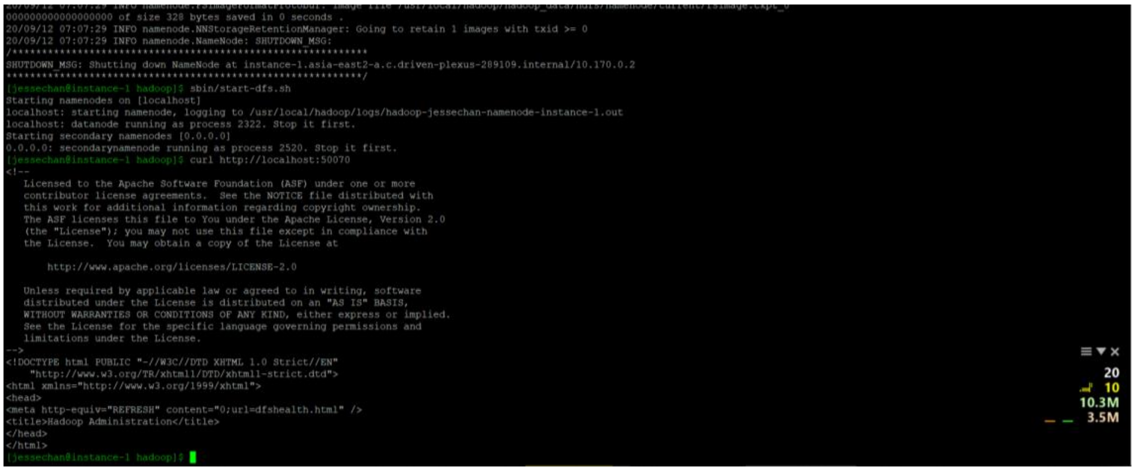
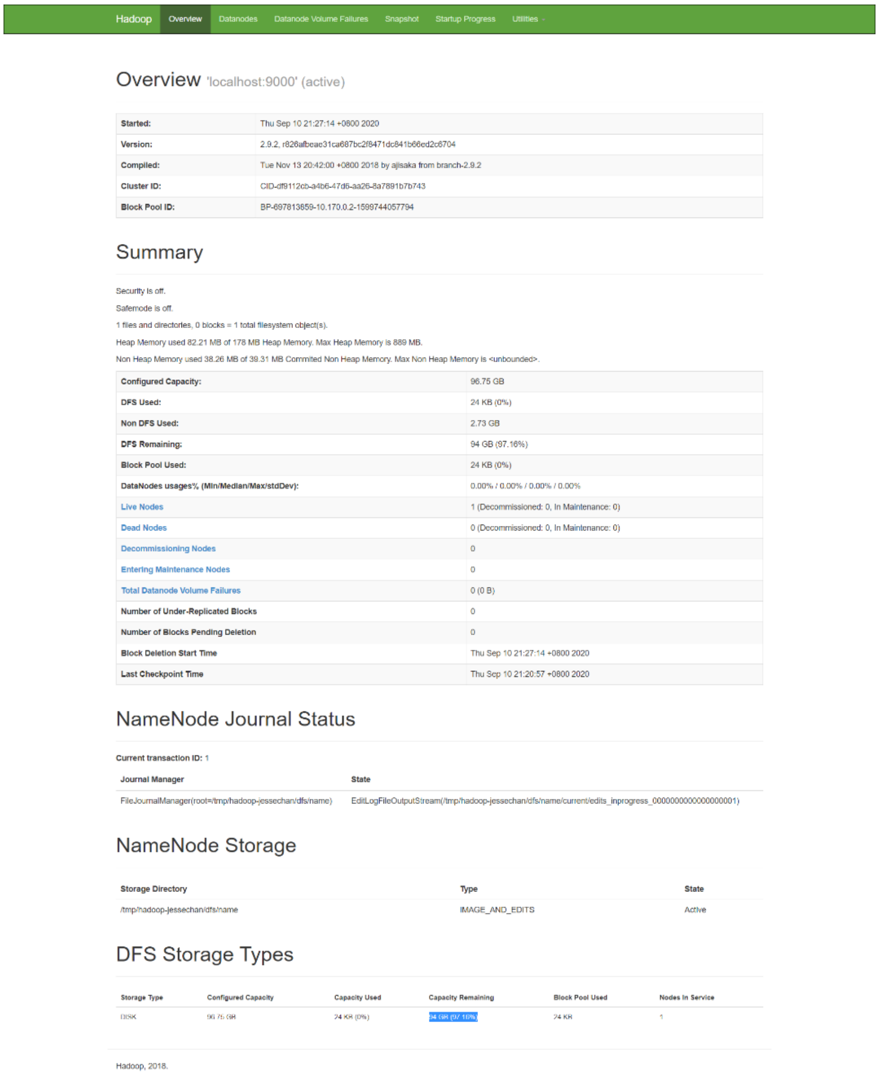
Single-node Hadoop Setup

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



```
20/09/12 07:07:29 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop/hadoop_data/hdfs/namenode/current/fsimage.ckpt_0
0000000000000000000 of size 328 bytes saved in 0 seconds .
20/09/12 07:07:29 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
20/09/12 07:07:29 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at instance-1.asia-east2-a.c.driven-plexus-289109.internal/10.170.0.2
************************************************************/
[jessechan@instance-1 hadoop]$ sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-jessechan-namenode-instance-1.out
localhost: datanode running as process 2322. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 2520. Stop it first.
[jessechan@instance-1 hadoop]$ 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>
[jessechan@instance-1 hadoop]$
```



Overview 'localhost:9000' (active)

Started:	Thu Sep 10 21:27:14 +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-df9112cb-a4b6-47d6-aa26-8a7891b7b743
Block Pool ID:	BP-697813659-10.170.0.2-1599744057794

Summary

Security is off.

Safemode is off.

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

Heap Memory used 82.21 MB of 178 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 38.26 MB of 39.31 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	96.75 GB
DFS Used:	24 KB (0%)
Non DFS Used:	2.73 GB
DFS Remaining:	94 GB (97.16%)
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 10 21:27:14 +0800 2020
Last Checkpoint Time	Thu Sep 10 21:20:57 +0800 2020

NameNode Journal Status

Current transaction ID: 1

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

NameNode Storage

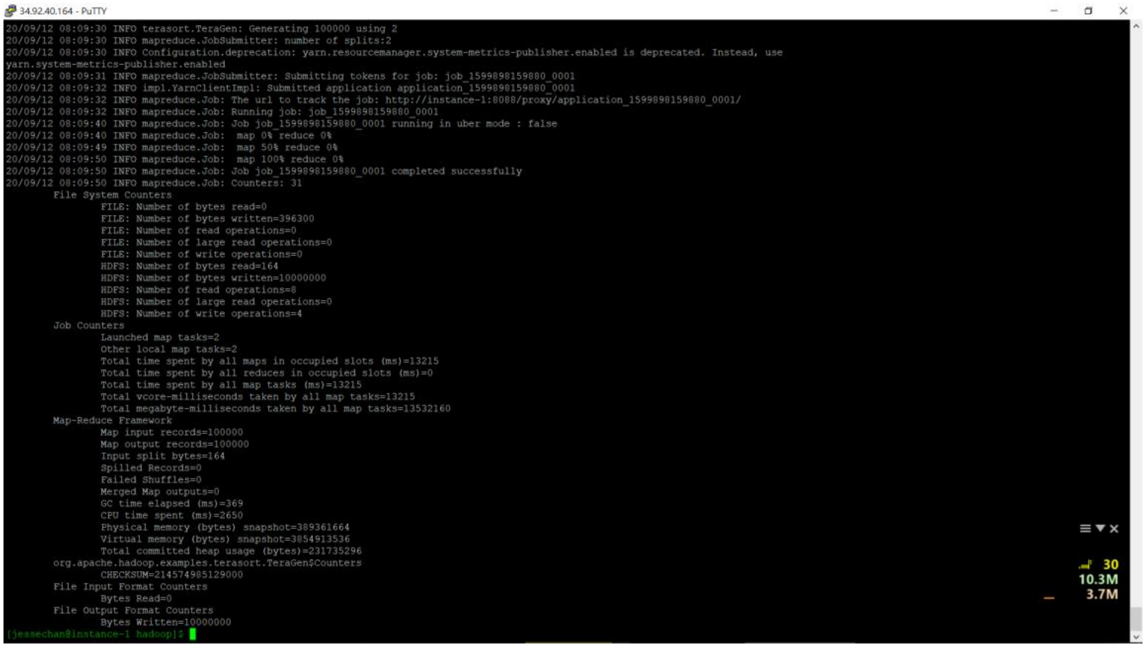
Storage Directory	Type	State
/tmp/hadoop-jessechan/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	96.75 GB	24 KB (0%)	94 GB (97.16%)	24 KB	1

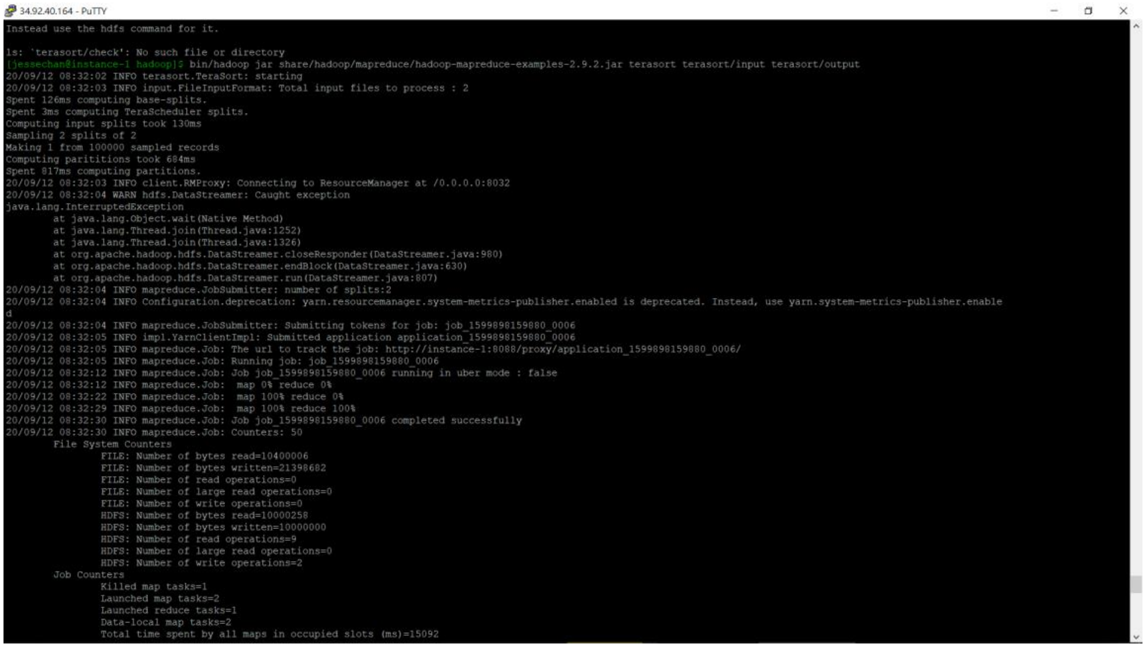
Hadoop, 2018.

We try to run the Terasort. First, we generate a list of random numbers.



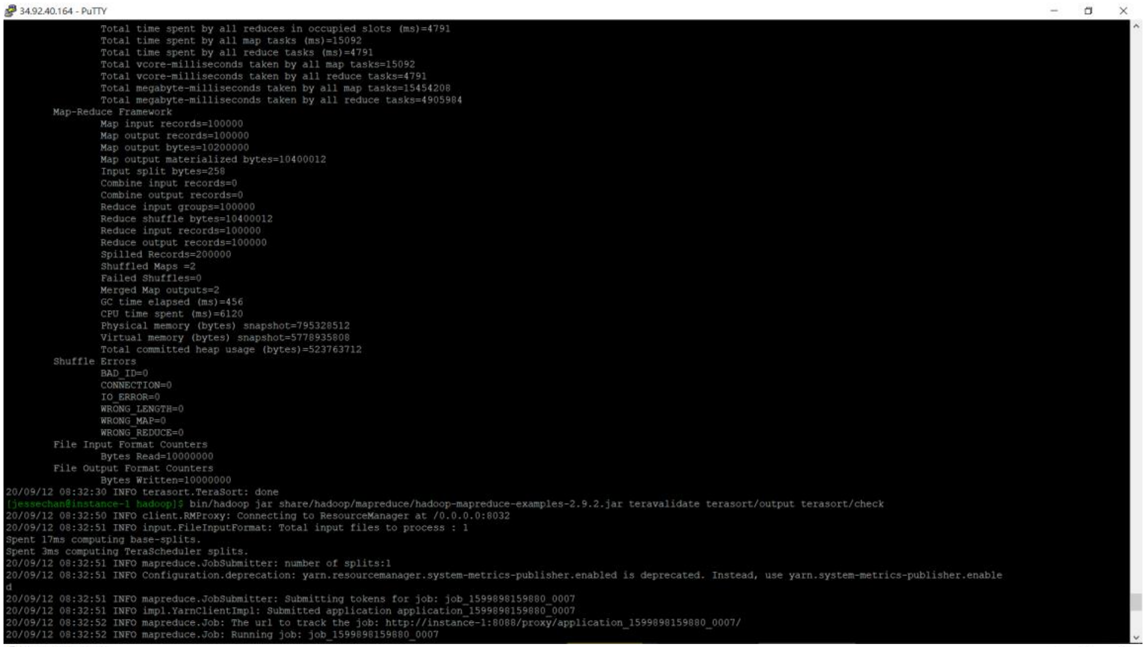
```
20/09/12 08:09:30 INFO terasort.TeraGen: Generating 100000 using 2
20/09/12 08:09:30 INFO mapreduce.JobSubmitter: number of splits:2
20/09/12 08:09:30 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
yarn.system-metrics-publisher.enabled
20/09/12 08:09:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1599898159880_0001
20/09/12 08:09:32 INFO impl.YarnClientImpl: Submitted application application_1599898159880_0001
20/09/12 08:09:32 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1599898159880_0001/
20/09/12 08:09:32 INFO mapreduce.Job: Running job: job_1599898159880_0001
20/09/12 08:09:40 INFO mapreduce.Job: Job job_1599898159880_0001 running in uber mode : false
20/09/12 08:09:40 INFO mapreduce.Job:  map 0% reduce 0%
20/09/12 08:09:49 INFO mapreduce.Job:  map 50% reduce 0%
20/09/12 08:09:50 INFO mapreduce.Job:  map 100% reduce 0%
20/09/12 08:09:50 INFO mapreduce.Job: Job job_1599898159880_0001 completed successfully
20/09/12 08:09:50 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396300
                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=10000000
                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)=13215
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=13215
                Total vcore-milliseconds taken by all map tasks=13215
                Total megabyte-milliseconds taken by all map tasks=13532160
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Input split bytes=164
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=369
                CPU time spent (ms)=2650
                Physical memory (bytes) snapshot=389361664
                Virtual memory (bytes) snapshot=3854913536
                Total committed heap usage (bytes)=231735296
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=214574985129000
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=10000000
[jessechan@instance-1 hadoop]$
```

Then, Sorting the data.



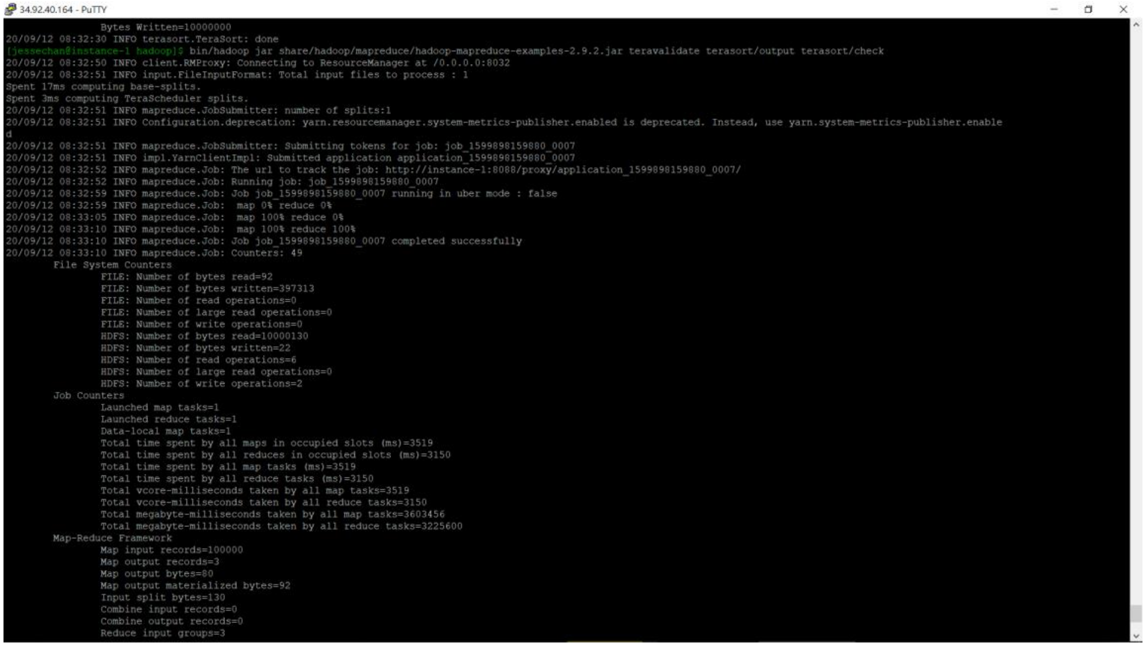
```
Instead use the hdfs command for it.

ls: `terasort/check': No such file or directory
[jessechan@instance-1 hadoop]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar terasort terasort/input terasort/output
20/09/12 08:32:02 INFO terasort.TeraSort: starting
20/09/12 08:32:03 INFO input.FileInputFormat: Total input files to process : 2
Spent 126ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
Computing input splits took 130ms
Sampling 2 splits of 2
Making 1 from 100000 sampled records
Computing parititions took 684ms
Spent 817ms computing partitions.
20/09/12 08:32:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/09/12 08:32:04 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)
20/09/12 08:32:04 INFO mapreduce.JobSubmitter: number of splits:2
20/09/12 08:32:04 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enable
d
20/09/12 08:32:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1599898159880_0006
20/09/12 08:32:05 INFO impl.YarnClientImpl: Submitted application application_1599898159880_0006
20/09/12 08:32:05 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1599898159880_0006/
20/09/12 08:32:05 INFO mapreduce.Job: Running job: job_1599898159880_0006
20/09/12 08:32:12 INFO mapreduce.Job: Job job_1599898159880_0006 running in uber mode : false
20/09/12 08:32:12 INFO mapreduce.Job:  map 0% reduce 0%
20/09/12 08:32:22 INFO mapreduce.Job:  map 100% reduce 0%
20/09/12 08:32:29 INFO mapreduce.Job:  map 100% reduce 100%
20/09/12 08:32:30 INFO mapreduce.Job: Job job_1599898159880_0006 completed successfully
20/09/12 08:32:30 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=10400006
                FILE: Number of bytes written=21398682
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000258
                HDFS: Number of bytes written=10000000
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=15092
```

Finish and now checking,



34.92.40.164 - PuTTY

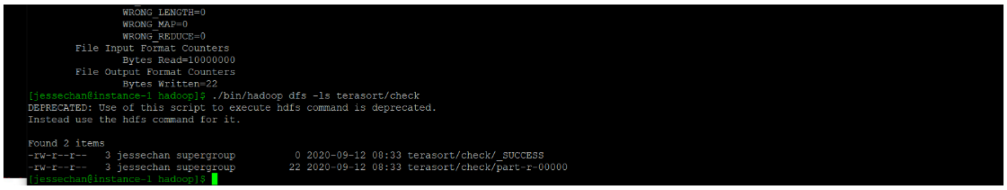
```
                Total time spent by all reduces in occupied slots (ms)=4791
                Total time spent by all map tasks (ms)=15092
                Total time spent by all reduce tasks (ms)=4791
                Total vcore-milliseconds taken by all map tasks=15092
                Total vcore-milliseconds taken by all reduce tasks=4791
                Total megabyte-milliseconds taken by all map tasks=15454208
                Total megabyte-milliseconds taken by all reduce tasks=4905984
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Map output bytes=10200000
                Map output materialized bytes=10400012
                Input split bytes=258
                Combine input records=0
                Combine output records=0
                Reduce input groups=100000
                Reduce shuffle bytes=10400012
                Reduce input records=100000
                Reduce output records=100000
                Spilled Records=200000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=456
                CPU time spent (ms)=6120
                Physical memory (bytes) snapshot=795328512
                Virtual memory (bytes) snapshot=5778935808
                Total committed heap usage (bytes)=523763712
        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
20/09/12 08:32:30 INFO terasort.TeraSort: done
[jessechan@instance-1 hadoop]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/output terasort/check
20/09/12 08:32:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/09/12 08:32:51 INFO input.FileInputFormat: Total input files to process : 1
Spent 17ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
20/09/12 08:32:51 INFO mapreduce.JobSubmitter: number of splits:1
20/09/12 08:32:51 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enable
d
20/09/12 08:32:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1599898159880_0007
20/09/12 08:32:51 INFO impl.YarnClientImpl: Submitted application application_1599898159880_0007
20/09/12 08:32:52 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1599898159880_0007/
20/09/12 08:32:52 INFO mapreduce.Job: Running job: job_1599898159880_0007
```



34.92.40.164 - PuTTY

```
                Bytes Written=10000000
20/09/12 08:32:30 INFO terasort.TeraSort: done
[jessechan@instance-1 hadoop]$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar teravalidate terasort/output terasort/check
20/09/12 08:32:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/09/12 08:32:51 INFO input.FileInputFormat: Total input files to process : 1
Spent 17ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
20/09/12 08:32:51 INFO mapreduce.JobSubmitter: number of splits:1
20/09/12 08:32:51 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enable
d
20/09/12 08:32:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1599898159880_0007
20/09/12 08:32:51 INFO impl.YarnClientImpl: Submitted application application_1599898159880_0007
20/09/12 08:32:52 INFO mapreduce.Job: The url to track the job: http://instance-1:8088/proxy/application_1599898159880_0007/
20/09/12 08:32:52 INFO mapreduce.Job: Running job: job_1599898159880_0007
20/09/12 08:32:59 INFO mapreduce.Job: Job job_1599898159880_0007 running in uber mode : false
20/09/12 08:32:59 INFO mapreduce.Job:  map 0% reduce 0%
20/09/12 08:33:05 INFO mapreduce.Job:  map 100% reduce 0%
20/09/12 08:33:10 INFO mapreduce.Job:  map 100% reduce 100%
20/09/12 08:33:10 INFO mapreduce.Job: Job job_1599898159880_0007 completed successfully
20/09/12 08:33:10 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=92
                FILE: Number of bytes written=397313
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000130
                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)=3519
                Total time spent by all reduces in occupied slots (ms)=3150
                Total time spent by all map tasks (ms)=3519
                Total time spent by all reduce tasks (ms)=3150
                Total vcore-milliseconds taken by all map tasks=3519
                Total vcore-milliseconds taken by all reduce tasks=3150
                Total megabyte-milliseconds taken by all map tasks=3603456
                Total megabyte-milliseconds taken by all reduce tasks=3225600
        Map-Reduce Framework
                Map input records=100000
                Map output records=3
                Map output bytes=80
                Map output materialized bytes=92
                Input split bytes=130
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
```

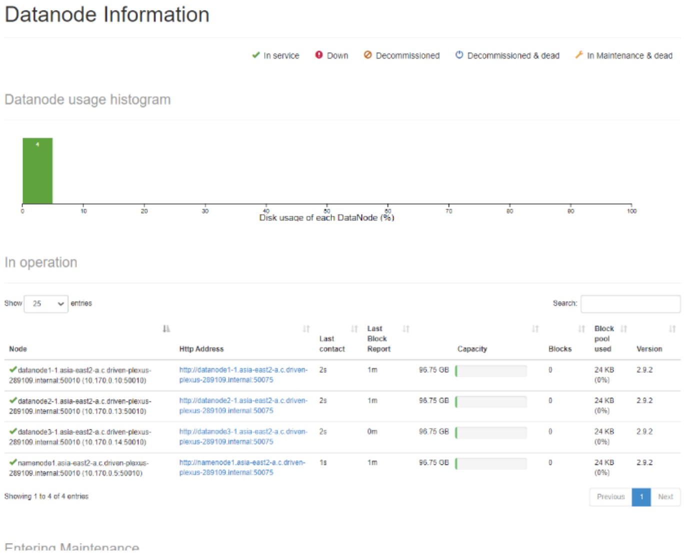
Use command to validate the sorting result,



```
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=10000000
        File Output Format Counters
                Bytes Written=22
[jessechan@instance-1 hadoop]$ ./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--   3 jessechan supergroup          0 2020-09-12 08:33 terasort/check/_SUCCESS
-rw-r--r--   3 jessechan supergroup         22 2020-09-12 08:33 terasort/check/part-r-00000
[jessechan@instance-1 hadoop]$
```

Multi-node Hadoop Cluster Setup

Using One name-node as master and 3 data-node as slave



Datanode Information

Datanode usage histogram

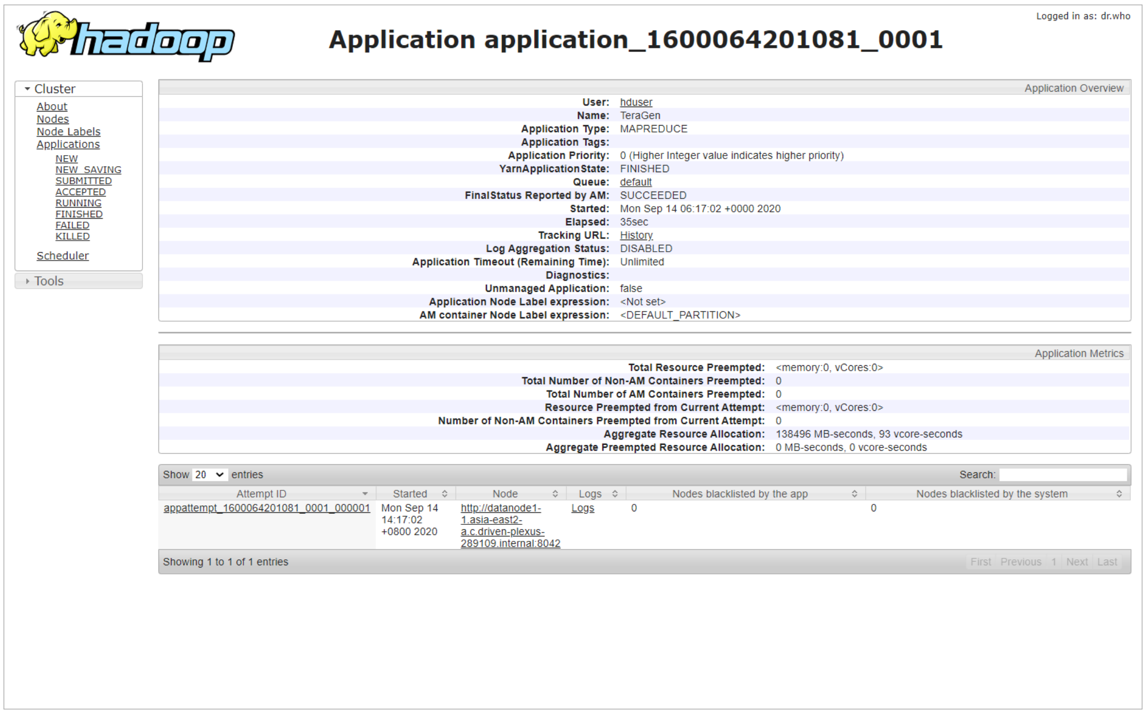
In operation

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ datanode1-1.asia-east2-a.c.driven-plexus-289109.internal:50010 (10.170.0.10:50010)	http://datanode1-1.asia-east2-a.c.driven-plexus-289109.internal:50075	2s	1m	96.75 GB	0	24 KB (0%)	2.9.2
✓ datanode2-1.asia-east2-a.c.driven-plexus-289109.internal:50010 (10.170.0.13:50010)	http://datanode2-1.asia-east2-a.c.driven-plexus-289109.internal:50075	2s	1m	96.75 GB	0	24 KB (0%)	2.9.2
✓ datanode3-1.asia-east2-a.c.driven-plexus-289109.internal:50010 (10.170.0.14:50010)	http://datanode3-1.asia-east2-a.c.driven-plexus-289109.internal:50075	2s	0m	96.75 GB	0	24 KB (0%)	2.9.2
✓ namenode1.asia-east2-a.c.driven-plexus-289109.internal:50010 (10.170.0.5:50010)	http://namenode1.asia-east2-a.c.driven-plexus-289109.internal:50075	1s	1m	96.75 GB	0	24 KB (0%)	2.9.2

Showing 1 to 4 of 4 entries

Entering Maintenance

In the 2GB TeraSort example:



Application application_1600064201081_0001

Application Overview

User:	hduser
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:	Mon Sep 14 06:17:02 +0000 2020
Elapsed:	35sec
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>

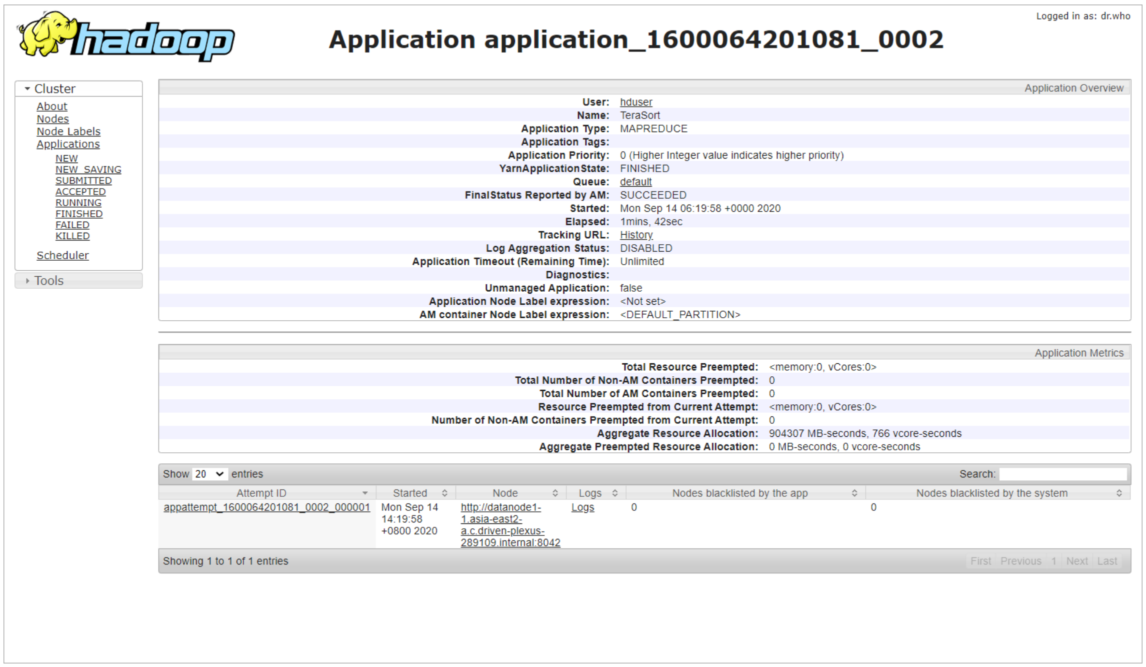
Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	138496 MB-seconds, 93 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600064201081_0001_000001	Mon Sep 14 14:17:02 +0800 2020	http://datanode1-1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

Showing 1 to 1 of 1 entries

Figure 1 2GB Gen.



Application application_1600064201081_0002

Application Overview

User:	hduser
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:	Mon Sep 14 06:19:58 +0000 2020
Elapsed:	1mins, 42sec
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>

Application Metrics

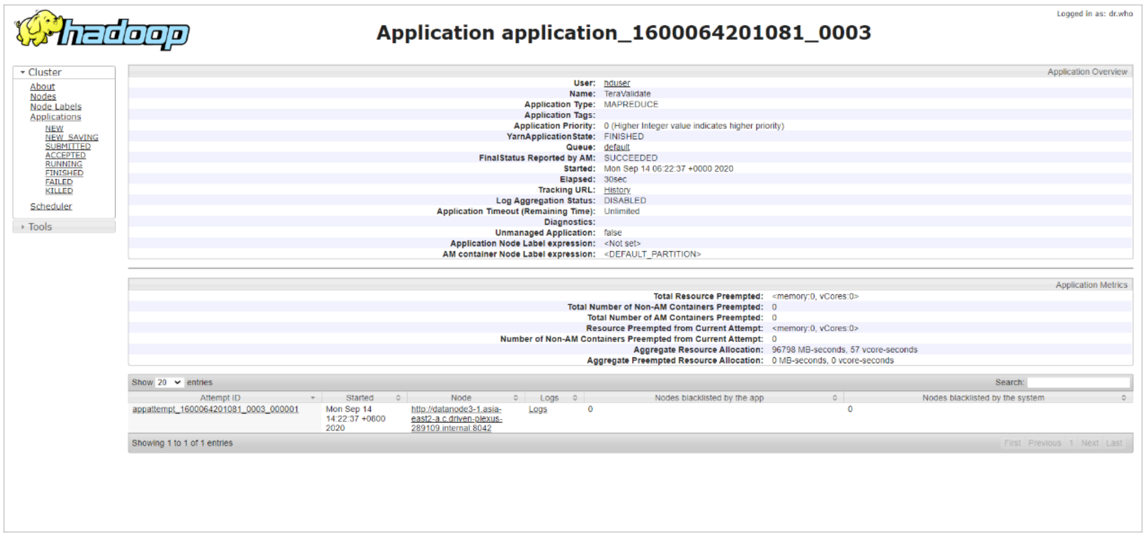
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	904307 MB-seconds, 766 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600064201081_0002_000001	Mon Sep 14 14:19:58 +0800 2020	http://datanode1-1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

Showing 1 to 1 of 1 entries

Figure 2 2GB Sort



Application application_1600064201081_0003

Application Overview

User:	hduser
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:	Mon Sep 14 06:22:37 +0000 2020
Elapsed:	30sec
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>

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	96798 MB-seconds, 57 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

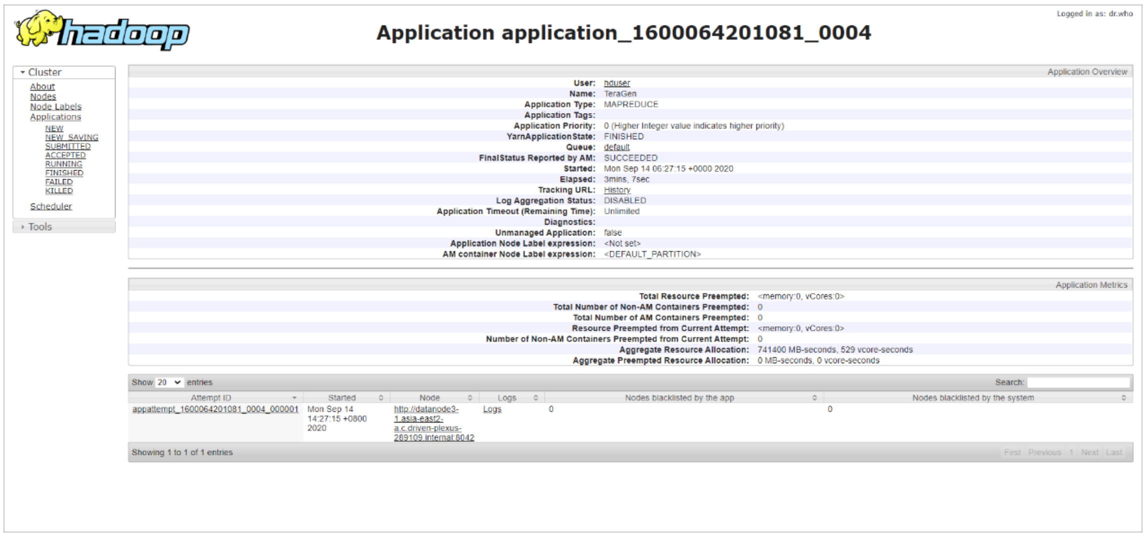
Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600064201081_0003_000001	Mon Sep 14 14:22:37 +0800 2020	http://datanode3-1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

Showing 1 to 1 of 1 entries

Figure 3 2GB Check

In the 20GB TeraSort example:



Application application_1600064201081_0004

User:	hduser
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:	Mon Sep 14 06:27:15 +0000 2020
Elapsed:	3mins, 7sec
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>

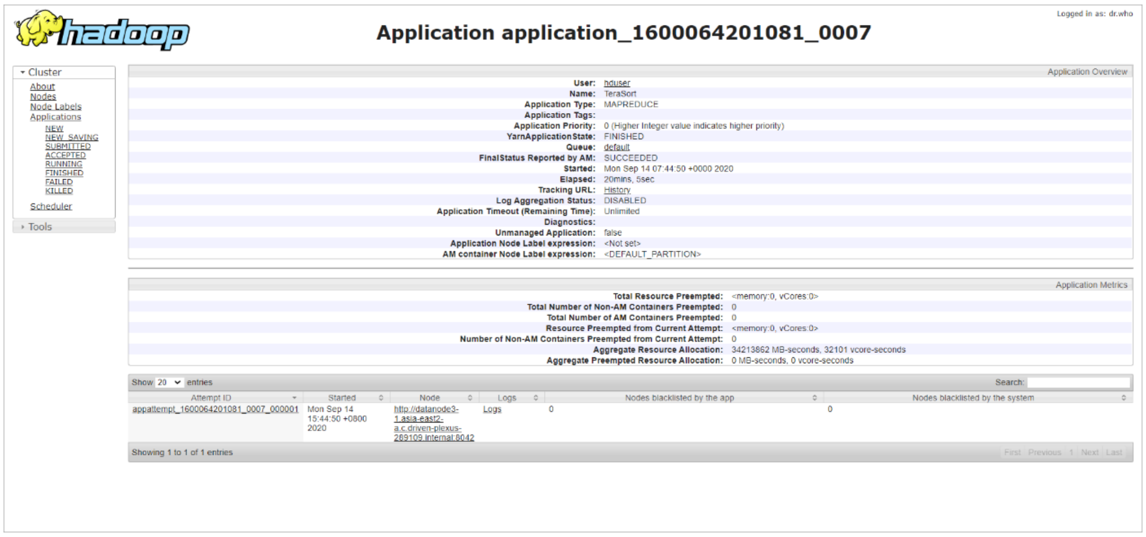
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	741400 MB-seconds, 529 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600064201081_0004_000001	Mon Sep 14 14:27:15 +0800 2020	http://datanode3-1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Figure 4 20GB Gen.



Application application_1600064201081_0007

User:	hduser
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:	Mon Sep 14 07:44:50 +0000 2020
Elapsed:	20mins, 5sec
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>

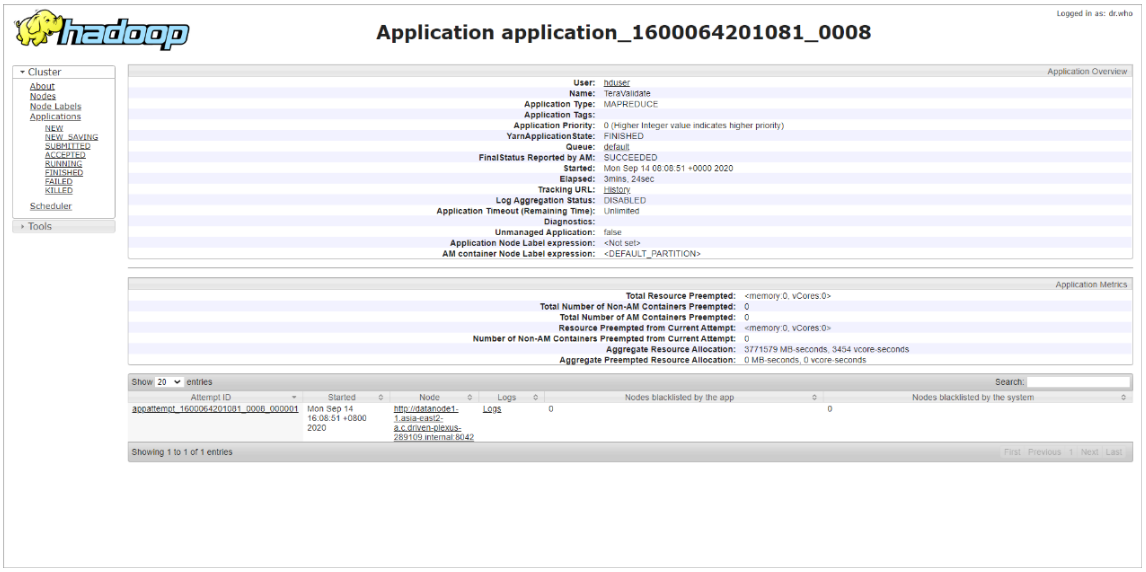
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	34213862 MB-seconds, 32101 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600064201081_0007_000001	Mon Sep 14 15:44:50 +0800 2020	http://datanode3-1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Figure 5 20GB Sort.



Logged in as: dr.who

Application application_1600064201081_0008

Application Overview

User: hduser
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: Mon Sep 14 08:08:51 +0000 2020
Elapsed: 3mins, 24sec
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>

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 3771579 MB-seconds, 3454 vcore-seconds
Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

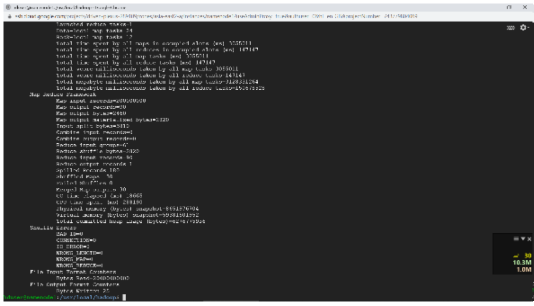
Show 20 entries Search:

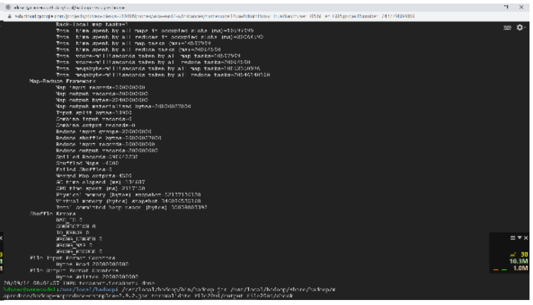
Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600064201081_0008_000001	Mon Sep 14 16:08:51 +0800 2020	http://datanode1-1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

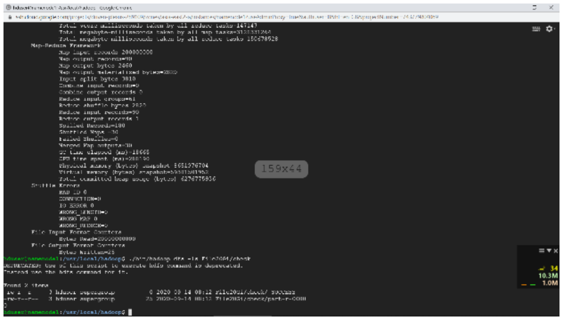
Showing 1 to 1 of 1 entries

Figure 6 20GB Check

As conclusion, The 20GB TeraSort program take much longer time than 2GB TeraSort program due to the data size. The gen., sort. and check process are similar to above, below figures attached as references.

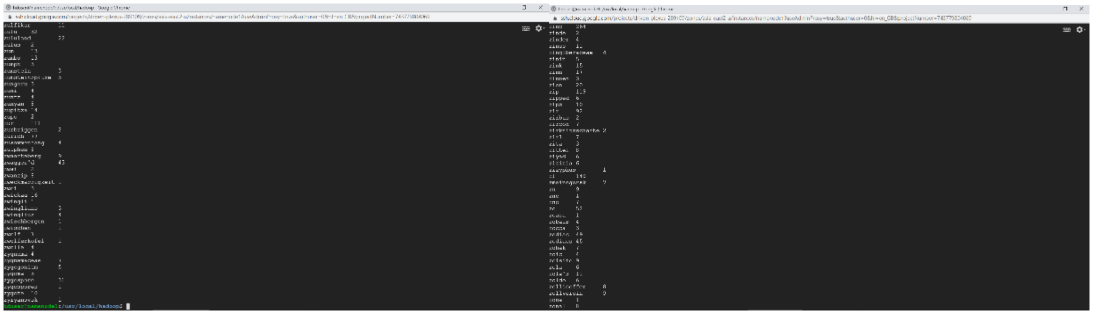


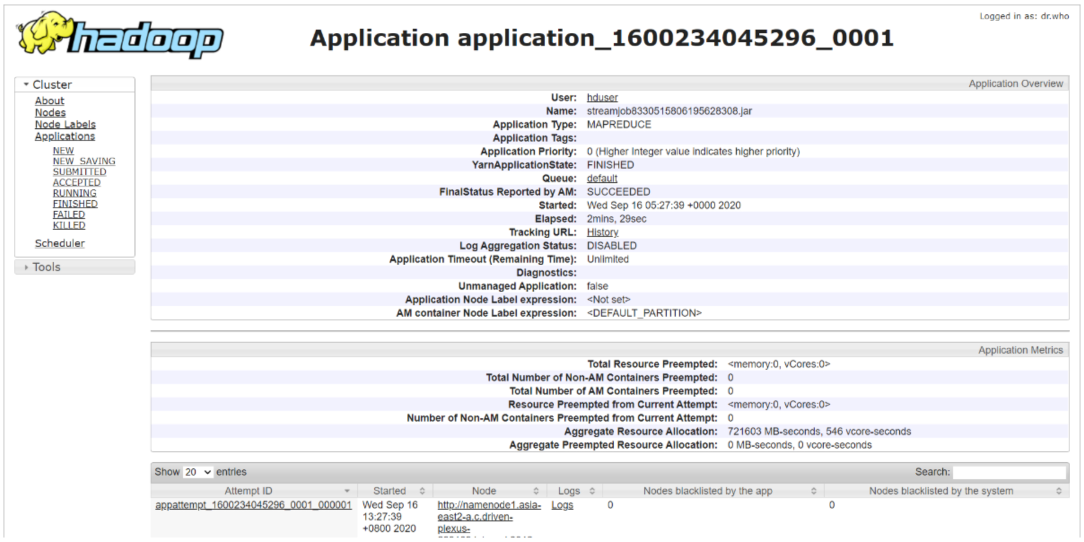




Running the Python Code on Hadoop

The result of the Python Wordcount Script.





Logged in as: dr.who

Application application_1600234045296_0001

Application Overview

User:	hduser
Name:	streamjob8330515806195628308.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:	Wed Sep 16 05:27:39 +0000 2020
Elapsed:	2mins, 29sec
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>

Application Metrics

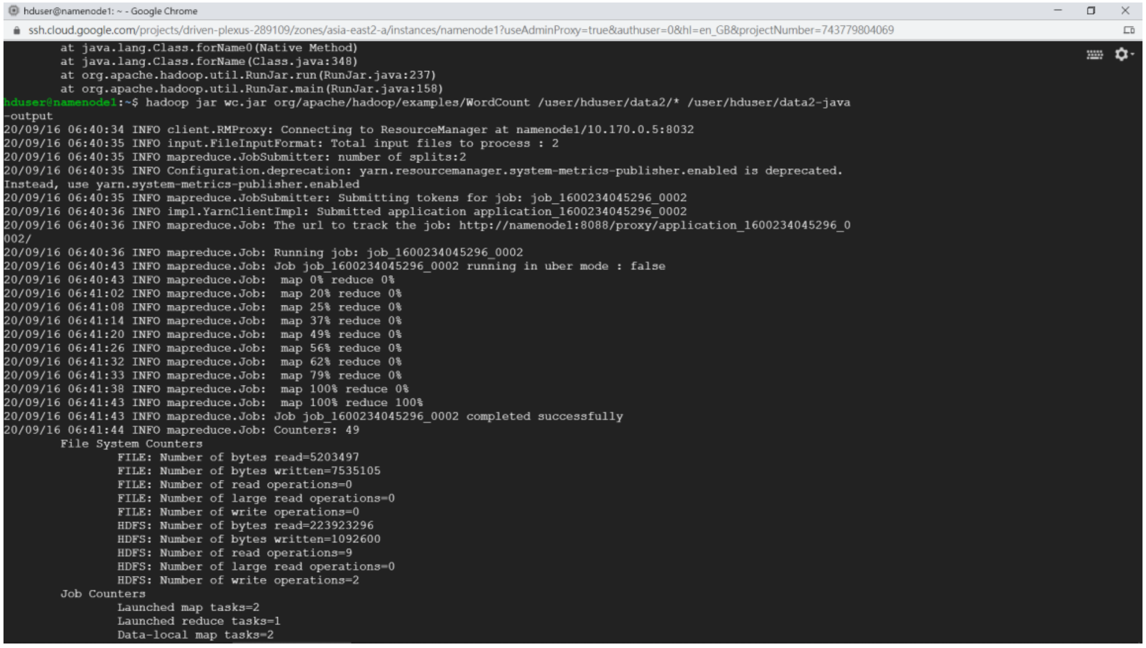
Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	721603 MB-seconds, 546 vcore-seconds
Aggregate Preempted Resource Allocation:	0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600234045296_0001_000001	Wed Sep 16 13:27:39 +0800 2020	http://namenode1.asia-east2-a.c.driven-plexus-	Logs	0	0

Figure 7 Python Wordcount Program

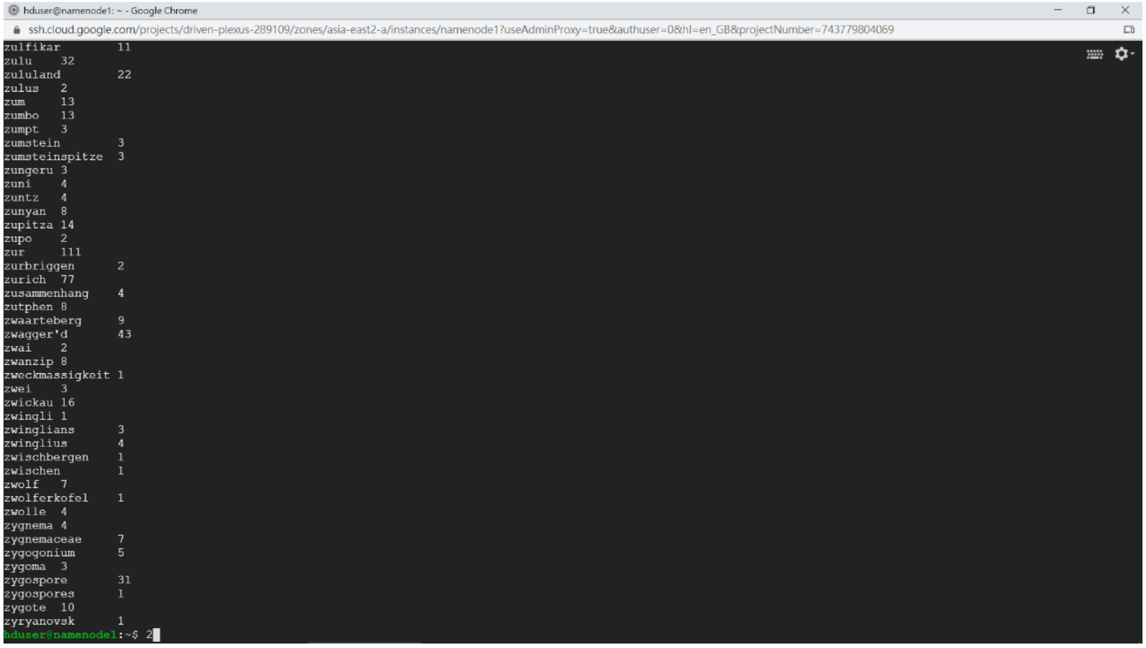
Compiling the Java WordCount program for MapReduce



hduser@namenode1: ~ - Google Chrome

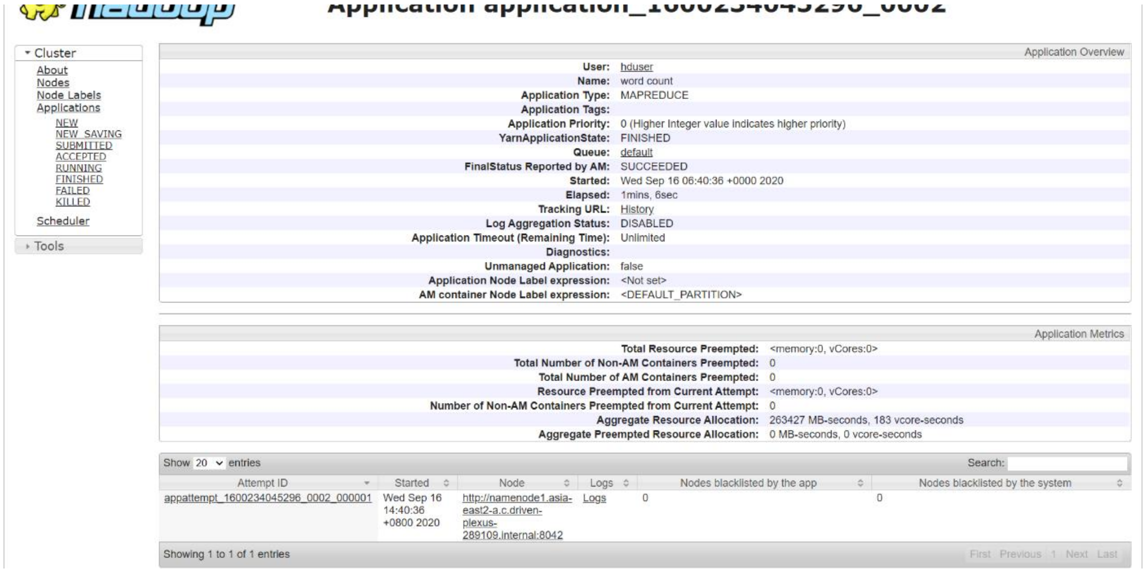
ssh.cloud.google.com/projects/driven-plexus-289109/zones/asia-east2-a/instances/namenode1?useAdminProxy=true&authuser=0&hl=en_GB&projectNumber=743779804069

```
        at java.lang.Class.forName0(Native Method)
        at java.lang.Class.forName(Class.java:348)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:237)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:158)
hduser@namenode1:~$ hadoop jar wc.jar org/apache/hadoop/examples/WordCount /user/hduser/data2/* /user/hduser/data2-java
-output
20/09/16 06:40:34 INFO client.RMProxy: Connecting to ResourceManager at namenode1/10.170.0.5:8032
20/09/16 06:40:35 INFO input.FileInputFormat: Total input files to process : 2
20/09/16 06:40:35 INFO mapreduce.JobSubmitter: number of splits:2
20/09/16 06:40:35 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enabled
20/09/16 06:40:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1600234045296_0002
20/09/16 06:40:36 INFO impl.YarnClientImpl: Submitted application application_1600234045296_0002
20/09/16 06:40:36 INFO mapreduce.Job: The url to track the job: http://namenode1:8088/proxy/application_1600234045296_0
002/
20/09/16 06:40:36 INFO mapreduce.Job: Running job: job_1600234045296_0002
20/09/16 06:40:43 INFO mapreduce.Job: Job job_1600234045296_0002 running in uber mode : false
20/09/16 06:40:43 INFO mapreduce.Job:  map 0% reduce 0%
20/09/16 06:41:02 INFO mapreduce.Job:  map 20% reduce 0%
20/09/16 06:41:08 INFO mapreduce.Job:  map 25% reduce 0%
20/09/16 06:41:14 INFO mapreduce.Job:  map 37% reduce 0%
20/09/16 06:41:20 INFO mapreduce.Job:  map 49% reduce 0%
20/09/16 06:41:26 INFO mapreduce.Job:  map 56% reduce 0%
20/09/16 06:41:32 INFO mapreduce.Job:  map 62% reduce 0%
20/09/16 06:41:33 INFO mapreduce.Job:  map 79% reduce 0%
20/09/16 06:41:38 INFO mapreduce.Job:  map 100% reduce 0%
20/09/16 06:41:43 INFO mapreduce.Job:  map 100% reduce 100%
20/09/16 06:41:43 INFO mapreduce.Job: Job job_1600234045296_0002 completed successfully
20/09/16 06:41:44 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=5203497
                FILE: Number of bytes written=7535105
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=223923296
                HDFS: Number of bytes written=1092600
                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
```


hduser@namenode1: ~ - Google Chrome
ssh.cloud.google.com/projects/driven-plexus-289109/zones/asia-east2-a/instances/namenode1?useAdminProxy=true&authuser=0&hl=en_GB&projectNumber=743779804069

```
zulfikar        11
zulu    32
zululand        22
zulus   2
zum     13
zumbo   13
zumpt   3
zumstein        3
zumsteinspitze  3
zungeru 3
zuni    4
zuntz   4
zunyan  8
zupitza 14
zupo    2
zur     111
zurbriggen      2
zurich  77
zusammenhang    4
zutphen 8
zwaarteberg     9
zwagger'd       43
zwai    2
zwanzip 8
zweckmassigkeit 1
zwei    3
zwickau 16
zwingli 1
zwinglians      3
zwinglius       4
zwischbergen    1
zwischen        1
zwolf   7
zwolferkofel    1
zwolle  4
zygnema 4
zygnemaceae     7
zygogonium      5
zygoma  3
zygospore       31
zygospores      1
zygote  10
zyryanovsk      1
hduser@namenode1:~$ 2
```

The result is identical to the above python program.



Application Overview

User: hduser
Name: word count
Application Type: MAPREDUCE
Application Tags:
Application Priority: 0 (Higher Integer value indicates higher priority)
YarnApplicationState: FINISHED
Queue: default
FinalStatus Reported by AM: SUCCEEDED
Started: Wed Sep 16 06:40:36 +0000 2020
Elapsed: 1mins, 6sec
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>

Application Metrics

Total Resource Preempted: <memory:0, vCores:0>
Total Number of Non-AM Containers Preempted: 0
Total Number of AM Containers Preempted: 0
Resource Preempted from Current Attempt: <memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt: 0
Aggregate Resource Allocation: 263427 MB-seconds, 183 vcore-seconds
Aggregate Preempted Resource Allocation: 0 MB-seconds, 0 vcore-seconds

Show 20 entries Search:

Attempt ID	Started	Node	Logs	Nodes blacklisted by the app	Nodes blacklisted by the system
appattempt_1600234045296_0002_000001	Wed Sep 16 14:40:36 +0800 2020	http://namenode1.asia-east2-a.c.driven-plexus-289109.internal:8042	Logs	0	0

Showing 1 to 1 of 1 entries

Figure 8 Java Wordcount Program

As a conclusion, the java version of Wordcount program run faster than python program. I guess it due to Java is less dynamic than Python and It makes it more efficient on VM.