

BIG DATA FRAMEWORKS

CSE6701

Prof. Ramesh Ragala

July 9, 2019

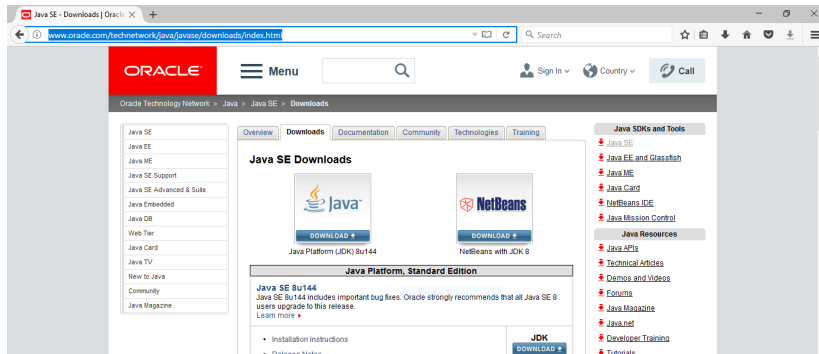
- **ssh**

- **Check whether ssh is working or not**
- **\$ ssh localhost**
- if the result like **connection refused on port 22** then start **sshd** service
- command to start sshd service in fedora
 - **\$ sudo systemctl start sshd** → it asks password
 - check ssh
 - **\$ ssh localhost** → if it shows any error, then install ssh
 - command to install ssh
 - **\$ sudo dnf install openssh-server***
 - check the ssh in machine

• java

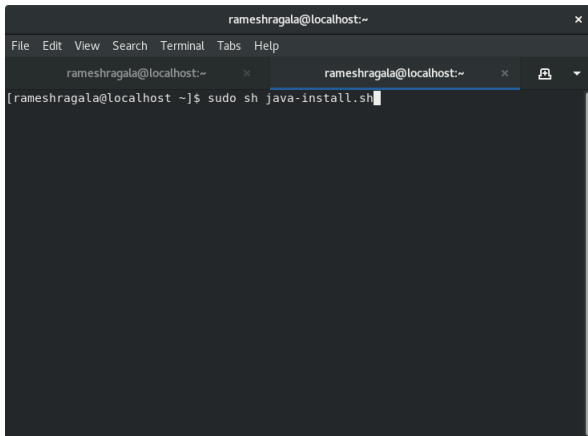
- Check java version on your local machine
- command is \$ **java -version**
- if the java version is **openjdk**, then install oracle JDK
- **procedure to install oracle java**
- Download JDK from Oracle jdk website

<http://www.oracle.com/technetwork/java/javase/downloads/index.html>



- create a directory **in** /opt directory
 - command is \$ **sudo mkdir -p /opt/java8**
- change the permission of the directory
 - command is \$ **sudo chmod 777 -R java**
- copy the jdk-x.x.x.gz into /opt/java8 directory
 - command is \$ **cp /home/vitchennai/Downloads/jdk-x.x.x.gz /opt/java8**
- 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/java8/jdk-x.x.x**

- path setting details are set in java-install.sh(personal file)
- Execute java-install.sh file in terminal



A screenshot of a terminal window titled "rameshragala@localhost:~". The window has a menu bar with "File", "Edit", "View", "Search", "Terminal", "Tabs", and "Help". Below the menu bar, there are two tabs, both labeled "rameshragala@localhost:~". The active tab shows a terminal prompt "[rameshragala@localhost ~]\$ sudo sh java-install.sh" with a cursor at the end of the command. The terminal background is dark, and the text is light gray.

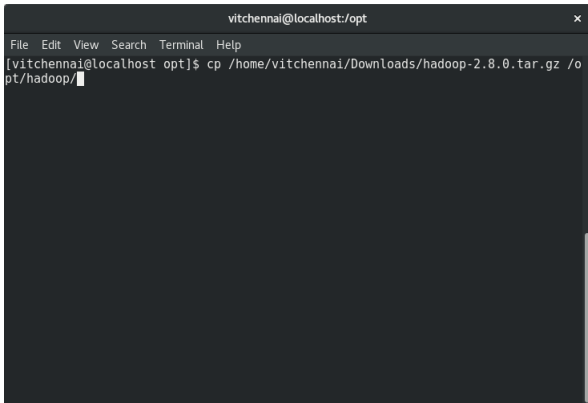
- Download latest hadoop binary file from **<http://hadoop.apache.org/releases.html>**
- Let us assume it have downloaded in Download directory

```
vitchennai@localhost:~/Downloads x
File Edit View Search Terminal Help
[vitchennai@localhost ~]$ cd /home/vitchennai/Downloads/
[vitchennai@localhost Downloads]$ ls
hadoop-2.8.0.tar.gz      pycharm-community-2017.1.4.tar.gz
jdk-8u131-linux-x64.tar.gz
[vitchennai@localhost Downloads]$
```

- create hadoop directory in /opt directory

```
vitchennai@localhost:/opt/
File Edit View Search Terminal Help
[vitchennai@localhost Downloads]$ cd /opt/
[vitchennai@localhost opt]$ ls
JAVA8
[vitchennai@localhost opt]$ sudo mkdir hadoop
[sudo] password for vitchennai:
[vitchennai@localhost opt]$ ls
hadoop JAVA8
[vitchennai@localhost opt]$ sudo chmod 777 -R hadoop/
[vitchennai@localhost opt]$ ls
hadoop JAVA8
[vitchennai@localhost opt]$
```

- change the permission of this directory
 - command is \$ **sudo chmod 777 -R /opt/hadoop**
- copy the hadoop-2.8.0.tar.gz into /opt/hadoop directory
 - command is \$ **cp**
/home/vitchennai/Downloads/hadoop-2.8.0.tar.gz
/opt/hadoop/

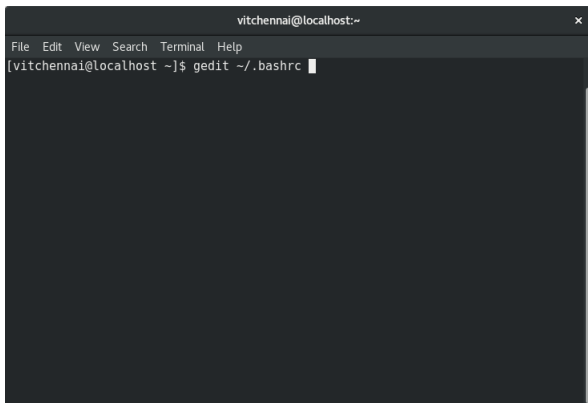


```
vitchennai@localhost:/opt
File Edit View Search Terminal Help
[vitchennai@localhost opt]$ cp /home/vitchennai/Downloads/hadoop-2.8.0.tar.gz /opt/hadoop/
```


- untar the file
 - command is \$ **tar -xvzf hadoop-2.8.0.tar.gz**
- change the permission of **hadoop** directory
 - command is \$ **sudo chmod 777 -R /opt/hadoop**

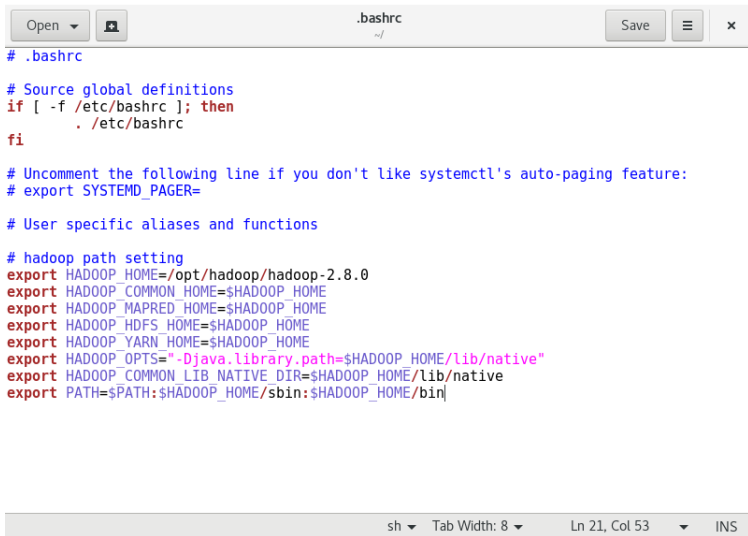
```
vitchennai@localhost:/opt/hadoop
File Edit View Search Terminal Help
[vitchennai@localhost hadoop]$ ls
hadoop-2.8.0  hadoop-2.8.0.tar.gz
[vitchennai@localhost hadoop]$ cd ..
[vitchennai@localhost opt]$ sudo chmod 777 -R hadoop/
[sudo] password for vitchennai:
Sorry, try again.
[sudo] password for vitchennai:
[vitchennai@localhost opt]$ ls
hadoop  JAVA8
[vitchennai@localhost opt]$ cd hadoop/
[vitchennai@localhost hadoop]$ ls
hadoop-2.8.0  hadoop-2.8.0.tar.gz
[vitchennai@localhost hadoop]$
```

- open bashrc file for hadoop path setting
 - command is \$ **gedit ~/.bashrc**



```
vitchennai@localhost:~  
File Edit View Search Terminal Help  
[vitchennai@localhost ~]$ gedit ~/.bashrc
```

- Append the Hadoop and java paths



```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

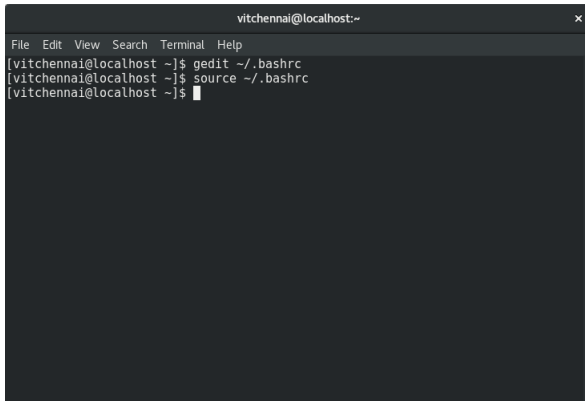
# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions

# hadoop path setting
export HADOOP_HOME=/opt/hadoop/hadoop-2.8.0
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$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
```

sh ▾ Tab Width: 8 ▾ Ln 21, Col 53 ▾ INS

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



```
vitchennai@localhost:~  
File Edit View Search Terminal Help  
[vitchennai@localhost ~]$ gedit ~/.bashrc  
[vitchennai@localhost ~]$ source ~/.bashrc  
[vitchennai@localhost ~]$
```

- 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.

```
vitchennai@localhost:~  
File Edit View Search Terminal Help  
[vitchennai@localhost ~]$ ssh-keygen -t rsa -P ""  
Generating public/private rsa key pair.  
Enter file in which to save the key (/home/vitchennai/.ssh/id_rsa):  
Your identification has been saved in /home/vitchennai/.ssh/id_rsa.  
Your public key has been saved in /home/vitchennai/.ssh/id_rsa.pub.  
The key fingerprint is:  
SHA256:BP+LYo20Y0mtRvZ6e9T7k73TxqwIQ45kkkbxMCbPmA vitchennai@localhost.localdom  
ain  
The key's randomart image is:  
+---[RSA 2048]-----+  
| ..0. . |  
| ... o |  
| oo o |  
|.Eo. = o . |  
|.o = + S + |  
| oo o +.o = |  
| +oo. o* + . . |  
|.o+B .+.o . + |  
|.ooo.B++o o.. |  
+----[SHA256]-----+  
[vitchennai@localhost ~]$
```

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

```
vit Chennai@localhost:~  
File Edit View Search Terminal Help  
[vit Chennai@localhost ~]$ cat /home/vit Chennai/.ssh/id_rsa.pub >> /home/vit Chennai/.ssh/authorized_keys
```

- Go to hadoop directory for Configuration purpose
- add java path in **hadoop_env.sh** file
- command to open hadoop_env.sh is \$ **gedit hadoop_env.sh**

```

hadoop-env.sh
/opt/hadoop/hadoop-2.8.0/etc/hadoop

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

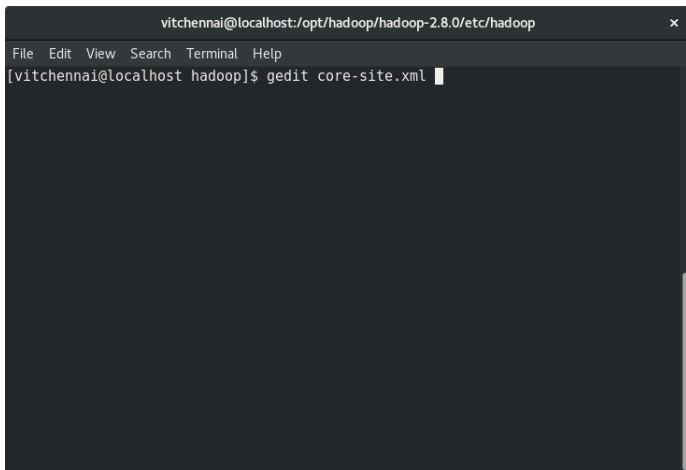
export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
for f in $(HADOOP_HOME/contrib/capacity-scheduler/*.jar); do
    if [ "$HADOOP_CLASSPATH" ]; then
        export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
    else
        export HADOOP_CLASSPATH=$f
    fi
done

# The maximum amount of heap to use in MB. Default is 1000.

```

- open **core-site.xml** for configuration purpose
- command to open core-site.xml is

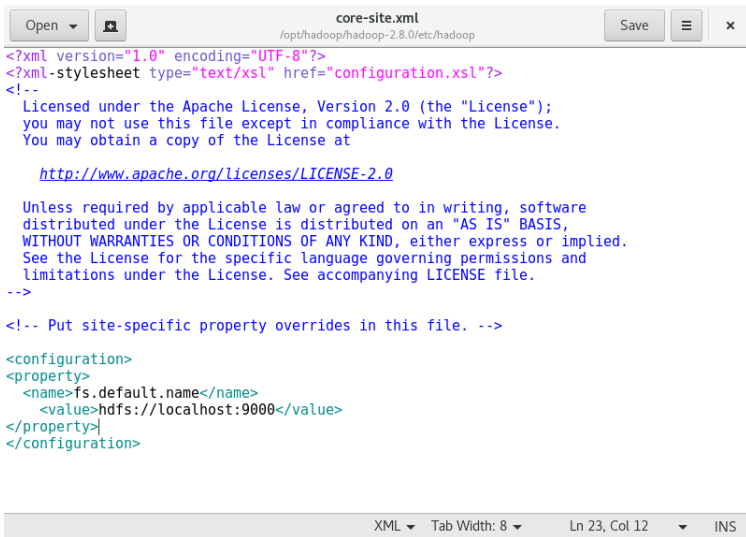


```
vitchennai@localhost:/opt/hadoop/hadoop-2.8.0/etc/hadoop
```

File Edit View Search Terminal Help

```
[vitchennai@localhost hadoop]$ gedit core-site.xml
```


- do the modification in core-site.xml as shown below, save and exit



```

<?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

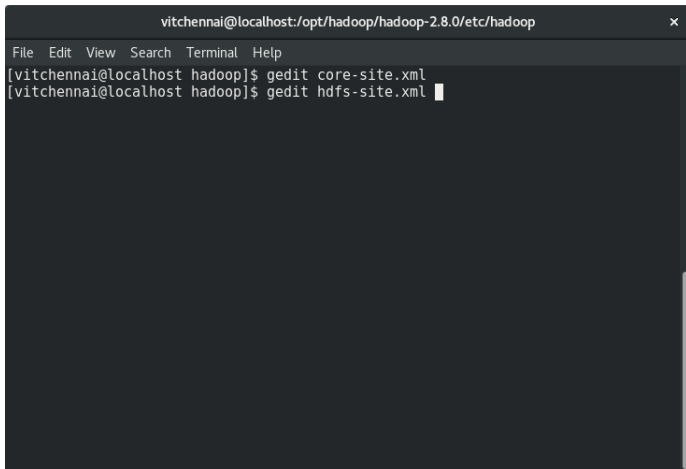
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under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

XML ▼ Tab Width: 8 ▼ Ln 23, Col 12 ▼ INS

- open **hdfs-site.xml** for configuration purpose
- command to open hdfs-site.xml is



```
vitchennai@localhost:/opt/hadoop/hadoop-2.8.0/etc/hadoop
```

File Edit View Search Terminal Help

```
[vitchennai@localhost hadoop]$ gedit core-site.xml  
[vitchennai@localhost hadoop]$ gedit hdfs-site.xml
```

- do the modification in `hdfs-site.xml` as shown below, save and exit



```

hdfs-site.xml
/opt/hadoop/hadoop-2.8.0/etc/hadoop

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.name.dir</name>
  <value>file:///opt/hadoop/hadoop_tmp/hdfs/namenode</value>
</property>

<property>
  <name>dfs.data.dir</name>
  <value>file:///opt/hadoop/hadoop_tmp/hdfs/datanode</value>
</property>

</configuration>
    
```

XML ▼ Tab Width: 8 ▼ Ln 32, Col 23 ▼ INS

- create **mapred-site.xml** from **mapred-site.xml.template** configuration purpose
- command to open mapred-site.xml is

```
vitchennai@localhost:/opt/hadoop/hadoop-2.8.0/etc/hadoop
```

```
File Edit View Search Terminal Help
```

```
[vitchennai@localhost hadoop]$ gedit core-site.xml
[vitchennai@localhost hadoop]$ gedit hdfs-site.xml
[vitchennai@localhost hadoop]$ cp mapred-site.xml.template mapred-site.xml
[vitchennai@localhost hadoop]$ gedit mapred-site.xml
[vitchennai@localhost hadoop]$
```

- do the modification in mapred-site.xml as shown below, save and exit

Open

mapred-site.xml

Save

/opt/hadoop/hadoop-2.8.0/etc/hadoop

```

<?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>
    
```

XML

Tab Width: 8

Ln 23, Col 13

INS

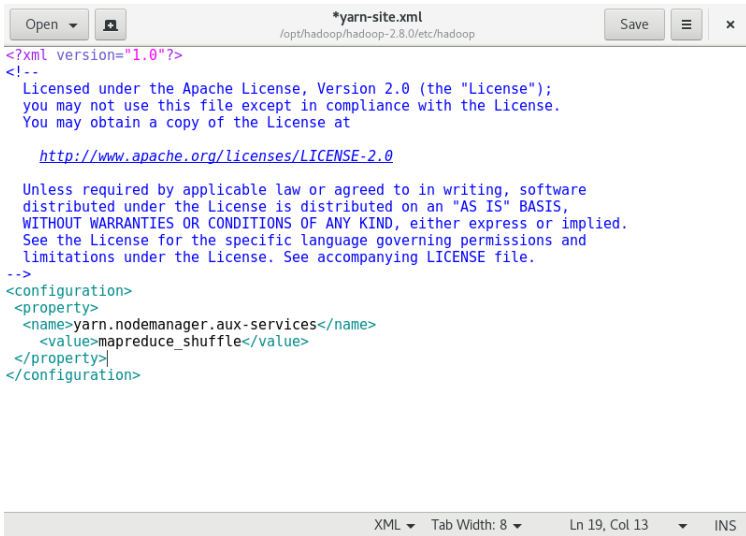
- open **yarn-site.xml** for configuration purpose
- command to open yarn-site.xml is

```
vitchennai@localhost:/opt/hadoop/hadoop-2.8.0/etc/hadoop
```

```
File Edit View Search Terminal Help
```

```
[vitchennai@localhost hadoop]$ gedit core-site.xml
[vitchennai@localhost hadoop]$ gedit hdfs-site.xml
[vitchennai@localhost hadoop]$ cp mapred-site.xml.template mapred-site.xml
[vitchennai@localhost hadoop]$ gedit mapred-site.xml
[vitchennai@localhost hadoop]$ gedit mapred-site.xml
[vitchennai@localhost hadoop]$ gedit yarn-site.xml
```

- do the modification in yarn-site.xml as shown below, save and exit



```

<?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>
</configuration>
    
```

XML ▾ Tab Width: 8 ▾ Ln 19, Col 13 ▾ INS

- 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

```
vitchennai@localhost: /opt/hadoop
File Edit View Search Terminal Help
[vitchennai@localhost ~]$ cd /opt/
[vitchennai@localhost opt]$ ls
hadoop JAVA8
[vitchennai@localhost opt]$ cd hadoop/
[vitchennai@localhost hadoop]$ sudo mkdir -p hadoop_tmp/hdfs/namenode
[sudo] password for vitchennai:

[vitchennai@localhost hadoop]$
[vitchennai@localhost hadoop]$
[vitchennai@localhost hadoop]$ sudo mkdir -p hadoop_tmp/hdfs/datanode
[vitchennai@localhost hadoop]$ sudo chmod 777 -R /opt/hadoop/
[vitchennai@localhost hadoop]$ sudo chmod 777 -R /opt/hadoop/hadoop_tmp/
[vitchennai@localhost hadoop]$
```


- 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-2.8.0 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 ~]$
```

- If Datanode is not started, then change the permissions of datanode
- The following command is used

```
sudo chown vitchennai:root  
/opt/Hadoop/hadoop_tmp/hdfs/datanode
```

- Now check the background process for hadoop distributed file system

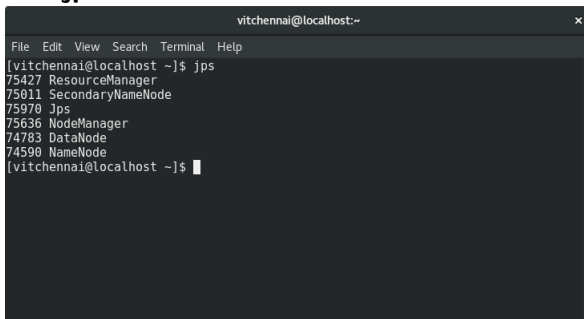
```
vitchennai@localhost:~  
File Edit View Search Terminal Help  
[vitchennai@localhost ~]$ jps  
75251 Jps  
75011 SecondaryNameNode  
74783 DataNode  
74590 NameNode  
[vitchennai@localhost ~]$
```

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

```

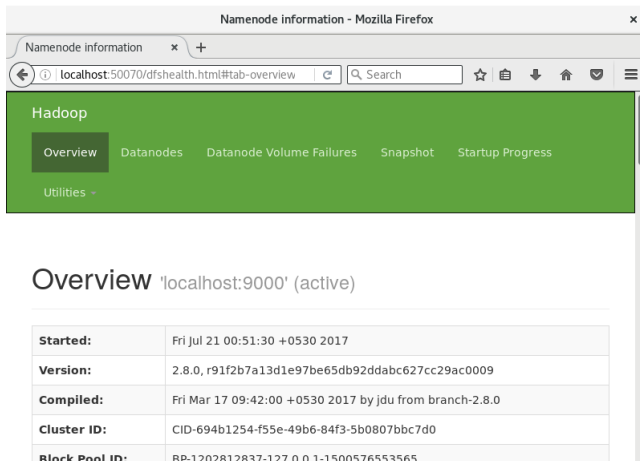
vitchennai@localhost:~
File Edit View Search Terminal Help
[vitchennai@localhost ~]$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /opt/hadoop/hadoop-2.8.0/logs/yarn-vitchennai-resource-manager-localhost.localdomain.out
vitchennai@localhost's password:
localhost: starting nodemanager, 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-2.8 on local machine are
- **NameNode**
- **DataNode**
- **SecondaryNameNode**
- **NodeManager**
- **ResourceManager**
- we have to use **jps** command to check



```
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 ~]$
```

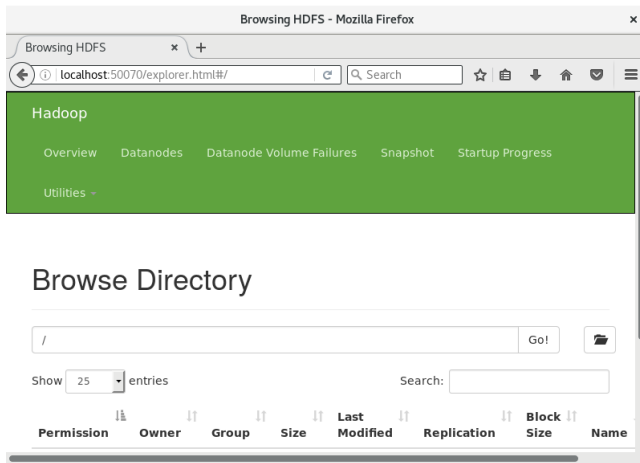
- **UI view of Hadoop**
- Open **http://localhost:50070** in browser → Namenode



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**



- command to stop hadoop are **\$stop-dfs.sh** and **\$stop-yarn.sh**

```
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 ~]$
```

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