Lab 1. Introduction to Hadoop Systems Administration.

Part 1. Install support tools for Hadoop.

You will be provided with several <distribution here> systems that you will use to install the Cloudera hadoop distribution package. Additionally, we will install some related tools that will make it easier to administer the cluster.

You will be provided with the following systems:

1. An administrative node. Use this node to issues commands to individual cluster nodes or to all nodes on the cluster
2. An HDFS node. This node will store all of the HDFS data that you will need for the labs.
3. A YARN node. This node will administer the Hadoop jobs that you will run.

Step 1. Install an ssh key on the nodes.

In order to avoid having to type in a password for every ssh command, you will provide an RSA key that ssh will use to authenticate you to the nodes..

On your administrative server, please generate an ssh key using ssh-keygen .

Click on this link to get instructions on how to generate an ssh key for your administrative node.

https://www.howtoforge.com/linux-basics-how-to-install-ssh-keys-on-the-shell

Note that it is not necessary to create a passphrase.

Once this is done, log in to the provided servers and insert the ssh key created to the /root/.ssh/authorized_keys file.

Verify that you can now log into the server without having to type in a password.

This process should be repeated for all of the nodes in your hadoop cluster.

Step 2. Configure the /etc/hosts file for your administrative node.

Please edit the /etc/hosts file in your administrative node and enter the IP addresses of all of the cluster nodes. :

```
<ip Address> hadoop-1
<ip Address> hadoop-2
<ip Address> hadoop-3
<ip Address> hadoop-4
```

Run *cat /etc/hosts* to verify that all of the IP addresses and node names are entered correctly.

Step 3. Create a student username

Using the adduser program, create a *student* user.

Once this is done, log out as root and re-login as the student. Note that you should repeat the key procedure above with the student login.

Repeat this for all of your hadoop cluster nodes.

Now create another ssh key

Step 4. Create a shell function cdo() as follows:

```
# cdo() { for n in {1..4}; do ssh hadoop-${n} $@; done; }
```

This will allow you to run commands on all nodes rather than having to do

Individual ssh commands.

For example:

```
# cdo tail -n20 /var/log/messages
```

Will print out the last 20 lines of the messages file for each node.

Step 5. Install Parallel SSH.

Click on the following link to obtain instructions for installing Parallel SShy on a CentOS system.

https://www.cyberciti.biz/cloud-computing/how-to-use-pssh-parallel-ssh-program-on-linux-unix/attachment/howto-install-pssh-on-rhel-redhat-centos-linux/

Click on the following link to obtain instructions for installing Parallel SSH on an Ubuntu system.

https://www.cyberciti.biz/cloud-computing/how-to-use-pssh-parallel-ssh-program-on-linux-unix/attachment/howto-install-pssh-on-debian-ubuntu-linux/

Here are some examples of using pssh.

```
pssh -H "node1 node2 node3" /usr/sbin/service sshd status
[1] 14:57:41 [SUCCESS] hadoop-1
[2] 14:57:41 [SUCCESS] hadoolp-2
[3] 14:57:41 [SUCCESS] hadoop-3

# echo -e "hadoop-1\nhadoop-2\nhadoop-3" >cluster_hosts.txt
# pssh -h clusterhosts.txt service sshd status
[1] 14:58:42 [SUCCESS] hadoop-1
[2] 14:58:42 [SUCCESS] hadoop-2
[3] 14:58:42 [SUCCESS] hadoop-3

The -I option returns the standard out or standard error rather than
the
return status code.

# pssh -ih clusterhosts.txt rpm -q pssh
[1] 14:50:51 [SUCCESS] hadoop-1
pssh-2.2.2-1.el6.noarch
[2] 14:50:51 [FAILURE] hadoop-3  Exited with error code 1
package pssh is not installed
[3] 14:50:51 [SUCCESS] hadoop-2
pssh-2.2.2-1.el6.noarch
```

Test that pssh is installed successfully.

```
This creates a file called pssh-hosts with the hostnames of your
nodes.
# echo "hadoop-2
```

```
>  hadoop-22
> hadoop-3 " > /root/pssh-hosts

Create an alias called allnodes that will run pssh on the hosts in the
pssh-hosts file.
# alias allnodes="pssh -h /root/pssh-hosts"

# allnodes ip addr list
[1] 18:11:28 [SUCCESS] hadoop-1
[2] 18:11:28 [SUCCESS] hadoop-2
[3] 18:11:28 [SUCCESS] hadoop-3

Add the alias to the .bashrc  file in the root directory  so that you
don't have to type this in
every time you log in.

 alias allnodes='pssh -h /root/pssh-hosts'
```

Run allnodes with the -i option to make sure that you get standard error/output as well.

Edit the hosts file on the other nodes and add the pssh-hosts file created to all the other nodes in the cluster.

Step 6.

(Optional). Install VNC on your administrative console.

Go to this link to install VNC on Ubuntu servers.

https://www.digitalocean.com/community/tutorials/how-to-install-and-configure-vnc-on-ubuntu-16-04

Go to this link to install VNC on CentOS servers.

https://www.digitalocean.com/community/tutorials/how-to-install-and-configure-vnc-remote-access-for-the-gnome-desktop-on-centos-7

You can now use any VNC client (such as TightVNC) on your client if you are running Windows. Or install a VNC client for Linux or the Macintosh as necessary.

Step 7. (Optional).

Install ClusterSSH on your administrative node.

Use the following link to install ClusterSSH on your server.

https://www.linux.com/learn/managing-multiple-linux-servers-clusterssh

Part 2. Install the Hadoop cluster.

Step 1. Install Java 10 for Linux.

The oracle binary rpm distribution for Java 10 will be made available to you on your first node in the /root directory.

Run the following command to install it:

```
yum localinstall
```

```
# java -show-version
java 10.0.1 2018-04-17
Java(TM) SE Runtime Environment 18.3 (build 10.0.1+10)
Java HotSpot(TM) 64-Bit Server VM 18.3 (build 10.0.1+10, mixed mode)
```

Additionally, we should install some JDK tools to help us administer the cluster.

```
allnodes yum  install -y java-1.8.0-open-jdk-devel.x86_64
```

Step 2. Install Hadoop.

For Ubuntu, run the following commands

```
1. sudo wget
   'https://archive.cloudera.com/cdh5/ubuntu/xenial/amd64/cdh/cloude
   ra.list' \ -O /etc/apt/sources.list.d/cloudera.list
2. sudo apt-get update
3. sudo apt-get install hadoop-yarn-resourcemanager
```

For CentOS , run the following commands.

```
1. yum clean all
2. yum install  yum-utils
3. yum-config-manager --add-repo
   https://archive.cloudera.com/cdh5/redhat/7/x86_64/cdh/cloudera-
   cdh5.repo
4. yum list hadoop
5. yum install hadoop
6. pscp.pssh -H "hadoop-2 hadoop-3 " /etc/yum.repos.d/cloudera-
   cdh5.repo /etc/yum.repos.d/
```

```
7. allnodes systemctl stop firewall
8. allnodes systemctl mask firewall
```