**Ans 1 - Setup for HDFS in linux**

a. Install java. I already had java installed in my laptop so not writing its step.
b. To install hadoop we have to create a user and for that go to system preferences User and group then unlock user and give password and then click on plus button and give user name hadoop and password

c . Now this user should be able to login remotely so for this we will go to system preference and go to sharing and click on remote login and add hadoop in only these user.

login with hadoop user

```
su - hadoop
password of user
```

d. Now do key generation

```
ssh-keygen -t rsa -P hadoop
```

It will ask for key generation path. I gave default path i.e. /Users/hadoop/.ssh/id_rsa

Output while performing operations

```
Generating public/private rsa key pair.
Enter file in which to save the key (/Users/hadoop/.ssh/id_rsa):
Created directory '/Users/hadoop/.ssh'.
Your identification has been saved in /Users/hadoop/.ssh/id_rsa.
Your public key has been saved in /Users/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:/Z7SgZepQtK3i4O+7HhW/0RnBbigpcERzCbEk39SybM
hadoop@Aditis-MacBook-Air.local
The key's randomart image is:
+---[RSA 2048]----+
|          o.=o+ . ..  |
|          = * B .  . |
|          = * + .  .|
|          =.E .  . |
|          .So...oo  |
|          . + oo=o   |
|          = o =o.   |
|          o+ o.=o..   |
|          .==..o.++         |
+----[SHA256]-----+
```

Aditis-MacBook-Air:~ hadoop$ cd .ssh
Aditis-MacBook-Air:.ssh hadoop$ ls -lrt
total 16

-rw-r--r--  1 hadoop  staff   413 Jan 26 14:55 id_rsa.pub
-rw-------  1 hadoop  staff  1766 Jan 26 14:55 id_rsa
 Now will copy content of id_rsa.pub to authorized_keys
cat /Users/hadoop/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
This key is generated so that only authorized user can acces it using

$ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ECDSA key fingerprint is SHA256:4AkQadf4pr7J+xwKZnhMB4WtmF9LOouCNBa07zxhIV0.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Enter passphrase for key '/Users/hadoop/.ssh/id_rsa':
passphrase will be password given while genrating key i.e. hadoop

E. Now install hadoop for doing that use below command

```
curl -O http://www.eu.apache.org/dist/hadoop/common/hadoop-2.7.4/hadoop-2.7.4-src.tar.gz
tar -xzvf hadoop-2.7.4-src.tar.gz
#Give owner rights to hadoop
chown -R hadoop hadoop-*
cd hadoop*
ls -lsrt
```

conf folder will be there in hadoop-mapreduce-project/conf

Then I edited files:
1.hadoop-common-project/hadoop-common/src/main/conf/hadoop-env.sh
edit JAVA_HOME with correct java path

Can see Java home directory from command
/usr/libexec/java_home
export JAVA_HOME={}(path of java home directory)
export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_121.jdk/Contents/Home

2.hadoop-common-project/hadoop-common/src/main/conf/core-site.xml

<configuration>
<property>
<name>hadoop.temp.dir</name>
<value>/Users/hadoop/hdfstmp</value>
<description>This is a bas for other temporary directory</description>
</property>

<property>
<name>fs.default.name</name>

<value>hdfs://localhost:54310</value>
<description>The name of default system</description>
</property>
</configuration>

3. hadoop-hdfs-project/hadoop-hdfs/src/main/conf/hdfs-site.xml

<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
<description> Default file replication. You can change replication at file level when file is created</description>
</property>
</configuration>

4../hadoop-mapreduce-project/conf/mapred-site.xml.template

<configuration>
<property>
<name>mapred.job.tracker</name>
<value>ocalhost:54311</value>
<description>The host and port of mapred job tracker</description>
</property>
</configuration>

This was my wrong steps.. I had taken src.tar file which I later realized was the wrong file. So after facing many issues I downloaded “hadoop-2.7.4.tar.gz” and repeated the steps

```
curl -O http://www.eu.apache.org/dist/hadoop/common/hadoop-2.7.4/hadoop-2.7.4.tar.gz
tar -xzvf hadoop-2.7.4-src.tar.gz
#Give owner rights to hadoop
chown -R hadoop hadoop-*
cd hadoop*
ls -lsrt
```

I deleted previous hadoop-2.7.4-src folder. And followed below steps.

```
Edited below files:
1.vi etc/hadoop/hadoop-env.sh
edit JAVA_HOME with correct java path
```

Can see Java home directory from command
/usr/libexec/java_home
export JAVA_HOME={}(path of java home directory)
export JAVA_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0_121.jdk/Contents/Home

2.vi etc/hadoop/core-site.xml

```
<configuration>
<property>
<name>hadoop.temp.dir</name>
<value>/Users/hadoop/hdfstmp</value>
<description>This is a bas for other temporary directory</description>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:54310</value>
<description> The name of default system</description>
</property>
</configuration>
```

#If fs.default is not given node doesn't starts.

3. vi etc/hadoop/hdfs-site.xml

```
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
<description> Default file replication. You can change replication at file level when file is
created</description>
</property>
</configuration>
```

4.vi etc/hadoop/mapred-site.xml.template

```
<configuration>
<property>
<name>mapred.job.tracker</name>
<value>localhost:54311</value>
<description>The host and port of mapred job tracker</description>
</property>
</configuration>
```

cp mapred-site.xml.template mapred-site.xml

cd ../..
bin/hadoop namenode -format
sbin/start-all.sh
#give rsa passphrase when ask
Jps # to see list of processes running in the JVM

```
9456 ResourceManager
9570 Jps
9351 SecondaryNameNode
9257 DataNode
9183 NameNode
```

bin/hadoop fs -ls / was giving warning
WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
fixed by setting path
I searched a lot but nothing was resolving my problem then I ignored this warning as HDFS was working after making many changes even with warning.

Setted variable

```
export HADOOP_HOME=/Users/hadoop/hadoop-2.7.4
export PATH=$HADOOP_HOME/bin:$PATH
export HADOOP_PREFIX=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_PREFIX
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_PREFIX/lib/native
export HADOOP_CONF_DIR=$HADOOP_PREFIX/etc/hadoop
export HADOOP_HDFS_HOME=$HADOOP_PREFIX
export HADOOP_MAPRED_HOME=$HADOOP_PREFIX
export HADOOP_YARN_HOME=$HADOOP_PREFIX
export JAVA_LIBRARY_PATH=$HADOOP_HOME/lib/native:$JAVA_LIBRARY_PATH
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="$HADOOP_OPTS -Djava.net.preferIPv4Stack=true
-Djava.library.path=$HADOOP_HOME/lib/native"
```

I realized that data node was not coming up because of which I was not able to create/copy file in HDFS so I made following changes

```
in etc/hadoop/hadoop-env.sh

export HADOOP_DATANODE_OPTS="-Dhadoop.security.logger=ERROR,RFAS
$HADOOP_DATANODE_OPTS"
```

```
Added below line in etc/hadoop/hdfs-site.xml also
<property>
 <name>dfs.datanode.data.dir</name>
 <value>file:/Users/hadoop/hadoop-2.7.4/hdfstmp/datanode/data</value>
</property>
<property>
 <name>dfs.permissions.superusergroup</name>
```

```
 <value>hadoop</value>
</property>
```

After adding above code datanode was up and I was able to create directory and copy a file to HDFS.

```
bin/hadoop fs -mkdir /TestingHDFS
bin/hadoop fs -copyFromLocal /Users/hadoop/edges.txt /TestingHDFS
```

Can check details about hadoop using below commands
Job Tracker:  http://localhost:8088/cluster
Name node/Resource Manager : http://localhost:50070

**Installing Spark**
brew install apache-spark