**Documentation For Hadoop Multi Node Cluster Setup**

Follow the following steps :

**DO THE FOLLWOWING STEPS FOR BOTH MASTER AND SLAVE NODE**

1. Install SSH using the following command:

sudo apt install ssh

1. Install PDSH using the following command:

sudo apt install pdsh

1. Open the .bashrc file with the following command:

sudo nano .bashrc

At the end of the file add the following line:

export PDSH\_RCMD\_TYPE=ssh

1. Now let’s configure SSH. Let’s create a new key using the following command:

ssh-keygen -t rsa -P ""

Just press Enter every time that is needed.

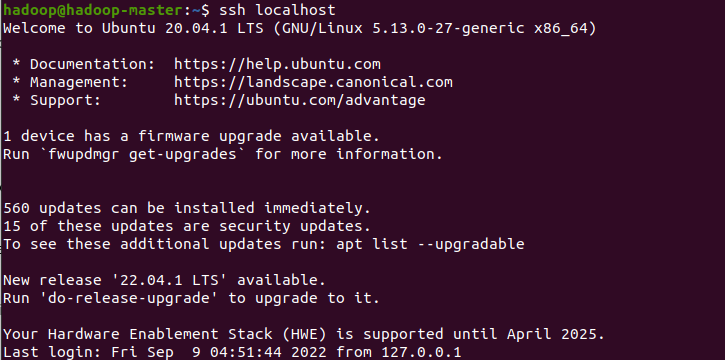
1. Now we need to copy the public key to the authorized\_keys file with the following command:

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

1. Now we can verify the SSH configuration by connecting to the localhost:

ssh localhost

Just type “yes” and press Enter when needed.



1. This is the step where we install Java 8. We use this command:

sudo apt install openjdk-8-jdk

1. Download Hadoop using the following command:

sudo wget -P ~ <https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz>

If the above command doesn’t work due to proxy, just manually download the hadoop binary version using the browser, just by clicking the above link.

1. Now, if go to the folder where this hadoop file is downloaded and apply the following command :

tar xzf hadoop-3.3.4.tar.gz

1. Now, change the hadoop-3.3.4 folder name to hadoop (this makes it easier to use). Use this command:

mv hadoop-3.3.4 hadoop

1. Open the hadoop-env.sh file in the nano editor to edit JAVA\_HOME. Make sure you are in the directory where hadoop folder is present.

nano hadoop/etc/hadoop/hadoop-env.sh

Paste this line to JAVA\_HOME:

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64/

1. Change the hadoop folder directory to /usr/local/hadoop. This is the command ( again in the same terminal ):

sudo mv hadoop /usr/local/hadoop

1. Open the environment file on nano with this command:

sudo nano /etc/environment

Then, add the following configurations:

PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/usr/local/hadoop/bin:/usr/local/hadoop/sbin"JAVA\_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"

1. Now we will add a user called hadoop, and we will set up it’s configurations:

sudo adduser hadoop

Provide the password and you can leave the rest blank, just press Enter.

Now type these commands:

sudo usermod -aG hadoop hadoop

sudo chown hadoop:root -R /usr/local/hadoop/

sudo chmod g+rwx -R /usr/local/hadoop/

sudo adduser hadoop sudo



**ONCE YOU HAVE ADDED THE USER, DO ALL THE FOLLOWING COMMANDS AFTER RUNNING THE FOLLOWING COMMAND (FOR BOTH THE MACHINES):**

su hadoop

1. Open the .bashrc file with the following command:

sudo nano .bashrc

At the end of the file add the following lines:

#Hadoop Related Options

export HADOOP\_HOME=”/usr/local/hadoop”

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

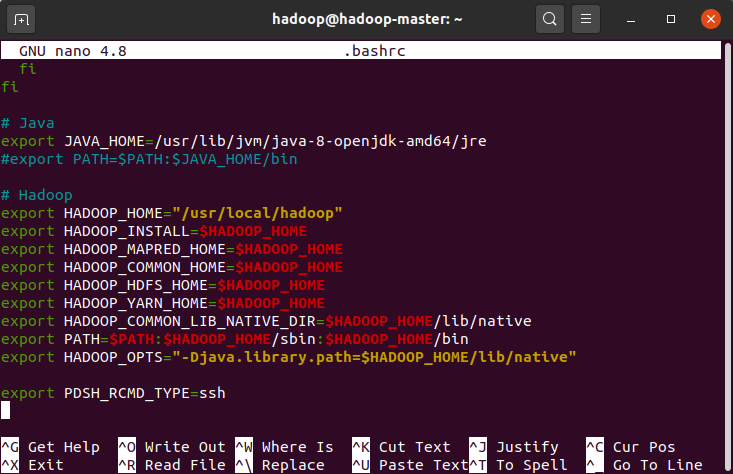
export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export 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”



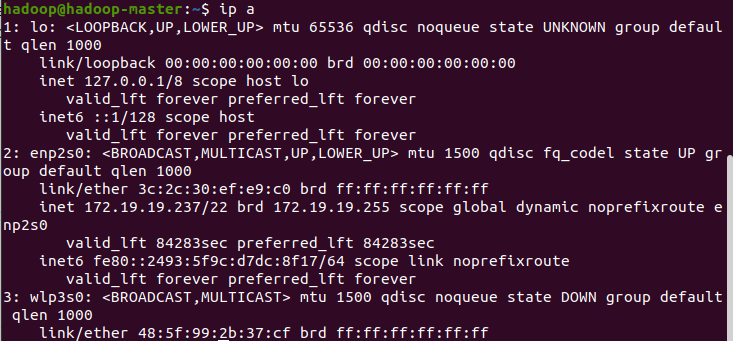
Save the file.

It is vital to apply the changes to the current running environment by using the following command:

source ~/.bashrc

1. Now we will add the machine ip addresses into the hosts file for both the nodes. Use the following command to get the ip address of each machine.

MAKE SURE BOTH THE MACHINES ARE ON THE SAME NETWORK



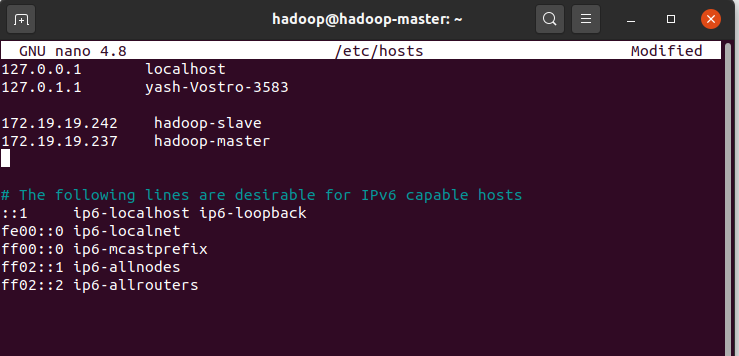
The ip addr of the above machine is 172.19.19.237

Write down the ip addresses of both the machines.

Now, run the following command on both machines:

sudo nano /etc/hosts

Now write down the ip addresses for the master and slave node as shown



Save the file.

1. Now, in the **master node**, apply the following command:

ssh-keygen -t rsa

Now, run the following commands:

ssh-copy-id -i ~/.ssh/id\_rsa.pub hadoop@hadoop-master

ssh-copy-id -i ~/.ssh/id\_rsa.pub hadoop@hadoop-slave

**DO THE FOLLWOWING STEPS FOR BOTH MASTER AND SLAVE NODE IN ORDER TO PROCESS FASTER** ( the process mentioned in the tutorials make changes only in the master node and the copies all the files into slave node which takes a lot of time )

1. Open core-site.xml file on nano:

sudo nano /usr/local/hadoop/etc/hadoop/core-site.xml

Then add the following configurations:

<configuration>

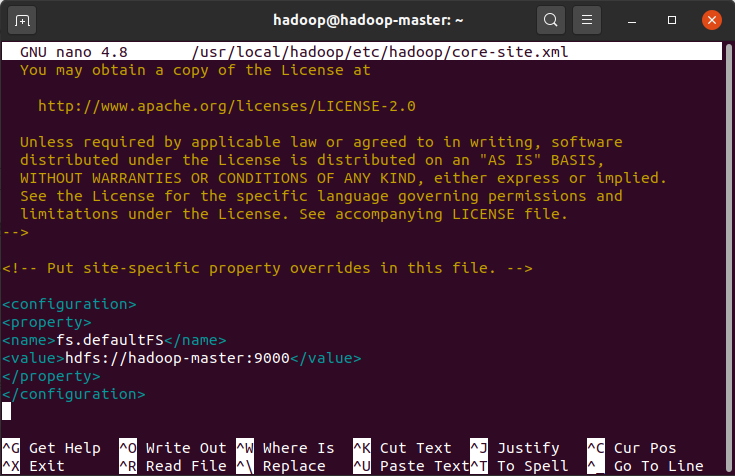
<property>

<name>fs.defaultFS</name>

<value>hdfs://hadoop-master:9000</value>

</property>

</configuration>



1. open the hdfs-site.xml file.

sudo nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml

Add the following configurations:

<configuration>

<property>

<name>dfs.name.dir</name><value>/usr/local/hadoop/data/nameNode</value>

</property>

<property>

<name>dfs.data.dir</name><value>/usr/local/hadoop/data/dataNode</value>

</property>

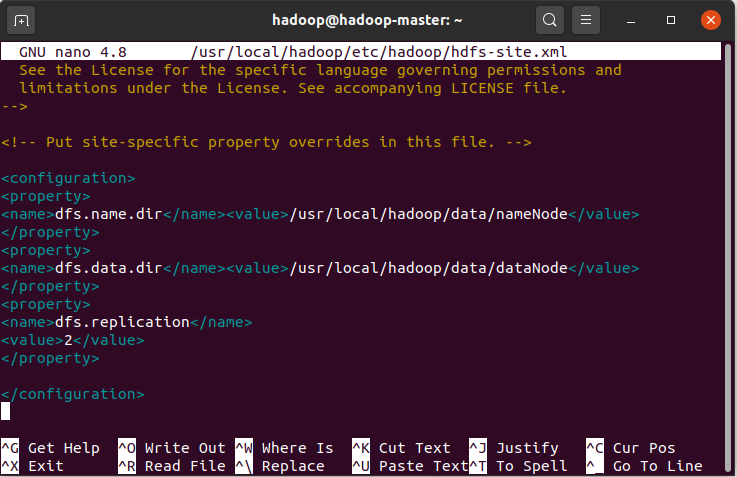
<property>

<name>dfs.replication</name>

<value>2</value>

</property>

</configuration>



1. Open the mapred-site.xml file.

sudo nano /usr/local/hadoop/etc/hadoop/mapred-site.xml

Add the following configurations:

<configuration>

<property>

<name>mapred.job.tracker</name>

<value>hadoop-master:9001</value>

</property>

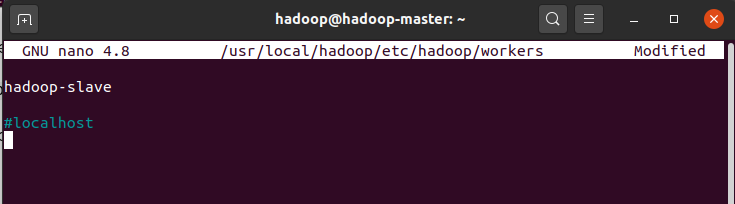
</configuration>

1. Now, on **hadoop-master**, let’s open the workers file:

sudo nano /usr/local/hadoop/etc/hadoop/workers

Add the following lines: (the slave name)

hadoop-slave



1. Now, in the **slave node**, open the yarn-site.xml file:

sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.xml

add the following configurations:

<property>

<name>yarn.resourcemanager.hostname</name>

<value>hadoop-master</value>

</property>

1. Now we need to format the HDFS file system. Run this command in the **master node:**:

hdfs namenode -format

1. Now, in the **master node,** start HDFS with this command:

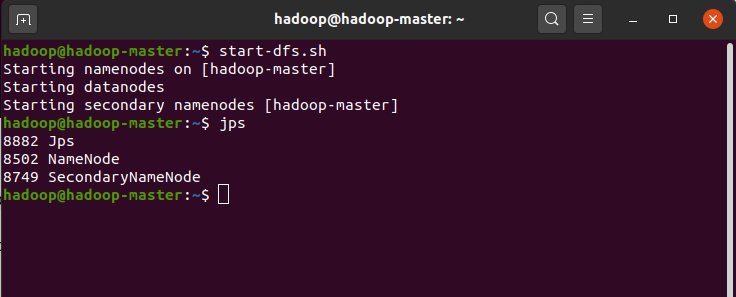
start-dfs.sh

Verify the same by applying the following command in both the machines:

jps

The output will look like this:

For master node:

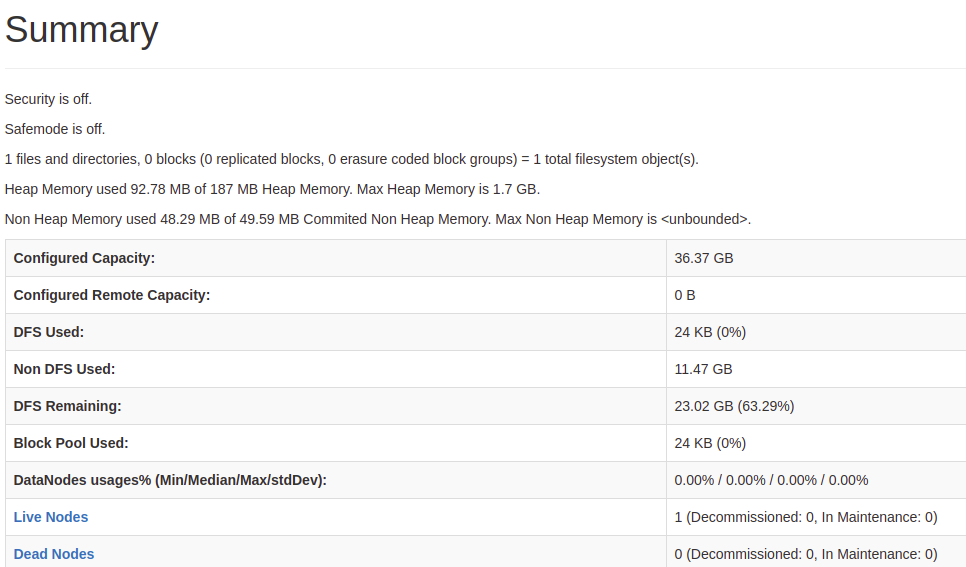


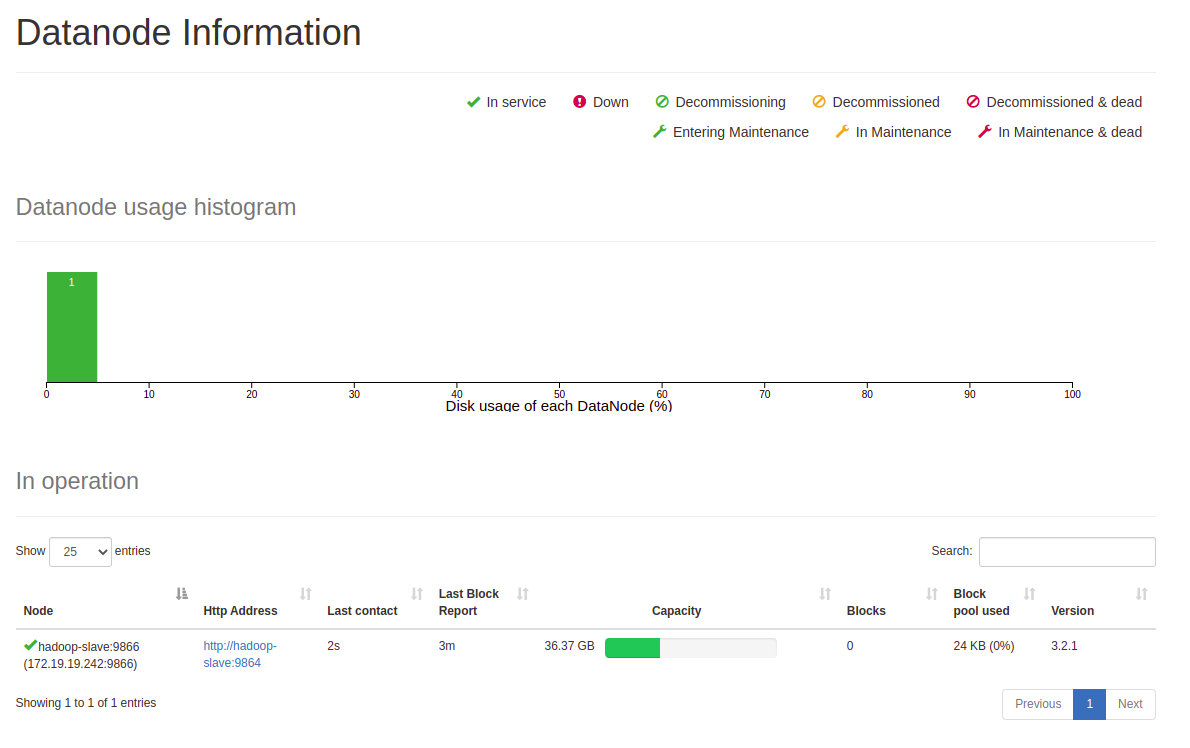
For slave node :



Also verify if everything is working fine by going to **localhost:9870** in the **master node.**

There it must be showing 1 live node



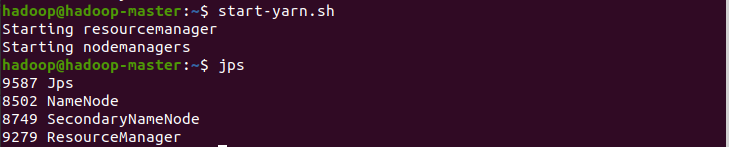


1. Now, on the **master node**, let’s start yarn. Use this command:

start-yarn.sh

Verify the same using the jps command on both the machines. The output will be :

For master node:

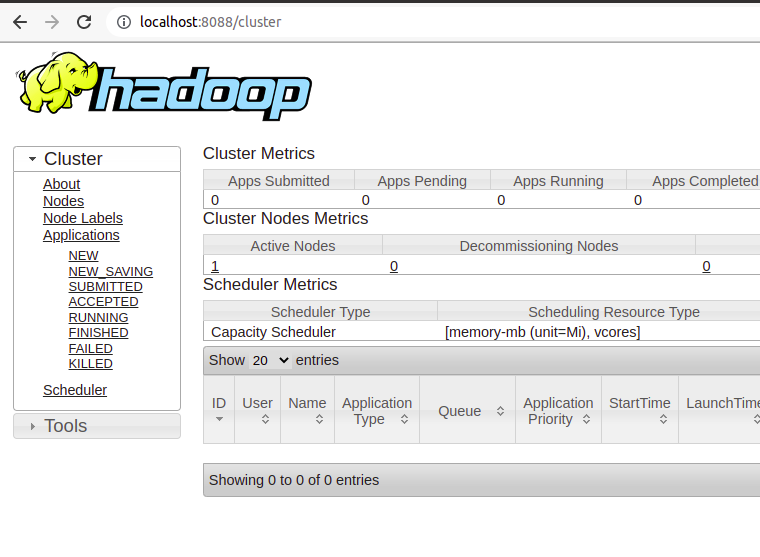


For slave node:



Also verify the same on the **master node,**  by going to **localhost:8088** in the browser.

You will find details of the network. There must be 1 active node, if everything went well.



1. Now, in order to stop the network, just use the following command on the master node:

stop-all.sh

1. Next time, whenever you want to use this network, just update the ip addresses in the /etc/hosts file (in both the machines) and you are good to go!!

Just apply the following command on the master node to start the network:

start-all.sh

**References :**

1. <https://medium.com/@jootorres_11979/how-to-set-up-a-hadoop-3-2-1-multi-node-cluster-on-ubuntu-18-04-2-nodes-567ca44a3b12>
2. <https://phoenixnap.com/kb/install-hadoop-ubuntu>
3. <https://www.tutorialspoint.com/hadoop/hadoop_multi_node_cluster.htm>