Exp.1 **210701144**

Downloading and installing Hadoop, Understanding different Hadoop modes,Startup scripts, Configuration files.

AIM:

To Download and install Hadoop, Understanding different Hadoop modes, Startup scripts, Configuration files.

Procedure:

Step 1 : Install Java Development Kit
The default Ubuntu repositories contain Java 8 and Java 11 both. But, Install Java 8 because hive only works on this version. Use the following command to install it.

$sudo apt update&&sudo apt install openjdk-8-jdk

Step 2 : Verify the Java version
Once installed, verify the installed version of Java with the following

command: **$ java -version Output:**

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ java -version
openjdk version "1.8.0_422"
OpenJDK Runtime Environment (build 1.8.0_422-8u422-b05-1~24.04-b05)
OpenJDK 64-Bit Server VM (build 25.422-b05, mixed mode)
```

Step 3: Install SSH

SSH (Secure Shell) installation is vital for Hadoop as it enables secure communication between nodes in the Hadoop cluster. This ensures data integrity, confidentiality, and allows for efficient distributed processing of data across the cluster. **$sudo apt install ssh**

Step 4 : Create the hadoop user :
All the Hadoop components will run as the user that you create for Apache Hadoop, and the user will also be used for logging in to Hadoop's web interface. Run the command to create user and set password:

$ sudo adduser hadoop

Output

```
sanjay@sanjay-VirtualBox:~$ sudo adduser hadoop
Adding user `hadoop' ...
Adding new group `hadoop' (1001) ...
Adding new user `hadoop' (1001) with group `hadoop (1001)' ...
adduser: The home directory '/home/hadoop' already exists.  Not copying from `/etc/skel'.
New password:
BAD PASSWORD: The password is shorter than 8 characters
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] y
Adding new user `hadoop' to supplemental / extra groups `users' ...
Adding user `hadoop' to group `users' ...
sanjay@sanjay-VirtualBox:~$ su - hadoop
Password:
hadoop@sanjay-VirtualBox:~$
```

Step 5 : Switch user
Switch to the newly created hadoop user:
$ su - hadoop

Step 6 : Configure SSH
Now configure password-less SSH access for the newly created hadoop user, so didn't enter the key to save file and passphrase. Generate an SSH keypair (generate Public and Private Key Pairs)first

$ssh-keygen -t rsa

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
/home/hadoop/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:uagHTeOndLi7HaKESN2yW0yRTjYCVe/sHubBEeTVde8 hadoop@makesh-HP-Laptop-15s-eq3xxx
The key's randomart image is:
+---[RSA 3072]----+
| ..... . .. .. . |
|  .   = .  .  . .|
|   . * +        .|
|  . * B ..     . |
| . o B *S       E|
|.. .* B.+.       |
|. ...=o@o        |
|   .ooBoo.       |
|   .o.o=.        |
+----[SHA256]-----+
```

Step 7 : Set permissions :

Next, append the generated public keys from id_rsa.pub to authorized_keys and set proper permission:

$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

$ chmod 640 ~/.ssh/authorized_keys

Step 8 : SSH to the localhost
Next, verify the password less SSH authentication with the following command:

$ ssh localhost

You will be asked to authenticate hosts by adding RSA keys to known hosts. Type yes and hit Enter to authenticate the localhost:

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ chmod 640 ~/.ssh/authorized_keys
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:MkKdpFyMhsIW8XAhD4zc6JL2YvSZp0q/f5jUuwFDnuo.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 24.04.1 LTS (GNU/Linux 6.8.0-41-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

1 update can be applied immediately.
To see these additional updates run: apt list --upgradable

8 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm


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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.
```

Step 9 : Switch user

Again switch to hadoop. So, First, change the user to hadoop with the following command: **$ su–hadoop**

Step 10 : Install hadoop

Next, download the latest version of Hadoop using the wget command:

$ wgethttps://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz

Once downloaded, extract the downloaded file:

$ tar -xvzf hadoop-3.3.6.tar.gz

Next, rename the extracted directory to hadoop:

$ mv hadoop-3.3.6 hadoop

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ mv hadoop-3.3.6 hadoop
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ ls
Desktop    Downloads  hadoop-3.3.6.tar.gz  Pictures  snap       Videos
Documents  hadoop     Music                Public    Templates
```

Next, you will need to configure Hadoop and Java Environment Variables on your system. Open the ~/.bashrc file in your favorite text editor. Use nano editior , to pasting the code we use ctrl+shift+v for saving the file ctrl+x and ctrl+y ,then hit enter:

Next, you will need to configure Hadoop and Java Environment Variables on your system.

Open the ~/.bashrc file in your favorite text editor:

$ nano ~/.bashrc

Append the below lines to file.

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

Save and close the file. Then, activate the environment variables with the following command:

s$ source ~/.bashrc

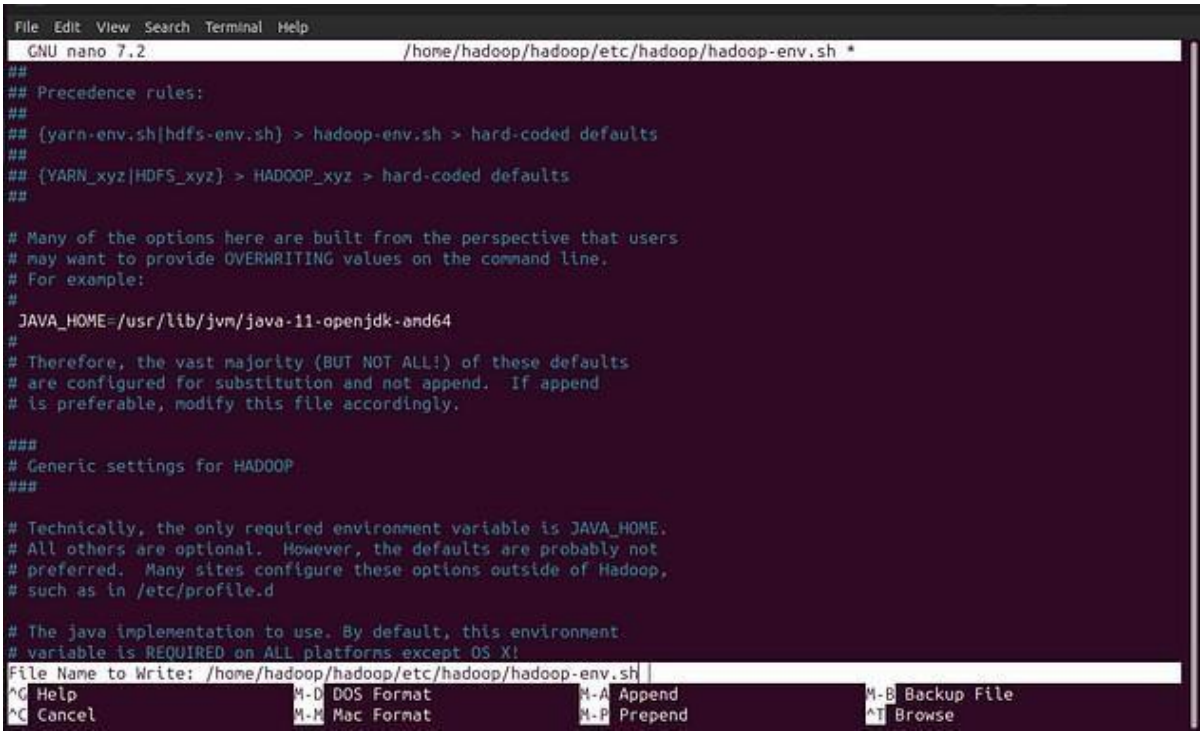
Next, open the Hadoop environment variable file:

$ nano $HADOOP_HOME/etc/hadoop/hadoop-

env.sh

Search for the “export JAVA_HOME” and configure it.

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



Save and close the file when you are finished.

Step 11 : Configuring Hadoop :

First, you will need to create the namenode and datanode directories inside the Hadoop user home directory. Run the following command to create both directories:

$ cd hadoop/
$mkdir -p ~/hadoopdata/hdfs/{namenode,datanode}

- Next, edit the core-site.xml file and

update with your system hostname:

$nano $HADOOP_HOME/etc/hadoop/core-site.xml

Change the following name as per your system hostname:

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

Save and close the file.

Then, edit the hdfs-site.xml file:

$nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

- Change the NameNode and DataNode directory paths as shown below:

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>

    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
    </property>

    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
    </property>
</configuration>
```

- Then, edit the mapred-site.xml file:
  $nano $HADOOP_HOME/etc/hadoop/mapred-site.xml

- Make the following changes:

```
<configuration>
   <property>
      <name>yarn.app.mapreduce.am.env</name>
      <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
   </property>
   <property>
      <name>mapreduce.map.env</name>
      <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
   </property>
   <property>
      <name>mapreduce.reduce.env</name>
      <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
   </property>
</configuration>
```

- Then, edit the yarn-site.xml file:
  $nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
- Make the following changes:

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

  Save the file and close it .

Step 12 – Start Hadoop Cluster

Before starting the Hadoop cluster. You will need to format the Namenode as a hadoop user.

Run the following command to format the Hadoop Namenode:

$hdfs namenode –format

Once the namenode directory is successfully formatted with hdfs file system, you will see the message “Storage directory /home/hadoop/hadoopdata/hdfs/namenode has been successfully formatted “

Then start the Hadoop cluster with the following command.

$ start-all.sh

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~/hadoop$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
Starting datanodes
Starting secondary namenodes [makesh-HP-Laptop-15s-eq3xxx]
makesh-HP-Laptop-15s-eq3xxx: Warning: Permanently added 'makesh-hp-laptop-15s-eq3xxx' (ED25
519) to the list of known hosts.
Starting resourcemanager
Starting nodemanagers
```

You can now check the status of all Hadoop services using the jps command:

$ jps

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ jps
14784 Jps
13921 ResourceManager
13442 DataNode
14066 NodeManager
13698 SecondaryNameNode
13295 NameNode
```

Step 13 – Access Hadoop Namenode and Resource Manager

- First we need to know our ipaddress, In Ubuntu we need to install net-tools to run ipconfig command,
  If you installing net-tools for the first time switch to default user:
  $sudo apt install net-tools
- Then run ifconfig command to know our ip address: **ifconfig**

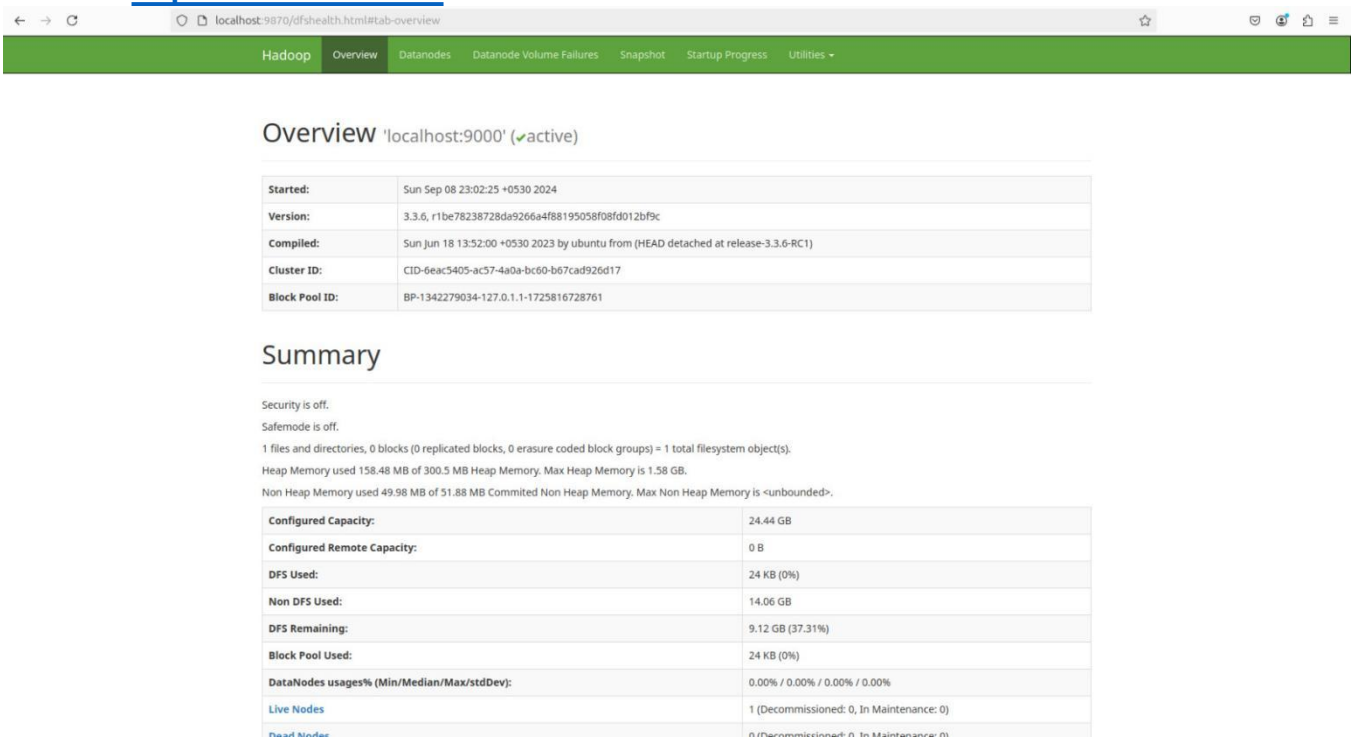
```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ ifconfig
lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 68758  bytes 9196628 (9.1 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 68758  bytes 9196628 (9.1 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

wlo1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.1.10  netmask 255.255.255.0  broadcast 192.168.1.255
        inet6 fe80::c962:7d5:707d:740d  prefixlen 64  scopeid 0x20<link>
        inet6 2401:4900:1cd0:fe15:ad36:a87c:831d:67ec  prefixlen 64  scopeid 0x0<global>
        inet6 2401:4900:1cd0:fe15:e5c7:8cd1:880c:9676  prefixlen 64  scopeid 0x0<global>
        ether 2c:3b:70:9f:0a:d1  txqueuelen 1000  (Ethernet)
        RX packets 581629  bytes 845869566 (845.8 MB)
        RX errors 0  dropped 2596  overruns 0  frame 0
        TX packets 68750  bytes 10990916 (10.9 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

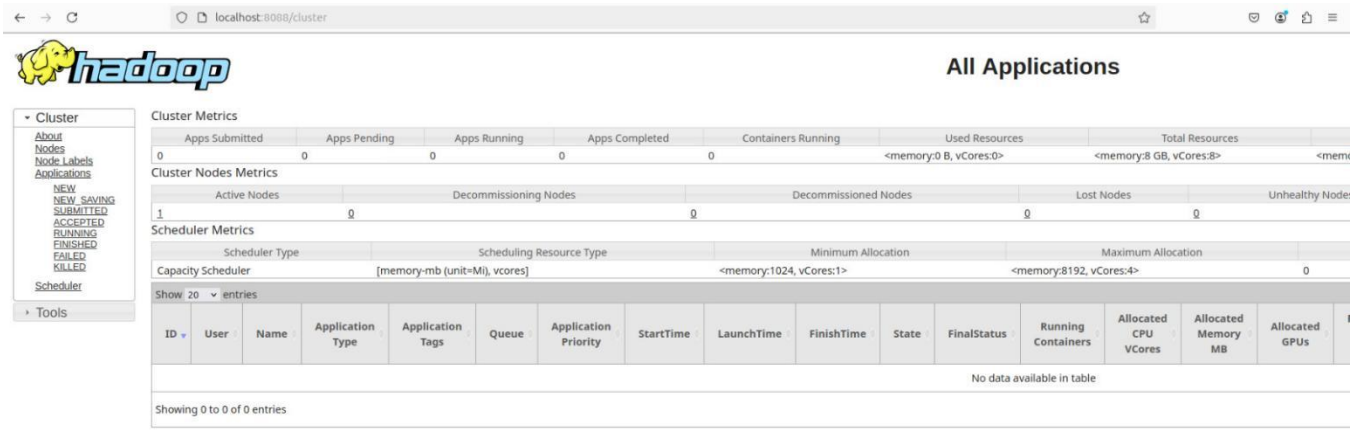
Here my ip address is 192.168.1.6.

- To access the Namenode, open your web browser and visit the URL http://your-serverip:9870.
- You should see the following screen:

http://192.168.1.6:9870



To access Resource Manage, open your web browser and visit the URL http://your-serverip:8088. You should see the following screen: http://192.168.16:8088



Step 14 – Verify the Hadoop Cluster

At this point, the Hadoop cluster is installed and configured. Next, we will create some directories in the HDFS filesystem to test the Hadoop.

Let's create some directories in the HDFS filesystem using the following command:

```
$ hdfsdfs -mkdir /test1
$ hdfsdfs -mkdir /logs
```

Next, run the following command to list the above directory:

```
$ hdfs dfs -ls /
```

You should get the following output:

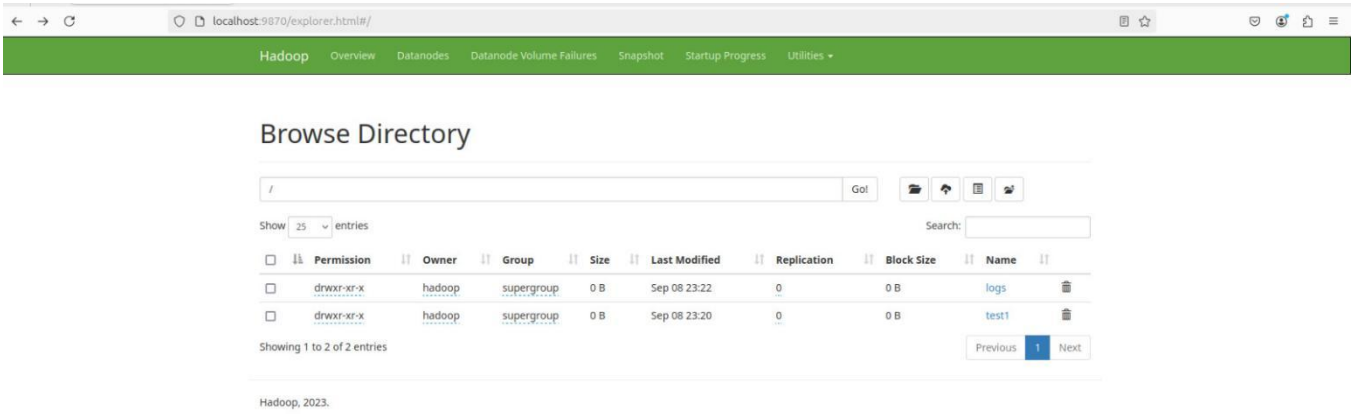
```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ hdfs dfs -mkdir /test1
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ hdfs dfs -mkdir /logs
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-08 23:20 /logs
drwxr-xr-x   - hadoop supergroup          0 2024-09-08 23:20 /test1
```

Also, put some files to hadoop file system. For the example, putting log files from host machine to hadoop file system.

```
$ hdfs dfs -put /var/log/* /logs/
```

You can also verify the above files and directory in the Hadoop Namenode web interface.

Go to the web interface, click on the Utilities => Browse the file system. You should see your directories which you have created earlier in the following screen:



Step 15 – Stop Hadoop Cluster

To stop the Hadoop all services, run the following command:

```
$ stop-all.sh
```

```
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [makesh-HP-Laptop-15s-eq3xxx]
Stopping nodemanagers
Stopping resourcemanager
hadoop@makesh-HP-Laptop-15s-eq3xxx:~$
```

Result:

The step-by-step installation and configuration of Hadoop on Ubuntu linux system have been successfully completed.