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, we will provided 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.

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> node1
```

```
<ip Address> node2
<ip Address> node3
<ip Address> node4
```

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.

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

```
# cdo() { for n in {1..4}; do ssh node${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" service sshd status
[1] 14:57:41 [SUCCESS] node1
[2] 14:57:41 [SUCCESS] node2
[3] 14:57:41 [SUCCESS] node3

# echo -e "node1\nnode2\nnode3" >cluster hosts.txt
```

```
# pssh -h clusterhosts.txt service sshd status
[1] 14:58:42 [SUCCESS] node1
[2] 14:58:42 [SUCCESS] node2
[3] 14:58:42 [SUCCESS] node3

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] node1
pssh-2.2.2-1.el6.noarch
[2] 14:50:51 [FAILURE] node3 Exited with error code 1
package pssh is not installed
[3] 14:50:51 [SUCCESS] node2
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 "node1
> node2
> node3
> node4" > /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] node1
[2] 18:11:28 [SUCCESS] node3
[3] 18:11:28 [SUCCESS] node4
[4] 18:11:28 [SUCCESS] node2

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 8 for Linux.

The oracle binary rpm distribution for Java 8 will be made available to you on your first node in the /root directory.

Run the following command to install it:

```
yum localinstall
```

```
# java -showversion
java version "1.8.0_144"
Java(TM) SE Runtime Environment (build 1.8.0_144-b01)
Java HotSpot(TM) 64-Bit Server VM (build 25.144-b01, 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/cloudera.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. pscp.pssh -H "node2 node3 node4" /etc/yum.repos.d/cloudera-cdh5.repo
   /etc/yum.repos.d/
6. allnodes systemctl stop firewall
7. allnodes systemctl mask firewall
```