# Installation of Hadoop 3.1.4 on ubuntu 18.04

## Step 1: Installation of openJDK-8

```
$ Sudo apt install openjdk-8-jdk openjdk-8-jre

$ java -version

$ sudo apt install vim openssh-server openssh-client
```

## Step 2: Adding the Jdk path to the path variable

Open ~/.bashrc and add

```
$ sudo vim ~/.bashrc

#go to the last line and add the following

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

export PATH=$PATH:$JAVA_HOME
```

**Inform the OS about the modification**

```
$ source ~/.bashrc
```

Type

```
$ echo $JAVA_HOME

$ echo $PATH
```

## Step 3: Add a dedicated user for the HADOOP

```
$ sudo adduser hadoop

$ sudo usermod -aG sudo hadoop
```

## Step 4: Once the user is added, login to the user “Hadoop” to generate the ssh key for passwordless login ( hadoop@machinename)

```
$ sudo su - hadoop

$ ssh-keygen -t rsa
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

```
$ chmod 0600 ~/.ssh/authorized_keys
```

Check the login to localhost using ssh is valid

```
$ ssh localhost
```

**Once the connection is made, logout from ssh**

```
$ exit
```

**Step 5: Download the latest binary from Hadoop site**

**“ hadoop-3.1.4.tar.gz “**

```
$ tar -xvzf hadoop-3.1.4.tar.gz
```

```
$ mv hadoop-3.1.4 /usr/local/hadoop
```

**Step 6: Setup the path variables for hadoop**

```
$ sudo vim /etc/profile.d/hadoop_java.sh
```

Add the following lines to it

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

export HADOOP_HOME=/usr/local/hadoop

export HADOOP_HDFS_HOME=$HADOOP_HOME

export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME

export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native

export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

exportHADOOP_OPTS="$HADOOP_OPTS -Djava.library.path=$HADOOP_HOME/lib/native”
```

**Save and exit. Then source the file**

```
$ source /etc/profile.d/hadoop_java.sh
```

**Confirm your hadoop and hdfs version**

```
$ hadoop version

$ hdfs version
```

### Step 7: Configuring Hadoop

Navigate to /usr/local/hadoop/etc/hadoop  and type ls

```
$ cd /usr/local/hadoop/etc/hadoop

$ hadoop@machine: /usr/local/hadoop/etc/hadoop: ls
```

**Give the permission for the hadoop folder to hadoop user**

```
$ sudo chown -R hadoop:hadoop /usr/local/hadoop
```

### Step 7a: Specify JAVA_HOME in hadoop-env.sh (/usr/local/hadoop/etc/hadoop)

```
$ vim hadoop-env.sh
```

Add the following line in java implementation

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64 (54 line)
```

Save and exit

### Step 7b: Modify core-site.xml to setup web portal for hadoop

Add the following lines to it

```
<configuration>
    <property>
   <name>fs.default.name</name>
   <value>hdfs://localhost:9000</value>
   <description>The default file system URI</description>
   </property>
   <property>

  <name>hadoop.tmp.dir</name>
  <value>/usr/local/hadoop/htemp</value>
   </property>
</configuration>
```

**Step 7c: Modify hdfs-site.xml to setup namenode and datanode path and replication factor**

**Create a folder for namenode and datanode usage**

```
$ ls
```

**Give the permission for the hdfs and htemp folder to hadoop user**

```
$ sudo chown -R hadoop:hadoop /usr/local/hadoop/hdfs
sudo chown -R hadoop:hadoop /usr/local/hadoop/htemp
```

**Modify hdfs-site.xml and add the following lines inside**

```
<configuration>
  <property>
  <name>dfs.replication</name>
  <value>1</value>
  </property>
  <property>
  <name>dfs.name.dir</name>
  <value>file:/usr/local/hadoop/hdfs/namenode</value>
  </property>


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

**Step 7d: Configure the mapreduce framework by editing the mapred-site.xml**

**Modify the mapred-site.xml and add the following lines**

```
 <configuration>
<property>
   <name>mapreduce.framework.name</name>
   <value>yarn</value>
</property>
```

```
<property>

<name>mapreduce.application.classpath</name>
<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>

</property>

</configuration>
```

**Step 7e: Configure the YARN resource manager by editing the yarn-site.xml**

```
<configuration>

<property>

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

    <value>mapreduce_shuffle</value>

</property>

<property>

    <name>yarn.nodemanager.env-whitelist</name>

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>

</property>

</configuration>
```

**Step 8: Format the namenode using the command**

```
$ hdfs namenode -format
```

**Test HDFS configuration (/usr/local/hadoop/sbin/)**

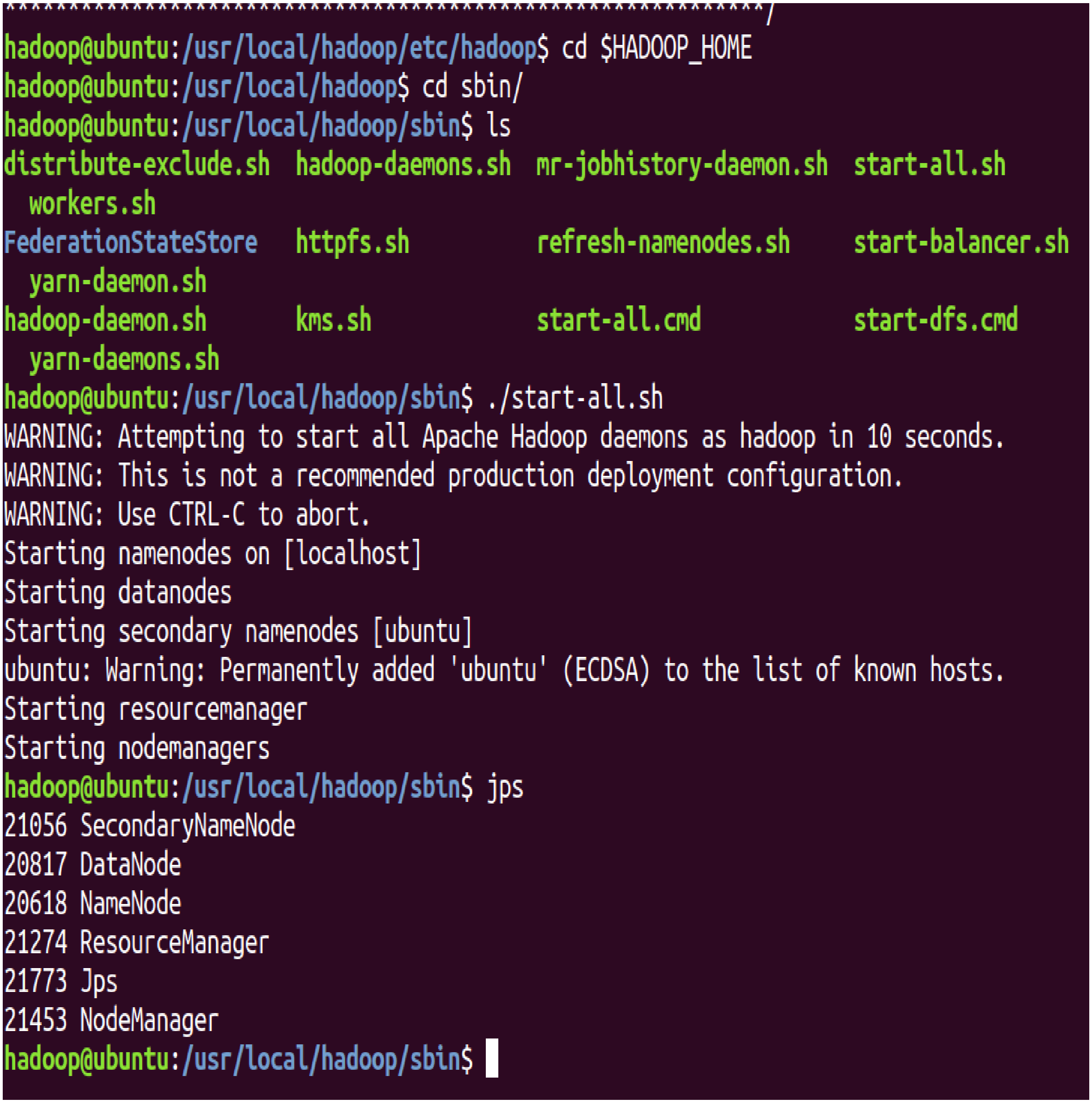
```
$ ./start-dfs.sh

$ ./start-yarn.sh

$./start-all.sh
```

**Check the availability of all the nodes by typing**

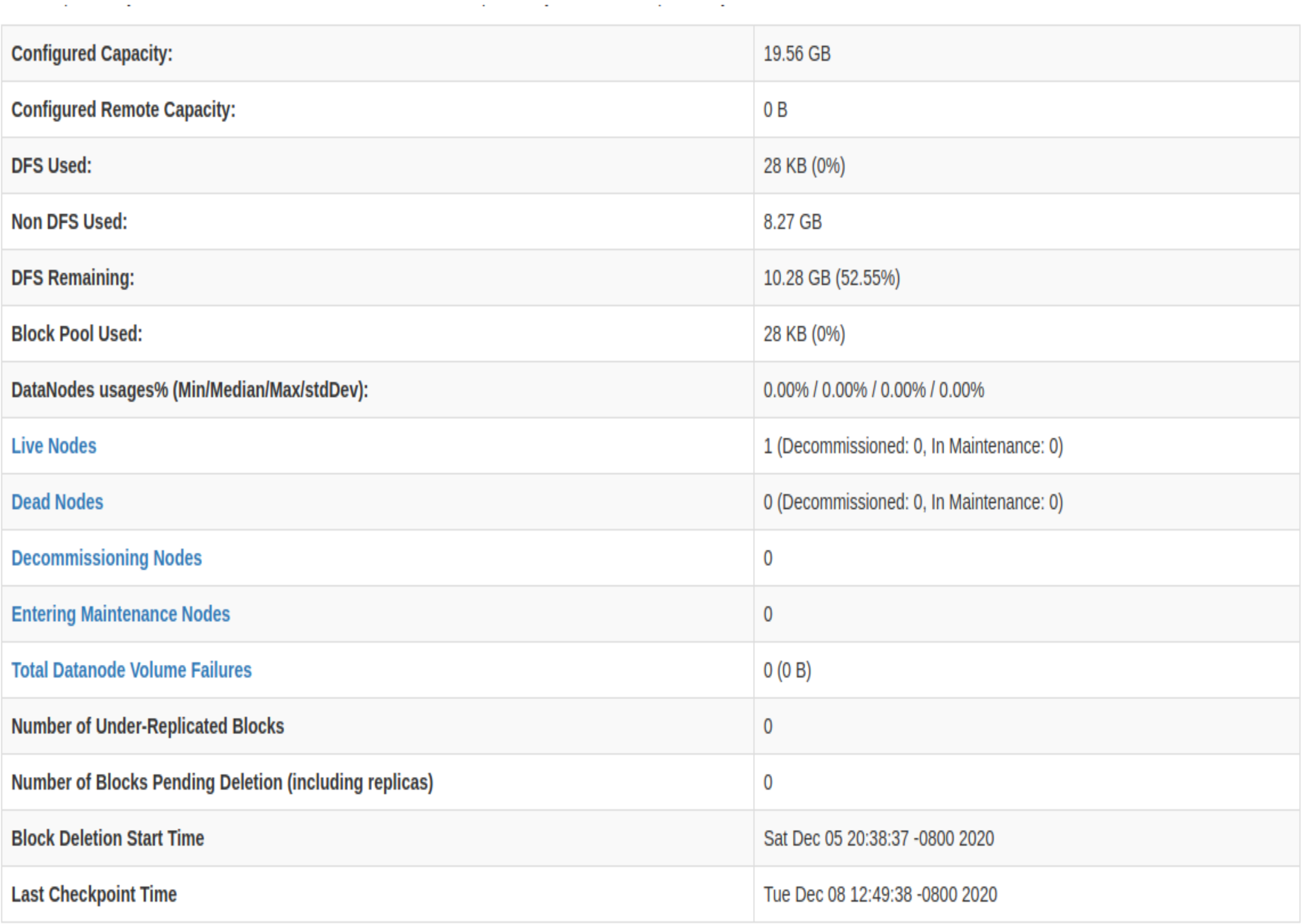
**$jps**



```
************************************************************/
hadoop@ubuntu:/usr/local/hadoop/etc/hadoop$ cd $HADOOP_HOME
hadoop@ubuntu:/usr/local/hadoop$ cd sbin/
hadoop@ubuntu:/usr/local/hadoop/sbin$ ls
distribute-exclude.sh  hadoop-daemons.sh  mr-jobhistory-daemon.sh  start-all.sh
  workers.sh
FederationStateStore   httpfs.sh          refresh-namenodes.sh     start-balancer.sh
  yarn-daemon.sh
hadoop-daemon.sh       kms.sh             start-all.cmd            start-dfs.cmd
  yarn-daemons.sh
hadoop@ubuntu:/usr/local/hadoop/sbin$ ./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]
Starting datanodes
Starting secondary namenodes [ubuntu]
ubuntu: Warning: Permanently added 'ubuntu' (ECDSA) to the list of known hosts.
Starting resourcemanager
Starting nodemanagers
hadoop@ubuntu:/usr/local/hadoop/sbin$ jps
21056 SecondaryNameNode
20817 DataNode
20618 NameNode
21274 ResourceManager
21773 Jps
21453 NodeManager
hadoop@ubuntu:/usr/local/hadoop/sbin$
```

**Step 9: Access the Web portal for hadoop management by typing in the following IP address in the browser**

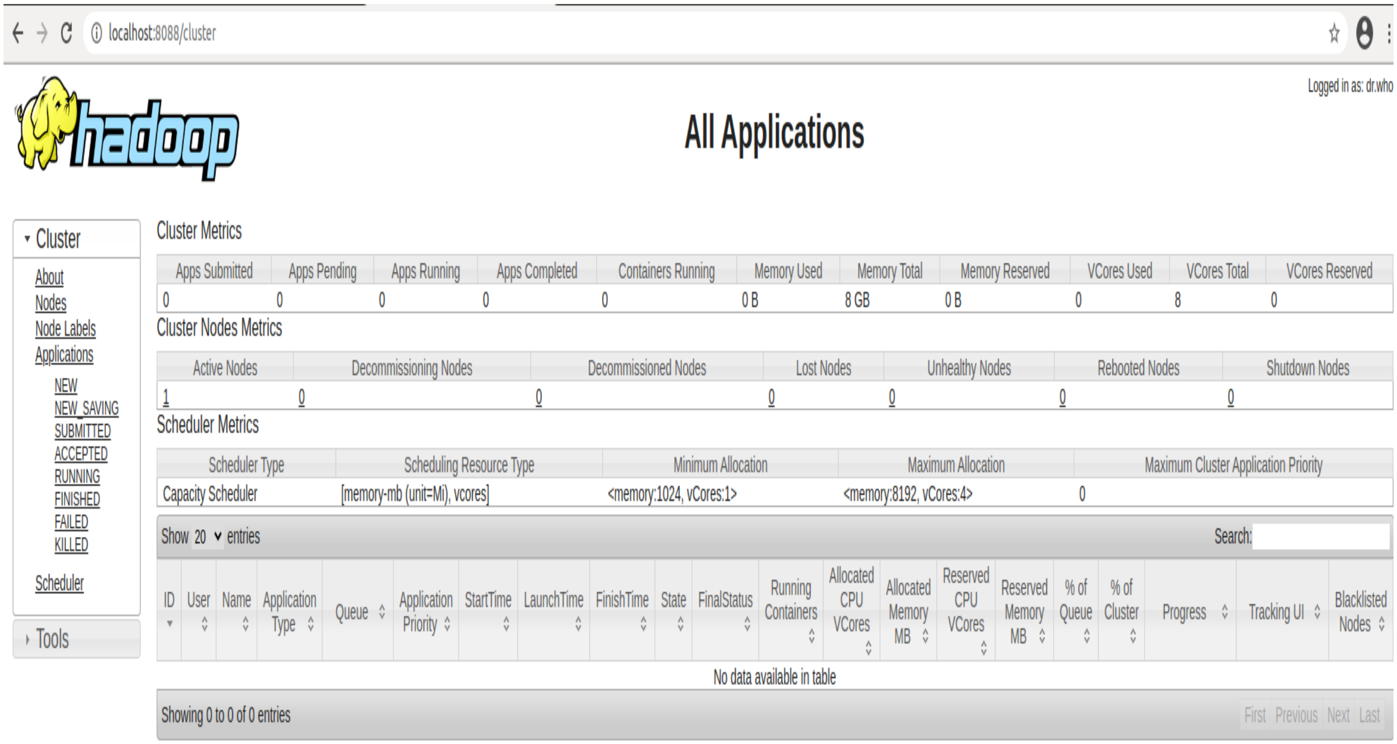
http://localhost:9870



| | |
|---|---|
| Configured Capacity: | 19.56 GB |
| Configured Remote Capacity: | 0 B |
| DFS Used: | 28 KB (0%) |
| Non DFS Used: | 8.27 GB |
| DFS Remaining: | 10.28 GB (52.55%) |
| Block Pool Used: | 28 KB (0%) |
| DataNodes usages% (Min/Median/Max/stdDev): | 0.00% / 0.00% / 0.00% / 0.00% |
| Live Nodes | 1 (Decommissioned: 0, In Maintenance: 0) |
| Dead Nodes | 0 (Decommissioned: 0, In Maintenance: 0) |
| Decommissioning Nodes | 0 |
| Entering Maintenance Nodes | 0 |
| Total Datanode Volume Failures | 0 (0 B) |
| Number of Under-Replicated Blocks | 0 |
| Number of Blocks Pending Deletion (including replicas) | 0 |
| Block Deletion Start Time | Sat Dec 05 20:38:37 -0800 2020 |
| Last Checkpoint Time | Tue Dec 08 12:49:38 -0800 2020 |

**Step 10: Check the hadoop cluster overview at**

http://localhost:8088



All Applications

Cluster Metrics

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory Reserved | VCores Used | VCores Total | VCores Reserved |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 B | 8 GB | 0 B | 0 | 8 | 0 |

Cluster Nodes Metrics

| Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes | Shutdown Nodes |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |

Scheduler Metrics

| Scheduler Type | Scheduling Resource Type | Minimum Allocation | Maximum Allocation | Maximum Cluster Application Priority |
|---|---|---|---|---|
| Capacity Scheduler | [memory-mb (unit=Mi), vcores] | <memory:1024, vCores:1> | <memory:8192, vCores:4> | 0 |

| ID | User | Name | Application Type | Queue | Application Priority | StartTime | LaunchTime | FinishTime | State | FinalStatus | Running Containers | Allocated CPU VCores | Allocated Memory MB | Reserved CPU VCores | Reserved Memory MB | % of Queue | % of Cluster | Progress | Tracking UI | Blacklisted Nodes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No data available in table | | | | | | | | | | | | | | | | | | | | |

Showing 0 to 0 of 0 entries

Execute $HADOOP_HOME/sbin - ./stop-all.sh