Scala, Hadoop installation guide on Linux

Oracle JDK 1.8

Execute the following command

```
sudo apt-get install openjdk-8-jdk
```

OR

Manual installation:
To install the JDK manually on a Linux system, follow these steps:

1. Download the .tar.gz archive from the Oracle website
2. Unpack the downloaded archive to a directory of your choice
3. Add the *bin/* directory of the extracted JDK to the PATH environment variable. Open the file *~/.bashrc* in an editor (create it if it doesn't exist) and add the following line:

```
export JAVA_HOME=/PATH/TO/YOUR/jdk1.8.0

export PATH="JAVA_HOME/bin:$PATH"

Examples:

export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64

export PATH="JAVA_HOME/bin:$PATH"
```

Verify java by executing the following command

```
java -version
```

SSH Setup and Key Generation

SSH setup is required to do different operations on a cluster such as starting, stopping, distributed daemon shell operations. To authenticate different users of Hadoop, it is required to provide a public/private key pair for a Hadoop user and share it with different users.

```
$ ssh-keygen -t rsa
```

Copy the public keys from id_rsa.pub to authorized_keys

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

$ chmod 0600 ~/.ssh/authorized_keys
```

SBT - Scala Build Tool

SBT is for building, testing, running and submitting assignments.
Installing SBT: execute the following commands

```
echo "deb https://dl.bintray.com/sbt/debian /" | sudo tee -a
/etc/apt/sources.list.d/sbt.list

curl -sL
"https://keyserver.ubuntu.com/pks/lookup?op=get&search=0x2EE0EA64E40A
89B84B2DF73499E82A75642AC823" | sudo apt-key add

sudo apt-get update

sudo apt-get install sbt
```

Scala IDE for Eclipse

download the Scala IDE for eclipse with the Scala Worksheet pre-installed from the following URL:
http://scala-ide.org/download/sdk.html
After downloading the zip file extract and start the eclipse.

Hadoop

Installation :

Download the hadoop using the following steps

```
# cd /usr/local
# wget
http://apache.claz.org/hadoop/common/hadoop-2.10.0/hadoop-2.10.
0.tar.gz
# tar xzf hadoop-2.10.0.tar.gz
# mv hadoop-2.10.0/*  hadoop/
```

Standalone Mode:

Set Hadoop environment variables by appending the following commands to **~/.bashrc** file

```
export HADOOP_HOME=/usr/local/hadoop
```

Before proceeding further, you need to make sure that Hadoop is working fine. Just issue the following command −

```
hadoop version
```

Pseudo Distributed Mode

Follow the steps given below to install pseudo distributed mode.

1. Setting up Environment variables - append following to ~/.bashrc file

```
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_INSTALL=$HADOOP_HOME
```

Now apply changes using below connad

```
$source ~/.bashrc
```

or

```
$ bash
```

2. Hadoop Configuration

All the configuration files are present in the location “$HADOOP_HOME/etc/hadoop” (i.e. /usr/local/hadoop/etc/hadoop).

cd $HADOOP_HOME/etc/hadoop

Set the java environment variable in **hadoop-env.sh** file by replacing JAVA_HOME value with the location of java in your system.

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
```

core-site.xml

Open the core-site.xml and add the following properties between <configuration> tags

```
<configuration>
     <property>
          <name>fs.default.name</name>
          <value>hdfs://debian:9000</value>
     </property>
</configuration>
```

hdfs-site.xml

Open hdfs-site.xml file and add the following properties in between <configuration>

```
<configuration>
     <property>
          <name>dfs.replication</name>
          <value>1</value>
     </property>
     <property>
          <name>dfs.user.home.base.dir</name>
          <value>/home/user</value>
          <description>Base directory of user
     home.</description>
     </property>
     <property>
          <name>dfs.name.dir</name>
          <value>file:///home/user/hadoopinfra/hdfs/namenode
     </value>
     </property>

      <property>
          <name>dfs.data.dir</name>
          <value>file:///home/user/hadoopinfra/hdfs/datanode
     </value>
      </property>
</configuration>
```

yarn-site.xml

Open the yarn-site.xml and add the following properties in between <configuration> tags.

```
<configuration>
     <property>
          <name>yarn.nodemanager.aux-services</name>
          <value>mapreduce_shuffle</value>
     </property>
</configuration>
```

mapred-site.xml

Copy mapred-site.xml.template to mapred-site.xml

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

Add the following properties in between the <configuration> tags.

```
<configuration>
     <property>
          <name>mapreduce.framework.name</name>
          <value>yarn</value>
     </property>
</configuration>
```

Name Node Setup

Set up the namenode using below command

```
$ hdfs namenode -format
```

Start Hadoop file system

Execute the following command to start your Hadoop file system

```
$ start-dfs.sh
```

Start Yarn daemons

Execute the following command to start yarn demons

```
$ start-yarn.sh
```

Hit the following url and verify your hadoop services.

http://localhost:50070/

http://localhost:8088/

Multi Node Cluster

Master node configuration

Change the node(Desktop) name to as below

Node1 -> hadoop-master

sudo hostnamectl set-hostname hadoop-master

Node2 -> hadoop-slave-1

sudo hostnamectl set-hostname hadoop-slave-1

Node3 -> hadoop-slave-2

sudo hostnamectl set-hostname hadoop-slave-2

Open /etc/hosts file and add the IP address of each system followed by their names

```
$ vi /etc/hosts
10.250.1.67 hadoop-master
10.250.1.68 hadoop-slave-1
10.250.1.69 hadoop-slave-2
```

Configure ssh key - login

Setup ssh in every node such that they can communicate with one another without any prompt for password.

```
$ ssh-keygen -t rsa
$ ssh-copy-id -i ~/.ssh/id_rsa.pub user@hadoop-master
$ ssh-copy-id -i ~/.ssh/id_rsa.pub user@hadoop-slave-1
$ ssh-copy-id -i ~/.ssh/id_rsa.pub user@hadoop-slave-2
$ chmod 0600 ~/.ssh/authorized_keys
```

Install Hadoop

To install the Hadoop follow the steps specified above.

Set the java environment variable in **hadoop-env.sh** file by replacing JAVA_HOME value with the location of java in your system.

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
```

Configure Hadoop in Master

core-site.xml

Open core-site.xml file and add the following configuration.

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

```
        <value>hdfs://hadoop-master:9000/</value>
    </property>
    <property>
        <name>dfs.permissions</name>
        <value>false</value>
    </property>
</configuration>
```

hdfs-site.xml

Open hdfs-site.xml file and add the following configuration.

```
<configuration>
    <property>
        <name>dfs.user.home.base.dir</name>
        <value>/home/user</value>
    </property>
    <property>
        <name>dfs.data.dir</name>
        <value>/home/user/hadoop_store/hdfs/datanode</value>
        <final>true</final>
    </property>
    <property>
        <name>dfs.name.dir</name>
        <value>/home/user/hadoop_store/hdfs/namenode</value>
        <final>true</final>
    </property>
</configuration>
```

mapred-site.xml

Open the mapred-site.xml file and add the following configuration.

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

Add the master node

```
vi etc/hadoop/masters

hadoop-master
```

Add the slave nodes details in slaves file

```
vi etc/hadoop/slaves
```

```
hadoop-slave-1
hadoop-slave-2
```

Slaves side configuration

Install hadoop

```
$ cd /opt/hadoop
$ scp -r hadoop hadoop-slave-1:/opt/hadoop
$ scp -r hadoop hadoop-slave-2:/opt/hadoop
```

Update the /etc/hosts file

```
$ vi /etc/hosts
10.250.1.67 hadoop-master
10.250.1.68 hadoop-slave-1
10.250.1.69 hadoop-slave-2
```

Add master and slave entry

Add the master node

```
vi etc/hadoop/masters

hadoop-master
```

Add the slave nodes details in slaves file

```
vi etc/hadoop/slaves

hadoop-slave-1
hadoop-slave-2
```

In master format name node

```
$ hadoop namenode –format
```

Start hdfs and yarn

```
$ start-hdfs.sh
$ start-yarn.sh
```