# **Setup Multi-Node Hadoop Cluster on CentOS and Oracle Virtualbox**

**Pre-Requisites:**

<https://www.centos.org/download/>, I have installed CentOS7

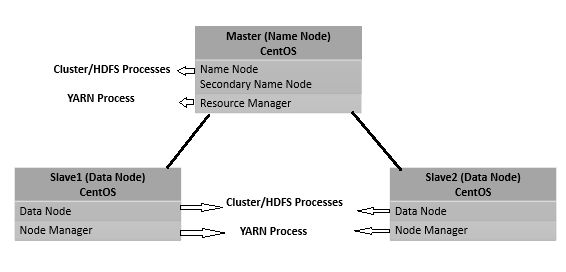
JDK 1.8

Hadoop 2.9.2

WinSCP – to copy files from local to virtual-machine

|  |
| --- |
| **About components of Hadoop cluster:** |
| **Master Node**, keeps the knowledge about the distributed system  **Master Node > Name Node**: It manages the distributed file system and knows where the store data blocks inside the cluster  **Master Node >** **ResourceManager**: which manages the yarn jobs and take care of the scheduling and executing processes on the slave nodes. **(job tracker** replaced with **ResourceManager**) |
| **Slave nodes,** is actually store the data and provide processing power to run the jobs.  **Slave -> NodeManager**: which manage the execution and task on the node ( **task tracker** replaced with **NodeManager**)  **Slave -> Datanode**: it manages the actual data physically stored on the node |

**Cluster Topology:**



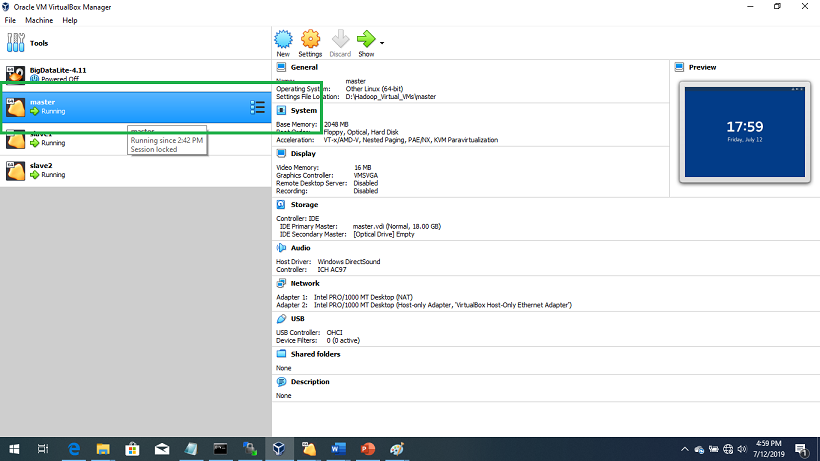
**Cluster Setup Information:**

Hadoop Name Node: 192.168.56.101 (master)

Hadoop Data Node: 192.168.56.102 (slave1)

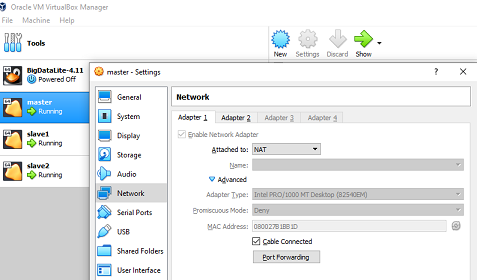
Hadoop Data Node: 192.168.56.103 (slave2)

**Step1: Create Virtual Machine with name “master” and install CentOS**

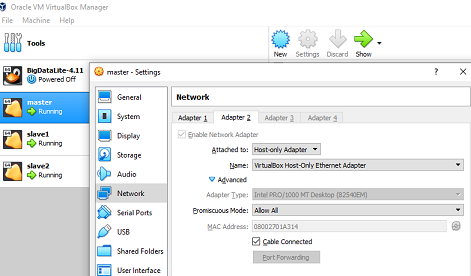


**Step2: Name Node basic configurations on Virtualbox settings:**

**2.1.** Make sure adapter 1 is set to NAT. Because user should be able to access the internet to download the required software directly on VM and also should be able to connect talk to a specific node as internal network

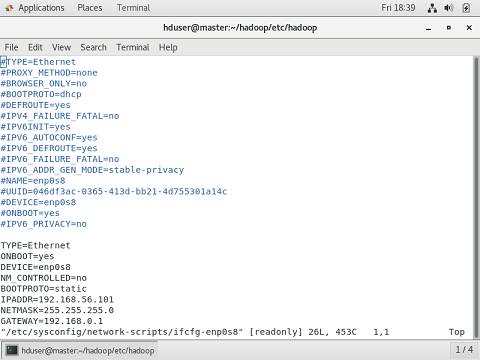


**2.2.** Adapter2, here we are creating local network of 3 machines(nodes). Select like showing down:

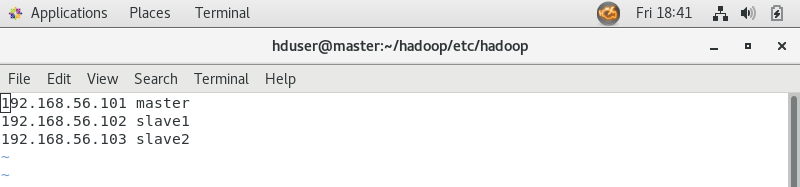


**Step3: Start master node from virtual box, here the master node up and running and logged in as root to make some configurations. Try ifconfig and check the IP Address is set or not. If not set then follow below steps:**

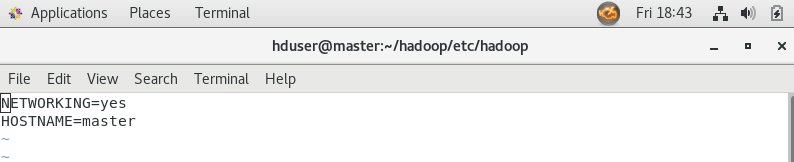
**3.1.** edit /etc/sysconfig/network-scripts/ifcfg-enp0s8



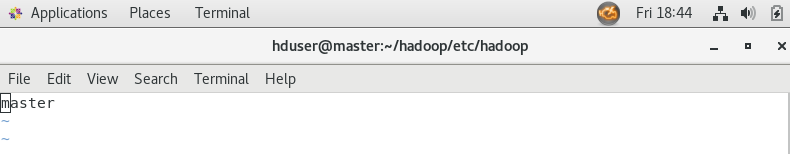
**3.2**. Edit /etc/hosts file like below:



**3.3**.Edit /etc/sysconfig/network like below:



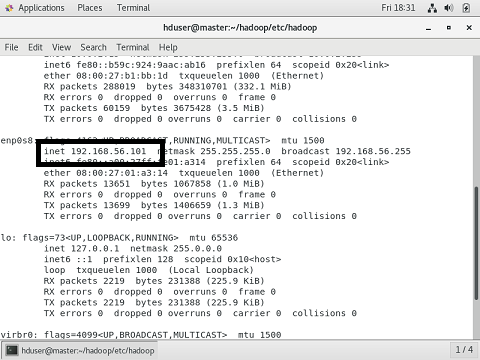
**3.4**. Edit /etc/hostname like below:

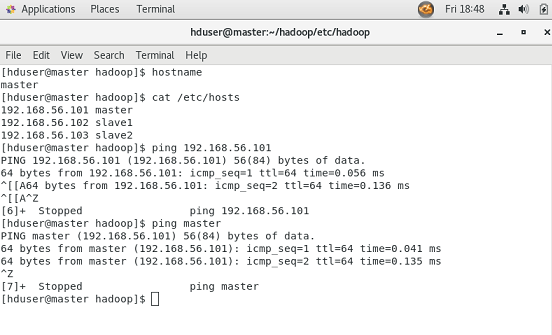


**3.5**. Restart network using below command

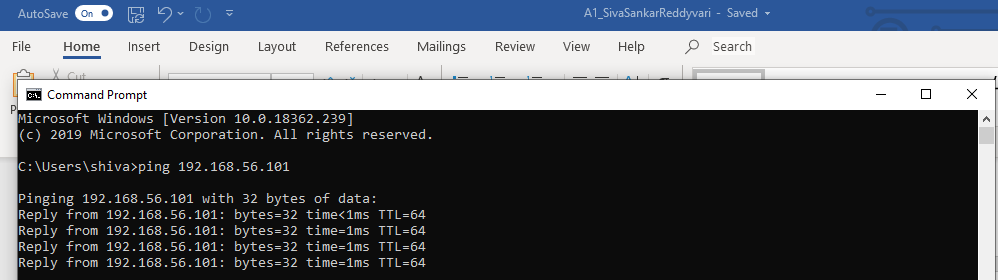
service network restart

**3.6.** Now cross check all network configurations like shown below:





No we are able to ping from windows command line as shown below:



**Step4: Recommended dedicated user for Hadoop to helps to separate Hadoop installation from other software applications and user accounts which are running from the same machine.**

# groupadd Hadoop

# useradd -G Hadoop hduser

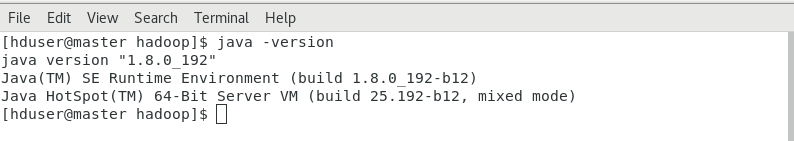
# passwd hduser

New password:

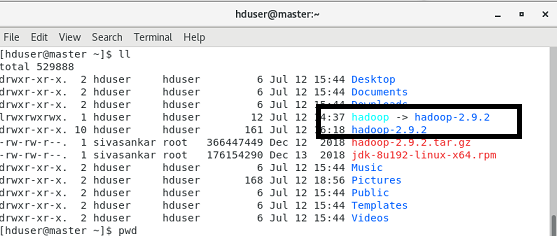
Retype new password:

**Step5: Login as hduser which created above, Install JDK8, download the rpm file and install using below command:**

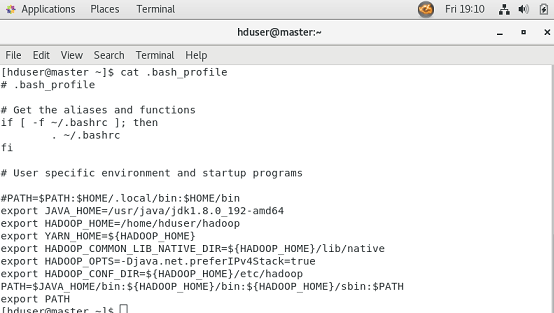
**rpm -i <jdk filename>**

**Step6: Install Hadoop2.9.2, download hadoop2.9.2 and extract and also create** **symlink for Hadoop (benefits: don’t have to type the longer folder name along with version and also in future will be able to replace the folder with an version without change the folder name in many places).**

**symlink example: # ln -s hadoop2.9.2 hadoop**



**Step7: Edit .bash\_profile file, setup paths**



After changes done, run to apply changes

# source .bash\_profile

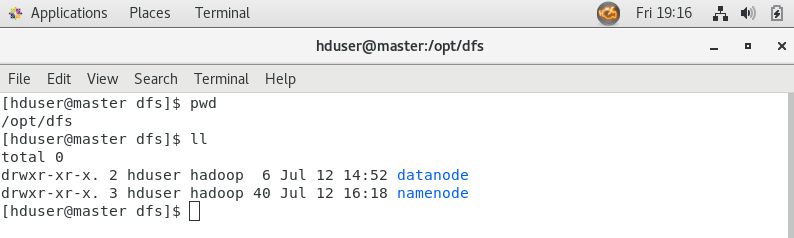
**Step8: Login as root, create namenode and datanode folders to store various blocks information store. I have created these folders under /opt/dfs/:**

root$master # cd /opt/

root$master # mkdir -p dfs/namenode

root$master # mkdir -p dfs/datanode

root$master # chown -R hduser:Hadoop dfs



**Step9: Configure Hadoop, login as hduser**

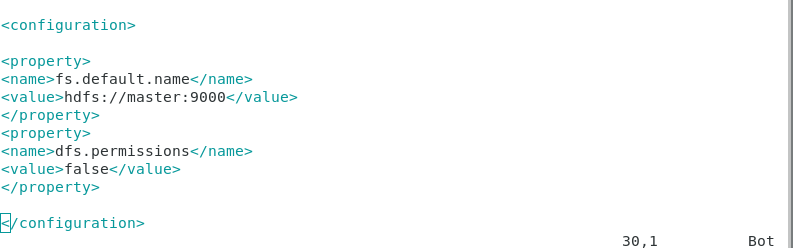
$ cd $HADOOP\_CONF\_DIR

Edit Hadoop-env.sh

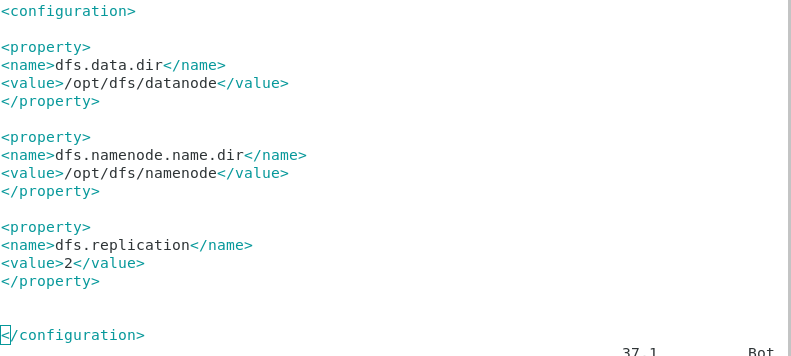
Set JAVA\_HOME=/usr/java/jdk…. <java path>

Set HADOOP\_LOG\_DIR=home/hduser/hadoop/logs

Edit core-site.xml



Edit hdfs-site.xml



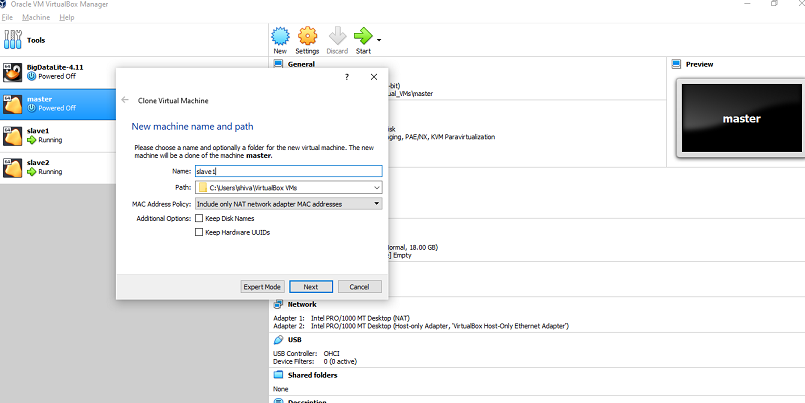
Edit mapred-site.xml



Edit yarn-site.xml



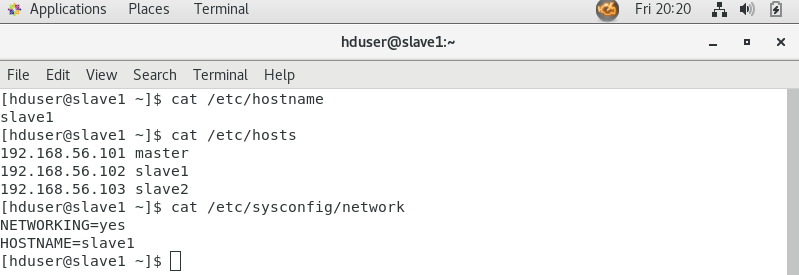
**Step10: clone slave1 and slave2 from master node as shown below:**



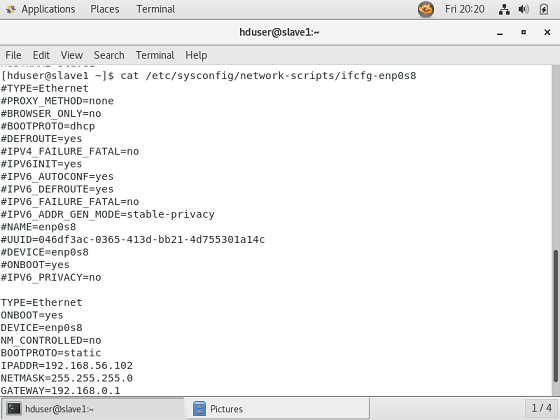
After clone both slave1 and slave2, start all master and slave1 and slave2 nodes.

**10.1.** Login as root and do the changes for both slave nodes.

Edit /etc/hostname , /etc/sysconfig/network as shown below:



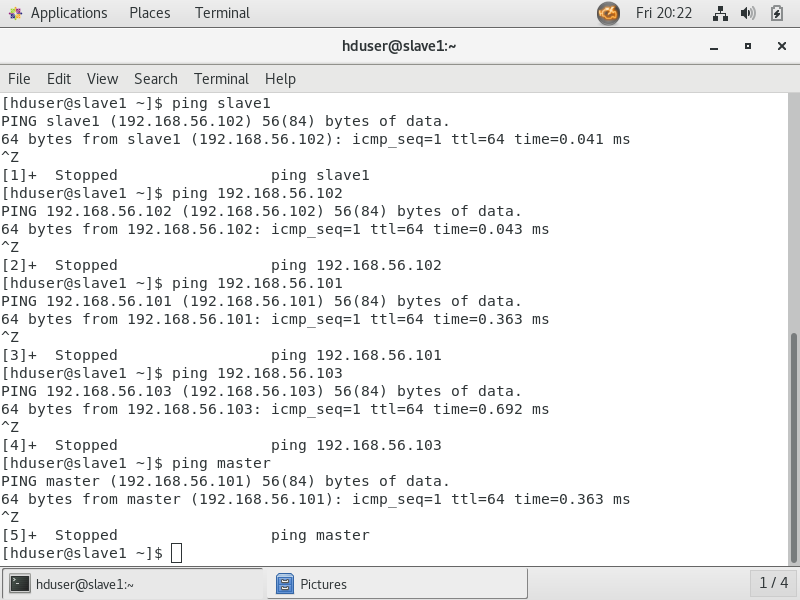
edit /etc/sysconfig/network-scripts/ifcfg-enp0s8



Restart network:

service network restart

Now cross check all network configurations like shown below:



**Step11: Configuring SSH key pair login**

[hduser@master]$ ssh-keygen -t rsa

[hduser@master]$ ssh-copy-id -i ~/.ssh/id\_rsa.pub hduser@master

[hduser@master]$ ssh-copy-id -i ~/.ssh/id\_rsa.pub hduser@slave1

[hduser@master]$ ssh-copy-id -i ~/.ssh/id\_rsa.pub hduser@slave2

[hduser@master]$ chmod 0600 ~/.ssh/authorized\_keys

Now, can login from one node to another node using ssh

**Step12: Format Name Node on Hadoop Master only**

[hduser@master]$ hadoop namenode -format

If you miss the logs folder creation, create under etc/hadoop folder

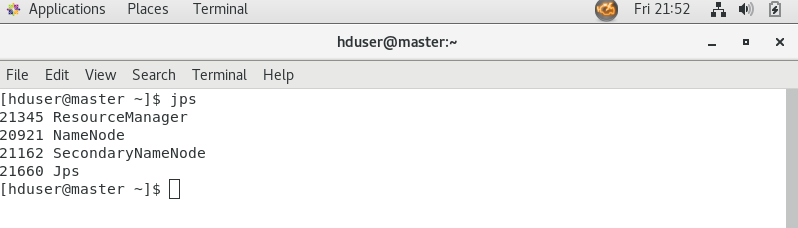
[hduser@master]$ mkdir logs

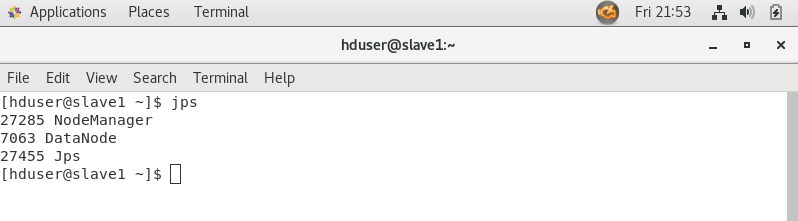
**Step13: Start Hadoop Services**

[hduser@master]$ start-dfs.sh (Starts the Hadoop DFS daemons, the namenode and datanodes)

[hduser@master]$ start-yarn.sh (starts ResourceManager and NodeManager)

Jobs running from master node:

Jobs running from slave1 node:



Jobs running from slave2 node:

