**Hadoop: Setting up a Single Node Cluster**

Hadoop software can be installed in three modes of operation:

* **Stand Alone Mode**: By default, Hadoop is configured to run in a non-distributed mode, as a single Java process. This is useful for debugging.
* **Pseudo Distributed Mode:** Hadoop can also be run on a single-node in a pseudo-distributed mode where each Hadoop daemon runs in a separate Java process.
* **Fully Distributed Mode**: In Fully Distributed Mode, the daemons NameNode, JobTracker, SecondaryNameNode (Optional and can be run on a separate node) run on the Master Node. The daemons DataNode and TaskTracker run on the Slave Node.

In this tutorial, I will show you how to install Hadoop on a single node cluster in CentOS 7 (Minimal ). This type of configuration is also referenced as **Hadoop Pseudo-Distributed Mode**. I will do the installation on a VMware virtual machine.

## ****Prerequisites****:

* VMware
* CentOS-7-x86\_64-Minimal-1611 (Refer “How to install CentOS 7 Minimal.docx”)
* jdk-8u161-linux-x64.tar.gz
* hadoop-2.8.1

Download Source:Paste the following link into web browser to download

https://drive.google.com/open?id=1WYIOJfupaq2X411KWEROJF1anAD5mfYQ

1. **Install java**

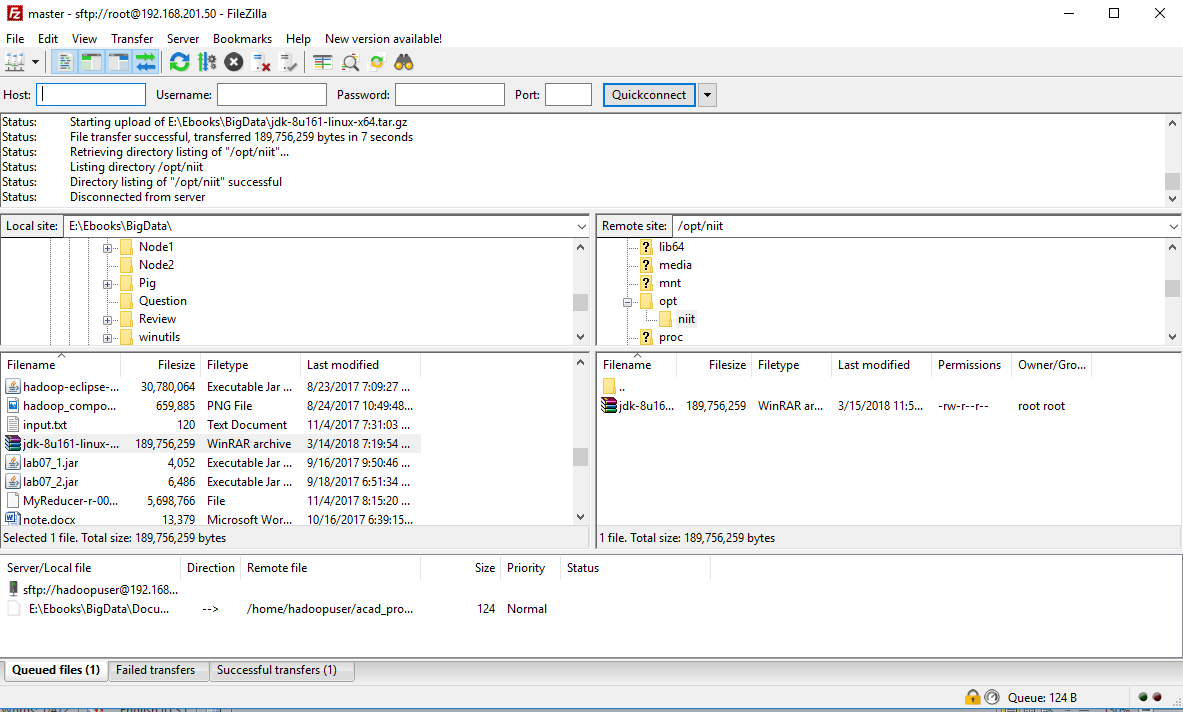
* Create folder to store program



* Move to this folder



* Copy jdk-8u161-linux-x64.tar.gz into VMware by using Filezilla (Refer “How to use Filezilla.docx”). It is copied to /opt/niit



* Extract the Java tar file



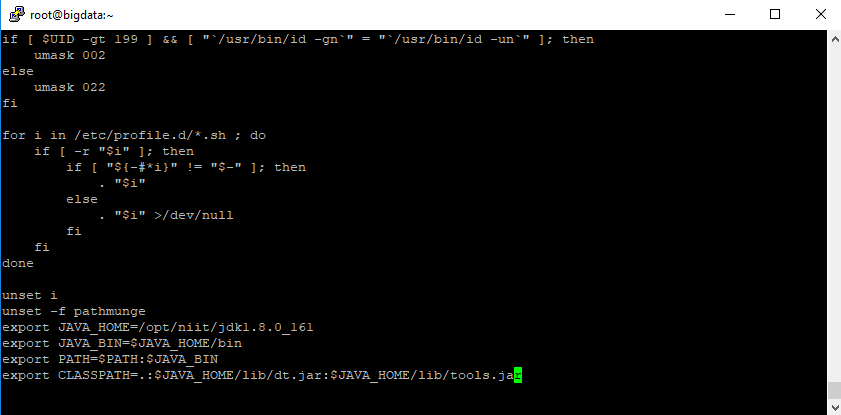
* Configure environment variables: Add these lines at the bottom of **/etc/profile**

export JAVA\_HOME=/opt/niit/jdk1.8.0\_161

export JAVA\_BIN=$JAVA\_HOME/bin

export PATH=$PATH:$JAVA\_BIN

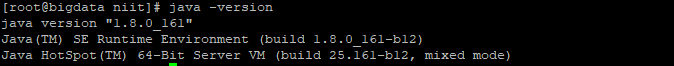
export CLASSPATH=.:$JAVA\_HOME/lib/dt.jar:$JAVA\_HOME/lib/tools.jar



* Reload /etc/profile



* Verify Java version

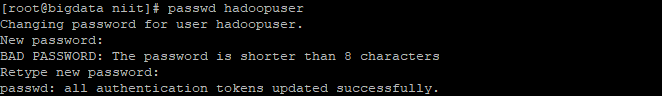


1. **Add a dedicated Hadoop system user**

* Create new user



* Set password for user



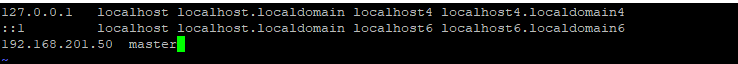
* Change owner **hadoopuser** for **/opt/niit** folder



1. **Set machine hostname**



1. **Add a new record in hosts file**



* Restart network



1. **Install Hadoop**

* Extract tar file hadoop-2.8.1.tar.gz into /opt/niit



* Configure environment variables: Add these lines at the bottom of **/etc/profile**

export HADOOP\_HOME=/opt/niit/hadoop-2.8.1

export PATH=$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin

* Reload /etc/profile

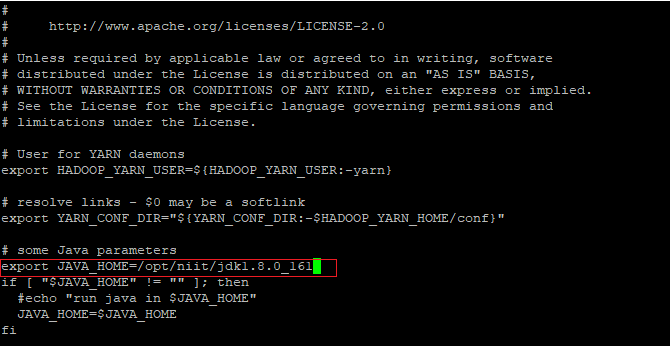


1. **Edit the Hadoop configuration files**

* Edit **$HADOOP\_HOME/etc/hadoop/hadoop-env.sh**



* Edit **$HADOOP\_HOME/etc/hadoop/yarn-env.sh**



* Create temporary folder



* Change owner of tmp folder



* Edit **$HADOOP\_HOME/etc/hadoop/core-site.xml**

<configuration>

<property>

<name>hadoop.tmp.dir</name>

<value>/opt/niit/tmp</value>

</property>

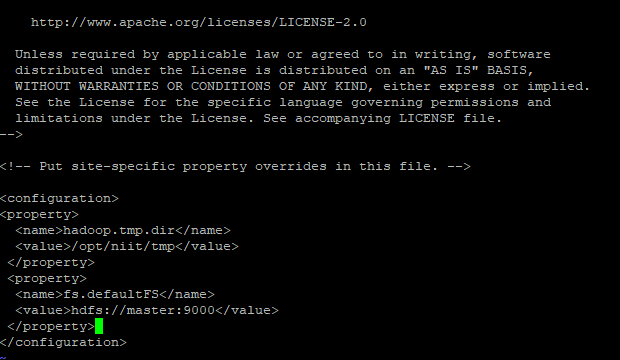
<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>



* In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.



* Edit **$HADOOP\_HOME/etc/hadoop/mapred-site.xml**

<configuration>

<property>

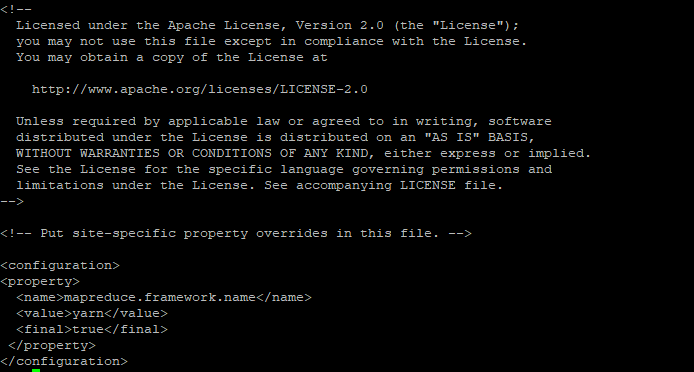
<name>mapreduce.framework.name</name>

<value>yarn</value>

<final>true</final>

</property>

</configuration>



* Create namenode and datanode folder



* Edit **$HADOOP\_HOME /etc/hadoop/hdfs-site.xml.** Add these lines at the bottom of file

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:///opt/niit/hadoop\_store/hdfs/namenode</value>

</property>

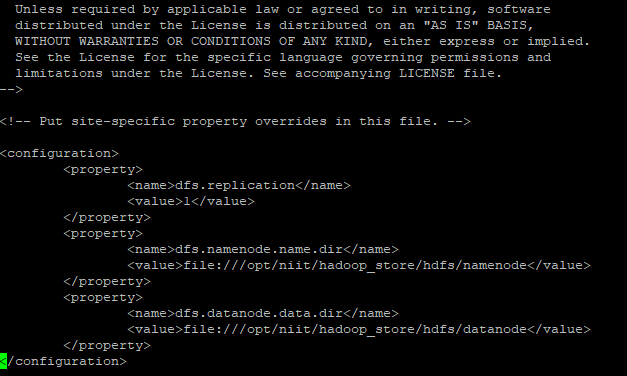
<property>

<name>dfs.datanode.data.dir</name>

<value>file:///opt/niit/hadoop\_store/hdfs/datanode</value>

</property>

</configuration>



* Edit **$HADOOP\_HOME /etc/hadoop/yarn-site.xml.** Add these lines at the bottom of file

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>master:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address</name>

<value>master:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address</name>

<value>master:8031</value>

</property>

<property>

<name>yarn.resourcemanager.admin.address</name>

<value>master:8033</value>

</property>

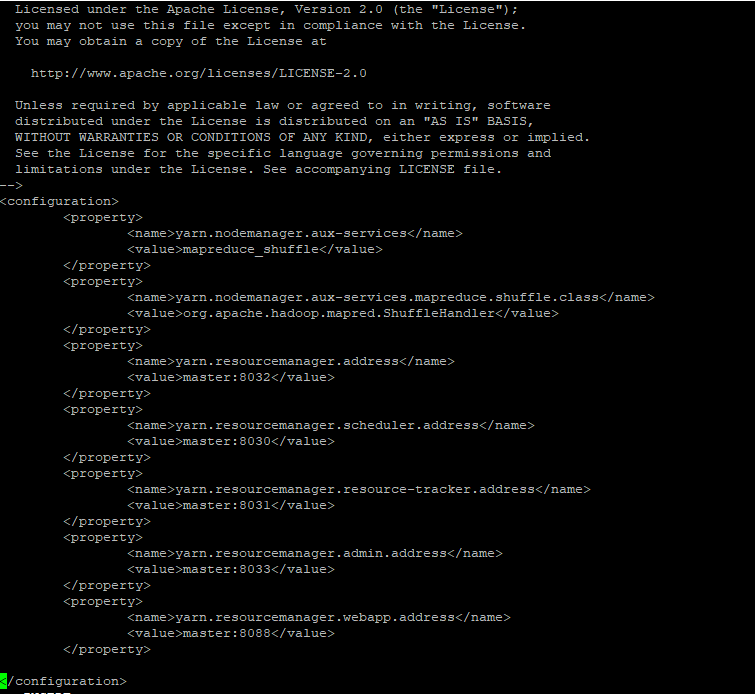
<property>

<name>yarn.resourcemanager.webapp.address</name>

<value>master:8088</value>

</property>

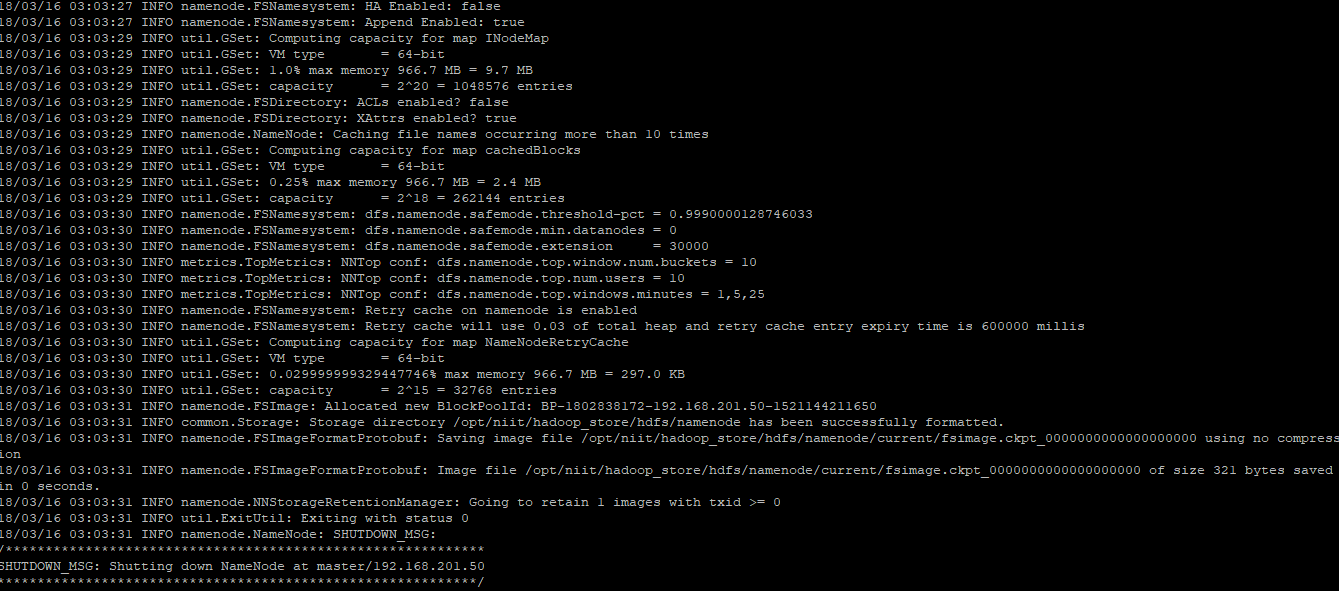
</configuration>



1. **Format Hadoop Namenode**

Enter the command as below

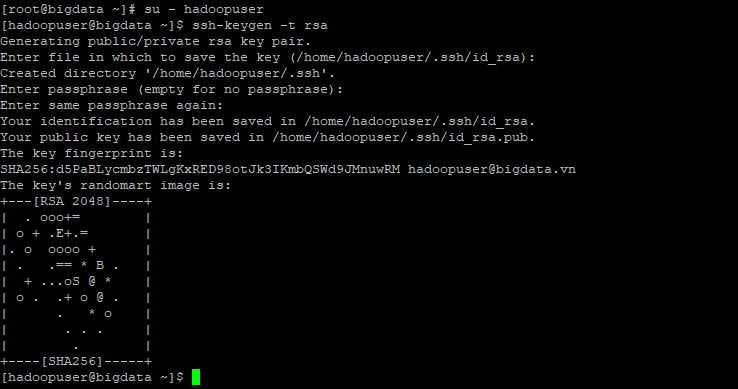


****

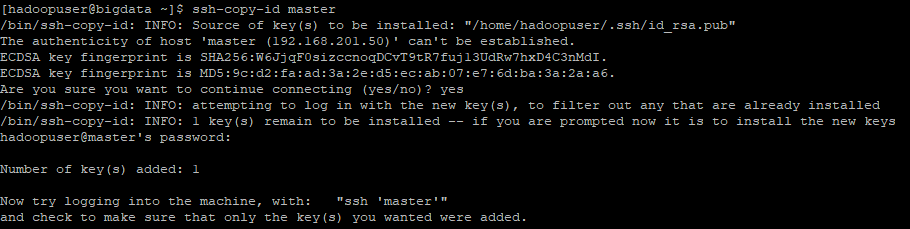
1. **Configure SSH**

SSH needs to be setup to allow password-less login for the hadoop user from machines in the cluster

* Login hadoopuser and enter the command **ssh-keygen –t rsa**



* Next we enter the command **ssh-copy-id master**



* Test SSH



1. **Start Hadoop**

First, we need to make sure hadoopuser have permission on **/opt/niit** and **/opt/niit/hadoop-2.8.1/logs**





Second, enter the following command to update time



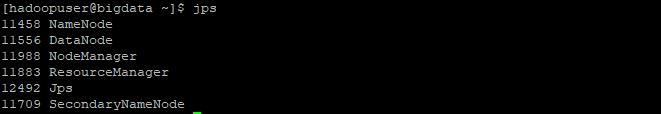




Enter the command: **start-all.sh**

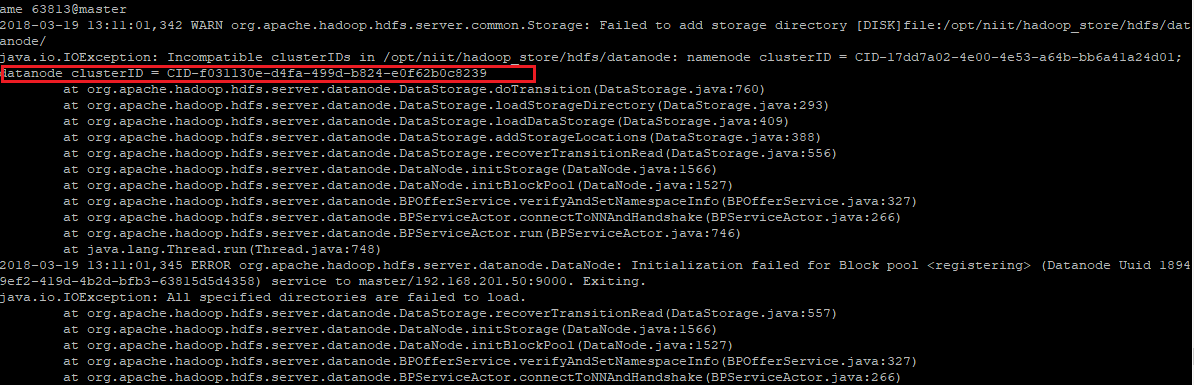


Check the service status with the following command



If DataNode service is not starting, we check the log and encounter the error “**java.io.IOException: Incompatible clusterIDs**”, we need to run the following command.

hdfs namenode -format -clusterId <clusterID of datanode>





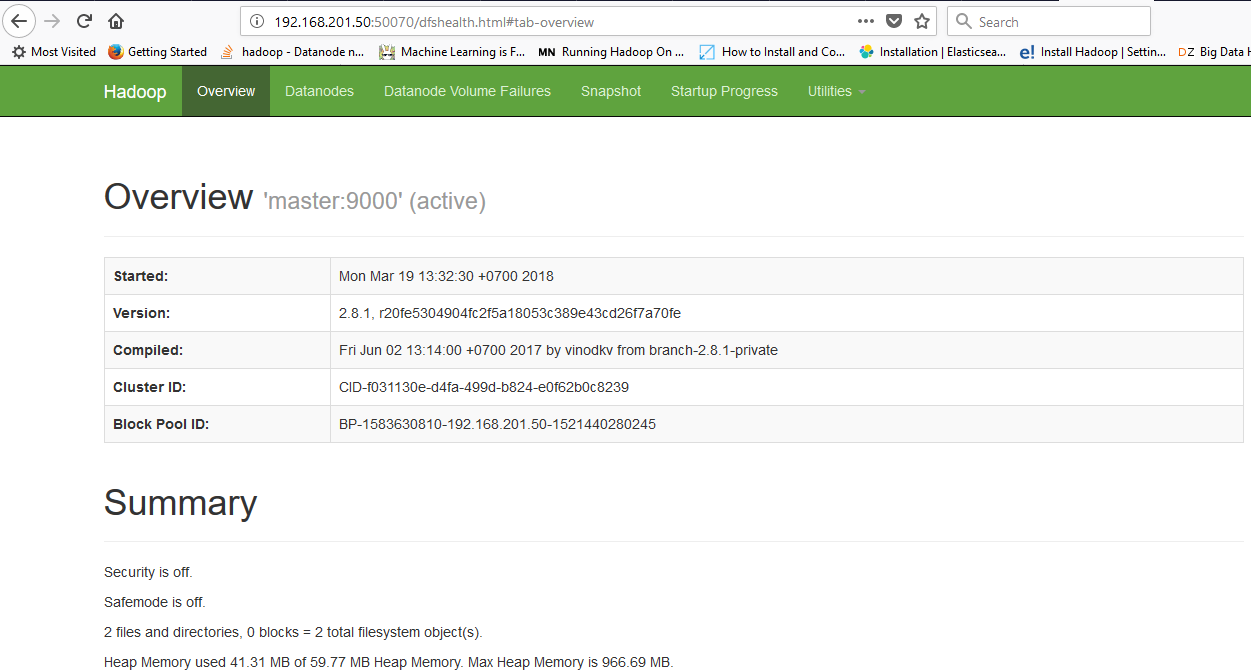
1. **Browse Hadoop Services**

Disable system firewall



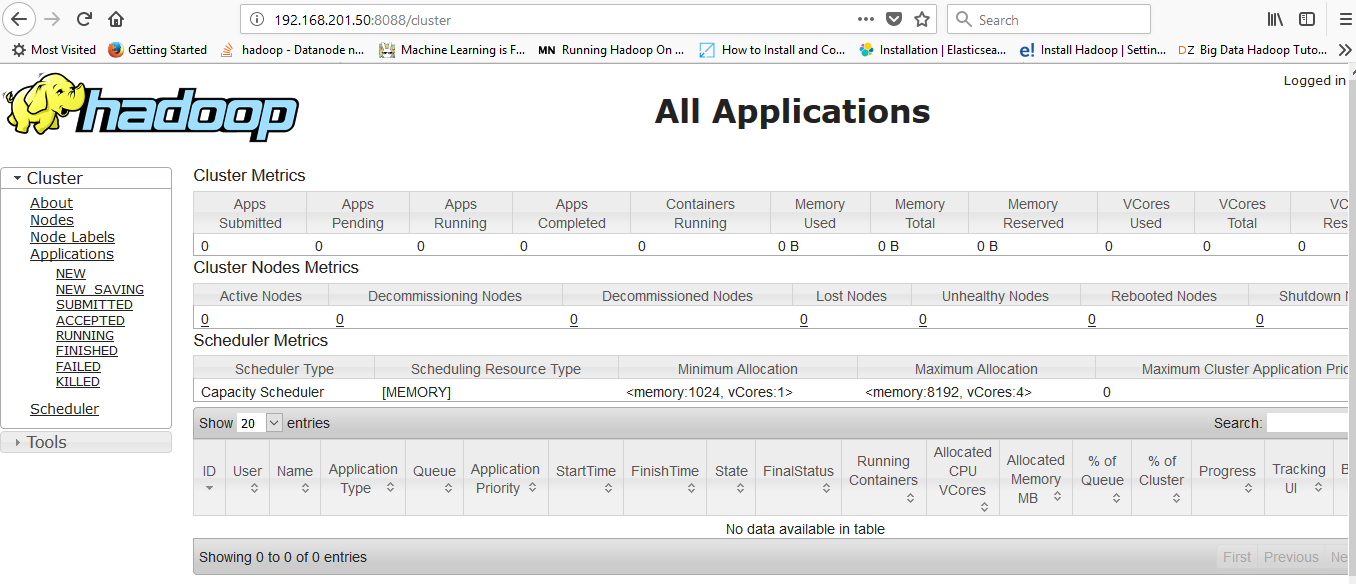
For Hadoop Overview of NameNode service.

**http://192.168.201.50:50070**



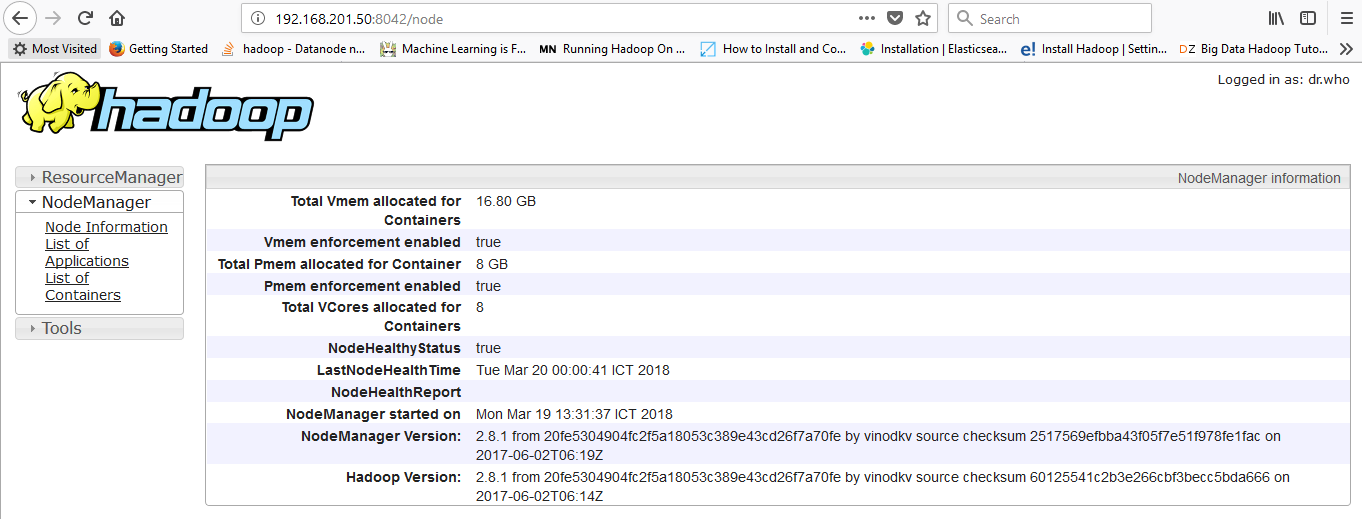
For Cluster and Apps Information (ResourceManager).

**http://192.168.201.50:8088**



For NodeManager Information.

**http://192.168.201.50:8042**



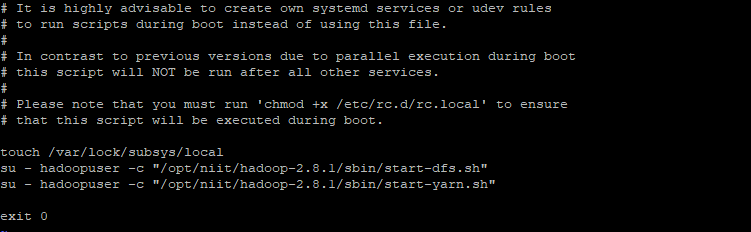
1. **Manage Hadoop Service**

In order to enable Hadoop daemons system-wide, login with root user, open /etc/rc.local file for editing and add the below lines:

su – hadoopuser –c “/opt/niit/hadoop-2.8.1/sbin/start-dfs.sh”

su – hadoopuser –c “/opt/niit/hadoop-2.8.1/sbin/start-yarn.sh”

exit 0



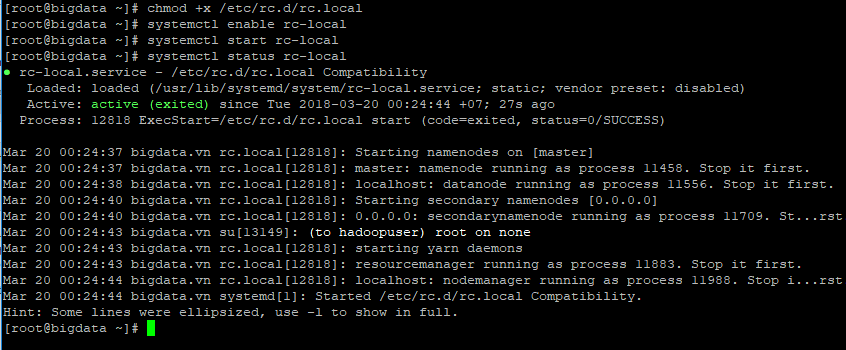
Then, add executable permissions for rc.local file and enable, start and check service status by issuing the below commands:

chmod +x /etc/rc.d/rc.local

systemctl enable rc-local

systemctl start rc-local

systemctl status rc-local



Next time you **reboot** your machine the **Hadoop** services will be automatically started for you

**Reference: Start/Stop Hadoop command**

|  |  |
| --- | --- |
| Start all daemons | start-all.sh |
| Stop all daemons | stop-all.sh |
| Start NameNode | hadoop-daemon.sh start namenode |
| Start DataNode | hadoop-daemon.sh start datanode |
| Start ResourceManager | yarn-daemon.sh start resourcemanager |
| Start NodeManager | yarn-daemon.sh start nodemanager |