Topics in Computing Lab Assignment 4 : Hadoop

Team R : Vandita Goyal (2016ucp1004)
Nidheesh Panchal (2016ucp1008)
G. Jahnvi (2016ucp1332)

Objective:

Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications. Hadoop can handle various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide.

A Multi Node Cluster in Hadoop contains two or more DataNodes in a distributed Hadoop environment.Hadoop has two main components. The first component, the Hadoop Distributed File System, helps split the data, put it on different nodes, replicate it and manage it. The second component, MapReduce, processes the data on each node in parallel and calculates the results of the job.

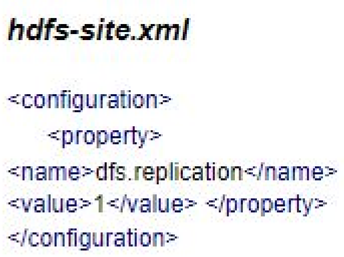
Hadoop multi-node cluster involves a minimum of two nodes - Master and Slave

Implementation:

1. Installation of Hadoop and setting up of Hadoop single node:

Steps followed:

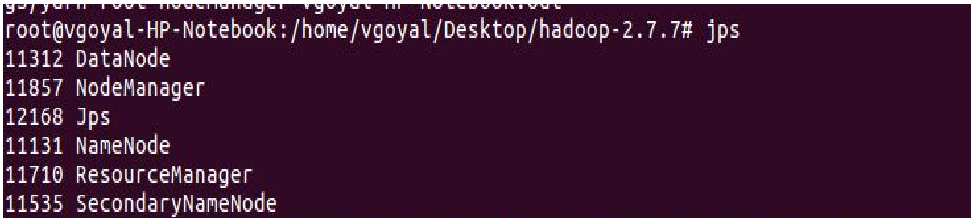
1. Install Java and set JAVA_HOME variable to **/usr/lib/jvm/java-1.7.0-openjdk-amd64**
2. Download Hadoop from one of the Apache Mirrors (We installed Hadoop 2.7.x)
3. Set HADOOP_CLASSPATH variable to **$JAVA_HOME/lib/tools.jar**
4. Check bin/hadoop
5. Edit core-site.xml and hdfs-site.xml residing in hadoop2.7.7/etc/hadoop



hdfs-site.xml

```
<configuration>
     <property>
<name>dfs.replication</name>
<value>1</value> </property>
</configuration>
```

6. Check ssh to localhost**: ssh localhost**
7. Format the file system - **bin/hdfs namenode –format**
8. Start the daemons **sbin/start-dfs.sh**
9. Check web interface for namenode at http://localhost:50070/

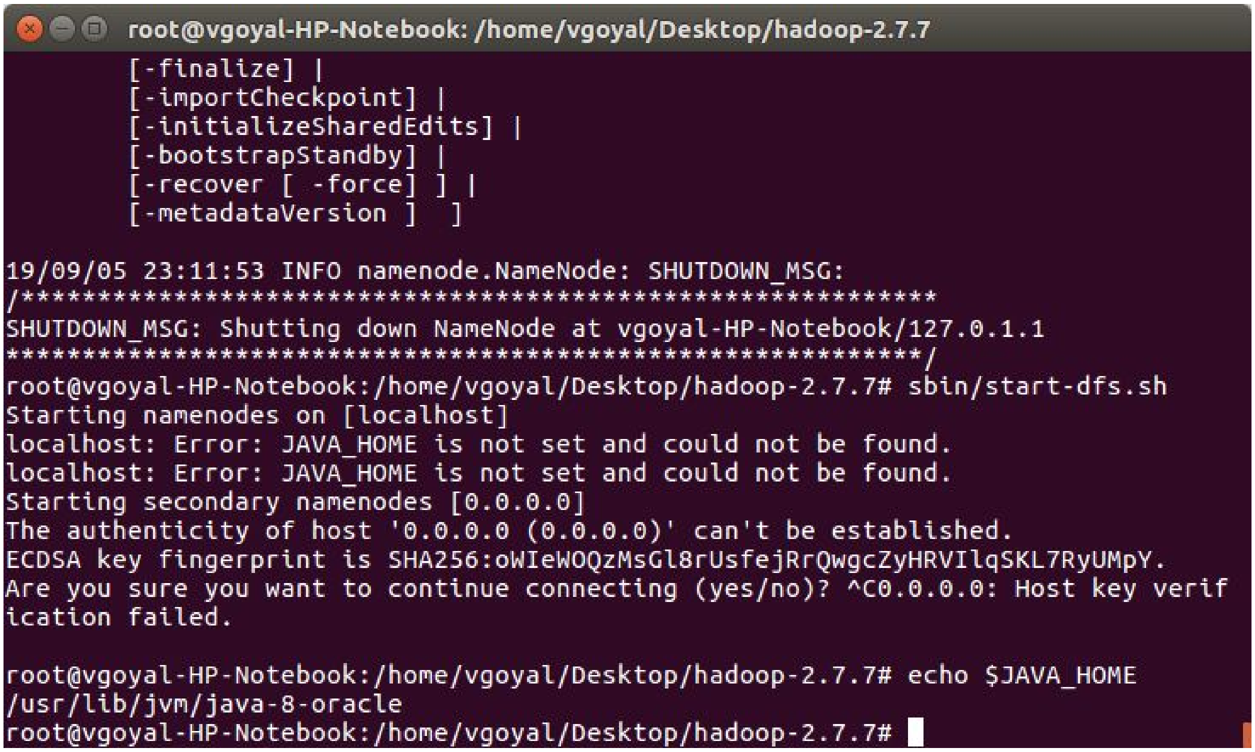


```
gs/yarn-root-nodemanager-vgoyal-HP-Notebook.out
root@vgoyal-HP-Notebook:/home/vgoyal/Desktop/hadoop-2.7.7# jps
11312 DataNode
11857 NodeManager
12168 Jps
11131 NameNode
11710 ResourceManager
11535 SecondaryNameNode
```

Note: If ssh to localhost is not successful run:
ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys

Errors occuring during the installation and setting up of Single node:

1. **JAVA_HOME not found even after being set**

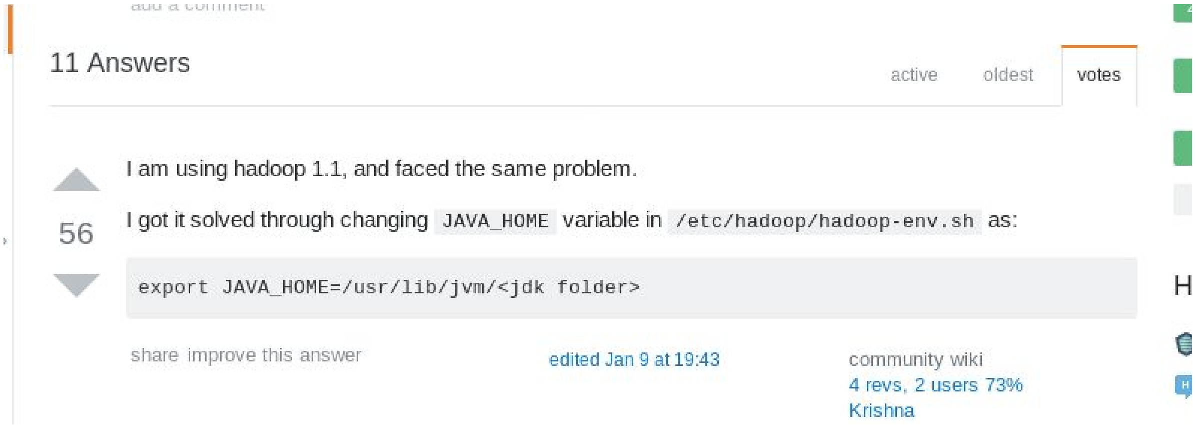


```
        [-finalize] |
        [-importCheckpoint] |
        [-initializeSharedEdits] |
        [-bootstrapStandby] |
        [-recover [ -force] ] |
        [-metadataVersion ]  ]

19/09/05 23:11:53 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at vgoyal-HP-Notebook/127.0.1.1
************************************************************/
root@vgoyal-HP-Notebook:/home/vgoyal/Desktop/hadoop-2.7.7# sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: Error: JAVA_HOME is not set and could not be found.
localhost: Error: JAVA_HOME is not set and could not be found.
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:oWIeWOQzMsGl8rUsfejRrQwgcZyHRVIlqSKL7RyUMpY.
Are you sure you want to continue connecting (yes/no)? ^C0.0.0.0: Host key verif
ication failed.

root@vgoyal-HP-Notebook:/home/vgoyal/Desktop/hadoop-2.7.7# echo $JAVA_HOME
/usr/lib/jvm/java-8-oracle
root@vgoyal-HP-Notebook:/home/vgoyal/Desktop/hadoop-2.7.7# 
```

Solution: Set the variable manually in Hadoop-env.sh and in the bash script



I am using hadoop 1.1, and faced the same problem.

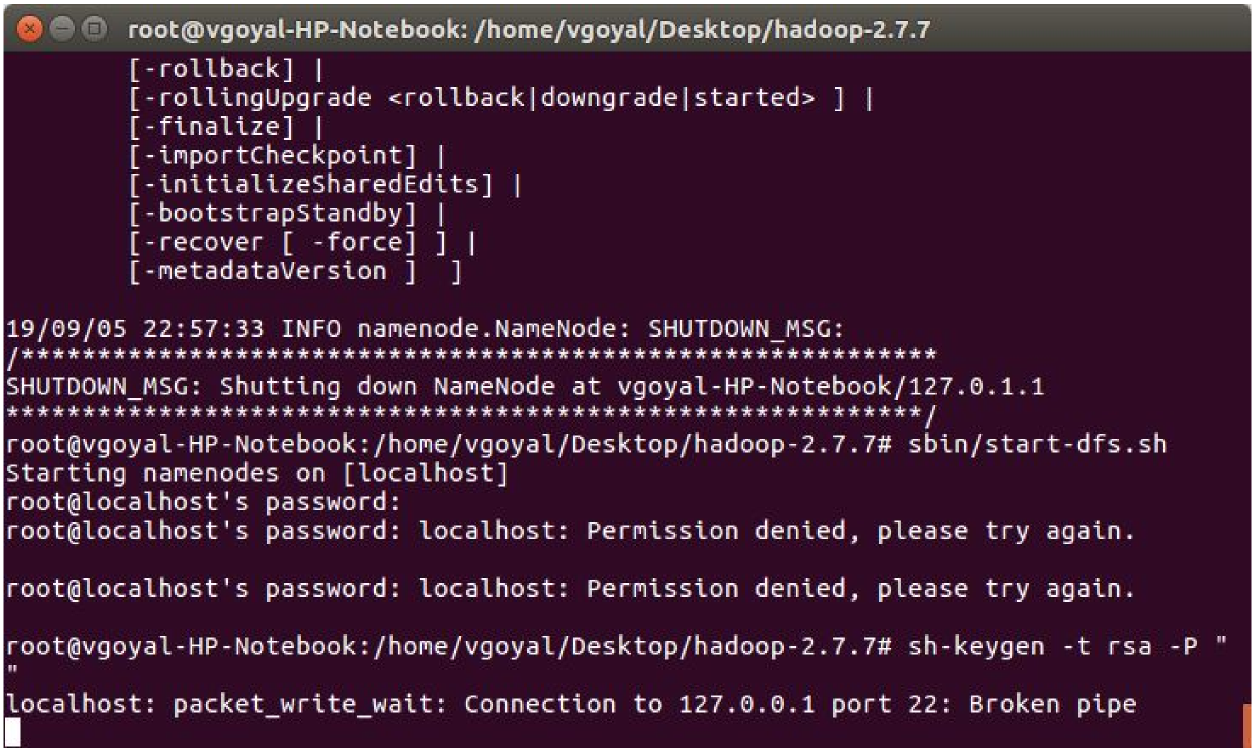
I got it solved through changing `JAVA_HOME` variable in `/etc/hadoop/hadoop-env.sh` as:

```
export JAVA_HOME=/usr/lib/jvm/<jdk folder>
```

edited Jan 9 at 19:43

community wiki
4 revs, 2 users 73%
Krishna

2. **Localhost permission denied**



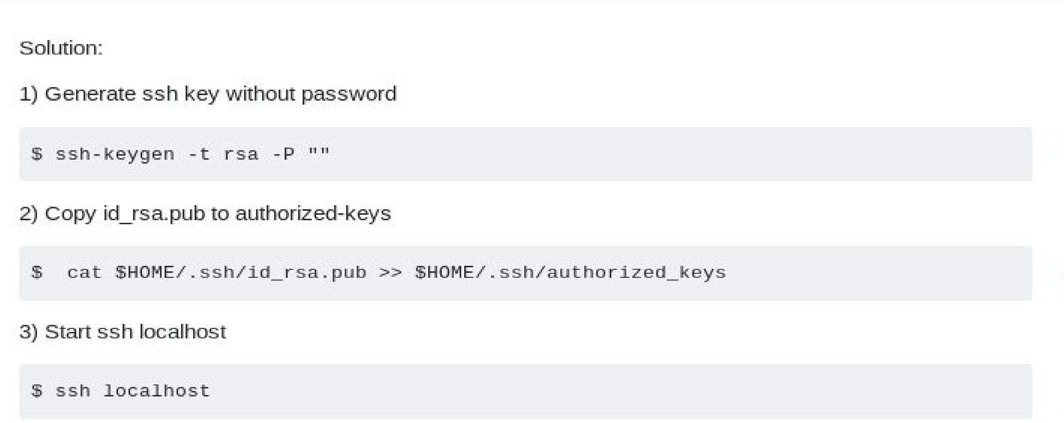
```
        [-rollback] |
        [-rollingUpgrade <rollback|downgrade|started> ] |
        [-finalize] |
        [-importCheckpoint] |
        [-initializeSharedEdits] |
        [-bootstrapStandby] |
        [-recover [ -force] ] |
        [-metadataVersion ]  ]

19/09/05 22:57:33 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at vgoyal-HP-Notebook/127.0.1.1
************************************************************/
root@vgoyal-HP-Notebook:/home/vgoyal/Desktop/hadoop-2.7.7# sbin/start-dfs.sh
Starting namenodes on [localhost]
root@localhost's password:
root@localhost's password: localhost: Permission denied, please try again.

root@localhost's password: localhost: Permission denied, please try again.

root@vgoyal-HP-Notebook:/home/vgoyal/Desktop/hadoop-2.7.7# sh-keygen -t rsa -P "
"
localhost: packet_write_wait: Connection to 127.0.0.1 port 22: Broken pipe
```

Solution:
Generate ssh key without password and append it to ida_rsa.pub



Solution:

1) Generate ssh key without password

```
$ ssh-keygen -t rsa -P ""
```

2) Copy id_rsa.pub to authorized-keys

```
$  cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

3) Start ssh localhost

```
$ ssh localhost
```

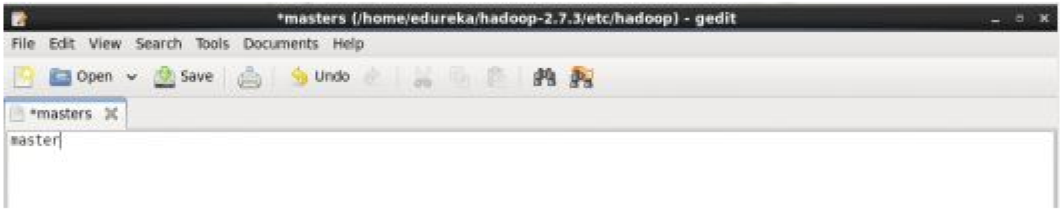
2. Setting up Multi-node system

Tutorial followed was :
https://www.edureka.co/blog/setting-up-a-multi-node-cluster-in-hadoop-2-x/

Due to error in connection (ssh) between 2 seperate systems, we created 2 virtual machines to create the multi-node system (Master & Slave)

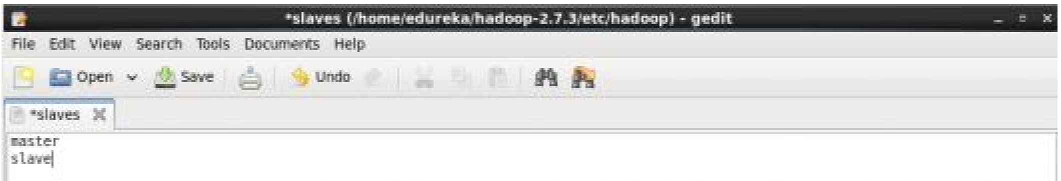
Steps followed:

1. Check IP addresses of both master and slave system
2. Disable the firewall restrictions. - **sudo ufw disable**
3. Open hosts file to add master and data node with their respective IP addresses. - **nano /etc/hosts**
4. Restart the sshd service.
5. Create the SSH Key in the master node.
6. Copy the generated ssh key to master node’s authorized keys.
7. Copy the master node’s ssh key to slave’s authorized keys. - **ssh-copy-id -i $HOME/.ssh/id_rsa.pub osboxes@slave’s_ip**
8. Create masters file and edit as follows in both master and slave machines as below:


*masters (/home/edureka/hadoop-2.7.3/etc/hadoop) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*masters

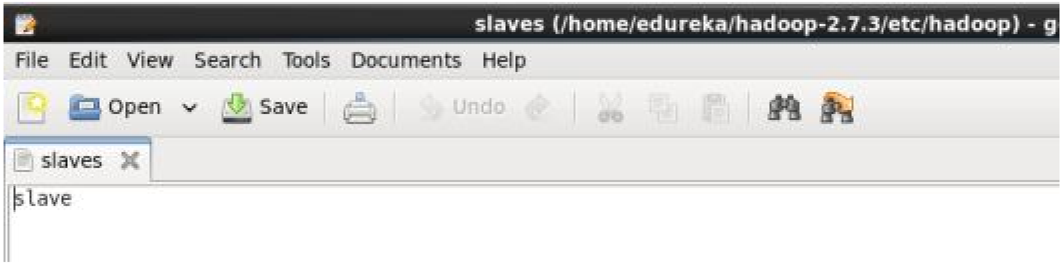
```
master
```

9. Edit slaves file in master machine as follows:


*slaves (/home/edureka/hadoop-2.7.3/etc/hadoop) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*slaves

```
master
slave
```

10. Edit slaves file in slave machine as follows:


slaves (/home/edureka/hadoop-2.7.3/etc/hadoop) - g
File Edit View Search Tools Documents Help
Open
Save
Undo
slaves

```
slave
```

11. Edit core-site.xml on both master and slave machines

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://master:9000</value>
</property>
</configuration>
```

12. Edit hdfs-site.xml on master

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/edureka/hadoop-2.7.3/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/edureka/hadoop-2.7.3/datanode</value>
</property>
</configuration>
```

13. Edit hdfs-site.xml on slave machine

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/edureka/hadoop-2.7.3/datanode</value>
</property>
</configuration>
```

14. Copy mapred-site from the template in configuration folder and the edit mapred-site.xml on both master and slave machines

```
<?xml version="1.0" encoding="UTF-8"?>
```

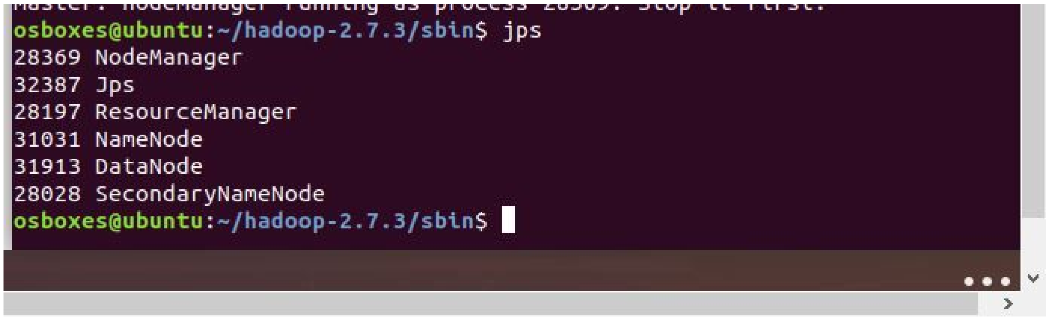
```
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

15. Edit yarn-site.xml on both master and slave machines

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

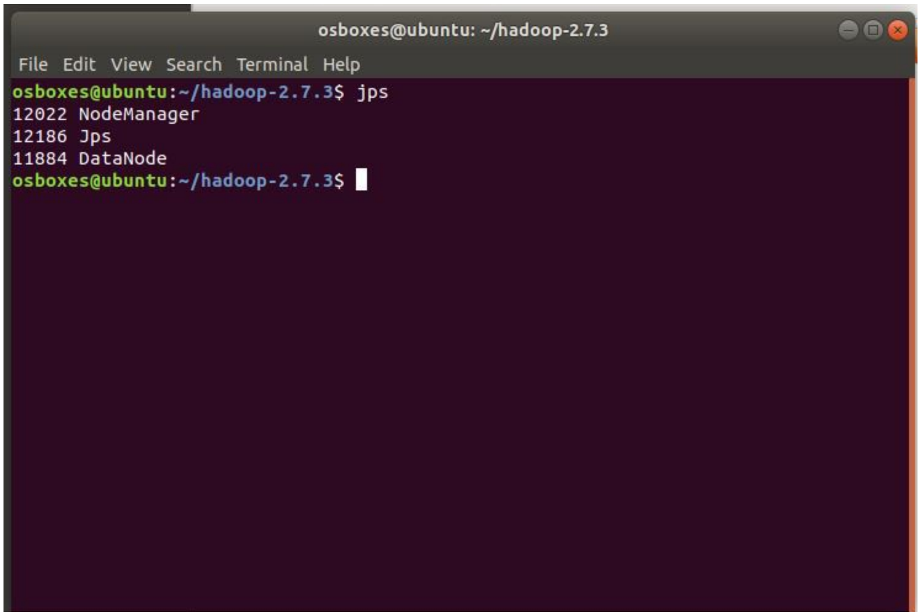
16. Format the namenode (Only on master machine).
17. Start all daemons (Only on master machine).
18. Check all the daemons running on both master and slave machines.

Master node:

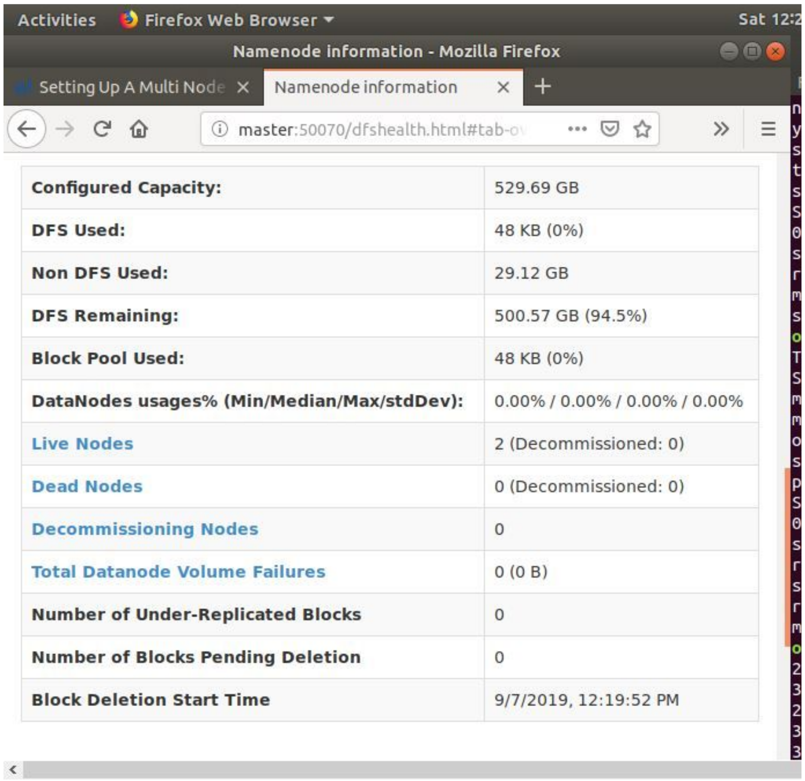


```
master: nodemanager running as process 28369. Stop it first.
osboxes@ubuntu:~/hadoop-2.7.3/sbin$ jps
28369 NodeManager
32387 Jps
28197 ResourceManager
31031 NameNode
31913 DataNode
28028 SecondaryNameNode
osboxes@ubuntu:~/hadoop-2.7.3/sbin$
```

Slave node:


osboxes@ubuntu: ~/hadoop-2.7.3
File Edit View Search Terminal Help

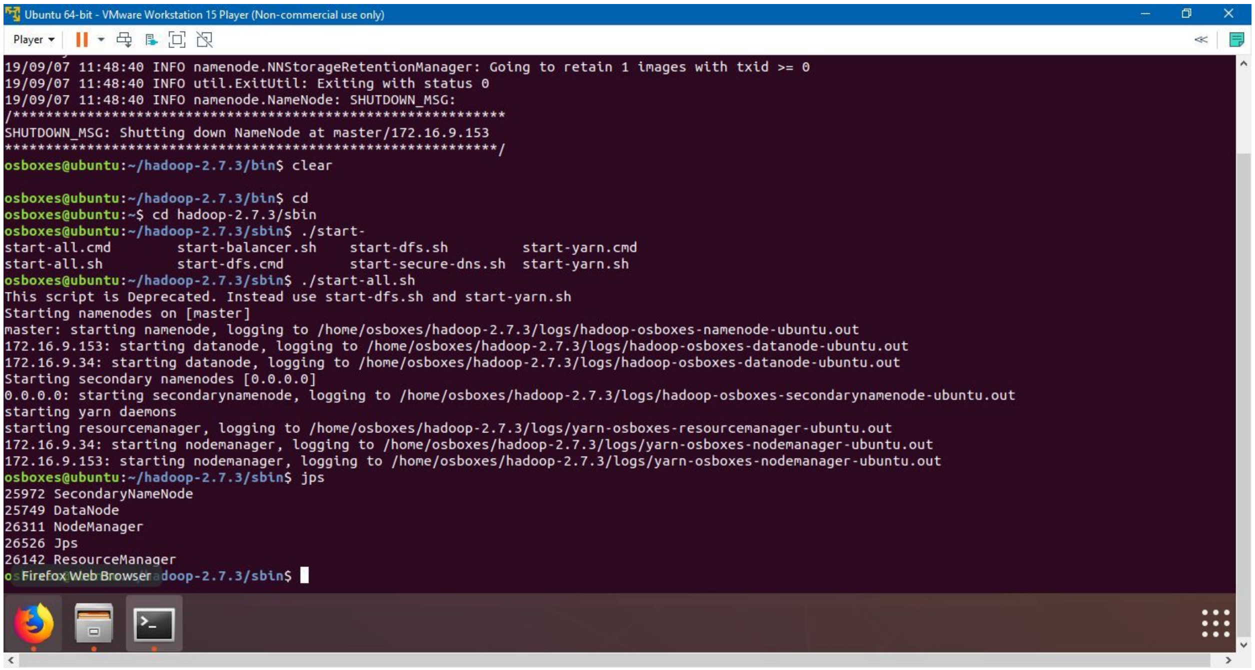
```
osboxes@ubuntu:~/hadoop-2.7.3$ jps
12022 NodeManager
12186 Jps
11884 DataNode
osboxes@ubuntu:~/hadoop-2.7.3$
```



Configured Capacity:	529.69 GB
DFS Used:	48 KB (0%)
Non DFS Used:	29.12 GB
DFS Remaining:	500.57 GB (94.5%)
Block Pool Used:	48 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	2 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	9/7/2019, 12:19:52 PM

Errors during setup of Multi-node cluster

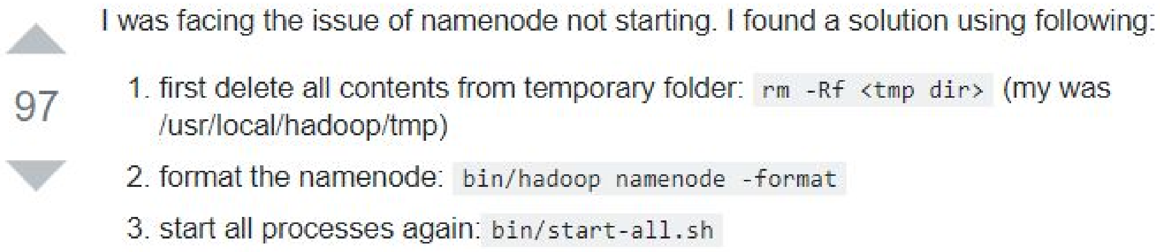
1. **Namenode /Datanode on master / Datanode on slave not found**


Ubuntu 64-bit - VMware Workstation 15 Player (Non-commercial use only)
Player
Firefox Web Browser

```
19/09/07 11:48:40 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
19/09/07 11:48:40 INFO util.ExitUtil: Exiting with status 0
19/09/07 11:48:40 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at master/172.16.9.153
************************************************************/
osboxes@ubuntu:~/hadoop-2.7.3/bin$ clear

osboxes@ubuntu:~/hadoop-2.7.3/bin$ cd
osboxes@ubuntu:~$ cd hadoop-2.7.3/sbin
osboxes@ubuntu:~/hadoop-2.7.3/sbin$ ./start-
start-all.cmd        start-balancer.sh    start-dfs.sh         start-yarn.cmd
start-all.sh         start-dfs.cmd        start-secure-dns.sh  start-yarn.sh
osboxes@ubuntu:~/hadoop-2.7.3/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/osboxes/hadoop-2.7.3/logs/hadoop-osboxes-namenode-ubuntu.out
172.16.9.153: starting datanode, logging to /home/osboxes/hadoop-2.7.3/logs/hadoop-osboxes-datanode-ubuntu.out
172.16.9.34: starting datanode, logging to /home/osboxes/hadoop-2.7.3/logs/hadoop-osboxes-datanode-ubuntu.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/osboxes/hadoop-2.7.3/logs/hadoop-osboxes-secondarynamenode-ubuntu.out
starting yarn daemons
starting resourcemanager, logging to /home/osboxes/hadoop-2.7.3/logs/yarn-osboxes-resourcemanager-ubuntu.out
172.16.9.34: starting nodemanager, logging to /home/osboxes/hadoop-2.7.3/logs/yarn-osboxes-nodemanager-ubuntu.out
172.16.9.153: starting nodemanager, logging to /home/osboxes/hadoop-2.7.3/logs/yarn-osboxes-nodemanager-ubuntu.out
osboxes@ubuntu:~/hadoop-2.7.3/sbin$ jps
25972 SecondaryNameNode
25749 DataNode
26311 NodeManager
26526 Jps
26142 ResourceManager
o                    doop-2.7.3/sbin$
```

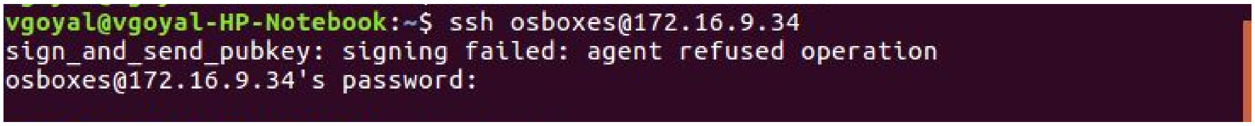
Solution:



I was facing the issue of namenode not starting. I found a solution using following:

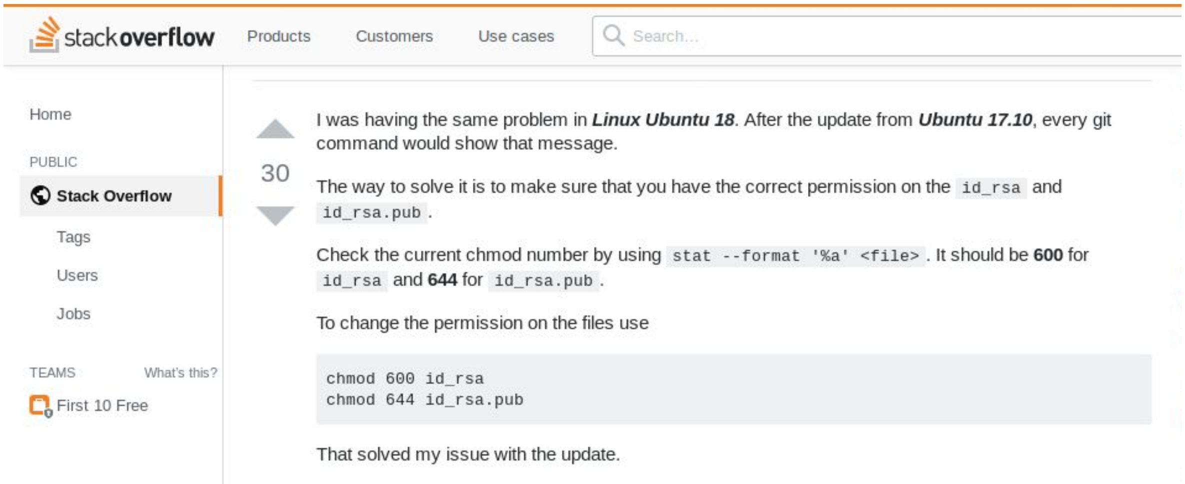
1. first delete all contents from temporary folder: `rm -Rf <tmp dir>` (my was /usr/local/hadoop/tmp)
2. format the namenode: `bin/hadoop namenode -format`
3. start all processes again: `bin/start-all.sh`

2. SSH - agent refused connection



```
vgoyal@vgoyal-HP-Notebook:~$ ssh osboxes@172.16.9.34
sign_and_send_pubkey: signing failed: agent refused operation
osboxes@172.16.9.34's password:
```

Solution:



I was having the same problem in ***Linux Ubuntu 18***. After the update from ***Ubuntu 17.10***, every git command would show that message.

The way to solve it is to make sure that you have the correct permission on the `id_rsa` and `id_rsa.pub` .

Check the current chmod number by using `stat --format '%a' <file>` . It should be **600** for `id_rsa` and **644** for `id_rsa.pub` .

To change the permission on the files use

```
chmod 600 id_rsa
chmod 644 id_rsa.pub
```

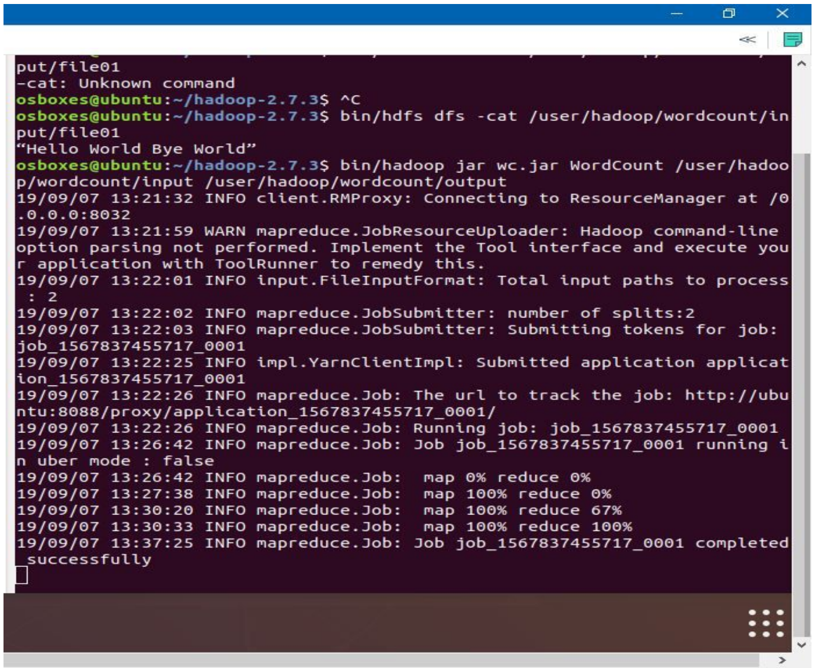
That solved my issue with the update.

3. Map-reduce to find word-count:

Steps followed:

1. Make local repository
2. Write a java file with code in it
3. Copy the repository in hadoop repository
4. Compile the code and create a jar file
5. Create directories in Hdfs folder
6. Create input files
7. Execute the jar files

Job running successfully



```
put/file01
-cat: Unknown command
osboxes@ubuntu:~/hadoop-2.7.3$ ^C
osboxes@ubuntu:~/hadoop-2.7.3$ bin/hdfs dfs -cat /user/hadoop/wordcount/in
put/file01
“Hello World Bye World”
osboxes@ubuntu:~/hadoop-2.7.3$ bin/hadoop jar wc.jar WordCount /user/hadoo
p/wordcount/input /user/hadoop/wordcount/output
19/09/07 13:21:32 INFO client.RMProxy: Connecting to ResourceManager at /0
.0.0.0:8032
19/09/07 13:21:59 WARN mapreduce.JobResourceUploader: Hadoop command-line
option parsing not performed. Implement the Tool interface and execute you
r application with ToolRunner to remedy this.
19/09/07 13:22:01 INFO input.FileInputFormat: Total input paths to process
 : 2
19/09/07 13:22:02 INFO mapreduce.JobSubmitter: number of splits:2
19/09/07 13:22:03 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1567837455717_0001
19/09/07 13:22:25 INFO impl.YarnClientImpl: Submitted application applicat
ion_1567837455717_0001
19/09/07 13:22:26 INFO mapreduce.Job: The url to track the job: http://ubu
ntu:8088/proxy/application_1567837455717_0001/
19/09/07 13:22:26 INFO mapreduce.Job: Running job: job_1567837455717_0001
19/09/07 13:26:42 INFO mapreduce.Job: Job job_1567837455717_0001 running i
n uber mode : false
19/09/07 13:26:42 INFO mapreduce.Job:  map 0% reduce 0%
19/09/07 13:27:38 INFO mapreduce.Job:  map 100% reduce 0%
19/09/07 13:30:20 INFO mapreduce.Job:  map 100% reduce 67%
19/09/07 13:30:33 INFO mapreduce.Job:  map 100% reduce 100%
19/09/07 13:37:25 INFO mapreduce.Job: Job job_1567837455717_0001 completed
 successfully
```