Installing Hadoop in Ubuntu

1. To update latest distributions of software:

- *sudo apt-get update*

2. Hadoop requires Java 8 to be installed. To install java version 8:

- *sudo apt-get install openssh-server*
- *sudo apt-get install openjdk-8-jdk*

3. Verifying JAVA 8 Installation:

- *java -version*

```
student@student-VirtualBox:~$ java -version
openjdk version "1.8.0_191"
OpenJDK Runtime Environment (build 1.8.0_191-8u191-b12-2ubuntu0.18.04.1-b12)
OpenJDK 64-Bit Server VM (build 25.191-b12, mixed mode)
```

4. Download and extract Hadoop:

- *wget https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz*

5. Unzip the Hadoop file:

- *tar -xzf hadoop-2.7.7.tar.gz*

```
student@student-VirtualBox:~$ hadoop version
Hadoop 2.7.7
Subversion Unknown -r c1aad84bd27cd79c3d1a7dd58202a8c3ee1ed3ac
Compiled by stevel on 2018-07-18T22:47Z
Compiled with protoc 2.5.0
From source with checksum 792e15d20b12c74bd6f19a1fb886490
This command was run using /home/WQD7007/hadoop/share/hadoop/common/hadoop-commo
n-2.7.7.jar
student@student-VirtualBox:~$
```

6. Move the file to your own directory:

- *sudo mkdir /home/{yourname}*
- *sudo mv hadoop-2.7.7 /home/{yourname}/hadoop/*

7. Set JAVA_HOME in /home/{yourname}/hadoop/etc/hadoop/hadoop-env.sh:

- *nano /home/{yourname}/hadoop/etc/hadoop/hadoop-env.sh*

```
student@student-VirtualBox: /home/WQD7007/hadoop/etc/hadoop
File  Edit  View  Search  Terminal  Help
  GNU nano 2.9.3                    hadoop-env.sh

# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
# export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre/
```

8. Export path for Hadoop in ~/.bashr to allow easy Hadoop function access:

- *export PATH=$PATH:/home/{yourname}/hadoop/bin*
- *export PATH=$PATH:/home/{yourname}/hadoop/sbin*

```
  GNU nano 2.9.3                    /home/student/.bashrc

if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export PATH=$PATH:/home/WQD7007/hadoop/bin
export PATH=$PATH:/home/WQD7007/hadoop/sbin
```

9. Run *hadoop*:

```
student@student-VirtualBox:/$ hadoop
Usage: hadoop [--config confdir] [COMMAND | CLASSNAME]
  CLASSNAME            run the class named CLASSNAME
 or
  where COMMAND is one of:
  fs                   run a generic filesystem user client
  version              print the version
  jar <jar>            run a jar file
                       note: please use "yarn jar" to launch
                             YARN applications, not this command.
  checknative [-a|-h]  check native hadoop and compression libraries availabilit
y
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archi
ve
  classpath            prints the class path needed to get the
  credential           interact with credential providers
                       Hadoop jar and the required libraries
  daemonlog            get/set the log level for each daemon
  trace                view and modify Hadoop tracing settings

Most commands print help when invoked w/o parameters.
```

10. Update hdfs-site.xml in /home/{yourname}/hadoop/etc/hadoop folder using nano. This file contains the configuration properties that Hadoop uses when starting up. Save and close this file.

- *nano /home/{yourname}/hadoop/etc/Hadoop/hdfs-site.xml*

```
  GNU nano 2.9.3                    hdfs-site.xml

  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.namenode.name.dir</name>
                <value>/home/WQD7007/hadoop/data/nameNode</value>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>/home/WQD7007/hadoop/data/dataNode</value>
        </property>
        <property>
                <name>dfs.replication</name>
```

11. Update core-site.xml in /home/{yourname}/hadoop/etc/hadoop using nano. You can change 'localhost' to your PC's IP address. Save and close this file.

- *nano /home/{yourname}/hadoop/etc/Hadoop/core-site.xml*

```
  GNU nano 2.9.3                    core-site.xml

  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>
```

12. Rename/copy the mapred-site.xml.template in /home/{yourname}/hadoop/etc/hadoop to mapred-site.xml. This file is used to specify which framework is being used for MapReduce. Then, update mapred-site.xml. Save and close the file.

- *cp /home/{yourname}/hadoop/etc/hadoop/mapred-site.xml.template /home/{yourname}/hadoop/etc/hadoop/mapred-site.xml*
- *nano /home/{yourname}/hadoop/etc/hadoop/mapred-site.xml*

```
File  Edit  View  Search  Terminal  Help
  GNU nano 2.9.3                                                    mapred-site.xml

  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>
        <property>
                <name>yarn.app.mapreduce.am.resource.mb</name>
                <value>512</value>
        </property>
        <property>
                <name>mapreduce.map.memory.mb</name>
                <value>256</value>
        </property>
        <property>
                <name>mapreduce.reduce.memory.mb</name>
                <value>256</value>
        </property>
</configuration>
```

13. Update yarn-site.xml in /home/{yourname}/hadoop/etc/hadoop folder using nano. This file contains the configuration properties that Mapreduce uses when starting up. Save and close this file.

- *nano /home/{yourname}/hadoop/etc/Hadoop/yarn-site.xml*

```
  GNU nano 2.9.3                                                    yarn-site.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>

<!-- Site specific YARN configuration properties -->
        <property>
                <name>yarn.acl.enable</name>
                <value>0</value>
        </property>
        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>localhost</value>
        </property>
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
        <property>
                <name>yarn.nodemanager.resources.memory-mb</name>
                <value>1536</value>
                </property>
        <property>
                <name>yarn.scheduler.maximum -allocation-mb</name>
                <value>1536</value>
        </property>
        <property>
                <name>yarn-scheduler.minimum-allocation-mb</name>
                <value>128</value>
        </property>
```

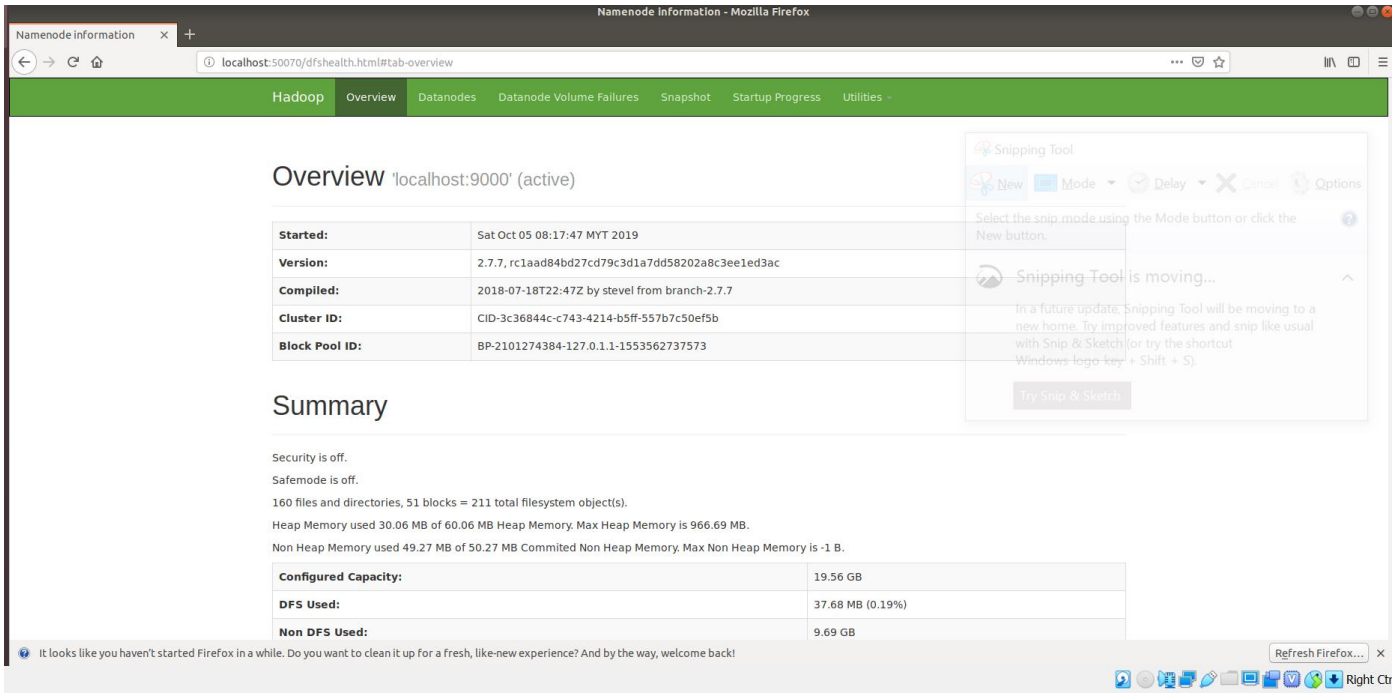
14. Run *hdfs namenode -format.*

```
19/10/12 23:27:26 INFO util.GSet: capacity      = 2^21 = 2097152 entries
19/10/12 23:27:26 INFO blockmanagement.BlockManager: dfs.block.access.token.enable=false
19/10/12 23:27:26 INFO blockmanagement.BlockManager: defaultReplication         = 1
19/10/12 23:27:26 INFO blockmanagement.BlockManager: maxReplication             = 512
19/10/12 23:27:26 INFO blockmanagement.BlockManager: minReplication             = 1
19/10/12 23:27:26 INFO blockmanagement.BlockManager: maxReplicationStreams      = 2
19/10/12 23:27:26 INFO blockmanagement.BlockManager: replicationRecheckInterval = 3000
19/10/12 23:27:26 INFO blockmanagement.BlockManager: encryptDataTransfer        = false
19/10/12 23:27:26 INFO blockmanagement.BlockManager: maxNumBlocksToLog          = 1000
19/10/12 23:27:26 INFO namenode.FSNamesystem: fsOwner             = student (auth:SIMPLE)
19/10/12 23:27:26 INFO namenode.FSNamesystem: supergroup          = supergroup
19/10/12 23:27:26 INFO namenode.FSNamesystem: isPermissionEnabled = true
19/10/12 23:27:26 INFO namenode.FSNamesystem: HA Enabled: false
19/10/12 23:27:26 INFO namenode.FSNamesystem: Append Enabled: true
19/10/12 23:27:26 INFO util.GSet: Computing capacity for map INodeMap
19/10/12 23:27:26 INFO util.GSet: VM type       = 64-bit
19/10/12 23:27:26 INFO util.GSet: 1.0% max memory 966.7 MB = 9.7 MB
19/10/12 23:27:26 INFO util.GSet: capacity      = 2^20 = 1048576 entries
19/10/12 23:27:26 INFO namenode.FSDirectory: ACLs enabled? false
19/10/12 23:27:26 INFO namenode.FSDirectory: XAttrs enabled? true
19/10/12 23:27:26 INFO namenode.FSDirectory: Maximum size of an xattr: 16384
19/10/12 23:27:26 INFO namenode.NameNode: Caching file names occuring more than 10 times
19/10/12 23:27:26 INFO util.GSet: Computing capacity for map cachedBlocks
19/10/12 23:27:26 INFO util.GSet: VM type       = 64-bit
19/10/12 23:27:26 INFO util.GSet: 0.25% max memory 966.7 MB = 2.4 MB
19/10/12 23:27:26 INFO util.GSet: capacity      = 2^18 = 262144 entries
19/10/12 23:27:26 INFO namenode.FSNamesystem: dfs.namenode.safemode.threshold-pct = 0.9990000128746033
19/10/12 23:27:26 INFO namenode.FSNamesystem: dfs.namenode.safemode.min.datanodes = 0
19/10/12 23:27:26 INFO namenode.FSNamesystem: dfs.namenode.safemode.extension     = 30000
19/10/12 23:27:26 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
19/10/12 23:27:26 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
19/10/12 23:27:26 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
19/10/12 23:27:26 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
19/10/12 23:27:26 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
19/10/12 23:27:26 INFO util.GSet: Computing capacity for map NameNodeRetryCache
19/10/12 23:27:26 INFO util.GSet: VM type       = 64-bit
19/10/12 23:27:26 INFO util.GSet: 0.029999999329447746% max memory 966.7 MB = 297.0 KB
19/10/12 23:27:26 INFO util.GSet: capacity      = 2^15 = 32768 entries
Re-format filesystem in Storage Directory /home/WQD7007/hadoop/data/nameNode ? (Y or N) N
Format aborted in Storage Directory /home/WQD7007/hadoop/data/nameNode
19/10/12 23:27:35 INFO util.ExitUtil: Exiting with status 1
19/10/12 23:27:35 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at student-VirtualBox/127.0.1.1
************************************************************/
student@student-VirtualBox:/$
```

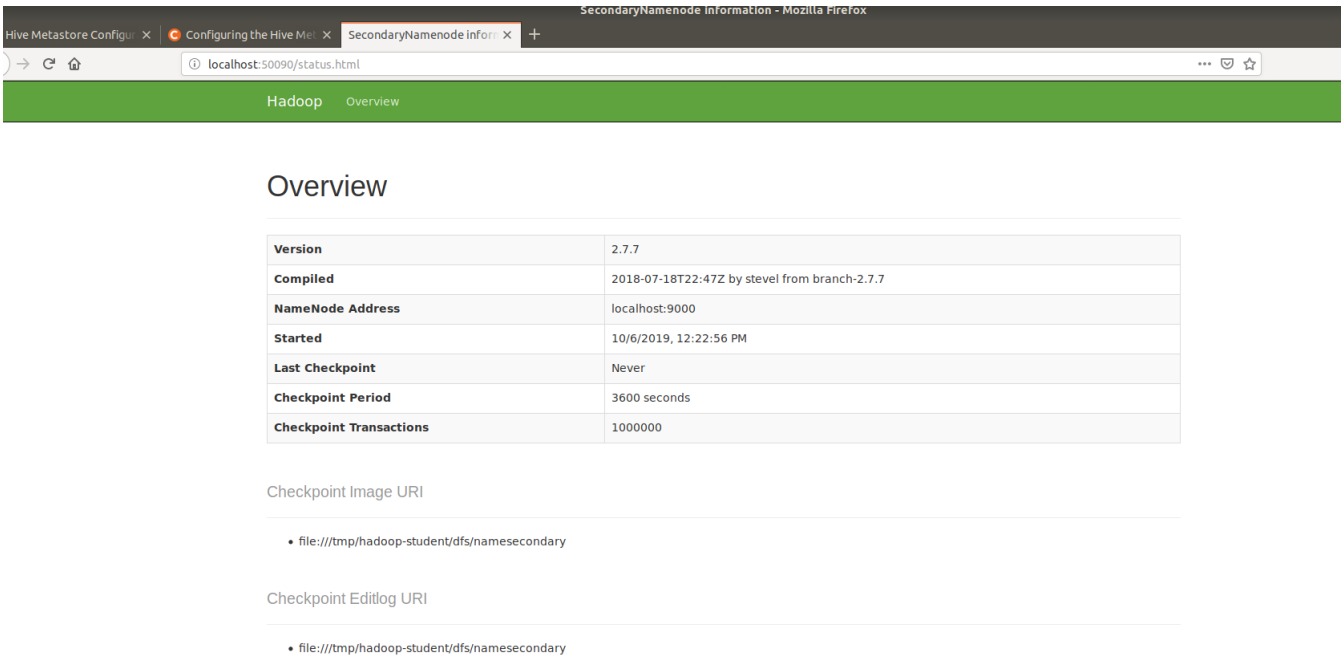
15. Run *start-all.sh* (or run *start-dfs.sh* and *start-yarn.sh* separately).

```
student@student-VirtualBox:/$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
student@localhost's password:
localhost: starting namenode, logging to /home/WQD7007/hadoop/logs/hadoop-student-namenode-student-VirtualBox.out
student@localhost's password:
localhost: datanode running as process 1956. Stop it first.
Starting secondary namenodes [0.0.0.0]
student@0.0.0.0's password:
0.0.0.0: secondarynamenode running as process 2168. Stop it first.
starting yarn daemons
resourcemanager running as process 2324. Stop it first.
student@localhost's password:
localhost: nodemanager running as process 2605. Stop it first.
```

16. Browse localhost:50070 in your browser for namenode:



17. Browse localhost:50090 in your browser for secondary namenode:



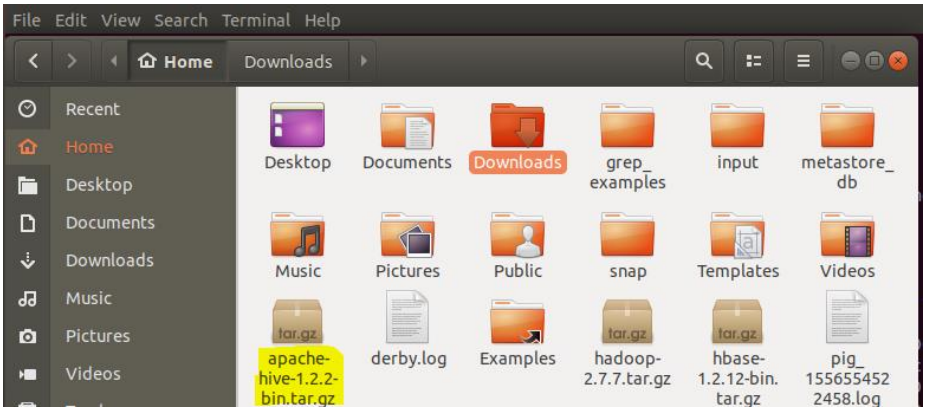
17. You can run *jps* to show all the packages installed in your terminal.

```
student@student-VirtualBox:~$ jps
3664 HRegionServer
3760 JobHistoryServer
3488 HQuorumPeer
2146 NameNode
2771 ResourceManager
2935 NodeManager
5066 Jps
2589 SecondaryNameNode
3551 HMaster
2319 DataNode
```

Installing Hive in Ubuntu

1. Download and install hive using:

- *wget https://www.apache.org/dist/hive/hive-1.2.2/apache-hive-1.2.2-bin.tar.gz*



2. Unzip the hive folder:

- *tar -xzf apache-hive-1.2.2-bin.tar.gz*

3. Move the hive folder to your own directory:

- *mv apache-hive-1.2.2-bin /home/{yourname}/hive/*

4. In *nano ~/.bashrc*, set export PATH=$PATH:/home/{yourname}/hive/bin

```
File  Edit  View  Search  Terminal  Help
  GNU nano 2.9.3                                        /home/student/.bashrc

fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export PATH=$PATH:/home/WQD7007/hadoop/bin
export PATH=$PATH:/home/WQD7007/hadoop/sbin

export PATH=$PATH:/home/WQD7007/hbase/bin

export PATH=$PATH:/home/WQD7007/hive/bin
```

5. In hive bin folder, *nano hive-config.sh* and add export HADOOP_HOME=/home/wlhoo/hadoop at the end of the file to connect hive with Hadoop.

```
File  Edit  View  Search  Terminal  Help
  GNU nano 2.9.3                                                    hive-config.sh

while [ $# -gt 0 ]; do    # Until you run out of parameters . . .
  case "$1" in
    --config)
        shift
        confdir=$1
        shift
        HIVE_CONF_DIR=$confdir
        ;;
    --auxpath)
        shift
        HIVE_AUX_JARS_PATH=$1
        shift
        ;;
    *)
        break;
        ;;
  esac
done


# Allow alternate conf dir location.
HIVE_CONF_DIR="${HIVE_CONF_DIR:-$HIVE_HOME/conf}"

export HIVE_CONF_DIR=$HIVE_CONF_DIR
export HIVE_AUX_JARS_PATH=$HIVE_AUX_JARS_PATH

# Default to use 256MB
export HADOOP_HEAPSIZE=${HADOOP_HEAPSIZE:-256}

export HADOOP_HOME=/home/WQD7007/hadoop
```

6. Run Hive and hive is successfully installed.

```
student@student-VirtualBox:~$ hive
ls: cannot access '/home/WQD7007/spark/lib/spark-assembly-*.jar': No such file o
r directory

Logging initialized using configuration in jar:file:/home/WQD7007/hive/lib/hive-
common-1.2.2.jar!/hive-log4j.properties
hive>
```