VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ▪ CHENNAI
www.vit.ac.in

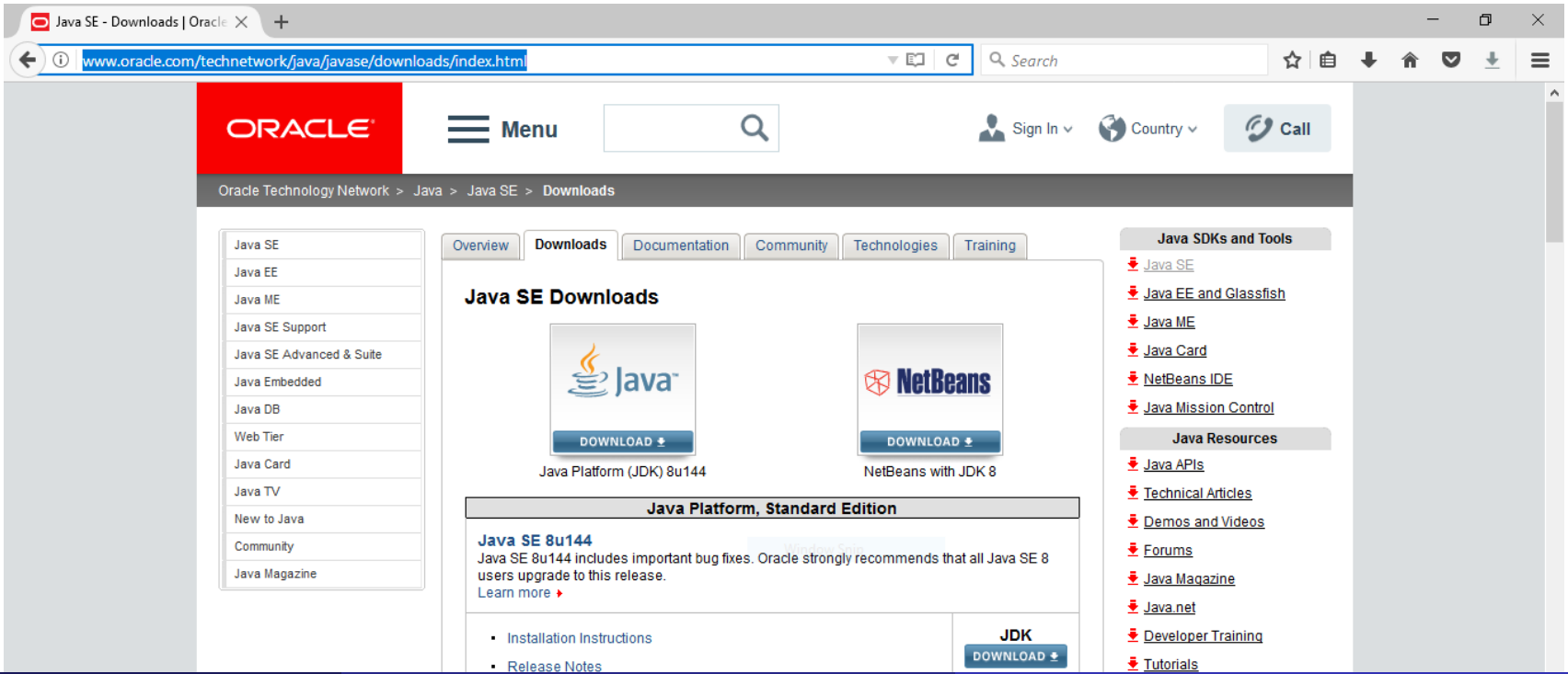
Big Data Frameworks CSE6701

Prof. Ramesh Ragala

November 7, 2017

- **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



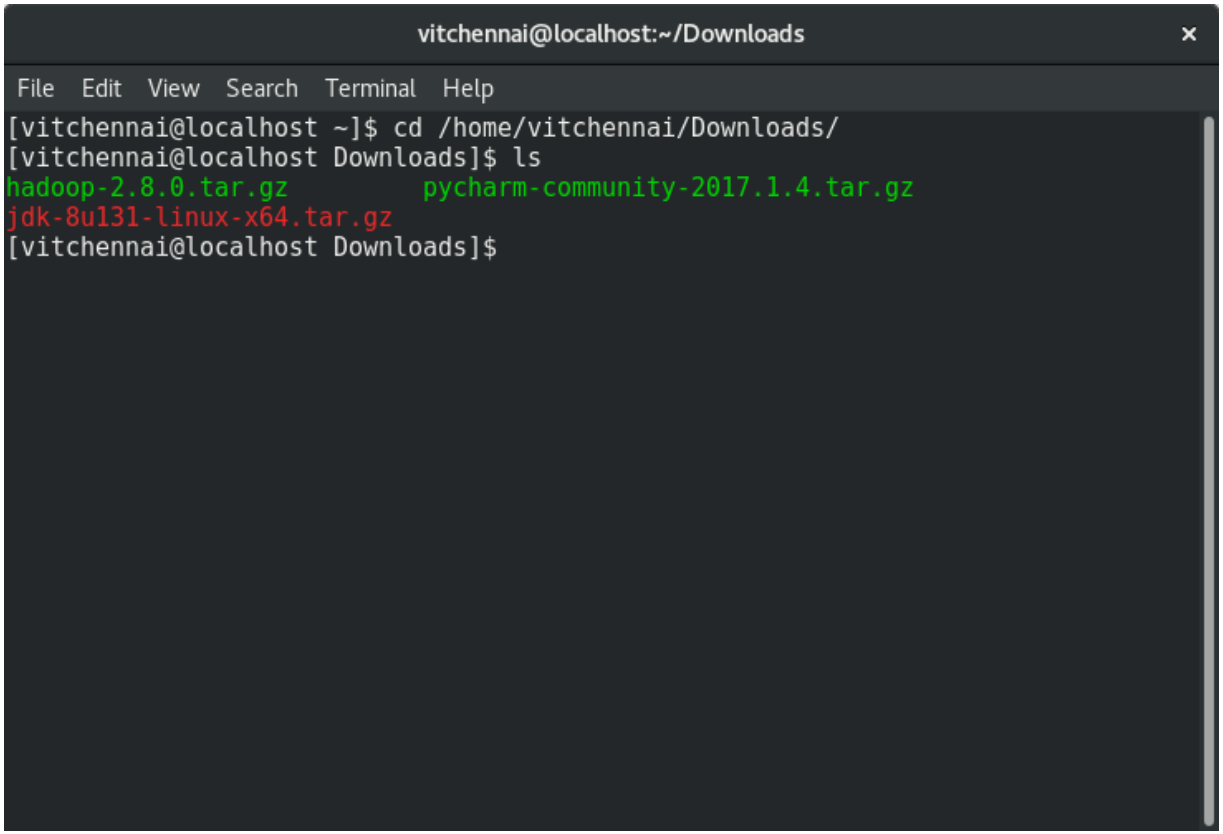
Hadoop Installation

- 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**

- set the path set for java in /etc/profile
  - command is $ **gedit /etc/profile**
- Append the following lines in /etc/profile
  - **export JAVA_HOME=/opt/java8/jdk-x.x.x**
  - **export PATH=$JAVA_HOME/bin:$PATH**
- restart the terminal
  - command is $ **sudo source /etc/profile**

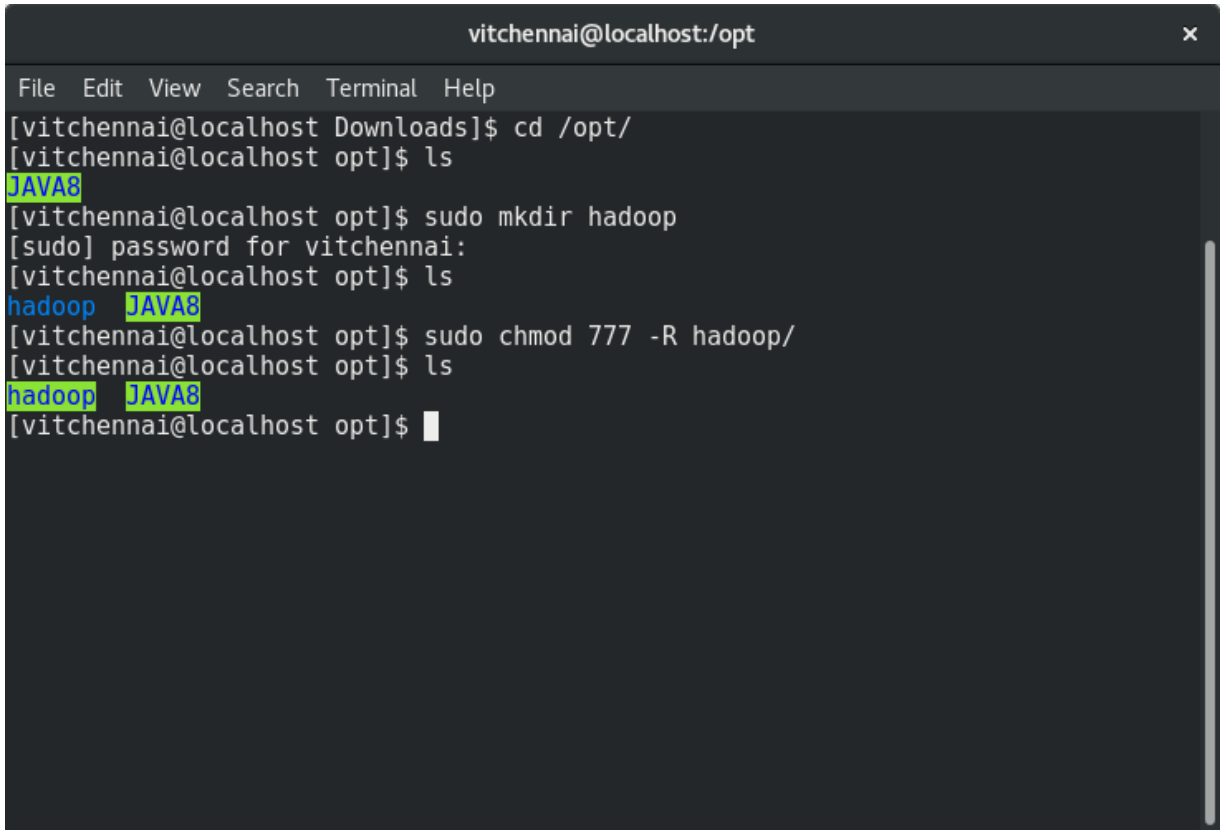
Hadoop Installation steps

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

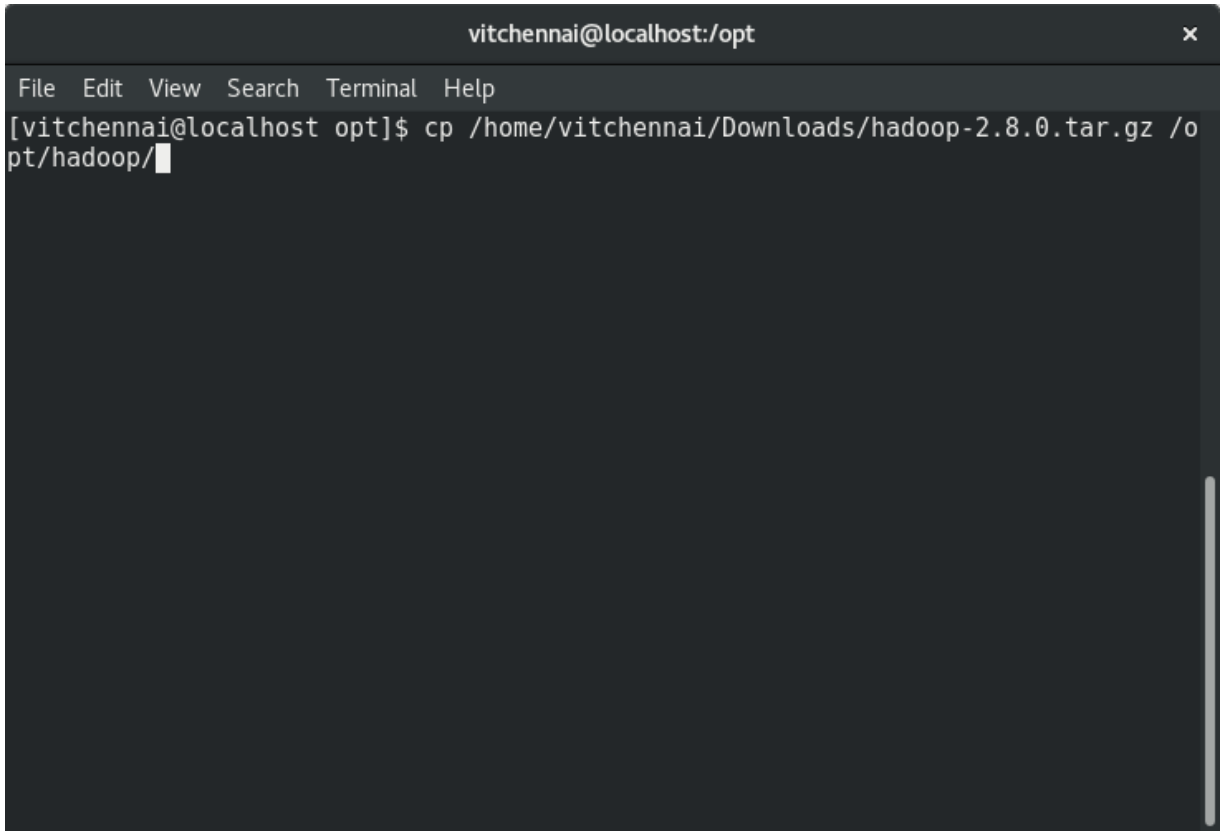


Hadoop Installation steps

- create hadoop directory in /opt directory

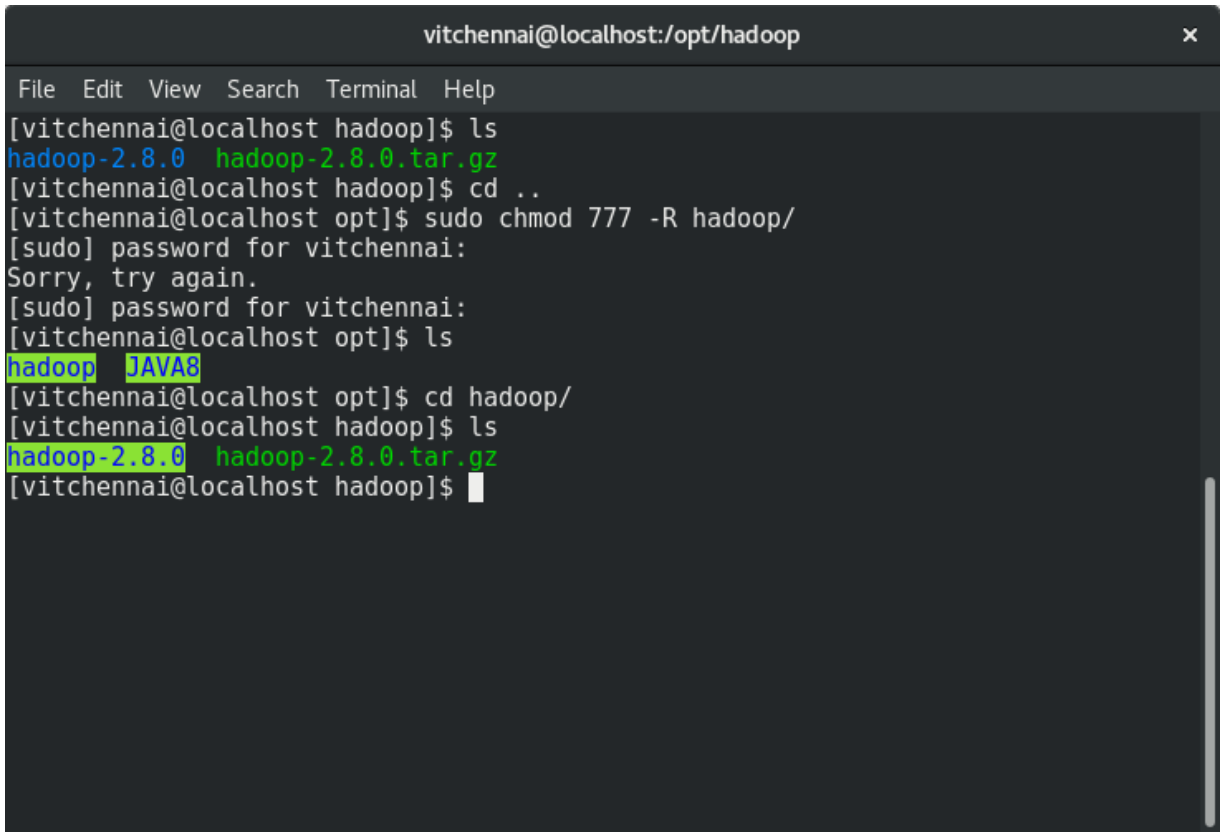


- change the permission of this directory
  - command is $ **sudo chmod 777 -R /opt/hadoop**
- copy the hadoop-2.8.0.tar.gz into /opt/hadoop directory
  - command is $ **cp /home/vitchennai/Downloads/hadoop-2.8.0.tar.gz /opt/hadoop/**

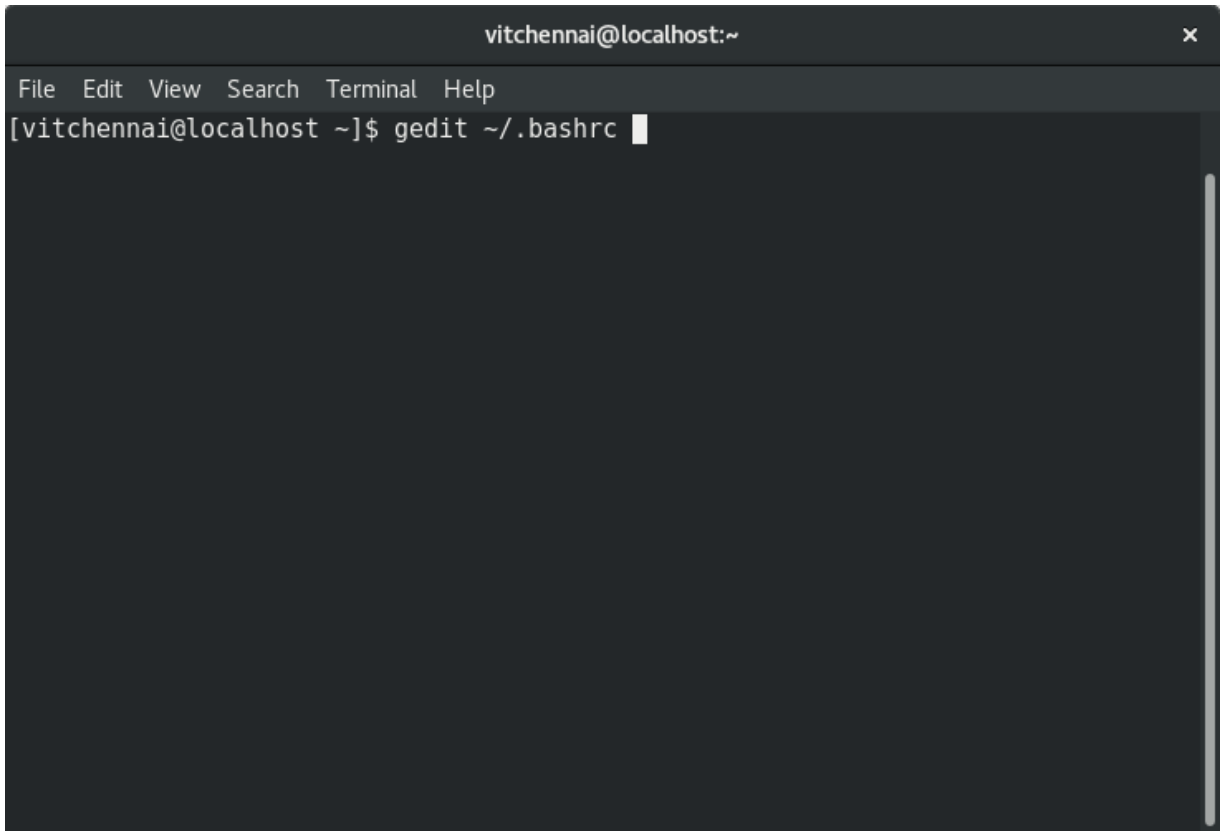


Hadoop Installation steps

- 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**



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



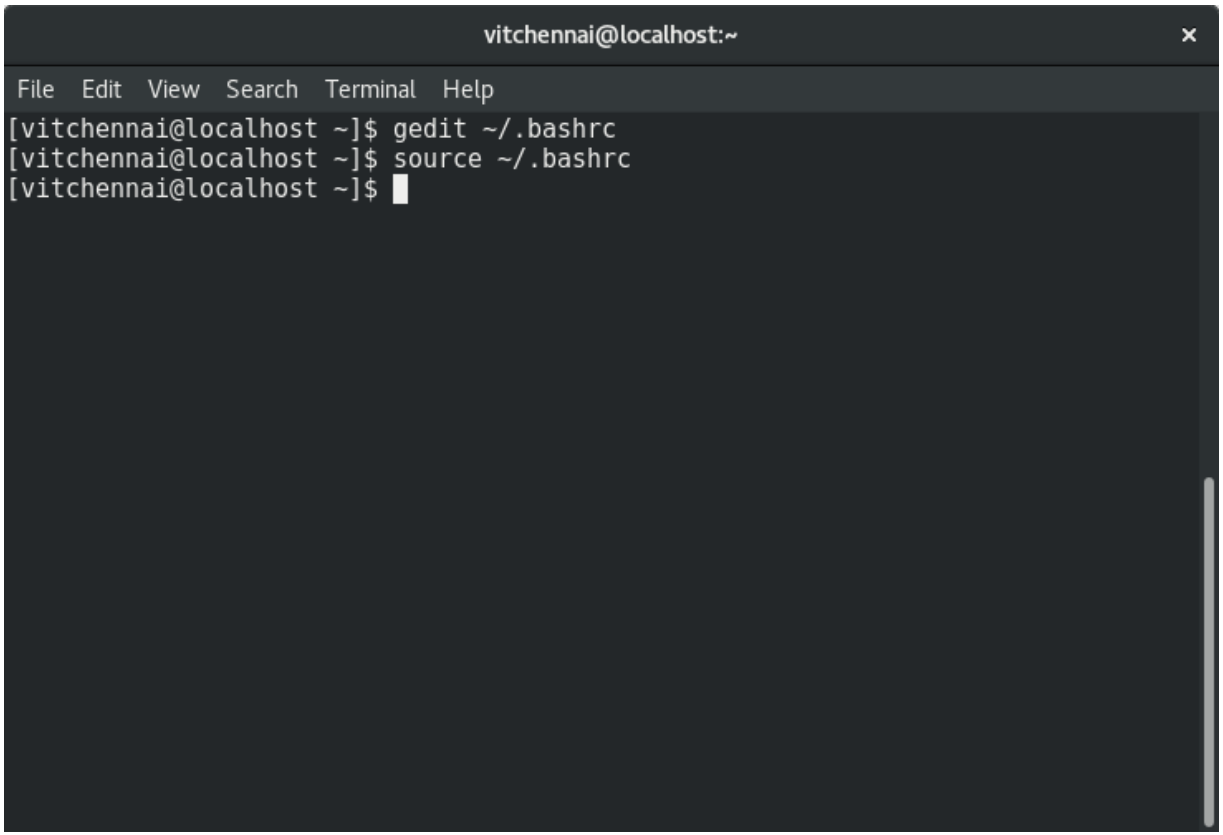
- Append the Hadoop and java paths

Open .bashrc ~/ Save

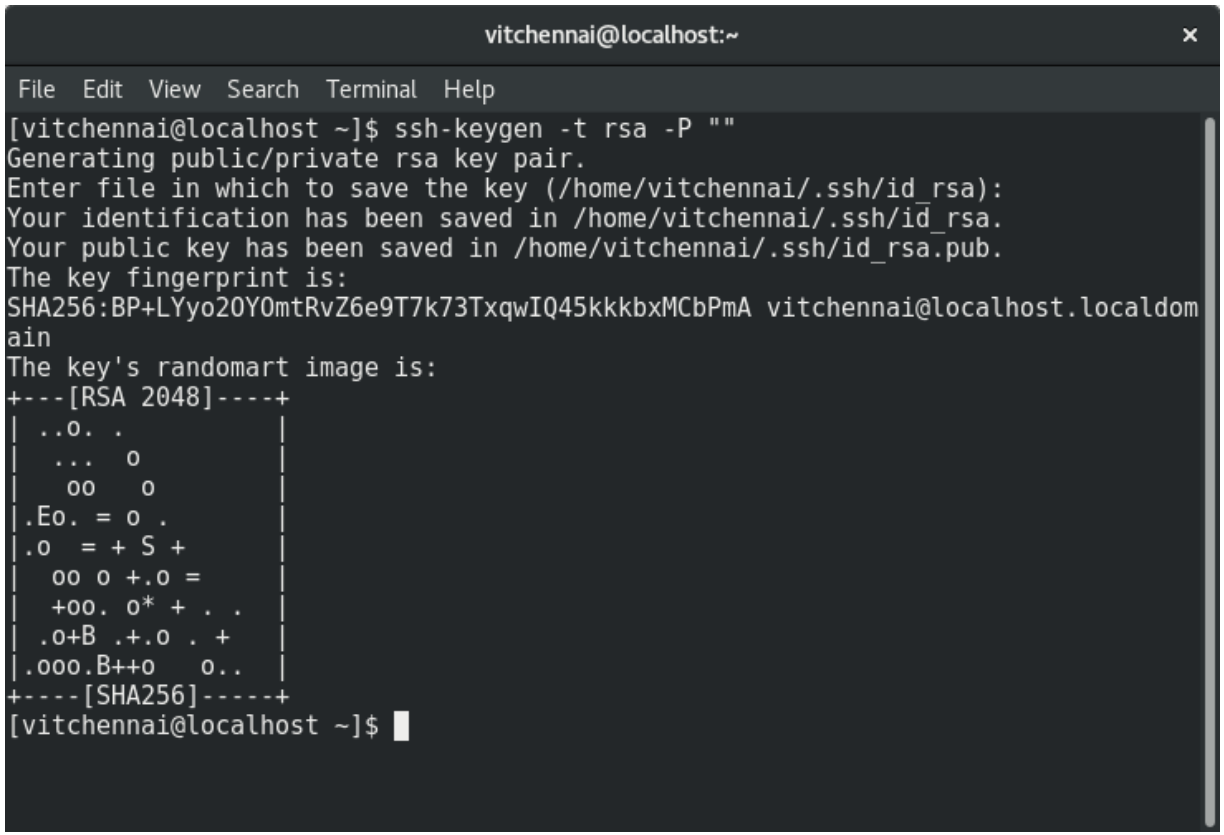
```
# .bashrc

# 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

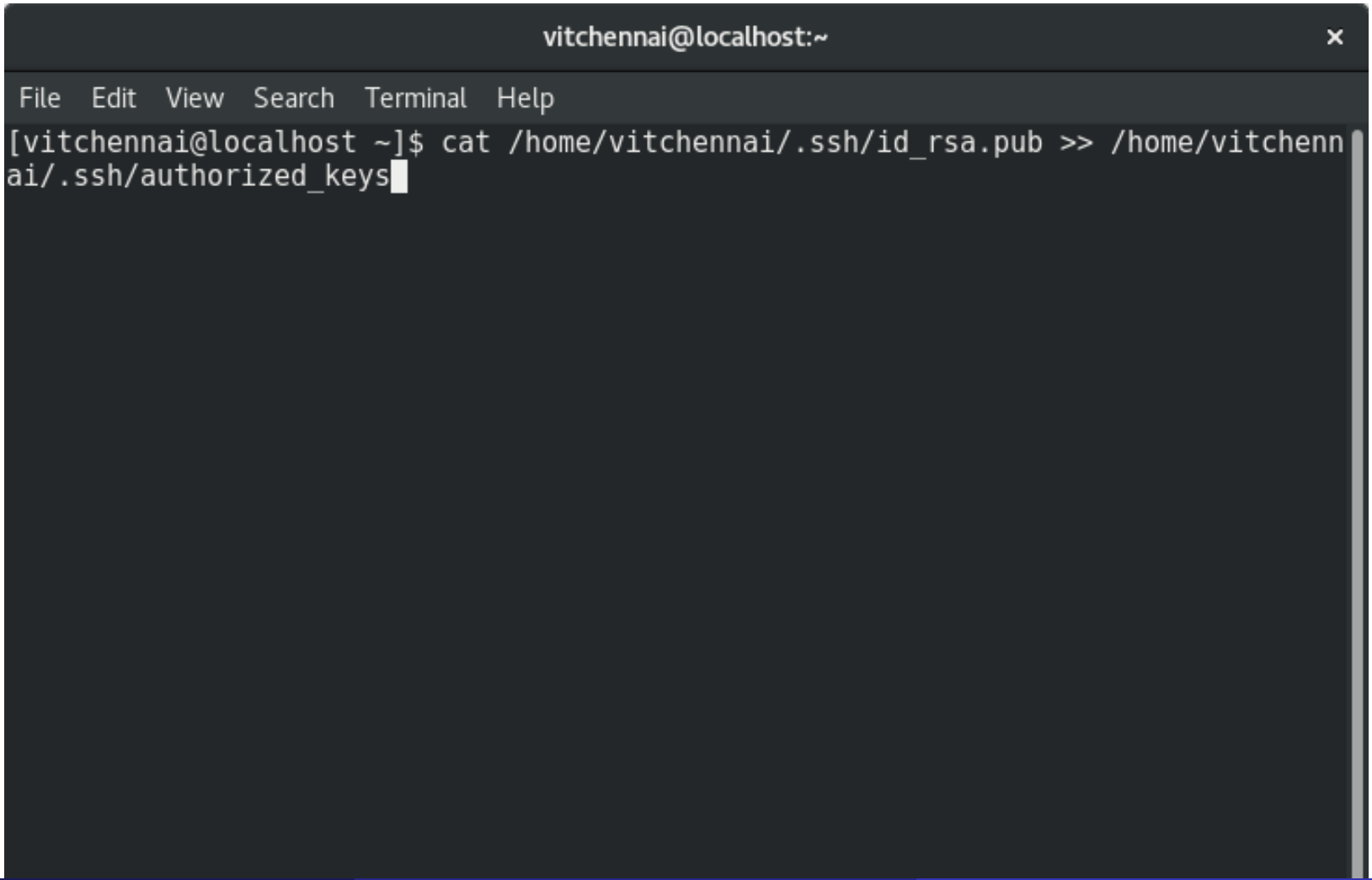
- Restart the terminal
  - command to restart is $**source ~/.bashrc**



- 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.
- Commands is



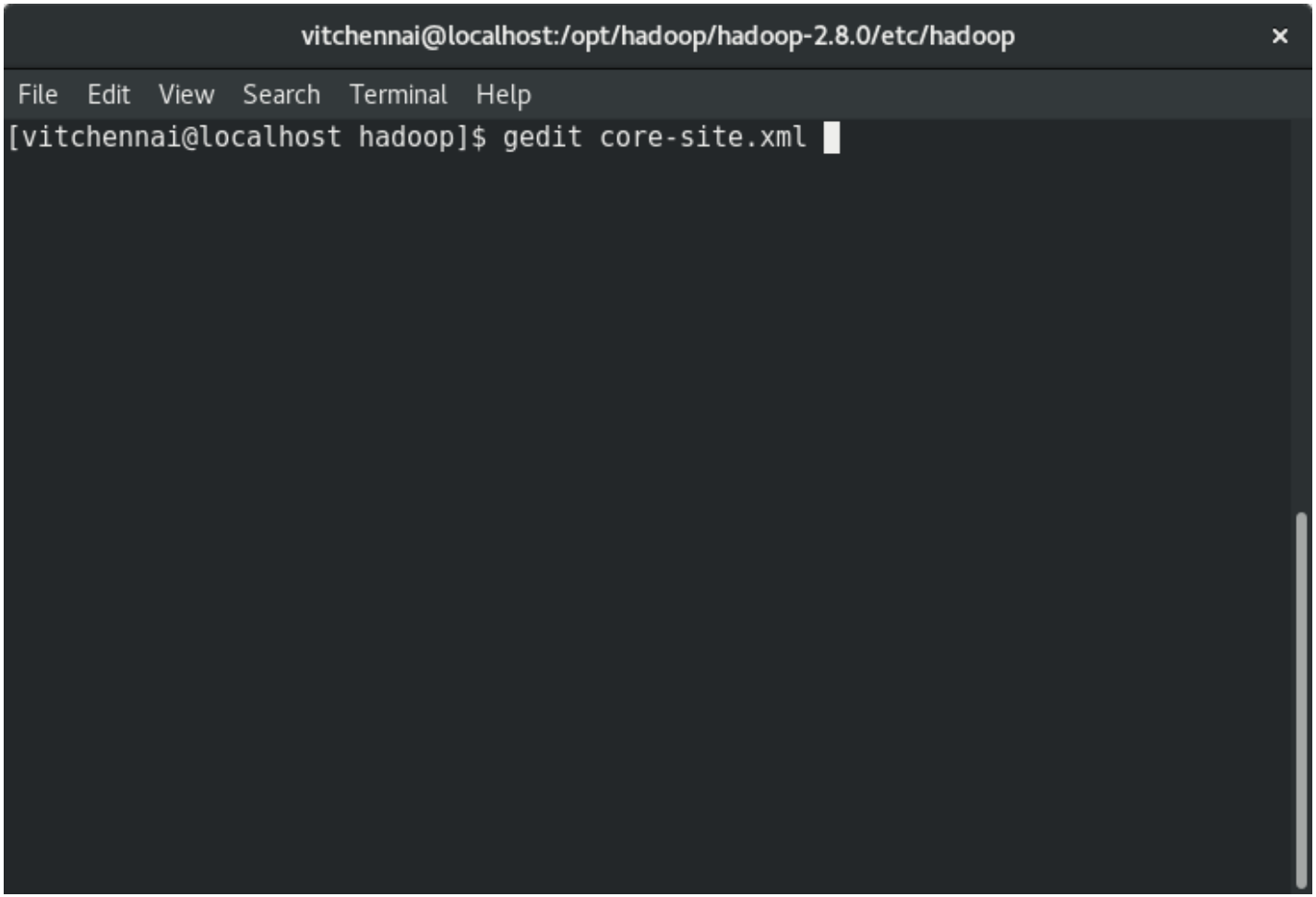
- Go to hadoop directory for Configuration purpose
- add java path in **hadoop_env.sh** file
- command to open hadoop_env.sh is $ **gedit hadoop_env.sh**

Open | hadoop-env.sh | /opt/hadoop/hadoop-2.8.0/etc/hadoop | Save

```
# 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.
export JAVA_HOME=/opt/JAVA8/jdk1.8.0_131

# The jsvc 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.
```

sh | Tab Width: 8 | Ln 25, Col 18 | INS

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



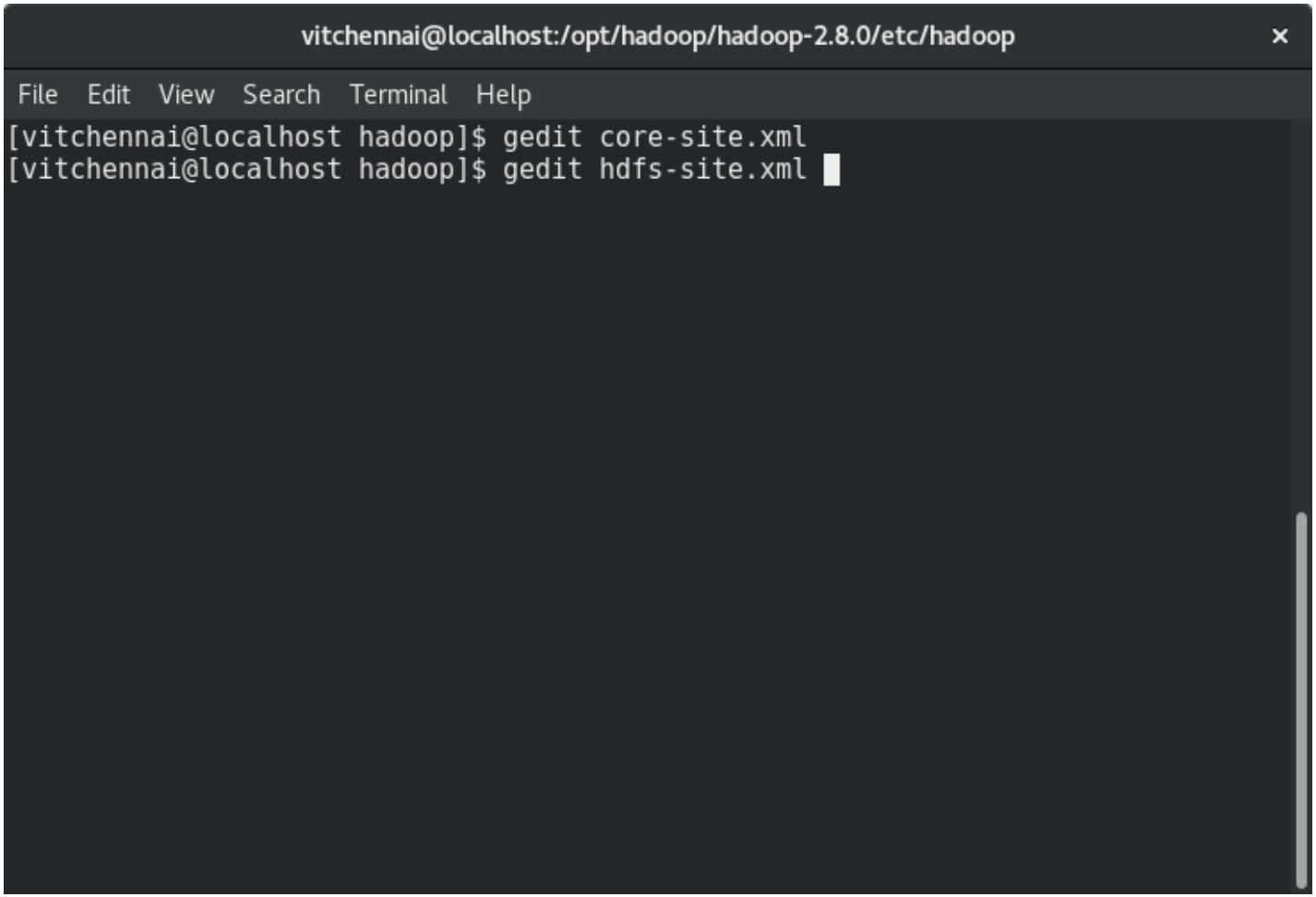
- do the modification in core-site.xml as shown below, save and exit

Open | core-site.xml — /opt/hadoop/hadoop-2.8.0/etc/hadoop | Save

```
<?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



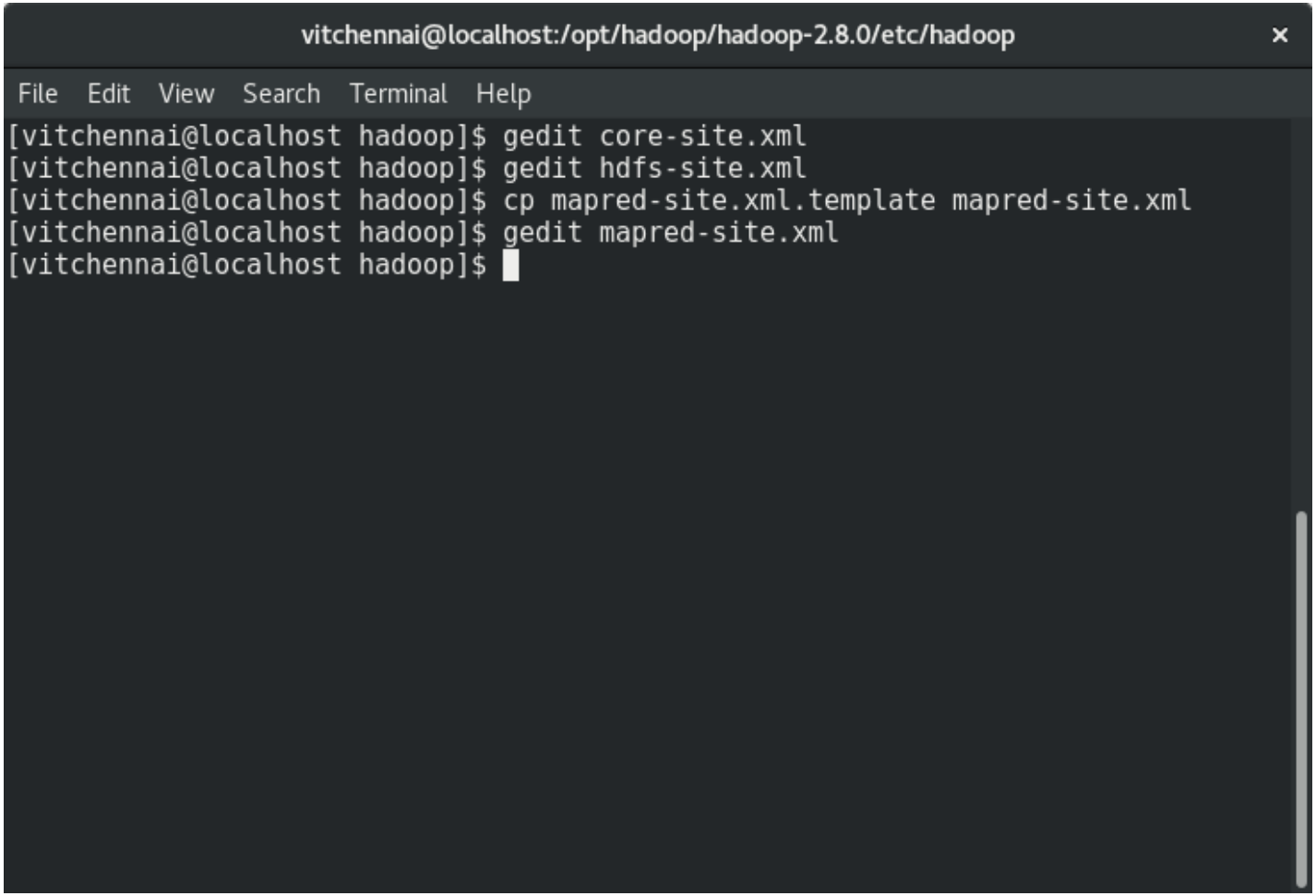
- do the modification in hdfs-site.xml as shown below, save and exit

Open | hdfs-site.xml /opt/hadoop/hadoop-2.8.0/etc/hadoop | Save

```
    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



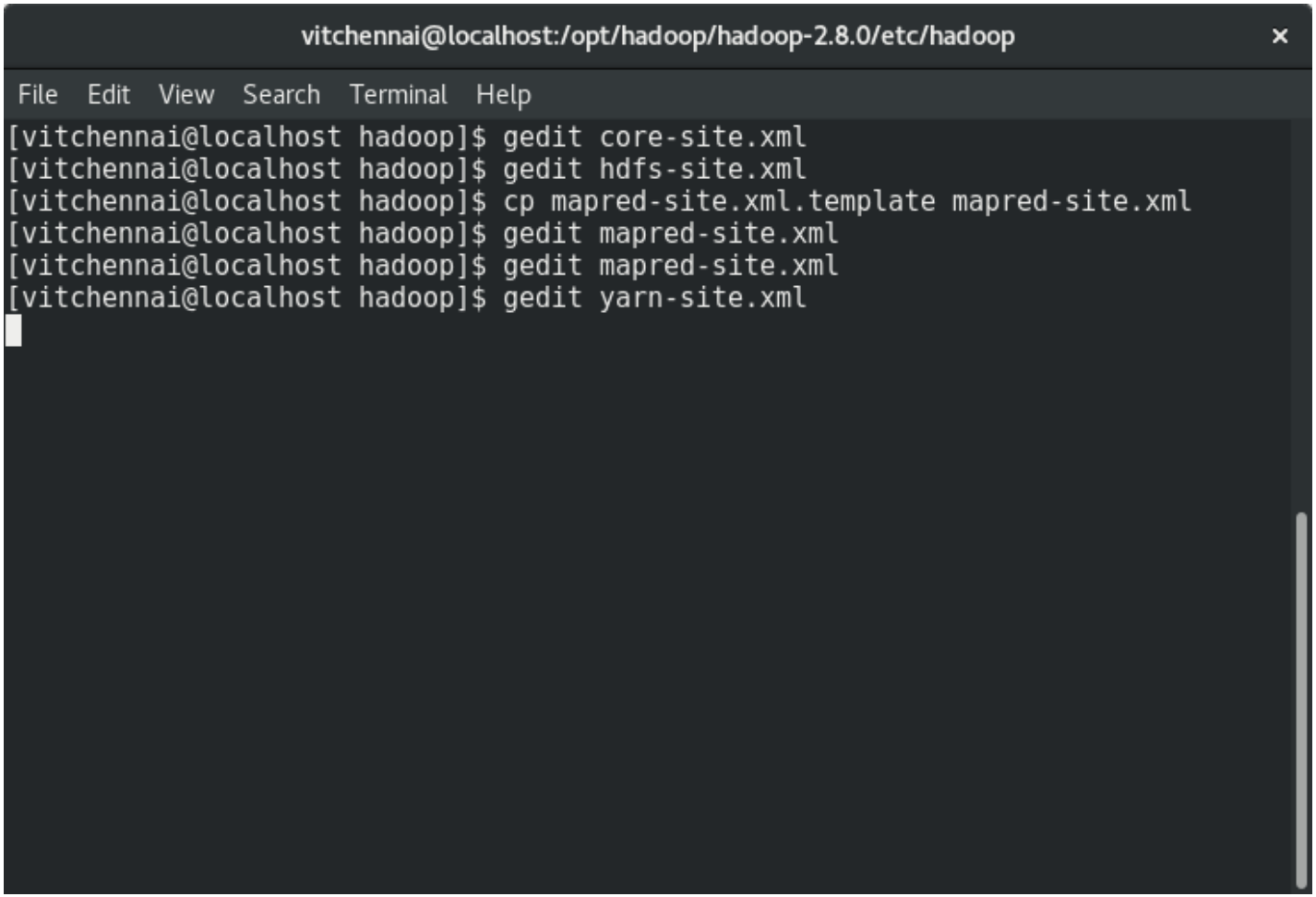
- do the modification in mapred-site.xml as shown below, save and exit

Open | mapred-site.xml /opt/hadoop/hadoop-2.8.0/etc/hadoop | Save

```
<?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



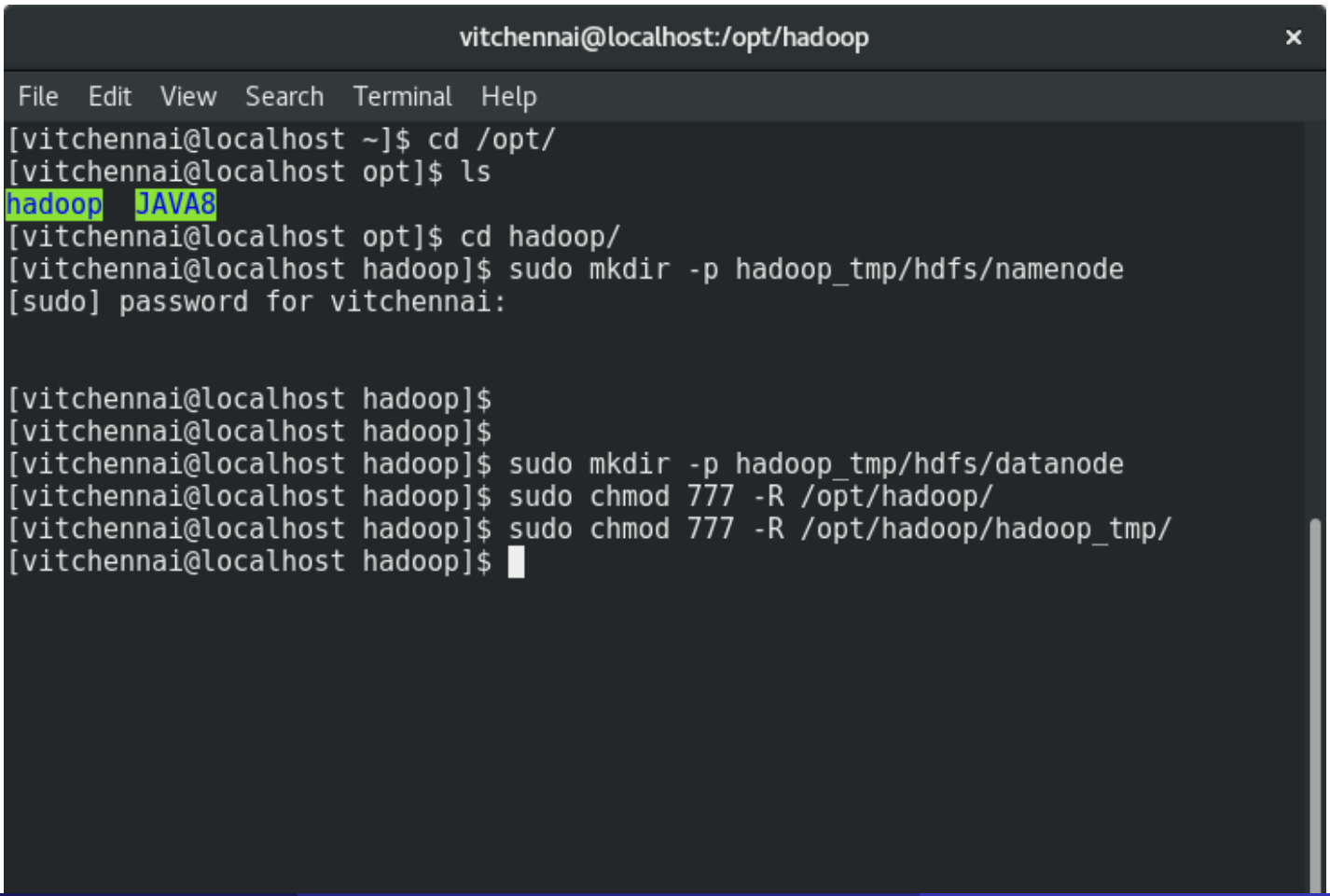
- do the modification in yarn-site.xml as shown below, save and exit

Open *yarn-site.xml /opt/hadoop/hadoop-2.8.0/etc/hadoop Save

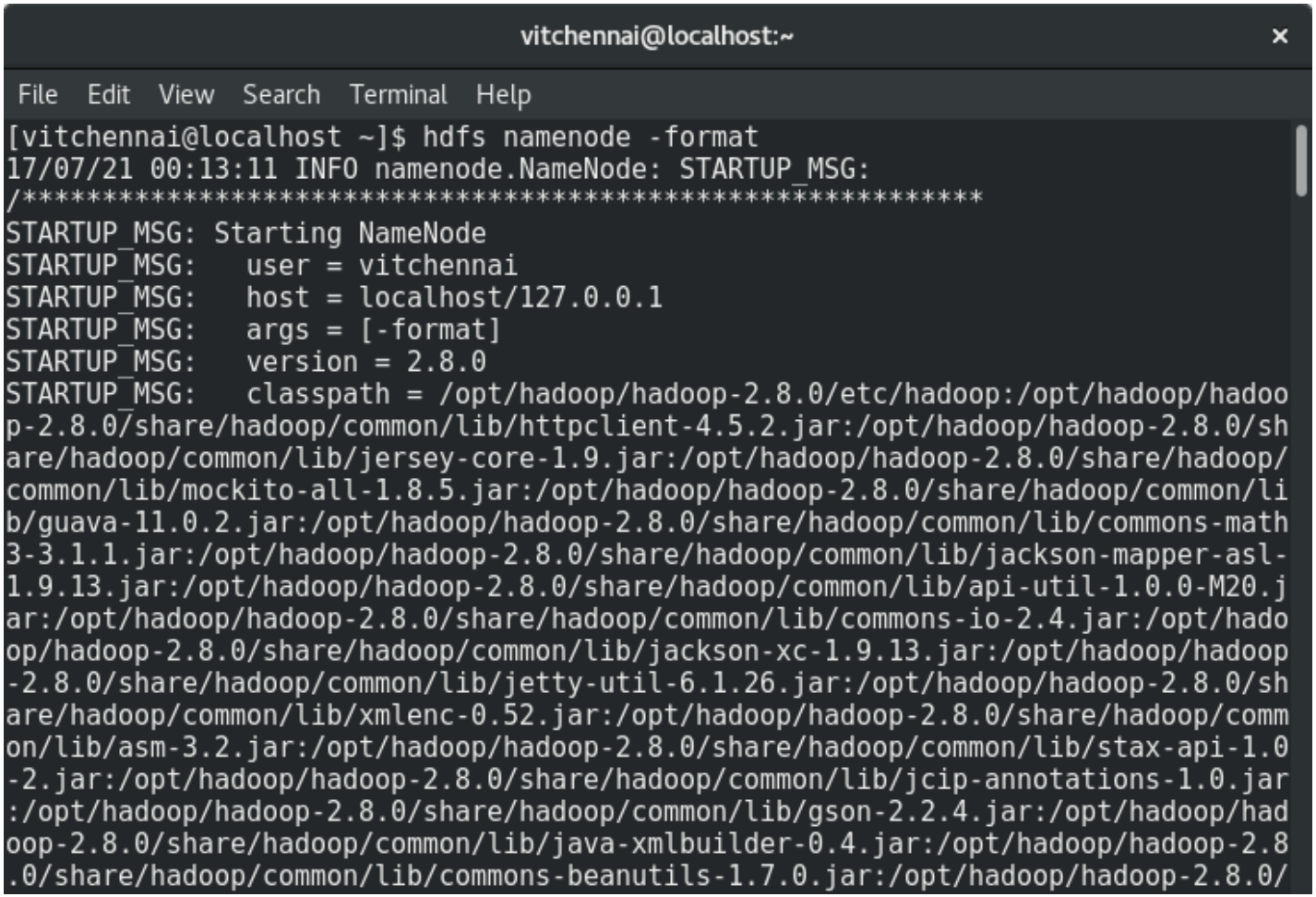
```
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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



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



- **Hadoop-2.8.0 installation has completed**
- **Check the Hadoop status**

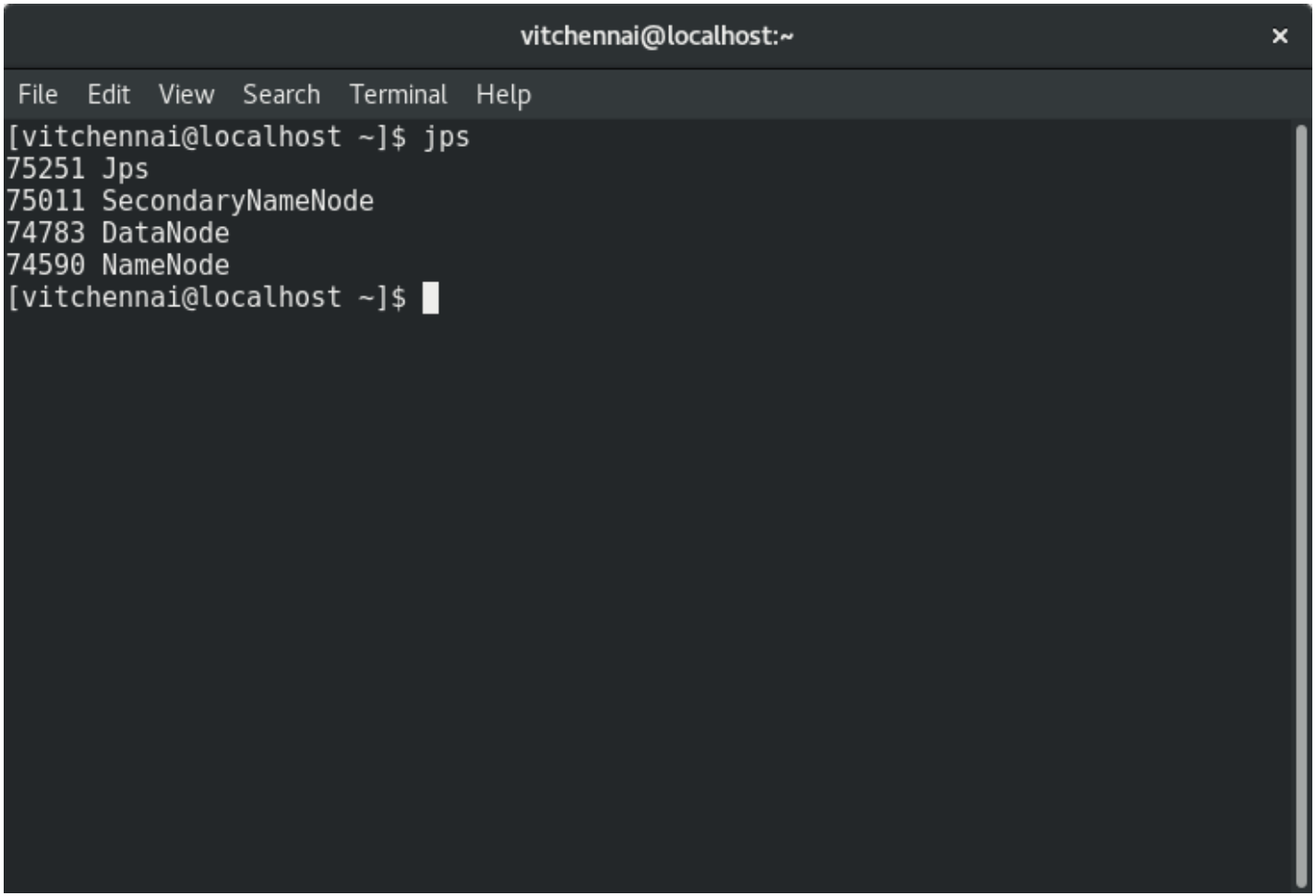
- Now start the Hadoop Distributed File System
- The following commands (**start-dfs.sh**) is used for this

```
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+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 ~]$
```

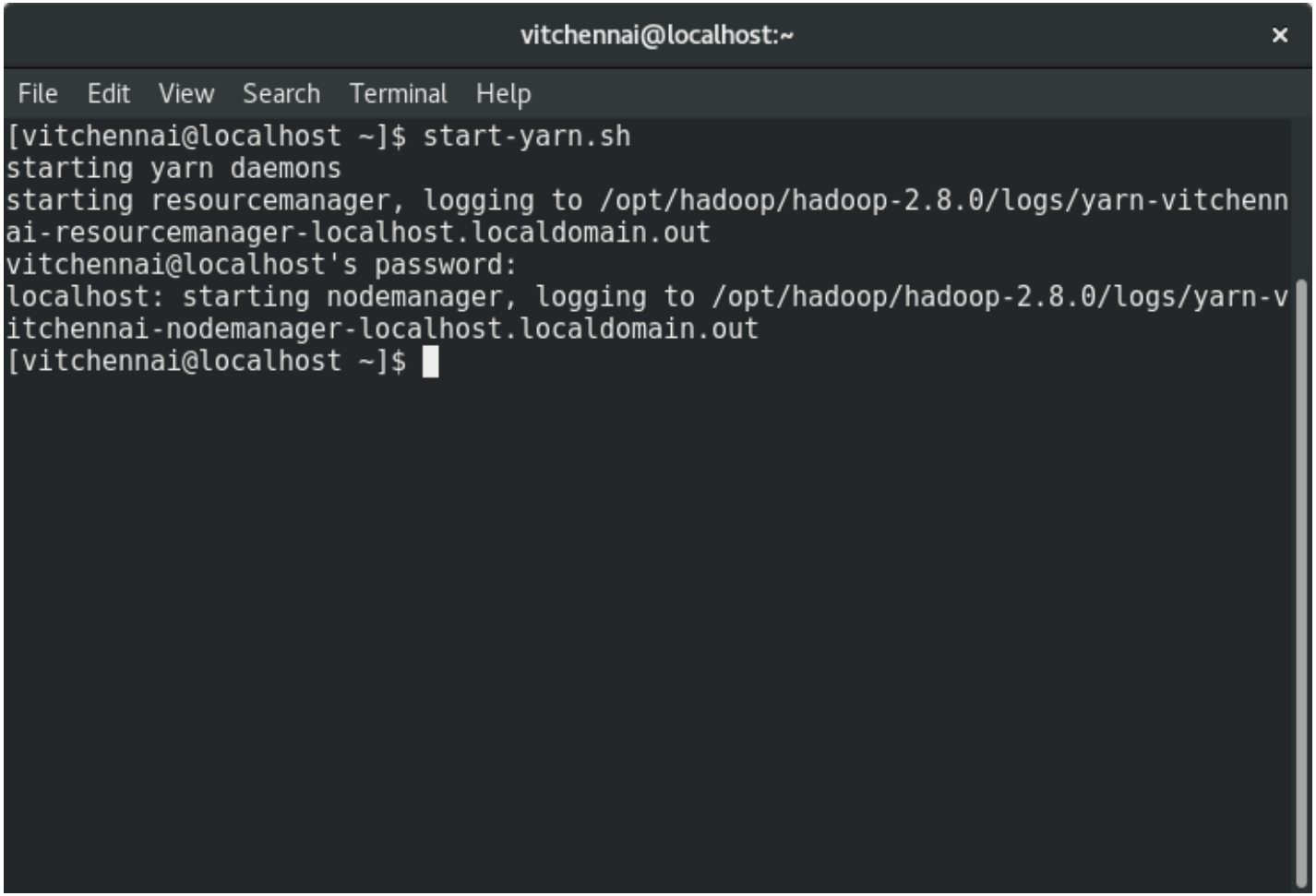
- If Datanode is not started, then change the permissions of datanode
- The following command is used

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

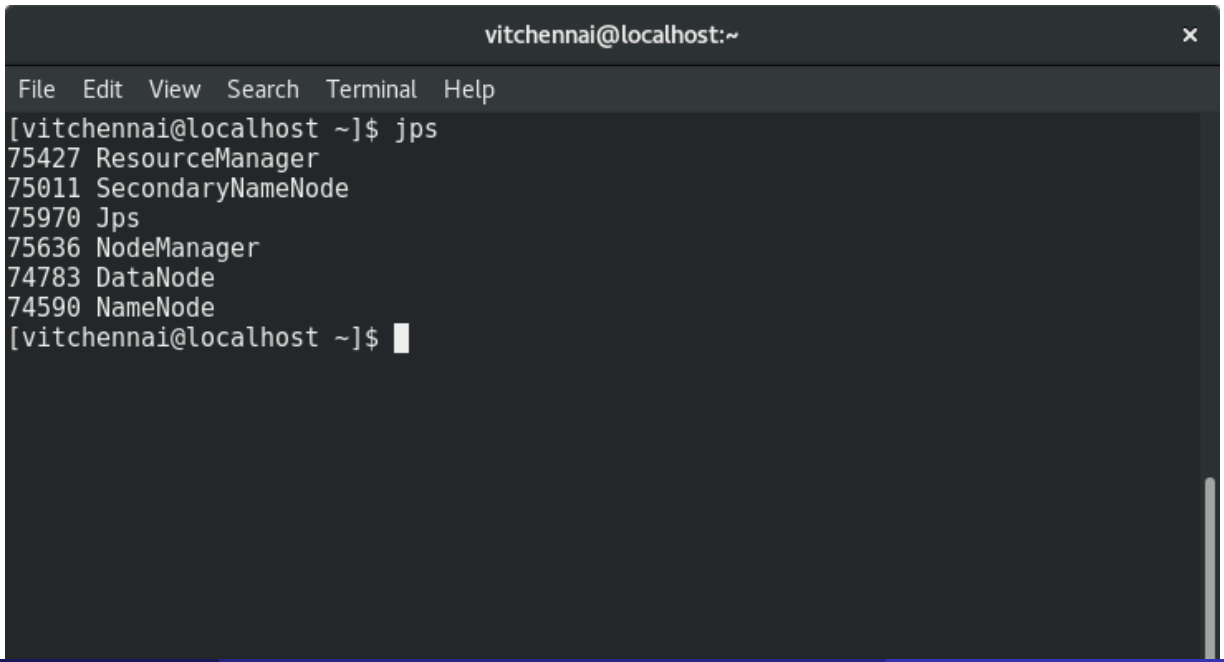
- Now check the background process for hadoop distributed file system



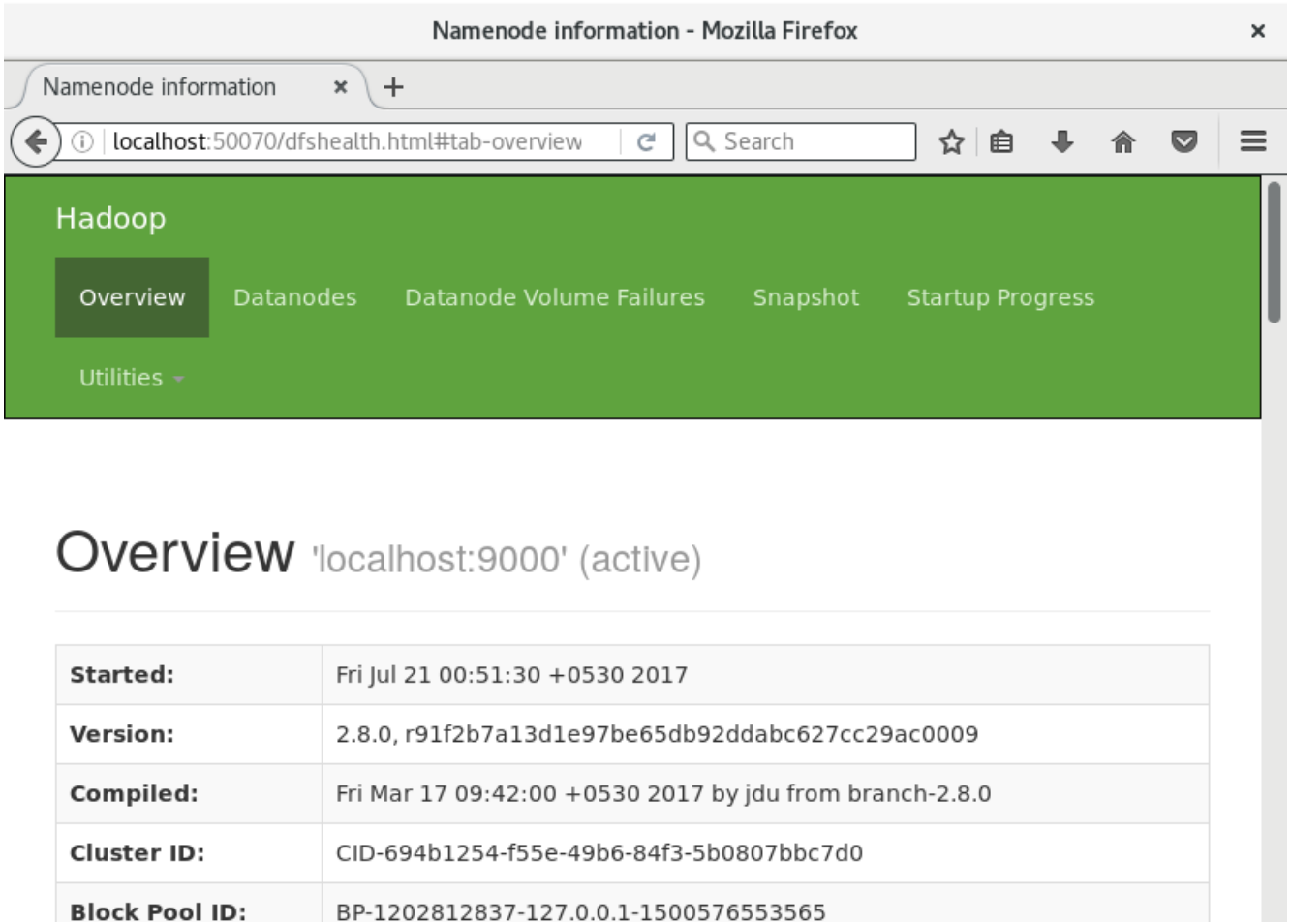
- Now start yarn resources for hadoop
- The following commands (**start-yarn.sh**) is used for this



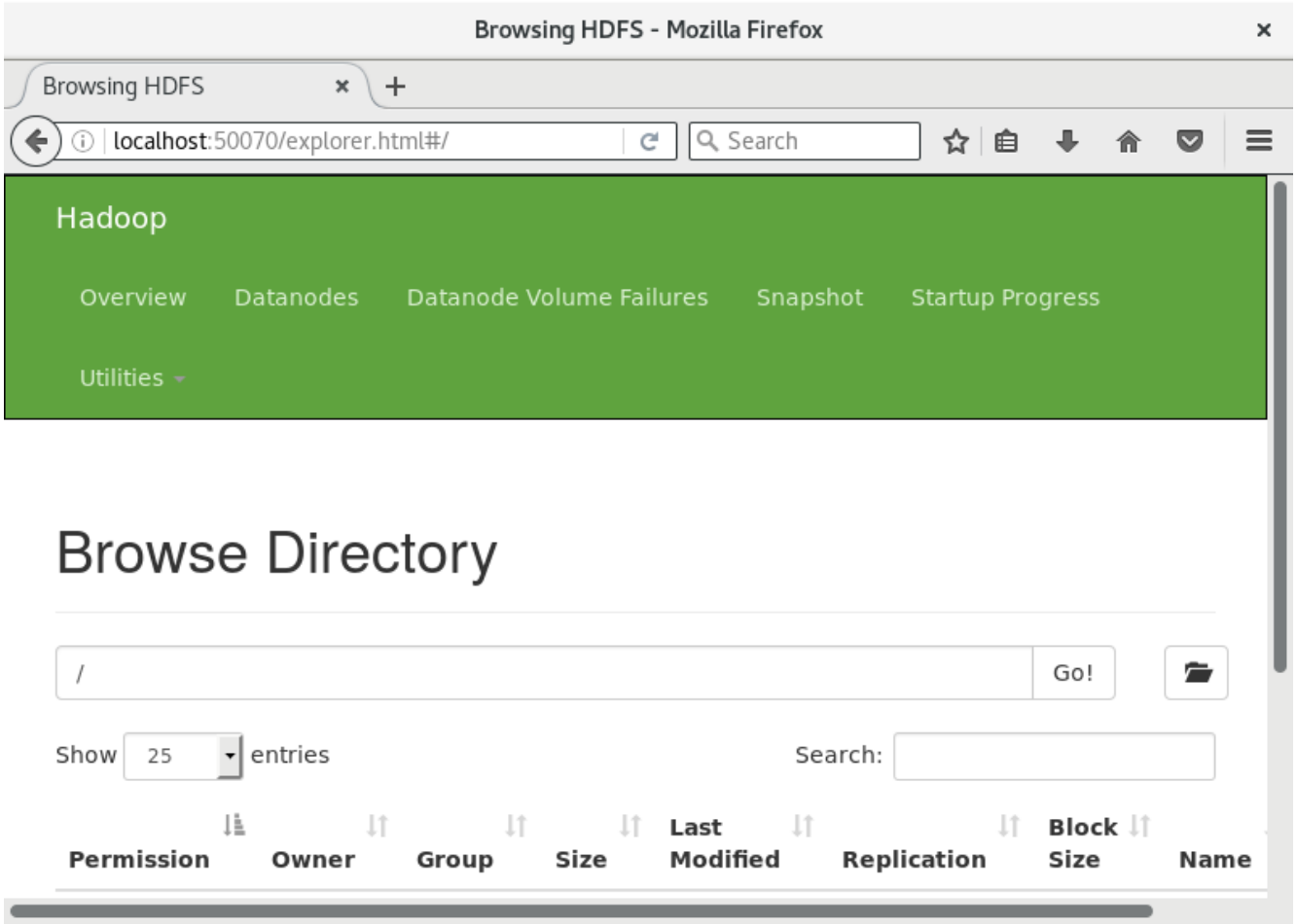
- The total number of daemons to execute hadoop-2.8 on local machine are
- **NameNode**
- **DataNode**
- **SecondaryNameNode**
- **NodeManager**
- **ResourceManager**
- we have to use **jps** command to check



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



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



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