Hadoop: Setting up a Single Node Cluster

Qi Song 08/25/2015

This instruction is derived from Hadoop: Setting up a Single Node Cluster. If you meet any errors during installation or execution, please try to google them and solve them!

Prerequisites:

Supported Platforms

Windows is also a supported platform but the followings steps are for Linux (Ubuntu 14.04 – 64bit) only. To set up Hadoop on Windows, see Build and Install Hadoop 2.x or newer on Windows.

Hadoop version

We highly recommend stable version of hadoop . We use hadoop-2.7.1 here. Different versions of hadoop can be download from Apache Hadoop Releases.

Java install:

Please choose java version based on Hadoop Java Versions. Oracle 1.7.0_45 is recommended here. Download jdk and jre from Java SE 7 Archive Downloads and then:

```
$ sudo cp -r jdk-7u45-linux-x64.tar.gz /usr/local/java
$ sudo cp -r jre-7u45-linux-x64.tar.gz /usr/local/java
$ sudo tar xvzf jdk-7u45-linux-x64.tar.gz
$ sudo tar xvzf jre-7u45-linux-x64.tar.gz
```

Then open */etc/profil*e and add:

```
JAVA_HOME=/usr/local/java/jdk1.7.0_45
PATH=$PATH:$HOME/bin:$JAVA_HOME/bin
JRE_HOME=/usr/local/java/jre1.7.0_45
PATH=$PATH:$HOME/bin:$JRE_HOME/bin
export JAVA_HOME
export JRE_HOME
export PATH
```

Run

```
$ source /etc/profile
$ java -version
```

Should see java version now.

SSH Install:

Install SSH and set up ssh to the localhost without a passphrase:

```
$ sudo apt-get install ssh
$ sudo apt-get install rsync
$ ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa
$ cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys
$ export HADOOP\_PREFIX=/usr/local/hadoop
```

Now you can ssh to localhost (*$ ssh localhost*) without typing a password.

Setup Pseudo-Distributed mode

Configuration:

Edit the file *etc/hadoop/hadoop-env.sh* to define some parameters as follows:

```
export JAVA_HOME="/usr/local/java/jdk1.7.0_45"
```

Use the following:

etc/hadoop/core-site.xml:

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

etc/hadoop/hdfs-site.xml:

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

/etc/profile:

```
#HADOOP
export HADOOP_HOME=/home/qsong/hadoop-2.7.1
export PATH=$PATH:$HADOOP_HOME/bin
export HADOOP_HOME_WARN_SUPPRESS=1
```

Execution:

1. Format the filesystem:

```
$ bin/hdfs namenode -format
```

If you meet "Error: Could not find or load main class org.apache.hadoop.hdfs.server.namenode.NameNode" , open *~/.bash_profile* and add *"export HADOOP_PREFIX=/home/qsong/hadoop-2.7.1".* Then run *$ source ~/.bash_profile*

2. Start NameNode daemon and DataNode daemon:

```
$ sbin/start-dfs.sh
```

3. Browse the web interface for the NameNode; by default it is available at:
   NameNode - http://localhost:50070/
4. Make the HDFS directories required to execute MapReduce jobs:

```
$ bin/hdfs dfs -mkdir /user
$ bin/hdfs dfs -mkdir /user/qsong
```

5. Copy the input files into the distributed filesystem:

```
$ bin/hdfs dfs -put etc/hadoop input
```

6. Run some of the examples provided:

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar grep input output 'dfs[a-z.]+'
```

7. View the output files on the distributed filesystem:

```
$ bin/hdfs dfs -cat output/*
```

8. When you're done, stop the daemons with:

```
$ sbin/stop-dfs.sh
```