cewiki

Hadoop Installation Setup in Psuedo-Distributed Mode on Linux (CentOS 6.5)

JAVA Installation

- Use "$java -version" to check java version already installed in the machine.
- Latest version of Oracle jdk is preferrable.
- If it is not installed, download it from http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html [http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html]

(Linux x64 176.95 MB jdk-8u131-linux-x64.tar.gz)

```
$ tar zxvf  jdk-8u131-linux-x64.tar.gz
```

- Preferred Location for installation: /usr/java/. Move extracted contents to this directory.
- Use command "alternatives –config java" to see the list of all java versions. It will ask to select anyone version as default. Select oracle jdk as default.

(Note : Oracle's RPMs are ignorant of the "alternatives" system. So use following command to solve the problem if alternatives command does not print the recently installed java version.

```
/usr/sbin/alternatives --install /usr/bin/java java /usr/java/default/bin/java 20000)
```

For setting up PATH and JAVA_HOME variables, add following commands to ~/.bashrc file.

```
export JAVA_HOME=/usr/java/
export PATH=$PATH:$JAVA_HOME/bin
```

Apply all changes into current running system.

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

SSH Setup and Key Generation

```
$rm -rf ~/.ssh //removes previous keys
$ ssh-keygen -t rsa
$ssh-copy-id -i ~/.ssh/id_rsa.pub hadoop@localhost
//Press enter when it asks for some input.
$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
$ chmod 0600 ~/.ssh/authorized_keys
```

Hadoop Installation

(Assuming currently working in hadoop user.)

- Download latest stable version. (e.g. hadoop-2.8.0)
- Move to Downloads folder.
- Execute following command.

```
# tar xzf hadoop-2.8.0.tar.gz
```

- Move to super user using su command.

- Create a hadoop directory in /opt
- Copy contents of extracted folder to /opt/hadoop/.

```
# su
#mkdir /opt/hadoop
#mv /home/hadoop/Downloads/hadoop-2.8.0 /opt/hadoop
// Using cd command, move to opt directiry.
#chown -R hadoop:hadoop hadoop/ //Changes owner and group of the directory.
# exit
```

Verify using ll command.

Setting Up Hadoop

Add following contents in ~/.bashrc file of hadoop user.

```
export JAVA_HOME=/usr/java/jdk1.8.0_91
export HADOOP_HOME=/opt/hadoop
export HADOOP_PREFIX=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
export PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

Verify installation by executing following command.

```
$hadoop version
```

Following contents should be added to configuration files: Location : /opt/hadoop/etc/hadoop

core-site.xml

```
<configuration>
<property>
    <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000/</value>
  </property>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/opt/hadoop/tmp</value>
  </property>
</configuration>
```

hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/opt/hadoop/tmp/dfs/name</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/opt/hadoop/tmp/dfs/data</value>
  </property>
  <property>
    <name>dfs.namenode.http-address</name>
    <value>localhost:50070</value>
  </property>

  <property>
    <name>dfs.namenode.secondary.http-address</name>
    <value>localhost:50090</value>
```

```
        </property>
</configuration>
```

This command needs to be executed before following next steps.

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

mapred-site.xml

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

    <property>
        <name>mapreduce.jobhistory.address</name>
        <value>localhost:10020</value>
    </property>

    <property>
        <name>mapreduce.jobhistory.webapp.address</name>
        <value>localhost:19888</value>
    </property>
</configuration>
```

yarn-site.xml

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

Format the namenode (only once)

```
$ mkdir -p /opt/hadoop/tmp/dfs/name
$ mkdir -p /opt/hadoop/tmp/dfs/data

$hdfs namenode -format
```

Start services. (Note : /etc/hosts file should have entry for localhost. /etc/sysconfig/network file should have localhost as HOSTNAME)

```
$start-dfs.sh
$start-yarn.sh
$mr-jobhistory-daemon.sh --config $HADOOP_CONF_DIR start historyserver
$jps
```

This command should list following daemons. If not, there may be errors in installation.

```
4807 ResourceManager
4343 NameNode
5273 JobHistoryServer
5321 Jps
4477 DataNode
4653 SecondaryNameNode
4911 NodeManager
```

Run your programs. Stop services.

```
$stop-dfs.sh
```

```
$stop-yarn.sh
$mr-jobhistory-daemon.sh --config $HADOOP_CONF_DIR stop historyserver
```