Exp.No: 1

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.

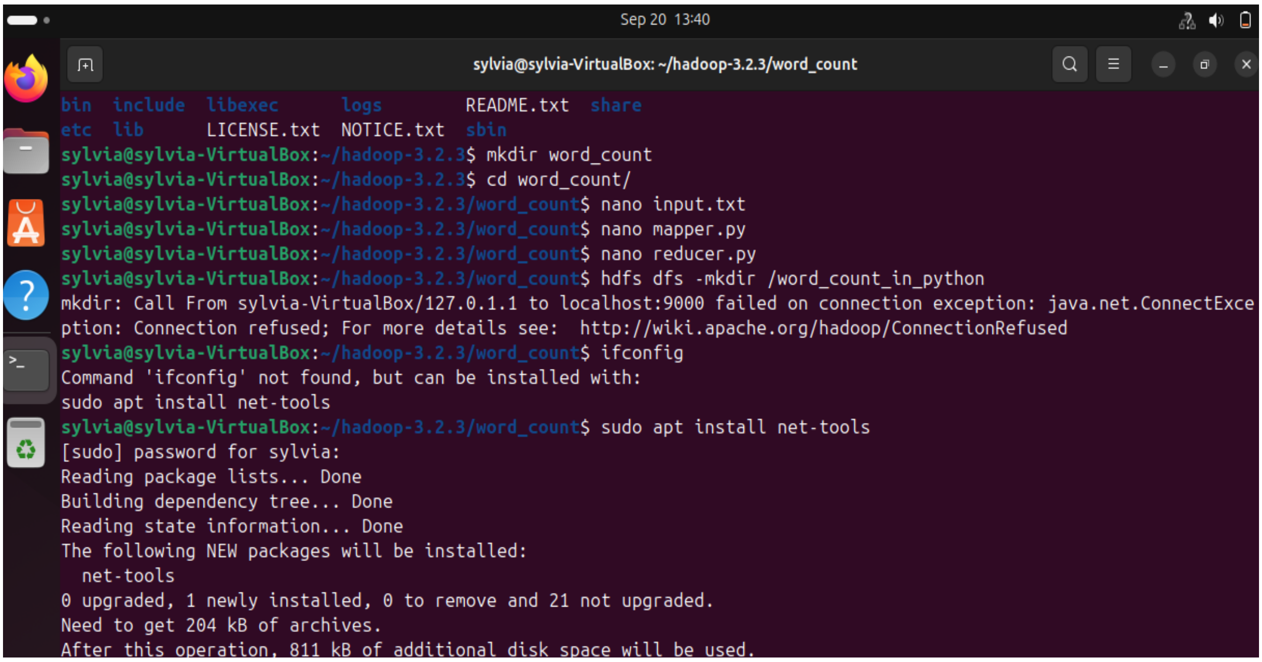
Procedure:

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:



```
bin  include  libexec      logs        README.txt  share
etc  lib      LICENSE.txt  NOTICE.txt  sbin
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3$ mkdir word_count
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3$ cd word_count/
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ nano input.txt
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ nano mapper.py
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ nano reducer.py
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ hdfs dfs -mkdir /word_count_in_python
mkdir: Call From sylvia-VirtualBox/127.0.1.1 to localhost:9000 failed on connection exception: java.net.ConnectExce
ption: Connection refused; For more details see:  http://wiki.apache.org/hadoop/ConnectionRefused
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ ifconfig
Command 'ifconfig' not found, but can be installed with:
sudo apt install net-tools
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ sudo apt install net-tools
[sudo] password for sylvia:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  net-tools
0 upgraded, 1 newly installed, 0 to remove and 21 not upgraded.
Need to get 204 kB of archives.
After this operation, 811 kB of additional disk space will be used.
```

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

Step 5 : Switch user

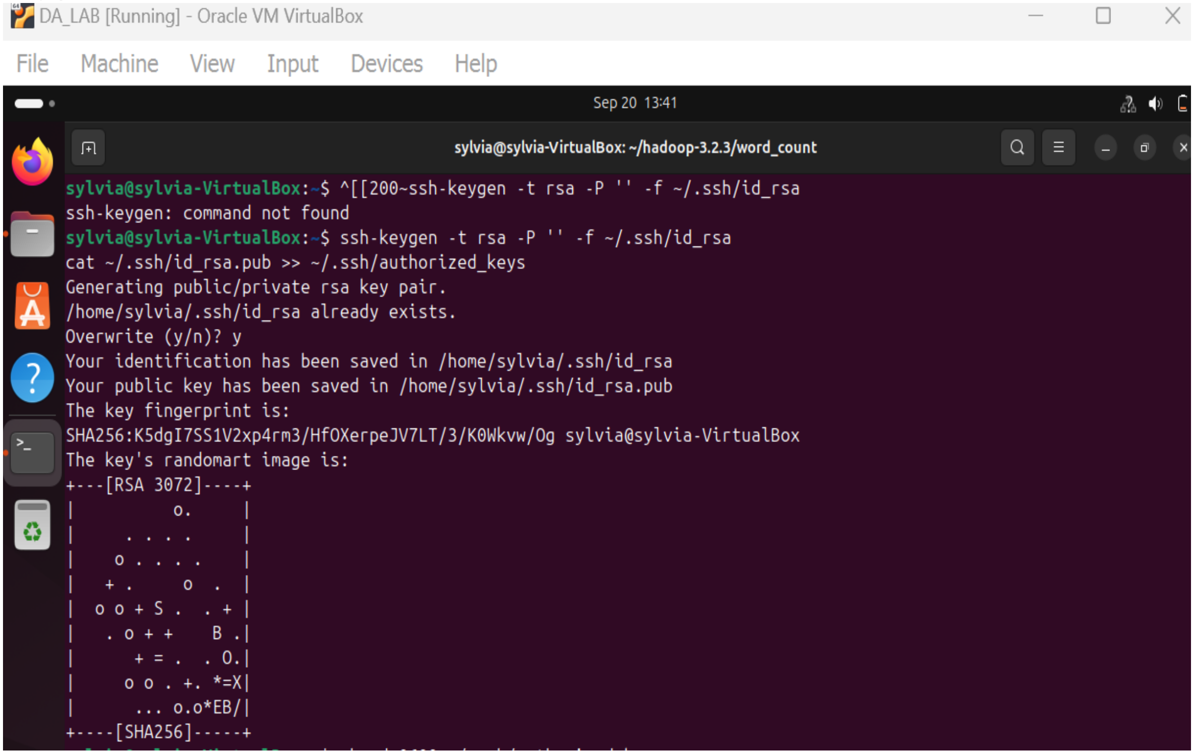
Switch to the newly created hadoop user:

$ 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


DA_LAB [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Sep 20 13:41
sylvia@sylvia-VirtualBox: ~/hadoop-3.2.3/word_count

```
sylvia@sylvia-VirtualBox:~$ ^[[200~ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
ssh-keygen: command not found
sylvia@sylvia-VirtualBox:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
Generating public/private rsa key pair.
/home/sylvia/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/sylvia/.ssh/id_rsa
Your public key has been saved in /home/sylvia/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:K5dgI7SS1V2xp4rm3/HfOXerpeJV7LT/3/K0Wkvw/Og sylvia@sylvia-VirtualBox
The key's randomart image is:
+---[RSA 3072]----+
|          o.     |
|     . . . .     |
|    o . . . .    |
|   + .     o  .  |
|  o o + S .  . + |
|   . o + +    B .|
|      + = .  . O.|
|     o o . +. *=X|
|      ... o.o*EB/|
+----[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

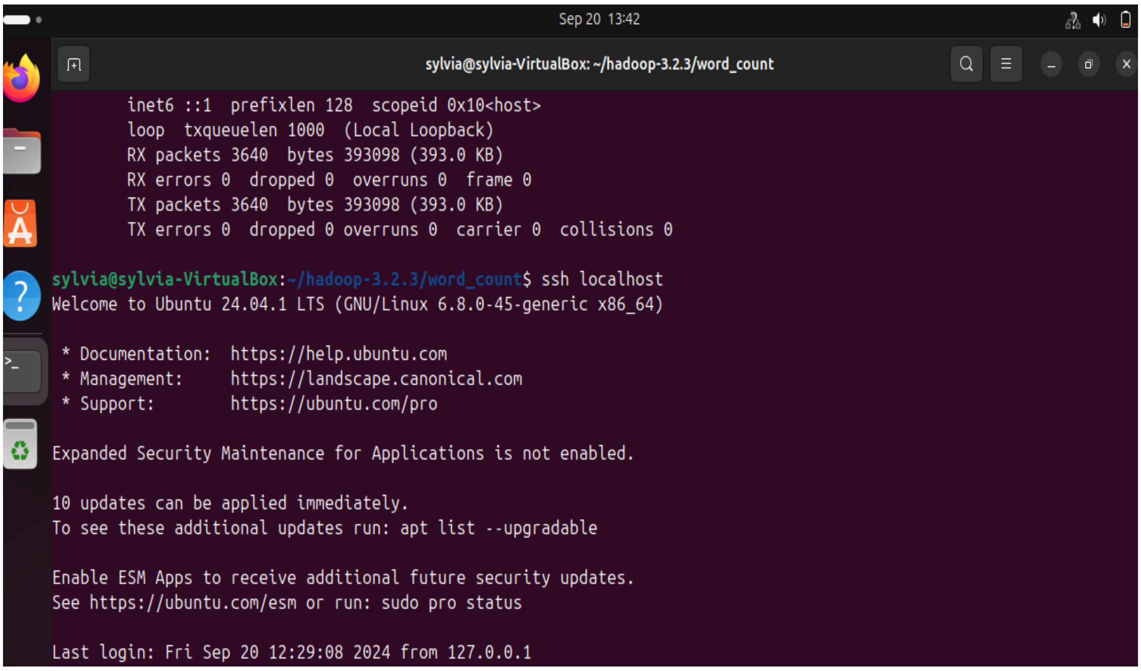
$ chmod 640 ~/.ssh/authorized_keys

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:



```
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 3640  bytes 393098 (393.0 KB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 3640  bytes 393098 (393.0 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ ssh localhost
Welcome to Ubuntu 24.04.1 LTS (GNU/Linux 6.8.0-45-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.

10 updates can be applied immediately.
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: Fri Sep 20 12:29:08 2024 from 127.0.0.1
```

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

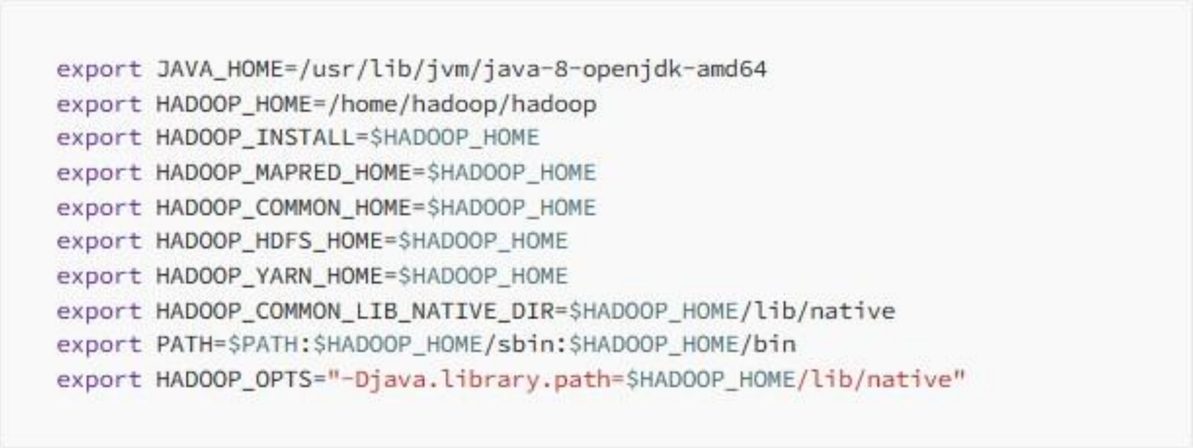
Next, you will need to configure Hadoop and Java Environment Variables on your system. Open the ~/.bashrc file in your favorite text editor. Use nano editior , to pasting the code we use ctrl+shift+v for saving the file ctrl+x and ctrl+y ,then hit enter:

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:

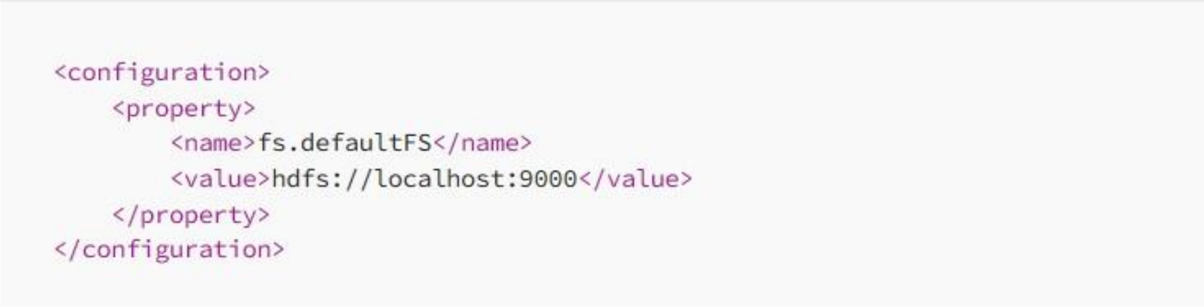
$ cd hadoop/

$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:



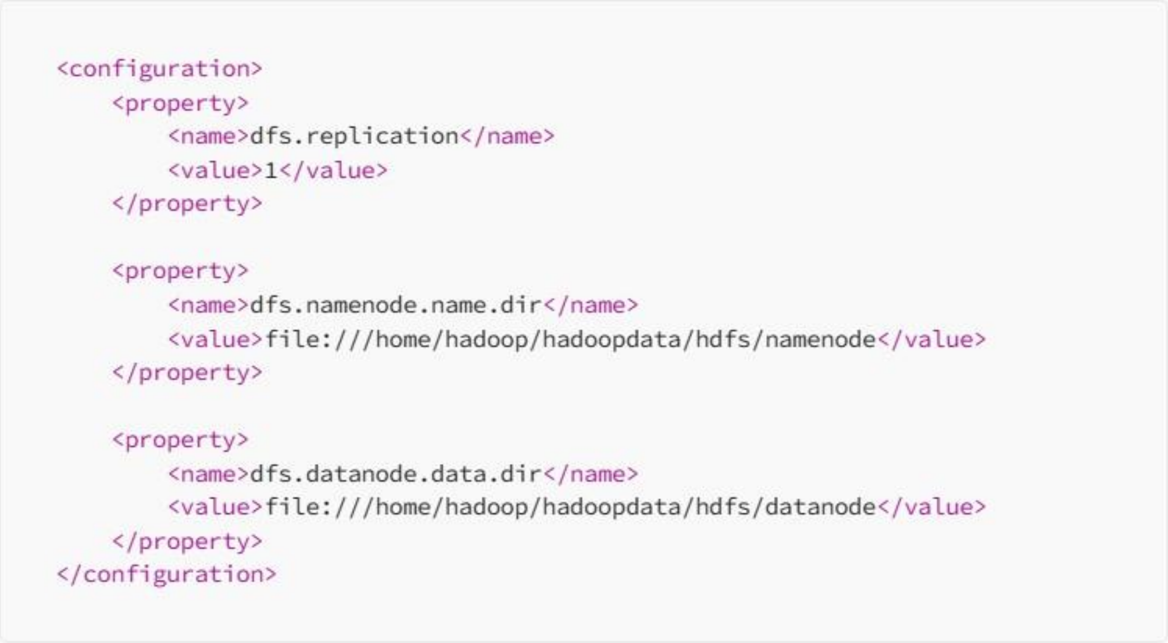
```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

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:



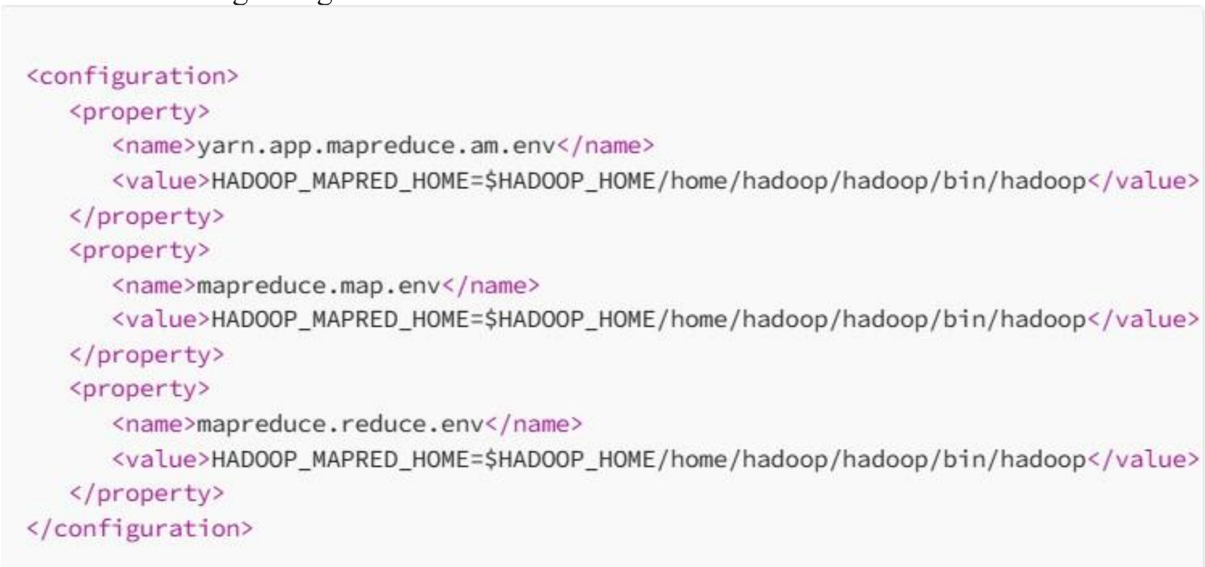
```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>

    <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>
```

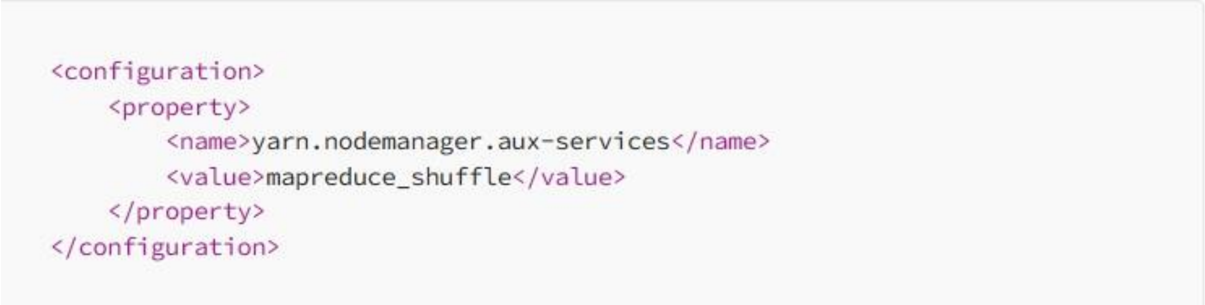
- Then, edit the mapred-site.xml file:
  $nano $HADOOP_HOME/etc/hadoop/mapred-site.xml

- 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.

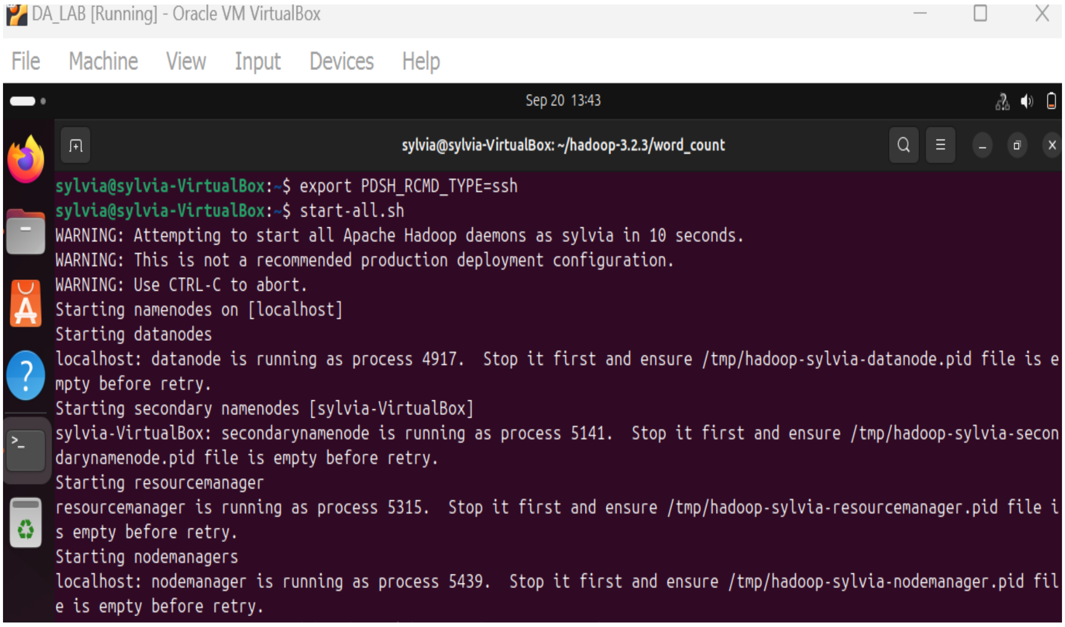
Run the following command to format the Hadoop Namenode:

$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.

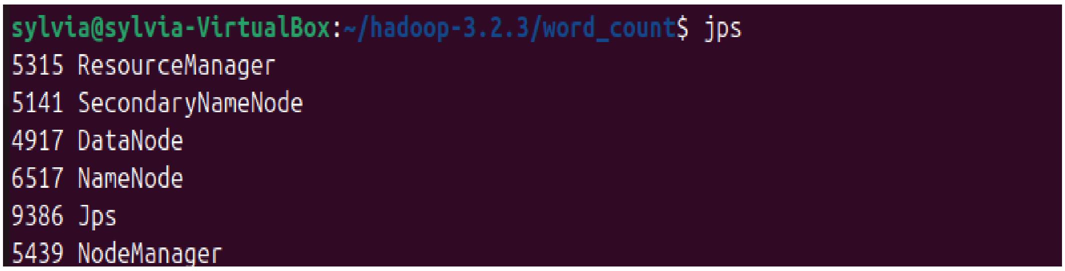
$ start-all.sh


DA_LAB [Running] - Oracle VM VirtualBox
File
Machine
View
Input
Devices
Help
Sep 20 13:43
sylvia@sylvia-VirtualBox: ~/hadoop-3.2.3/word_count

```
sylvia@sylvia-VirtualBox:~$ export PDSH_RCMD_TYPE=ssh
sylvia@sylvia-VirtualBox:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as sylvia in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
localhost: datanode is running as process 4917.  Stop it first and ensure /tmp/hadoop-sylvia-datanode.pid file is e
mpty before retry.
Starting secondary namenodes [sylvia-VirtualBox]
sylvia-VirtualBox: secondarynamenode is running as process 5141.  Stop it first and ensure /tmp/hadoop-sylvia-secon
darynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 5315.  Stop it first and ensure /tmp/hadoop-sylvia-resourcemanager.pid file i
s empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 5439.  Stop it first and ensure /tmp/hadoop-sylvia-nodemanager.pid fil
e is empty before retry.
```

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

$ jps



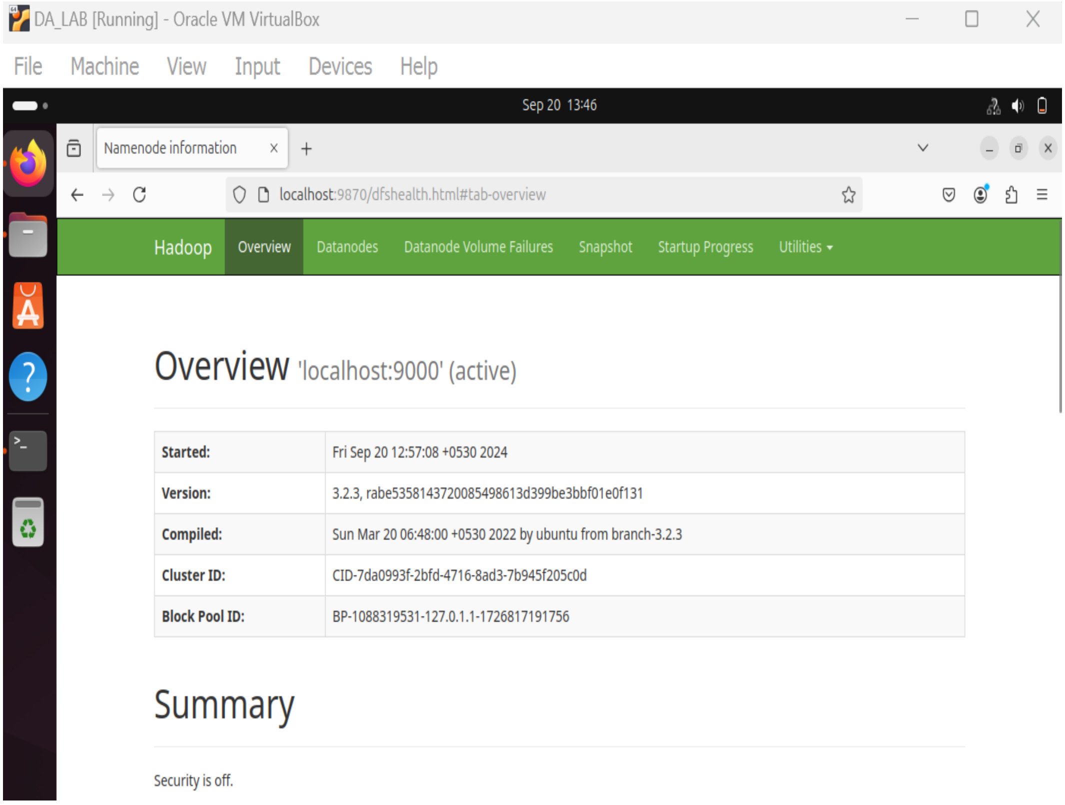
```
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ jps
5315 ResourceManager
5141 SecondaryNameNode
4917 DataNode
6517 NameNode
9386 Jps
5439 NodeManager
```

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**

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


DA_LAB [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Sep 20 13:46
Namenode information
localhost:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities

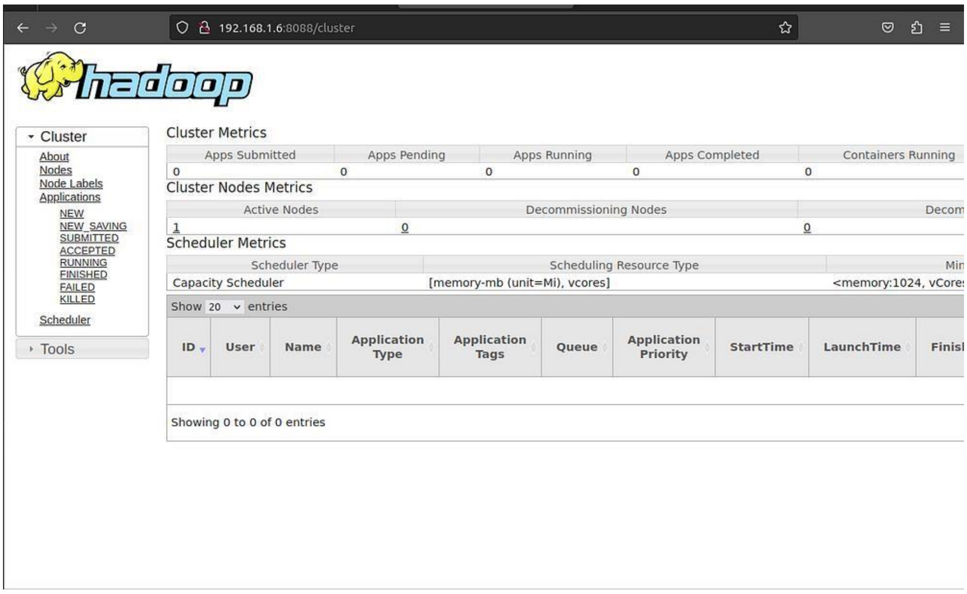
Overview 'localhost:9000' (active)

Started:	Fri Sep 20 12:57:08 +0530 2024
Version:	3.2.3, rabe5358143720085498613d399be3bbf01e0f131
Compiled:	Sun Mar 20 06:48:00 +0530 2022 by ubuntu from branch-3.2.3
Cluster ID:	CID-7da0993f-2bfd-4716-8ad3-7b945f205c0d
Block Pool ID:	BP-1088319531-127.0.1.1-1726817191756

Summary

Security is off.

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



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decom
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Min
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCore:

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	Finis

Showing 0 to 0 of 0 entries

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
```

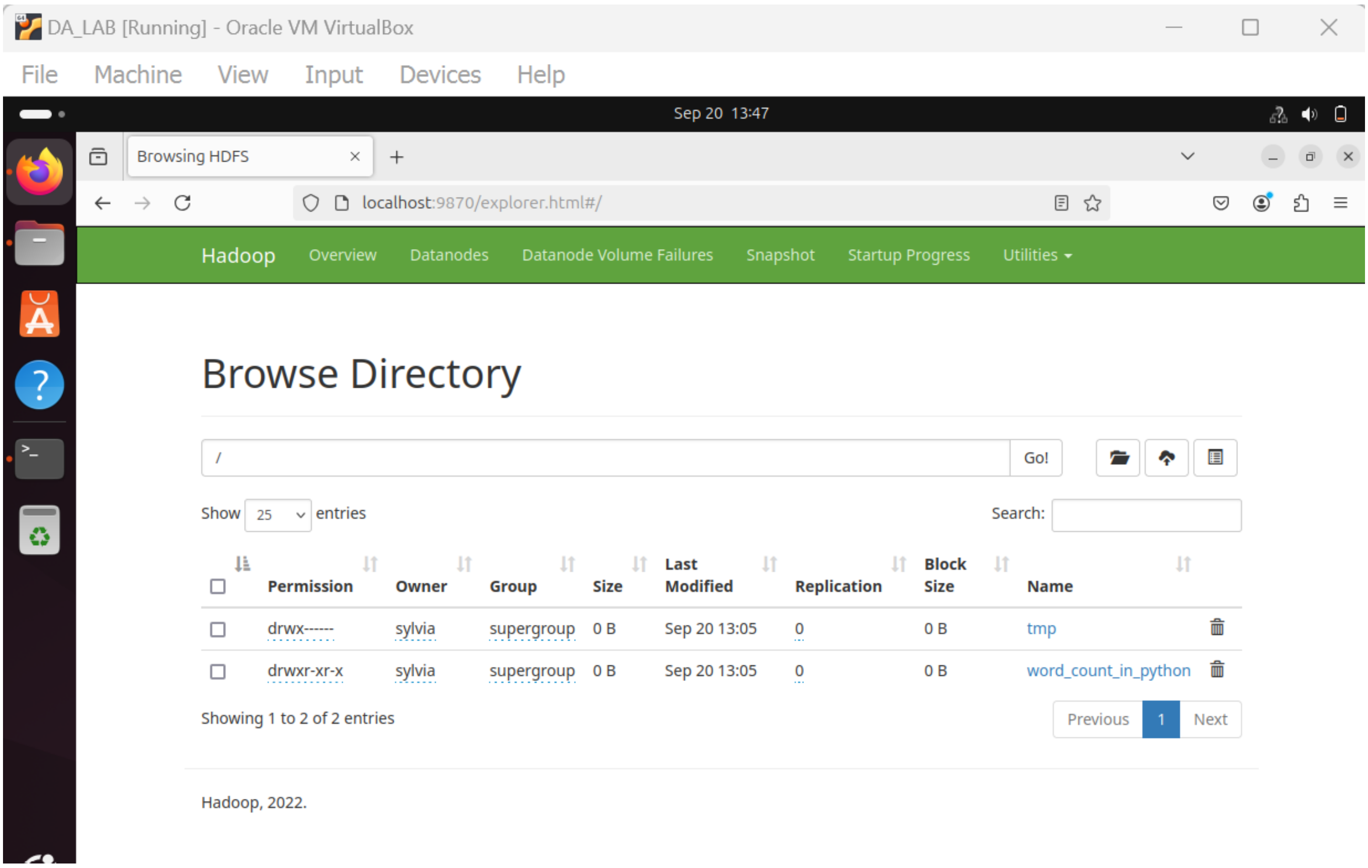
Next, run the following command to list the above directory:

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:



Browse Directory

/

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	drwx------	sylvia	supergroup	0 B	Sep 20 13:05	0	0 B	tmp
☐	drwxr-xr-x	sylvia	supergroup	0 B	Sep 20 13:05	0	0 B	word_count_in_python

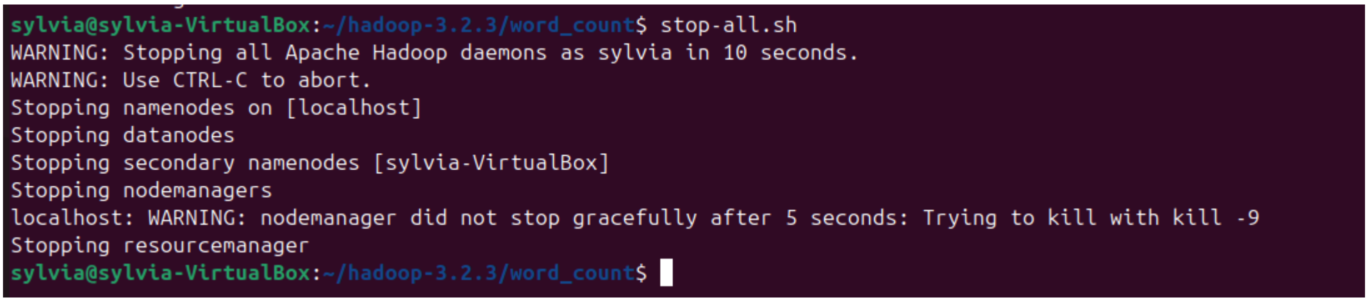
Showing 1 to 2 of 2 entries

Hadoop, 2022.

Step 15 – Stop Hadoop Cluster

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

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



```
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as sylvia in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [sylvia-VirtualBox]
Stopping nodemanagers
localhost: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
Stopping resourcemanager
sylvia@sylvia-VirtualBox:~/hadoop-3.2.3/word_count$
```

Result:

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