Building a Hadoop cluster

1. Checked whether your Hadoop cluster is running correctly or not.

<u>Firstly, logging in and installing java on all 3 nodes:</u>

Accessing cscluster.uis.edu

Master node IP:10.92.128.52

ssh csc@10.92.128.52

```
csc@CSC534BD-HM: ~
PS C:\Users\Swathi>  ssh csc@10.92.128.52
The authenticity of host '10.92.128.52 (10.92.128.52)' can't be established.
ECDSA key fingerprint is SHA256:dXhKfHsYIXe/53hvU+HOK2V6fVrTbz/QxmUhpnPXpzA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '10.92.128.52' (ECDSA) to the list of known hosts.
csc@10.92.128.52's password:
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 5.4.0-42-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Wed Sep 18 14:36:24 UTC 2024

  System load:  0.07               Processes:               322
  Usage of /:   9.8% of 97.93GB    Users logged in:         0
  Memory usage: 16%                IPv4 address for ens160: 10.92.128.52
  Swap usage:   0%

 * Are you ready for Kubernetes 1.19? It's nearly here! Try RC3 with
   sudo snap install microk8s --channel=1.19/candidate --classic

   https://www.microk8s.io/ has docs and details.

0 updates can be installed immediately.
0 of these updates are security updates.


The list of available updates is more than a week old.
To check for new updates run: sudo apt update

Last login: Fri Jun  5 13:55:11 2020
csc@CSC534BD-HM:~$
```

```
csc@CSC534BD-HM: ~
csc@CSC534BD-HM:~$ sudo apt update
[sudo] password for csc:
Hit:1 http://us.archive.ubuntu.com/ubuntu focal InRelease
Get:2 http://us.archive.ubuntu.com/ubuntu focal-updates InRelease [128 kB]
Get:3 http://us.archive.ubuntu.com/ubuntu focal-backports InRelease [128 kB]
Get:4 http://us.archive.ubuntu.com/ubuntu focal-security InRelease [128 kB]
Get:5 http://us.archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [3597 kB]
Get:6 http://us.archive.ubuntu.com/ubuntu focal-updates/main Translation-en [553 kB]
```

```
csc@CSC534BD-HM: ~
csc@CSC534BD-HM:~$ sudo apt install openjdk-8-jdk
eading package lists... Done
uilding dependency tree
eading state information... Done
he following additional packages will be installed:
 ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni lib
 libxdmcp-dev libxt-dev openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-core-de
uggested packages:
 default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-8-demo openjdk-8-source vi
 fonts-wqy-zenhei
he following NEW packages will be installed:
 ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni lib
 libxt-dev openjdk-8-jdk openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-core-d
he following packages will be upgraded:
 libx11-6
 upgraded, 21 newly installed, 0 to remove and 736 not upgraded.
```

```
csc@CSC534BD-HM: ~
csc@CSC534BD-HM: $ sudo update-alternatives --config java
There is only one alternative in link group java (providing /usr/bin/java): /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java
Nothing to configure.
csc@CSC534BD-HM: $
```

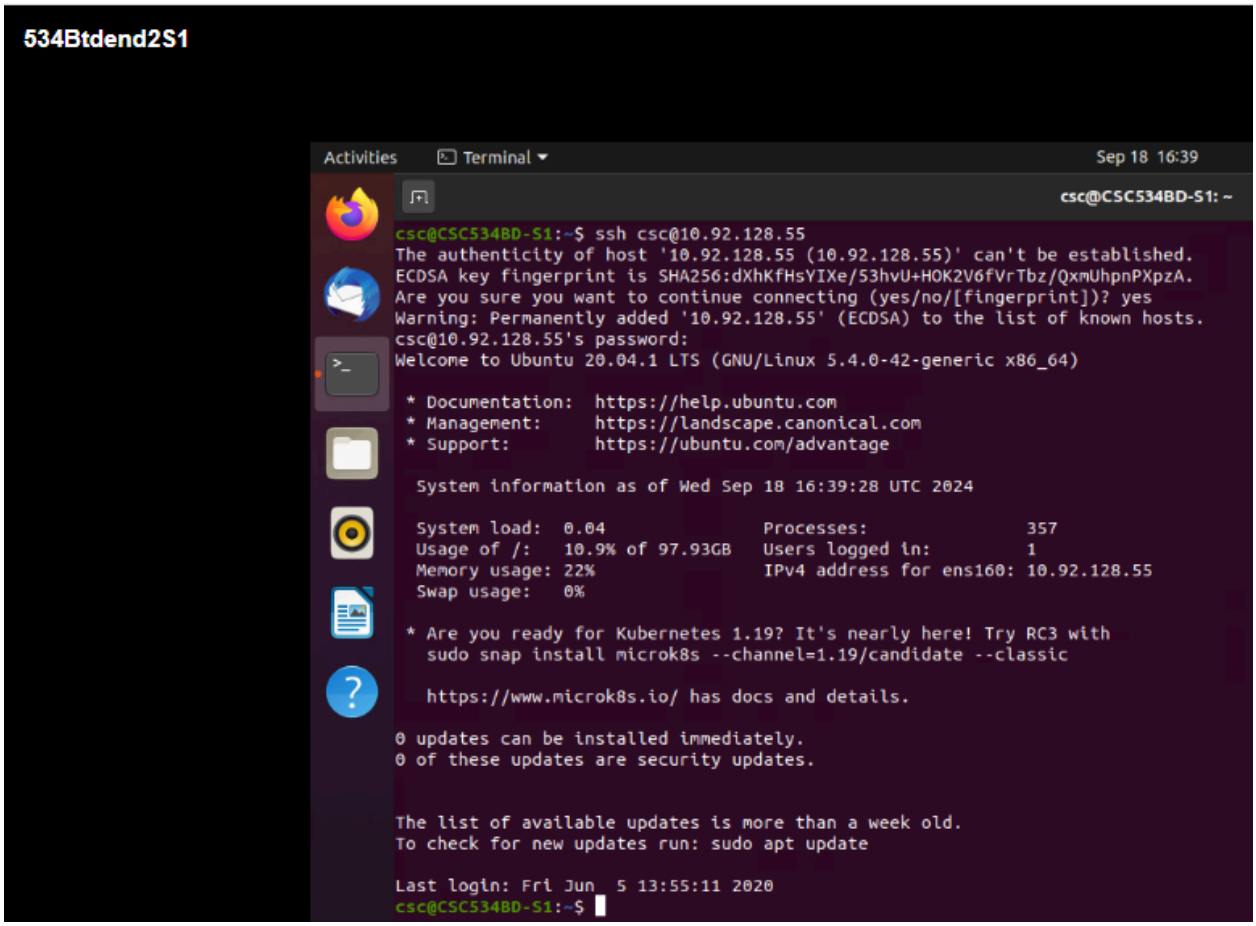
Java -version:

```
534Btdend2HM
Activities   Terminal   Sep 18 16:29
csc@CSC534BD-HM: ~
csc@CSC534BD-HM:~$ java -version
openjdk version "1.8.0_422"
OpenJDK Runtime Environment (build 1.8.0_422-8u422-b05-1~20.04-b05)
OpenJDK 64-Bit Server VM (build 25.422-b05, mixed mode)
```

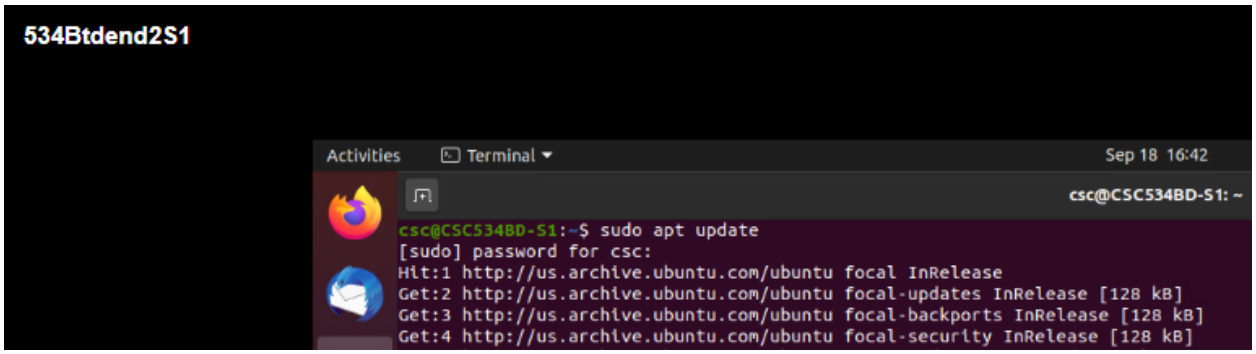
Hadoop version on master node after downloading hadoop & setting the path of hadoop and java in bashrc:

```
534Btdend2HM
Activities   Terminal   Sep 18 16:30
csc@CSC534BD-HM: ~
csc@CSC534BD-HM:~$ nano ~/.bashrc
csc@CSC534BD-HM:~$ source  ~/.bashrc
csc@CSC534BD-HM:~$ hadoop version
Hadoop 2.10.2
Subversion Unknown -r 965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled by ubuntu on 2022-05-24T22:35Z
Compiled with protoc 2.5.0
From source with checksum d3ab737f7788f05d467784f0a86573fe
This command was run using /home/csc/tdend2/hadoop-2.10.2/share/hadoop/common/hadoop-common-2.10.2.jar
csc@CSC534BD-HM:~$
```

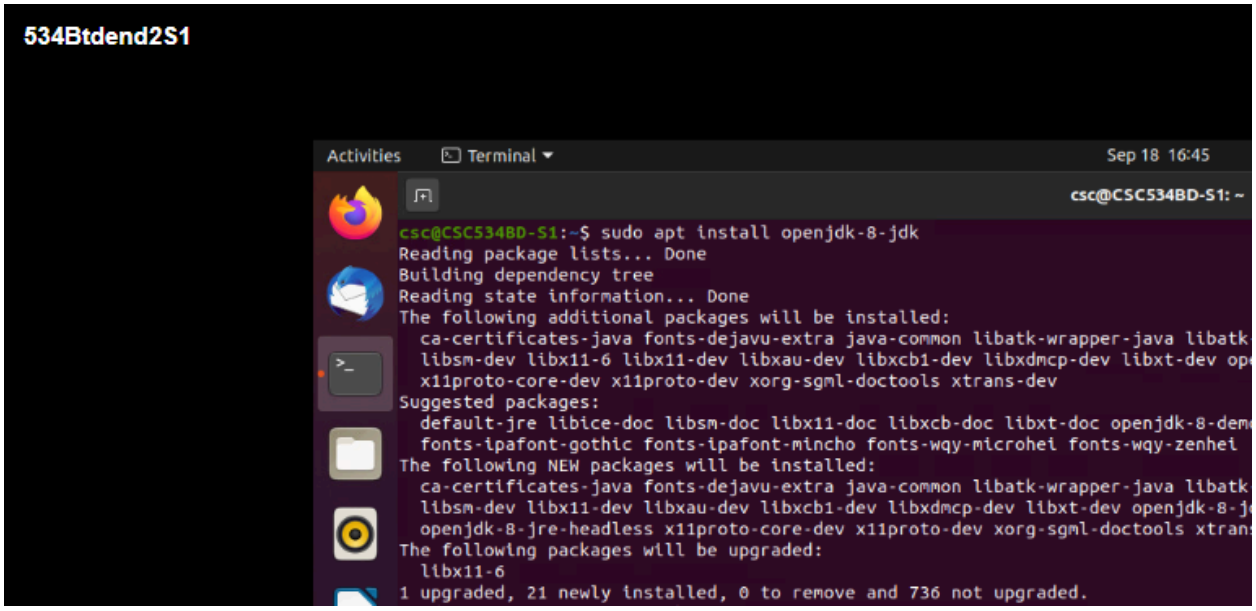
Logged into worker node -1:



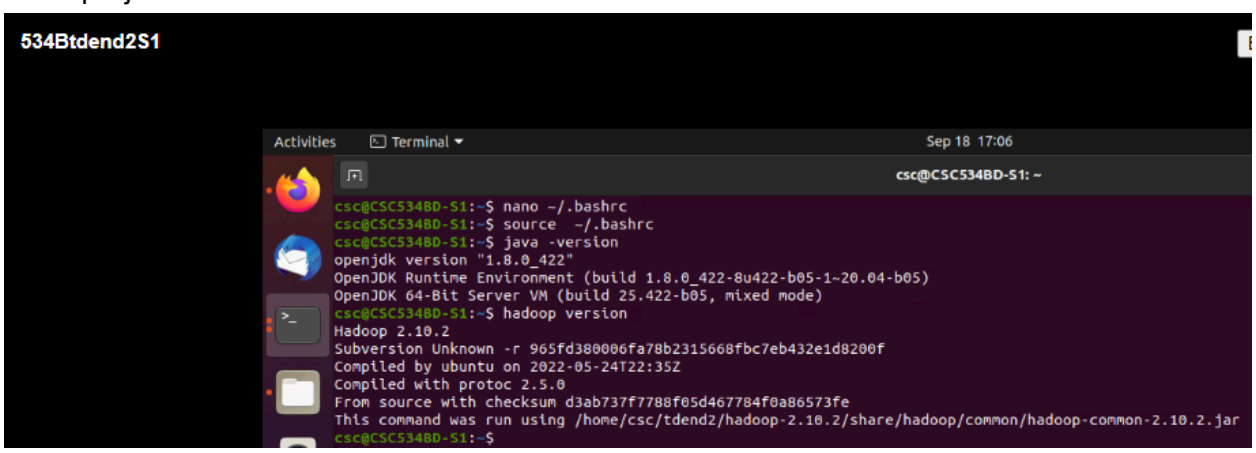
Updating the worker node1:



Installed java(jdk8) on worker node1:



Set the hadoop and java path in bashrc and hadoop-env.shfile and checked the versions of hadoop &java:



<u>Logged into worker node2 using VM and checked the versions of java and hadoop:</u>

534Btdend2S2

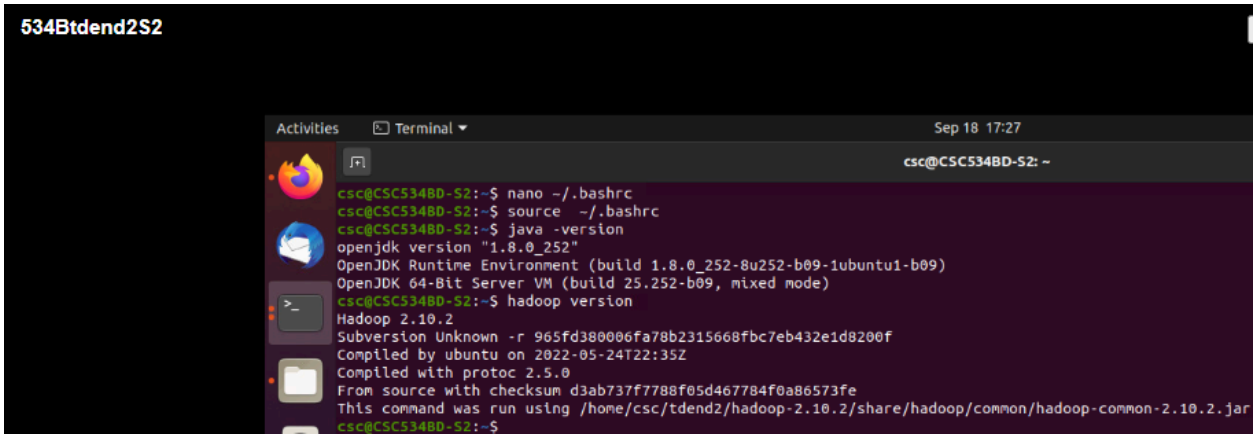
```
Activities   Terminal ▾                                   Sep 18 17:16
                                                          csc@CSC534BD-S2: ~
csc@CSC534BD-S2:~$ ssh csc@10.92.128.57
The authenticity of host '10.92.128.57 (10.92.128.57)' can't be established.
ECDSA key fingerprint is SHA256:dXhKfHsYIXe/53hvU+HOK2V6fVrTbz/QxmUhpnPXpzA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '10.92.128.57' (ECDSA) to the list of known hosts.
csc@10.92.128.57's password:
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 5.4.0-42-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Wed Sep 18 17:16:19 UTC 2024

  System load:  0.04               Processes:               329
  Usage of /:   9.9% of 97.93GB    Users logged in:         1
  Memory usage: 42%                IPv4 address for ens160: 10.92.128.57
  Swap usage:   0%

 * Are you ready for Kubernetes 1.19? It's nearly here! Try RC3 with
   sudo snap install microk8s --channel=1.19/candidate --classic

   https://www.microk8s.io/ has docs and details.

0 updates can be installed immediately.
0 of these updates are security updates.


The list of available updates is more than a week old.
To check for new updates run: sudo apt update

Last login: Fri Jun  5 13:55:11 2020
csc@CSC534BD-S2:~$
```

Installed java:

534Btdend2S2

```
Activities   Terminal ▾                                   Sep 18 17:29
                                                          csc@CSC534BD-S2: ~
csc@CSC534BD-S2:~$ sudo apt install openjdk-8-jdk
[sudo] password for csc:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-
  libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-8-jdk-headle
  x11proto-core-dev x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-8-demo openjdk
  fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-
  libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-8-jdk openjd
  openjdk-8-jre-headless x11proto-core-dev x11proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 21 newly installed, 0 to remove and 6 not upgraded.
Need to get 43.3 MB of archives.
After this operation, 160 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://us.archive.ubuntu.com/ubuntu focal/main amd64 java-common all 0.72 [6816 B]
```

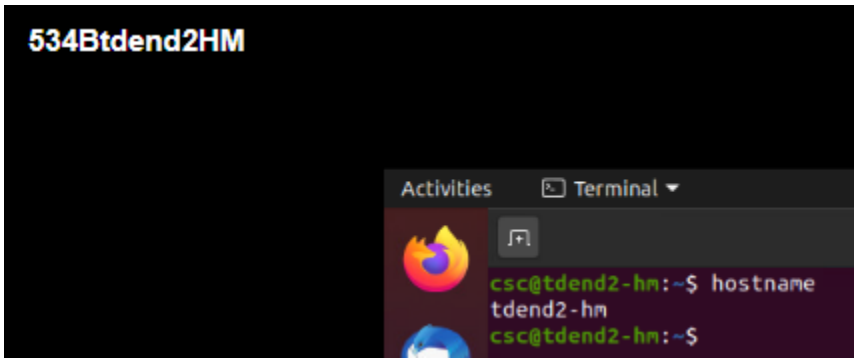
Hadoop version-2.10.2 :



a. Changed and showed the hostnames of three nodes.

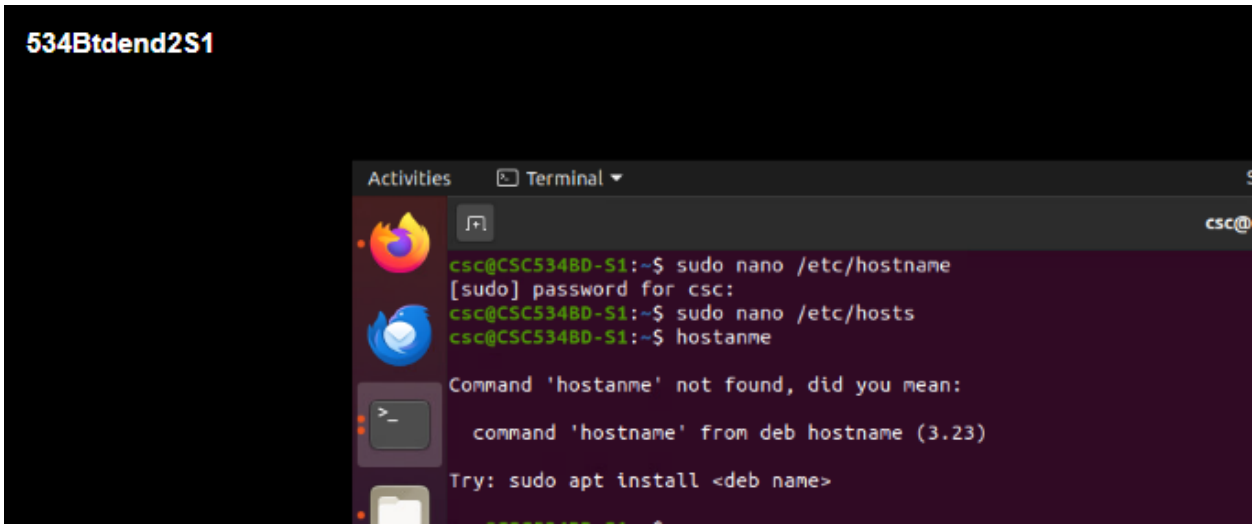
i. Hostnames

1. Master:

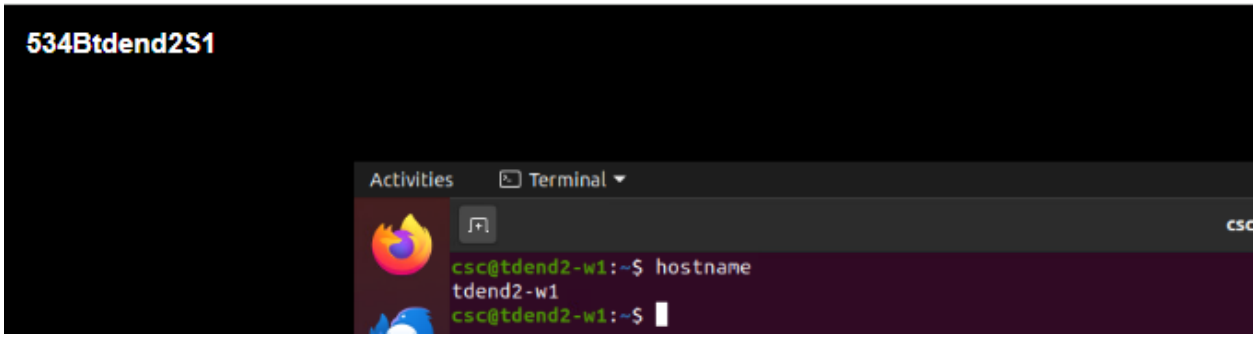


2. Worker1:

Before the host name got reflected:

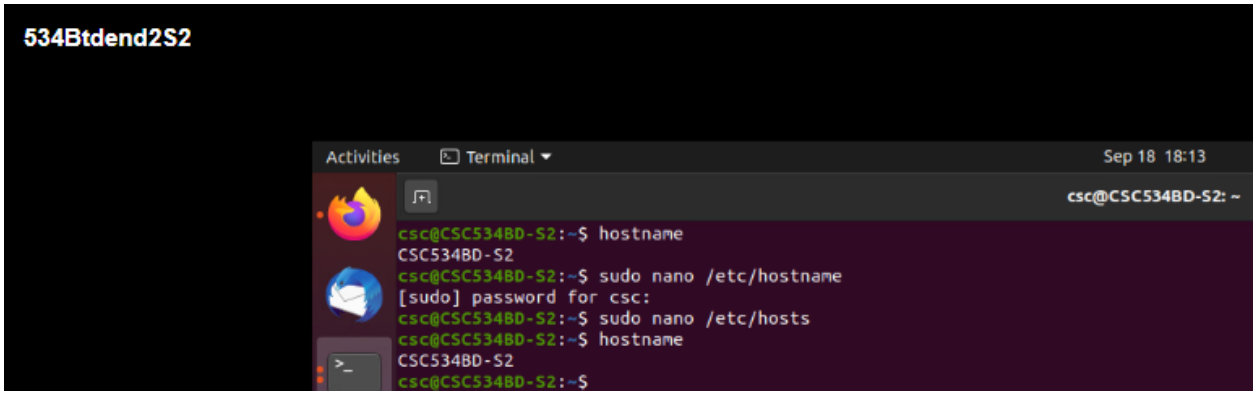


After rebooting,the host name changed successfully:

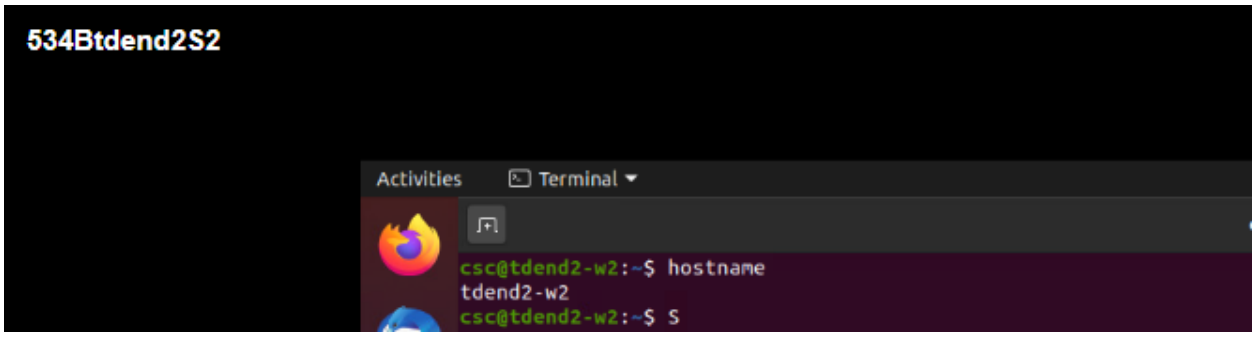


3. Worker2:

Before the name change reflected:



After rebooting, the hostname changed as:



Successfully, hostnames of all 3 nodes changed to tdend2-hm, tdend2-w1, and tdend2-w2 respectively & also IP addresses remained the same after rebooting too.

Reference: https://www.cyberciti.biz/faq/ubuntu-change-hostname-command/

Passwordless SSH connection set up on all nodes:

534Btdend2HM

Enforce US

Activities Terminal Sep 18 19:11

csc@tdend2-hm: ~

```
csc@tdend2-hm:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/csc/.ssh/id_rsa): /home/csc/.ssh/id_rsa
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/csc/.ssh/id_rsa
Your public key has been saved in /home/csc/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:3SlHg6aH/Hlwtk/k7ms8VhnU+BF5zDjLZ2ikpuuRhy0 csc@tdend2-hm
The key's randomart image is:
+---[RSA 3072]----+
|             *+|
|            .  =+=|
|           o o+.=o|
|        . = ooo=.+|
|         S =o*..oo|
|          o.@ + ..|
|           E.=.o. |
|            .= +=  |
|           ..  +=o |
+----[SHA256]-----+
csc@tdend2-hm:~$ ssh-copy-id -i ~/.ssh/id_rsa.pub tdend2-hm
/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/csc/.ssh/id_rsa.pub"
The authenticity of host 'tdend2-hm (10.92.128.52)' can't be established.
ECDSA key fingerprint is SHA256:dXhKfHsYIXe/53hvU+HOK2V6fVrTbz/QxmUhpnPXpzA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any that are already installed
/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it is to install the new keys
csc@tdend2-hm's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'tdend2-hm'"
and check to make sure that only the key(s) you wanted were added.
```

```
csc@tdend2-hm:~$ chmod 0600 ~/.ssh/authorized_keys
csc@tdend2-hm:~$ ls -l ~/.ssh/authorized_keys
-rw------- 1 csc csc 1134 Sep 19 14:17 /home/csc/.ssh/authorized_keys
```

Without password logging master node as " ***ssh 'tdend2-hm'*** "

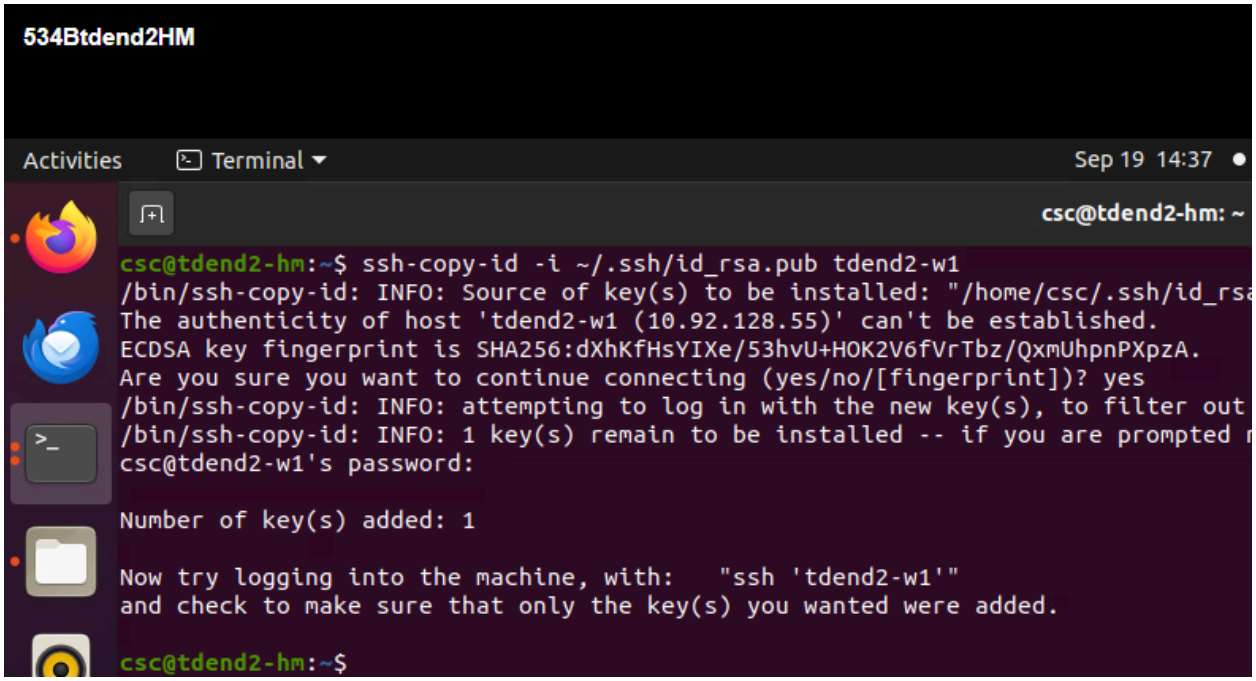
534Btdend2HM

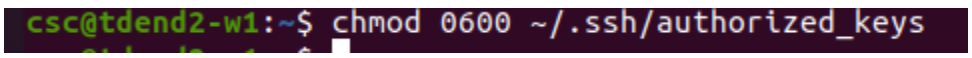
Activities Terminal Sep 18 19:23

csc@tdend2-hm: ~

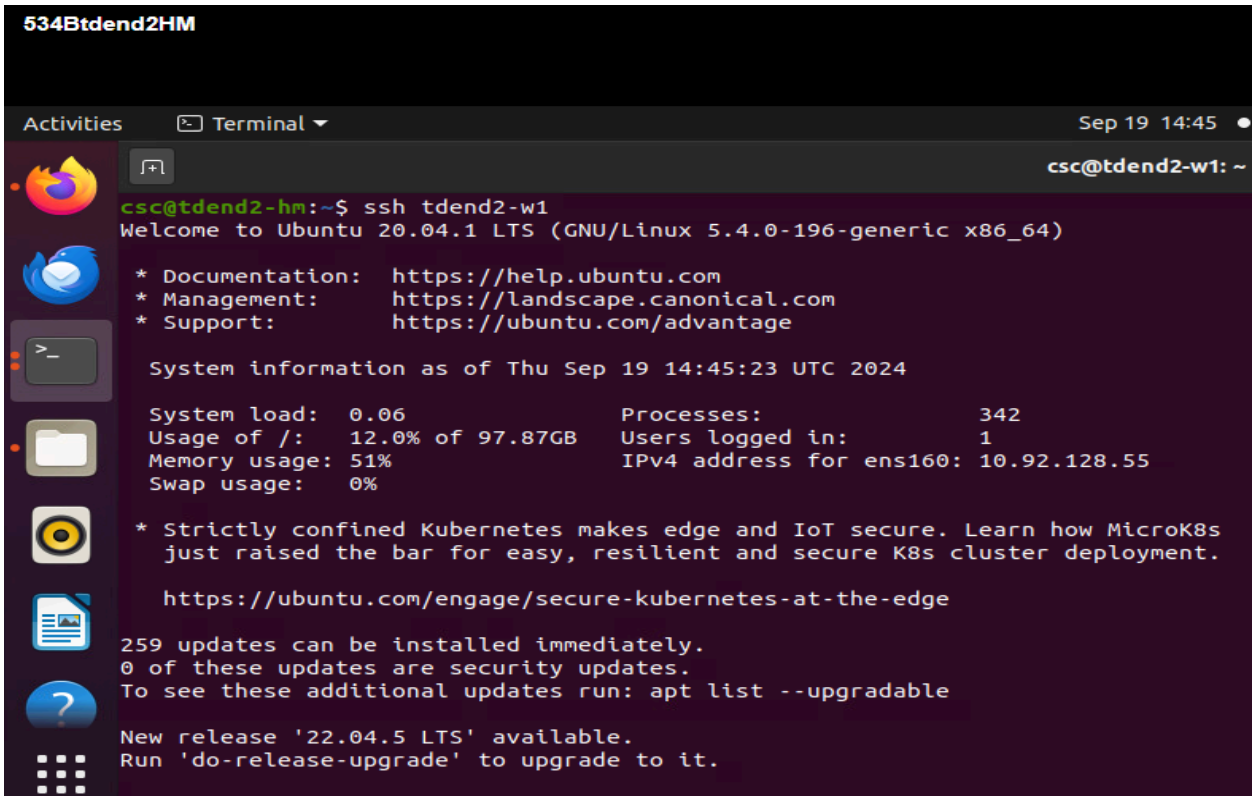
```
csc@tdend2-hm:~$ ssh 'tdend2-hm'
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 5.4.0-196-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Wed Sep 18 19:23:29 UTC 2024

  System load:  0.02               Processes:               323
  Usage of /:   13.4% of 97.87GB   Users logged in:         1
  Memory usage: 42%                IPv4 address for ens160: 10.92.128.52
  Swap usage:   0%

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

444 updates can be installed immediately.
212 of these updates are security updates.
To see these additional updates run: apt list --upgradable

New release '22.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Wed Sep 18 14:36:24 2024 from 192.168.23.239
csc@tdend2-hm:~$
```

Worker node1:





Without password logging master node as " ***ssh 'tdend2-w1'*** "



Workernode2 ssh set up:

```
534Btdend2HM

Activities   Terminal ▾                                   Sep 19 14:39 ●
                                                          csc@tdend2-hm: ~
csc@tdend2-hm:~$ ssh-copy-id -i ~/.ssh/id_rsa.pub tdend2-w2
/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/csc/.ssh/id_rsa.pub"
The authenticity of host 'tdend2-w2 (10.92.128.57)' can't be established.
ECDSA key fingerprint is SHA256:dXhKfHsYIXe/53hvU+HOK2V6fVrTbz/QxmUhpnPXpzA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter out any th
/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted now it
csc@tdend2-w2's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'tdend2-w2'"
and check to make sure that only the key(s) you wanted were added.
```

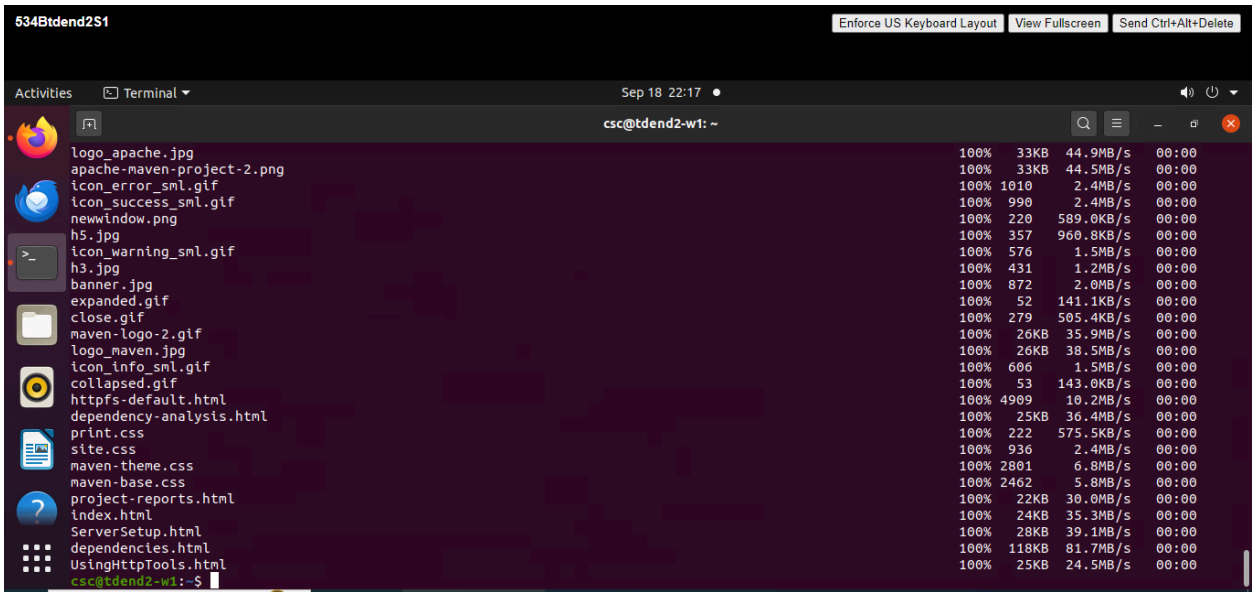
```
csc@tdend2-w2:~$ chmod 0600 ~/.ssh/authorized_keys
```

Without password logging master node as " ***ssh 'tdend2-w2'*** "

```
534Btdend2HM

Activities   Terminal ▾                                   Sep 19 14:52 ●
                                                          csc@tdend2-w2: ~
csc@tdend2-hm:~$ ssh tdend2-w2
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 5.4.0-42-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Thu Sep 19 14:52:15 UTC 2024

  System load:  0.0                Processes:               334
  Usage of /:   12.6% of 97.93GB   Users logged in:         1
  Memory usage: 47%                IPv4 address for ens160: 10.92.128.57
  Swap usage:   1%

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

259 updates can be installed immediately.
0 of these updates are security updates.
To see these additional updates run: apt list --upgradable

New release '22.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
```

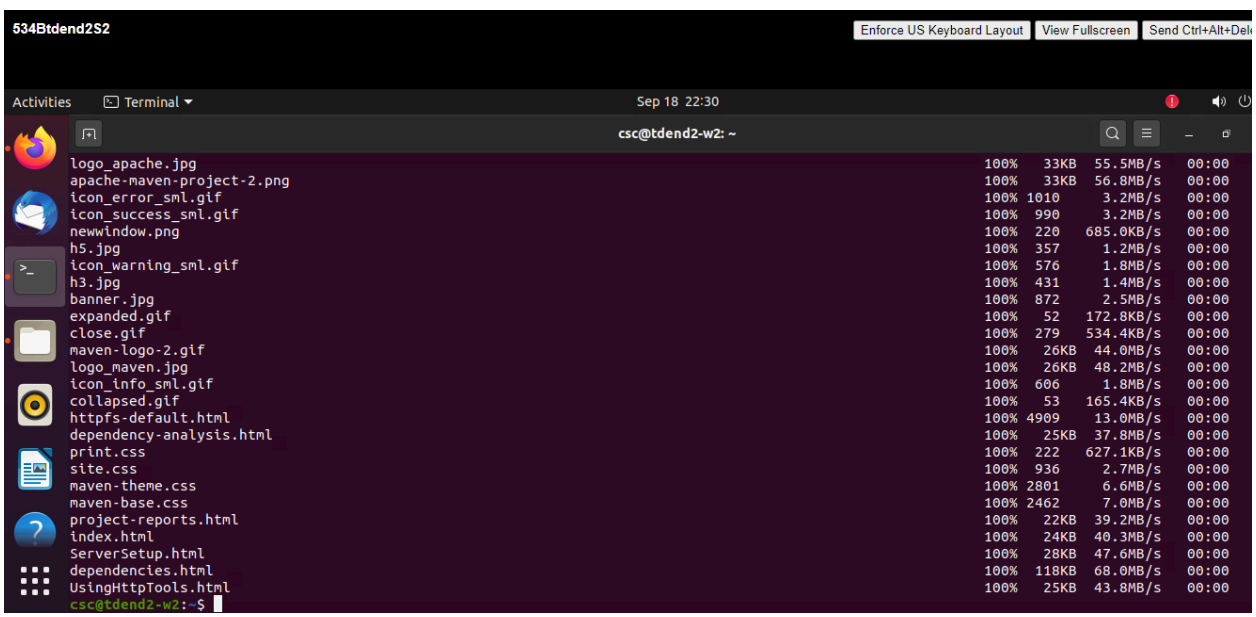
Setting up hadoop on workers - for example, the configuration files etc.

Using ***scp –r /home/csc/tdend2/hadoop-2.10.2 tdend2-w1:/home/csc/tdend2***



On worker node 2:

```
csc@tdend2-w2:~$ scp -r  /home/csc/tdend2/hadoop-2.10.2  tdend2-w2:/home/csc/tdend2
```

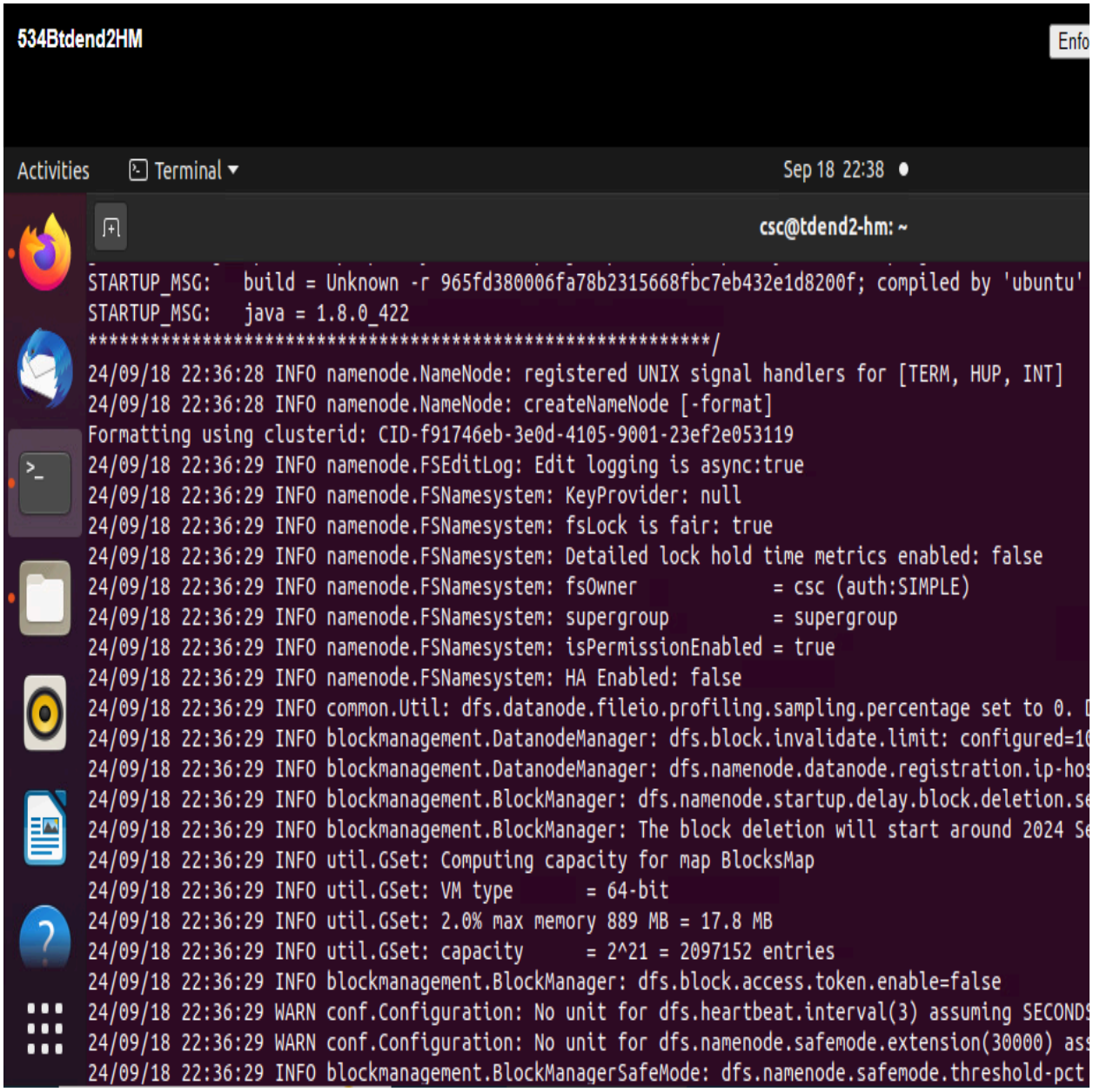


Config files are copied successfully.

Formatting HDFS file system via the namemode

• On master node, typed the following command: ***hadoop namenode -format***

This will format HDFS and creates the directory specified by *dfs.name.dir property* in the *hdfs-site.xml file.* We need to do this only the first time you run a new cluster. Otherwise, if we format a running cluster we will lose all the data in HDFS



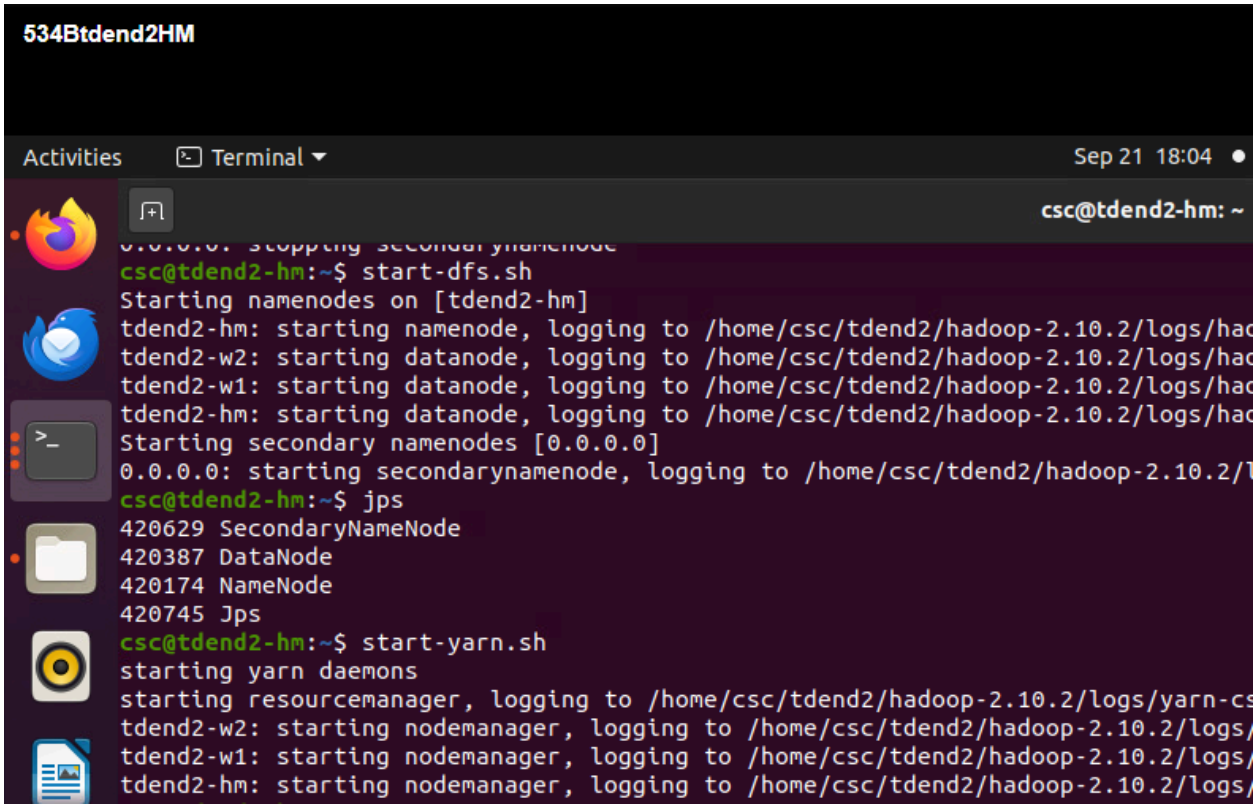
For the cluster to start successfully, java_home , hadoop_home, hadoop_conf_dir should be set with correct paths onhadoop_env.sh file.

>Starting HDFS

• To start hdfs, ran the command ***start-dfs.sh*** on your master node:

Data node worked without exiting in few seconds on port 50030 which is after trying to start it on multiple ports like 50010, 50020.

For data node to run on worker nodes, name node which is on master node should be referred correctly in core-site.xml.



• The namenode and secondary namenode daemons will be running on the master node and the datanode daemon will run on all nodes. To verify , used **jps** command to get all the running java processes on your master and worker node:

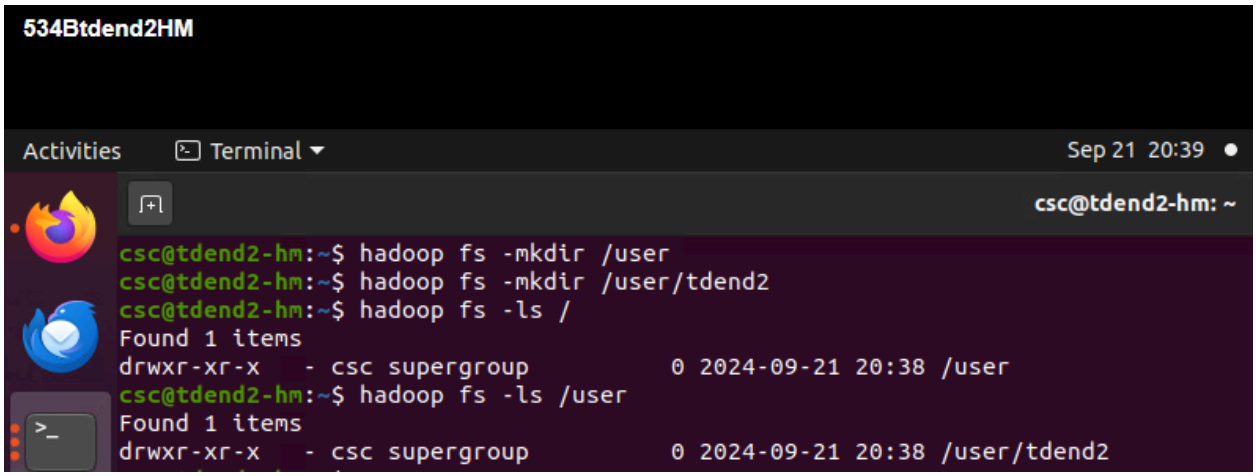
After starting yarn on master node:

On workernodes ran jps:

On worker node2 :

---

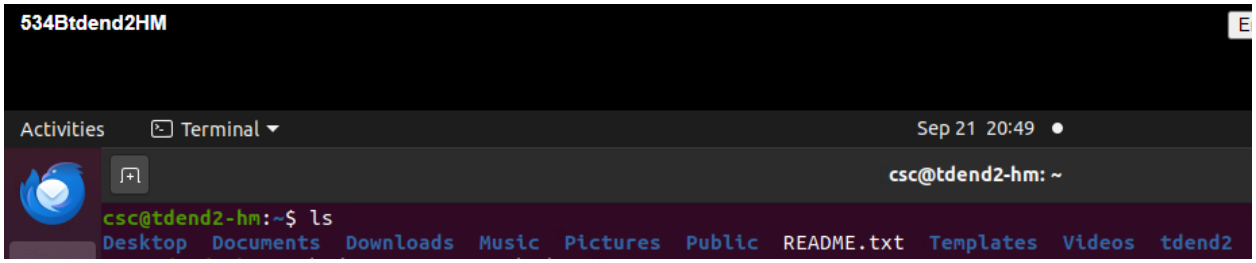
b. By creating your user directory



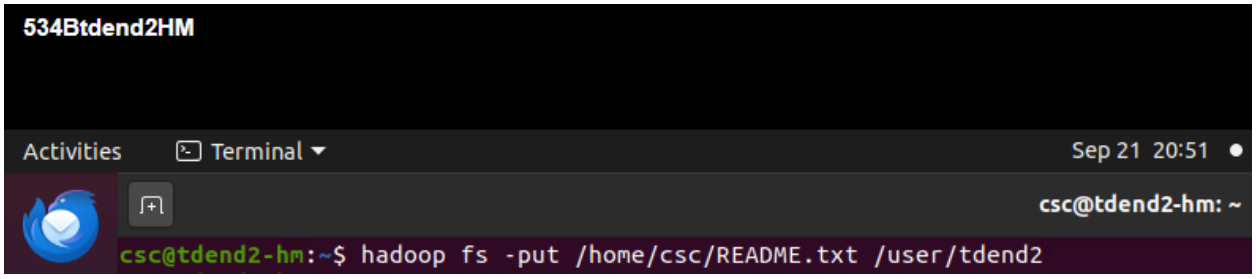
Created a user directory and tdend2 within it and verified using ls command and found that it has been created with supergroup permissions and made accessible to owner , group and others to read and execute.

---

c. Uploaded /put a file from the local (VM's,not laptop) Linux filesystem to HDFS.



Created README.txt on VM and uploaded to HDFS as shown below:

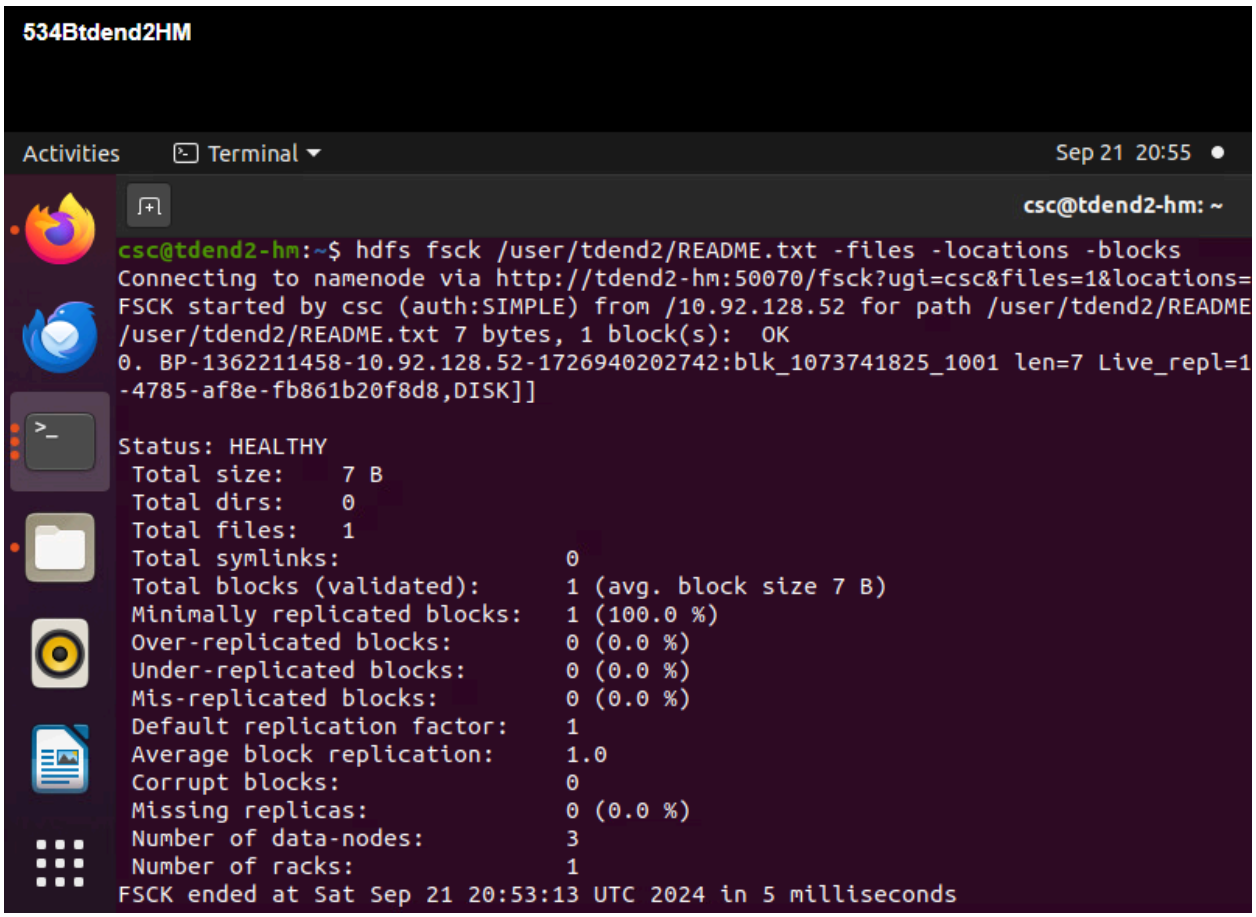


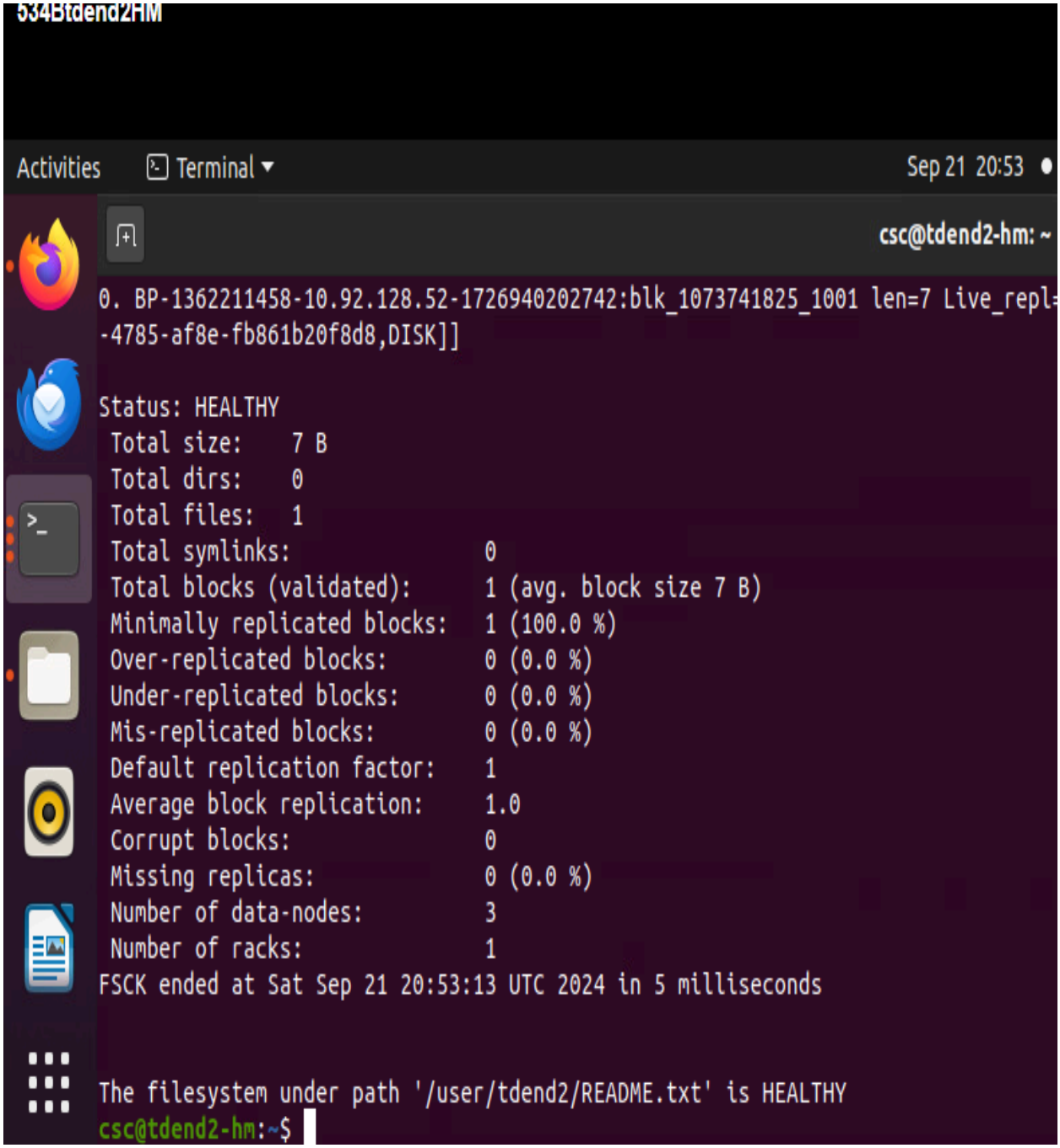
—----------

d. By running the fsck command. Checked the file.

The hdfs fsck command is a tool used in HDFS (Hadoop Distributed File System) to check the health of the file system and diagnose any issues. hdfs fsck can be used to check the consistency of file system metadata, such as block placement and replication, and to detect and correct any inconsistencies.

$ hdfs fsck /user/tdend2/README.txt -files -locations -blocks





The file at the given path is healthy as shown above and found that there are 3 data nodes, no missing replicas, no corrupt blocks, with total number of blocks accounting to 1 with 1 rack availability.

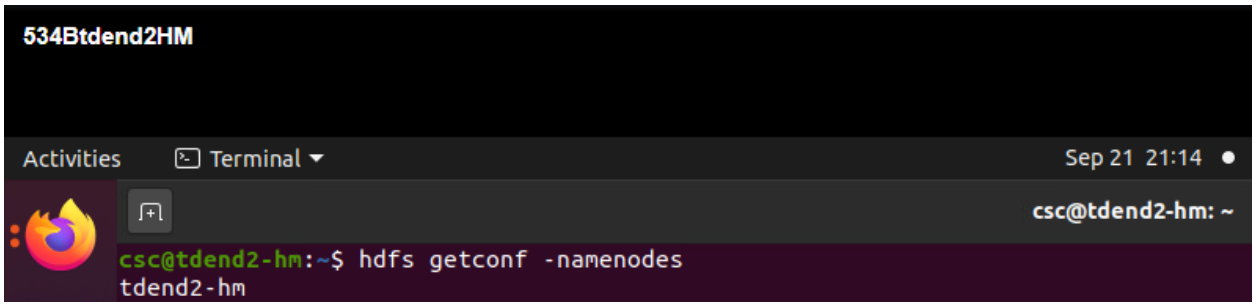
Files, blocks, locations option display detailed report of each file, blocks within HDFS and the location of each block respectively.

2.Exploring Hadoop User Interfaces using a web browser. If we start Hadoop, we see several web user interfaces (UIs).

a. Namenode web UI

i. Checked the number of namenode(s) my cluster has as shown below.

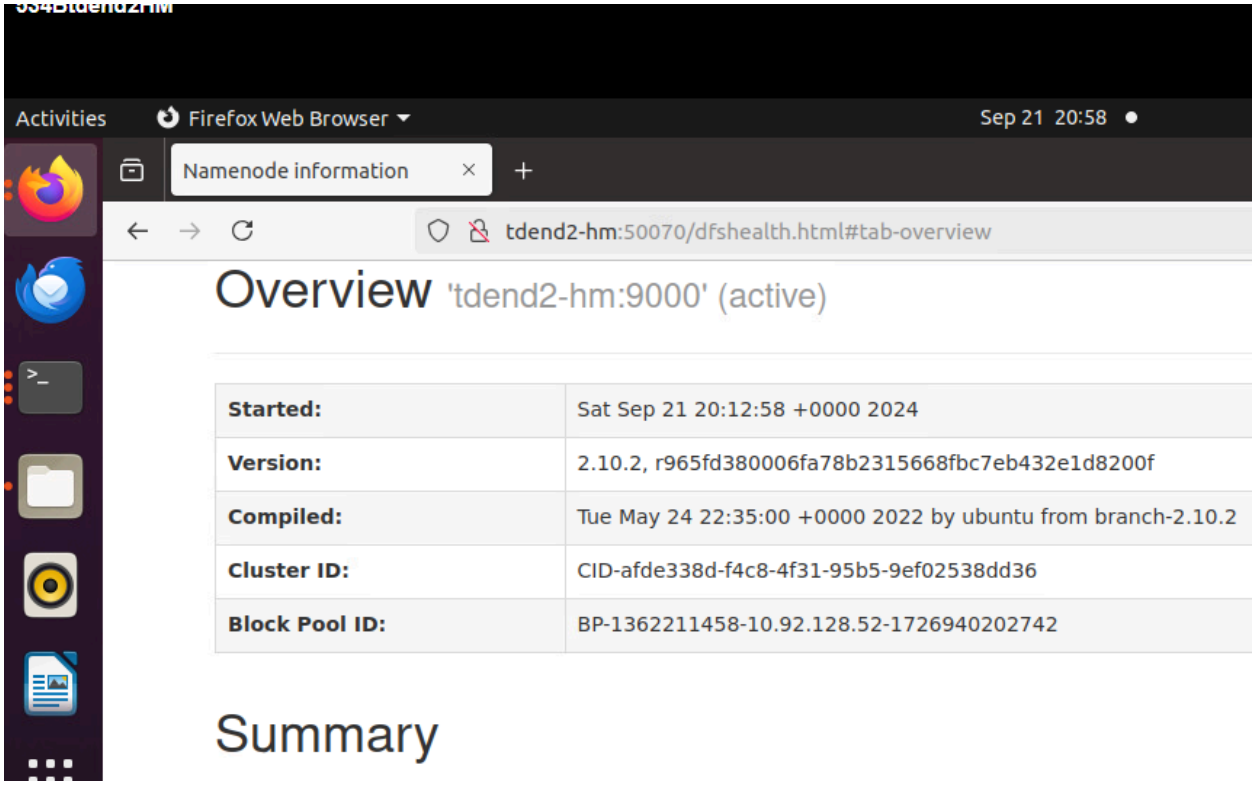
hdfs getconf -namenodes displays the list of active name nodes in a cluster.

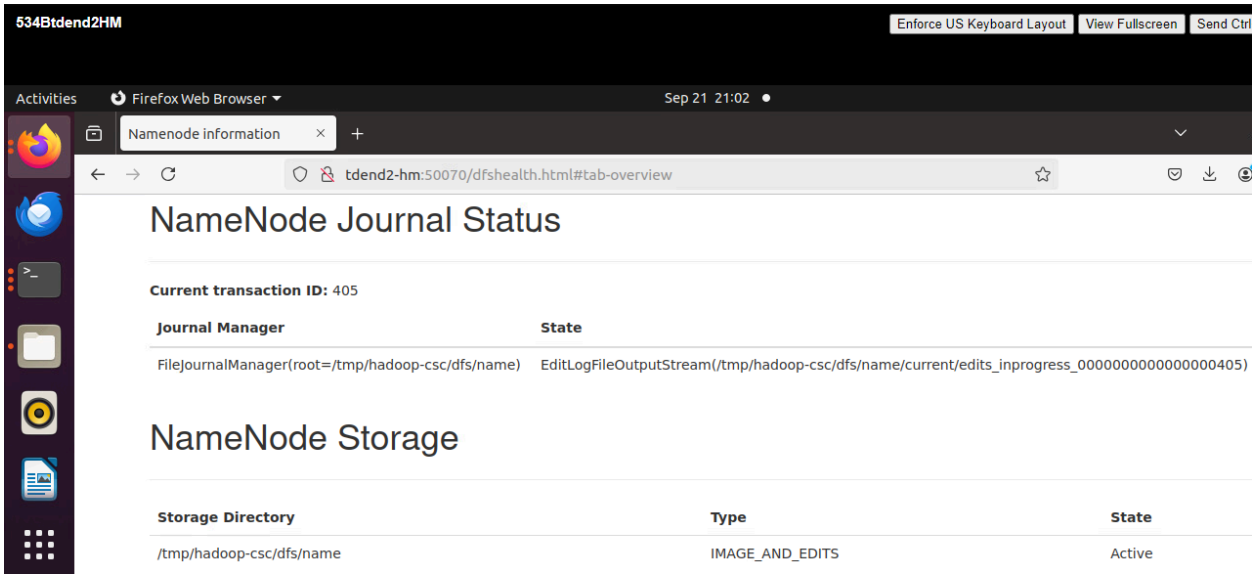


One name node is found on the cluster which is running on master node named tdend2-hm at port 9000 as shown below.

ii. Navigated to http://<nodename(s)>:50070 as ***http://tdend2-hm:50070***

The UI at the above url displayed the overview and summary of namenode information of cluster.



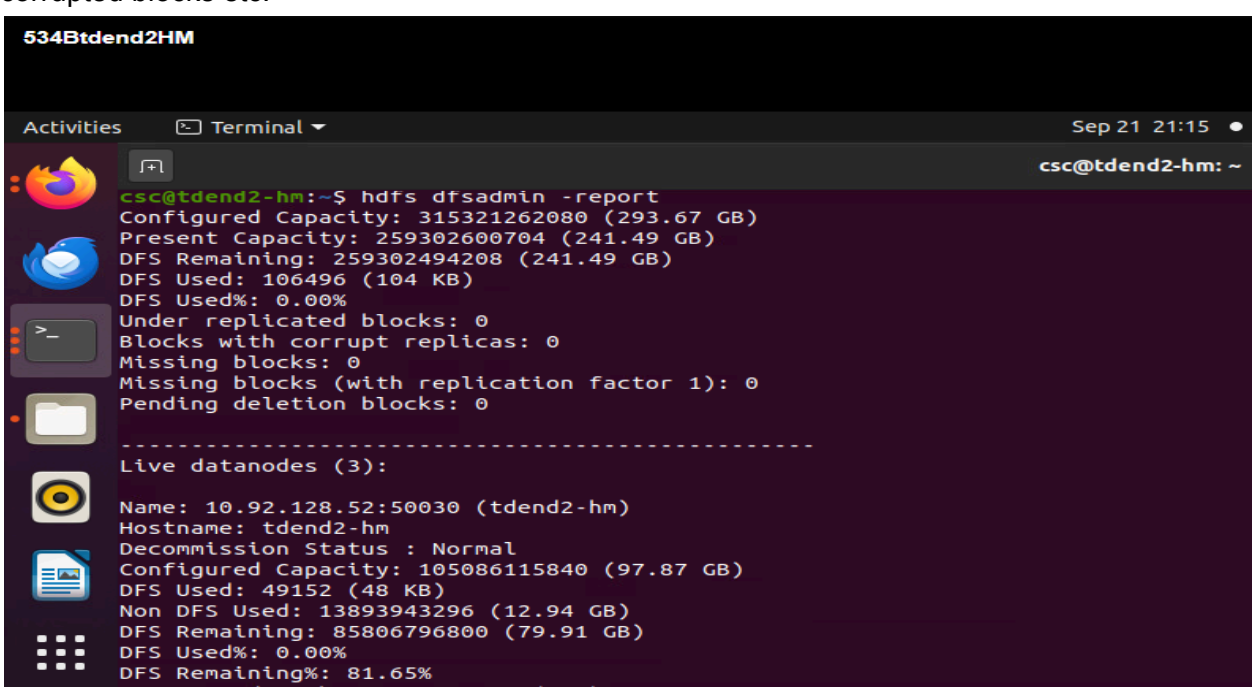


---

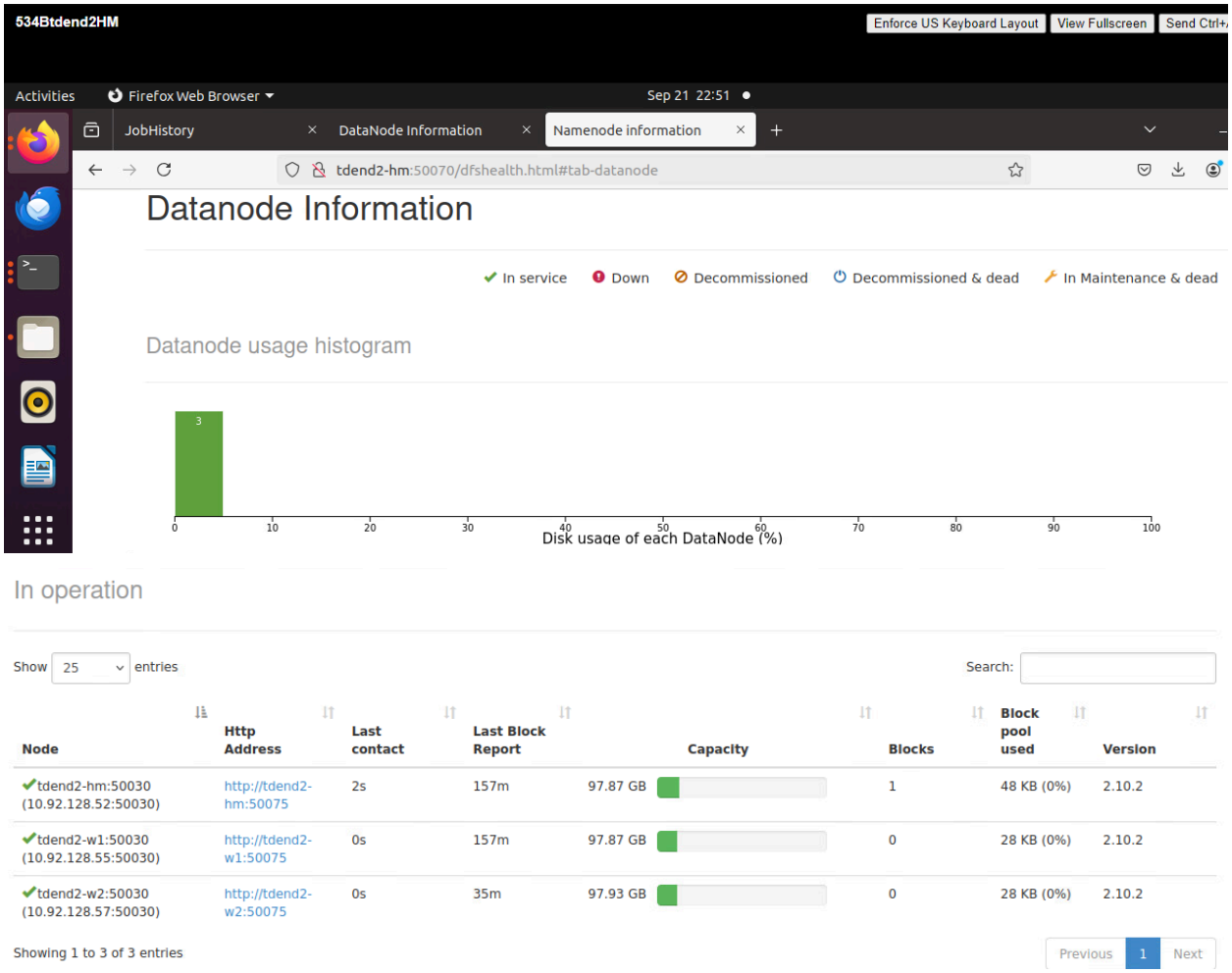
b. Datanode web UI

i. Checked the number of datanode(s) that my cluster has which is shown as below:

hdfs dfsadmin -report outputs a brief report on the overall HDFS filesystem. It's a useful command to quickly view how much disk is available, how many DataNodes are running, corrupted blocks etc.

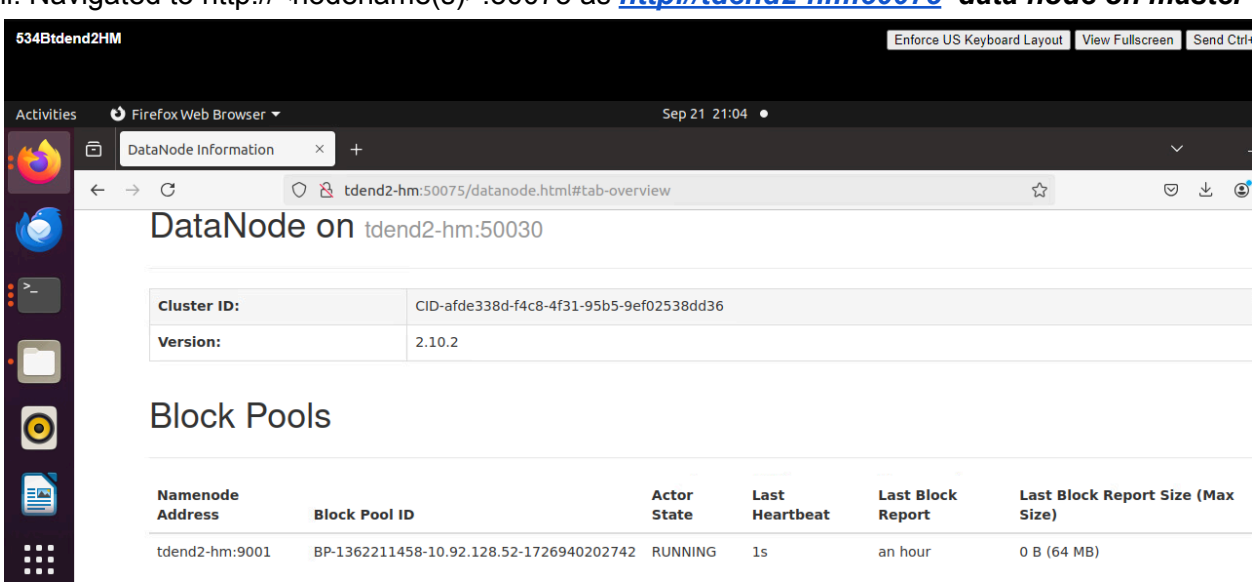


The cluster has 3 active data nodes - one on master, one on each worker node.

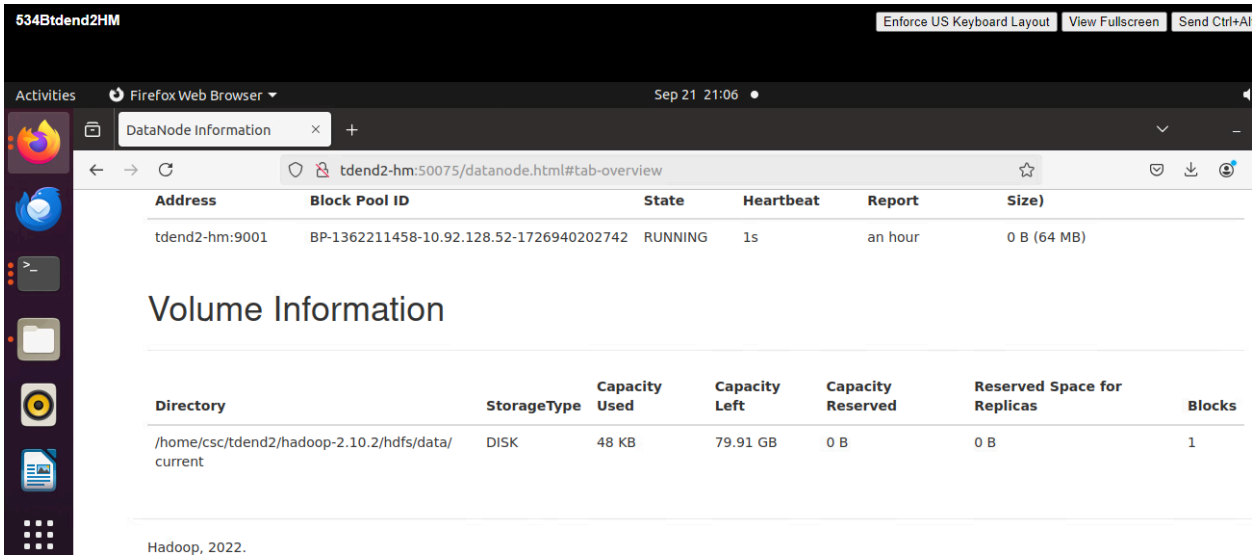


---

ii. Navigated to http://<nodename(s)>:50075 as ***http://tdend2-hm:50075*** ***data node on master***

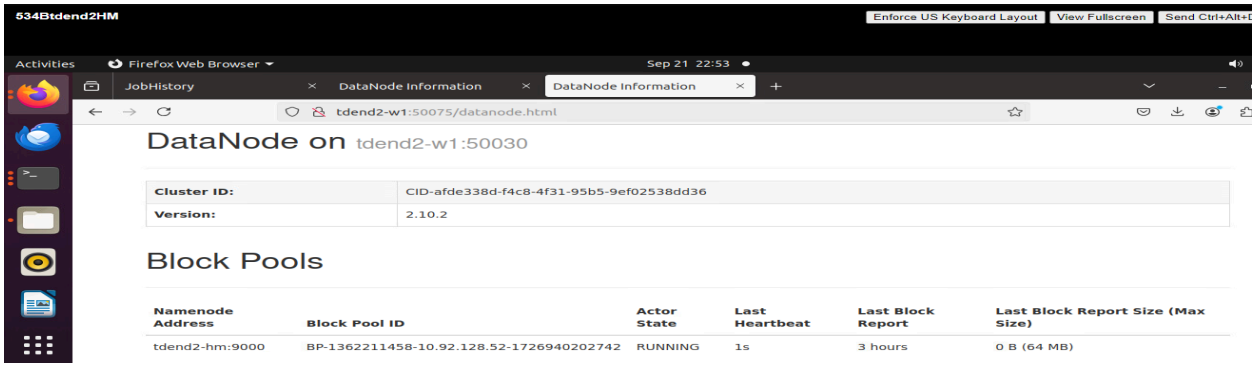


Last half of overview of data node information:

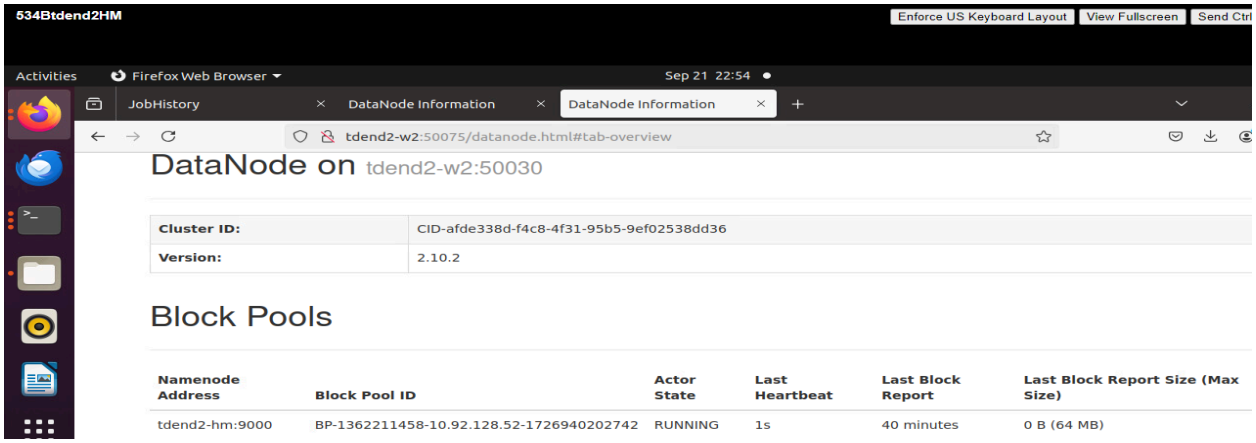


On master node node, 1 data node is available.

Data node UI interface of worker node 1: at http://tdend2-w1:50075



Data node WEB UI of worker node 2: **http://tdend2-w1:50075**

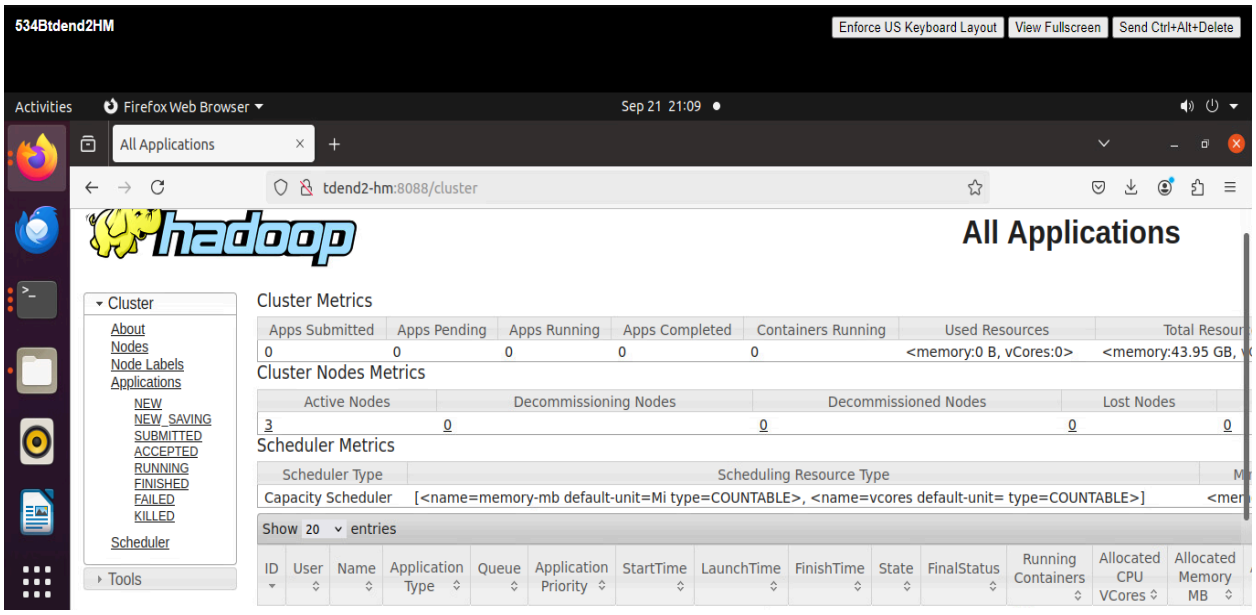


---

c. Resource Manager UI

i. Checked the master node's name

ii. Navigated to http://<nodename>:8088 which is ***http://tdend2-hm:8088***



Resource manager shows 3 active nodes in the cluster as seen from the cluster nodes metrics.

The Resource Manager (RM) is a Java process that manages resources in a Hadoop cluster, such as CPU, memory, and disk. It's the master daemon of YARN, which stands for "Yet Another Resource Negotiator". The RM works with other components to manage resources, including:

NodeManagers

These per-node frameworks manage resources on a single node and take instructions from the RM.

<u>ApplicationMasters</u>

These per-application components negotiate resources with the RM and work with NodeManagers to start containers.

<u>The RM's responsibilities include:</u>

Resource allocation

The RM receives resource requests from application masters and allocates resources to run the applications.

Cluster health monitoring

The RM monitors the health of nodes in the cluster and manages resource failover if a node fails.

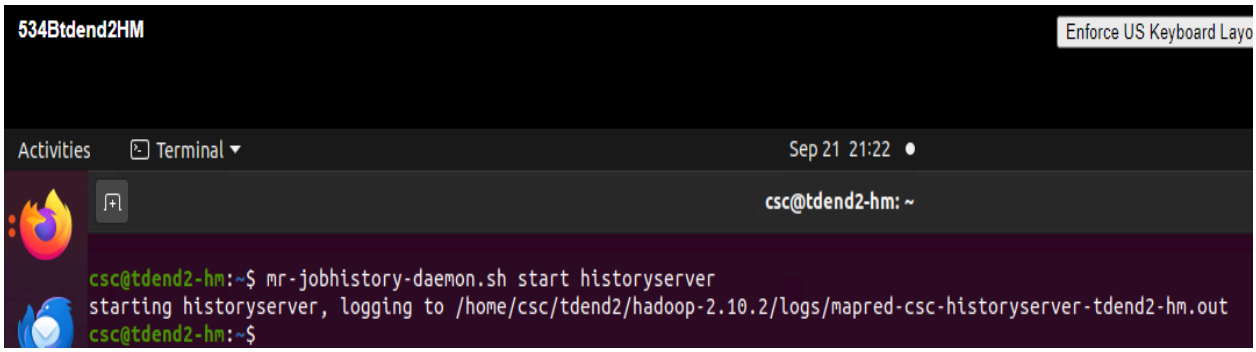
Cluster resource tracking

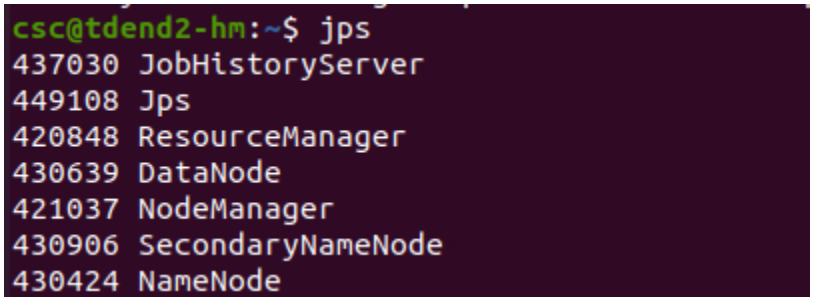
The RM keeps track of available resources on each node and maintains a global view of the cluster. The RM usually runs on the head node of the Hadoop cluster.

---

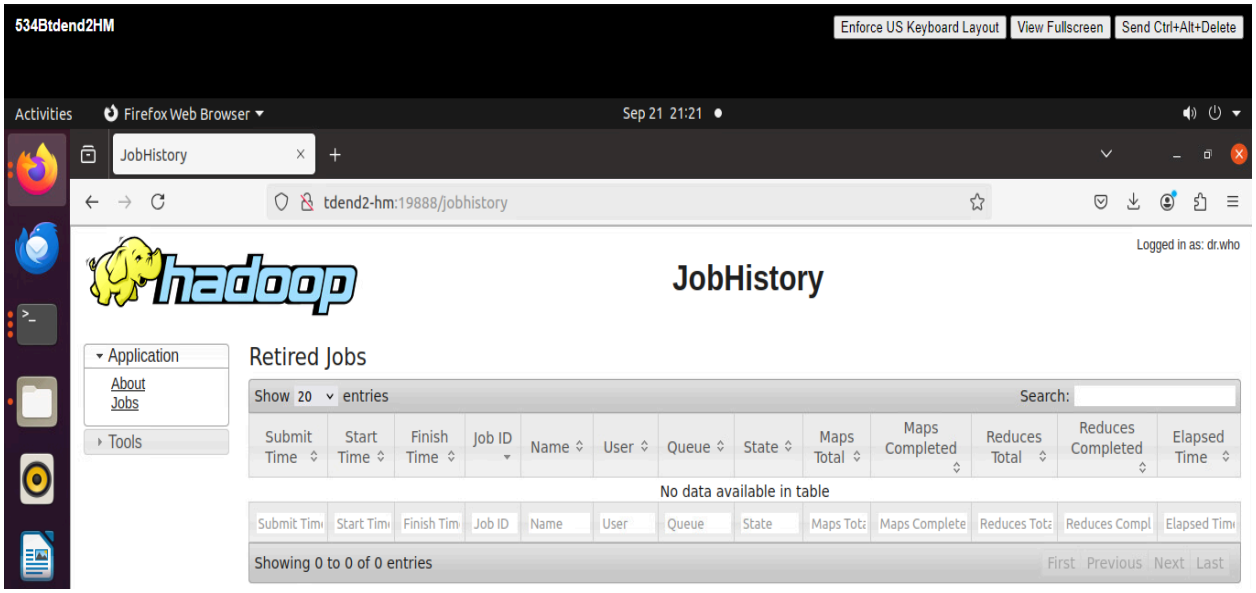
d. Started **MapReduce JobHistory Server UI** using the below command:
mr-jobhistory-daemon.sh start historyserver

i. Checked the master node's name





ii. Navigated to http:// <nodename>:19888 as ***http://tdend2-hm:19888***



JobHistoryServer is responsible for servicing all job history related requests from client.

The history server REST API's allow the user to get status on finished applications.

Verification if the cluster is running correctly:

```
csc@tdend2-hm:~$ yarn node -list
24/09/21 23:10:07 INFO client.RMProxy: Connecting to ResourceManager at tdend2-hm/10.92.128.52:8032
Total Nodes:3
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
 tdend2-w2:45555               RUNNING    tdend2-w2:8042                                   0
 tdend2-hm:45049               RUNNING    tdend2-hm:8042                                   0
 tdend2-w1:34257               RUNNING    tdend2-w1:8042                                   0
```

Finally stopped df,yarn,job history servers.

—----------THE END—----------