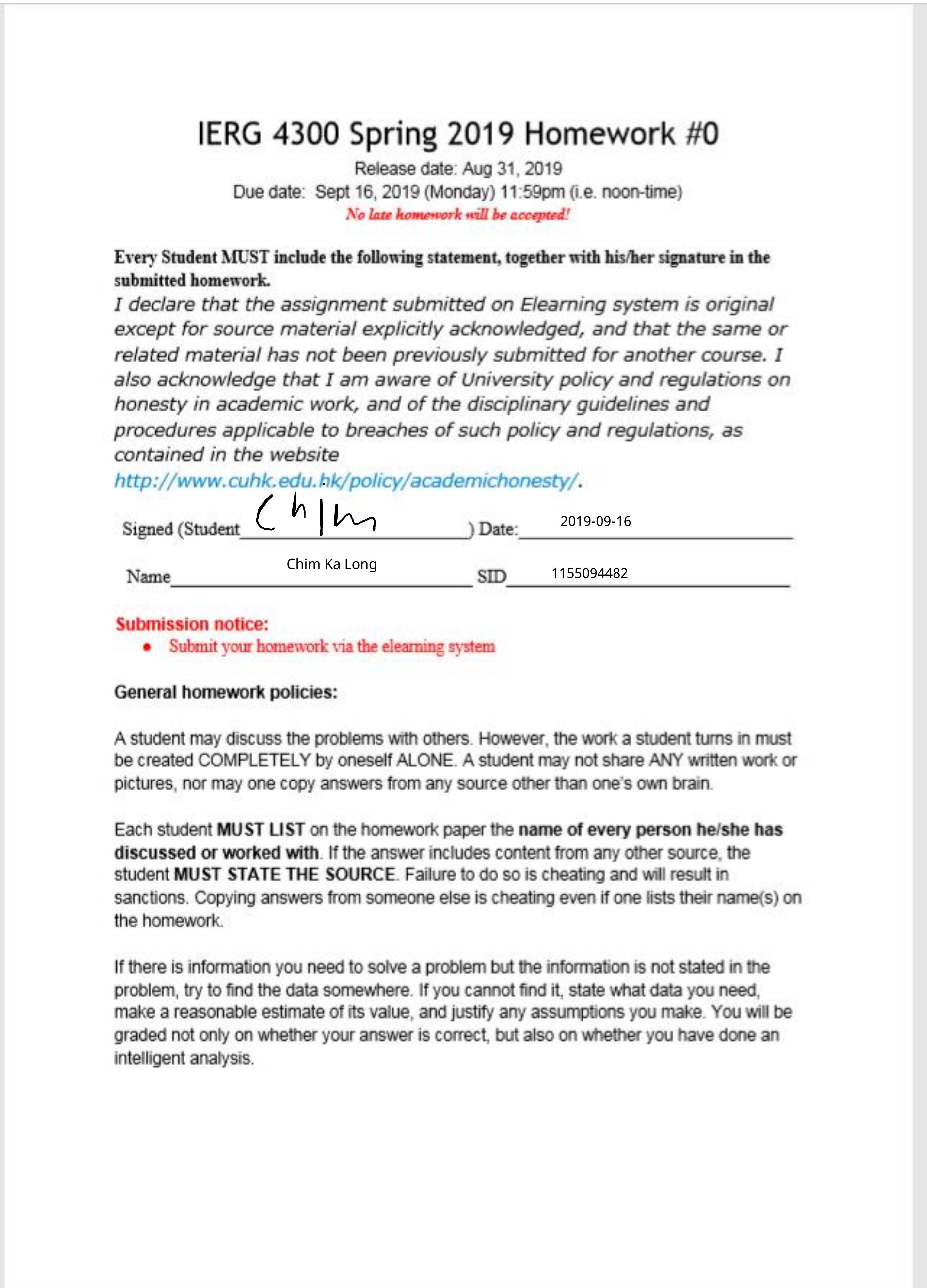


IERG 4300 Spring 2019 Homework #0

Release date: Aug 31, 2019

Due date: Sept 16, 2019 (Monday) 11:59pm (i.e. noon-time)

No late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

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 Chim) Date: 2019-09-16

Name Chim Ka Long SID 1155094482

Submission notice:

- Submit your homework via the elearning system

General homework policies:

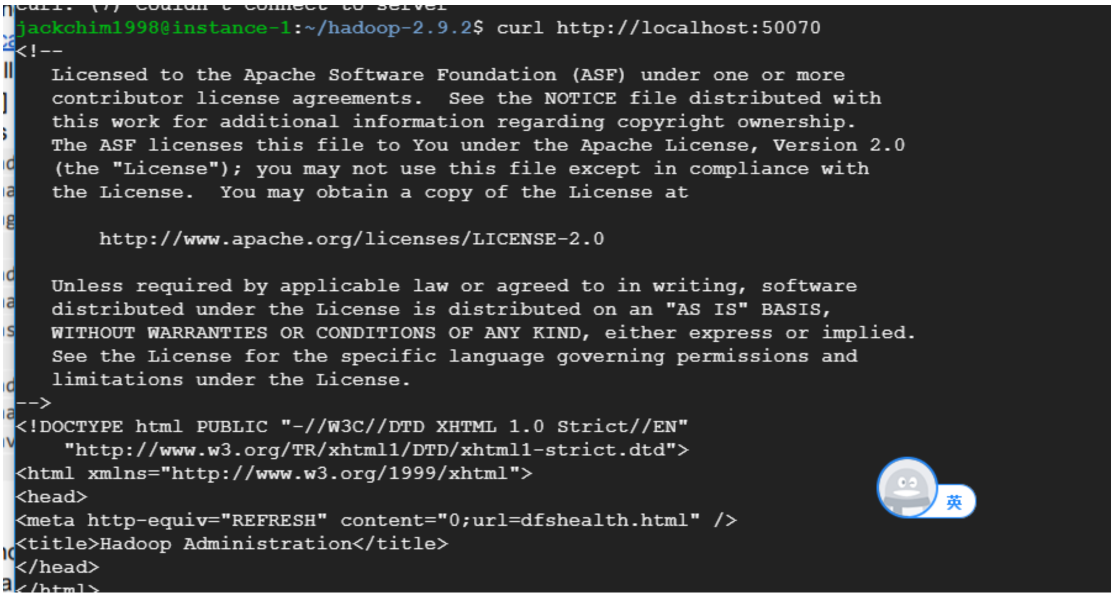
A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

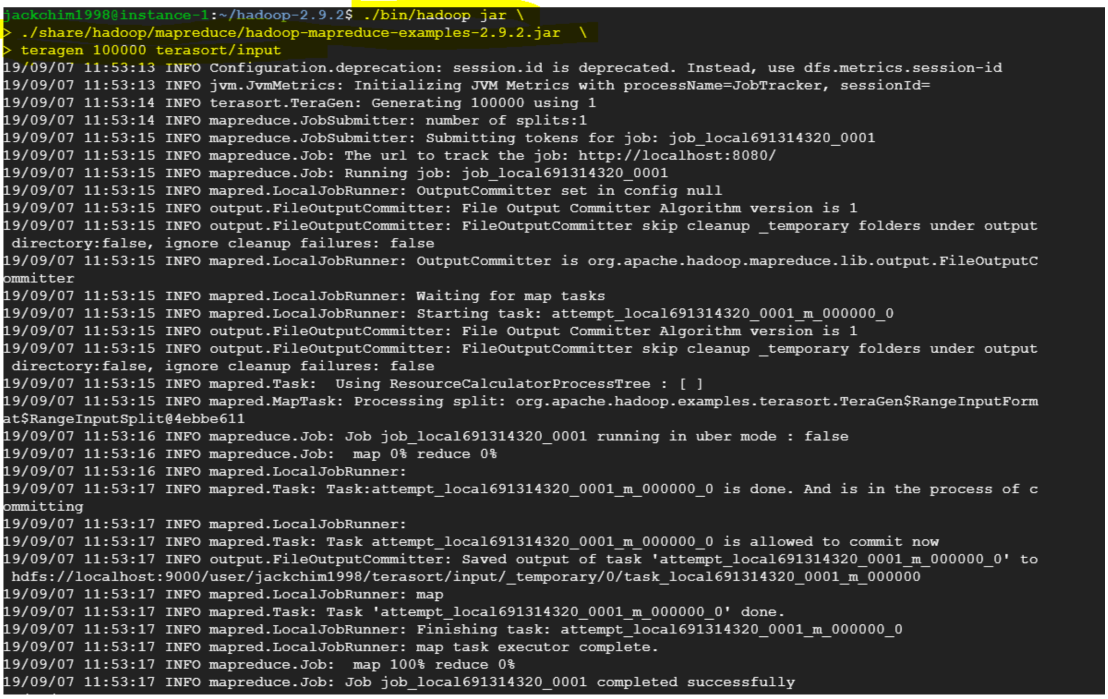
[20 marks] Single-node Hadoop Setup

After installation, we browse the port 50070 of localhost successfully.

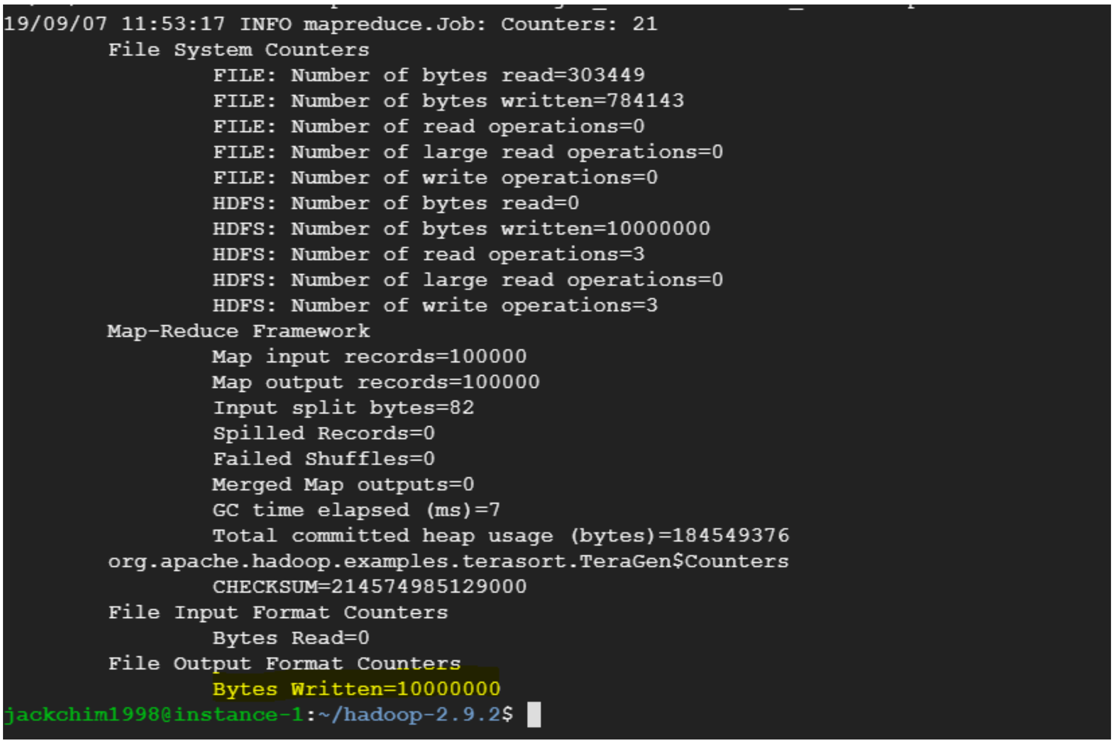


```
curl: (7) Couldn't connect to server
jackchim1998@instance-1:~/hadoop-2.9.2$ curl http://localhost:50070
<!--
   Licensed to the Apache Software Foundation (ASF) under one or more
   contributor license agreements.  See the NOTICE file distributed with
   this work for additional information regarding copyright ownership.
   The ASF licenses this file to You under the Apache License, Version 2.0
   (the "License"); you may not use this file except in compliance with
   the License.  You may obtain a copy of the License at

       http://www.apache.org/licenses/LICENSE-2.0

   Unless required by applicable law or agreed to in writing, software
   distributed under the License is distributed on an "AS IS" BASIS,
   WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
   See the License for the specific language governing permissions and
   limitations under the License.
-->
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="REFRESH" content="0;url=dfshealth.html" />
<title>Hadoop Administration</title>
</head>
</html>
```

We try to run terasort example. Firstly, generate input file of random number.

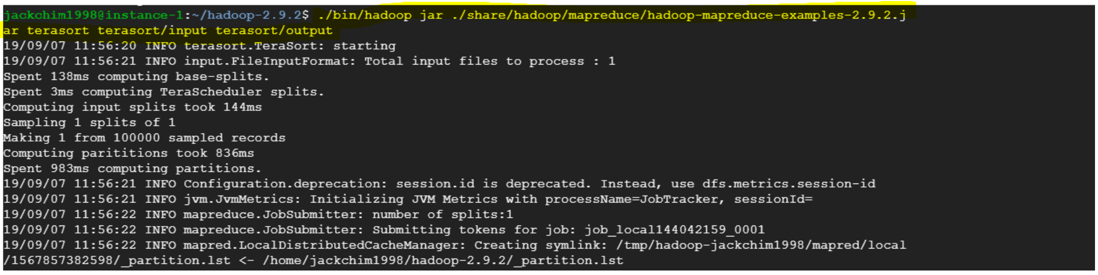


```
jackchim1998@instance-1:~/hadoop-2.9.2$ ./bin/hadoop jar \
> ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar  \
> teragen 100000 terasort/input
19/09/07 11:53:13 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/09/07 11:53:13 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/09/07 11:53:14 INFO terasort.TeraGen: Generating 100000 using 1
19/09/07 11:53:14 INFO mapreduce.JobSubmitter: number of splits:1
19/09/07 11:53:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local691314320_0001
19/09/07 11:53:15 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
19/09/07 11:53:15 INFO mapreduce.Job: Running job: job_local691314320_0001
19/09/07 11:53:15 INFO mapred.LocalJobRunner: OutputCommitter set in config null
19/09/07 11:53:15 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/09/07 11:53:15 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output
 directory:false, ignore cleanup failures: false
19/09/07 11:53:15 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputC
ommitter
19/09/07 11:53:15 INFO mapred.LocalJobRunner: Waiting for map tasks
19/09/07 11:53:15 INFO mapred.LocalJobRunner: Starting task: attempt_local691314320_0001_m_000000_0
19/09/07 11:53:15 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/09/07 11:53:15 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output
 directory:false, ignore cleanup failures: false
19/09/07 11:53:15 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
19/09/07 11:53:15 INFO mapred.MapTask: Processing split: org.apache.hadoop.examples.terasort.TeraGen$RangeInputForm
at$RangeInputSplit@4ebbe611
19/09/07 11:53:16 INFO mapreduce.Job: Job job_local691314320_0001 running in uber mode : false
19/09/07 11:53:16 INFO mapreduce.Job:  map 0% reduce 0%
19/09/07 11:53:16 INFO mapred.LocalJobRunner:
19/09/07 11:53:17 INFO mapred.Task: Task:attempt_local691314320_0001_m_000000_0 is done. And is in the process of c
ommitting
19/09/07 11:53:17 INFO mapred.LocalJobRunner:
19/09/07 11:53:17 INFO mapred.Task: Task attempt_local691314320_0001_m_000000_0 is allowed to commit now
19/09/07 11:53:17 INFO output.FileOutputCommitter: Saved output of task 'attempt_local691314320_0001_m_000000_0' to
 hdfs://localhost:9000/user/jackchim1998/terasort/input/_temporary/0/task_local691314320_0001_m_000000
19/09/07 11:53:17 INFO mapred.LocalJobRunner: map
19/09/07 11:53:17 INFO mapred.Task: Task 'attempt_local691314320_0001_m_000000_0' done.
19/09/07 11:53:17 INFO mapred.LocalJobRunner: Finishing task: attempt_local691314320_0001_m_000000_0
19/09/07 11:53:17 INFO mapred.LocalJobRunner: map task executor complete.
19/09/07 11:53:17 INFO mapreduce.Job:  map 100% reduce 0%
19/09/07 11:53:17 INFO mapreduce.Job: Job job_local691314320_0001 completed successfully
```

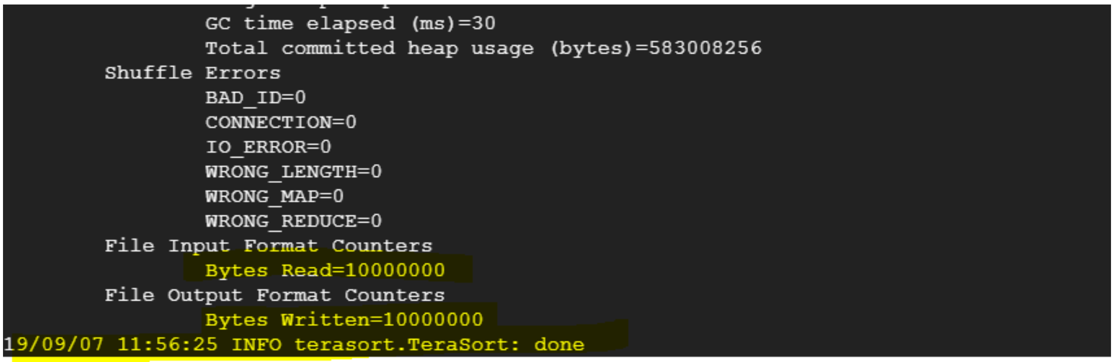


```
19/09/07 11:53:17 INFO mapreduce.Job: Counters: 21
        File System Counters
                FILE: Number of bytes read=303449
                FILE: Number of bytes written=784143
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=0
                HDFS: Number of bytes written=10000000
                HDFS: Number of read operations=3
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Input split bytes=82
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=7
                Total committed heap usage (bytes)=184549376
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=214574985129000
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=10000000
jackchim1998@instance-1:~/hadoop-2.9.2$
```

Then sorting it to output file.

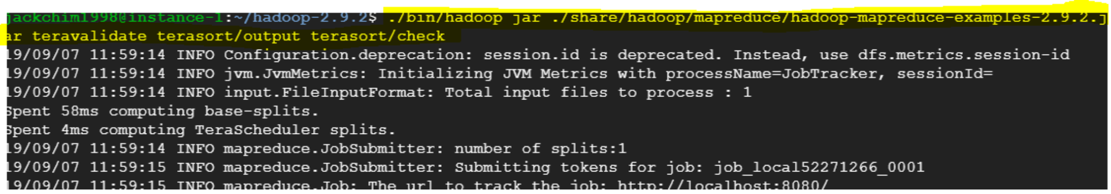


```
jackchim1998@instance-1:~/hadoop-2.9.2$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.j
ar terasort terasort/input terasort/output
19/09/07 11:56:20 INFO terasort.TeraSort: starting
19/09/07 11:56:21 INFO input.FileInputFormat: Total input files to process : 1
Spent 138ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
Computing input splits took 144ms
Sampling 1 splits of 1
Making 1 from 100000 sampled records
Computing paritions took 836ms
Spent 983ms computing partitions.
19/09/07 11:56:21 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/09/07 11:56:21 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/09/07 11:56:22 INFO mapreduce.JobSubmitter: number of splits:1
19/09/07 11:56:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local144042159_0001
19/09/07 11:56:22 INFO mapred.LocalDistributedCacheManager: Creating symlink: /tmp/hadoop-jackchim1998/mapred/local
/1567857382598/_partition.lst <- /home/jackchim1998/hadoop-2.9.2/_partition.lst
```

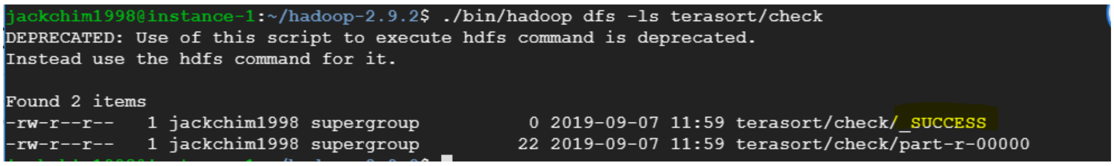


```
                GC time elapsed (ms)=30
                Total committed heap usage (bytes)=583008256
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=10000000
        File Output Format Counters
                Bytes Written=10000000
19/09/07 11:56:25 INFO terasort.TeraSort: done
```

Finally, run teravalidate to validate the sorting result.



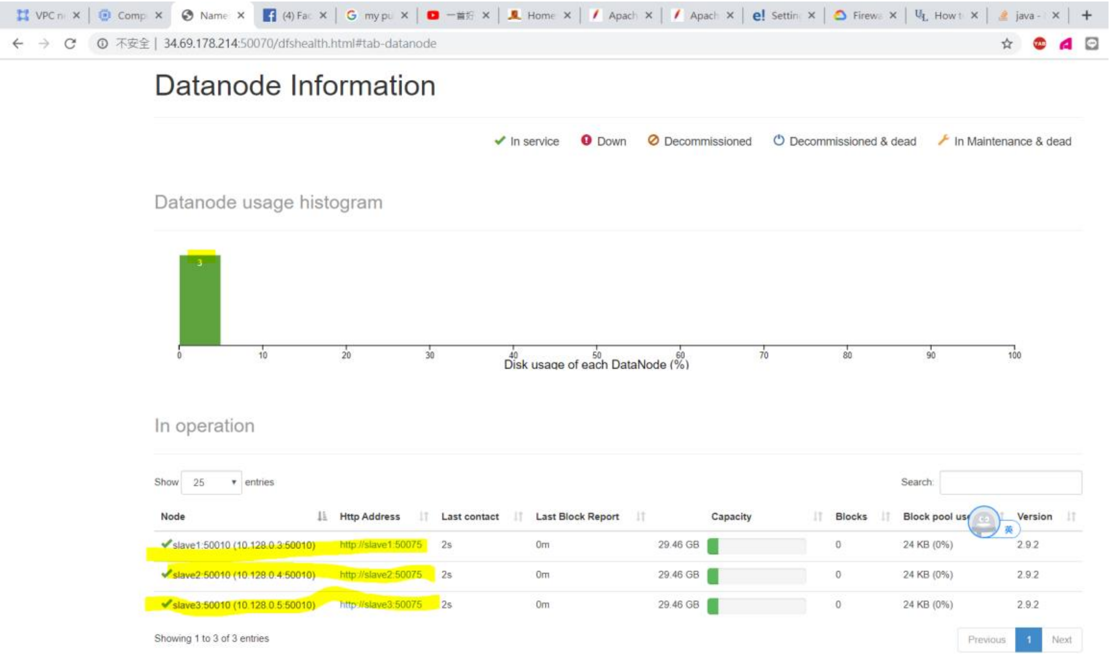
```
jackchim1998@instance-1:~/hadoop-2.9.2$ ./bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.j
ar teravalidate terasort/output terasort/check
19/09/07 11:59:14 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/09/07 11:59:14 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/09/07 11:59:14 INFO input.FileInputFormat: Total input files to process : 1
Spent 58ms computing base-splits.
Spent 4ms computing TeraScheduler splits.
19/09/07 11:59:14 INFO mapreduce.JobSubmitter: number of splits:1
19/09/07 11:59:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local52271266_0001
19/09/07 11:59:15 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
```



```
jackchim1998@instance-1:~/hadoop-2.9.2$ ./bin/hadoop dfs -ls terasort/check
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 2 items
-rw-r--r--   1 jackchim1998 supergroup          0 2019-09-07 11:59 terasort/check/_SUCCESS
-rw-r--r--   1 jackchim1998 supergroup         22 2019-09-07 11:59 terasort/check/part-r-00000
```

[40 marks] Multi-node Hadoop Cluster Setup

Set up 3 slaves successfully



Datanode Information

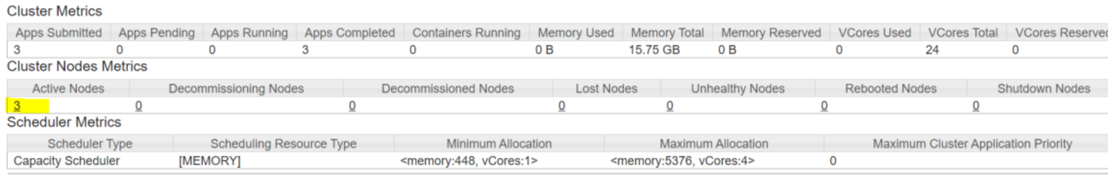
Datanode usage histogram

In operation

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool us	Version
✓slave1:50010 (10.128.0.3:50010)	http://slave1:50075	2s	0m	29.46 GB	0	24 KB (0%)	2.9.2
✓slave2:50010 (10.128.0.4:50010)	http://slave2:50075	2s	0m	29.46 GB	0	24 KB (0%)	2.9.2
✓slave3:50010 (10.128.0.5:50010)	http://slave3:50075	2s	0m	29.46 GB	0	24 KB (0%)	2.9.2

Showing 1 to 3 of 3 entries

They are all active nodes



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reservec
3	0	0	3	0	0 B	15.75 GB	0 B	0	24	0

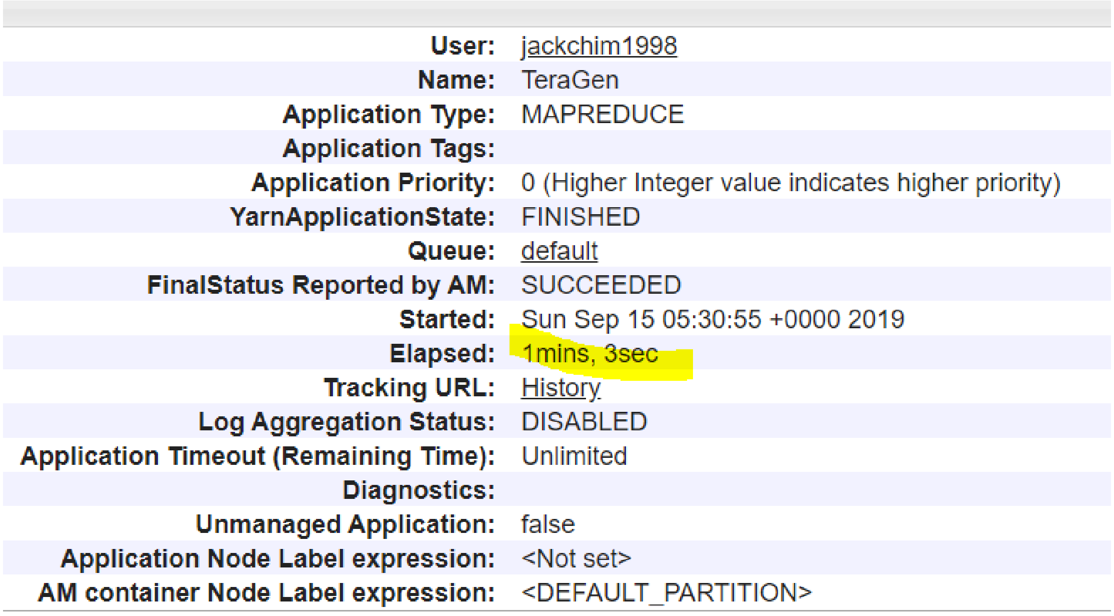
Cluster Nodes Metrics

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

Scheduler Metrics

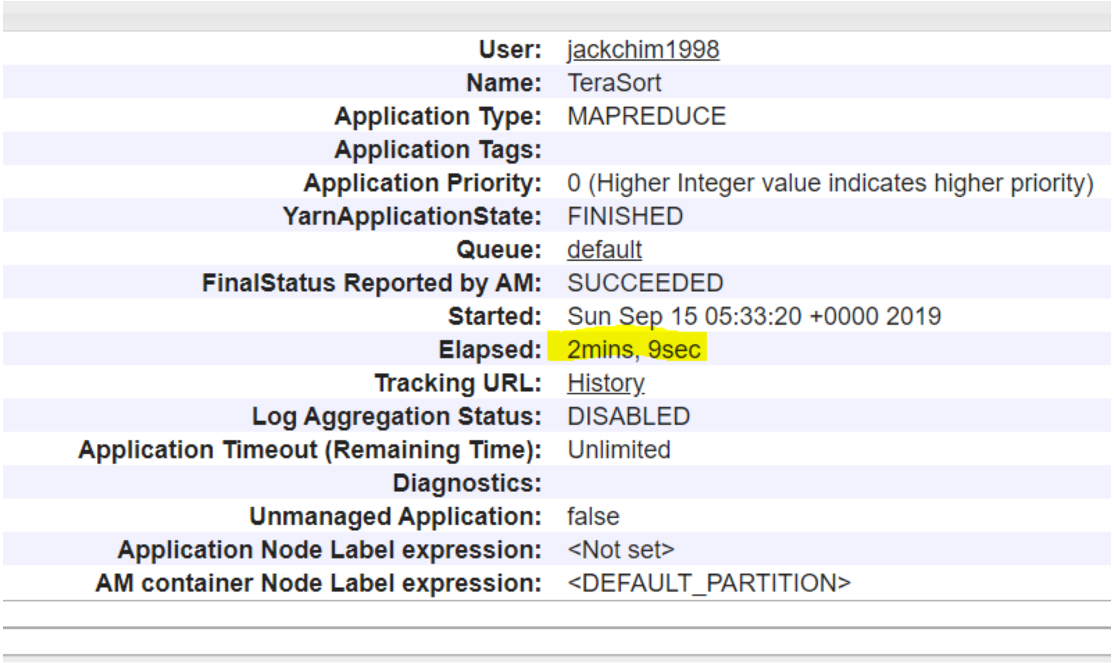
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:448, vCores:1>	<memory:5376, vCores:4>	0

We teragen 2GB data with 1min 3 sec



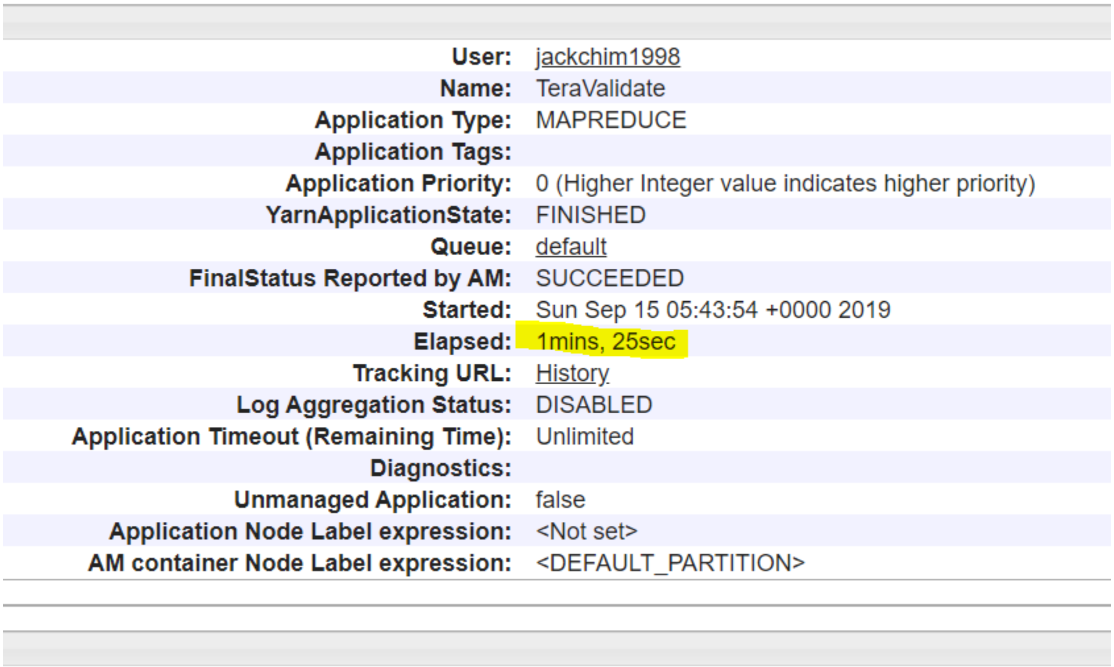
User:	jackchim1998
Name:	TeraGen
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Sep 15 05:30:55 +0000 2019
Elapsed:	1mins, 3sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Terasort 2GB data with 2 mins and 9 sec



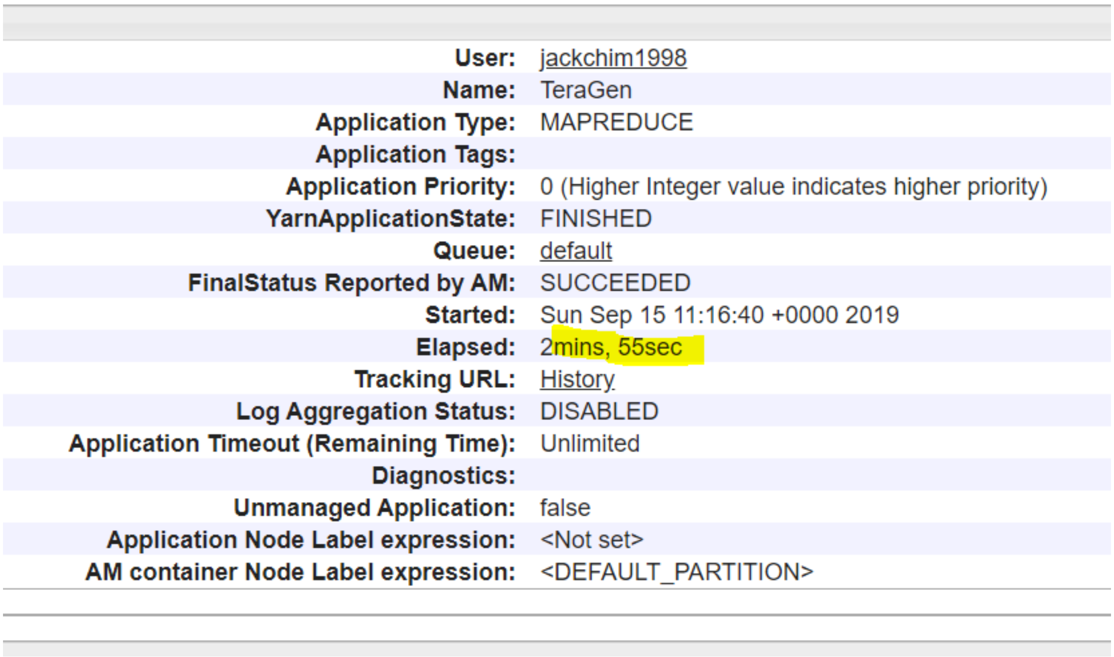
User:	jackchim1998
Name:	TeraSort
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Sep 15 05:33:20 +0000 2019
Elapsed:	2mins, 9sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

We teravalidate 2 GB data with 1 mins 25 sec



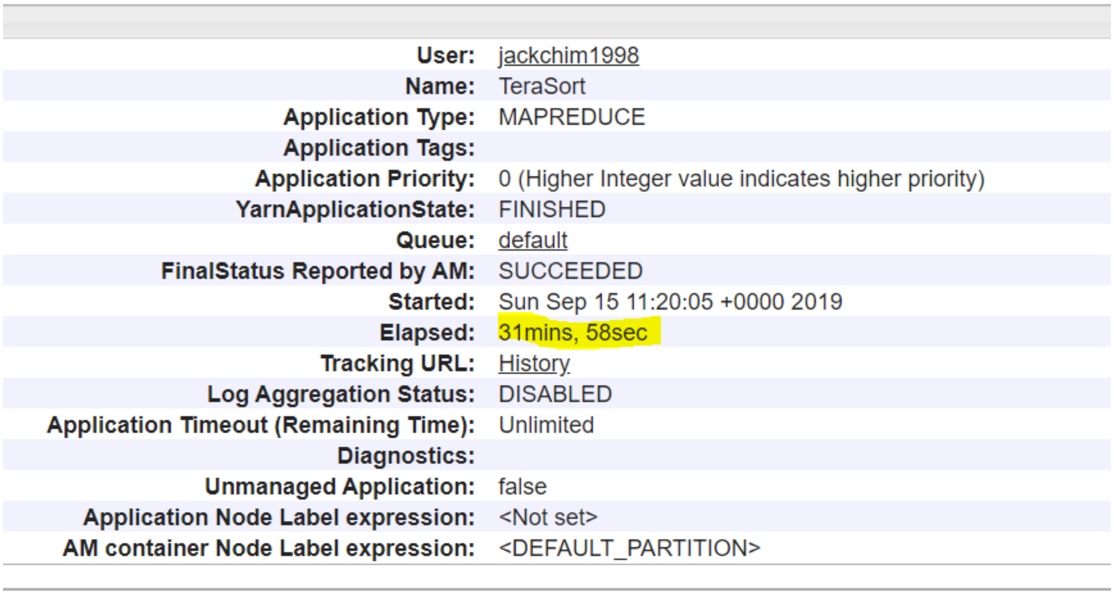
User:	jackchim1998
Name:	TeraValidate
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Sep 15 05:43:54 +0000 2019
Elapsed:	1mins, 25sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

We teragen 20GB data with 2min 55 sec



User:	jackchim1998
Name:	TeraGen
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Sep 15 11:16:40 +0000 2019
Elapsed:	2mins, 55sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Terasort 20GB data with 31 mins and 58 sec

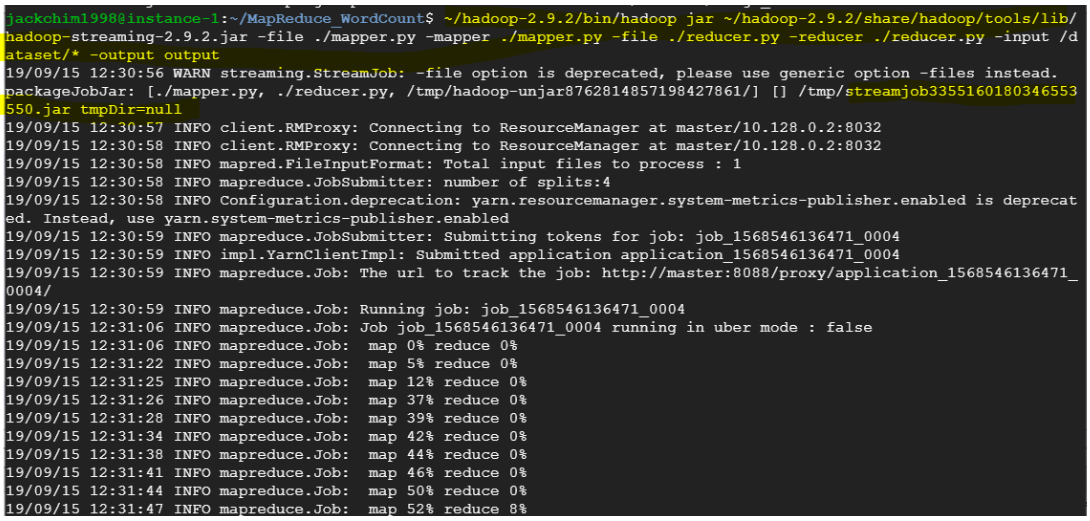


User:	jackchim1998
Name:	TeraSort
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Sep 15 11:20:05 +0000 2019
Elapsed:	31mins, 58sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

It is time-consuming, so that I would not do TearValidate.

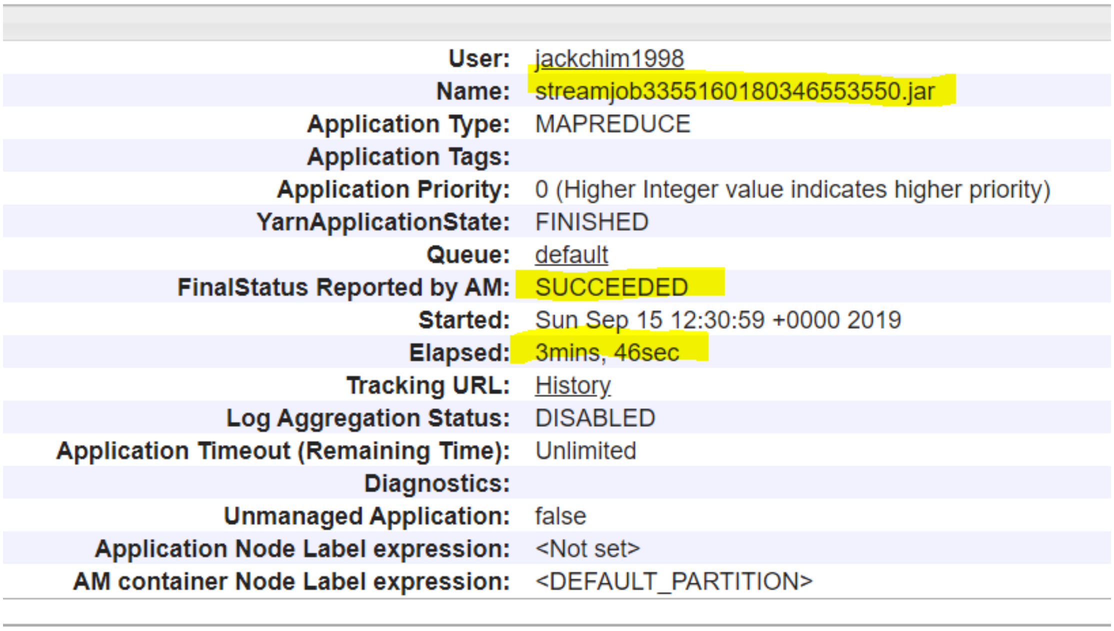
[40 marks] Running the Python Code on Hadoop

The below screen is the command, we can see the job name is streamjob3355160……….



```
jackchim1998@instance-1:~/MapReduce_WordCount$ ~/hadoop-2.9.2/bin/hadoop jar ~/hadoop-2.9.2/share/hadoop/tools/lib/
hadoop-streaming-2.9.2.jar -file ./mapper.py -mapper ./mapper.py -file ./reducer.py -reducer ./reducer.py -input /d
ataset/* -output output
19/09/15 12:30:56 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [./mapper.py, ./reducer.py, /tmp/hadoop-unjar8762814857198427861/] [] /tmp/streamjob3355160180346553
550.jar tmpDir=null
19/09/15 12:30:57 INFO client.RMProxy: Connecting to ResourceManager at master/10.128.0.2:8032
19/09/15 12:30:58 INFO client.RMProxy: Connecting to ResourceManager at master/10.128.0.2:8032
19/09/15 12:30:58 INFO mapred.FileInputFormat: Total input files to process : 1
19/09/15 12:30:58 INFO mapreduce.JobSubmitter: number of splits:4
19/09/15 12:30:58 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecat
ed. Instead, use yarn.system-metrics-publisher.enabled
19/09/15 12:30:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1568546136471_0004
19/09/15 12:30:59 INFO impl.YarnClientImpl: Submitted application application_1568546136471_0004
19/09/15 12:30:59 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1568546136471_
0004/
19/09/15 12:30:59 INFO mapreduce.Job: Running job: job_1568546136471_0004
19/09/15 12:31:06 INFO mapreduce.Job: Job job_1568546136471_0004 running in uber mode : false
19/09/15 12:31:06 INFO mapreduce.Job:  map 0% reduce 0%
19/09/15 12:31:22 INFO mapreduce.Job:  map 5% reduce 0%
19/09/15 12:31:25 INFO mapreduce.Job:  map 12% reduce 0%
19/09/15 12:31:26 INFO mapreduce.Job:  map 37% reduce 0%
19/09/15 12:31:28 INFO mapreduce.Job:  map 39% reduce 0%
19/09/15 12:31:34 INFO mapreduce.Job:  map 42% reduce 0%
19/09/15 12:31:38 INFO mapreduce.Job:  map 44% reduce 0%
19/09/15 12:31:41 INFO mapreduce.Job:  map 46% reduce 0%
19/09/15 12:31:44 INFO mapreduce.Job:  map 50% reduce 0%
19/09/15 12:31:47 INFO mapreduce.Job:  map 52% reduce 8%
```

The below screen shows that it elapsed 3mins 46sec.



User:	jackchim1998
Name:	streamjob3355160180346553550.jar
Application Type:	MAPREDUCE
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	FINISHED
Queue:	default
FinalStatus Reported by AM:	SUCCEEDED
Started:	Sun Sep 15 12:30:59 +0000 2019
Elapsed:	3mins, 46sec
Tracking URL:	History
Log Aggregation Status:	DISABLED
Application Timeout (Remaining Time):	Unlimited
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>