Exploring the Hadoop Ecosystem through Linux commands

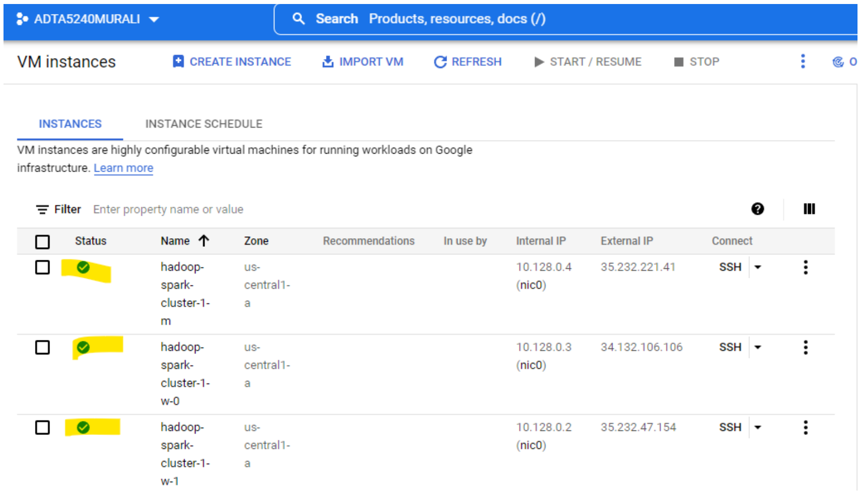
We will access the Hadoop ecosystem using Linuz commands through Dataproc on Google Cloud Platform.

You must have your GCP account set up and ready with the following:

1) Active GCP account
2) A project on Hadoop-Spark cluster
3) A storage bucket ready for use (we added user dataset, refer to previous document on how to create clusters and bucket)
4) The cluster must contain 1 master node and 2 worker nodes. (Make sure to turn on/off during and after completion of assignment-refer to previous document to start and stop clusters)

TURN ON ALL CLUSTERS

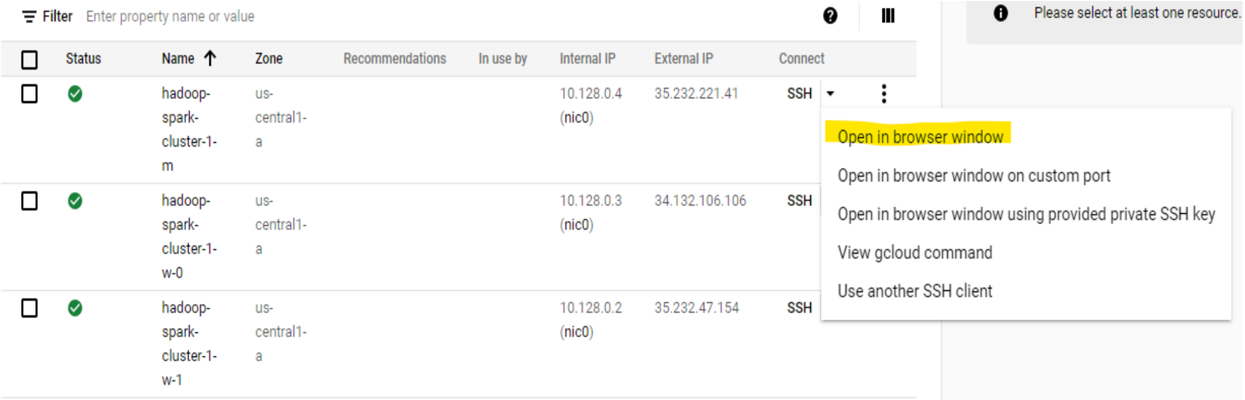
1) Start all the nodes in the cluster. Your window must look like image below.



VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

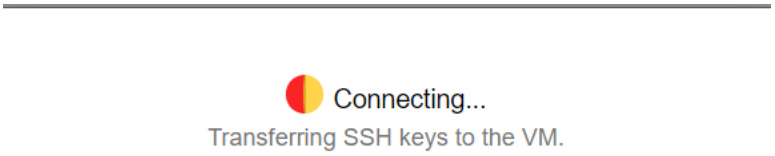
□	Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
□	✓	hadoop-spark-cluster-1-m	us-central1-a			10.128.0.4 (nic0)	35.232.221.41	SSH
□	✓	hadoop-spark-cluster-1-w-0	us-central1-a			10.128.0.3 (nic0)	34.132.106.106	SSH
□	✓	hadoop-spark-cluster-1-w-1	us-central1-a			10.128.0.2 (nic0)	35.232.47.154	SSH

2) Click on the SSH dropdown menu and select “Open in browser window”.


Filter Enter property name or value
Please select at least one resource.
Open in browser window
Open in browser window on custom port
Open in browser window using provided private SSH key
View gcloud command
Use another SSH client

□	Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
□	✓	hadoop-spark-cluster-1-m	us-central1-a			10.128.0.4 (nic0)	35.232.221.41	SSH
□	✓	hadoop-spark-cluster-1-w-0	us-central1-a			10.128.0.3 (nic0)	34.132.106.106	SSH
□	✓	hadoop-spark-cluster-1-w-1	us-central1-a			10.128.0.2 (nic0)	35.232.47.154	SSH

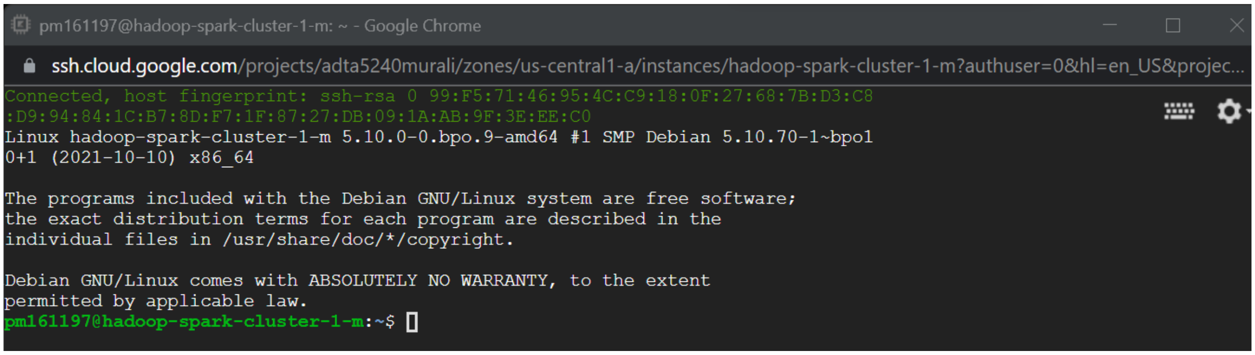
3) It will ask you to establish connection with VM. Select “Connect”



Connecting...

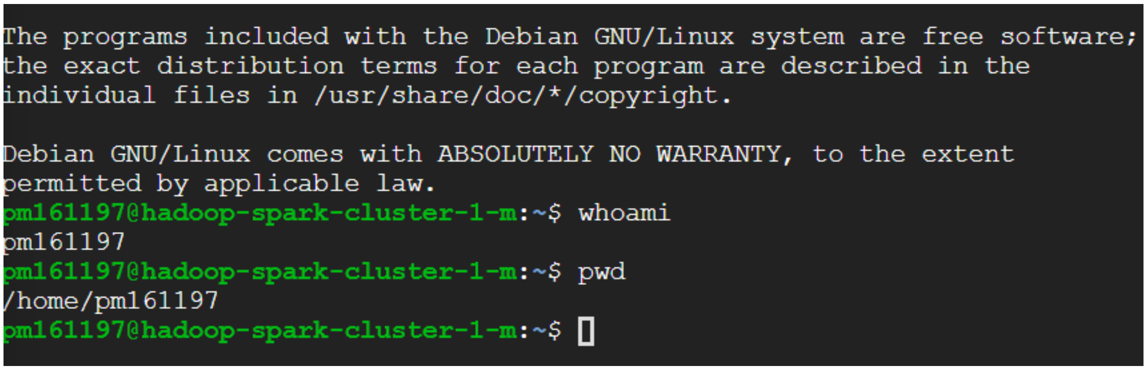
Transferring SSH keys to the VM.

4) It should open a terminal that looks like the below picture.


pm161197@hadoop-spark-cluster-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/adta5240murali/zones/us-central1-a/instances/hadoop-spark-cluster-1-m?authuser=0&hl=en_US&projec...

```
Connected, host fingerprint: ssh-rsa 0 99:F5:71:46:95:4C:C9:18:0F:27:68:7B:D3:C8
:D9:94:84:1C:B7:8D:F7:1F:87:27:DB:09:1A:AB:9F:3E:EE:C0
Linux hadoop-spark-cluster-1-m 5.10.0-0.bpo.9-amd64 #1 SMP Debian 5.10.70-1~bpo1
0+1 (2021-10-10) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
pm161197@hadoop-spark-cluster-1-m:~$ 
```

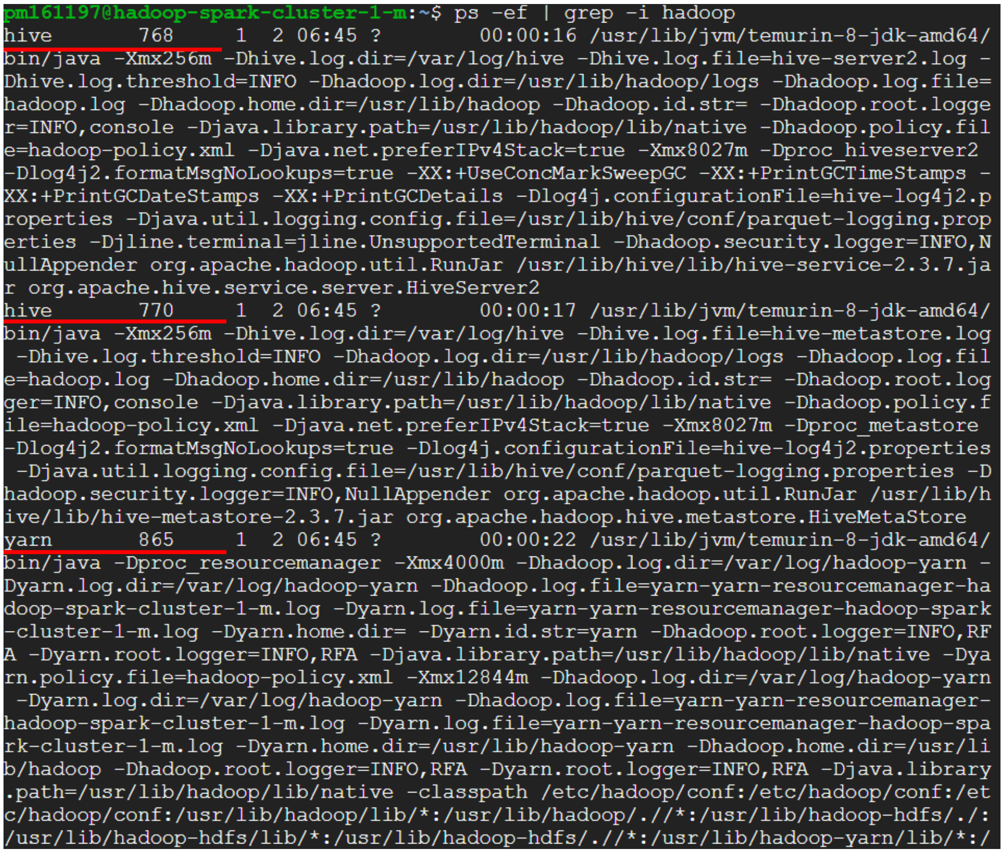
5) Type the following – whoami and pwd one after the other.It will show you the username and working directory as shown below.



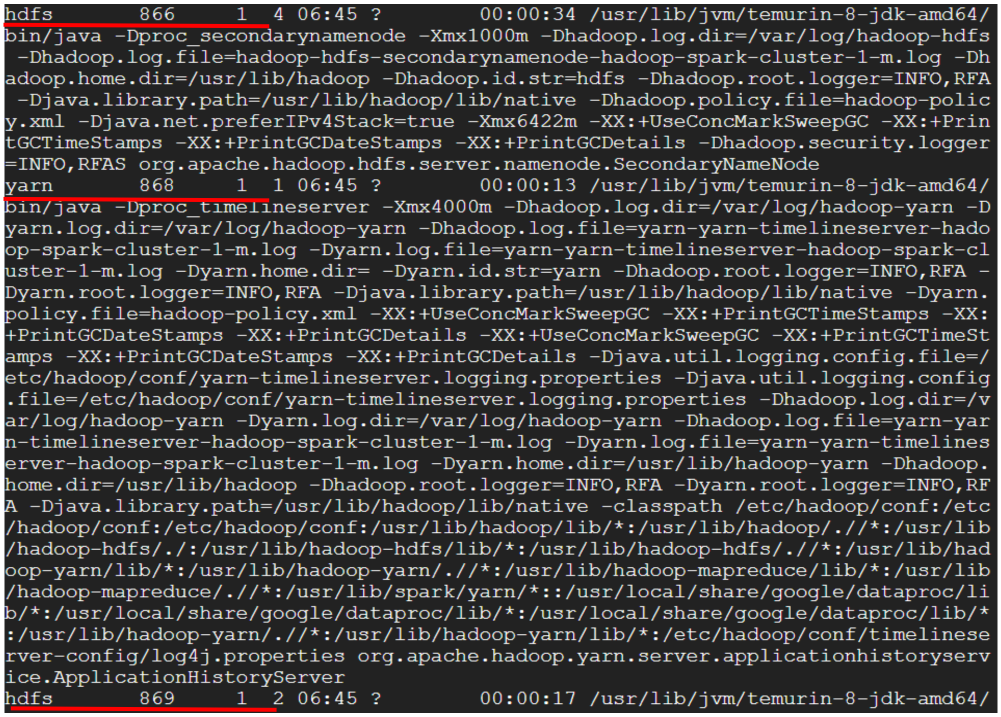
```
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
pm161197@hadoop-spark-cluster-1-m:~$ whoami
pm161197
pm161197@hadoop-spark-cluster-1-m:~$ pwd
/home/pm161197
pm161197@hadoop-spark-cluster-1-m:~$ 
```

6) Type the following: ps -ef | grep -i Hadoop

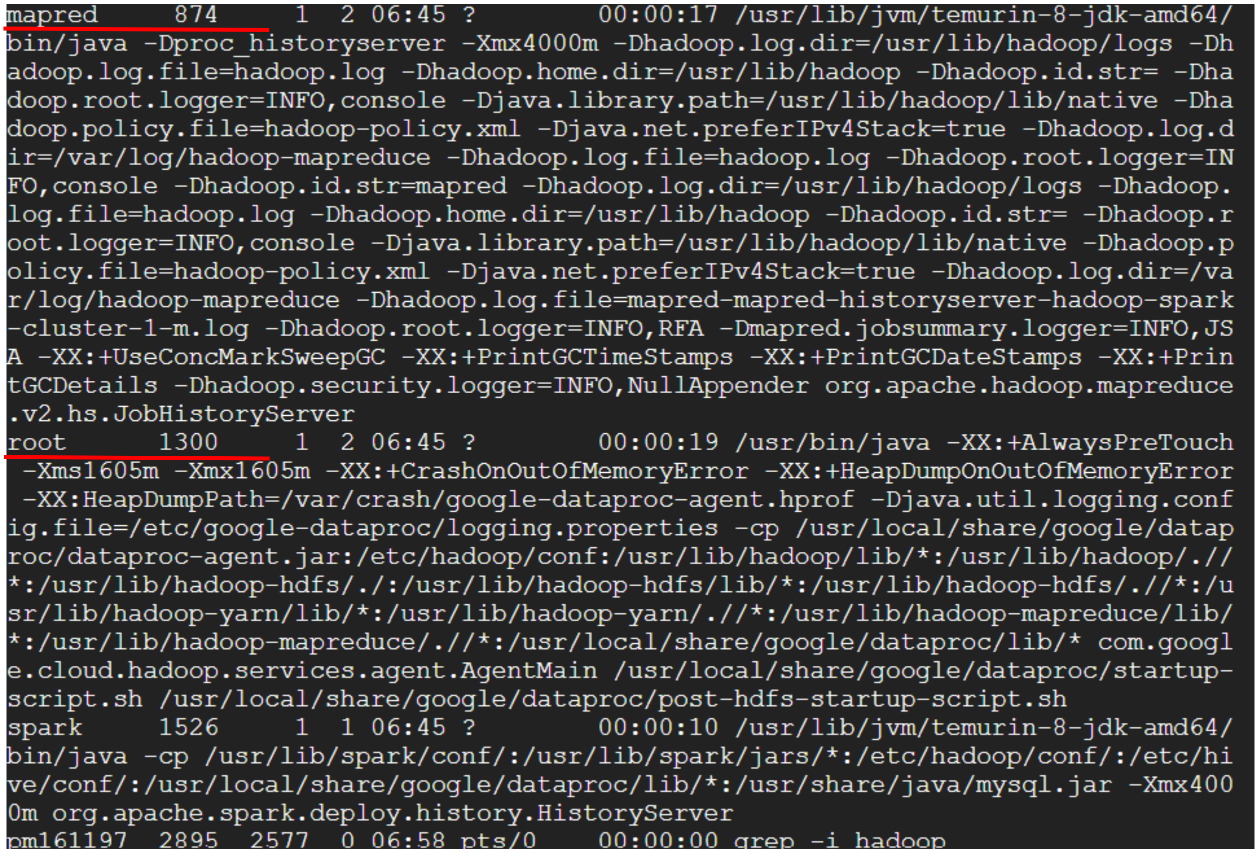
EXTRA – ps stands for process status, ps -e is used to select all and ps -f provides full information about the process. When written together it shows all the processes with corresponding information.



```
pm161197@hadoop-spark-cluster-1-m:~$ ps -ef | grep -i hadoop
hive         768       1  2 06:45 ?        00:00:16 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -Xmx256m -Dhive.log.dir=/var/log/hive -Dhive.log.file=hive-server2.log -
Dhive.log.threshold=INFO -Dhadoop.log.dir=/usr/lib/hadoop/logs -Dhadoop.log.file=
hadoop.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str= -Dhadoop.root.logge
r=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.policy.fil
e=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -Xmx8027m -Dproc_hiveserver2
-Dlog4j2.formatMsgNoLookups=true -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -
XX:+PrintGCDateStamps -XX:+PrintGCDetails -Dlog4j.configurationFile=hive-log4j2.p
roperties -Djava.util.logging.config.file=/usr/lib/hive/conf/parquet-logging.prop
erties -Djline.terminal=jline.UnsupportedTerminal -Dhadoop.security.logger=INFO,N
ullAppender org.apache.hadoop.util.RunJar /usr/lib/hive/lib/hive-service-2.3.7.ja
r org.apache.hive.service.server.HiveServer2
hive         770       1  2 06:45 ?        00:00:17 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -Xmx256m -Dhive.log.dir=/var/log/hive -Dhive.log.file=hive-metastore.log
 -Dhive.log.threshold=INFO -Dhadoop.log.dir=/usr/lib/hadoop/logs -Dhadoop.log.fil
e=hadoop.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str= -Dhadoop.root.log
ger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.policy.f
ile=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -Xmx8027m -Dproc_metastore
-Dlog4j2.formatMsgNoLookups=true -Dlog4j.configurationFile=hive-log4j2.properties
 -Djava.util.logging.config.file=/usr/lib/hive/conf/parquet-logging.properties -D
hadoop.security.logger=INFO,NullAppender org.apache.hadoop.util.RunJar /usr/lib/h
ive/lib/hive-metastore-2.3.7.jar org.apache.hadoop.hive.metastore.HiveMetaStore
yarn         865       1  2 06:45 ?        00:00:22 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -Dproc_resourcemanager -Xmx4000m -Dhadoop.log.dir=/var/log/hadoop-yarn -
Dyarn.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=yarn-yarn-resourcemanager-ha
doop-spark-cluster-1-m.log -Dyarn.log.file=yarn-yarn-resourcemanager-hadoop-spark
-cluster-1-m.log -Dyarn.home.dir= -Dyarn.id.str=yarn -Dhadoop.root.logger=INFO,RF
A -Dyarn.root.logger=INFO,RFA -Djava.library.path=/usr/lib/hadoop/lib/native -Dya
rn.policy.file=hadoop-policy.xml -Xmx12844m -Dhadoop.log.dir=/var/log/hadoop-yarn
 -Dyarn.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=yarn-yarn-resourcemanager-
hadoop-spark-cluster-1-m.log -Dyarn.log.file=yarn-yarn-resourcemanager-hadoop-spa
rk-cluster-1-m.log -Dyarn.home.dir=/usr/lib/hadoop-yarn -Dhadoop.home.dir=/usr/li
b/hadoop -Dhadoop.root.logger=INFO,RFA -Dyarn.root.logger=INFO,RFA -Djava.library
.path=/usr/lib/hadoop/lib/native -classpath /etc/hadoop/conf:/etc/hadoop/conf:/et
c/hadoop/conf:/usr/lib/hadoop/lib/*:/usr/lib/hadoop/.//*:/usr/lib/hadoop-hdfs/./:
/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/.//*:/usr/lib/hadoop-yarn/lib/*:/
```



```
hdfs         866       1  4 06:45 ?        00:00:34 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -Dproc_secondarynamenode -Xmx1000m -Dhadoop.log.dir=/var/log/hadoop-hdfs
 -Dhadoop.log.file=hadoop-hdfs-secondarynamenode-hadoop-spark-cluster-1-m.log -Dh
adoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=hdfs -Dhadoop.root.logger=INFO,RFA
 -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.policy.file=hadoop-polic
y.xml -Djava.net.preferIPv4Stack=true -Xmx6422m -XX:+UseConcMarkSweepGC -XX:+Prin
tGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -Dhadoop.security.logger
=INFO,RFAS org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
yarn         868       1  1 06:45 ?        00:00:13 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -Dproc_timelineserver -Xmx4000m -Dhadoop.log.dir=/var/log/hadoop-yarn -D
yarn.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=yarn-yarn-timelineserver-hado
op-spark-cluster-1-m.log -Dyarn.log.file=yarn-yarn-timelineserver-hadoop-spark-cl
uster-1-m.log -Dyarn.home.dir= -Dyarn.id.str=yarn -Dhadoop.root.logger=INFO,RFA -
Dyarn.root.logger=INFO,RFA -Djava.library.path=/usr/lib/hadoop/lib/native -Dyarn.
policy.file=hadoop-policy.xml -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:
+PrintGCDateStamps -XX:+PrintGCDetails -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeSt
amps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -Djava.util.logging.config.file=/
etc/hadoop/conf/yarn-timelineserver.logging.properties -Djava.util.logging.config
.file=/etc/hadoop/conf/yarn-timelineserver.logging.properties -Dhadoop.log.dir=/v
ar/log/hadoop-yarn -Dyarn.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=yarn-yar
n-timelineserver-hadoop-spark-cluster-1-m.log -Dyarn.log.file=yarn-yarn-timelines
erver-hadoop-spark-cluster-1-m.log -Dyarn.home.dir=/usr/lib/hadoop-yarn -Dhadoop.
home.dir=/usr/lib/hadoop -Dhadoop.root.logger=INFO,RFA -Dyarn.root.logger=INFO,RF
A -Djava.library.path=/usr/lib/hadoop/lib/native -classpath /etc/hadoop/conf:/etc
/hadoop/conf:/etc/hadoop/conf:/usr/lib/hadoop/lib/*:/usr/lib/hadoop/.//*:/usr/lib
/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/.//*:/usr/lib/had
oop-yarn/lib/*:/usr/lib/hadoop-yarn/.//*:/usr/lib/hadoop-mapreduce/lib/*:/usr/lib
/hadoop-mapreduce/.//*:/usr/lib/spark/yarn/*::/usr/local/share/google/dataproc/li
b/*:/usr/local/share/google/dataproc/lib/*:/usr/local/share/google/dataproc/lib/*
:/usr/lib/hadoop-yarn/.//*:/usr/lib/hadoop-yarn/lib/*:/etc/hadoop/conf/timelinese
rver-config/log4j.properties org.apache.hadoop.yarn.server.applicationhistoryserv
ice.ApplicationHistoryServer
hdfs         869       1  2 06:45 ?        00:00:17 /usr/lib/jvm/temurin-8-jdk-amd64/
```



```
mapred      874       1  2 06:45 ?        00:00:17 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -Dproc_historyserver -Xmx4000m -Dhadoop.log.dir=/usr/lib/hadoop/logs -Dh
adoop.log.file=hadoop.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str= -Dha
doop.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dha
doop.policy.file=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -Dhadoop.log.d
ir=/var/log/hadoop-mapreduce -Dhadoop.log.file=hadoop.log -Dhadoop.root.logger=IN
FO,console -Dhadoop.id.str=mapred -Dhadoop.log.dir=/usr/lib/hadoop/logs -Dhadoop.
log.file=hadoop.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str= -Dhadoop.r
oot.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.p
olicy.file=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -Dhadoop.log.dir=/va
r/log/hadoop-mapreduce -Dhadoop.log.file=mapred-mapred-historyserver-hadoop-spark
-cluster-1-m.log -Dhadoop.root.logger=INFO,RFA -Dmapred.jobsummary.logger=INFO,JS
A -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+Prin
tGCDetails -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.mapreduce
.v2.hs.JobHistoryServer
root       1300       1  2 06:45 ?        00:00:19 /usr/bin/java -XX:+AlwaysPreTouch
 -Xms1605m -Xmx1605m -XX:+CrashOnOutOfMemoryError -XX:+HeapDumpOnOutOfMemoryError
 -XX:HeapDumpPath=/var/crash/google-dataproc-agent.hprof -Djava.util.logging.conf
ig.file=/etc/google-dataproc/logging.properties -cp /usr/local/share/google/datap
roc/dataproc-agent.jar:/etc/hadoop/conf:/usr/lib/hadoop/lib/*:/usr/lib/hadoop/.//
*:/usr/lib/hadoop-hdfs/./:/usr/lib/hadoop-hdfs/lib/*:/usr/lib/hadoop-hdfs/.//*:/u
sr/lib/hadoop-yarn/lib/*:/usr/lib/hadoop-yarn/.//*:/usr/lib/hadoop-mapreduce/lib/
*:/usr/lib/hadoop-mapreduce/.//*:/usr/local/share/google/dataproc/lib/* com.googl
e.cloud.hadoop.services.agent.AgentMain /usr/local/share/google/dataproc/startup-
script.sh /usr/local/share/google/dataproc/post-hdfs-startup-script.sh
spark      1526       1  1 06:45 ?        00:00:10 /usr/lib/jvm/temurin-8-jdk-amd64/
bin/java -cp /usr/lib/spark/conf/:/usr/lib/spark/jars/*:/etc/hadoop/conf/:/etc/hi
ve/conf/:/usr/local/share/google/dataproc/lib/*:/usr/share/java/mysql.jar -Xmx400
0m org.apache.spark.deploy.history.HistoryServer
pm161197   2895    2577  0 06:58 pts/0    00:00:00 grep -i hadoop
```

7) **TIP:** To access previous commands click on the up key on the keyboard.

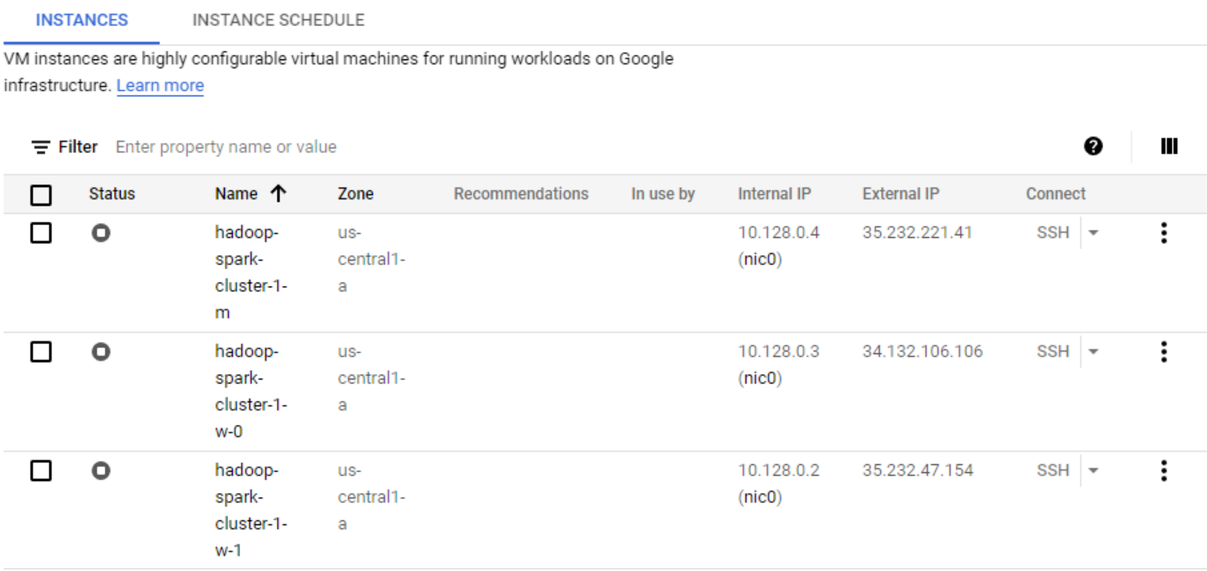
8) To access certain documentation like “How to copy/paste”, click on the settings icon on the top right corner of the command window.

What do the numbers mean in the above snippet?

It is used to unique identify an active process.

Process Name	**Process ID**	**Node name**
Hive	768	HiveServer2
Hive	770	HiveServer2
Yarn	865	ResourceManager
HDFS	866	SecondaryNameNode
Yarn	868	ApplicationHistoryServer
HDFS	869	NameNode
Root	1300	HistoryServer
Mapred	874	JobHistoryServer

TURN OFF YOUR CLUSTERS!



VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

□	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect	
□	○	hadoop-spark-cluster-1-m	us-central1-a			10.128.0.4 (nic0)	35.232.221.41	SSH ▾	⋮
□	○	hadoop-spark-cluster-1-w-0	us-central1-a			10.128.0.3 (nic0)	34.132.106.106	SSH ▾	⋮
□	○	hadoop-spark-cluster-1-w-1	us-central1-a			10.128.0.2 (nic0)	35.232.47.154	SSH ▾	⋮