

ASSIGNMENT

Name-Tushar Raj Verma

1.Setup 2 apache hadoop clusters with 2 nodes each i.e.

Cluster 1 : M1 - NN,RM,DN,NM

M2 - DN,NM,SNN

Cluster 2 : M3 - NN,RM,DN,NM

M4 - DN,NM,SNN

Now demonstrate writing data (such as directory containing files from Cluster 1 to Cluster 2. Also demonstrate updating the data if data in Cluster 1 changes (say new files added to directory that was copied earlier) and overwriting data.

Hint: Usage of Distcp

Ans-1

The Cluster1 having machines-m1,m2 with hadoop running on it

```
hdu@m1:/usr/local/hadoop$ jps
5992 NodeManager
5434 NameNode
6204 Jps
5838 ResourceManager
5615 DataNode
hdu@m1:/usr/local/hadoop$
```

```
hdu@m2:~$ jps
2304 DataNode
2822 Jps
2620 NodeManager
2444 SecondaryNameNode
hdu@m2:~$
```

Cluster-2 having machines-u1,u2 with hadoop running on it

```
hdu@u1: /usr/local/hadoop
File Edit View Search Terminal Help
hdu@u1:/usr/local/hadoop$ jps
11523 NameNode
13204 Jps
10886 ResourceManager
11046 NodeManager
11709 DataNode
hdu@u1:/usr/local/hadoop$
```

```
hdu@u2: /usr/local/hadoop$ jps
2864 Jps
2244 DataNode
2748 NodeManager
2444 SecondaryNameNode
hdu@u2: /usr/local/hadoop$
```

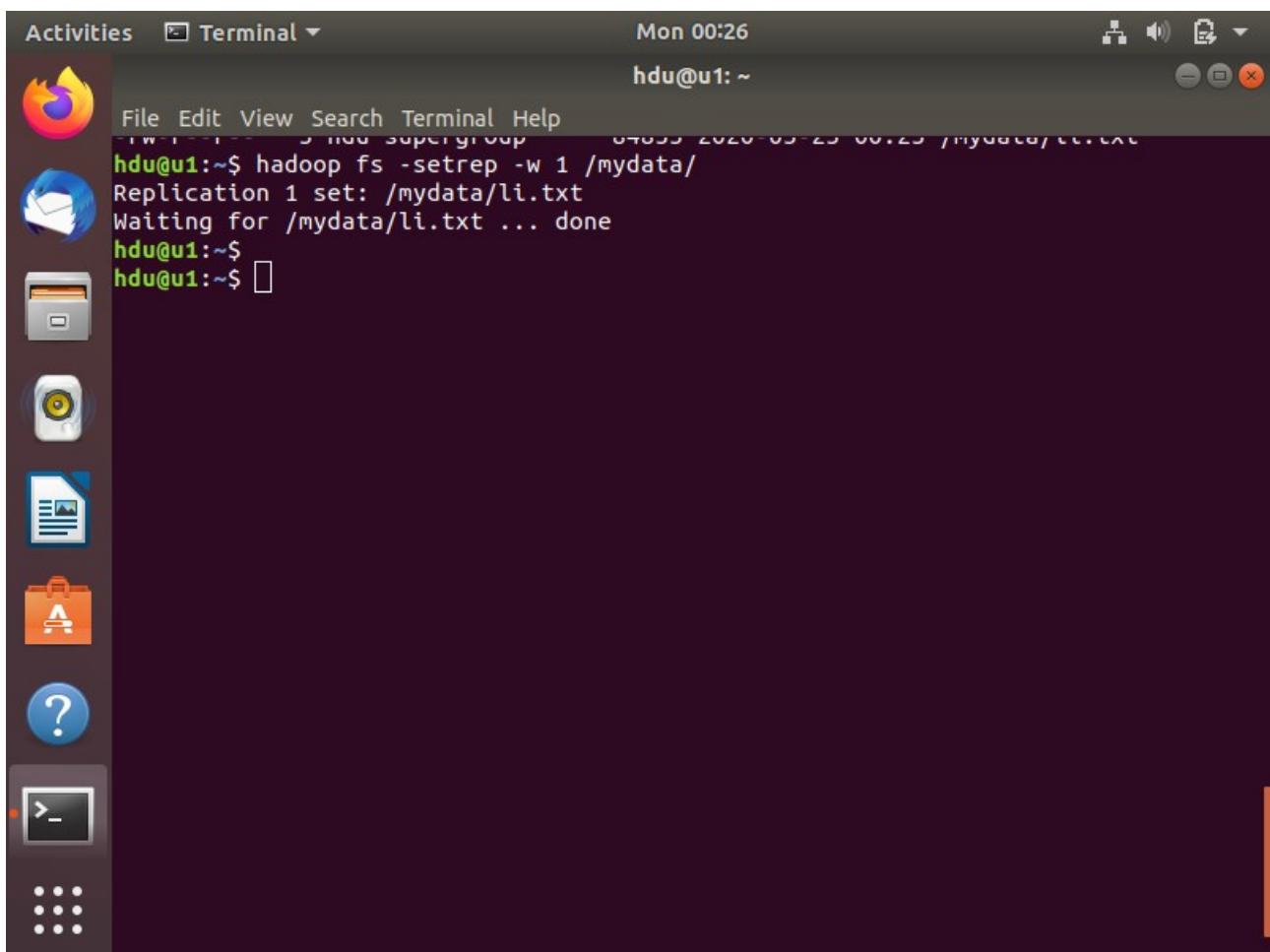
Distcp:-

```
hdu@m1:~$ hadoop distcp hdfs://m1:9000/mydata hdfs://u1:9000/mydatacopy12
20/03/21 21:52:35 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, sync
ures=false, maxMaps=20, sslConfigurationFile='null', copyStrategy='uniformsize', sourceFil
data], targetPath=hdfs://u1:9000/mydatacopy12, targetPathExists=false, preserveRawXattrs=fa
20/03/21 21:52:35 INFO client.RMPProxy: Connecting to ResourceManager at m1/192.168.56.104:
20/03/21 21:52:36 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use m
20/03/21 21:52:36 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, u
20/03/21 21:52:37 INFO client.RMPProxy: Connecting to ResourceManager at m1/192.168.56.104:
20/03/21 21:52:39 INFO mapreduce.JobSubmitter: number of splits:4
20/03/21 21:52:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_158480459898
20/03/21 21:52:41 INFO impl.YarnClientImpl: Submitted application application_158480459898
```

2. Write data to hdfs with a different replication than set in configuration file and simulate a Under replicated situation and then fix this using a HDFS command.

Ans-2

hadoop fs -setrep -w 1 /mydata



```
Activities Terminal Mon 00:26
hdu@u1: ~
File Edit View Search Terminal Help
hdu@u1:~$ hadoop fs -setrep -w 1 /mydata/
Replication 1 set: /mydata/li.txt
Waiting for /mydata/li.txt ... done
hdu@u1:~$
hdu@u1:~$
```

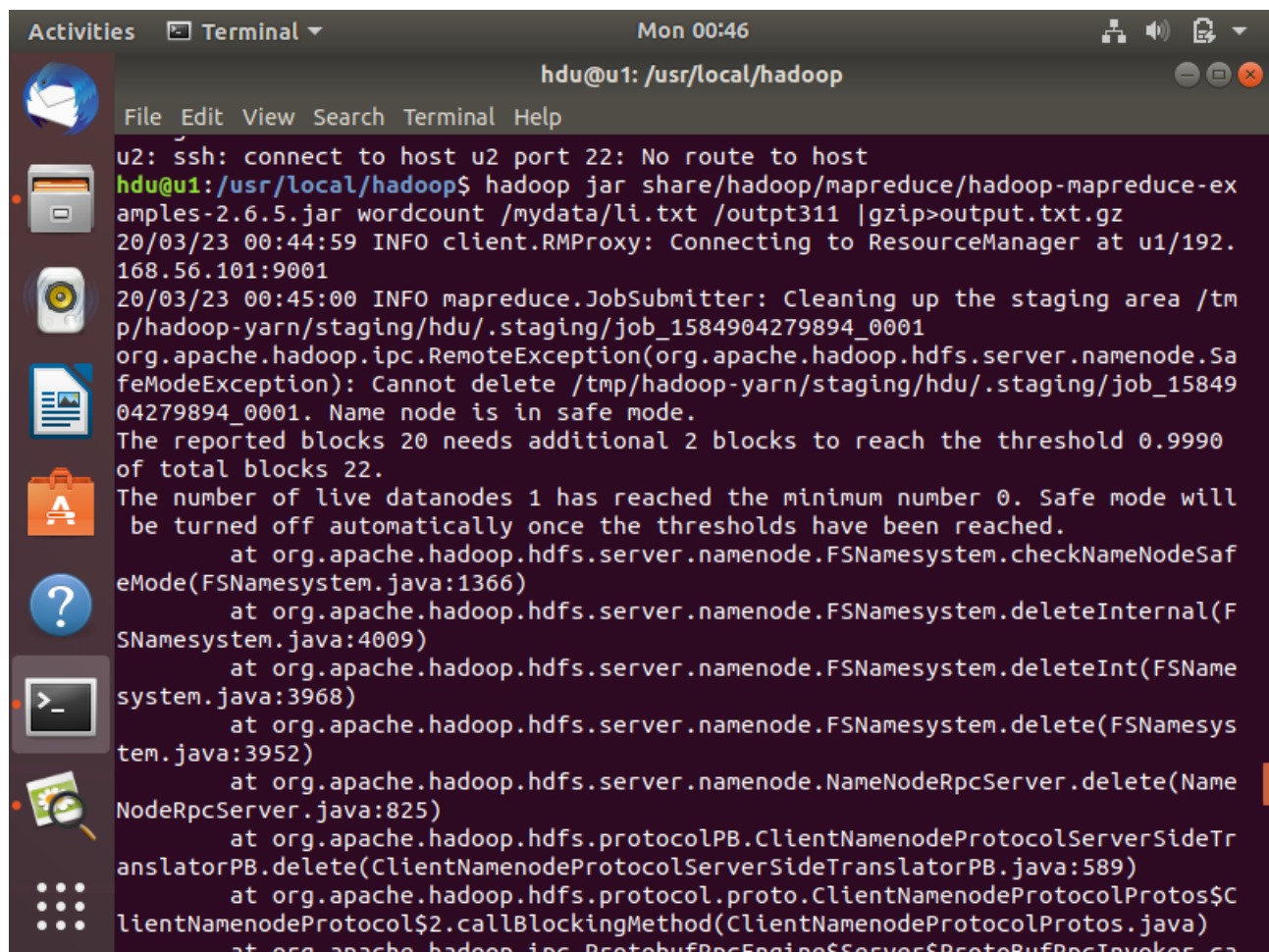
```
Activities Terminal Mon 00:33
hdu@u1: ~
File Edit View Search Terminal Help
hdu@u1:~$ hdfs fsck /
Connecting to namenode via http://u1:50070
FSCK started by hdu (auth:SIMPLE) from /192.168.56.101 for path / at Mon Mar 23
00:31:43 IST 2020
..
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.jar: Under re
licated BP-183376081-192.168.56.101-1583999280591:blk_1073741849_1025. Target
Replicas is 10 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.split: CORRUPT
blockpool BP-183376081-192.168.56.101-1583999280591 block blk_1073741850
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.split: MISSING
1 blocks of total size 189 B..
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.splitmetainfo:
Under replicated BP-183376081-192.168.56.101-1583999280591:blk_1073741851_102
7. Target Replicas is 3 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.xml: Under re
licated BP-183376081-192.168.56.101-1583999280591:blk_1073741852_1028. Target
Replicas is 3 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job_1584035052670_
0004_1_conf.xml: Under replicated BP-183376081-192.168.56.101-1583999280591:bl
k_1073741857_1033. Target Replicas is 3 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0005/job.jar: Under re
licated BP-183376081-192.168.56.101-1583999280591:blk_1073741853_1029. Target
```

```
Activities Terminal Mon 01:21
hdu@u1: /usr/local/hadoop
File Edit View Search Terminal Help
The filesystem under path '/' is CORRUPT
hdu@u1:~$ hdfs fsck / | grep 'Under replicated' | awk -F ':' '{print $1}' >> /t
mp/files
Connecting to namenode via http://u1:50070
hdu@u1:~$ hdfs fsck /
Connecting to namenode via http://u1:50070
FSCK started by hdu (auth:SIMPLE) from /192.168.56.101 for path / at Mon Mar 23
00:37:07 IST 2020
..
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.jar: Under re
licated BP-183376081-192.168.56.101-1583999280591:blk_1073741849_1025. Target
Replicas is 10 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.split: CORRUPT
blockpool BP-183376081-192.168.56.101-1583999280591 block blk_1073741850
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.split: MISSING
1 blocks of total size 189 B..
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.splitmetainfo:
Under replicated BP-183376081-192.168.56.101-1583999280591:blk_1073741851_102
7. Target Replicas is 3 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job.xml: Under re
licated BP-183376081-192.168.56.101-1583999280591:blk_1073741852_1028. Target
Replicas is 3 but found 1 replica(s).
/tmp/hadoop-yarn/staging/hdu/.staging/job_1584035052670_0004/job_1584035052670_
0004_1_conf.xml: Under replicated BP-183376081-192.168.56.101-1583999280591:bl
```

3. Demonstrate running a mapreduce job on hadoop cluster (wordcount) where the output written to HDFS should be compressed (say Gzip or Snappy). Also demonstrate running Mapreduce job in local mode instead of using YARN.

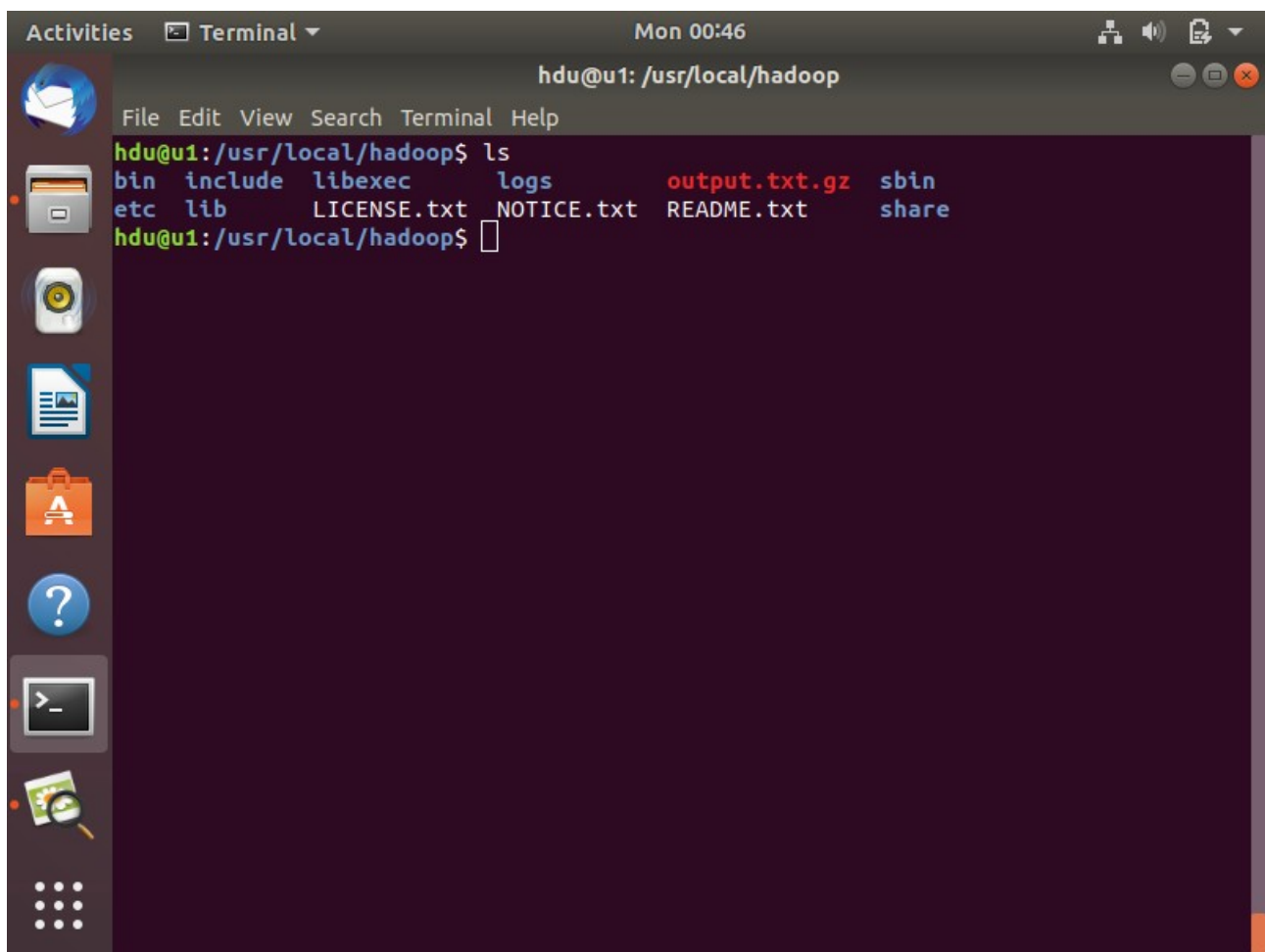
Ans-3

mapreduce of file li.txt to output.txt.gz



```
Activities Terminal Mon 00:46
hdu@u1: /usr/local/hadoop
File Edit View Search Terminal Help
u2: ssh: connect to host u2 port 22: No route to host
hdu@u1: /usr/local/hadoop$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-ex
amples-2.6.5.jar wordcount /mydata/li.txt /outpt311 |gzip>output.txt.gz
20/03/23 00:44:59 INFO client.RMProxy: Connecting to ResourceManager at u1/192.
168.56.101:9001
20/03/23 00:45:00 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tm
p/hadoop-yarn/staging/hdu/.staging/job_1584904279894_0001
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.server.namenode.Sa
feModeException): Cannot delete /tmp/hadoop-yarn/staging/hdu/.staging/job_15849
04279894_0001. Name node is in safe mode.
The reported blocks 20 needs additional 2 blocks to reach the threshold 0.9990
of total blocks 22.
The number of live datanodes 1 has reached the minimum number 0. Safe mode will
be turned off automatically once the thresholds have been reached.
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkNameNodeSaf
eMode(FSNamesystem.java:1366)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.deleteInternal(F
SNamesystem.java:4009)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.deleteInt(FSName
system.java:3968)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.delete(FSNamesys
tem.java:3952)
    at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.delete(Name
NodeRpcServer.java:825)
    at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTr
anslatorPB.delete(ClientNamenodeProtocolServerSideTranslatorPB.java:589)
    at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$C
lientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.ca
```


file created

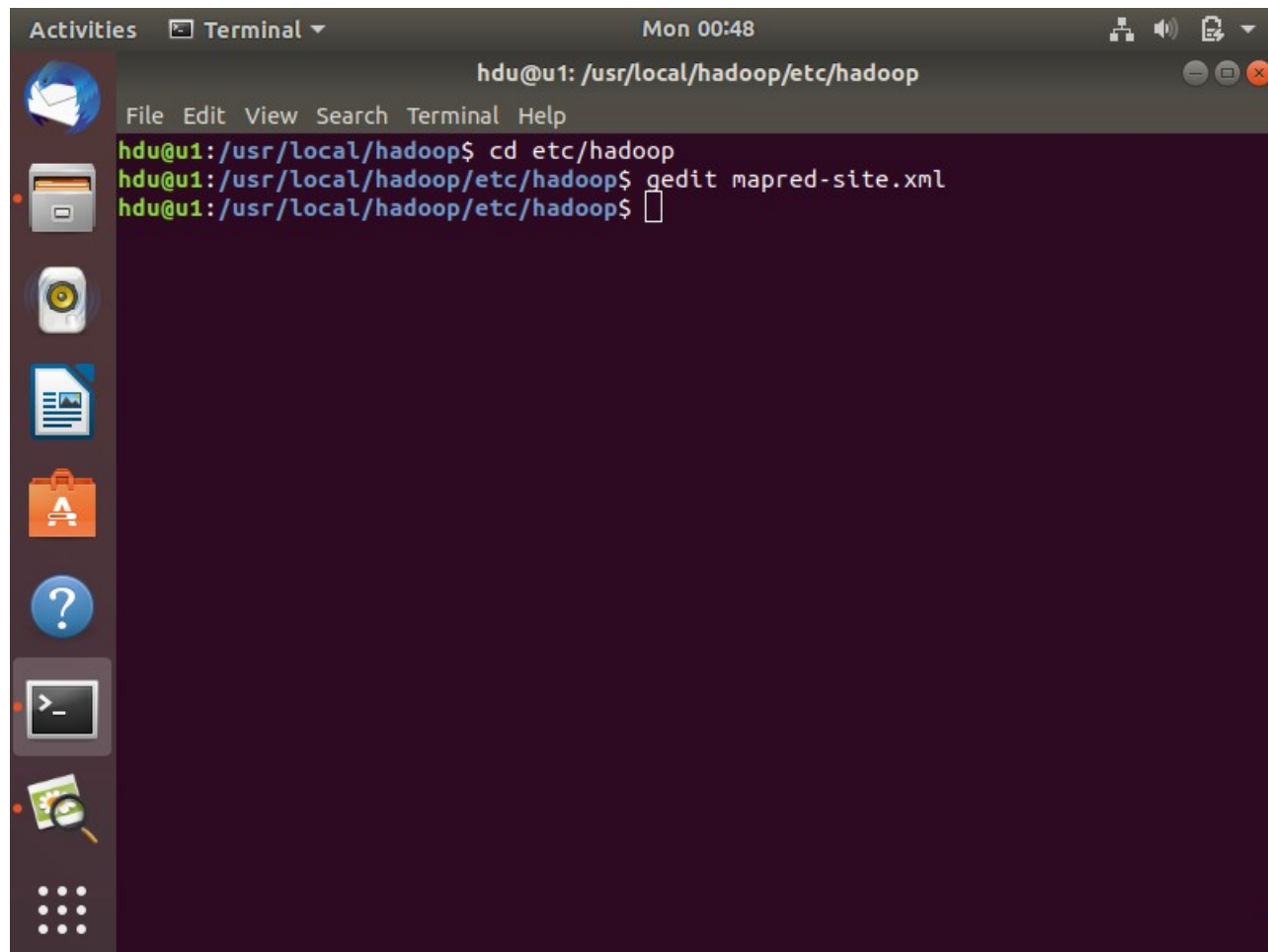


A terminal window titled "Terminal" with a subtitle "hdu@u1: /usr/local/hadoop". The window shows the output of the `ls` command, listing the contents of the `/usr/local/hadoop` directory. The output is as follows:

```
hdu@u1:/usr/local/hadoop$ ls
bin  include  libexec  logs  output.txt.gz  sbin
etc  lib      LICENSE.txt  NOTICE.txt  README.txt  share
hdu@u1:/usr/local/hadoop$
```

Also demonstrate running Mapreduce job in local mode instead of using YARN.

Editing the mapred-site.xml to run it in local mode

A screenshot of a Linux terminal window. The window title is "Terminal" and it shows the current time as "Mon 00:48". The terminal prompt is "hdu@u1: /usr/local/hadoop/etc/hadoop". The user has entered the command "cd etc/hadoop" and then "gedit mapred-site.xml". The terminal output shows the current directory as "/usr/local/hadoop/etc/hadoop". The terminal window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The terminal window is open on a desktop environment with a sidebar containing icons for "Activities", "Terminal", "Files", "Applications", "Help", "Terminal", and "Dash".

```
hdu@u1: /usr/local/hadoop/etc/hadoop
File Edit View Search Terminal Help
hdu@u1:/usr/local/hadoop$ cd etc/hadoop
hdu@u1:/usr/local/hadoop/etc/hadoop$ gedit mapred-site.xml
hdu@u1:/usr/local/hadoop/etc/hadoop$
```

changing default(yarn to local)



```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed 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

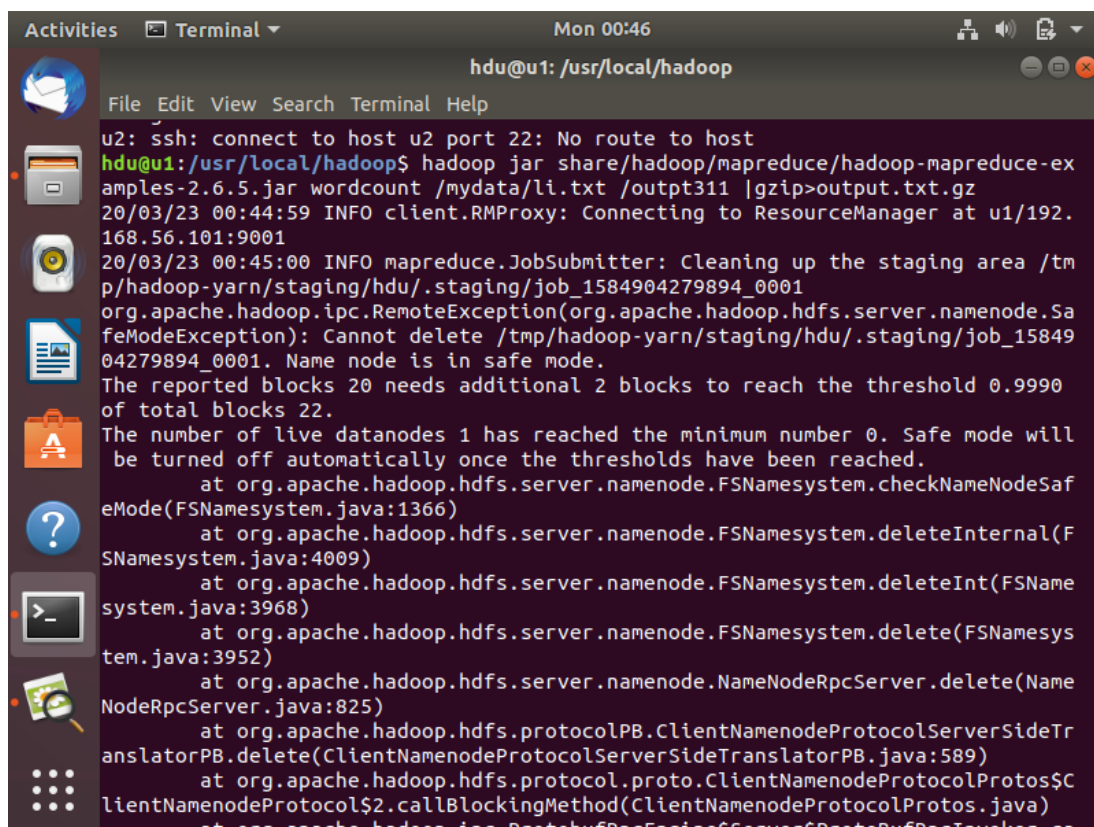
    http://www.apache.org/licenses/LICENSE-2.0

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. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>local</value>
</property>

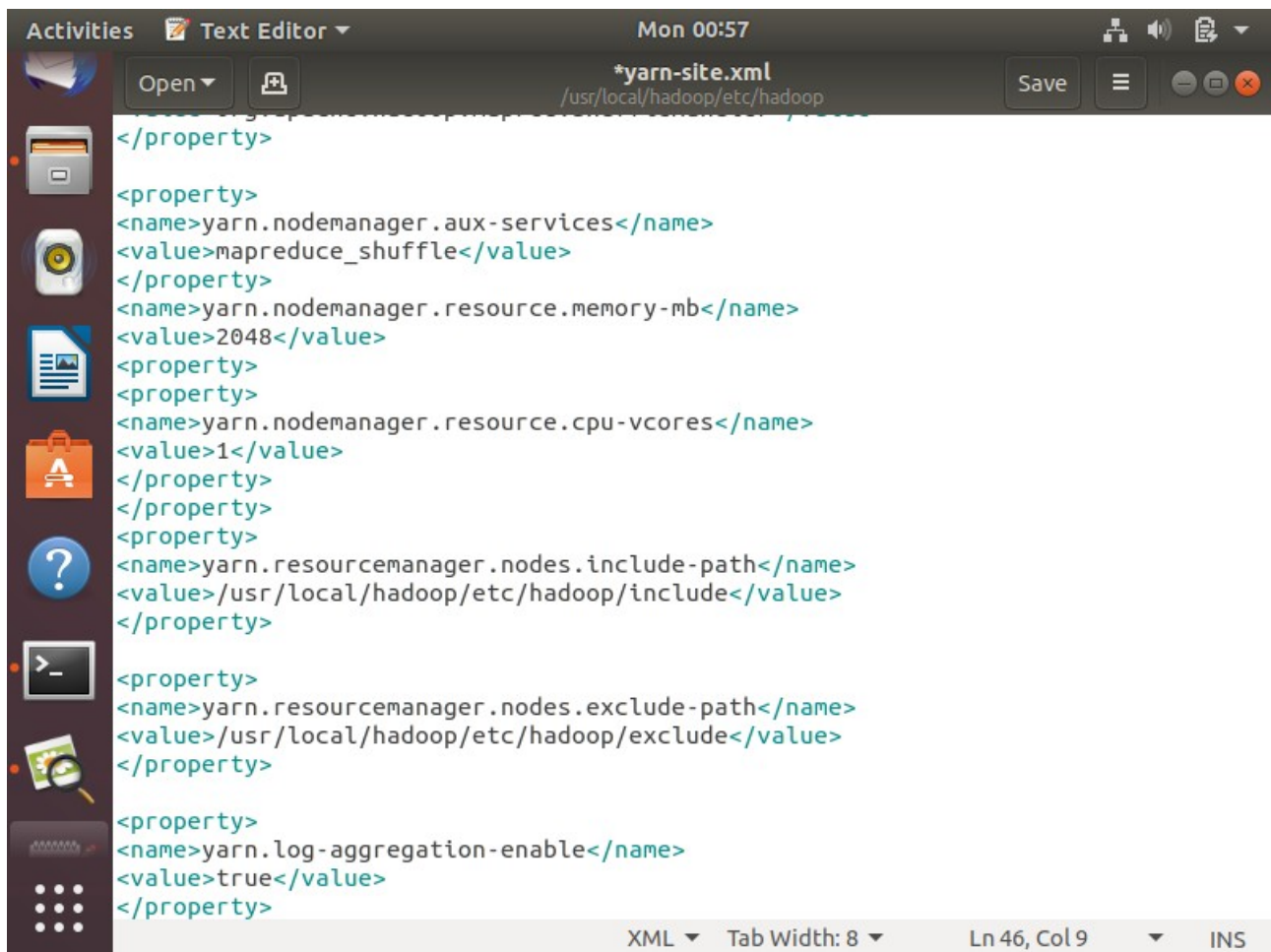
</configuration>
```



```
hdu@u1: /usr/local/hadoop
File Edit View Search Terminal Help
u2: ssh: connect to host u2 port 22: No route to host
hdu@u1: /usr/local/hadoop$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-ex
amples-2.6.5.jar wordcount /mydata/li.txt /outpt311 |gzip>output.txt.gz
20/03/23 00:44:59 INFO client.RMProxy: Connecting to Resource Manager at u1/192.
168.56.101:9001
20/03/23 00:45:00 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tm
p/hadoop-yarn/staging/hdu/.staging/job_1584904279894_0001
org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.server.namenode.Sa
feModeException): Cannot delete /tmp/hadoop-yarn/staging/hdu/.staging/job_15849
04279894_0001. Name node is in safe mode.
The reported blocks 20 needs additional 2 blocks to reach the threshold 0.9990
of total blocks 22.
The number of live datanodes 1 has reached the minimum number 0. Safe mode will
be turned off automatically once the thresholds have been reached.
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkNameNodeSaf
eMode(FSNamesystem.java:1366)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.deleteInternal(F
SNamesystem.java:4009)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.deleteInt(FSName
system.java:3968)
    at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.delete(FSNamesys
tem.java:3952)
    at org.apache.hadoop.hdfs.server.namenode.NameNodeRpcServer.delete(Name
NodeRpcServer.java:825)
    at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolServerSideTr
anslatorPB.delete(ClientNamenodeProtocolServerSideTranslatorPB.java:589)
    at org.apache.hadoop.hdfs.protocol.proto.ClientNamenodeProtocolProtos$C
lientNamenodeProtocol$2.callBlockingMethod(ClientNamenodeProtocolProtos.java)
    at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$ProtoBufRpcInvoker.ca
```

4. Assign specific amount of ram and cpu cores to Nodemanager by editing properties in YARN-site.xml. (for ex: node has 4gb ram and 2 cpu cores, then assign 2gb ram and 1 cpu core to the Nodemanager).This should show up in YARN UI.

Ans-4



```
</property>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<name>yarn.nodemanager.resource.memory-mb</name>
<value>2048</value>
<property>
<property>
<name>yarn.nodemanager.resource.cpu-vcores</name>
<value>1</value>
</property>
</property>
<property>
<name>yarn.resourcemanager.nodes.include-path</name>
<value>/usr/local/hadoop/etc/hadoop/include</value>
</property>
<property>
<name>yarn.resourcemanager.nodes.exclude-path</name>
<value>/usr/local/hadoop/etc/hadoop/exclude</value>
</property>
<property>
<name>yarn.log-aggregation-enable</name>
<value>true</value>
</property>
```


YARN-UI


Activities Firefox Web Browser Mon 01:09

Mozilla Firefox

u1:8042/node/node

u1:8042/node/node

Logged in as: dr.who



ResourceManager

NodeManager

Node Information

List of Applications

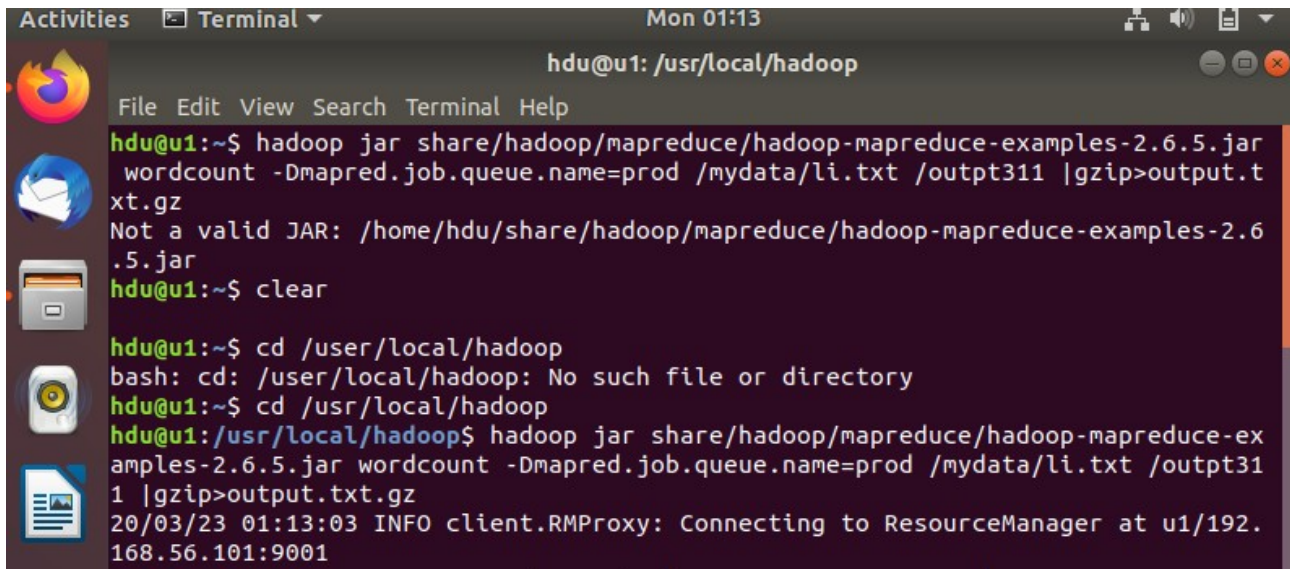
List of Containers

Tools

NodeManager information	
Total Vmem allocated for Containers	4.20 GB
Vmem enforcement enabled	true
Total Pmem allocated for Container	2 GB
Pmem enforcement enabled	true
Total VCores allocated for Containers	1
NodeHealthyStatus	true
LastNodeHealthTime	Mon Mar 23 01:08:47 IST 2020
NodeHealthReport	
Node Manager	2.6.5 from

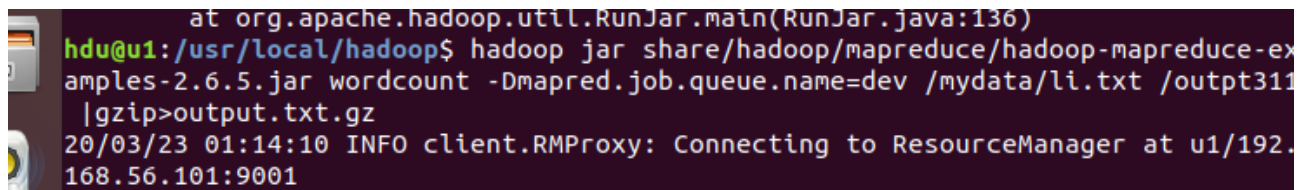
5. Demonstrate setting up of capacity scheduler with 2 queues (prod n dev) and assign 40% resources to each and run a mapreduce job in each queue.

Ans-5



A terminal window titled 'Terminal' with a dark background. The prompt is 'hdu@u1: /usr/local/hadoop'. The user enters a Hadoop command to run a wordcount job with a specific queue name. An error message appears: 'Not a valid JAR: /home/hdu/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar'. The user then clears the screen and navigates to the correct directory. The command is re-executed, and a log message is visible at the bottom: '20/03/23 01:13:03 INFO client.RMProxy: Connecting to ResourceManager at u1/192.168.56.101:9001'.

```
hdu@u1: /usr/local/hadoop
File Edit View Search Terminal Help
hdu@u1:~$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar
wordcount -Dmapred.job.queue.name=prod /mydata/li.txt /outpt311 |gzip>output.t
xt.gz
Not a valid JAR: /home/hdu/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6
.5.jar
hdu@u1:~$ clear
hdu@u1:~$ cd /user/local/hadoop
bash: cd: /user/local/hadoop: No such file or directory
hdu@u1:~$ cd /usr/local/hadoop
hdu@u1:/usr/local/hadoop$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-ex
amples-2.6.5.jar wordcount -Dmapred.job.queue.name=prod /mydata/li.txt /outpt31
1 |gzip>output.txt.gz
20/03/23 01:13:03 INFO client.RMProxy: Connecting to ResourceManager at u1/192.
168.56.101:9001
```



A terminal window showing the execution of a Hadoop wordcount job for the 'dev' queue. The command is: 'hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar wordcount -Dmapred.job.queue.name=dev /mydata/li.txt /outpt311 |gzip>output.txt.gz'. A log message is visible at the bottom: '20/03/23 01:14:10 INFO client.RMProxy: Connecting to ResourceManager at u1/192.168.56.101:9001'.

```
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
hdu@u1:/usr/local/hadoop$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-ex
amples-2.6.5.jar wordcount -Dmapred.job.queue.name=dev /mydata/li.txt /outpt311
|gzip>output.txt.gz
20/03/23 01:14:10 INFO client.RMProxy: Connecting to ResourceManager at u1/192.
168.56.101:9001
```

Activities

Firefox Web Browser

Thu 23:46

NEW,NEW_SAVING,SUBMITTED,ACCEPTED,RUNNING Applications - Mozilla Firefox

NEW,NEW_SAVING,SUBMIT X


+

← → ↺ 🏠

u2:8088/cluster/scheduler

… 🛡️ ☆

📄 📁 👤 ☰

NEW,NEW_SAVING,SUBM

Cluster

About

Nodes

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	M
2	0	1	1	1	2 GB	2

Application Queues

Legend: Capacity Used Used (over capacity)

root

default

dev

prod

Show 20 entries

ID	User	Name	Application Type	Q
----	------	------	------------------	---