How to install Hadoop distribution from Bigtop 0.5.0

Introduction

Installing Bigtop Hadoop distribution artifacts lets you have an up and running Hadoop cluster complete with various Hadoop ecosystem projects in just a few minutes. Be it a single node pseudo-distributed configuration, or a fully distributed cluster, just make sure you install the packages, install the JDK, format the namenode and have fun! If Bigtop is not supported on your OS, you can install one of the supported 64-bit OSes on a virtual machine. There are known issues with 32-bit OSes.

WARNING 1: We are supporting only Oracle JDK/JRE 6. Bigtop distribution may or may not run with OpenJDK. Please got to the Oracle's web site and download the JDK6 package for the Linux distro that you're using: http://www.oracle.com/technetwork/java/javasebusiness/downloads/java-archive-downloads-javase6-419409.html before installing Bigtop

WARNING 2: We are supporting only 64bit Linux OSes with our binary convenience artifacts. If you happen to have an older 32bit one you can rebuild Bigtop from source but you won't be able to install from our repositories, so the rest of the instructions won't apply to you.

WARNING 3: BigTop involves installing several Hadoop-related services on your machine that are enabled by default (i.e., started upon system reboot). If you would rather have the services start only when you specifically activate them, check your OS-specific documentation for disabling specific services at startup, see here, here and here for some options.

NOTE 1: We support Fedora 16 and tested Bigtop 0.5.0 on it; however, we haven't released any binary artifacts for it.

Getting the packages onto your box

CentOS 5, CentOS 6, Fedora 17, RHEL5, RHEL6

1. Make sure to grab the repo file:

```
wget -O /etc/yum.repos.d/bigtop.repo http://archive.apache.org/dist/bigtop/bigtop-0.5.0/repos/
[centos5|centos6|fedora17]/bigtop.repo
```

2. Browse through the artifacts

```
yum search mahout
```

3. Install the full Hadoop stack (or parts of it)

```
sudo yum install hadoop\* flume\* mahout\* oozie\* whirr\* hbase\* hive\*
```

SLES 11, OpenSUSE

1. Make sure to grab the repo file:

```
wget http://archive.apache.org/dist/bigtop/bigtop-0.5.0/repos/[sles11|opensuse12]/bigtop.repo
#mv bigtop.repo  /etc/zypp/repos.d/bigtop.repo
```

2. Refresh zypper to start looking at the newly added repo

```
#zypper refresh
```

3. Browse through the artifacts

```
zypper search mahout
```

4. Install the full Hadoop stack (or parts of it)

```
zypper install hadoop\* flume\* mahout\* oozie\* whirr\* hive\*
```

Ubuntu (64 bit, lucid, precise, quantal)

1. Install the Apache Bigtop GPG key

```
wget -O- http://archive.apache.org/dist/bigtop/bigtop-0.5.0/repos/GPG-KEY-bigtop | sudo apt-key add -
```

2. Make sure to grab the repo file:

```
sudo wget -O /etc/apt/sources.list.d/bigtop-0.5.0.list http://archive.apache.org/dist/bigtop/bigtop-0.5.0
/repos/`lsb_release --codename --short`/bigtop.list
```

3. Update the apt cache

```
sudo apt-get update
```

4. Browse through the artifacts

```
apt-cache search mahout
```

5. Install bigtop-utils

```
sudo apt-get install bigtop-utils
```

6. Make sure that you have the latest JDK installed on your system as well. You can either get it from the official Oracle website (http://www.oracle.com/technetwork/java/javase/downloads/jdk-6u29-download-513648.html) or follow the advice given by your Linux distribution. If your JDK is installed in a non-standard location, make sure to add the line below to the /etc/default/bigtop-utils file

```
export JAVA_HOME=XXXX
```

7. Install the full Hadoop stack (or parts of it)

```
sudo apt-get install hadoop\* flume-* mahout\* oozie\* whirr-* hive\*
```

Running Hadoop

After installing Hadoop packages onto your Linux box, make sure that:

1. You have the latest JDK installed on your system as well. You can either get it from the official Oracle website (http://www.oracle.com/technetwork/java/javase/downloads/jdk-6u29-download-513648.html) or follow the advice given by your Linux distribution (e.g. some Debian based Linux distributions have JDK packaged as part of their extended set of packages). If your JDK is installed in a non-standard location, make sure to add the line below to the /etc/default/bigtop-utils file

```
export JAVA_HOME=XXXX
```

2. Format the namenode

```
sudo /etc/init.d/hadoop-hdfs-namenode init
```

3. Start the necessary Hadoop services. E.g. for the pseudo distributed Hadoop installation you can simply do:

```
for i in hadoop-hdfs-namenode hadoop-hdfs-datanode ; do sudo service $i start ; done
```

4. Make sure to create a sub-directory structure in HDFS before running any daemons:

```
sudo -u hdfs hadoop fs -mkdir -p /user/$USER
sudo -u hdfs hadoop fs -chown $USER:$USER /user/$USER
sudo -u hdfs hadoop fs -chmod 770 /user/$USER

sudo -u hdfs hadoop fs -mkdir /tmp
sudo -u hdfs hadoop fs -chmod -R 1777 /tmp

sudo -u hdfs hadoop fs -mkdir -p /var/log/hadoop-yarn
sudo -u hdfs hadoop fs -chown yarn:mapred /var/log/hadoop-yarn

sudo -u hdfs hadoop fs -mkdir -p /user/history
sudo -u hdfs hadoop fs -chown mapred:mapred /user/history
sudo -u hdfs hadoop fs -chmod 770 /user/history

sudo -u hdfs hadoop fs -mkdir -p /tmp/hadoop-yarn/staging
sudo -u hdfs hadoop fs -chmod -R 1777 /tmp/hadoop-yarn/staging

sudo -u hdfs hadoop fs -mkdir -p /tmp/hadoop-yarn/staging/history/done_intermediate
sudo -u hdfs hadoop fs -chmod -R 1777 /tmp/hadoop-yarn/staging/history/done_intermediate
sudo -u hdfs hadoop fs -chown -R mapred:mapred /tmp/hadoop-yarn/staging
```

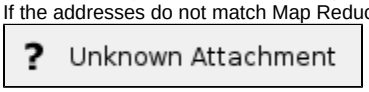
5. Now start YARN daemons:

```
sudo service hadoop-yarn-resourcemanager start
sudo service hadoop-yarn-nodemanager start
```

6. Enjoy your cluster

```
hadoop fs -lsr /
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples*.jar pi 10 1000
```

7. If you are using Amazon AWS it is important the IP address in /etc/hostname matches the Private IP Address in the AWS Management Console. If the addresses do not match Map Reduce programs will not complete.




```
ubuntu@ip-10-224-113-68:~$ cat /etc/hostname
ip-10-224-113-68
```

8. If the IP address in /etc/hostname does not match then open the hostname file in a text editor, change and reboot

Running Hadoop Components

Linky -> Step-by-step instructions on running Hadoop Components!

One of the advantages of Bigtop is the ease of installation of the different Hadoop Components without having to hunt for a specific Hadoop Component distribution and matching it with a specific Hadoop version.
Please visit the link above to run some easy examples from the Bigtop distribution !
Provided at the link above are examples to run Hadoop 1.0.1 and nine other components from the Hadoop ecosystem (hive/hbase/zookeeper/pig/sqoop /oozie/mahout/whirr and flume).
See the

Bigtop Make File

for a list of Hadoop components , officially available from the Bigtop distribution.

Where to go from here

It is highly recommended that you read documentation provided by the Hadoop project itself

1. https://hadoop.apache.org/common/docs/r1.0.1/ for Bigtop 0.3
2. or, http://hadoop.apache.org/common/docs/r0.20.205.0/) Bigtop 0.2
   and that you browse through the Puppet deployment code that is shipped as part of the Bigtop release (bigtop-deploy/puppet/modules, bigtop-deploy/puppet/manifests).