

Building a Hadoop cluster

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

Firstly, logging in and installing java on all 3 nodes:

Accessing csccluster.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
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-java-jni lib
  libxdmcp-dev libxt-dev openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-core-de
Suggested packages:
  default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-8-demo openjdk-8-source vis
  fonts-wqy-zenhei
The 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
The following packages will be upgraded:
  libx11-6
1 upgraded, 21 newly installed, 0 to remove and 736 not upgraded.

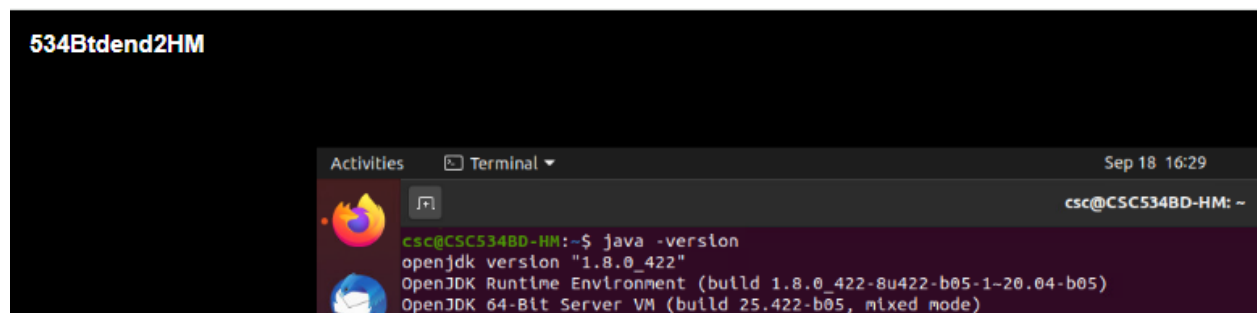
```

```

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:

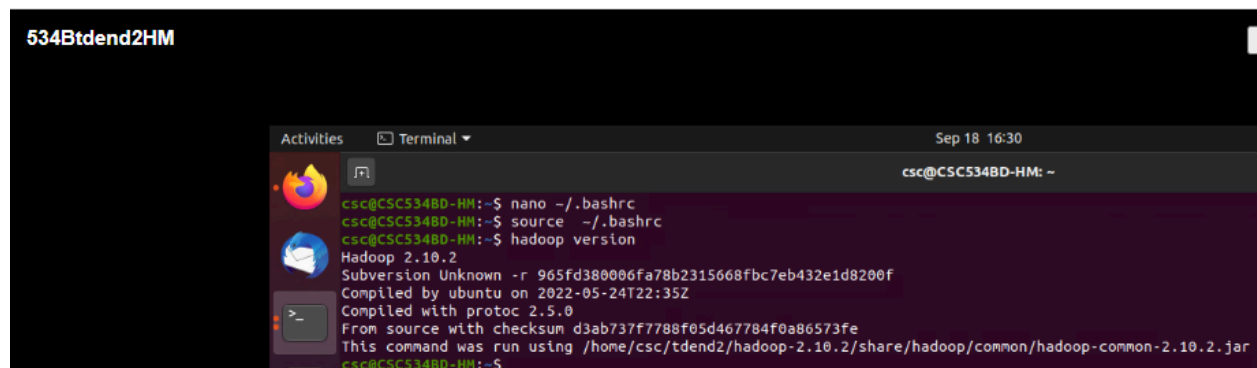


```

534Btdend2HM
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
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:

534Btdend2S1

```
Activities Terminal Sep 18 16:39
csc@CSC534BD-S1: ~
csc@CSC534BD-S1:~$ ssh csc@10.92.128.55
The authenticity of host '10.92.128.55 (10.92.128.55)' can't be established.
ECDSA key fingerprint is SHA256:dXhKfHsYIXe/53hvU+HOK2V6fVrTbz/QxmUhpNXPzA.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '10.92.128.55' (ECDSA) to the list of known hosts.
csc@10.92.128.55'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 16:39:28 UTC 2024

System load:  0.04               Processes:    357
Usage of /:   10.9% of 97.93GB   Users logged in: 1
Memory usage: 22%               IPv4 address for ens160: 10.92.128.55
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-S1:~$
```

Updating the worker node1:

534Btdend2S1

```
Activities Terminal Sep 18 16:42
csc@CSC534BD-S1: ~
csc@CSC534BD-S1:~$ 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]
```

Installed java(jdk8) on worker node1:

534Btdend2S1

```
Activities Terminal Sep 18 16:45
csc@CSC5348D-S1: ~
csc@CSC5348D-S1:~$ sudo apt install openjdk-8-jdk
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-
  libasm-dev libx11-6 libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev op
  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
  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-
  libasm-dev libx11-6 libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev op
  openjdk-8-jre-headless x11proto-core-dev x11proto-dev xorg-sgml-doctools xtrans-
The following packages will be upgraded:
  libx11-6
1 upgraded, 21 newly installed, 0 to remove and 736 not upgraded.
```

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

534Btdend2S1

```
Activities Terminal Sep 18 17:06
csc@CSC5348D-S1: ~
csc@CSC5348D-S1:~$ nano ~/.bashrc
csc@CSC5348D-S1:~$ source ~/.bashrc
csc@CSC5348D-S1:~$ 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)
csc@CSC5348D-S1:~$ hadoop version
Hadoop 2.10.2
Subversion Unknown -r 965fd38006fa78b2315668fbc7eb432e1d8200f
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@CSC5348D-S1:~$
```

Logged into worker node2 using VM and checked the versions of java and hadoop:

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-java-gtk
  libasm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-8-jdk-headless
  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-8-jre
  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-java-gtk
  libasm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-8-jdk openjdk-8-jdk-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 :

534Btdend2S2

```
Activities Terminal Sep 18 17:27
csc@CSC534BD-S2: ~
csc@CSC534BD-S2:~$ nano ~/.bashrc
csc@CSC534BD-S2:~$ source ~/.bashrc
csc@CSC534BD-S2:~$ java -version
openjdk version "1.8.0_252"
OpenJDK Runtime Environment (build 1.8.0_252-8u252-b09-1ubuntu1-b09)
OpenJDK 64-Bit Server VM (build 25.252-b09, mixed mode)
csc@CSC534BD-S2:~$ hadoop version
Hadoop 2.10.2
Subversion Unknown -r 965fd38006fa78b2315668fbc7eb432e1d8200f
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-S2:~$
```

a. Changed and showed the hostnames of three nodes.

i. Hostnames

1. Master:

534Btdend2HM

```
Activities Terminal
csc@tdend2-hm:~$ hostname
tdend2-hm
csc@tdend2-hm:~$
```

2. Worker1:

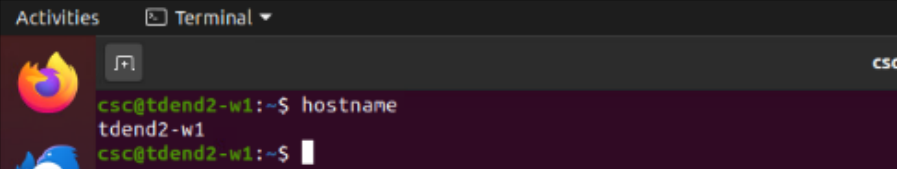
Before the host name got reflected:

534Btdend2S1

```
Activities Terminal
csc@CSC534BD-S1:~$ sudo nano /etc/hostname
[sudo] password for csc:
csc@CSC534BD-S1:~$ sudo nano /etc/hosts
csc@CSC534BD-S1:~$ hostname
Command 'hostname' not found, did you mean:
  command 'hostname' from deb hostname (3.23)
Try: sudo apt install <deb name>
```

After rebooting, the host name changed successfully:

534Btdend2S1

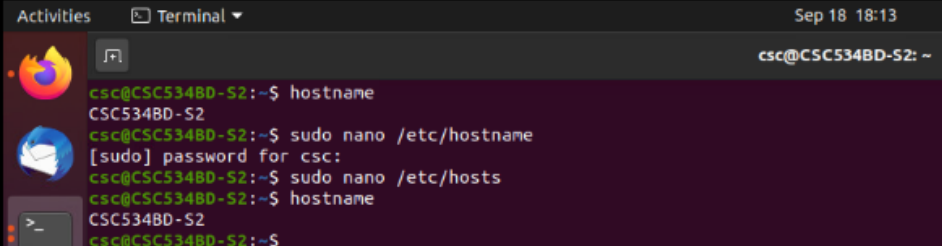


```
Activities Terminal
csc@tdend2-w1:~$ hostname
tdend2-w1
csc@tdend2-w1:~$
```

3. Worker2:

Before the name change reflected:

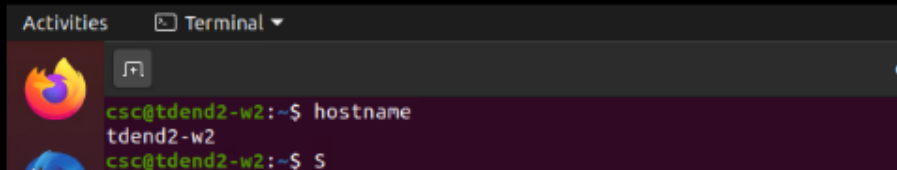
534Btdend2S2



```
Activities Terminal Sep 18 18:13
csc@CSC5348D-S2:~$ hostname
CSC5348D-S2
csc@CSC5348D-S2:~$ sudo nano /etc/hostname
[sudo] password for csc:
csc@CSC5348D-S2:~$ sudo nano /etc/hosts
csc@CSC5348D-S2:~$ hostname
CSC5348D-S2
csc@CSC5348D-S2:~$
```

After rebooting, the hostname changed as:

534Btdend2S2



```
Activities Terminal
csc@tdend2-w2:~$ hostname
tdend2-w2
csc@tdend2-w2:~$
```

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:


```

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:
SHA256:35LHg6aH/Hlwtk/k7ms8VhnU+BF5zDjLZ2ikpuuRhy0 csc@tdend2-hm
The key's randomart image is:
+----[RSA 3072]-----+
  .          +
  .  =+      +
  O 0+.=0    +
  . = 000=.+
  S =0*..00
  O.@ + ..
  E.=.0.
  . = +
  ..  +=0
+----[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’**”

```

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:

```
534Btdend2HM

Activities Terminal Sep 19 14:37
csc@tdend2-hm: ~

csc@tdend2-hm:~$ ssh-copy-id -i ~/.ssh/id_rsa.pub tdend2-w1
/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/csc/.ssh/id_rsa"
The authenticity of host 'tdend2-w1 (10.92.128.55)' 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
/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted
csc@tdend2-w1's password:

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'tdend2-w1'"
and check to make sure that only the key(s) you wanted were added.

csc@tdend2-hm:~$

csc@tdend2-w1:~$ chmod 0600 ~/.ssh/authorized_keys
```

Without password logging master node as “ssh ‘tdend2-w1’”

```
534Btdend2HM

Activities Terminal Sep 19 14:45
csc@tdend2-w1: ~

csc@tdend2-hm:~$ ssh tdend2-w1
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 Thu Sep 19 14:45:23 UTC 2024

System load:  0.06               Processes:            342
Usage of / :  12.0% of 97.87GB   Users logged in:     1
Memory usage: 51%               IPv4 address for ens160: 10.92.128.55
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

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.
```

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+H0K2V6fVrTbz/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.

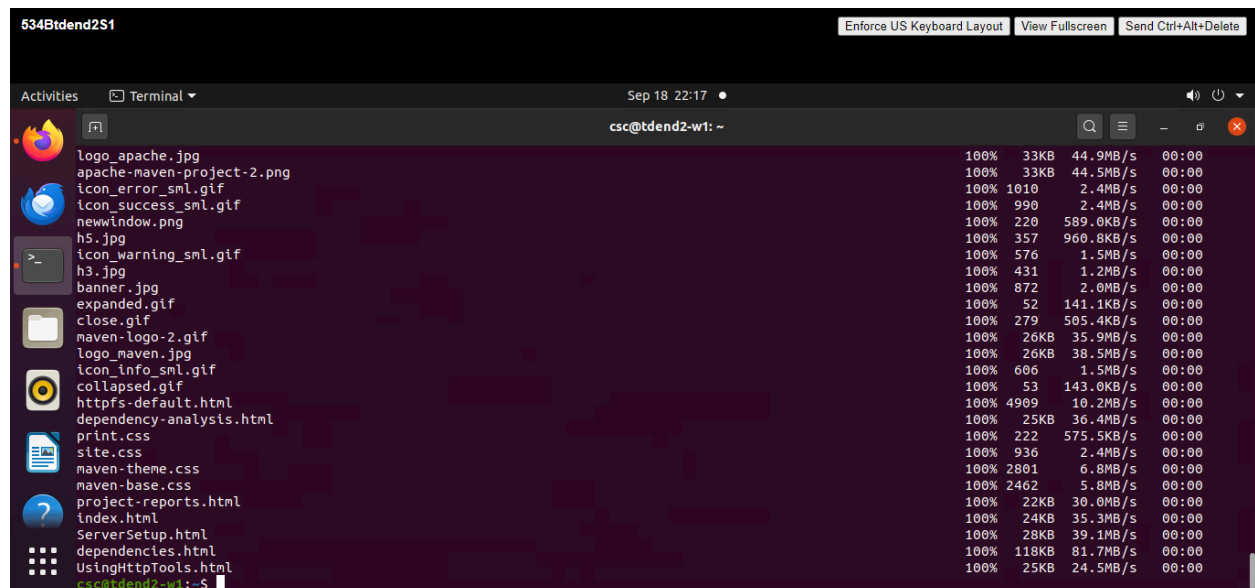
https://ubuntu.com/engage/secure-kubernetes-at-the-edge

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

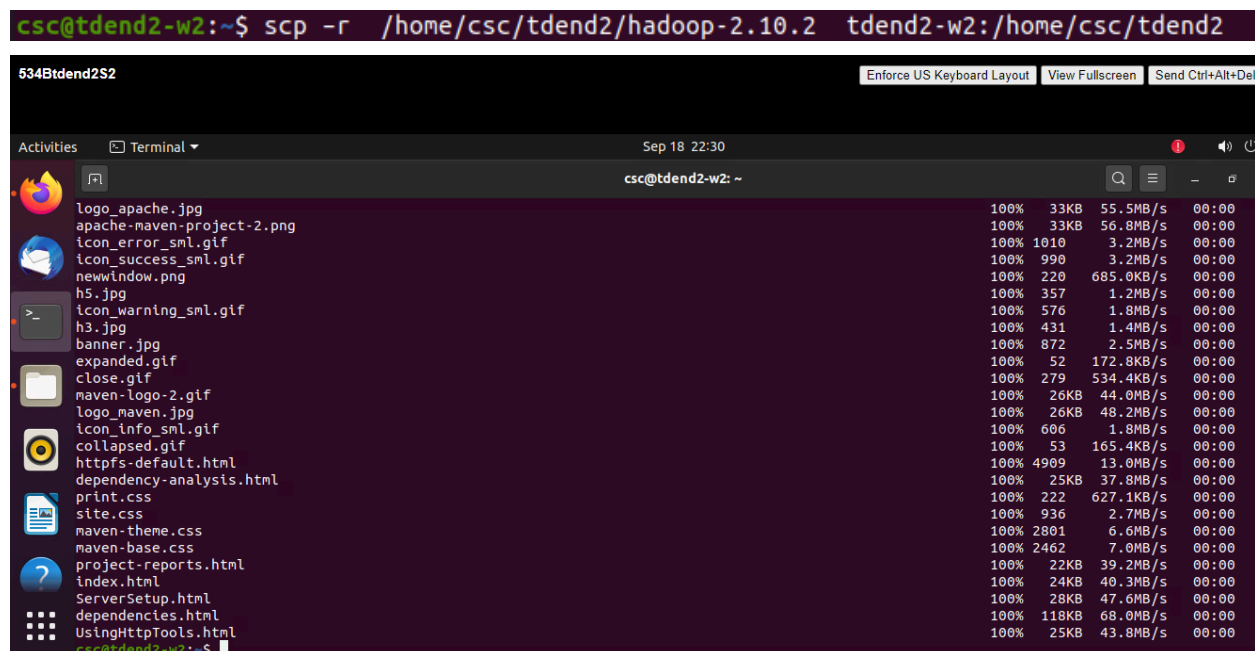
New release '22.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
```

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:



Config files are copied successfully.

Formatting HDFS file system via the namenode

- 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

```
534Btdend2HM Info
Activities Terminal ▾ Sep 18 22:38 ●
csc@tdend2-hm: ~
STARTUP_MSG: build = Unknown -r 965fd380006fa78b2315668fbc7eb432e1d8200f; compiled by 'ubuntu'
STARTUP_MSG: java = 1.8.0_422
*****/
24/09/18 22:36:28 INFO namenode.NameNode: registered UNIX signal handlers for [TERM, HUP, INT]
24/09/18 22:36:28 INFO namenode.NameNode: createNameNode [-format]
Formatting using clusterid: CID-f91746eb-3e0d-4105-9001-23ef2e053119
24/09/18 22:36:29 INFO namenode.FSEditLog: Edit logging is async:true
24/09/18 22:36:29 INFO namenode.FSNamesystem: KeyProvider: null
24/09/18 22:36:29 INFO namenode.FSNamesystem: fsLock is fair: true
24/09/18 22:36:29 INFO namenode.FSNamesystem: Detailed lock hold time metrics enabled: false
24/09/18 22:36:29 INFO namenode.FSNamesystem: fsOwner = csc (auth:SIMPLE)
24/09/18 22:36:29 INFO namenode.FSNamesystem: supergroup = supergroup
24/09/18 22:36:29 INFO namenode.FSNamesystem: isPermissionEnabled = true
24/09/18 22:36:29 INFO namenode.FSNamesystem: HA Enabled: false
24/09/18 22:36:29 INFO common.Util: dfs.datanode.fileio.profiling.sampling.percentage set to 0.0
24/09/18 22:36:29 INFO blockmanagement.DatanodeManager: dfs.block.invalidate.limit: configured=10
24/09/18 22:36:29 INFO blockmanagement.DatanodeManager: dfs.namenode.datanode.registration.ip-hostname-check.enabled: configured=false
24/09/18 22:36:29 INFO blockmanagement.BlockManager: dfs.namenode.startup.delay.block.deletion.se
24/09/18 22:36:29 INFO blockmanagement.BlockManager: The block deletion will start around 2024 Se
24/09/18 22:36:29 INFO util.GSet: Computing capacity for map BlocksMap
24/09/18 22:36:29 INFO util.GSet: VM type = 64-bit
24/09/18 22:36:29 INFO util.GSet: 2.0% max memory 889 MB = 17.8 MB
24/09/18 22:36:29 INFO util.GSet: capacity = 2^21 = 2097152 entries
24/09/18 22:36:29 INFO blockmanagement.BlockManager: dfs.block.access.token.enable=false
24/09/18 22:36:29 WARN conf.Configuration: No unit for dfs.heartbeat.interval(3) assuming SECONDS
24/09/18 22:36:29 WARN conf.Configuration: No unit for dfs.namenode.safemode.extension(30000) ass
24/09/18 22:36:29 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.threshold-pct
```

For the cluster to start successfully, `java_home` , `hadoop_home`, `hadoop_conf_dir` should be set with correct paths on `hadoop_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`.

```
534Btdend2HM

Activities  Terminal  Sep 21 18:04  csc@tdend2-hm: ~

0.0.0.0: stopping secondarynamenode
csc@tdend2-hm:~$ start-dfs.sh
Starting namenodes on [tdend2-hm]
tdend2-hm: starting namenode, logging to /home/csc/tdend2/hadoop-2.10.2/logs/hadoop-namenode-csc@tdend2-hm.out
tdend2-w2: starting datanode, logging to /home/csc/tdend2/hadoop-2.10.2/logs/hadoop-datanode-csc@tdend2-w2.out
tdend2-w1: starting datanode, logging to /home/csc/tdend2/hadoop-2.10.2/logs/hadoop-datanode-csc@tdend2-w1.out
tdend2-hm: starting datanode, logging to /home/csc/tdend2/hadoop-2.10.2/logs/hadoop-datanode-csc@tdend2-hm.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/csc/tdend2/hadoop-2.10.2/logs/hadoop-secondarynamenode-csc@tdend2-hm.out
csc@tdend2-hm:~$ jps
420629 SecondaryNameNode
420387 DataNode
420174 NameNode
420745 Jps
csc@tdend2-hm:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/csc/tdend2/hadoop-2.10.2/logs/yarn-csc@tdend2-hm.out
tdend2-w2: starting nodemanager, logging to /home/csc/tdend2/hadoop-2.10.2/logs/yarn-nodemanager-csc@tdend2-w2.out
tdend2-w1: starting nodemanager, logging to /home/csc/tdend2/hadoop-2.10.2/logs/yarn-nodemanager-csc@tdend2-w1.out
tdend2-hm: starting nodemanager, logging to /home/csc/tdend2/hadoop-2.10.2/logs/yarn-nodemanager-csc@tdend2-hm.out
```

- 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:

```
csc@tdend2-hm:~$ jps
420629 SecondaryNameNode
420387 DataNode
420848 ResourceManager
421361 Jps
420174 NameNode
421037 NodeManager
csc@tdend2-hm:~$
```

On workernodes ran jps:

```
534Btdend2S1

Activities  Terminal  Sep 21 20:23  csc@tdend2-w1: ~

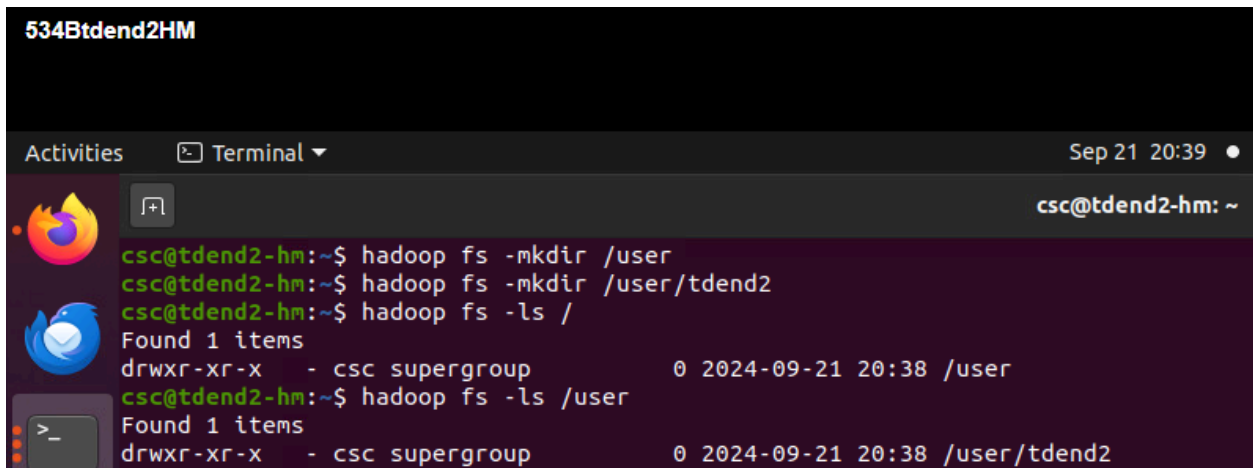
csc@tdend2-w1:~$ jps
376677 Jps
368212 NodeManager
376091 DataNode
```


On worker node2 :



```
534Btdend2S2
Activities Terminal Sep 21 20:24
csc@tdend2-w2: ~
csc@tdend2-w2:~$ jps
449543 Jps
444548 NodeManager
448991 DataNode
```

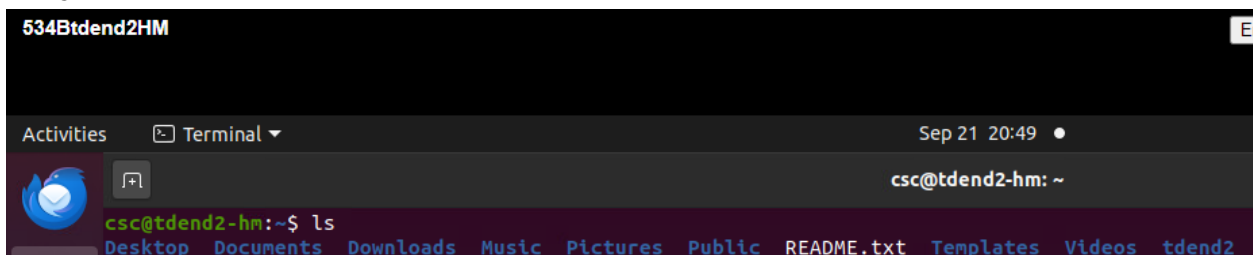
b. By creating your user directory



```
534Btdend2HM
Activities Terminal Sep 21 20:39
csc@tdend2-hm: ~
csc@tdend2-hm:~$ hadoop fs -mkdir /user
csc@tdend2-hm:~$ hadoop fs -mkdir /user/tdend2
csc@tdend2-hm:~$ hadoop fs -ls /
Found 1 items
drwxr-xr-x - csc supergroup 0 2024-09-21 20:38 /user
csc@tdend2-hm:~$ hadoop fs -ls /user
Found 1 items
drwxr-xr-x - csc supergroup 0 2024-09-21 20:38 /user/tdend2
```

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.



```
534Btdend2HM
Activities Terminal Sep 21 20:49
csc@tdend2-hm: ~
csc@tdend2-hm:~$ ls
Desktop Documents Downloads Music Pictures Public README.txt Templates Videos tdend2
```

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

```
534Btdend2HM

Activities  Terminal  Sep 21 20:51  ●
csc@tdend2-hm: ~
csc@tdend2-hm:~$ hadoop fs -put /home/csc/README.txt /user/tdend2
```

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`

```
534Btdend2HM

Activities  Terminal  Sep 21 20:55  ●
csc@tdend2-hm: ~
csc@tdend2-hm:~$ hdfs fsck /user/tdend2/README.txt -files -locations -blocks
Connecting to namenode via http://tdend2-hm:50070/fsck?ugi=csc&files=1&locations=
FSCK started by csc (auth:SIMPLE) from /10.92.128.52 for path /user/tdend2/README
/user/tdend2/README.txt 7 bytes, 1 block(s): OK
0. BP-1362211458-10.92.128.52-1726940202742:blk_1073741825_1001 len=7 Live_repl=1
-4785-af8e-fb861b20f8d8,DISK]]

Status: HEALTHY
Total size:      7 B
Total dirs:      0
Total files:     1
Total symlinks:   0
Total blocks (validated): 1 (avg. block size 7 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 3
Number of racks: 1
FSCK ended at Sat Sep 21 20:53:13 UTC 2024 in 5 milliseconds
```



```
534Btdend2HM

Activities  Terminal  Sep 21 20:53  ●

csc@tdend2-hm: ~

0. BP-1362211458-10.92.128.52-1726940202742:blk_1073741825_1001 len=7 Live_repl-
-4785-af8e-fb861b20f8d8,DISK]]

Status: HEALTHY
Total size: 7 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 1 (avg. block size 7 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 3
Number of racks: 1
FSCK ended at Sat Sep 21 20:53:13 UTC 2024 in 5 milliseconds

The filesystem under path '/user/tdend2/README.txt' is HEALTHY
csc@tdend2-hm:~$
```

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.

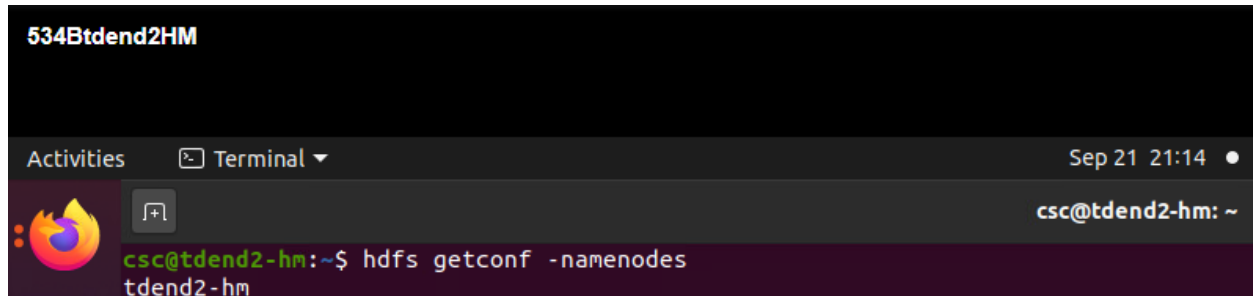
Files, blocks, locations option display detailed report of each file, blocks within HDFS and the location of each block respectively.

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.

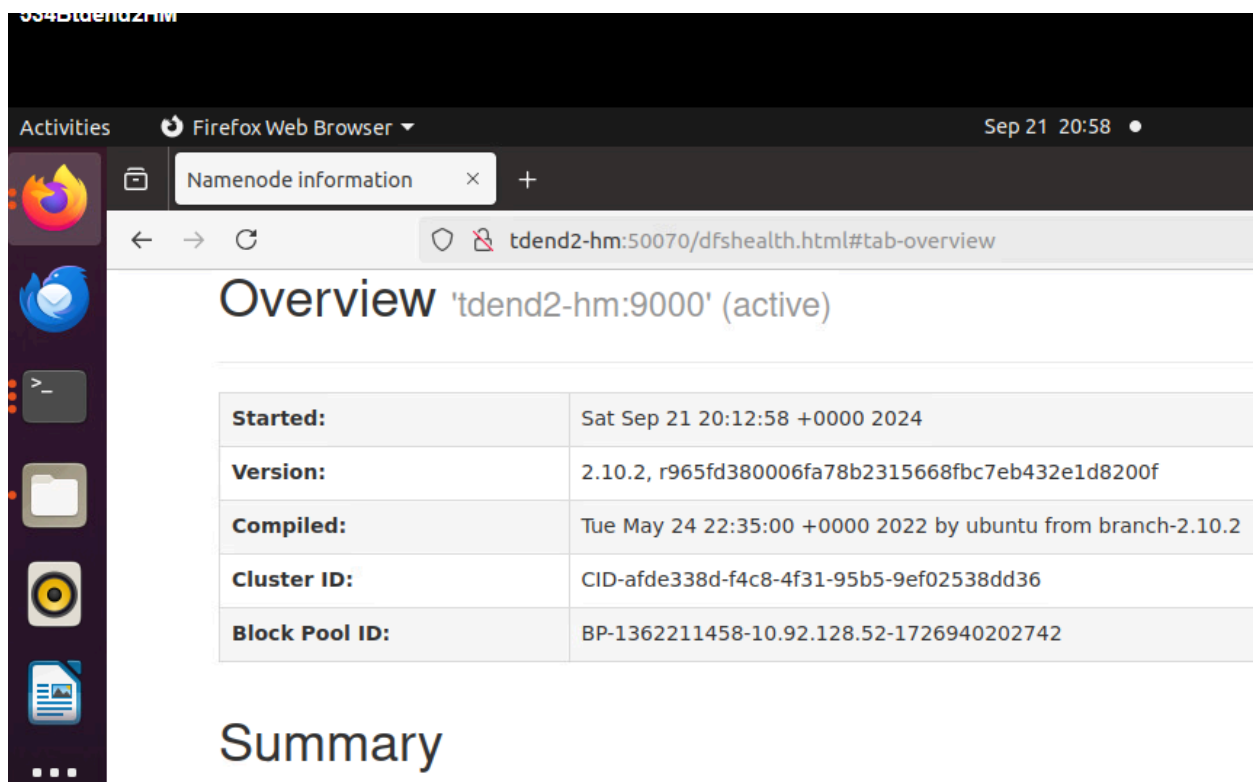


```
534Btdend2HM
Activities Terminal Sep 21 21:14
csc@tdend2-hm: ~
csc@tdend2-hm:~$ hdfs getconf -namenodes
tdend2-hm
```

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](http://<nodename(s)>:50070) as <http://tdend2-hm:50070>

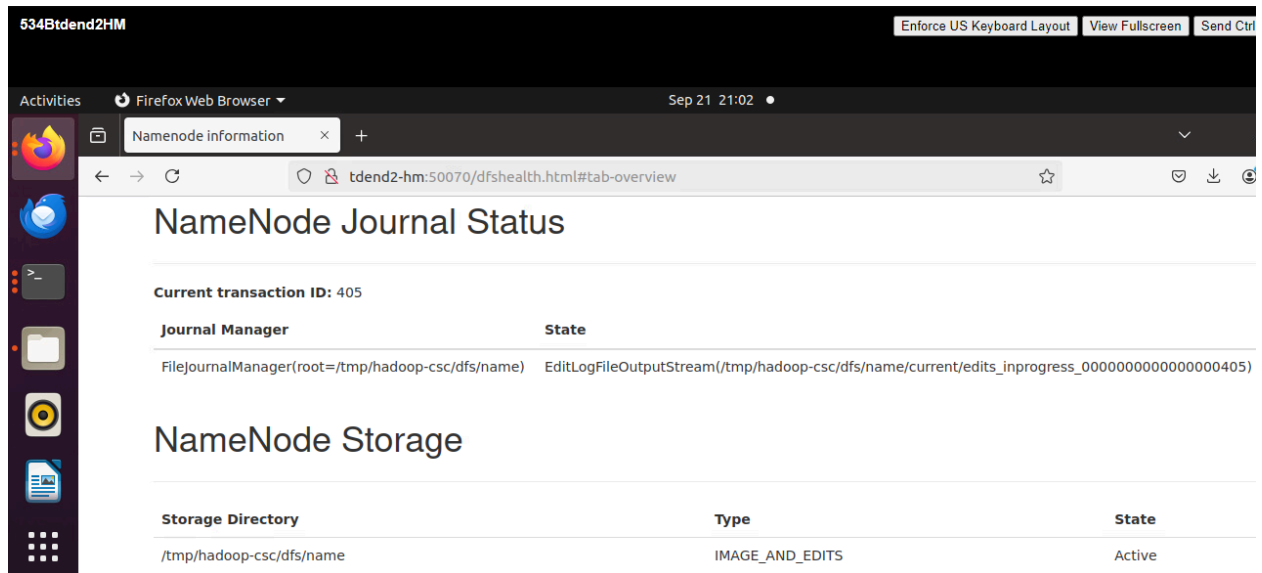
The UI at the above url displayed the overview and summary of namenode information of cluster.



The screenshot shows the Hadoop Namenode web UI in a Firefox browser window. The address bar shows the URL `tdend2-hm:50070/dfshealth.html#tab-overview`. The page title is "Overview 'tdend2-hm:9000' (active)". Below the title is a table with the following information:

Started:	Sat Sep 21 20:12:58 +0000 2024
Version:	2.10.2, r965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled:	Tue May 24 22:35:00 +0000 2022 by ubuntu from branch-2.10.2
Cluster ID:	CID-afde338d-f4c8-4f31-95b5-9ef02538dd36
Block Pool ID:	BP-1362211458-10.92.128.52-1726940202742

Below the table is a section titled "Summary".



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.

```

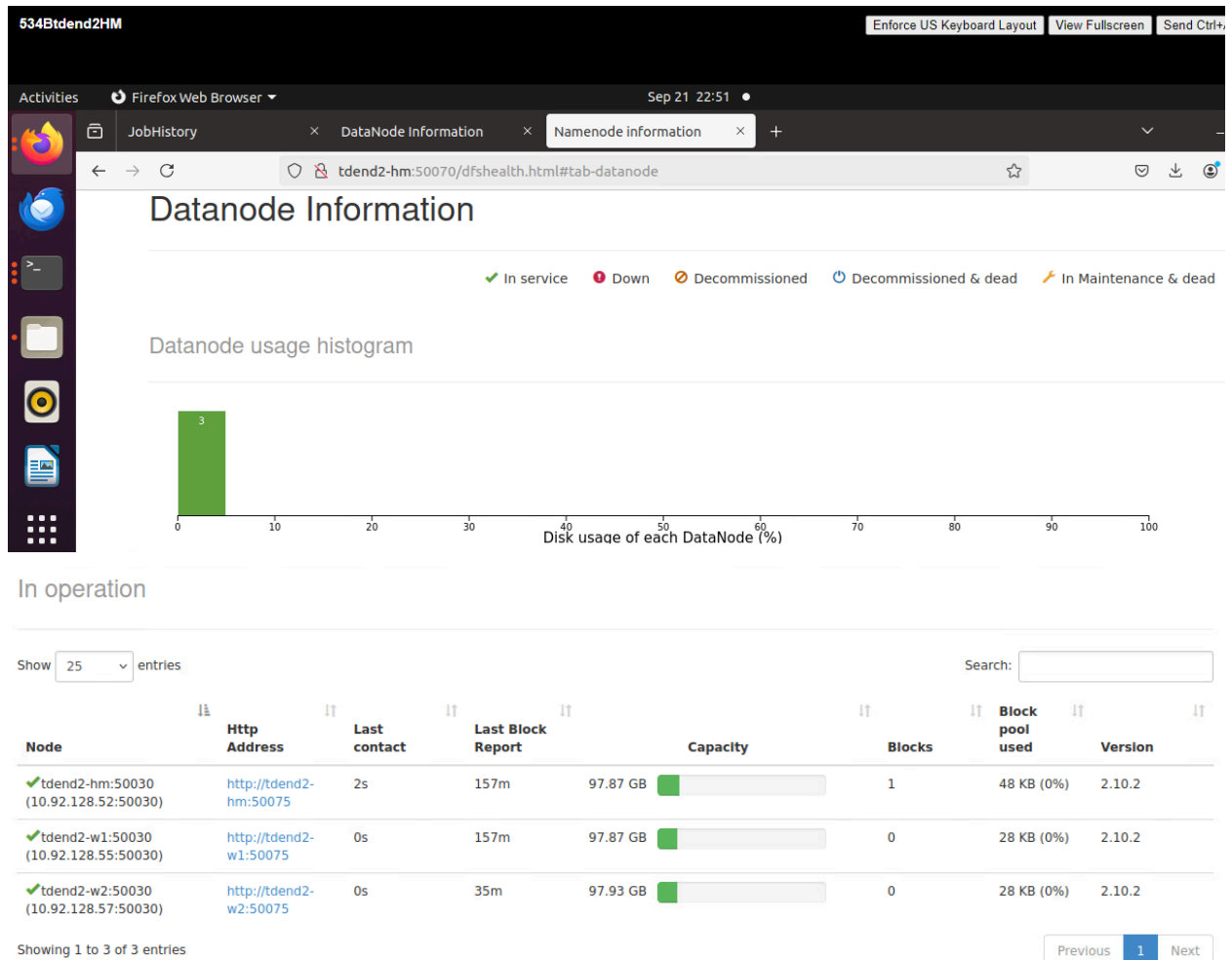
534Btdend2HM
Activities Terminal Sep 21 21:15
csc@tdend2-hm: ~
csc@tdend2-hm:~$ hdfs dfsadmin -report
Configured Capacity: 315321262080 (293.67 GB)
Present Capacity: 259302600704 (241.49 GB)
DFS Remaining: 259302494208 (241.49 GB)
DFS Used: 106496 (104 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0
Pending deletion blocks: 0

-----
Live datanodes (3):

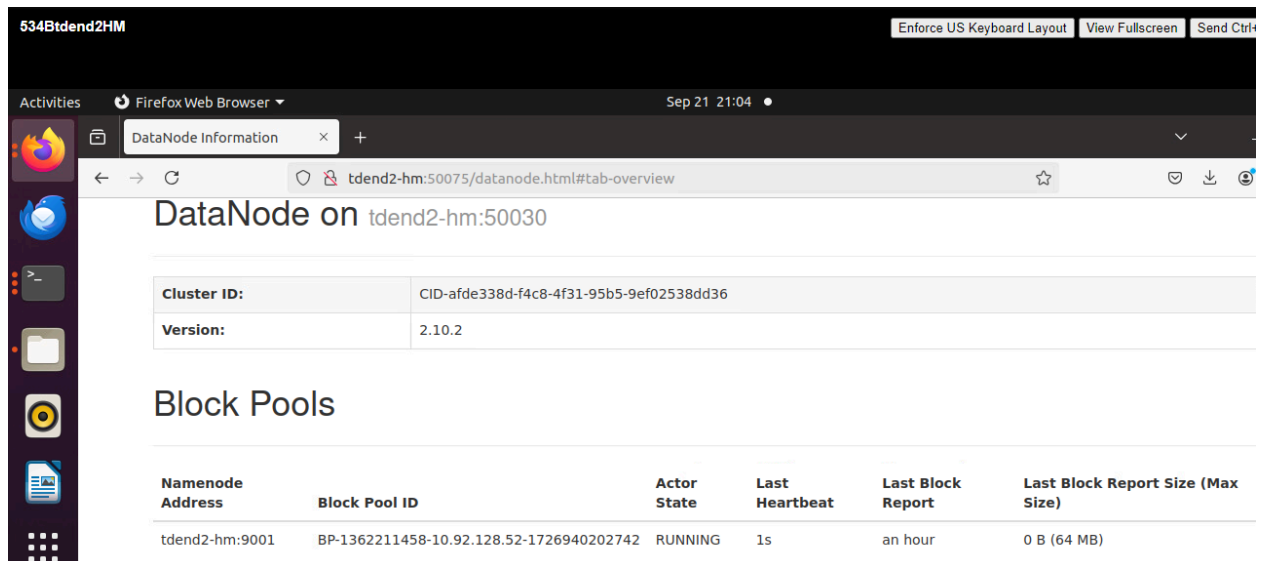
Name: 10.92.128.52:50030 (tdend2-hm)
Hostname: tdend2-hm
Decommission Status : Normal
Configured Capacity: 105086115840 (97.87 GB)
DFS Used: 49152 (48 KB)
Non DFS Used: 13893943296 (12.94 GB)
DFS Remaining: 85806796800 (79.91 GB)
DFS Used%: 0.00%
DFS Remaining%: 81.65%

```

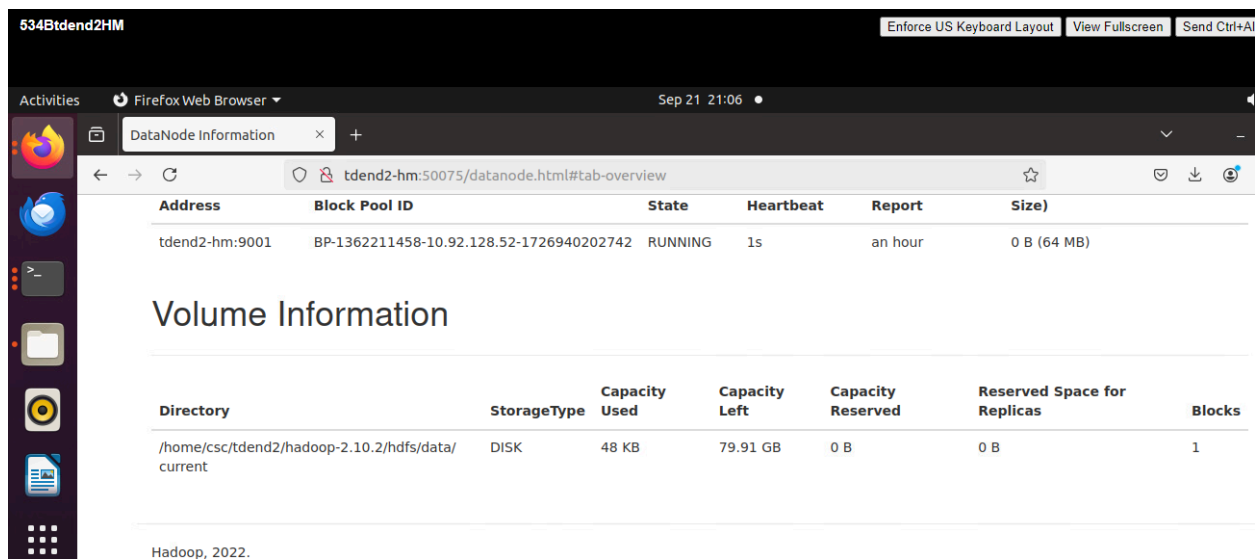
The cluster has 3 active data nodes - one on master, one on each worker node.



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



Last half of overview of data node information:



Address	Block Pool ID	State	Heartbeat	Report	Size
tdend2-hm:9001	BP-1362211458-10.92.128.52-1726940202742	RUNNING	1s	an hour	0 B (64 MB)

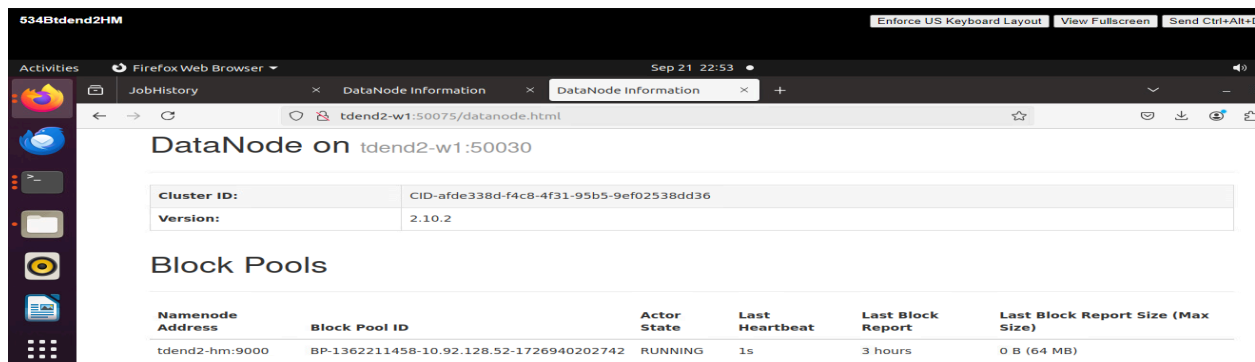
Volume Information

Directory	StorageType	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/home/csc/tdend2/hadoop-2.10.2/hdfs/data/current	DISK	48 KB	79.91 GB	0 B	0 B	1

Hadoop, 2022.

On master node node, 1 data node is available.

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



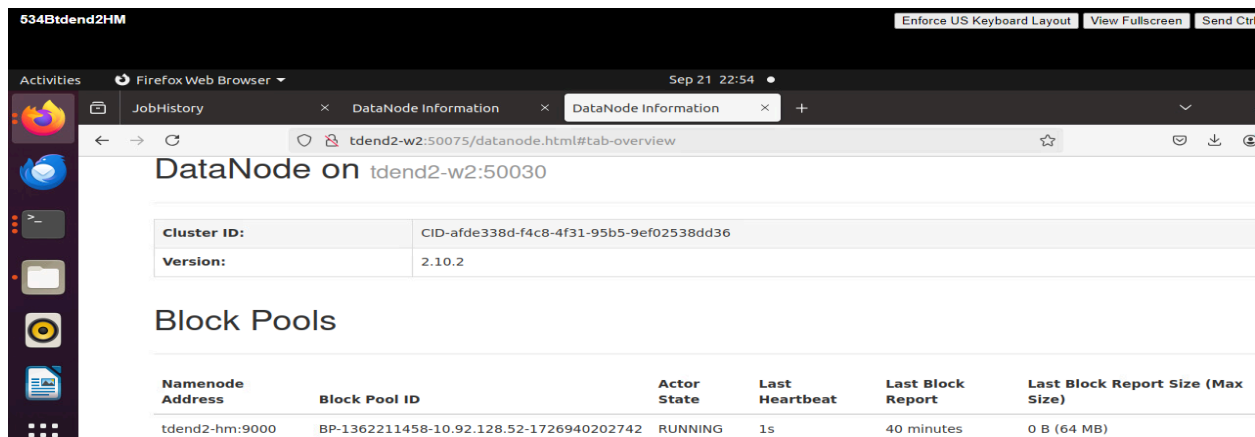
DataNode on tdend2-w1:50030

Cluster ID:	CID-afde338d-f4c8-4f31-95b5-9ef02538dd36
Version:	2.10.2

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
tdend2-hm:9000	BP-1362211458-10.92.128.52-1726940202742	RUNNING	1s	3 hours	0 B (64 MB)

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



DataNode on tdend2-w2:50030

Cluster ID:	CID-afde338d-f4c8-4f31-95b5-9ef02538dd36
Version:	2.10.2

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
tdend2-hm:9000	BP-1362211458-10.92.128.52-1726940202742	RUNNING	1s	40 minutes	0 B (64 MB)

c. Resource Manager UI

i. Checked the master node's name

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

The screenshot displays the Hadoop Resource Manager (RM) web interface in a Firefox browser. The address bar shows `tdend2-hm:8088/cluster`. The page title is "All Applications". On the left, there is a sidebar with a "Cluster" menu containing links for "About", "Nodes", "Node Labels", "Applications", and "Scheduler". The "Nodes" link is selected. The main content area shows "Cluster Metrics" with a table of application statistics. Below this, "Cluster Nodes Metrics" shows a table with 3 active nodes. The "Scheduler Metrics" section shows the scheduler type as "Capacity Scheduler" and the scheduling resource type as "Memory". At the bottom, there is a table of application entries with columns for ID, User, Name, Application Type, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU, and Allocated Memory.

Cluster Metrics	
Apps Submitted	Apps Pending
0	0

Cluster Nodes Metrics	
Active Nodes	Decommissioning Nodes
3	0

Scheduler Metrics	
Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU	Allocated Memory

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.

ApplicationMasters

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

The RM's responsibilities include:

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:

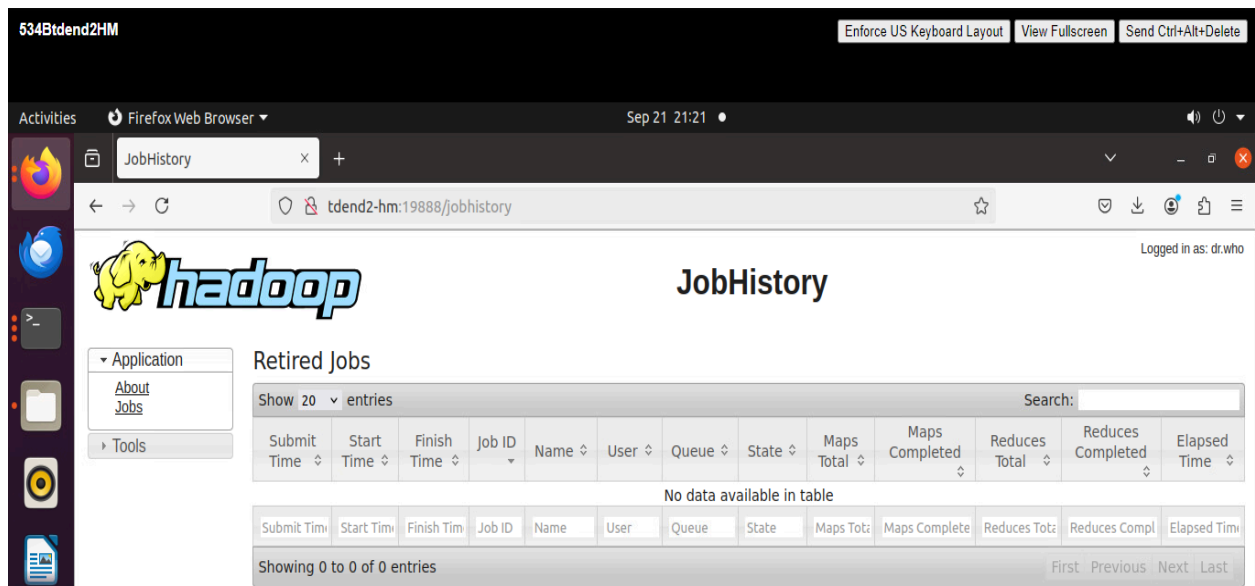
mr-jobhistory-daemon.sh start historyserver

i. Checked the master node's name

```
534Btdend2HM Enforce US Keyboard Layo
Activities Terminal Sep 21 21:22
csc@tdend2-hm: ~
csc@tdend2-hm:~$ mr-jobhistory-daemon.sh start historyserver
starting historyserver, logging to /home/csc/tdend2/hadoop-2.10.2/logs/mapred-csc-historyserver-tdend2-hm.out
csc@tdend2-hm:~$
```

```
csc@tdend2-hm:~$ jps
437030 JobHistoryServer
449108 Jps
420848 ResourceManager
430639 DataNode
421037 NodeManager
430906 SecondaryNameNode
430424 NameNode
```

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-----