

Large Scale Data Processing

CSE3025

Dr. Ramesh Ragala

School of Computer Science and Engineering
VIT Chennai

March 2, 2021



1 Hadoop Installation

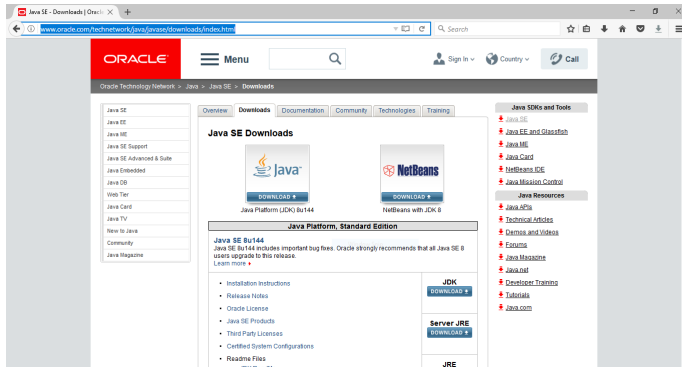
- **ssh**

- ▶ **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`

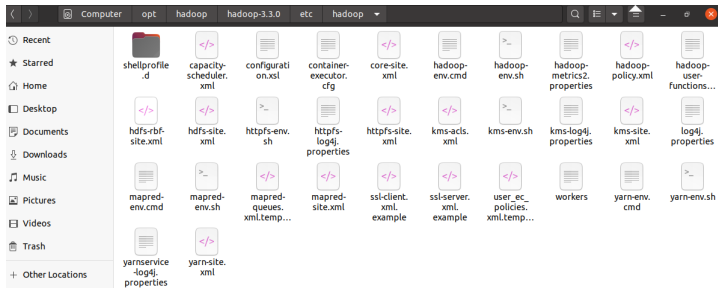
- Download latest hadoop (3.3.0) binary file from <http://hadoop.apache.org/releases.html>
- Let us assume it have downloaded in Download directory.
- create hadoop directory in /opt directory
 - ▶ The command is \$ **sudo mkdir hadoop**
- Change the permissions of the directory
 - ▶ The command is \$ **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

```
1 #Hadoop Related Options
2 export HADOOP_HOME=/opt/hadoop/hadoop-3.3.0
3 export HADOOP_INSTALL=$HADOOP_HOME
4 export HADOOP_MAPRED_HOME=$HADOOP_HOME
5 export HADOOP_COMMON_HOME=$HADOOP_HOME
6 export HADOOP_HDFS_HOME=$HADOOP_HOME
7 export YARN_HOME=$HADOOP_HOME
8 export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
9 export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
10 export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
11
12
13 #Java Related Options
14 export JAVA_HOME=/opt/java/jdk-11.0.9
15 export PATH=$PATH:$JAVA_HOME/bin
```

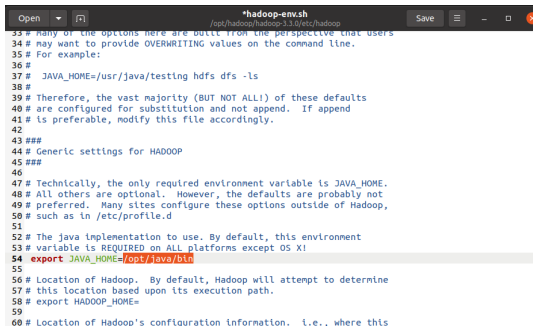
- Restart the terminal
 - ▶ command to restart is `$source ~/.bashrc`

- 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`

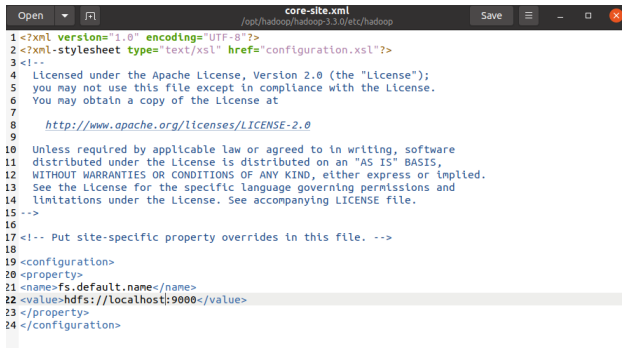
- add java path in **hadoop_env.sh** file
- command to open **hadoop_env.sh** is **\$ gedit hadoop_env.sh** or you can directly open in notepad



```
*hadoop-env.sh
/opt/hadoop/hadoop-3.3.0/etc/hadoop

33 # Many of the options here are built from the perspective that users
34 # may want to provide OVERWRITING values on the command line.
35 # For example:
36 #
37 #   JAVA_HOME=/usr/java/testing hdfs dfs -ls
38 #
39 # Therefore, the vast majority (BUT NOT ALL!) of these defaults
40 # are configured for substitution and not append. If append
41 # is preferable, modify this file accordingly.
42
43 ###
44 # Generic settings for HADOOP
45 ###
46
47 # Technically, the only required environment variable is JAVA_HOME.
48 # All others are optional. However, the defaults are probably not
49 # preferred. Many sites configure these options outside of Hadoop,
50 # such as in /etc/profile.d
51
52 # The java implementation to use. By default, this environment
53 # variable is REQUIRED on ALL platforms except OS X!
54 export JAVA_HOME=/opt/java/bin
55
56 # Location of Hadoop. By default, Hadoop will attempt to determine
57 # this location based upon its execution path.
58 export HADOOP_HOME=
59
60 # Location of Hadoop's configuration information. i.e., where this
```

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



```
core-site.xml
/opt/hadoop/hadoop-3.3.0/etc/hadoop

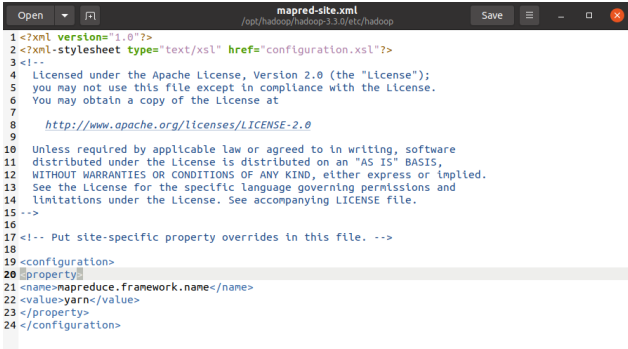
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20 <property>
21 <name>fs.default.name</name>
22 <value>hdfs://localhost:9000</value>
23 </property>
24 </configuration>
```

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



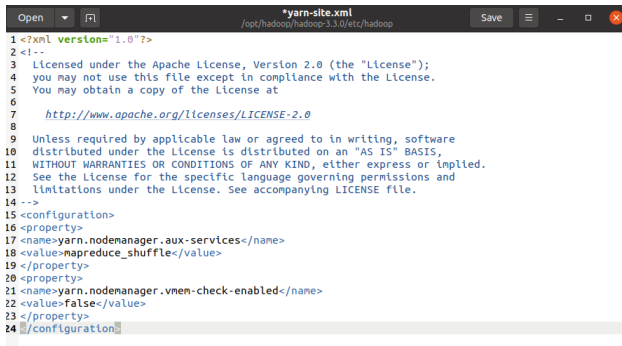
```
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8   http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20 <property>
21 <name>dfs.replication</name>
22 <value>1</value>
23 </property>
24 <property>
25 <name>dfs.namenode.name.dir</name>
26 <value>file:///opt/hadoop/hadoop_tmp/hdfs/namenode</value>
27 </property>
28 <property>
29 <name>dfs.datanode.data.dir</name>
30 <value>file:///opt/hadoop/hadoop_tmp/hdfs/datanode</value>
31 </property>
32 </configuration>
```

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



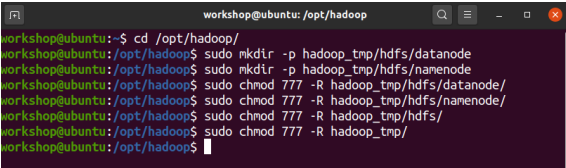
```
1 <?xml version="1.0"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4 Licensed under the Apache License, Version 2.0 (the "License");
5 you may not use this file except in compliance with the License.
6 You may obtain a copy of the License at
7
8 http://www.apache.org/licenses/LICENSE-2.0
9
10 Unless required by applicable law or agreed to in writing, software
11 distributed under the License is distributed on an "AS IS" BASIS,
12 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13 See the License for the specific language governing permissions and
14 limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20 <property>
21 <name>mapreduce.framework.name</name>
22 <value>yarn</value>
23 </property>
24 </configuration>
```

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



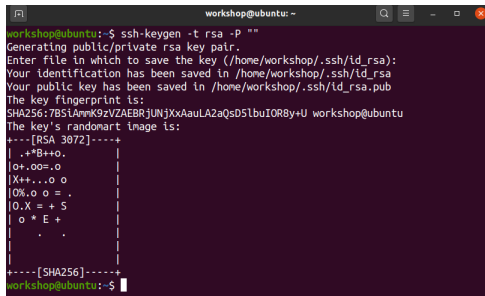
```
1 <?xml version="1.0"?>
2 <!--
3 Licensed under the Apache License, Version 2.0 (the "License");
4 you may not use this file except in compliance with the License.
5 You may obtain a copy of the License at
6
7 http://www.apache.org/licenses/LICENSE-2.0
8
9 Unless required by applicable law or agreed to in writing, software
10 distributed under the License is distributed on an "AS IS" BASIS,
11 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12 See the License for the specific language governing permissions and
13 limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16 <property>
17 <name>yarn.nodemanager.aux-services</name>
18 <value>mapreduce_shuffle</value>
19 </property>
20 <property>
21 <name>yarn.nodemanager.vmen-check-enabled</name>
22 <value>false</value>
23 </property>
24 </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

A terminal window titled 'workshop@ubuntu: /opt/hadoop' with search, menu, and window control icons. It displays a series of commands to create Hadoop directories and set permissions. The commands are: 'cd /opt/hadoop/', 'sudo mkdir -p hadoop_tmp/hdfs/datanode', 'sudo mkdir -p hadoop_tmp/hdfs/namenode', 'sudo chmod 777 -R hadoop_tmp/hdfs/datanode/', 'sudo chmod 777 -R hadoop_tmp/hdfs/namenode/', 'sudo chmod 777 -R hadoop_tmp/hdfs/', and 'sudo chmod 777 -R hadoop_tmp/'. The prompt returns to 'workshop@ubuntu: /opt/hadoop\$' after the last command.

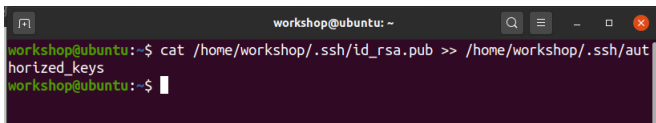
```
workshop@ubuntu: /opt/hadoop$ cd /opt/hadoop/
workshop@ubuntu: /opt/hadoop$ sudo mkdir -p hadoop_tmp/hdfs/datanode
workshop@ubuntu: /opt/hadoop$ sudo mkdir -p hadoop_tmp/hdfs/namenode
workshop@ubuntu: /opt/hadoop$ sudo chmod 777 -R hadoop_tmp/hdfs/datanode/
workshop@ubuntu: /opt/hadoop$ sudo chmod 777 -R hadoop_tmp/hdfs/namenode/
workshop@ubuntu: /opt/hadoop$ sudo chmod 777 -R hadoop_tmp/hdfs/
workshop@ubuntu: /opt/hadoop$ sudo chmod 777 -R hadoop_tmp/
workshop@ubuntu: /opt/hadoop$
```

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



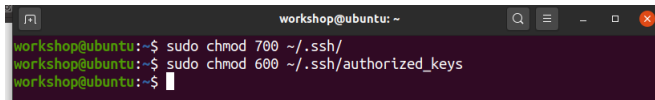
```
workshop@ubuntu: ~  
workshop@ubuntu:~$ ssh-keygen -t rsa -P ""  
Generating public/private rsa key pair.  
Enter file in which to save the key (/home/workshop/.ssh/id_rsa):  
Your identification has been saved in /home/workshop/.ssh/id_rsa  
Your public key has been saved in /home/workshop/.ssh/id_rsa.pub  
The key fingerprint is:  
SHA256:7B5iAmmK9zVZAEBrjUNjXxAauLA2aQsD5lbuI0R8y+U workshop@ubuntu  
The key's randomart image is:  
+---[RSA 3072]-----+  
| .+*B++o. |  
|o+.oo=.o |  
|X+...o o |  
|0%.o o = . |  
|0.X = + S |  
| o * E + |  
| . . |  
+-----[SHA256]-----+  
workshop@ubuntu:~$
```


- 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/authorized_keys  
workshop@ubuntu:~$
```

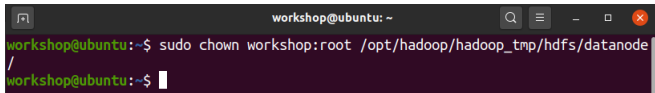
- command to passwordless prompt
- **\$ sudo chmod 700 ~/.ssh/**
- **\$ sudo chmod 600 ~/.ssh/authorized_keys**



```
workshop@ubuntu: ~  
workshop@ubuntu:~$ sudo chmod 700 ~/.ssh/  
workshop@ubuntu:~$ sudo chmod 600 ~/.ssh/authorized_keys  
workshop@ubuntu:~$
```

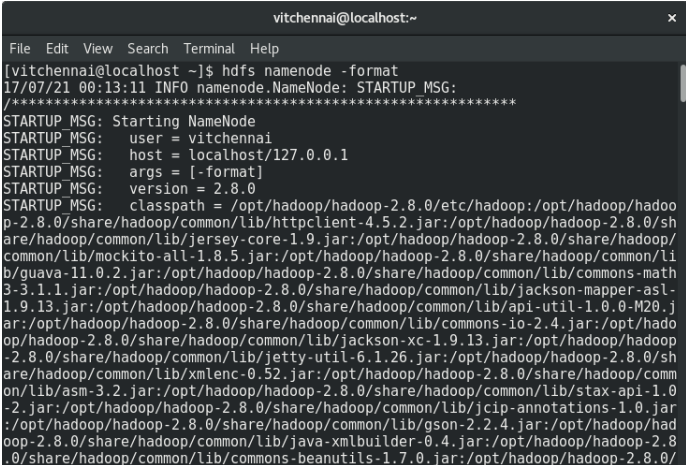
- We need to change the permissions of datanode
- The following command is used

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

A screenshot of a terminal window titled 'workshop@ubuntu: ~'. The terminal shows the command 'sudo chown workshop:root /opt/hadoop/hadoop_tmp/hdfs/datanode /' being entered and executed. The prompt 'workshop@ubuntu:~\$' is visible before and after the command. The output is a single slash '/' on the next line.

```
workshop@ubuntu: ~  
workshop@ubuntu:~$ sudo chown workshop:root /opt/hadoop/hadoop_tmp/hdfs/datanode  
/  
workshop@ubuntu:~$
```

- 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/hadoop-2.8.0/share/hadoop/common/lib/httpclient-4.5.2.jar:/opt/hadoop/hadoop-2.8.0/share/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/lib/guava-11.0.2.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/commons-math3-3.1.1.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.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/lib/commons-io-2.4.jar:/opt/hadoop/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/share/hadoop/common/lib/xmlenc-0.52.jar:/opt/hadoop/hadoop-2.8.0/share/hadoop/common/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/hadoop-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

```

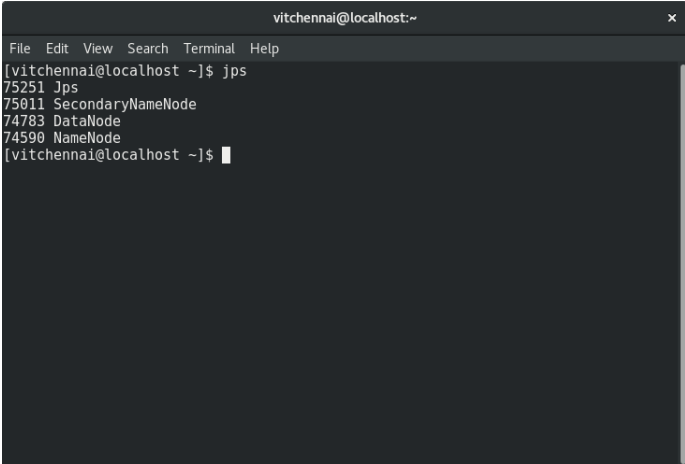
vitchennai@localhost:~
File Edit View Search Terminal Help
SHUTDOWN MSG: Shutting down NameNode at localhost/127.0.0.1
*****/
[vitchennai@localhost ~]$ clear

[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+LJ0wDpd/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+LJ0wDpd/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 ~]$ █

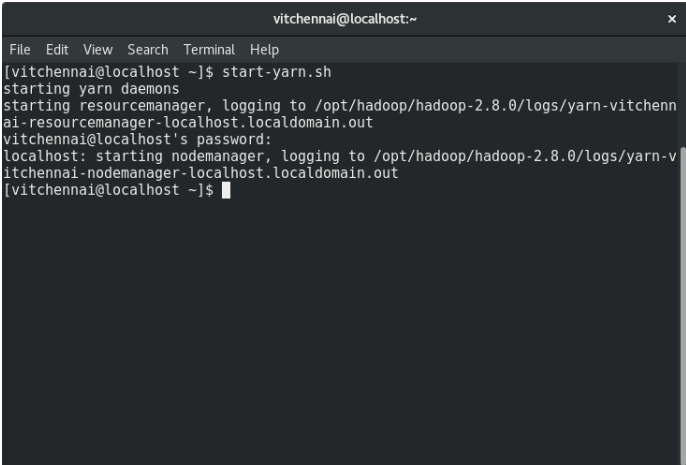
```

- Now check the background process for hadoop distributed file system

A terminal window titled 'vitchennai@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The command '[vitchennai@localhost ~]\$ jps' has been executed, resulting in the following output:

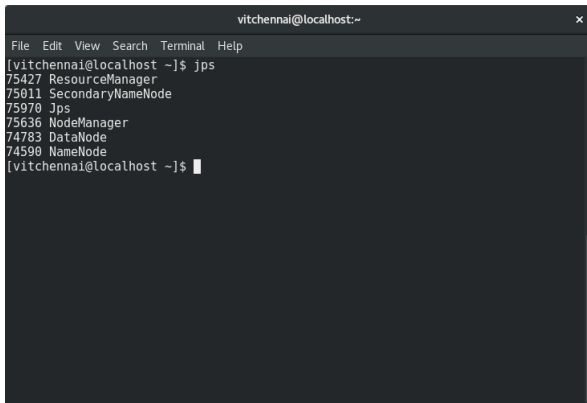
```
[vitchennai@localhost ~]$ jps
75251 Jps
75011 SecondaryNameNode
74783 DataNode
74590 NameNode
[vitchennai@localhost ~]$
```

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

A terminal window titled 'vitchennai@localhost:~' with a standard menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the command '[vitchennai@localhost ~]\$ start-yarn.sh' being executed. The output is: 'starting yarn daemons', 'starting resourcemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-vitchennai-resourcemanager-localhost.localdomain.out', 'vitchennai@localhost's password:', 'localhost: starting nodemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-vitchennai-nodemanager-localhost.localdomain.out', and finally '[vitchennai@localhost ~]\$' with a cursor.

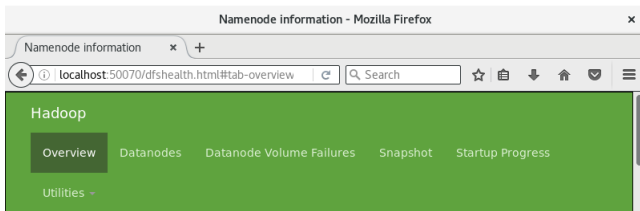
```
vitchennai@localhost:~
File Edit View Search Terminal Help
[vitchennai@localhost ~]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-vitchennai-resourcemanager-localhost.localdomain.out
vitchennai@localhost's password:
localhost: starting nodemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-vitchennai-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

A terminal window titled 'vitchennai@localhost:~' with a standard menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the command '[vitchennai@localhost ~]\$ jps' and its output: '75427 ResourceManager', '75011 SecondaryNameNode', '75970 Jps', '75636 NodeManager', '74783 DataNode', and '74590 NameNode'. The prompt '[vitchennai@localhost ~]\$' is followed by a cursor.

```
vitchennai@localhost:~
File Edit View Search Terminal Help
[vitchennai@localhost ~]$ jps
75427 ResourceManager
75011 SecondaryNameNode
75970 Jps
75636 NodeManager
74783 DataNode
74590 NameNode
[vitchennai@localhost ~]$
```

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



Overview 'localhost:9000' (active)

Started:	Fri Jul 21 00:51:30 +0530 2017
Version:	2.8.0, r91f2b7a13d1e97be65db92ddabc627cc29ac0009
Compiled:	Fri Mar 17 09:42:00 +0530 2017 by jdu from branch-2.8.0
Cluster ID:	CID-694b1254-f55e-49b6-84f3-5b0807bbc7d0
Block Pool ID:	BP-1202812837-127.0.0.1-1500576553565

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

Browsing HDFS - Mozilla Firefox

Browsing HDFS

localhost:50070/explorer.html#/

Hadoop

Overview Datanodes Datanode Volume Failures Snapshot Startup Progress

Utilities

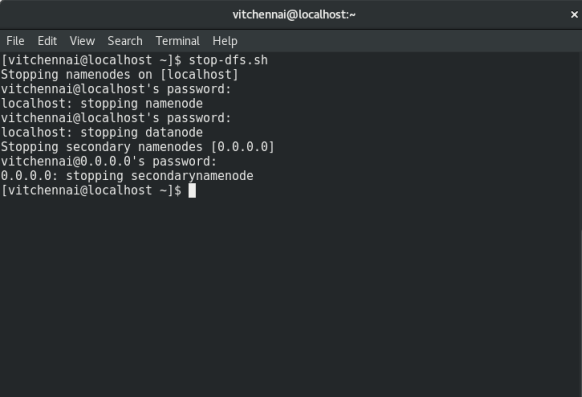
Browse Directory

/ Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
------------	-------	-------	------	---------------	-------------	------------	------

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

A terminal window titled 'vitchennai@localhost:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the execution of 'stop-dfs.sh'. It prompts for a password, then shows 'localhost: stopping namenode', 'localhost: stopping datanode', and 'Stopping secondary namenodes [0.0.0.0]'. It then prompts for a password for '0.0.0.0', followed by '0.0.0.0: stopping secondarynamenode'. The prompt returns to '[vitchennai@localhost ~]\$' with a cursor.

```
vitchennai@localhost:~  
File Edit View Search Terminal Help  
[vitchennai@localhost ~]$ stop-dfs.sh  
Stopping namenodes on [localhost]  
vitchennai@localhost's password:  
localhost: stopping namenode  
vitchennai@localhost's password:  
localhost: stopping datanode  
Stopping secondary namenodes [0.0.0.0]  
vitchennai@0.0.0.0's password:  
0.0.0.0: stopping secondarynamenode  
[vitchennai@localhost ~]$
```