hduser@namenode-1:/usr/local/hadoop$ bin/hadoop com.sun.tools.javac.Main WordCount.java
hduser@namenode-1:/usr/local/hadoop$ jar cf wc.jar WordCount*.class
```

Run the map-reduce for Shakespeare dataset

```
hduser@namenode-1:/usr/local/hadoop$ bin/hadoop jar wc.jar WordCount /user/hduser/shakespeare /user/hduser/wordcount/output
23/09/16 08:32:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/09/16 08:32:20 INFO client.RMProxy: Connecting to ResourceManager at namenode-1/10.142.0.10:8032
23/09/16 08:32:21 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to
 remedy this.
23/09/16 08:32:21 INFO input.FileInputFormat: Total input files to process : 1
23/09/16 08:32:21 INFO mapreduce.JobSubmitter: number of splits:1
23/09/16 08:32:21 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
23/09/16 08:32:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694843566323_0001
23/09/16 08:32:22 INFO impl.YarnClientImpl: Submitted application application_1694843566323_0001
23/09/16 08:32:22 INFO mapreduce.Job: The url to track the job: http://namenode-1:8088/proxy/application_1694843566323_0001/
23/09/16 08:32:22 INFO mapreduce.Job: Running job: job_1694843566323_0001
23/09/16 08:32:30 INFO mapreduce.Job: Job job_1694843566323_0001 running in uber mode : false
23/09/16 08:32:30 INFO mapreduce.Job:  map 0% reduce 0%
23/09/16 08:32:37 INFO mapreduce.Job:  map 100% reduce 0%
23/09/16 08:32:44 INFO mapreduce.Job:  map 100% reduce 100%
23/09/16 08:32:44 INFO mapreduce.Job: Job job_1694843566323_0001 completed successfully
23/09/16 08:32:44 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1192234
                FILE: Number of bytes written=2781415
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4583910
                HDFS: Number of bytes written=878840
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=5282
                Total time spent by all reduces in occupied slots (ms)=4186
                Total time spent by all map tasks (ms)=5282
                Total time spent by all reduce tasks (ms)=4186
                Total vcore-milliseconds taken by all map tasks=5282
                Total vcore-milliseconds taken by all reduce tasks=4186
                Total megabyte-milliseconds taken by all map tasks=5408768
                Total megabyte-milliseconds taken by all reduce tasks=4286464
        Map-Reduce Framework
                Map input records=111396
                Map output records=814671
                Map output bytes=7841067
                Map output materialized bytes=1192234
                Input split bytes=112
                Combine input records=814671
                Combine output records=80364
                Reduce input groups=80364
                Reduce shuffle bytes=1192234
                Reduce input records=80364
                Reduce output records=80364
                Spilled Records=160728
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=165
                CPU time spent (ms)=3960
                Physical memory (bytes) snapshot=492679168
                Virtual memory (bytes) snapshot=3764699136
                Total committed heap usage (bytes)=305659904
        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=4583798
        File Output Format Counters
                Bytes Written=878840
```



It uses 21s to run the program.

Compared to part c output, it is 1 second faster than the c part output. (21s vs 22s)