# INSTRUCTIONAL MATERIAL

## DEPARTMENT OF INFORMATION TECHNOLOGY

## CLOUD BASED MULTINODE CLUSTER SETUP USING HADOOP

**Dr. R. BEULAH JAYAKUMARI**
Professor and Head
Department of Information Technology


**TAGORE ENGINEERING COLLEGE**
Rathinamangalam, Chennai-600 127

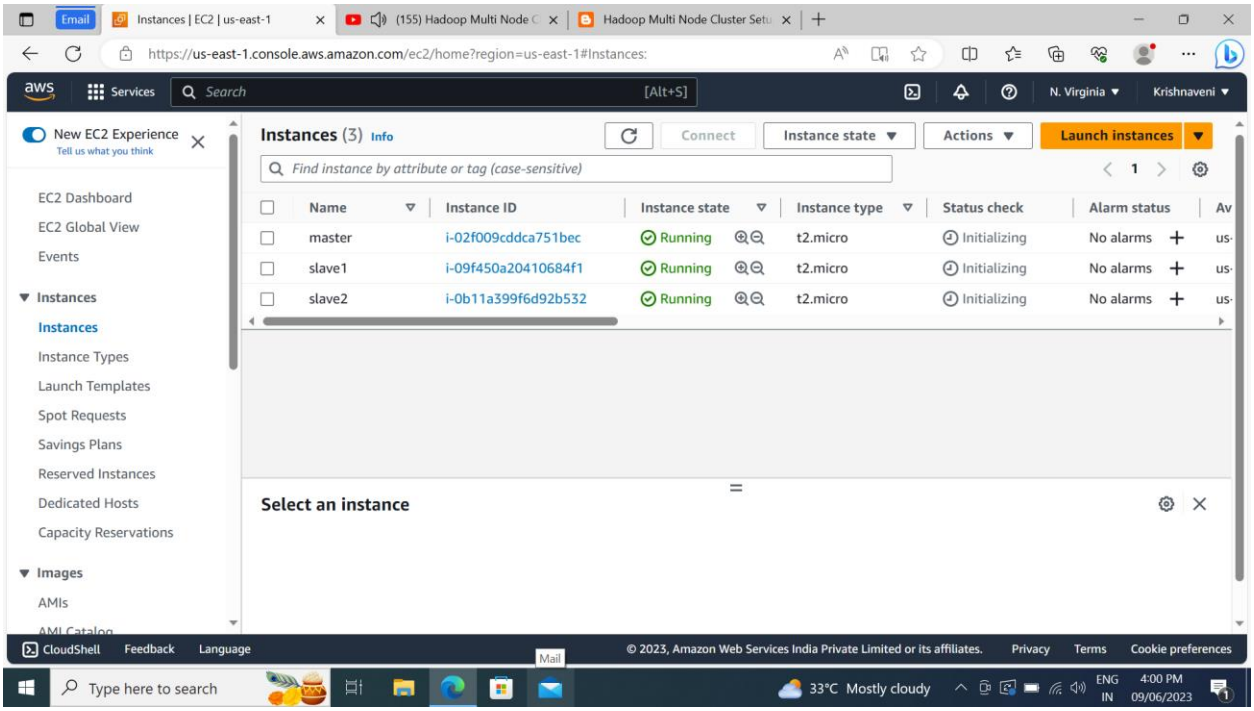
# CLOUD BASED MULTINODE CLUSTER SET UP USING HADOOP

Step 1: Create 3 EC2 instances with ubuntu OS in amazon web services named master, slave1,slave2.

**Master:** PublicIPs: 34.230.55.172 PrivateIPs: 172.31.37.21

**Slave1:** PublicIPs: 3.85.215.222 PrivateIPs: 172.31.38.178

**Slave2:** PublicIPs: 52.90.103.202 PrivateIPs: 172.31.44.172




Step2: connect the instances and execute the following commands in all nodes

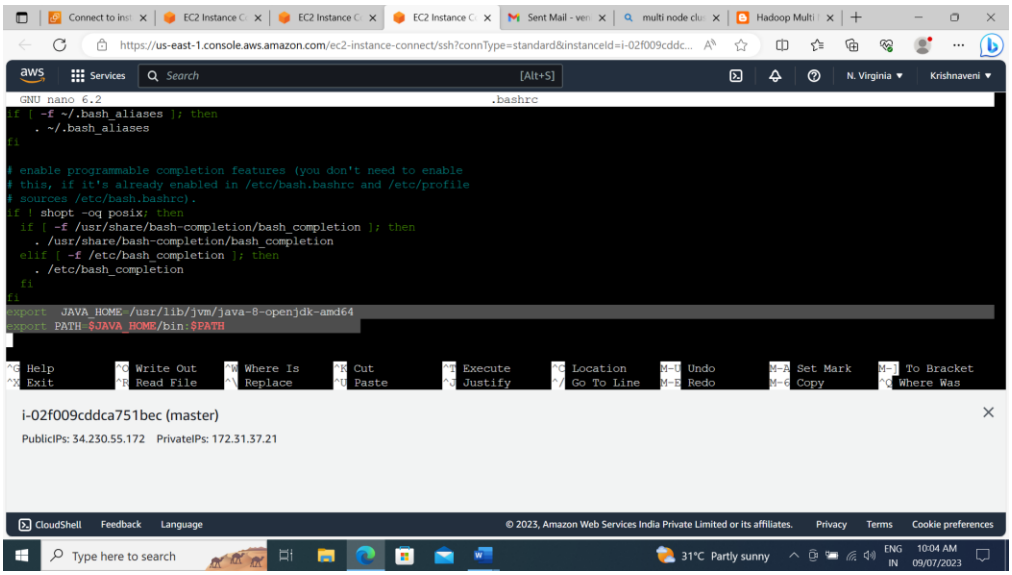
Installing java and hadoop in all nodes

1. Sudo apt update -y
2. sudo apt install openjdk-8-jdk -y
3. sudo wget https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-3.3.1.tar.gz
4. tar -xvzf hadoop-3.3.1.tar.gz

open .bashrc file in all nodes and edit add the below content

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$JAVA_HOME/bin:$PATH
```

5. sudo nano .bashrc



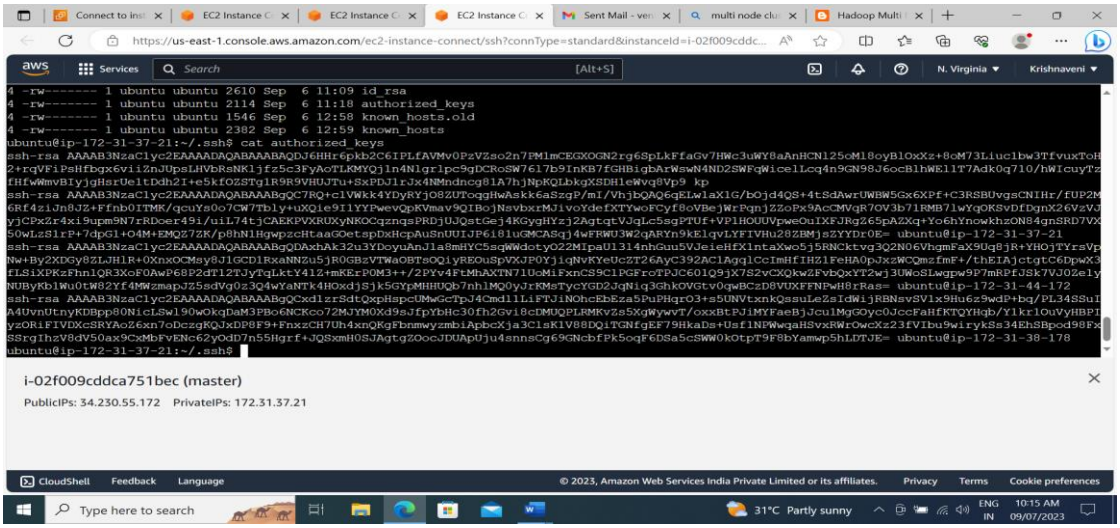

6.source .bashrc

7. ssh-keygen -t rsa

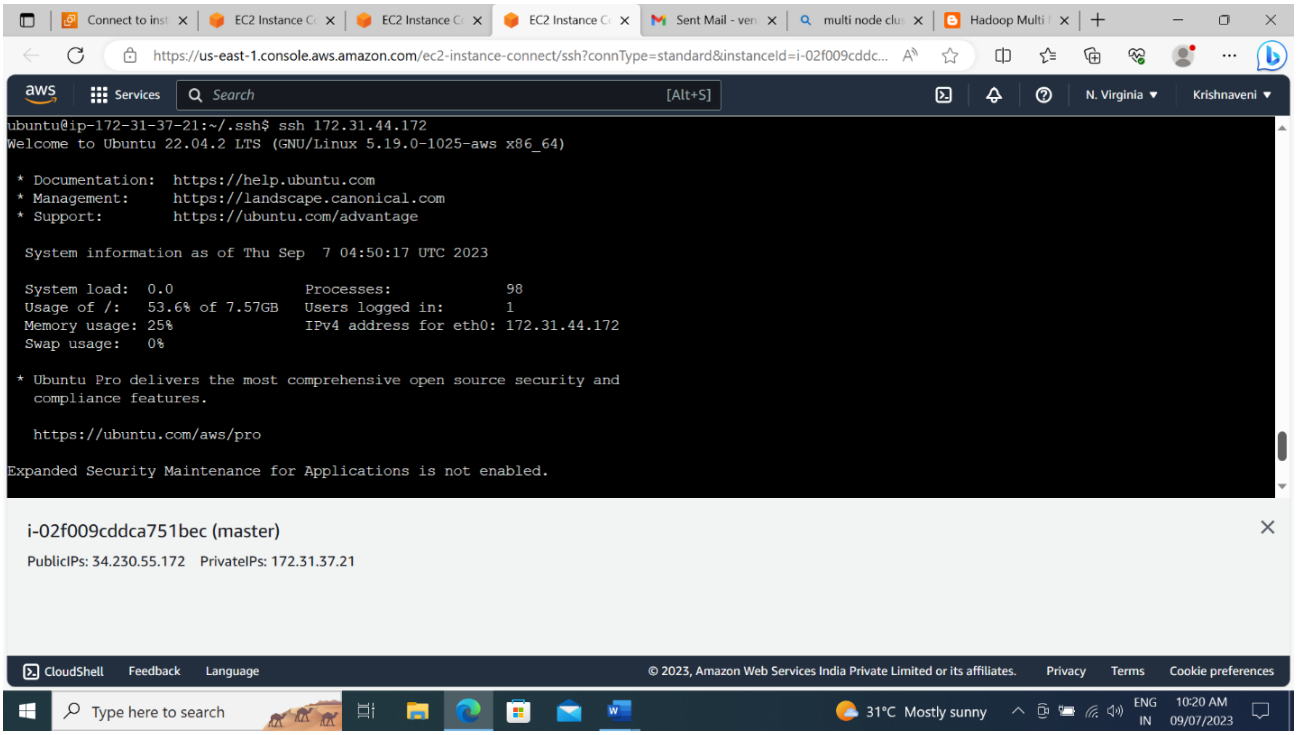
Generate the key (ssh-keygen -t rsa) - 4 times press enter

Generate the ssh key in all nodes and copy the keys of all nodes in .ssh/authorized_keys

Do the same step for all nodes



Ensure for password less connection through ssh




8. ssh 172.31.44.172

Master node is connected with slave node withoutpassword

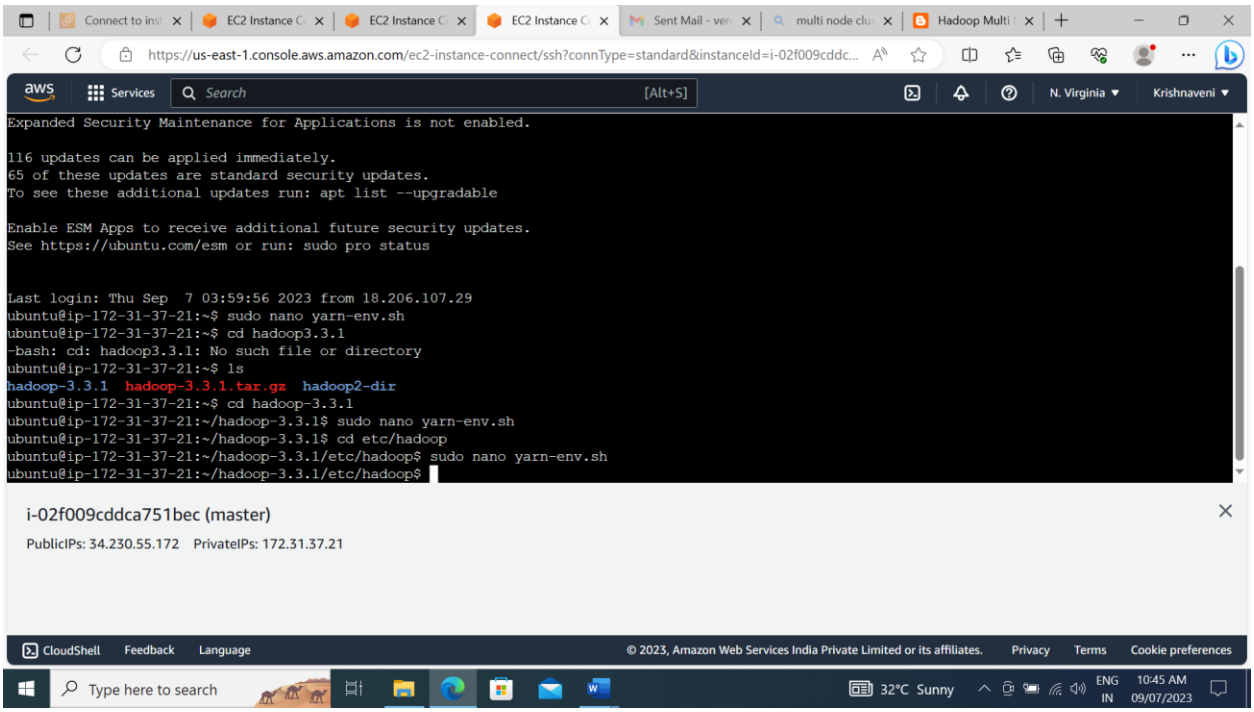
9.sudo nano mapred-env.sh

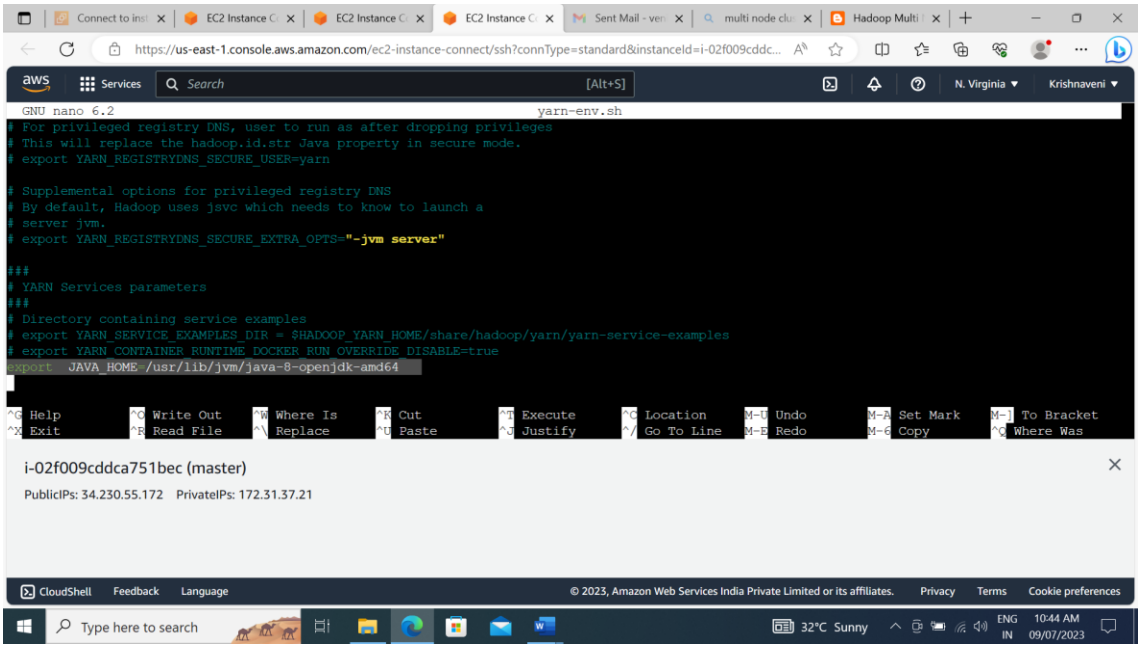
10. sudo nano hadoop-env.sh

11. sudo nano yarn-env.sh

Open all the files in all nodes and add the java home path

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64



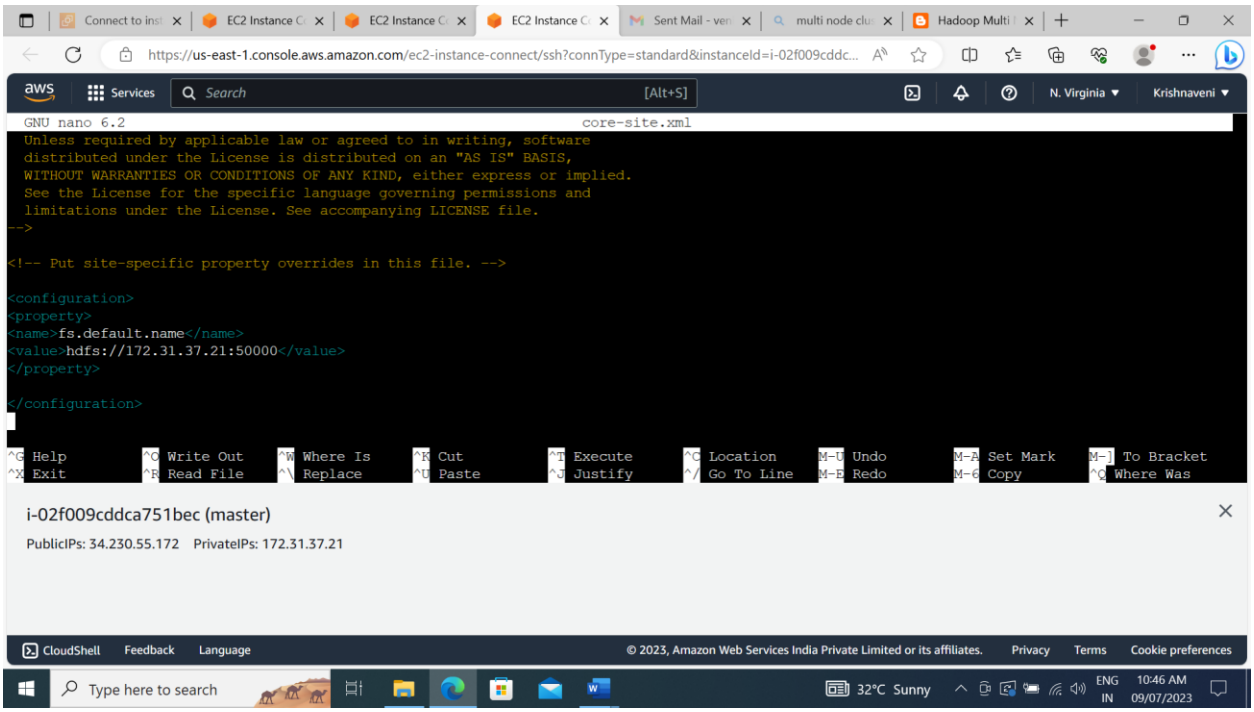

12. sudo nano core-site.xml

Open the file in all nodes and add the following content

<configuration>
<property>
<name>fs.default.name</name>

<value>hdfs://172.31.37.21:50000</value>
</property>

</configuration>




13.sudo nano yarn-site.xml

Open the file in all nodes and add the following content

<property>

<name>yarn.nodemanager.aux-services</name> <value>mapreduce_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

<property>

<description>The hostname of the RM.</description>

<name>yarn.resourcemanager.hostname</name>

<value>172.31.37.21</value>

</property>

<property>

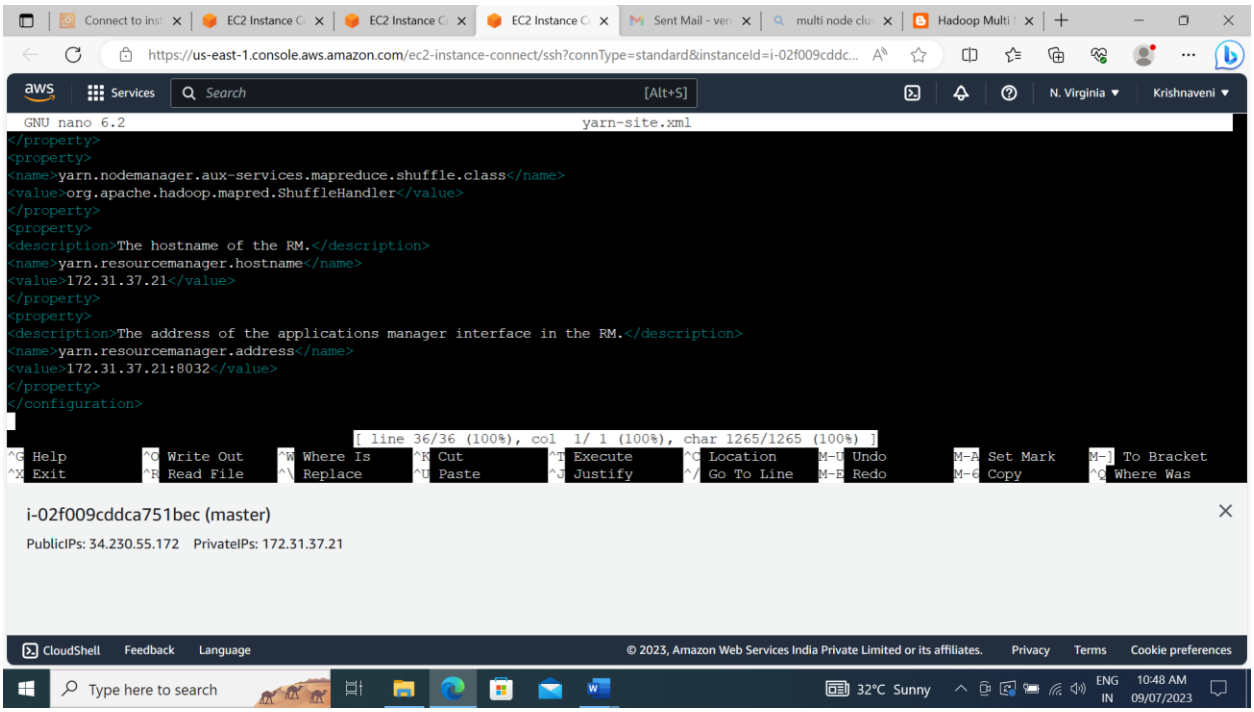
<description>The address of the applications manager interface in the RM.</description>

<name>yarn.resourcemanager.address</name>

<value>172.31.37.21:8032</value>

</property>

</configuration>




14. sudo nano mapred-site.xml

Open the file in all nodes and add the following content

<configuration>

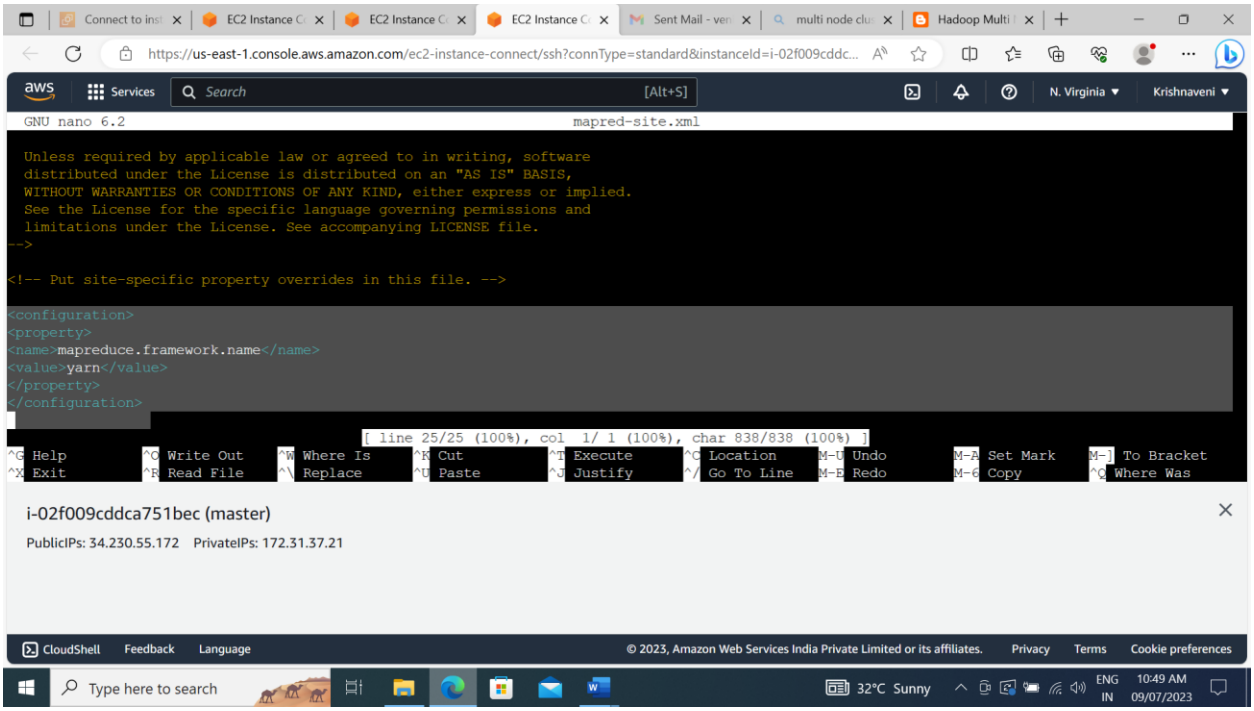
<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>




15. sudo nano hdfs-site.xml in all nodes and add the following content

<configuration>

<property>

<name>dfs.namenode.name.dir</name>

<value>/home/ubuntu/hadoop2-dir/namenode-dir</value>

</property>

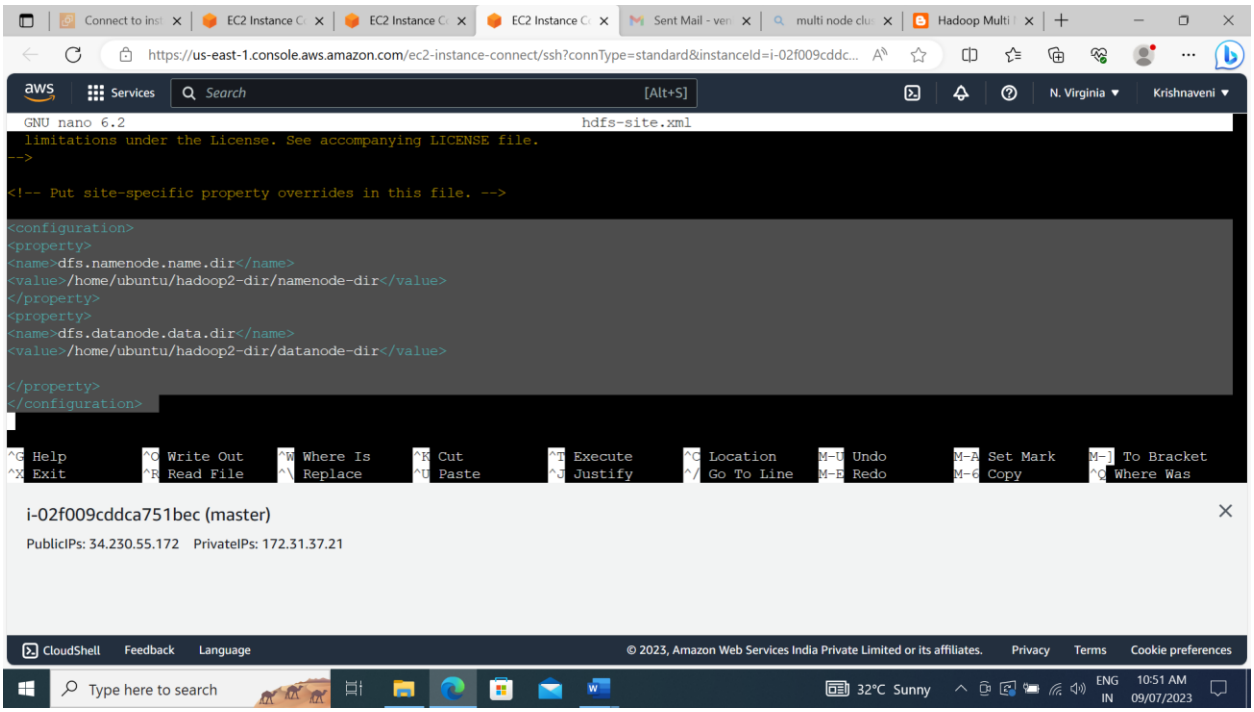
<property>

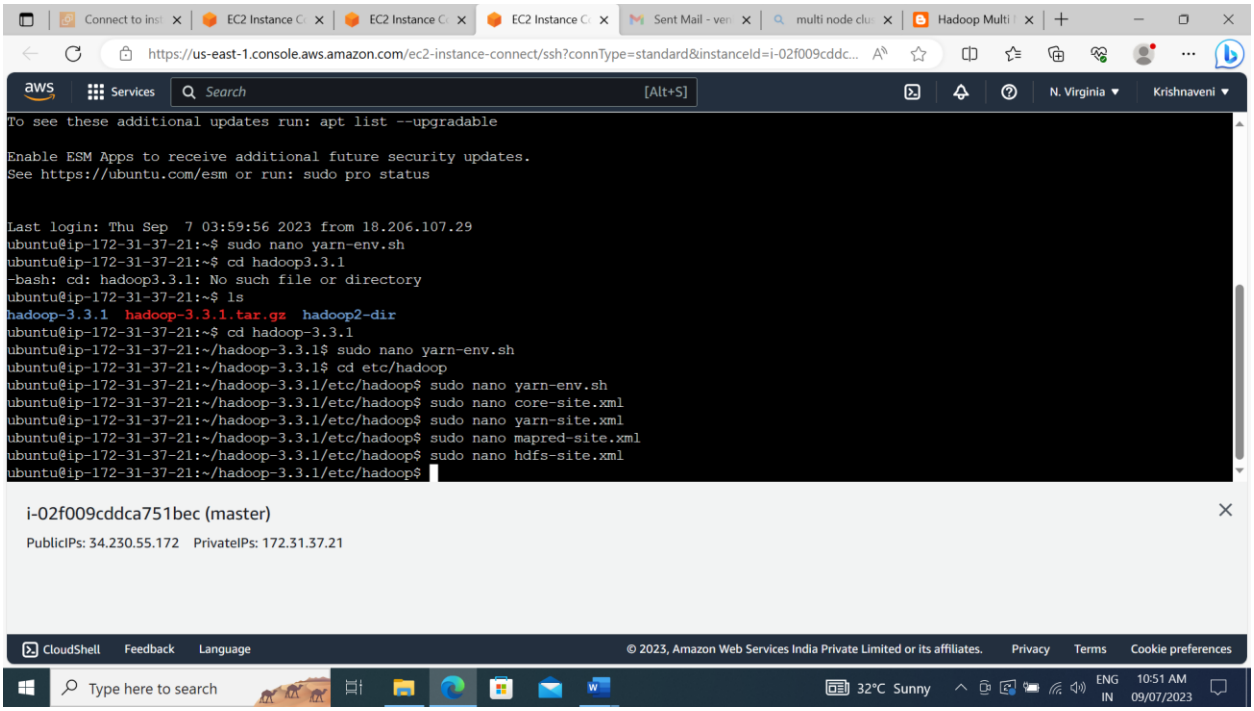
<name>dfs.datanode.data.dir</name>

<value>/home/ubuntu/hadoop2-dir/datanode-dir</value>

</property>

</configuration>




16 . ls -lstr

**Only in master node:**

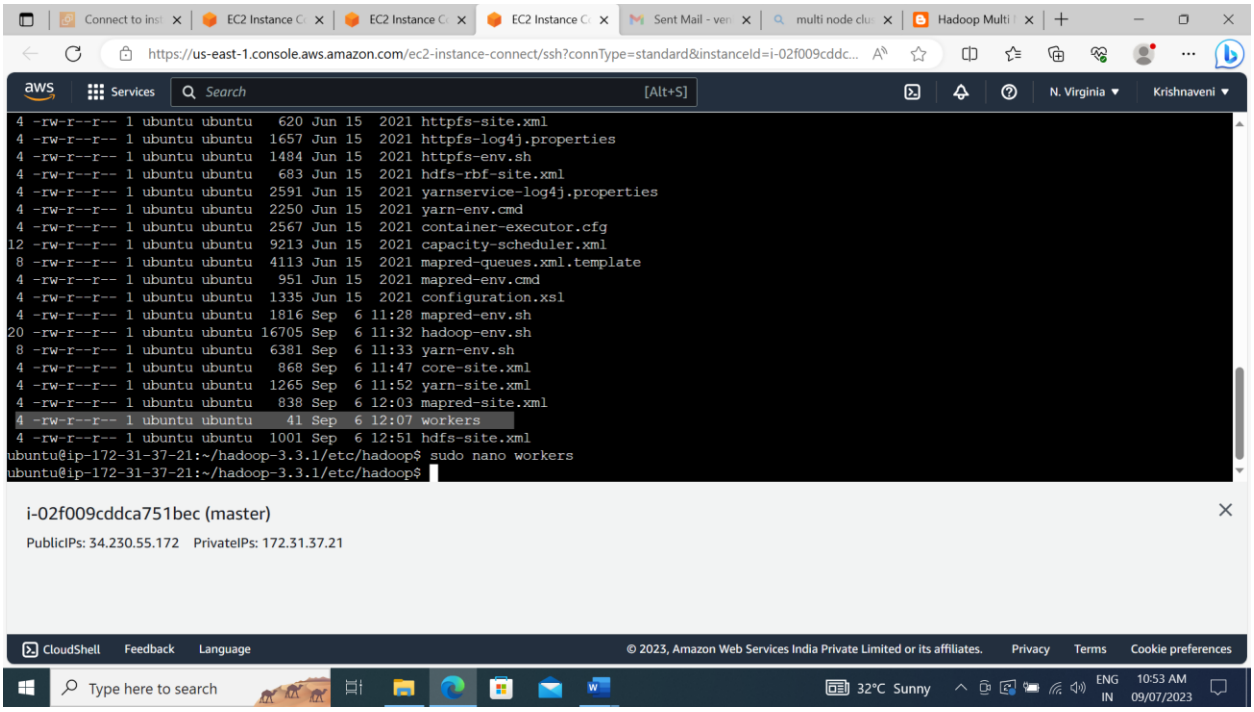
17. sudo nano workers

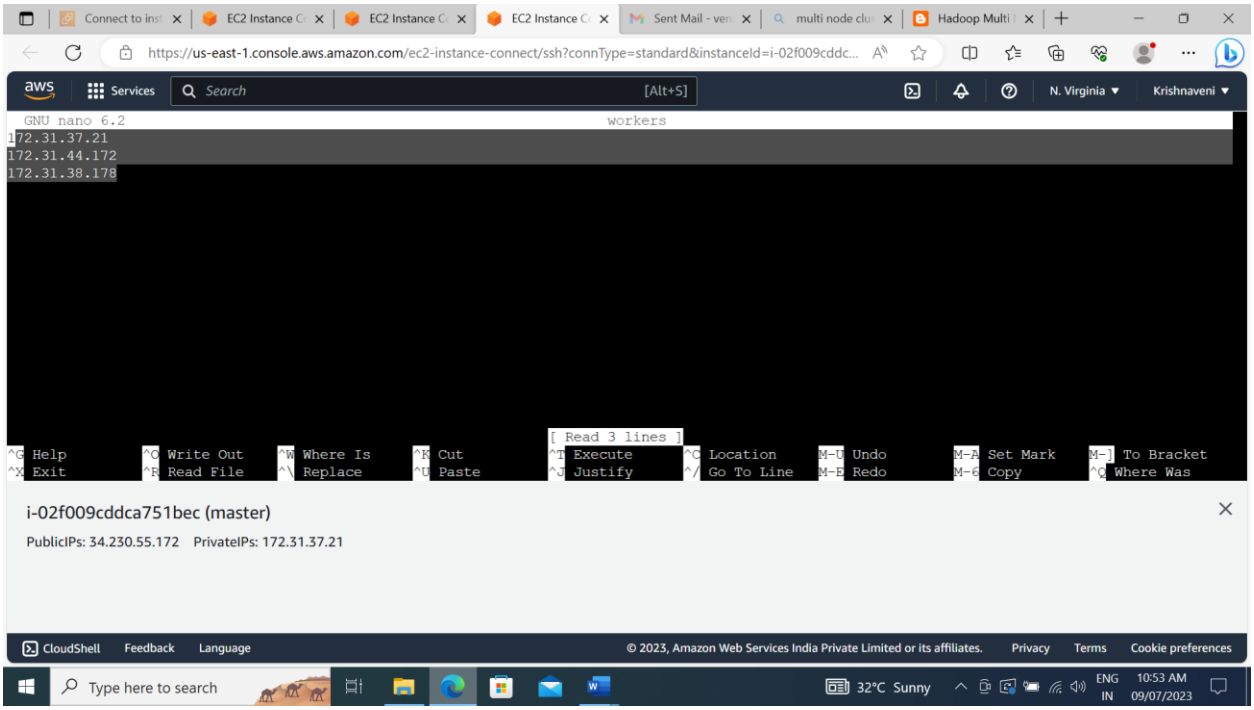
Only in master node and add the ip address of all node.

172.31.37.21

172.31.44.172

172.31.38.178



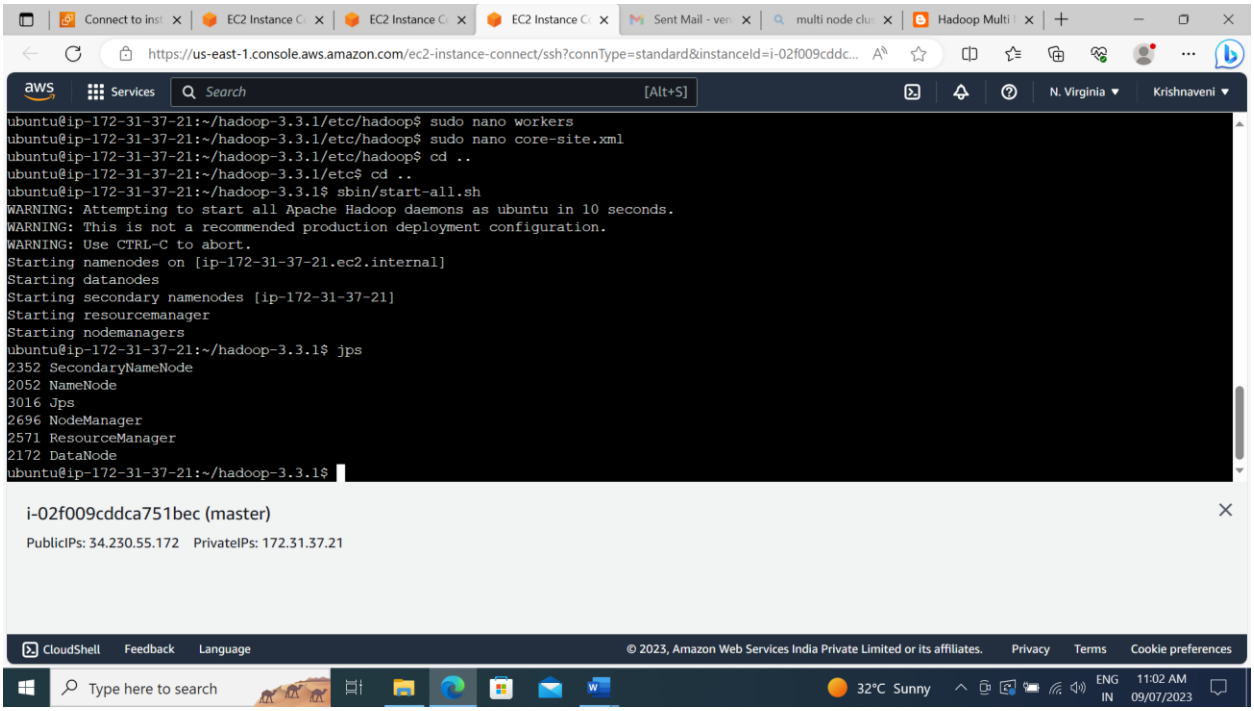



18. bin/hadoop namenode -format

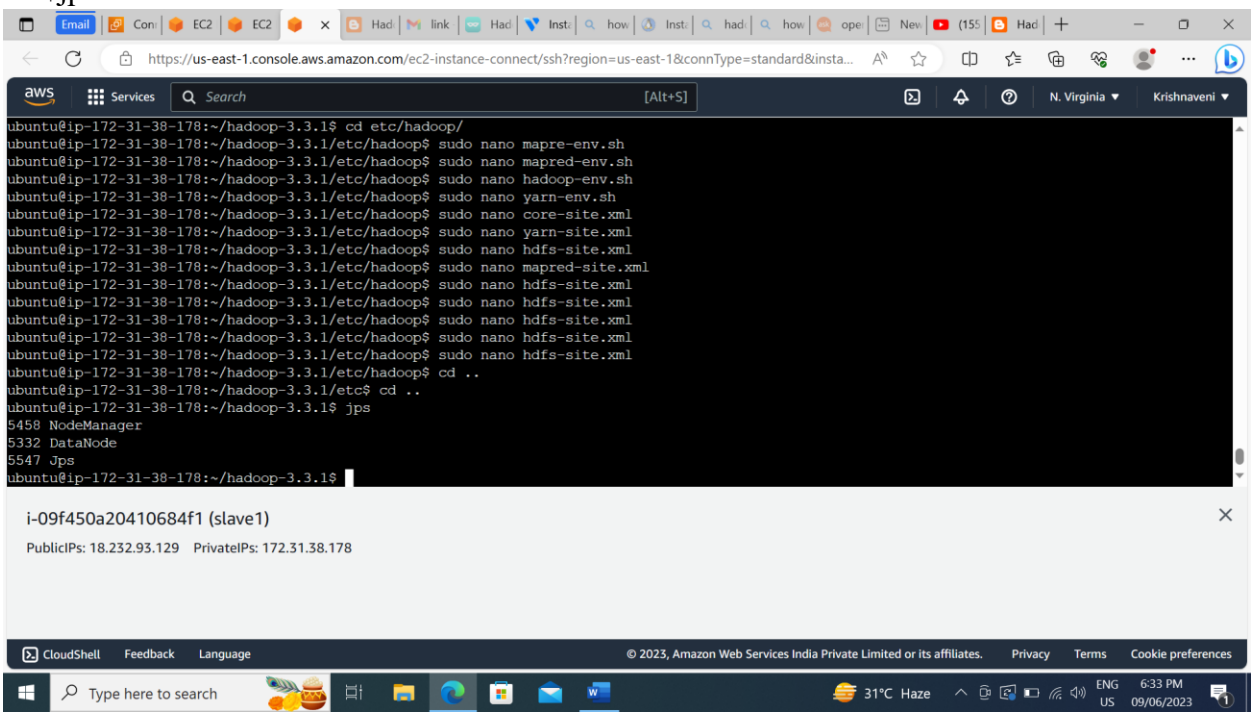
19. sbin/start-all.sh

20. jps

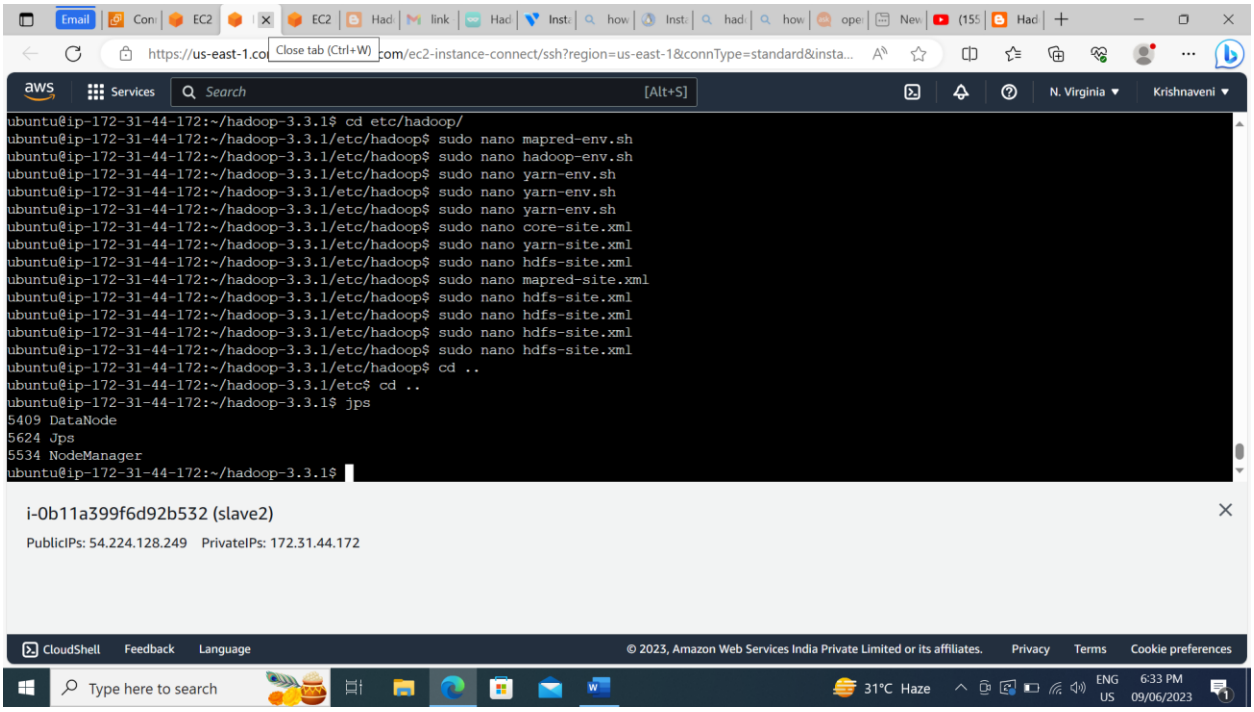
21.sbin/stop-all.sh



22.$jps -slavenode1

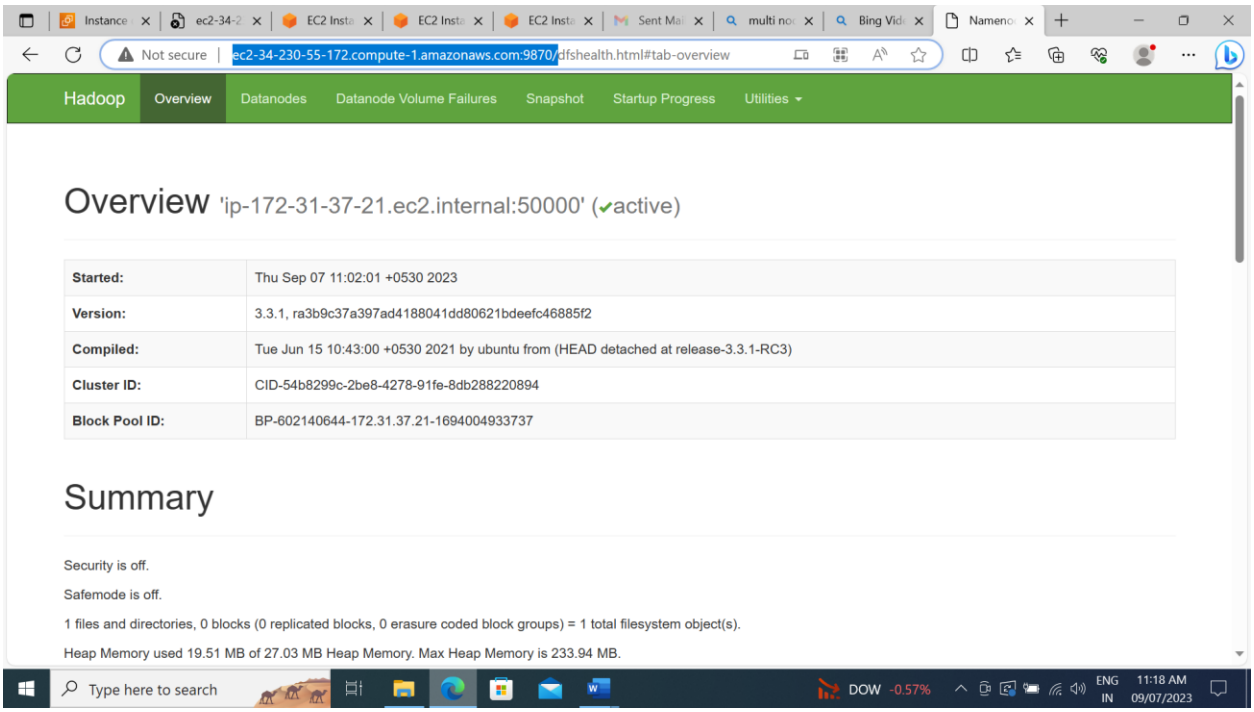


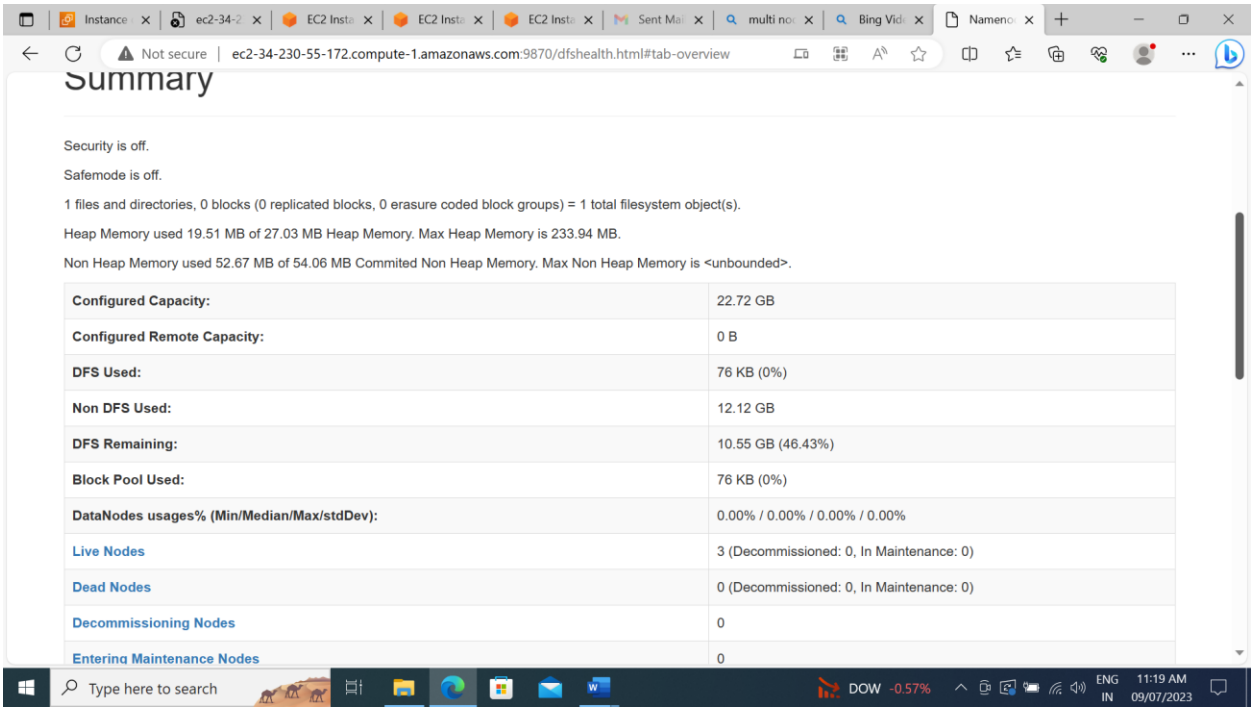

23.$jps -slavenode2

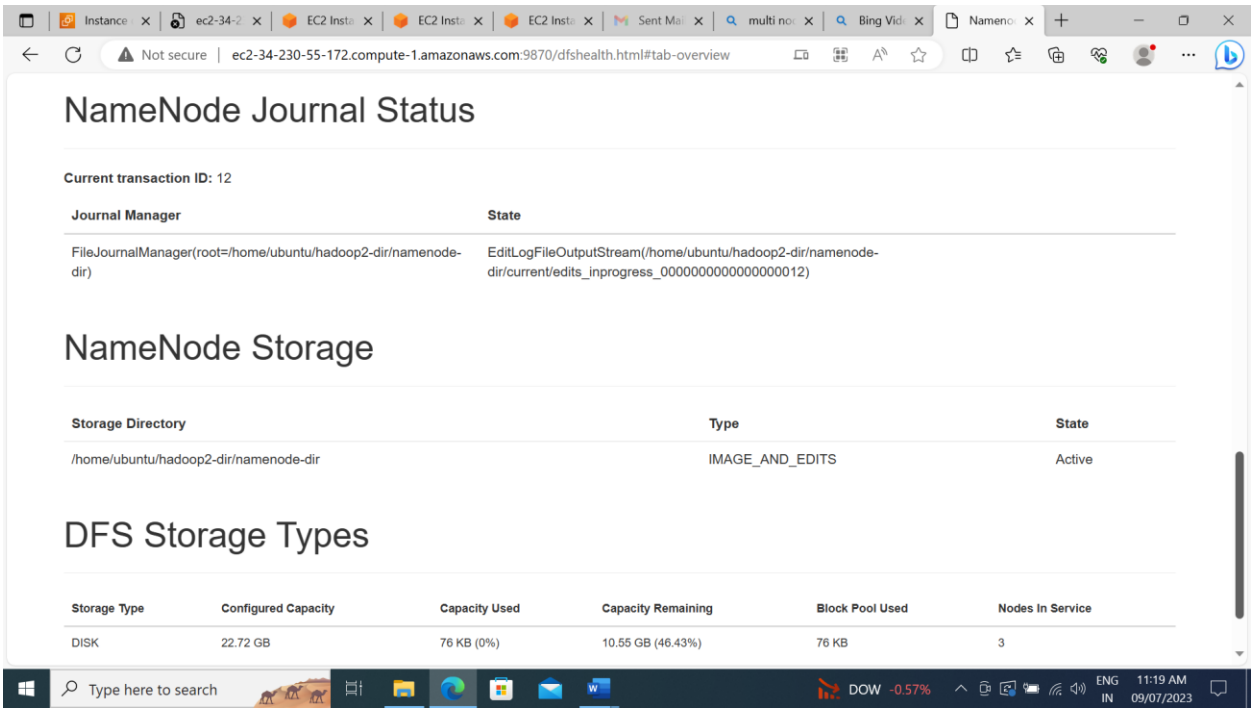



Open the link in browser

http://ec2-34-230-55-172.compute-1.amazonaws.com:9870/


Not secure
ec2-34-230-55-172.compute-1.amazonaws.com:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'ip-172-31-37-21.ec2.internal:50000' (active)
Started: Thu Sep 07 11:02:01 +0530 2023
Version: 3.3.1, ra3b9c37a397ad4188041dd80621bdeefc46885f2
Compiled: Tue Jun 15 10:43:00 +0530 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID: CID-54b8299c-2be8-4278-91fe-8db288220894
Block Pool ID: BP-602140644-172.31.37.21-1694004933737
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 19.51 MB of 27.03 MB Heap Memory. Max Heap Memory is 233.94 MB.
Type here to search
DOW -0.57%
ENG IN
11:18 AM
09/07/2023


Not secure
ec2-34-230-55-172.compute-1.amazonaws.com:9870/dfshealth.html#tab-overview
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 19.51 MB of 27.03 MB Heap Memory. Max Heap Memory is 233.94 MB.
Non Heap Memory used 52.67 MB of 54.06 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 22.72 GB
Configured Remote Capacity: 0 B
DFS Used: 76 KB (0%)
Non DFS Used: 12.12 GB
DFS Remaining: 10.55 GB (46.43%)
Block Pool Used: 76 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
Type here to search
DOW -0.57%
ENG IN
11:19 AM
09/07/2023


Not secure | ec2-34-230-55-172.compute-1.amazonaws.com:9870/dfshealth.html#tab-overview
NameNode Journal Status
Current transaction ID: 12
Journal Manager
State
FileJournalManager(root=/home/ubuntu/hadoop2-dir/namenode-dir)
EditLogFileOutputStream(/home/ubuntu/hadoop2-dir/namenode-dir/current/edits_inprogress_0000000000000000012)
NameNode Storage
Storage Directory
Type
State
/home/ubuntu/hadoop2-dir/namenode-dir
IMAGE_AND_EDITS
Active
DFS Storage Types
Storage Type
Configured Capacity
Capacity Used
Capacity Remaining
Block Pool Used
Nodes In Service
DISK
22.72 GB
76 KB (0%)
10.55 GB (46.43%)
76 KB
3
Type here to search
DOW -0.57%
ENG
IN
11:19 AM
09/07/2023

**The history of Commands Executed at Master node:**

1. Sudo apt update -y
2. sudo apt install openjdk-8-jdk -y
3. sudo wget https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop 3.3.1.tar.gz
4. ls -lstr
5. tar -xvzf hadoop-3.3.1.tar.gz
6. $JAVA_HOME
7. whichc java
8. java -version
9. ls -a
10. sudo nano .bashrc
11. source .bashrc
12. ssh-keygen -t rsa
13. ls -a
14. cd .ssh
15. ls -a
16. cat id_rsa.pub
17. ls -lstr
18. nano authorized_keys
19. ssh 172.31.44.172
20. cd ..
21. ls -a
22. cd hadoop-3.3.1
23. cd etc
24. ls
25. cd hadoop
26. ls -lstr
27. sudo nano mapred-env.sh
28. sudo nano hadoop-env.sh
29. sudo nano yarn-env.sh
30. sudo nano core-site.sh
31. sudo nano core-site.xml
32. sudo nano yarn-site.xml
33. sudo nano hdfs-site.xml
34. sudo nano mapred-site.xml
35. ls -lstr
36. sudo nano workers
37. bin/hadoop namenode -format
38. sbin/start-all.sh
39. jps
40. sbin/stop-all.sh