I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website

http://www.cuhk.edu.hk/policy/academichonesty/

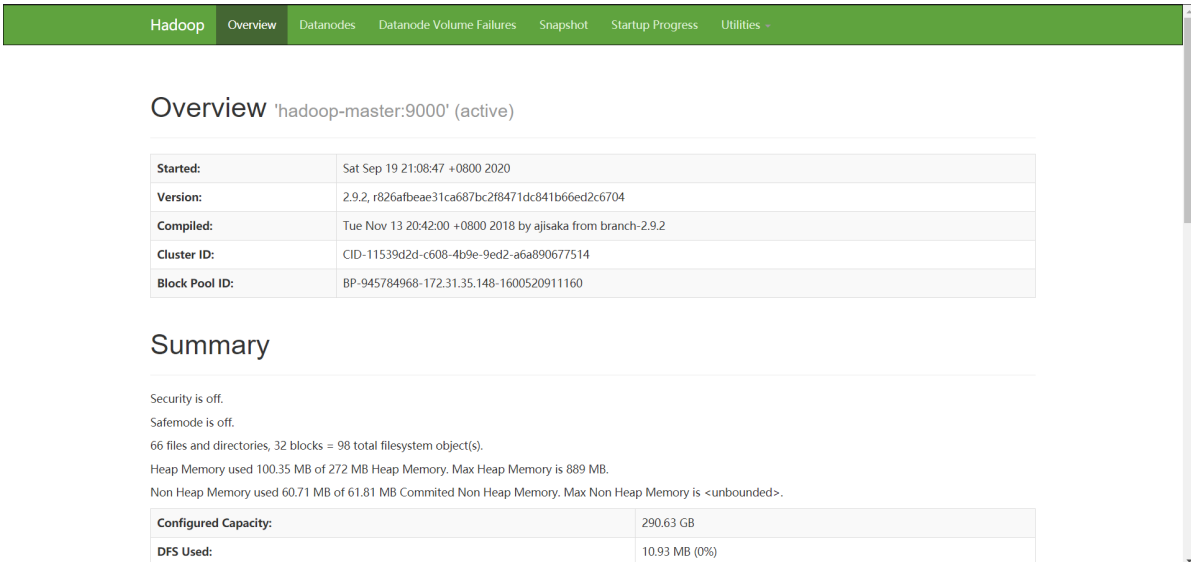
Signed (Student Wu Yan) Date: 2020.09.25

Name WU Yan SID 1155148594

a. Single-node Hadoop Setup

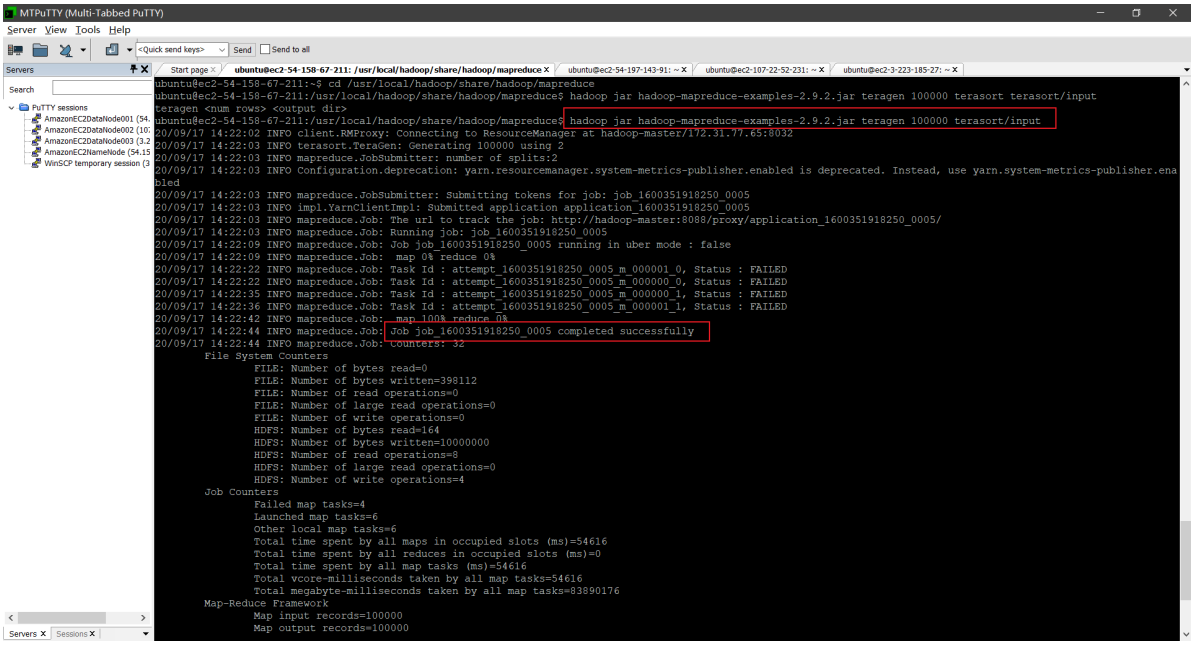
i. The namenode web page(attached in the end)

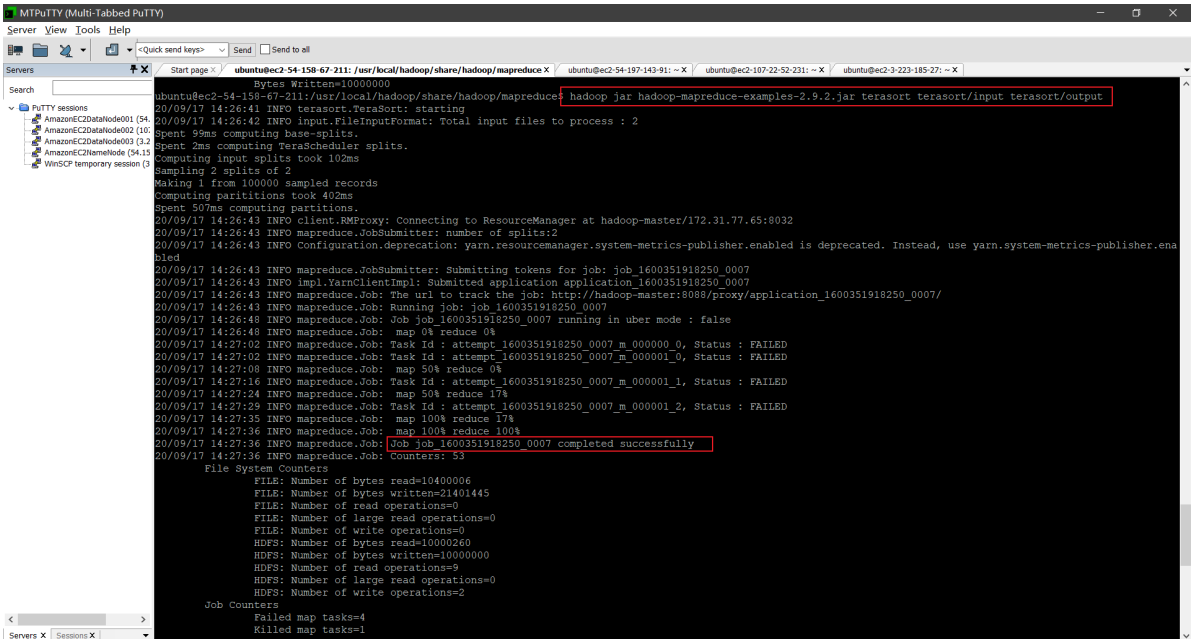
You can also check the print page at the end of this PDF file

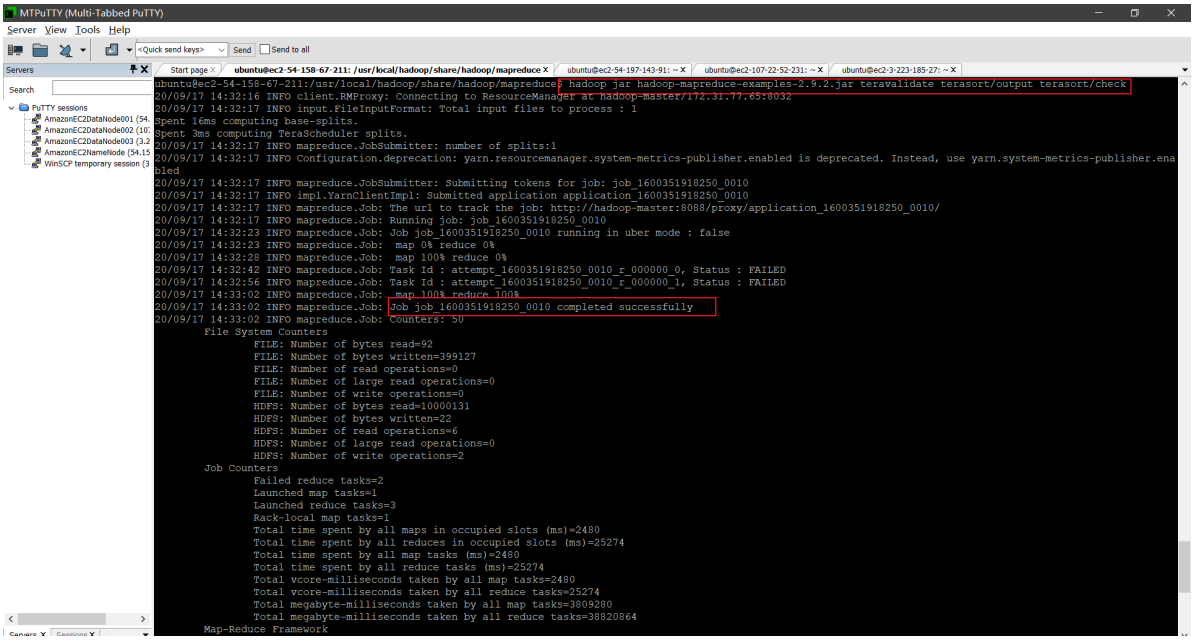


ii. The Terasort example running process

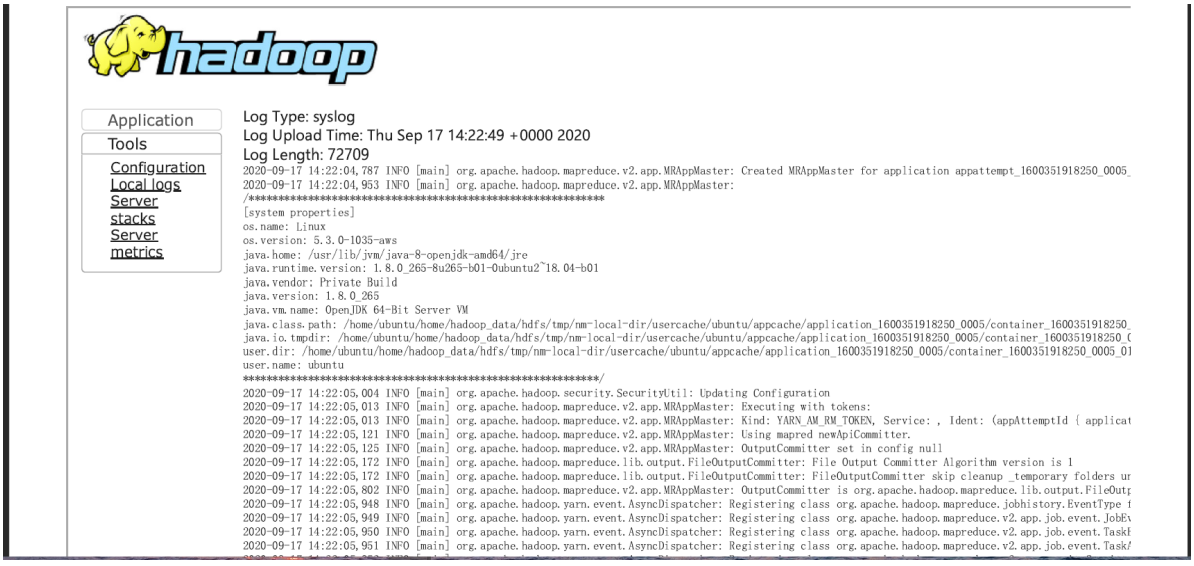
1. Run the commands in Putty

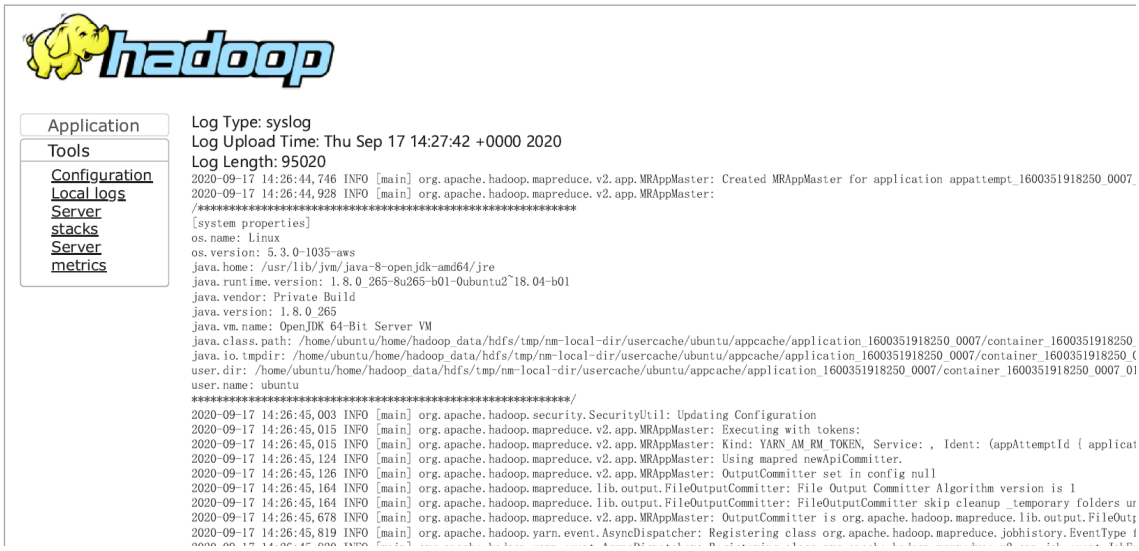


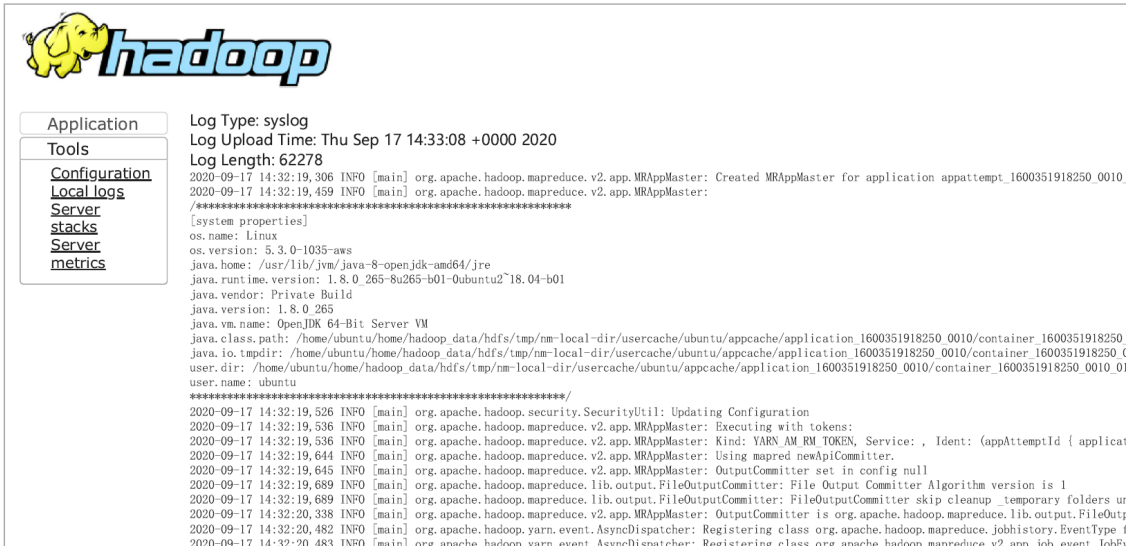




2. Check the logs in jobhistory web page

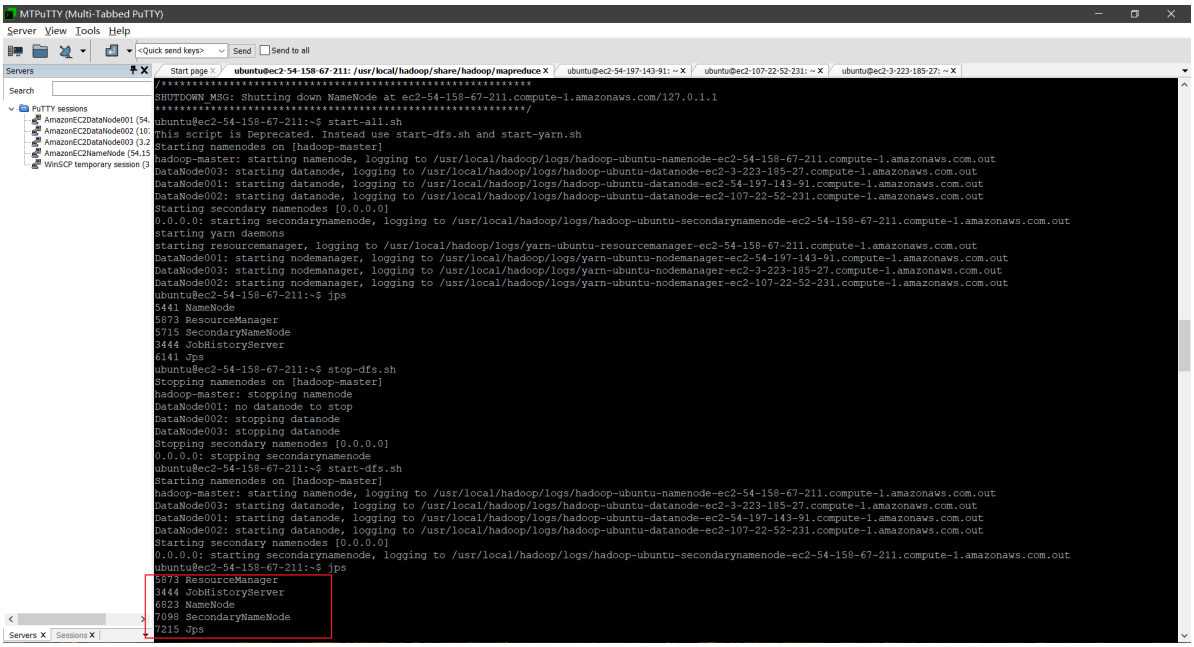


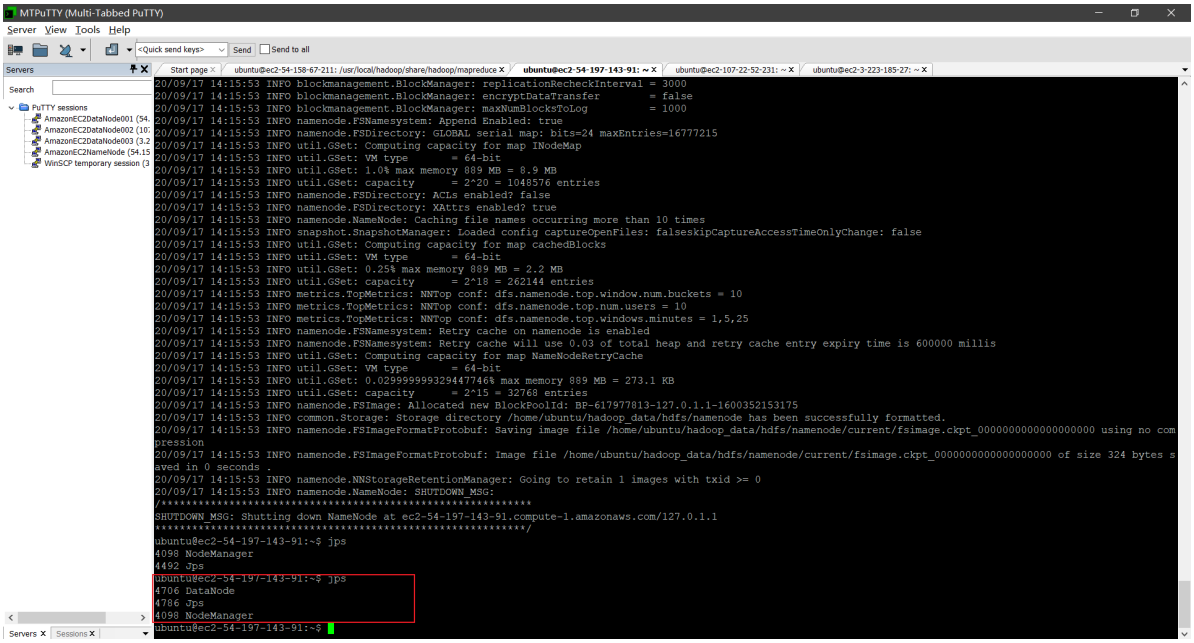


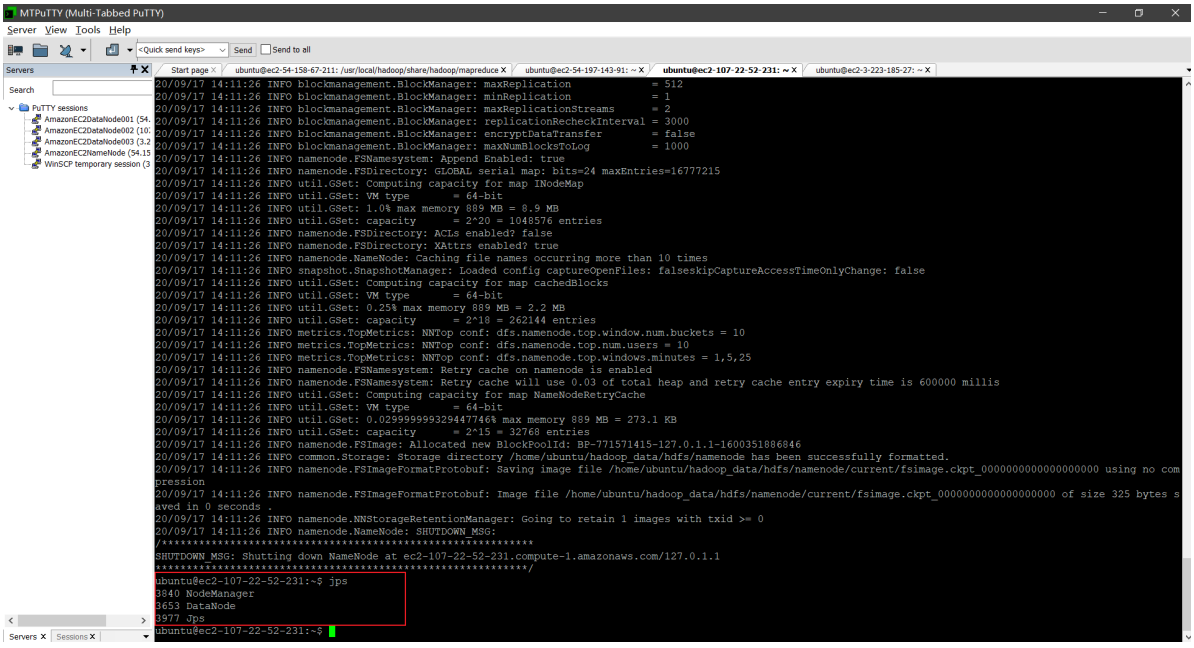


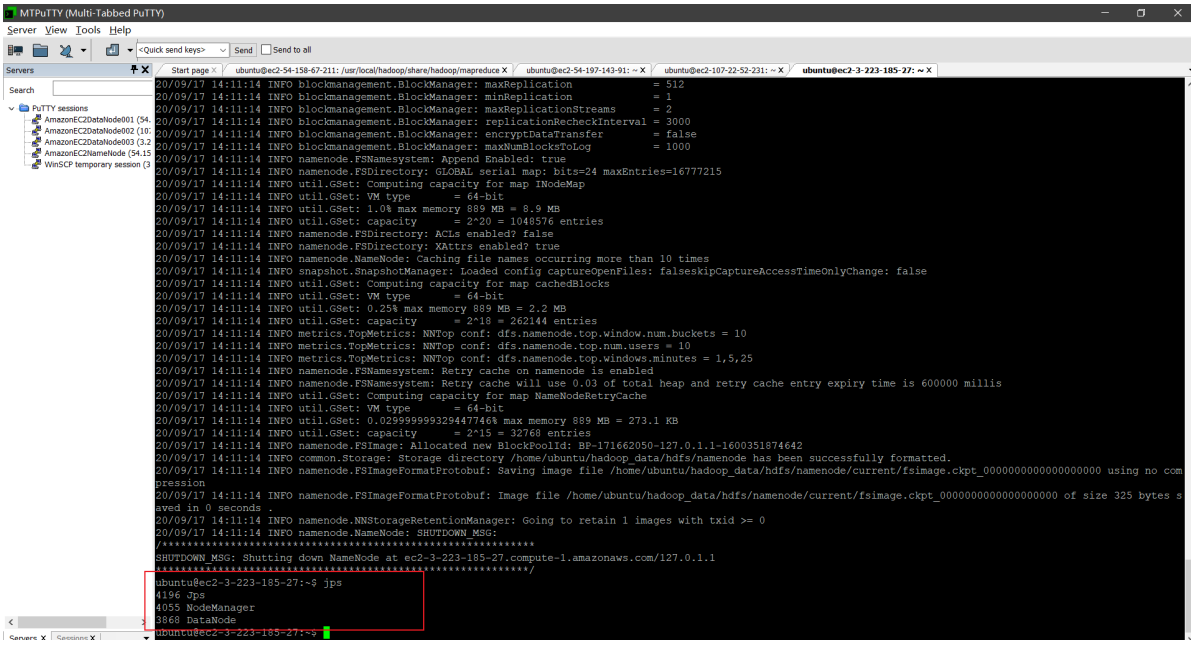
b. Multi-node Hadoop Cluster Setup

i. "jps" command on 4 VMs



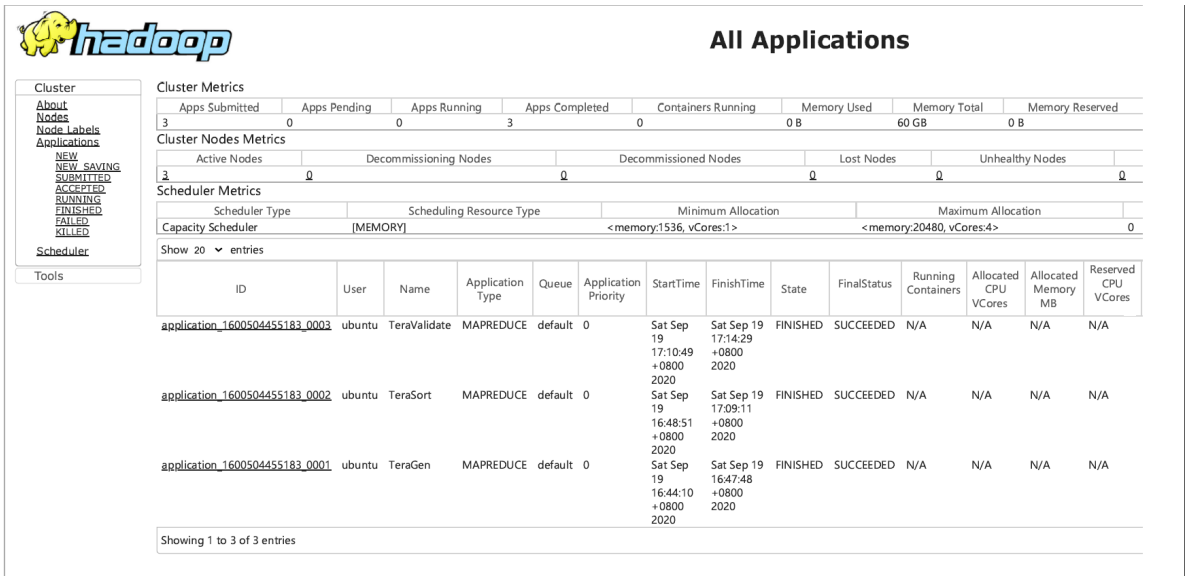






ii. Run Terasort example with 2GB data

check the time comsumed(about 27 minutes)



All Applications

Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved
3	0	0	3	0	0 B	60 GB	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	
3	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	
Capacity Scheduler	[MEMORY]	<memory:1536, vCores:1>	<memory:20480, vCores:4>	0

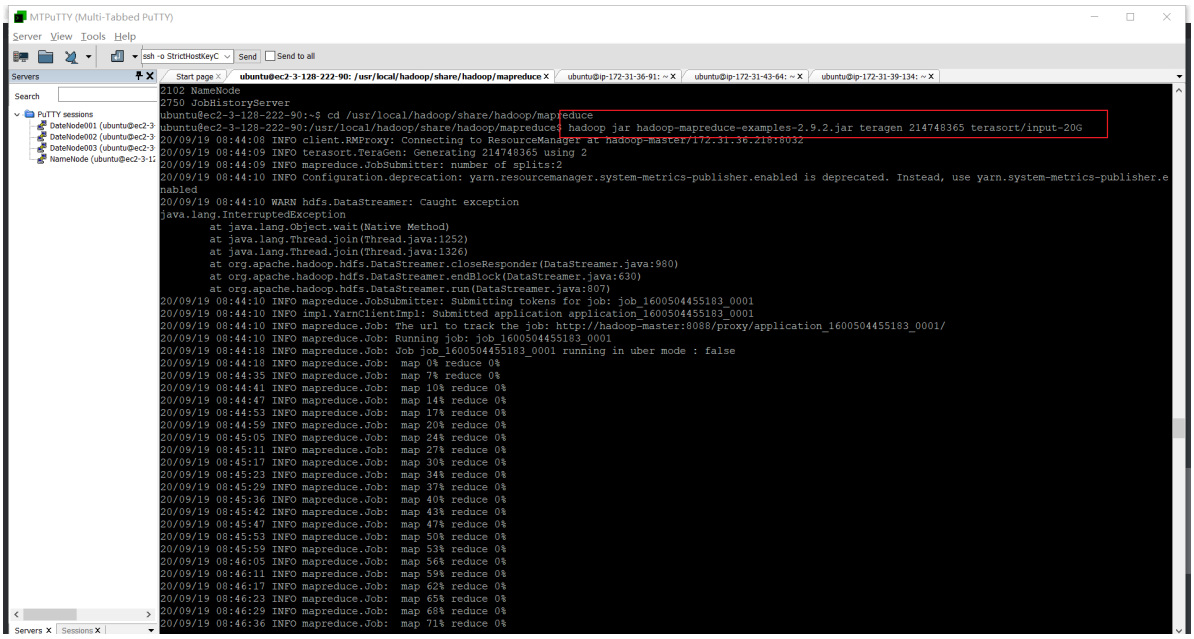
Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores
application_1600504455183_0003	ubuntu	TeraValidate	MAPREDUCE	default	0	Sat Sep 19 17:10:49 +0800 2020	Sat Sep 19 17:14:29 +0800 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A
application_1600504455183_0002	ubuntu	TeraSort	MAPREDUCE	default	0	Sat Sep 19 16:48:51 +0800 2020	Sat Sep 19 17:09:11 +0800 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A
application_1600504455183_0001	ubuntu	TeraGen	MAPREDUCE	default	0	Sat Sep 19 16:44:10 +0800 2020	Sat Sep 19 16:47:48 +0800 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A

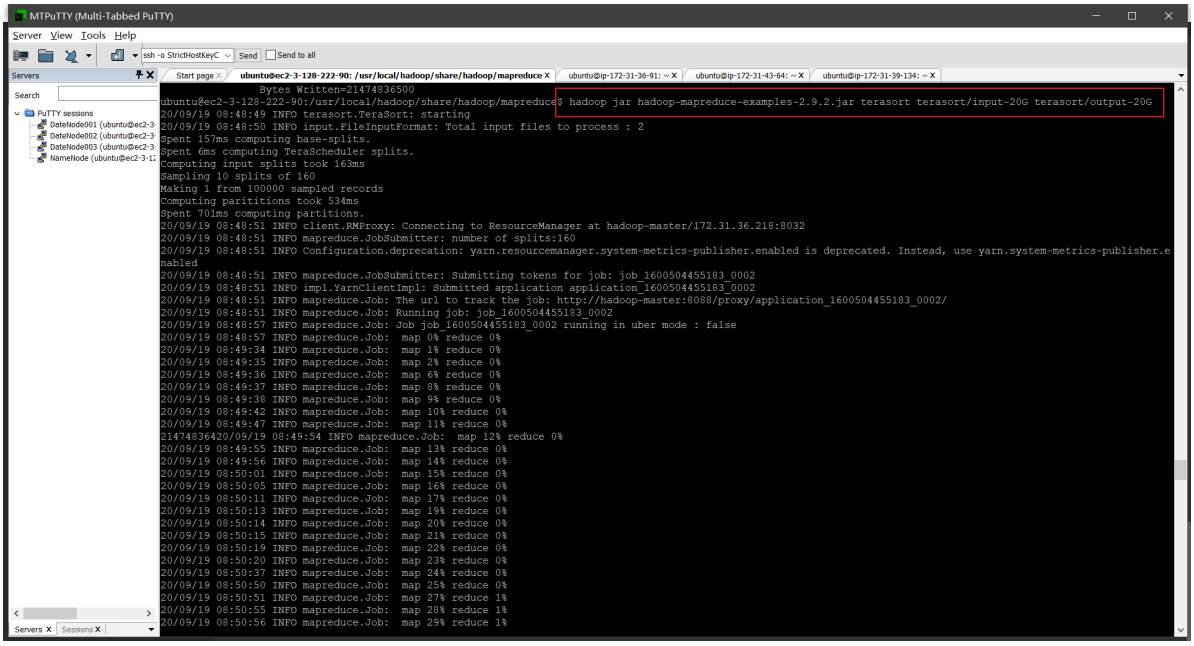
Showing 1 to 3 of 3 entries

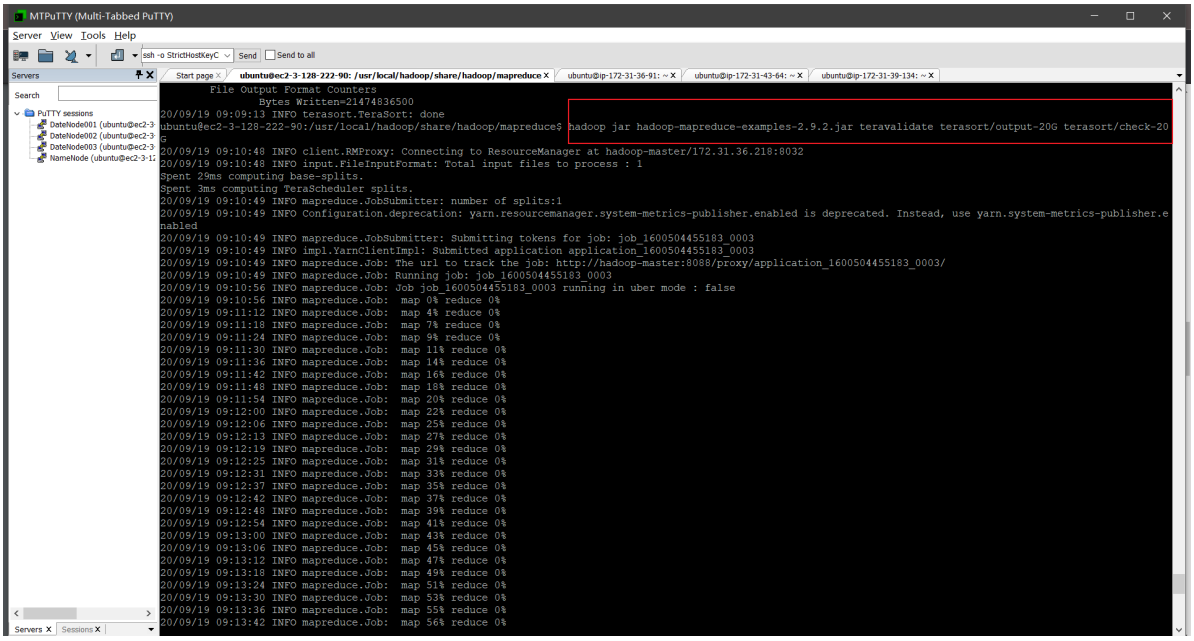
iii. Run Terasort example with 20 GB data

1. Run commands in Putty(the same as above)

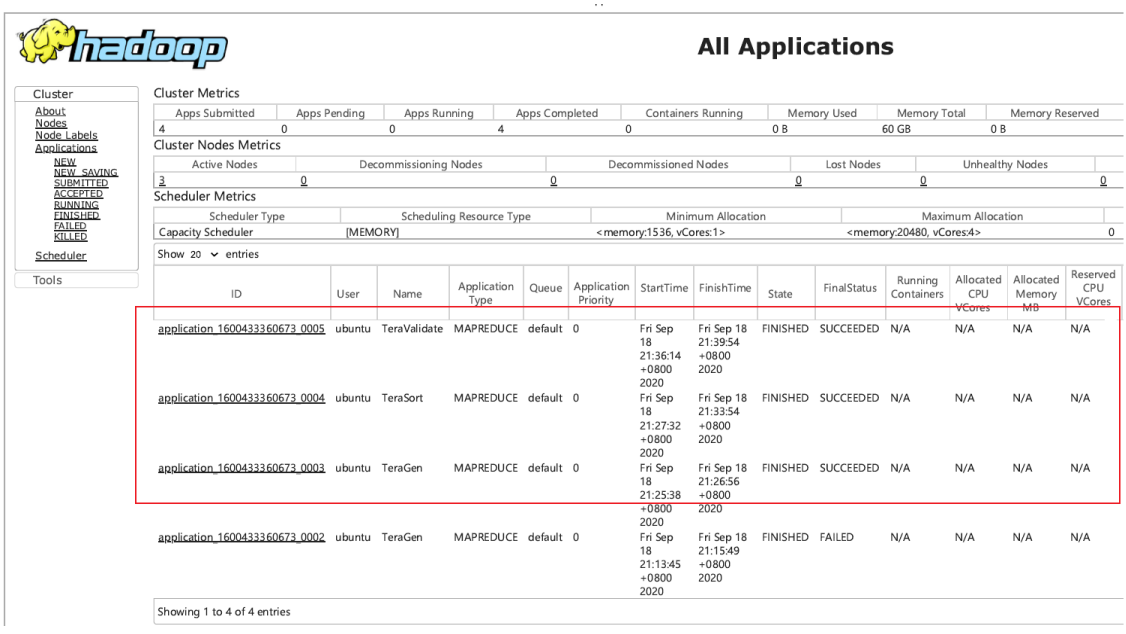


```
2102 NameNode
2750 JobHistoryServer
ubuntu@ec2-3-128-222-90:~$ cd /usr/local/hadoop/share/hadoop/mapreduce
ubuntu@ec2-3-128-222-90:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.9.2.jar teragen 214748365 terasort/input-20G
20/09/19 08:44:08 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.31.36.218:8032
20/09/19 08:44:09 INFO terasort.TeraGen: Generating 214748365 using 2
20/09/19 08:44:09 INFO mapreduce.JobSubmitter: number of splits:2
20/09/19 08:44:10 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.e
nabled
20/09/19 08:44:10 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:980)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:630)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:807)
20/09/19 08:44:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1600504455183_0001
20/09/19 08:44:10 INFO impl.YarnClientImpl: Submitted application application_1600504455183_0001
20/09/19 08:44:10 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1600504455183_0001/
20/09/19 08:44:10 INFO mapreduce.Job: Running job: job_1600504455183_0001
20/09/19 08:44:18 INFO mapreduce.Job: Job job_1600504455183_0001 running in uber mode : false
20/09/19 08:44:18 INFO mapreduce.Job:  map 0% reduce 0%
20/09/19 08:44:35 INFO mapreduce.Job:  map 7% reduce 0%
20/09/19 08:44:41 INFO mapreduce.Job:  map 10% reduce 0%
20/09/19 08:44:47 INFO mapreduce.Job:  map 14% reduce 0%
20/09/19 08:44:53 INFO mapreduce.Job:  map 17% reduce 0%
20/09/19 08:44:59 INFO mapreduce.Job:  map 20% reduce 0%
20/09/19 08:45:05 INFO mapreduce.Job:  map 24% reduce 0%
20/09/19 08:45:11 INFO mapreduce.Job:  map 27% reduce 0%
20/09/19 08:45:17 INFO mapreduce.Job:  map 30% reduce 0%
20/09/19 08:45:23 INFO mapreduce.Job:  map 34% reduce 0%
20/09/19 08:45:29 INFO mapreduce.Job:  map 37% reduce 0%
20/09/19 08:45:36 INFO mapreduce.Job:  map 40% reduce 0%
20/09/19 08:45:42 INFO mapreduce.Job:  map 43% reduce 0%
20/09/19 08:45:47 INFO mapreduce.Job:  map 47% reduce 0%
20/09/19 08:45:53 INFO mapreduce.Job:  map 50% reduce 0%
20/09/19 08:45:59 INFO mapreduce.Job:  map 53% reduce 0%
20/09/19 08:46:05 INFO mapreduce.Job:  map 56% reduce 0%
20/09/19 08:46:11 INFO mapreduce.Job:  map 59% reduce 0%
20/09/19 08:46:17 INFO mapreduce.Job:  map 62% reduce 0%
20/09/19 08:46:23 INFO mapreduce.Job:  map 65% reduce 0%
20/09/19 08:46:29 INFO mapreduce.Job:  map 68% reduce 0%
20/09/19 08:46:36 INFO mapreduce.Job:  map 71% reduce 0%
```





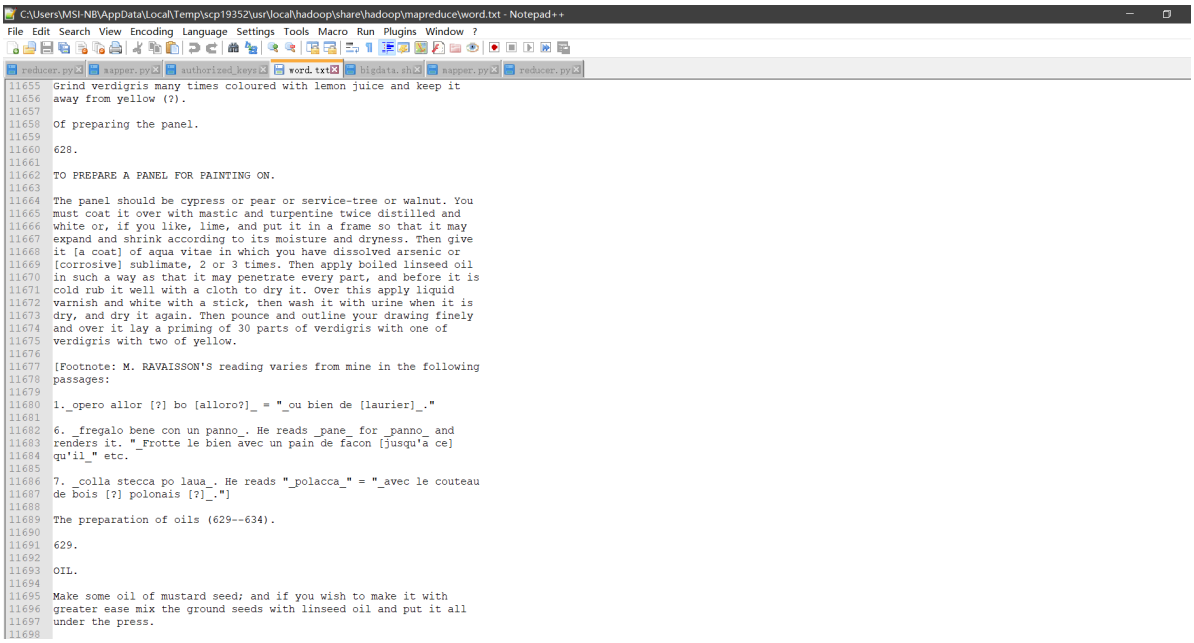
2. Check the time consumed(about 10 minutes)



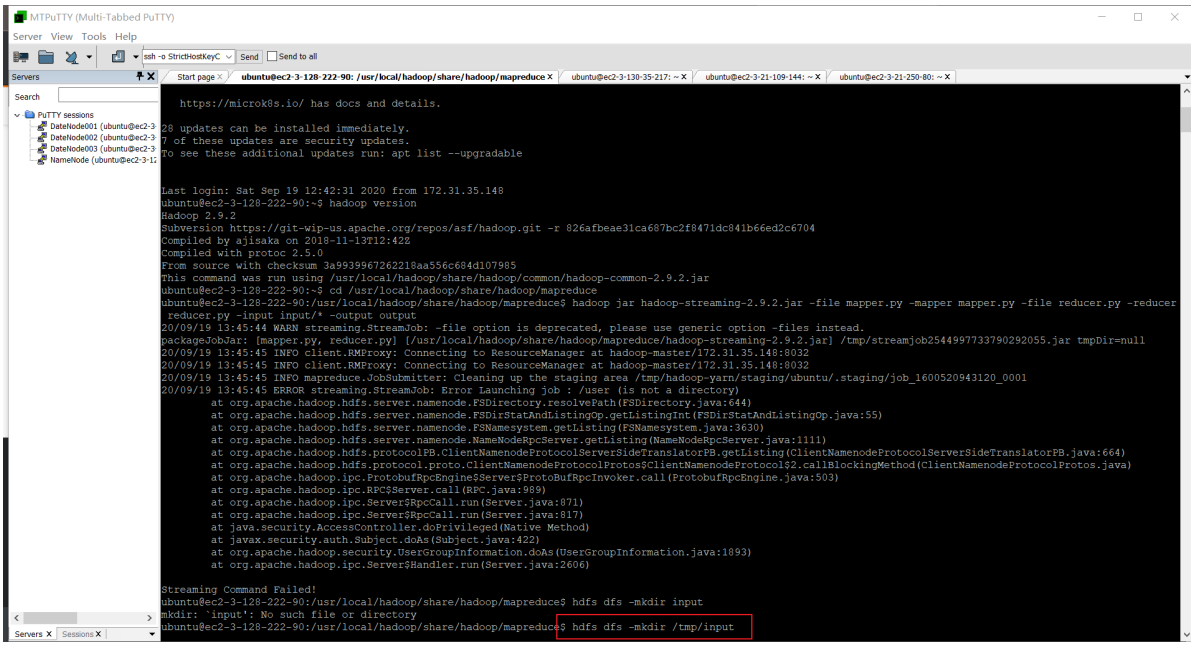
c. Running the Python Code on Hadoop

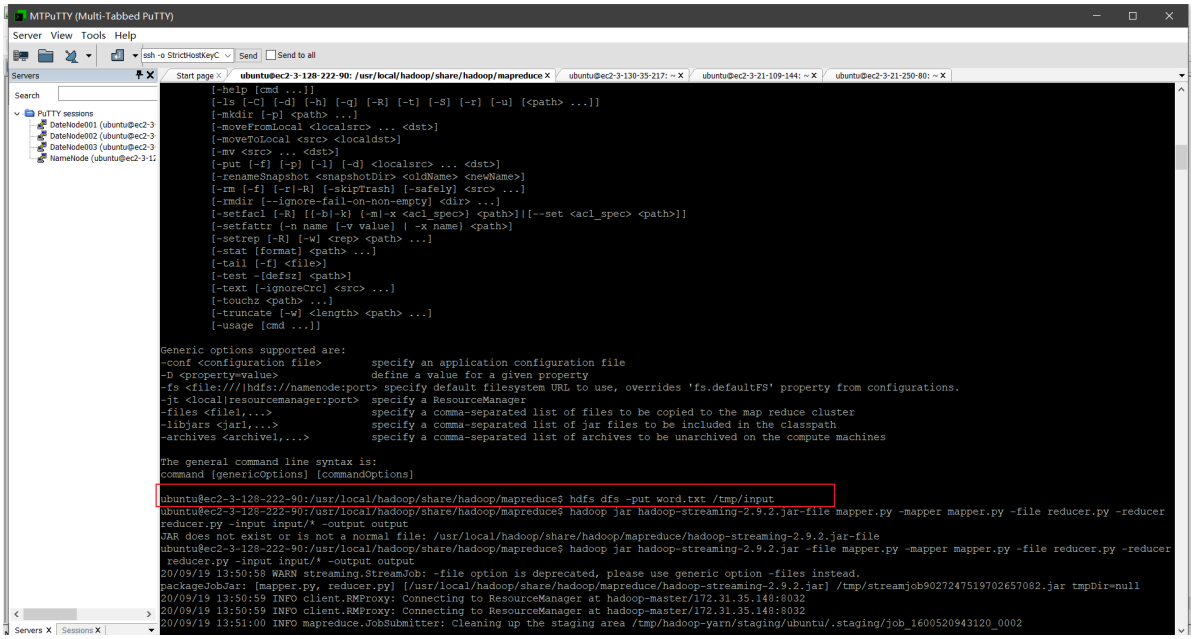
i. Necessary modifications

1. Download the Lemon Juice's artical



2. Create a directory on HDFS



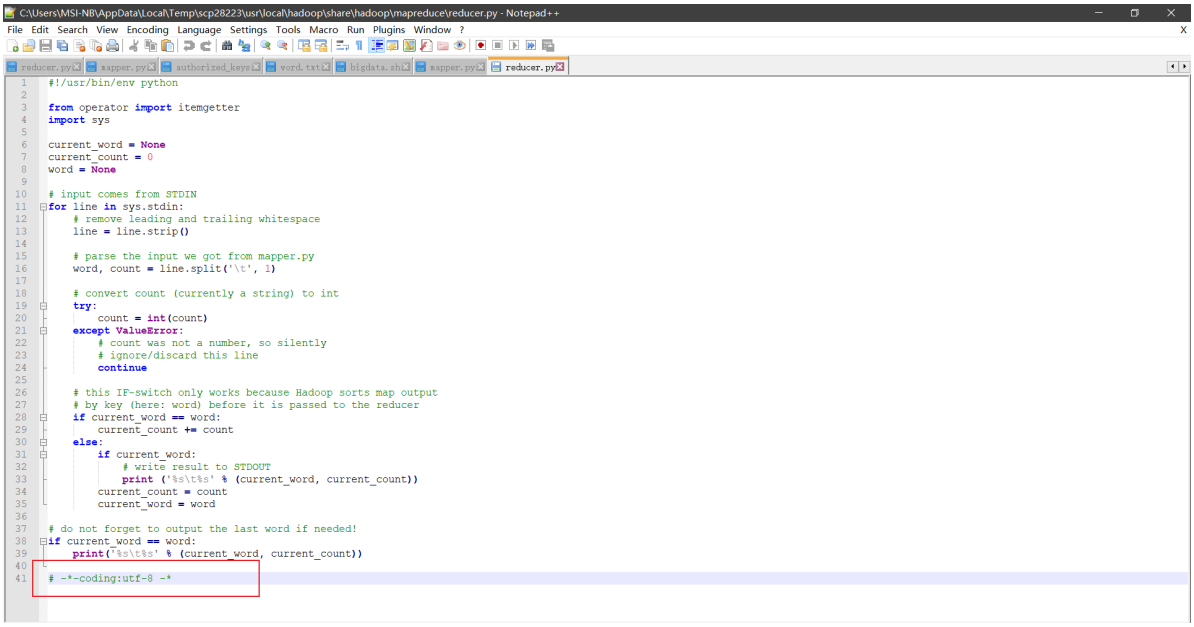


3. Modify the .py files' authorty and format

```
$ chmod +x mapper.py

$ chmod +x reducer.py
```

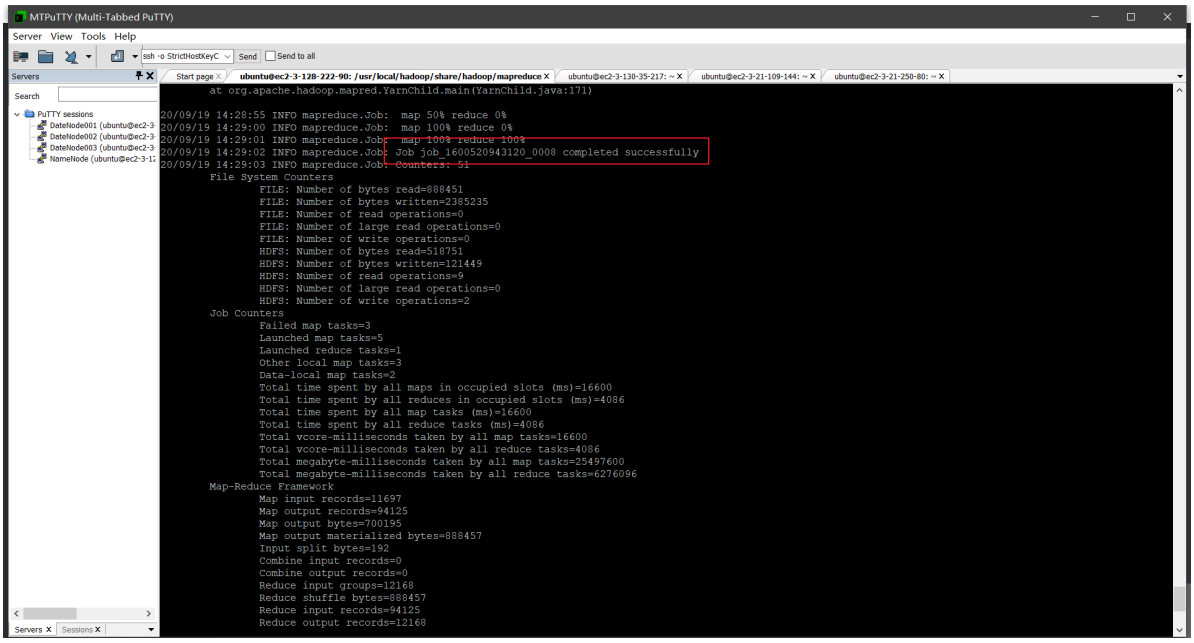
`add on the second line :` *# --coding:utf-8 -*``



4. Run the following codes

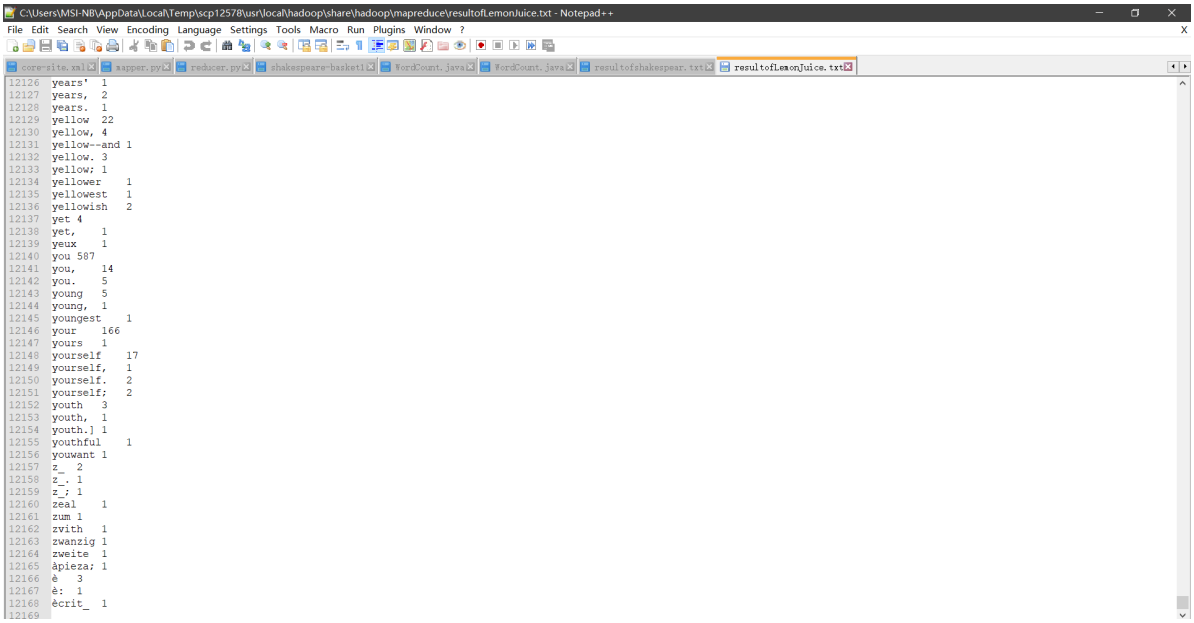
```
$ hadoop jar hadoop-streaming-2.9.2.jar -file
/usr/local/hadoop/share/hadoop/mapreduce/mapper.py -mapper
/usr/local/hadoop/share/hadoop/mapreduce/mapper.py -file
/usr/local/hadoop/share/hadoop/mapreduce/reducer.py -reducer
/usr/local/hadoop/share/hadoop/mapreduce/reducer.py -input /tmp/input/* -
output output4
```

5. Check the result



6. Cat the file to VM and check the result

```
$ hadoop fs -cat /user/ubuntu/output4/part-00000 > resultofLemonJuice.txt
```



d. Compiling the Java WordCount program for MapReduce

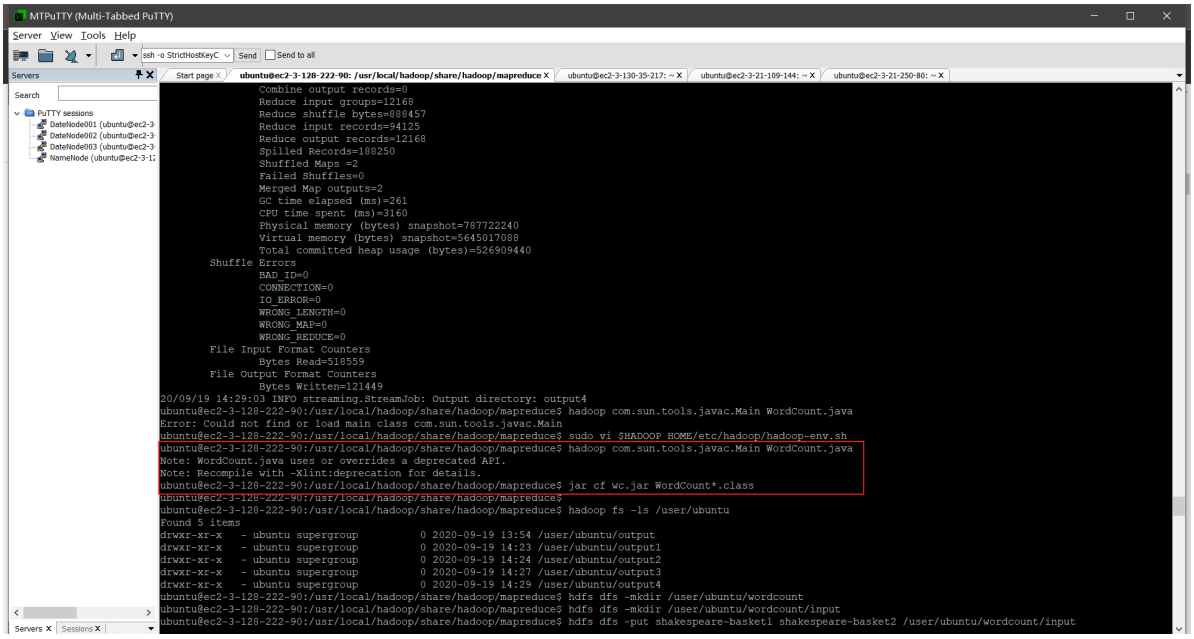
i. Make a preparation for running this job

1. Enter these path variables into $HADOOP_HOME/etc/hadoop/hadoop-env.sh

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

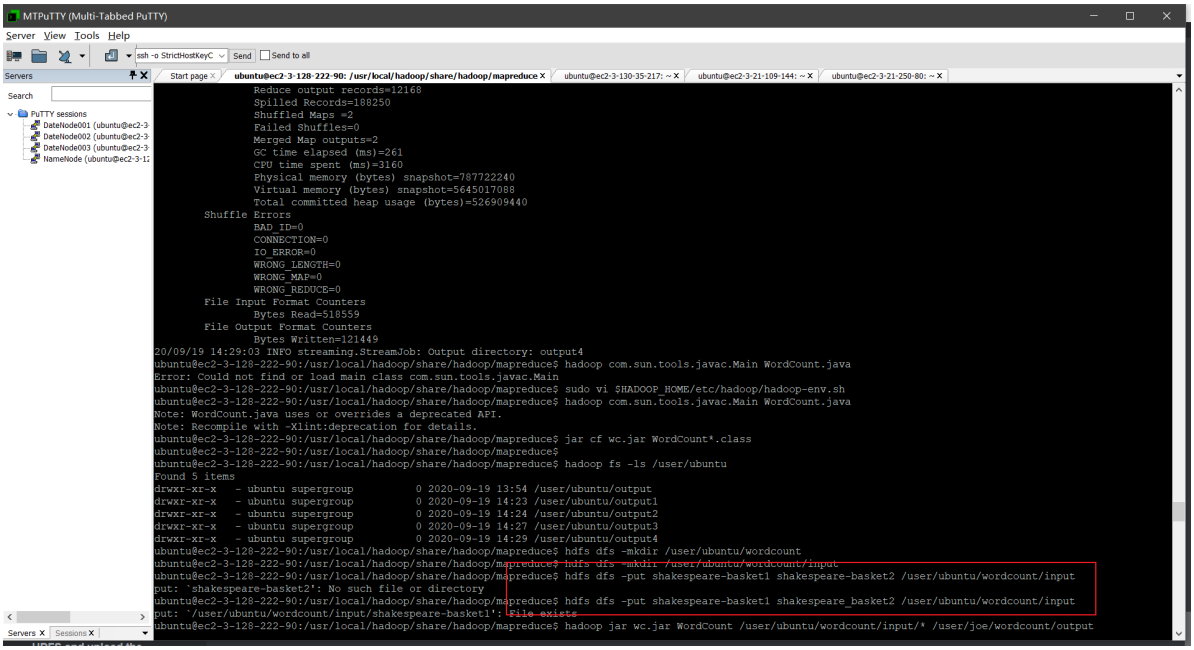
2. Compile `WordCount.java` and create a jar

```
$ hadoop com.sun.tools.javac.Main WordCount.java
$ jar cf wc.jar WordCount*.class
```



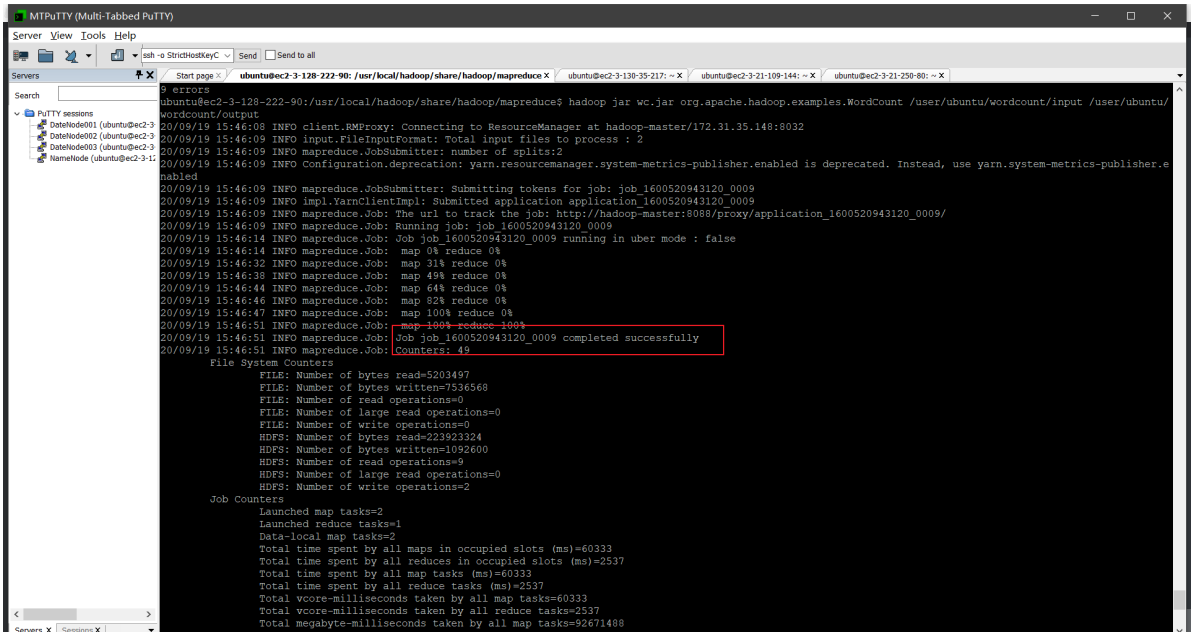
3. Create input directory in HDFS and upload the shakespear_basket1/2 onto it

```
$ hdfs dfs -put shakespeare-basket1 shakespeare_basket2
/user/ubuntu/wordcount/input
```



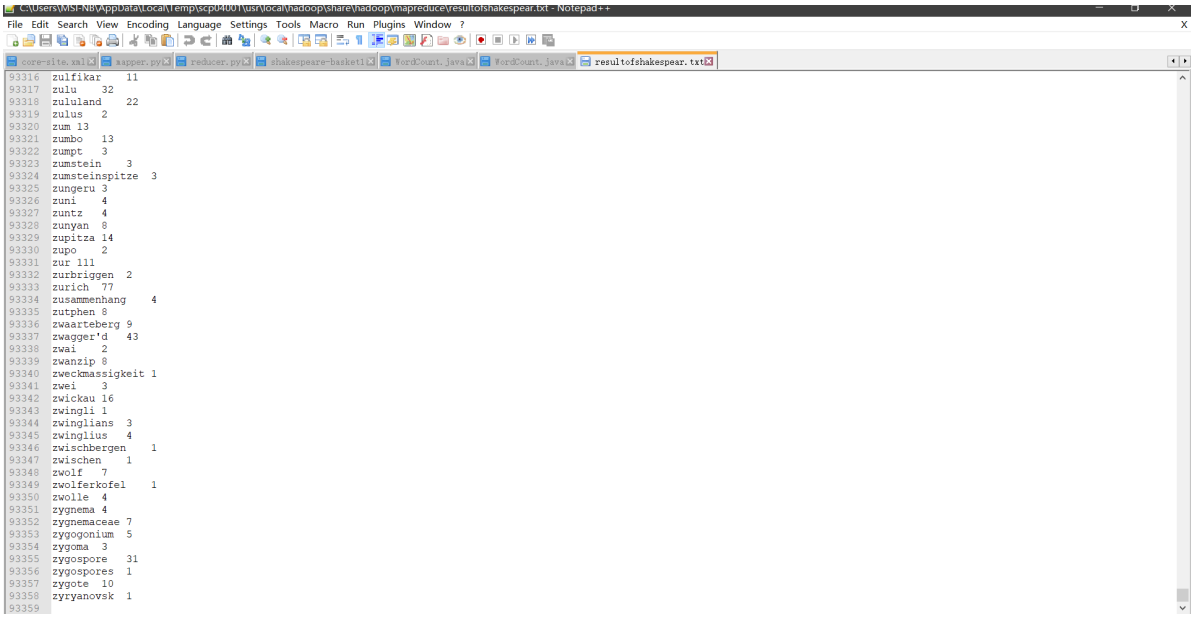
4. Run the application

```
$ hadoop jar wc.jar org.apache.hadoop.examples.WordCount
/user/ubuntu/wordcount/input /user/ubuntu/wordcount/output
```



5. Duplicate the output into a .txt file

```
$ hadoop fs -cat /user/ubuntu/wordcount/output/part-r-00000 >
resultofshakespear.txt
```



6. Compare the running time with task c(almost the same)

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved
7	0	0	7	0	0 B	80 GB	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	
4	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	
Capacity Scheduler	[MEMORY]	<memory:1536, vCores:1>	<memory:20480, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserv CPU VCore
application_1600520943120_0009	ubuntu	word count	MAPREDUCE	default	0	Sat Sep 19 23:46:09 +0800 2020	Sat Sep 19 23:46:50 +0800 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A
application_1600520943120_0008	ubuntu	streamjob4069683088770483485.jar	MAPREDUCE	default	0	Sat Sep 19 22:28:39 +0800 2020	Sat Sep 19 22:29:01 +0800 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A
application_1600520943120_0007	ubuntu	streamjob3317272319078999419.jar	MAPREDUCE	default	0	Sat Sep 19 22:26:43 +0800	Sat Sep 19 22:27:08 +0800 2020	FINISHED	FAILED	N/A	N/A	N/A	N/A

Overview 'hadoop-master:9000' (active)

Started:	Thu Sep 17 22:16:30 +0800 2020
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 20:42:00 +0800 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-d9947b4f-c252-4b78-8245-385ad99223e6
Block Pool ID:	BP-1931485071-127.0.1.1-1600351832416

Summary

Security is off.

Safemode is off.

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

Heap Memory used 76.7 MB of 185.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 38.43 MB of 39.5 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	23.08 GB
DFS Used:	72 KB (0%)
Non DFS Used:	10.04 GB
DFS Remaining:	12.99 GB (56.27%)
Block Pool Used:	72 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	3 (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 17 22:16:30 +0800 2020
Last Checkpoint Time	Thu Sep 17 22:12:53 +0800 2020

NameNode Journal Status

Current transaction ID: 4

Journal Manager	State
FileJournalManager(root=/home/ubuntu/hadoop_data/hdfs/namenode)	EditLogFileOutputStream(/home/ubuntu/hadoop_data/hdfs/namenode/current/edits_inprc

NameNode Storage

Storage Directory	Type	State
/home/ubuntu/hadoop_data/hdfs/namenode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	23.08 GB	72 KB (0%)	12.99 GB (56.27%)	72 KB	3

Hadoop, 2018.