

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

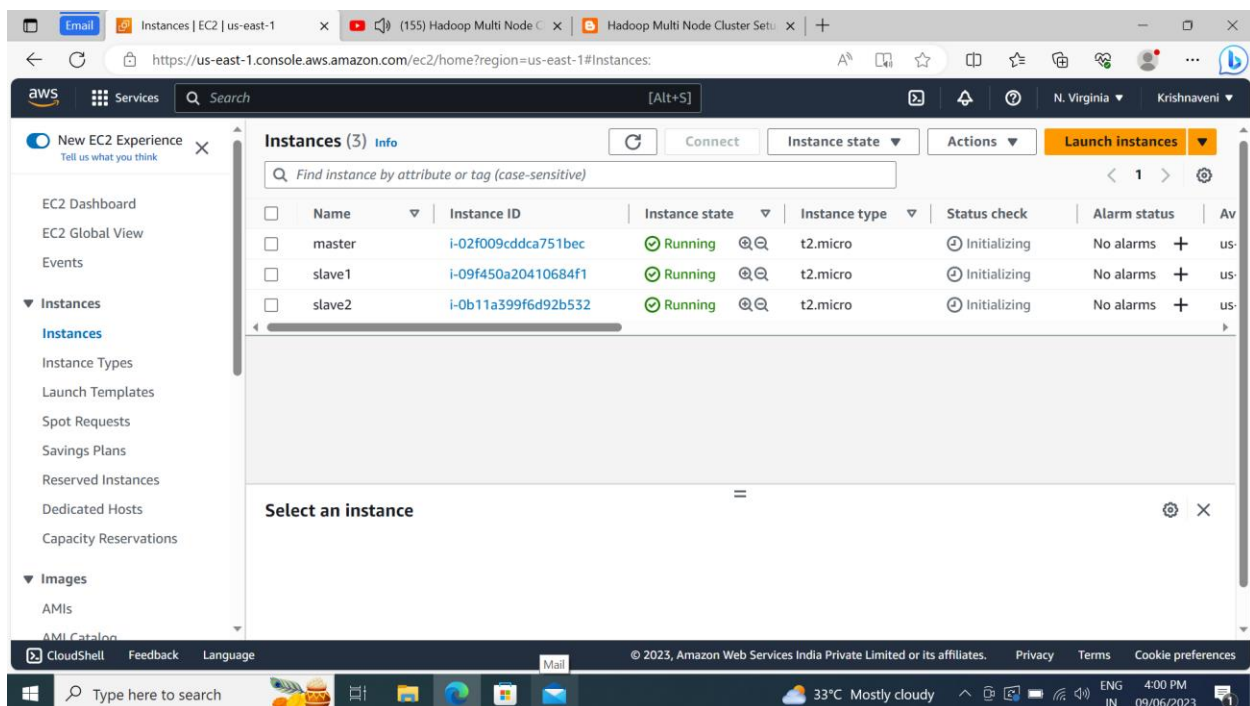
CLOUD BASED MULTINODE CLUSTER SET UP USING HADOOP

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

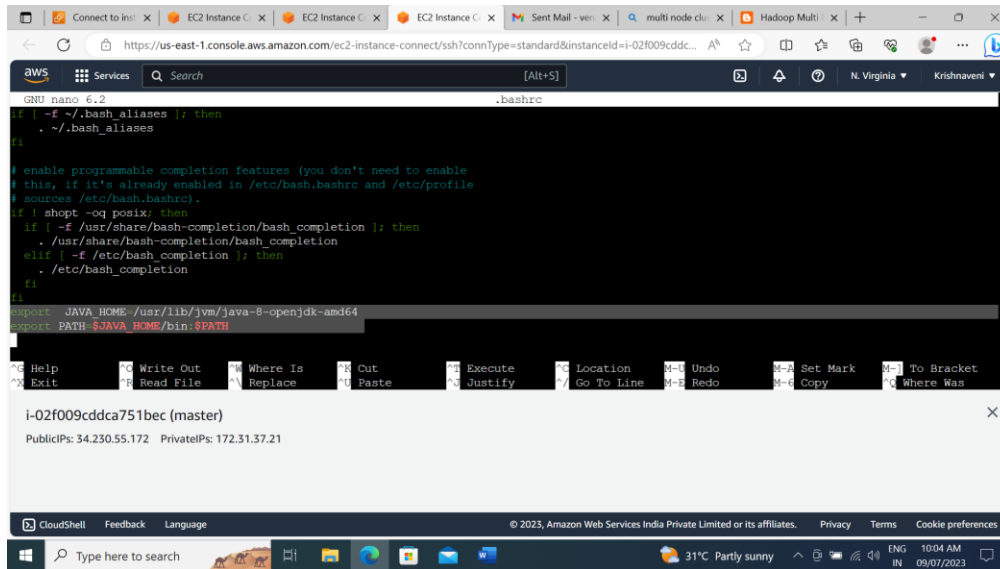
Installing java and hadoop in all nodes

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



The screenshot shows a terminal window with the nano editor open to the .bashrc file. The file content includes comments about enabling programmable completion features, sourcing /etc/bash.bashrc, and setting environment variables like JAVA_HOME and PATH. The terminal window title is "i-02f009cddca751bec (master)" and it shows public and private IP addresses. The bottom status bar indicates the date and time as 10:04 AM on 09/07/2023.

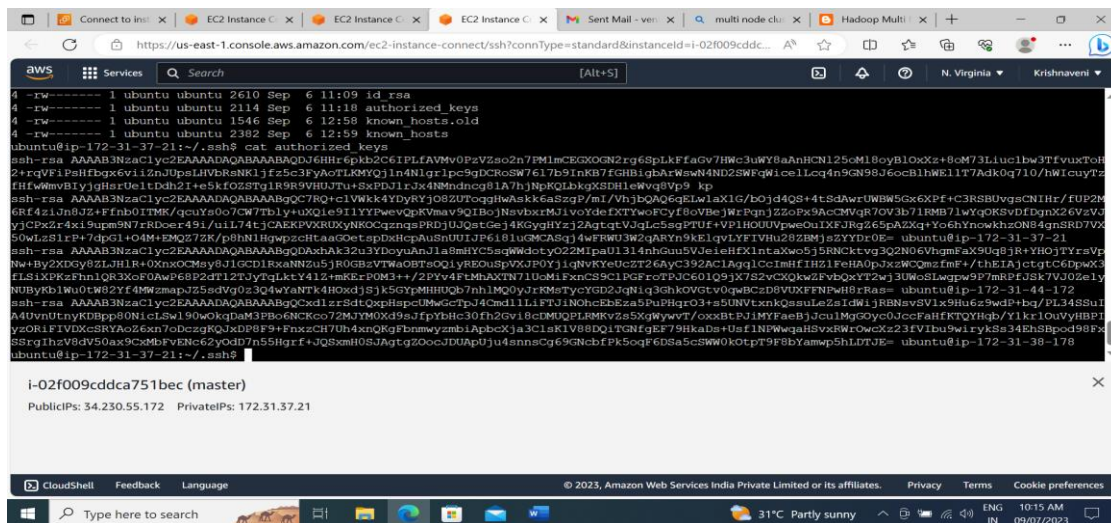
6.source .bashrc

7. ssh-keygen -t rsa

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

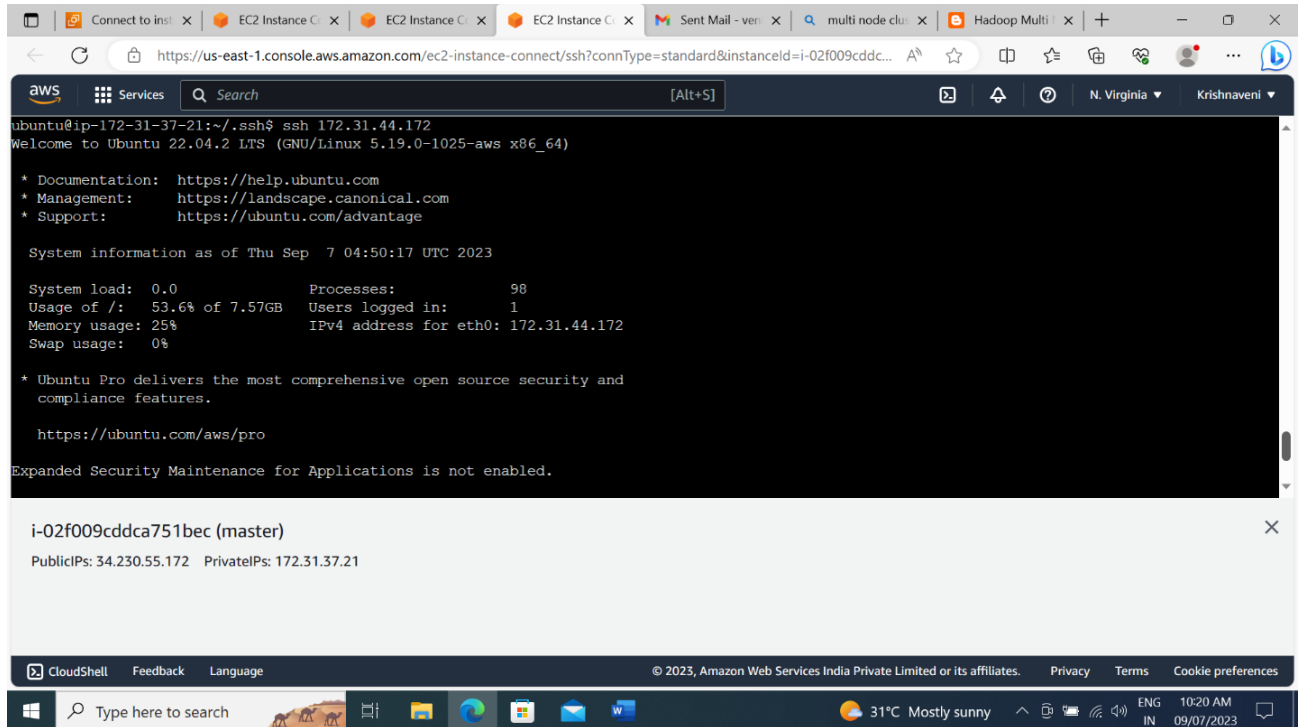
Generate the ssh key in all nodes and copy the keys of all nodes in .ssh/authorized_keys

Do the same step for all nodes



The screenshot shows a terminal window with the output of the ssh-keygen -t rsa command. The output displays the file permissions, the user (ubuntu), the host (ubuntu), the date and time (2610 Sep 6 11:09), and the key type (rsa). The key is displayed in a long, single line. The terminal window title is "i-02f009cddca751bec (master)" and it shows public and private IP addresses. The bottom status bar indicates the date and time as 10:15 AM on 09/07/2023.

Ensure for password less connection through ssh



8. ssh 172.31.44.172

Master node is connected with slave node without password

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
```

The screenshot shows the AWS CloudShell interface. The terminal window displays the following output:

```
Expanded Security Maintenance for Applications is not enabled.

116 updates can be applied immediately.
65 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

Last login: Thu Sep  7 03:59:56 2023 from 18.206.107.29
ubuntu@ip-172-31-37-21:~$ sudo nano yarn-env.sh
ubuntu@ip-172-31-37-21:~$ cd hadoop3.3.1
-bash: cd: hadoop3.3.1: No such file or directory
ubuntu@ip-172-31-37-21:~$ ls
hadoop-3.3.1  hadoop-3.3.1.tar.gz  hadoop2-dir
ubuntu@ip-172-31-37-21:~$ cd hadoop-3.3.1
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$ sudo nano yarn-env.sh
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$ cd etc/hadoop
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-env.sh
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$
```

Below the terminal, a file explorer shows the file `i-02f009cddca751bec (master)` with public and private IP addresses.

The screenshot shows the AWS CloudShell interface with the `yarn-env.sh` file open in the `nano` editor. The file contains the following content:

```
GNU nano 6.2 yarn-env.sh
# For privileged registry DNS, user to run as after dropping privileges
# This will replace the hadoop.id.str Java property in secure mode.
# export YARN_REGISTRYDNS_SECURE_USER=yarn

# Supplemental options for privileged registry DNS
# By default, Hadoop uses javc which needs to know to launch a
# server jvm.
# export YARN_REGISTRYDNS_SECURE_EXTRA_OPTS="-jvm server"

###
# YARN Services parameters
###
# Directory containing service examples
# export YARN_SERVICE_EXAMPLES_DIR = $HADOOP_YARN_HOME/share/hadoop/yarn/yarn-service-examples
# export YARN_CONTAINER_RUNTIME_DOCKER_RUN_OVERRIDE_DISABLE=true
# export JAVA_HOME /usr/lib/jvm/java-8-openjdk-amd64
```

Below the terminal, a file explorer shows the file `i-02f009cddca751bec (master)` with public and private IP addresses.

12. sudo nano core-site.xml

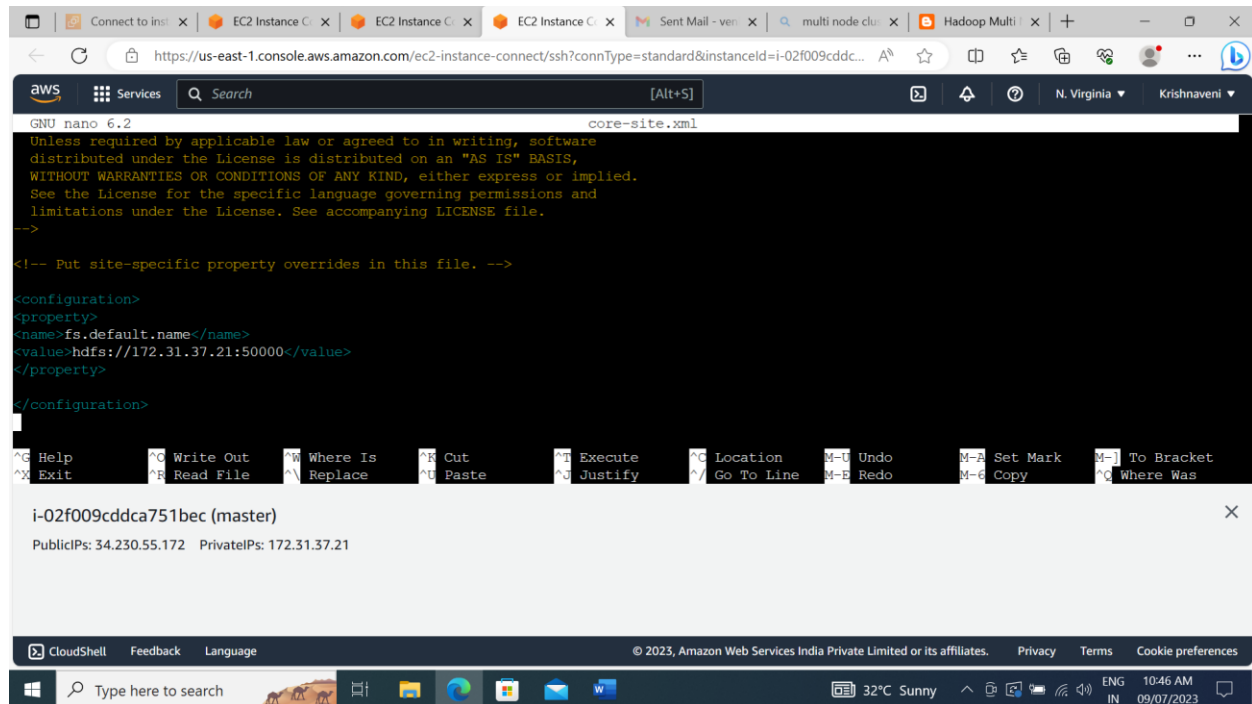
Open the file in all nodes and add the following content

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

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

```
</property>
```

```
</configuration>
```



```
GNU nano 6.2 core-site.xml
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>fs.default.name</name>
<value>hdfs://172.31.37.21:50000</value>
</property>
</configuration>
^G Help      ^O Write Out ^R Where Is  ^K Cut      ^T Execute  ^C Location M-U Undo     M-A Set Mark M-I To Bracket
^X Exit      ^P Read File ^\ Replace  ^U Paste    ^J Justify  ^_ Go To Line M-E Redo     M-C Copy     ^C Where Was
i-02f009cddca751bec (master)
PublicIPs: 34.230.55.172 PrivateIPs: 172.31.37.21
CloudShell Feedback Language
© 2023, Amazon Web Services India Private Limited or its affiliates. Privacy Terms Cookie preferences
Type here to search 32°C Sunny ENG 10:46 AM
IN 09/07/2023
```

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>
```

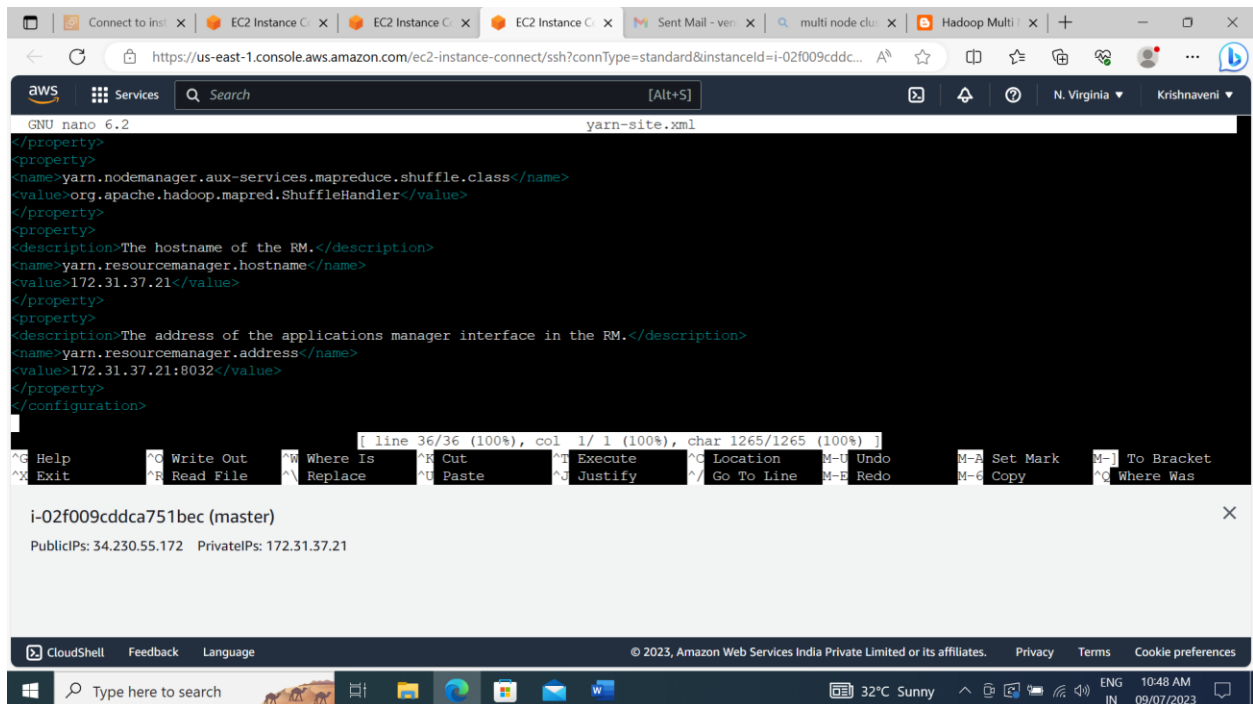
```
<description>The address of the applications manager interface in the RM.</description>
```

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

```
<value>172.31.37.21:8032</value>
```

```
</property>
```

```
</configuration>
```

A screenshot of an AWS EC2 console window. The browser tabs show 'Connect to ins...', 'EC2 Instance C...', 'EC2 Instance C...', 'EC2 Instance C...', 'Sent Mail - ven...', 'multi node clus...', and 'Hadoop Multi...'. The address bar shows 'https://us-east-1.console.aws.amazon.com/ec2-instance-connect/ssh?connType=standard&instanceId=i-02f009cddc...'. The AWS console header shows 'aws', 'Services', a search bar, and the user 'Krishnaveni' in 'N. Virginia'. The terminal window shows the 'GNU nano 6.2' editor editing 'yarn-site.xml'. The content of the file is:

```
</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>
```

 The terminal window also shows a status bar with 'i-02f009cddca751bec (master)' and 'PublicIPs: 34.230.55.172 PrivateIPs: 172.31.37.21'. The bottom of the screenshot shows a Windows taskbar with the search bar, task icons, and system tray showing '32°C Sunny' and '10:48 AM 09/07/2023'.

14. sudo nano mapred-site.xml

Open the file in all nodes and add the following content

```
<configuration>
```

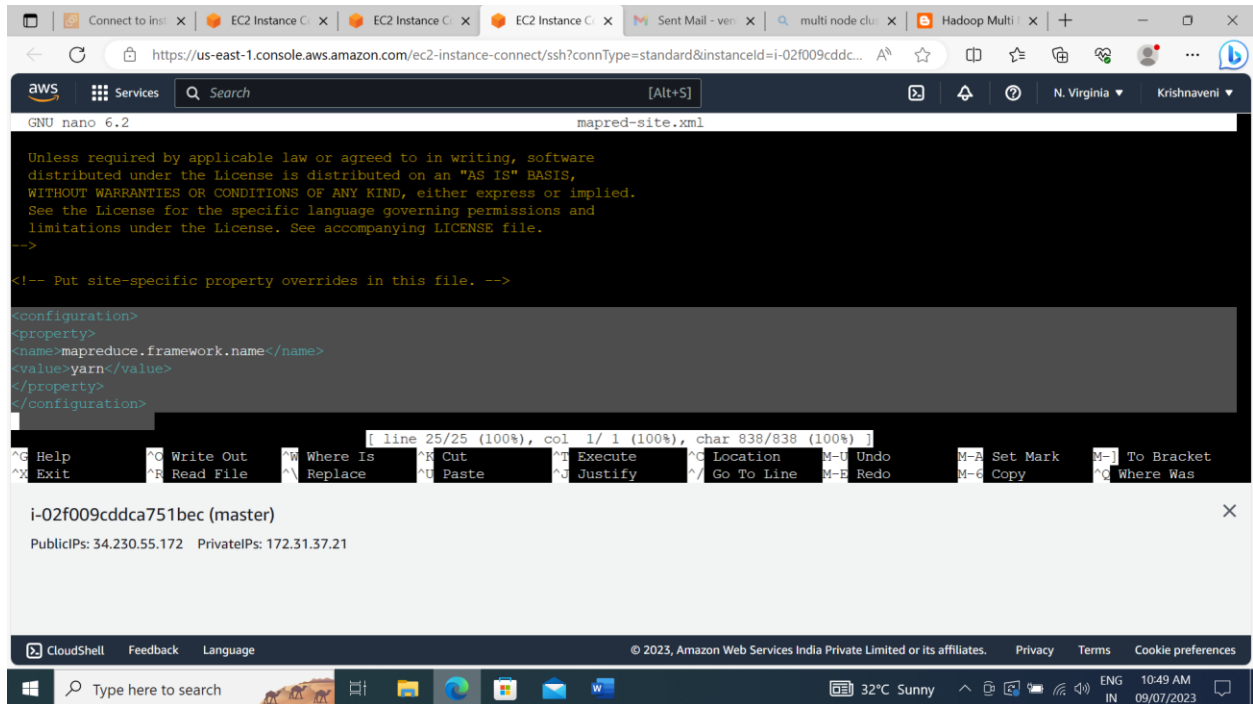
```
<property>
```

```
<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>

The screenshot shows the AWS CloudShell interface with a terminal window running GNU nano 6.2. The file being edited is `hdfs-site.xml`. The terminal content includes comments about license limitations and site-specific property overrides, followed by an XML configuration block for HDFS. The configuration sets the NameNode directory to `/home/ubuntu/hadoop2-dir/namenode-dir` and the DataNode directory to `/home/ubuntu/hadoop2-dir/datanode-dir`. Below the terminal, a metadata box for instance `i-02f009cddca751bec (master)` is visible, showing public and private IP addresses. The bottom of the interface includes the AWS CloudShell navigation bar and a Windows taskbar at the very bottom.

```
GNU nano 6.2 hdfs-site.xml
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->

<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>
```

i-02f009cddca751bec (master)
PublicIPs: 34.230.55.172 PrivateIPs: 172.31.37.21

This screenshot shows the AWS CloudShell terminal where Hadoop 3.3.1 is being installed and configured. The terminal output shows the user running `apt list --upgradable`, then downloading the Hadoop tarball and extracting it. Subsequently, the user navigates to the `etc/hadoop` directory and configures several files: `yarn-env.sh`, `core-site.xml`, `yarn-site.xml`, `mapred-site.xml`, and `hdfs-site.xml`. The terminal also shows the user's login history and the current directory structure. The instance metadata box for `i-02f009cddca751bec (master)` is visible below the terminal. The interface includes the AWS CloudShell navigation bar and a Windows taskbar at the bottom.

```
To see these additional updates run: apt list --upgradable
Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

Last login: Thu Sep  7 03:59:56 2023 from 18.206.107.29
ubuntu@ip-172-31-37-21:~$ sudo nano yarn-env.sh
ubuntu@ip-172-31-37-21:~$ cd hadoop3.3.1
-bash: cd: hadoop3.3.1: No such file or directory
ubuntu@ip-172-31-37-21:~$ ls
hadoop-3.3.1  hadoop-3.3.1.tar.gz  hadoop2-dir
ubuntu@ip-172-31-37-21:~$ cd hadoop-3.3.1
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$ sudo nano yarn-env.sh
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$ cd etc/hadoop
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-env.sh
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano core-site.xml
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-site.xml
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano mapred-site.xml
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$
```

i-02f009cddca751bec (master)
PublicIPs: 34.230.55.172 PrivateIPs: 172.31.37.21

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

The screenshot shows the AWS CloudShell interface with a terminal window. The terminal displays a list of files with their permissions, owner, group, size, date, and filename. The files are:

- 4 -rw-r--r-- 1 ubuntu ubuntu 620 Jun 15 2021 https-site.xml
- 4 -rw-r--r-- 1 ubuntu ubuntu 1657 Jun 15 2021 https-log4j.properties
- 4 -rw-r--r-- 1 ubuntu ubuntu 1484 Jun 15 2021 https-env.sh
- 4 -rw-r--r-- 1 ubuntu ubuntu 683 Jun 15 2021 hdfs-rbf-site.xml
- 4 -rw-r--r-- 1 ubuntu ubuntu 2591 Jun 15 2021 yarnservice-log4j.properties
- 4 -rw-r--r-- 1 ubuntu ubuntu 2250 Jun 15 2021 yarn-env.cmd
- 4 -rw-r--r-- 1 ubuntu ubuntu 2567 Jun 15 2021 container-executor.cfg
- 12 -rw-r--r-- 1 ubuntu ubuntu 9213 Jun 15 2021 capacity-scheduler.xml
- 8 -rw-r--r-- 1 ubuntu ubuntu 4113 Jun 15 2021 mapred-queues.xml.template
- 4 -rw-r--r-- 1 ubuntu ubuntu 951 Jun 15 2021 mapred-env.cmd
- 4 -rw-r--r-- 1 ubuntu ubuntu 1335 Jun 15 2021 configuration.xml
- 4 -rw-r--r-- 1 ubuntu ubuntu 1816 Sep 6 11:28 mapred-env.sh
- 20 -rw-r--r-- 1 ubuntu ubuntu 16705 Sep 6 11:32 hadoop-env.sh
- 8 -rw-r--r-- 1 ubuntu ubuntu 6381 Sep 6 11:33 yarn-env.sh
- 4 -rw-r--r-- 1 ubuntu ubuntu 868 Sep 6 11:47 core-site.xml
- 4 -rw-r--r-- 1 ubuntu ubuntu 1265 Sep 6 11:52 yarn-site.xml
- 4 -rw-r--r-- 1 ubuntu ubuntu 838 Sep 6 12:03 mapred-site.xml
- 4 -rw-r--r-- 1 ubuntu ubuntu 41 Sep 6 12:07 workers
- 4 -rw-r--r-- 1 ubuntu ubuntu 1001 Sep 6 12:51 hdfs-site.xml

The terminal prompt is `ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano workers`. Below the terminal, the instance details are shown: `i-02f009cddca751bec (master)` with PublicIPs: 34.230.55.172 and PrivateIPs: 172.31.37.21.

The screenshot shows the AWS CloudShell interface with a terminal window. The terminal displays the `workers` file being edited in the `nano` editor. The file content is:

```
172.31.37.21
172.31.44.172
172.31.38.178
```

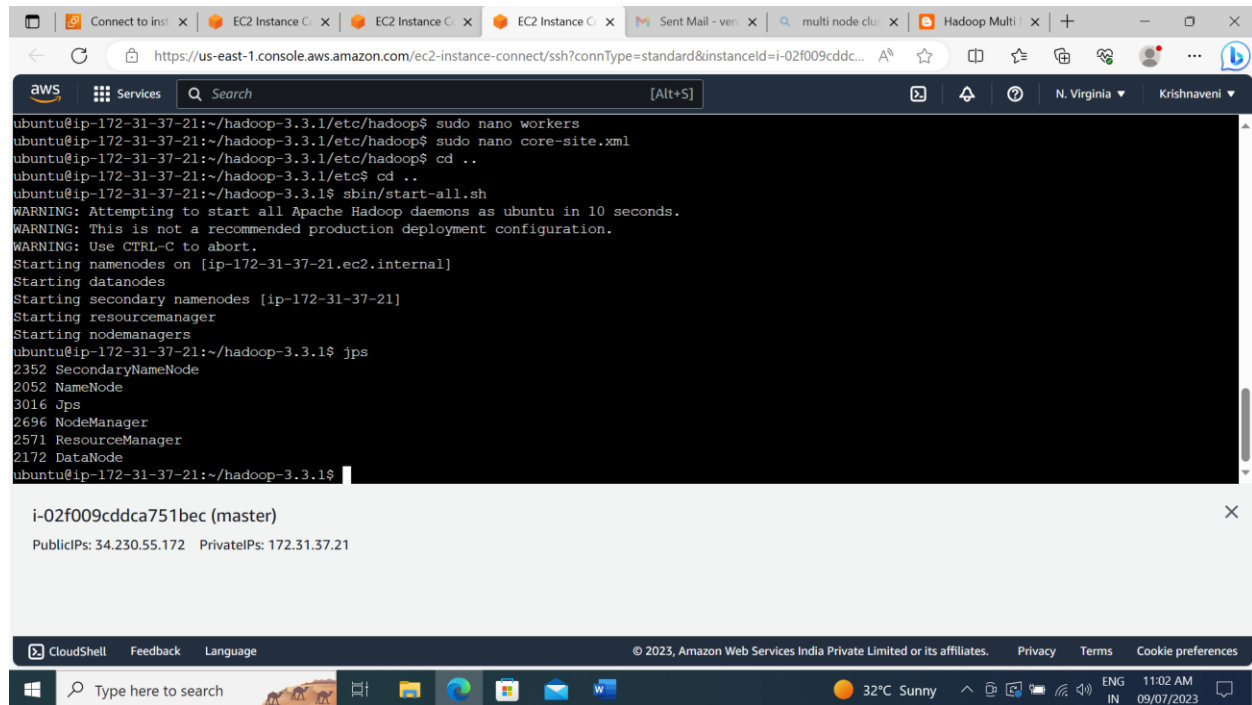
The terminal prompt is `ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$`. Below the terminal, the instance details are shown: `i-02f009cddca751bec (master)` with PublicIPs: 34.230.55.172 and PrivateIPs: 172.31.37.21.

18. bin/hadoop namenode -format

19. sbin/start-all.sh

20. jps

21. sbin/stop-all.sh



The screenshot shows an AWS CloudShell terminal window with the following content:

```
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano workers
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ sudo nano core-site.xml
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc/hadoop$ cd ..
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1/etc$ cd ..
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$ sbin/start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as ubuntu in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [ip-172-31-37-21.ec2.internal]
Starting datanodes
Starting secondary namenodes [ip-172-31-37-21]
Starting resourcemanager
Starting nodemanagers
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$ jps
2352 SecondaryNameNode
2052 NameNode
3016 Jps
2696 NodeManager
2571 ResourceManager
2172 DataNode
ubuntu@ip-172-31-37-21:~/hadoop-3.3.1$
```

Below the terminal output, a box displays instance information:

```
i-02f009cddca751bec (master)
PublicIPs: 34.230.55.172 PrivateIPs: 172.31.37.21
```

The bottom of the screenshot shows the AWS CloudShell interface with the text "© 2023, Amazon Web Services India Private Limited or its affiliates." and a Windows taskbar at the very bottom.

22.\$jps -slavenode1

The screenshot shows the AWS CloudShell interface for an EC2 instance. The terminal window displays the following commands and output:

```
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1$ cd etc/hadoop/
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano mapred-env.sh
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano mapred-env.sh
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hadoop-env.sh
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-env.sh
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano core-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano mapred-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc/hadoop$ cd ..
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1/etc$ cd ..
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1$ jps
5458 NodeManager
5332 DataNode
5547 Jps
ubuntu@ip-172-31-38-178:~/hadoop-3.3.1$
```

Below the terminal, a summary card for the instance **i-09f450a20410684f1 (slave1)** is shown, listing PublicIPs: 18.232.93.129 and PrivateIPs: 172.31.38.178.

23.\$jps -slavenode2

The screenshot shows the AWS CloudShell interface for an EC2 instance. The terminal window displays the following commands and output:

```
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano mapred-env.sh
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano hadoop-env.sh
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-env.sh
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-env.sh
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-env.sh
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano core-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano yarn-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano mapred-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ sudo nano hdfs-site.xml
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc/hadoop$ cd ..
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1/etc$ cd ..
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1$ jps
5409 DataNode
5624 Jps
5534 NodeManager
ubuntu@ip-172-31-44-172:~/hadoop-3.3.1$
```

Below the terminal, a summary card for the instance **i-0b11a399f6d92b532 (slave2)** is shown, listing PublicIPs: 54.224.128.249 and PrivateIPs: 172.31.44.172.

Open the link in browser

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

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