Large Scale Data Processing
CSE3025

Dr. Ramesh Ragala

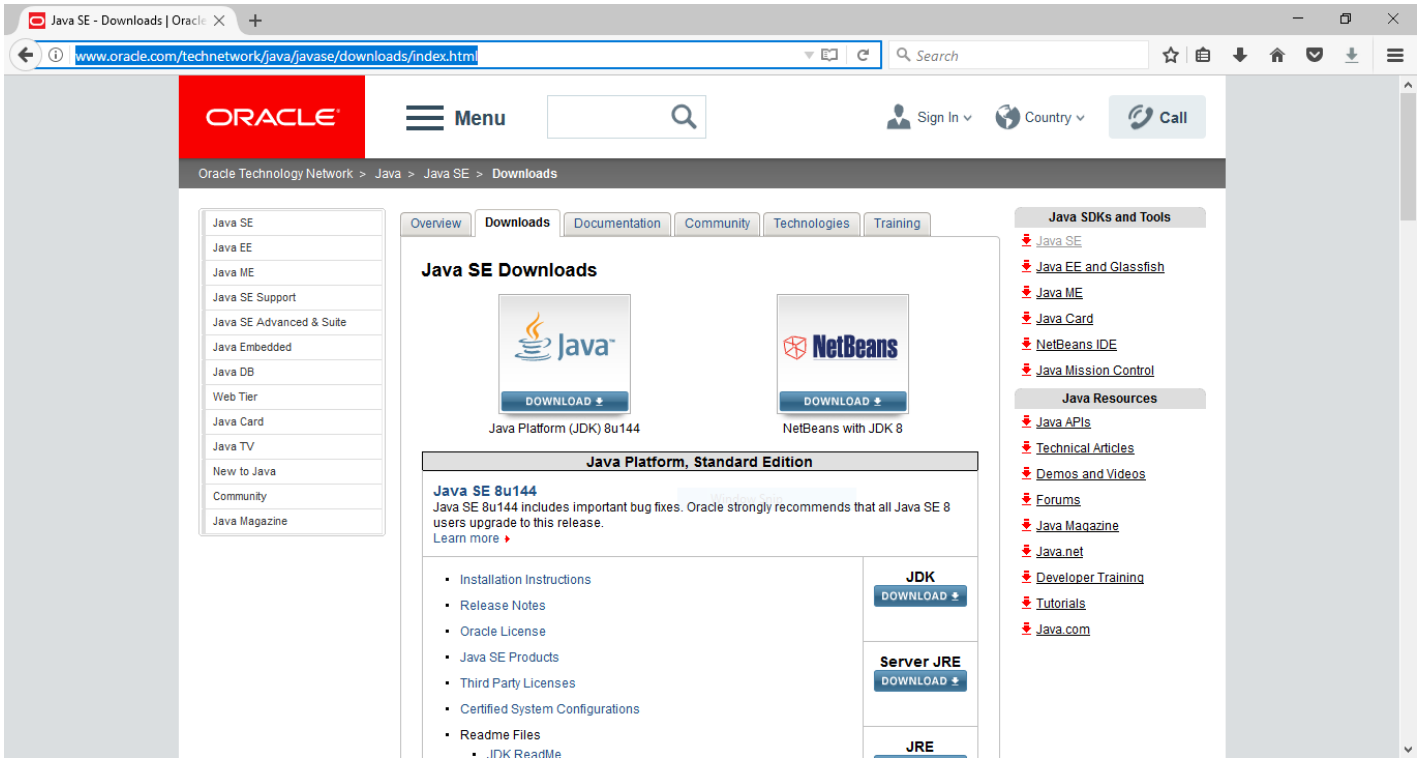
School of Computer Science and Engineering
VIT Chennai

March 2, 2021

1. Hadoop Installation

- **ssh**
  - **Check whether ssh is working or not**
  - $ **ssh localhost**
  - if the result like connection refused on port 22 then start **ssh** service
  - command to check ssh service status in ubuntu
    - $ sudo systemctl status ssh
    - Now, check ssh
    - $ **ssh localhost** → if it shows any error, then install ssh
    - command to install ssh in ubuntu
    - $ **sudo apt install openssh-server**
    - check the ssh in machine

- **java**
  - **Check java version on your local machine**
  - command is $ **java -version**
  - if the java version is **openjdk**, then install oracle JDK
  - **procedure to install oracle java**
  - Download JDK-11 from Oracle jdk website http://www.oracle.com/technetwork/java/javase/downloads/index.html



- create a directory **in** /opt directory
  - command is $ **sudo mkdir -p /opt/java**
- change the permission of the directory
  - command is $ **sudo chmod 777 -R java**
- copy the jdk-x.x.x.gz into /opt/java directory
  - command is $ **cp /home/ragalayathvisra/Downloads/jdk-x.x.x.gz /opt/java**
- untar the jdk file
  - command is $ **tar -xvzf jdk-x.x.x.gz**
- change the permission of java directory
  - command is $ **sudo chmod 777 -R /opt/java/jdk-x.x.x**

- set the path for java in /etc/profile
  - command is $ **sudo gedit /etc/profile**
- Append the following lines in /etc/profile
  - **export JAVA_HOME=/opt/java/jdk-x.x.x**
  - **export PATH=$JAVA_HOME/bin:$PATH**
- restart the terminal
  - command is $ **. /etc/profile**
- check java version
  - command is $ **java -version**
- path setting details are set in java-install.sh **(personal file)** → It have the symbolic links of java
- Execute java-install.sh file in terminal
  - command to change the permission $ **sudo chmod 777 -R java-install.sh**
  - command to execute the file $ **sudo ./java-install.sh**
  - Then check the java version. The command is $ **java -version**

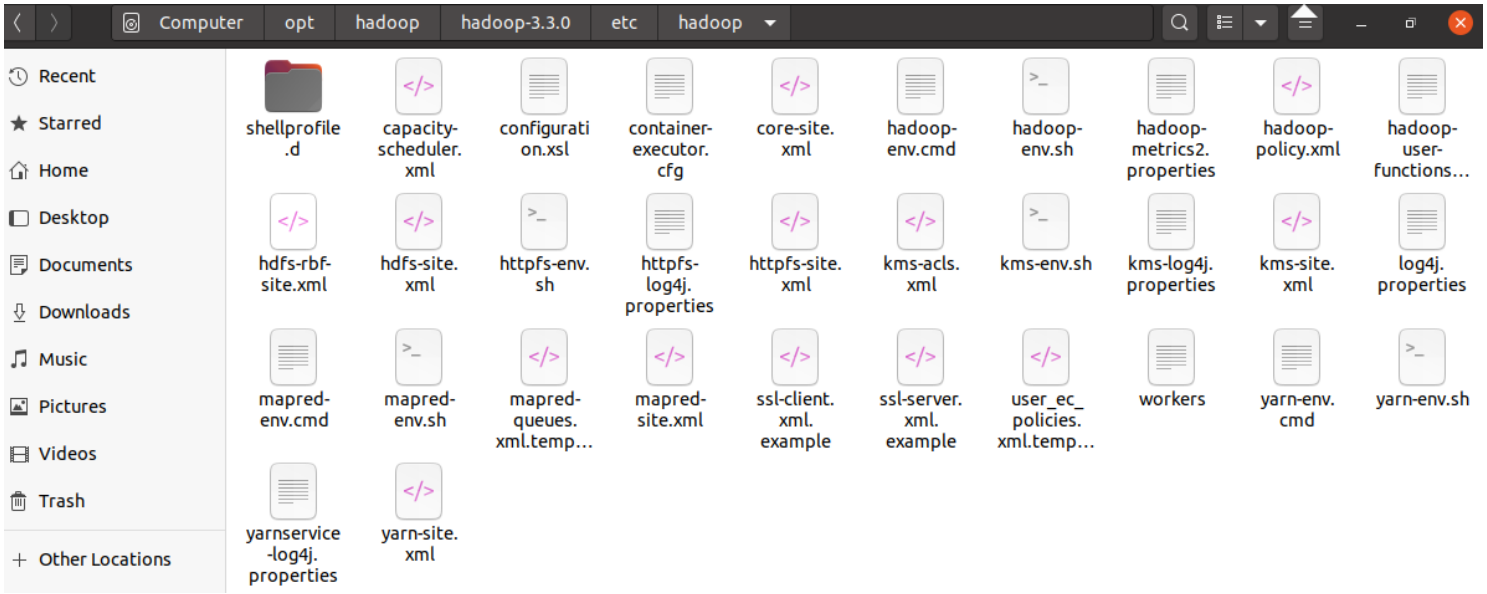
- Download latest hadoop (3.3.0) binary file from **http://hadoop.apache.org/releases.html**
- Let us assume it have downloaded in Download directory.
- create hadoop directory in /opt directory
  - The command is $ **sudo mkdir hadoop**
- Change the permissions of the directory
  - The command is $ **sudo chmod 777 -R hadoop**
- copy the hadoop-x.x.x.tar.gz into /opt/hadoop directory
  - The command is $ **cp /home/ragalayathvisra/Downloads/hadoop-x.x.x.tar.gz /opt/hadoop/**
- untar the file
  - command is $ **tar -xvzf hadoop-x.x.x.tar.gz**
- change the permission of **hadoop** directory
  - command is $ **sudo chmod 777 -R /opt/hadoop**

- open bashrc file for hadoop path setting
  - The command is $ **sudo gedit ~/.bashrc**
  - Append the Hadoop and java paths

```
#Hadoop Related Options
export HADOOP_HOME=/opt/hadoop/hadoop-3.3.0
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=$HADOOP_HOME
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin


#Java Related Options
export JAVA_HOME=/opt/java/jdk-11.0.9
export PATH=$PATH:$JAVA_HOME/bin
```

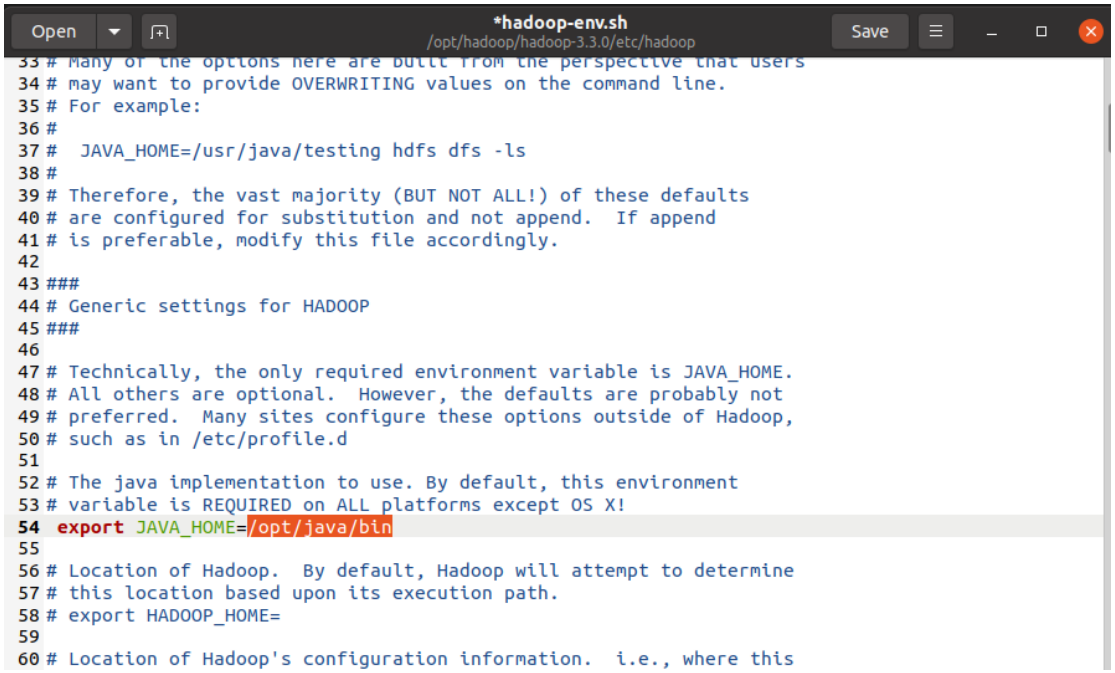
- Restart the terminal
  - command to restart is $**source ~/.bashrc**

- Go to hadoop directory for Configuration purpose. i.e /opt/hadoop/hadoop-3.3.0/etc/hadoop



- add java path in **hadoop_env.sh** file
- command to open hadoop_env.sh is $ **gedit hadoop_env.sh**

- add java path in **hadoop_env.sh** file
- command to open hadoop_env.sh is $ **gedit hadoop_env.sh** or you can directly open in notepad



- open **core-site.xml** for configuration purpose using notepad
- do the modification in core-site.xml as shown below, save and exit

Open core-site.xml /opt/hadoop/hadoop-3.3.0/etc/hadoop Save

```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

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

- open **hdfs-site.xml** for configuration purpose using notepad
- do the modification in hdfs-site.xml as shown below, save and exit

Open | *hdfs-site.xml /opt/hadoop/hadoop-3.3.0/etc/hadoop | Save

```
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at

   http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:///opt/hadoop/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:///opt/hadoop/hadoop_tmp/hdfs/datanode</value>
</property>
</configuration>
```

- open **mapred-site.xml** for configuration purpose using notepad
- do the modification in mapred-site.xml as shown below, save and exit

Open | mapred-site.xml /opt/hadoop/hadoop-3.3.0/etc/hadoop | Save

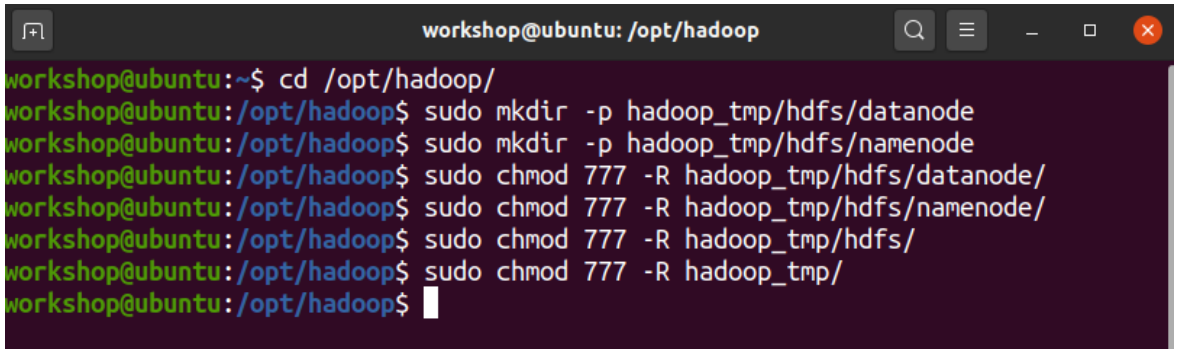
```xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

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

- open **yarn-site.xml** for configuration purpose using notepad
- do the modification in yarn-site.xml as shown below, save and exit

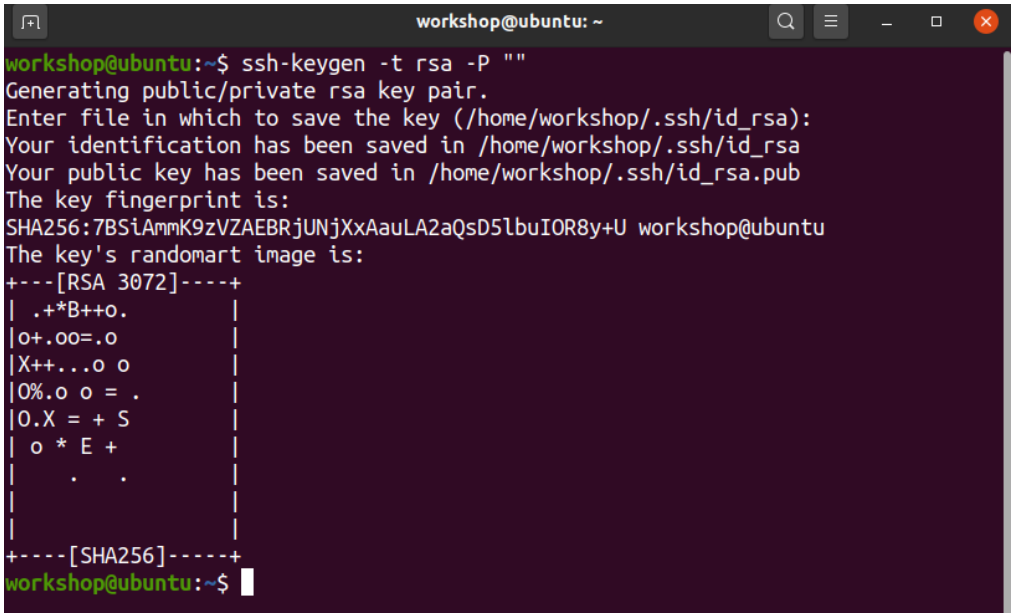
Open | *yarn-site.xml /opt/hadoop/hadoop-3.3.0/etc/hadoop | Save

```xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.vmem-check-enabled</name>
<value>false</value>
</property>
</configuration>
```

- we need to create directories for namenode and datanode, which are specified in hdfs-site.xml
- The following commands are for creating namenode, Datanode and permission settings



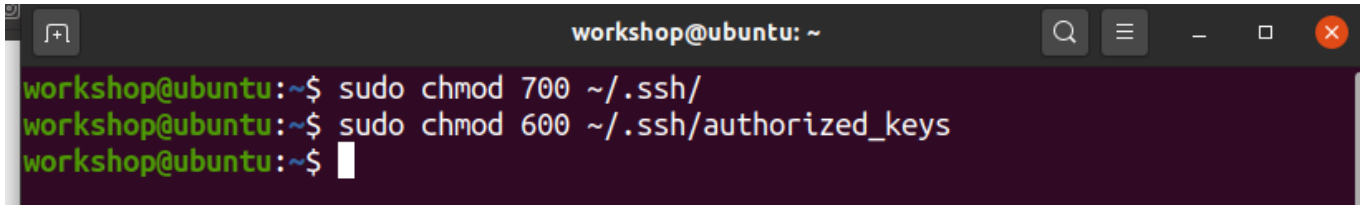
- Hadoop requires SSH access to manage its nodes, i.e. remote machines plus your local machine if you want to use Hadoop on it
- Commands is to create an RSA key pair with an empty password.



- we have to enable SSH access to local machine with this newly created key.
- The commands is

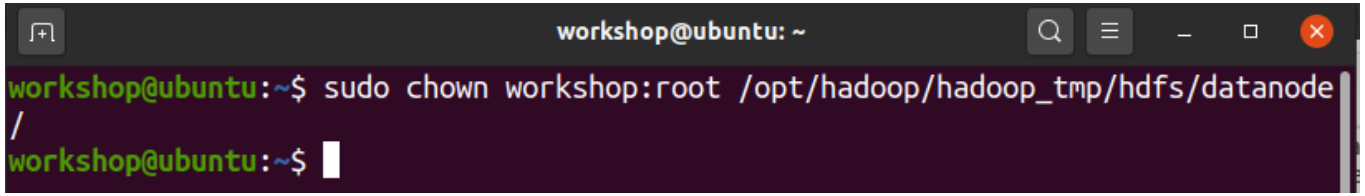
```
workshop@ubuntu: ~
workshop@ubuntu:~$ cat /home/workshop/.ssh/id_rsa.pub >> /home/workshop/.ssh/aut
horized_keys
workshop@ubuntu:~$
```

- command to passwordless promt
- $ **sudo chmod 700 ∽ ⁄.ssh⁄**
- $ **sudo chmod 600 ∽ ⁄.ssh⁄authorized_keys**



- We need to change the permissions of datanode
- The following command is used

sudo chown workshop:root /opt/Hadoop/hadoop_tmp/hdfs/datanode



- Now we need to format the namenode
- The following commands is used to format the namenode

vitchennai@localhost:~

File Edit View Search Terminal Help

```
[vitchennai@localhost ~]$ hdfs namenode -format
17/07/21 00:13:11 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   user = vitchennai
STARTUP_MSG:   host = localhost/127.0.0.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.8.0
STARTUP_MSG:   classpath = /opt/hadoop/hadoop-2.8.0/etc/hadoop:/opt/hadoop/hadoo
p-2.8.0/share/hadoop/common/lib/httpclient-4.5.2.jar:/opt/hadoop/hadoop-2.8.0/sh
are/hadoop/common/lib/jersey-core-1.9.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/
common/lib/mockito-all-1.8.5.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/li
b/guava-11.0.2.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/commons-math
3-3.1.1.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/jackson-mapper-asl-
1.9.13.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/api-util-1.0.0-M20.j
ar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/commons-io-2.4.jar:/opt/hado
op/hadoop-2.8.0/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/opt/hadoop/hadoop
-2.8.0/share/hadoop/common/lib/jetty-util-6.1.26.jar:/opt/hadoop/hadoop-2.8.0/sh
are/hadoop/common/lib/xmlenc-0.52.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/comm
on/lib/asm-3.2.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/stax-api-1.0
-2.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/jcip-annotations-1.0.jar
:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/gson-2.2.4.jar:/opt/hadoop/had
oop-2.8.0/share/hadoop/common/lib/java-xmlbuilder-0.4.jar:/opt/hadoop/hadoop-2.8
.0/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/opt/hadoop/hadoop-2.8.0/
```

- **Hadoop-3.3.0 installation has completed**
- **Check the Hadoop status**

- Now start the Hadoop Distributed File System
- The following commands (**start-dfs.sh**) is used for this

```
vitchennai@localhost:~
File  Edit  View  Search  Terminal  Help
SHUTDOWN_MSG: Shutting down NameNode at localhost/127.0.0.1
************************************************************/
[vitchennai@localhost ~]$ clear

[vitchennai@localhost ~]$ start-dfs.sh
Starting namenodes on [localhost]
The authenticity of host 'localhost (::1)' can't be established.
ECDSA key fingerprint is SHA256:T+LJOwDpd/6JnpGRbHJo0pru46r8l7kYDJM+tomWm3E.
ECDSA key fingerprint is MD5:bb:a4:e7:bb:13:1f:59:f1:18:41:9f:c8:60:93:5f:bd.
Are you sure you want to continue connecting (yes/no)? yes
localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
vitchennai@localhost's password:
localhost: starting namenode, logging to /opt/hadoop/hadoop-2.8.0/logs/hadoop-vitchenna
i-namenode-localhost.localdomain.out
vitchennai@localhost's password:
localhost: starting datanode, logging to /opt/hadoop/hadoop-2.8.0/logs/hadoop-vitchenna
i-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:T+LJOwDpd/6JnpGRbHJo0pru46r8l7kYDJM+tomWm3E.
ECDSA key fingerprint is MD5:bb:a4:e7:bb:13:1f:59:f1:18:41:9f:c8:60:93:5f:bd.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
vitchennai@0.0.0.0's password:

0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/hadoop-2.8.0/logs/hadoop-vi
tchennai-secondarynamenode-localhost.localdomain.out
[vitchennai@localhost ~]$
```

- Now check the background process for hadoop distributed file system

vitchennai@localhost:~

File Edit View Search Terminal Help

```
[vitchennai@localhost ~]$ jps
75251 Jps
75011 SecondaryNameNode
74783 DataNode
74590 NameNode
[vitchennai@localhost ~]$
```

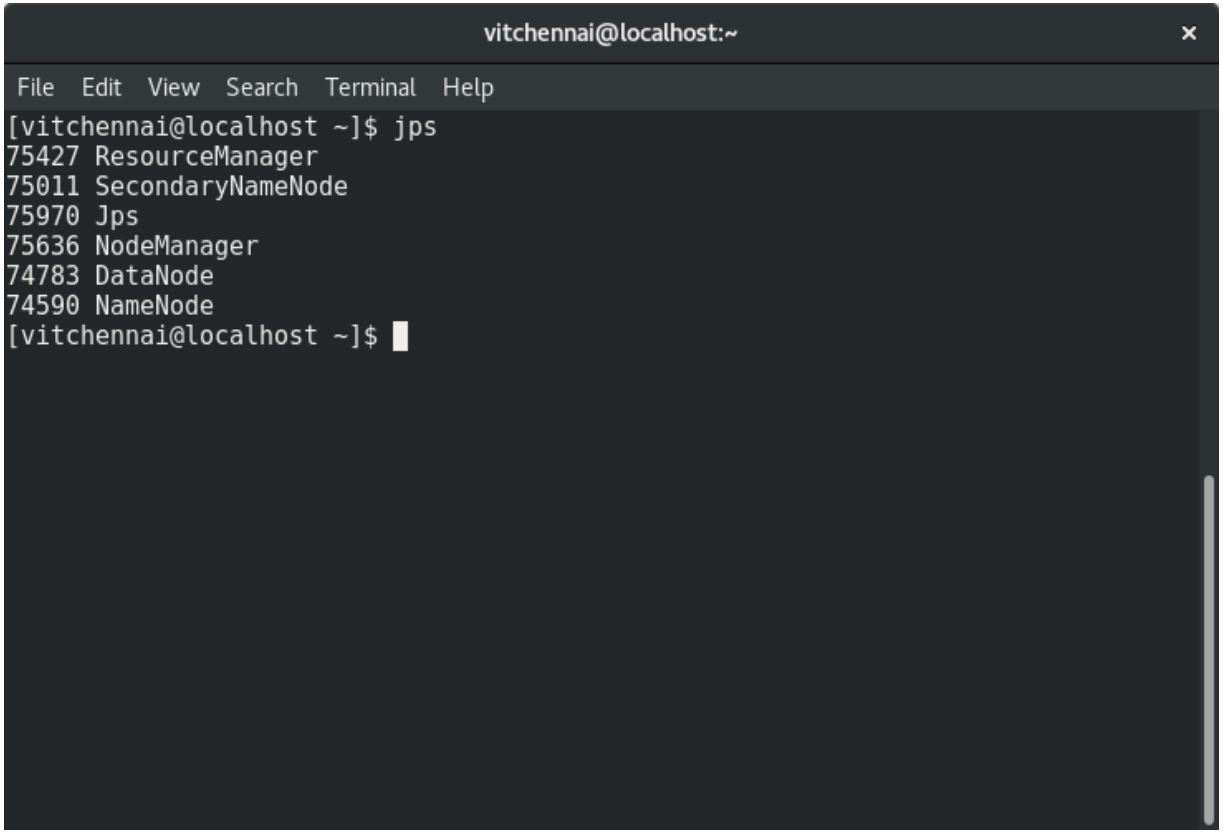
- Now start yarn resources for hadoop
- The following commands (**start-yarn.sh**) is used for this

vitchennai@localhost:~

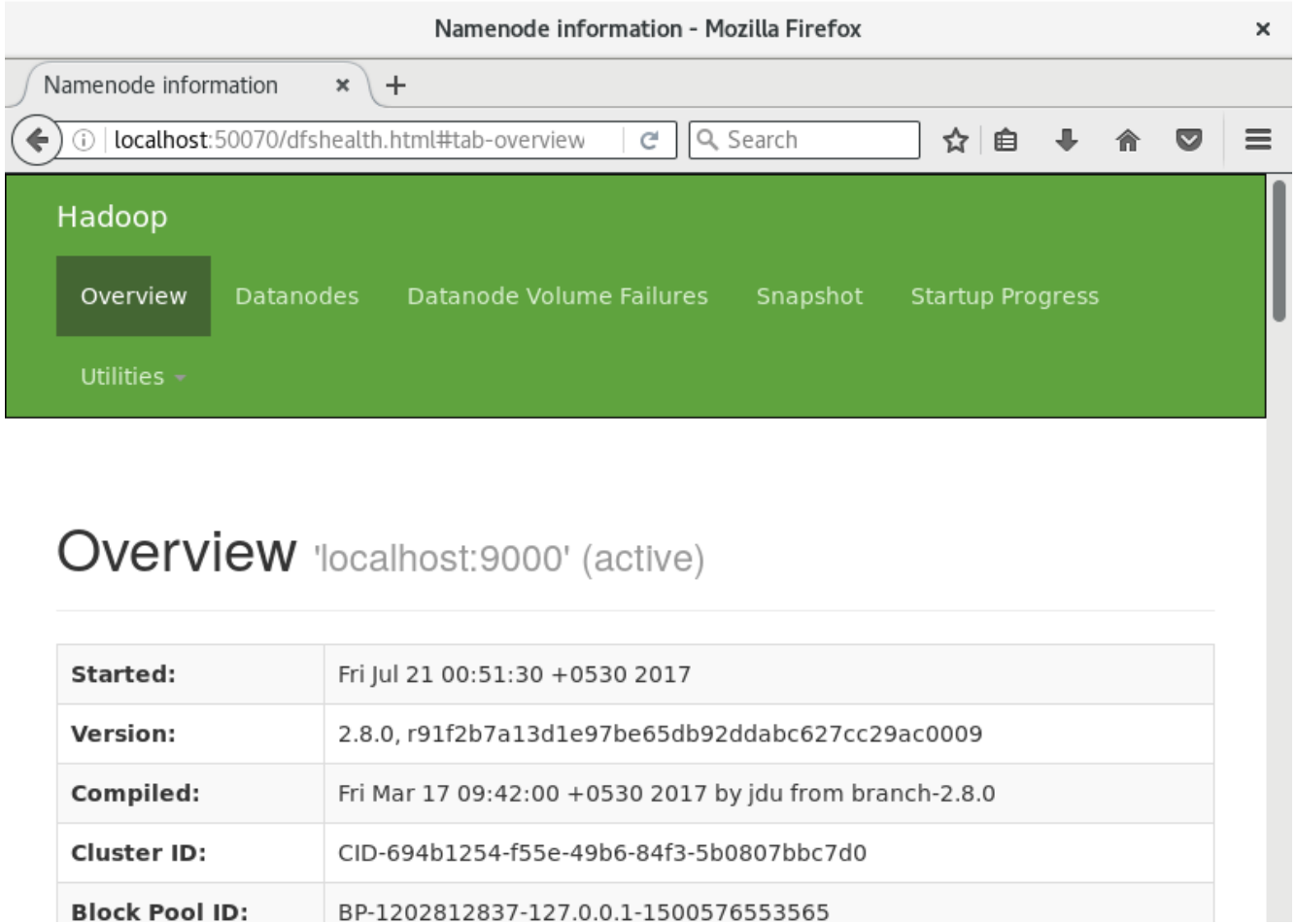
File Edit View Search Terminal Help

```
[vitchennai@localhost ~]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-vitchenn
ai-resourcemanager-localhost.localdomain.out
vitchennai@localhost's password:
localhost: starting nodemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-v
itchennai-nodemanager-localhost.localdomain.out
[vitchennai@localhost ~]$
```

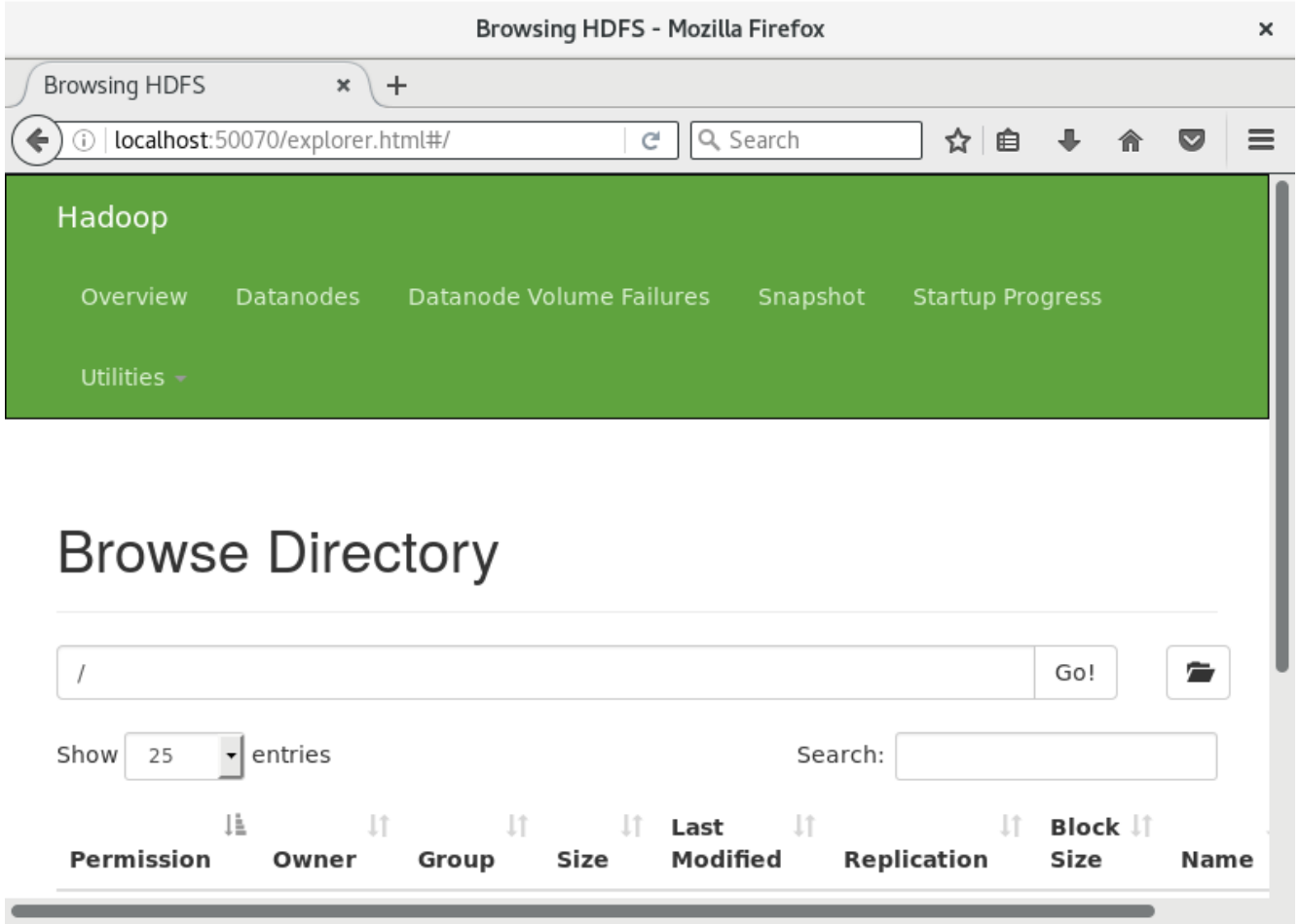
- The total number of daemons to execute hadoop-3.3.0 on local machine are
- **NameNode**
- **DataNode**
- **SecondaryNameNode**
- **NodeManager**
- **ResourceManager**
- we have to use **jps** command to check



- **UI view of Hadoop**
- Open **http://localhost:9870** in browser → Namenode



- Click on **Utilities** menu bar and then click on **Browse File System**



- The commands that are used to stop hadoop: $**stop-dfs.sh** and $**stop-yarn.sh**