How to install Bigtop 0.8.0 Hadoop on CentOS 6 with Puppet

Thanks to Martin Bukatovi for this excellent writeup !

This page expects that you have already provisioned cluster of CentOS 6.x machines where you would like to deploy 0.8.0 release of Bigtop Hadoop distribution using Puppet recipes from `bigtop-deploy`.

Requirements of 0.8.0 release:

- x86_64 is the only supported architecture
- deployment scripts requires puppet 2.7+/3.x (see details below)
- Java JDK 1.6

Quick overview

To install Bigtop Hadoop distribuion on your cluster of machines running CentOS 6.x, you would need 2 Bigtop deliverables:

- rpm packages of all hadoop components
- puppet depyloyment scripts which will install and setup your cluster

So where do I find the packages? Bigtop home page contains link to `bigtop-0.8.0/repos/centos6/bigtop.repo` file which uses

```
baseurl=http://bigtop.s3.amazonaws.com/releases/0.8.0/redhat/6/x86_64
```

You can use this url when you will configure puppet config file (as shown in detail below), or you can go to Bigtop jenkins and download all packages so that you can setup you own local repository. Another option would be to build it yourself, but this topic is not covered in this page.

Puppet deployment scrips are not packaged and you can get it from either:

- source tarball so that you use puppet recipes which are part of 0.8.0 release
- puppet recipes from current master branch of bigtop repository

While it's safer to use recipes from the release in theory, most people uses puppet scripts from git. The reasoning behind this is that while current recipes changed a lot, it still can deploy Bigtop Hadoop from 0.8.0 release and so it makes sense to learn and use the newest version from the start.

Note: the only 0.8.0 Bigtop Hadoop component which can't be deployed with scripts from git is sqoop.

Installation of requirements

There are basically only two requirements:

- install puppet on all machines of the cluster:
  - version 2.7+ if you are using recipes from the release
  - version 3.x if you are using scripts from git
- have packages with java 1.6 available in the repos of your distro

For the CentOS 6.x machines, this means:

- openjdk 1.6 which is shipped with the distro
- puppet package (depends on which recipes you are going to use):
  - either puppet 2.7 from EPEL (so I need to enable EPEL first)
  - or puppet 3.x from the upstream

Note that:

- You don't have to setup bigtop yum repository or install java - all these steps and more are automated via puppet.
- The only catch is that if you have java installed already on the machines, you would need to check that correct java is activated via alternatives. Pupped wouldn't notice that another java version is activated. See details below.

The deployment

Make sure that content of Bigtop 0.8.0 release tarball or bigtop repo (as discussed above) is available in `/opt/bigtop` on all machines. You may like to use NFS share for this.

Now it's the time to configure puppet `site.csv` file (the only puppet file you need to touch). It defines cluster roles, hadoop components installed and other details:

```
hadoop_head_node,master.bigtop.lab.example.com
hadoop_storage_dirs,/data/1
bigtop_yumrepo_uri,http://some-local-mirror.example.com/bigtop/bigtop-0.8.0/output/
components,hadoop,yarn,mapred-app
jdk_package_name,java-1.6.0-openjdk
```

Where:

- `hadoop_head_node` is a master (it runns eg. YARN Resource Manager), you need to specify fqdn there (otherwise no node will be configured as master).
- `bigtop_yumrepo_uri` url of bigtop repo you will use, puppet will create yum repofile for it. I'm using local mirror here, but it should be possible to use repo hosted on s3 (as shown in previous section).
- `components` list of hadoop components to install. The example shows minimal list of components for you to be able execute mapreduce jobs. See puppet recipes for list of all components.
- `jdk_package_name` - name of the package with java, puppet will install it

Bigtop uses puppet in masterless mode, so you need to distribute new version of site.csv and then run puppet locally on all machines.

So on each machine of the cluster:

1. push new site.csv `/opt/bigtop/bigtop-deploy/puppet/config/site.csv`
2. run puppet:

```
cd /opt/bigtop/bigtop-deploy/puppet
puppet apply -d --modulepath=modules --confdir=. manifests/site.pp
```

Check java version

If you have multiple versions of java installed, check which one is active:

```
# alternatives --display java
```

The puppet wouldn't notice that different version of java is active.

And install and/or change it to 1.6 if needed:

```
# yum install java-1.6.0-openjdk java-1.6.0-openjdk-devel
# alternatives --set javac /usr/lib/jvm/java-1.6.0-openjdk.x86_64/bin/javac
# alternatives --set java /usr/lib/jvm/jre-1.6.0-openjdk.x86_64/bin/java
```