Hands-On Exercise: Installing Hadoop

In this exercise you will install and test Hadoop using the CDH RPMs.

Introduction

The Virtual Machine contains a local yum repository of CDH4.1. We have created the local repository to save download time; typically, you would use the online repository following the instructions found at `http://docs.cloudera.com/`.

You will initially install CDH in *pseudo-distributed* mode. This is a way of installing a Hadoop cluster on a single machine; the machine runs all five Hadoop daemons (NameNode, Secondary NameNode, DataNode, JobTracker and TaskTracker). Because there is only a single DataNode, the replication factor must be set to 1.

Install Hadoop in Pseudo-Distributed Mode From CDH

1. If it has not already started, launch the Virtual Machine by double-clicking the `Cloudera_admin_training_4.1.vmx` file in the Virtual Machine directory. If you are asked whether you moved or copied the VM, choose **I copied it**.

2. **VERY IMPORTANT:** You need to give your VM a unique IP address, which will be specified by your instructor. Once your instructor has given you the IP number to use, you should run the following command and enter the appropriate value.

   ```
   $ ./clouderanetworking
   ```

   This scripts sets your IP address and adds a line to the `/etc/hosts` file.

Copyright © 2010-2013 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

3. Check you can ping another student's machine: using the number that your instructor gave a neighbor, try

```
$ ping your_neighbor's_IP_address
```

If you cannot ping other machines in the class, please let your instructor know!

4. Now we will install Hadoop. From the terminal window, type:

```
$ sudo yum install hadoop-0.20-conf-pseudo
```

This installs the core Hadoop package, init scripts for the Hadoop daemons, and the configuration files required for Hadoop to operate in pseudo-distributed mode.

5. You must now format the NameNode.

```
$ sudo -u hdfs hdfs namenode -format
```

(Note: In CDH3, installing Hadoop in pseudo-distributed mode would automatically format the NameNode. This is no longer the case in CDH4.)

The NameNode metadata is stored in `/var/lib/hadoop-hdfs/cache/hdfs/dfs/name`

6. CDH uses the Linux *alternatives* system, which is a way of retaining multiple configuration settings. You will find the configuration files in `/etc/hadoop/conf`, which is a symbolic link from `/etc/alternatives/hadoop-conf`.

Change directory to `/etc/hadoop/conf` and inspect the `*-site.xml` files.

Copyright © 2010-2013 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

7. Start Hadoop's HDFS daemons.

```
$ for service in /etc/init.d/hadoop-hdfs-*
do
sudo $service start
done
```

8. You now need to create a /tmp directory in HDFS. Some components of Hadoop will need this directory.

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

9. Now we need to create the directories used by the MapReduce components of Hadoop.

```
$ sudo -u hdfs hadoop fs -mkdir \
/var/lib/hadoop-hdfs/cache/mapred/mapred/staging
$ sudo -u hdfs hadoop fs -chmod 1777 \
/var/lib/hadoop-hdfs/cache/mapred/mapred/staging
$ sudo -u hdfs hadoop fs -chown -R \
mapred /var/lib/hadoop-hdfs/cache/mapred
```

10. Then start Hadoop's MapReduce daemons.

```
$ for service in /etc/init.d/hadoop-0.20-mapreduce-*
do
sudo $service start
done
```

11. Check to ensure all five daemons are running.

```
$ sudo jps
```


Copyright © 2010-2013 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

You should see the five Hadoop daemons running. If you do not, ask your instructor for assistance.

12. Finally, create a user directory for the user 'training'.

```
$ sudo -u hdfs hadoop fs -mkdir /user/training
$ sudo -u hdfs hadoop fs -chown training /user/training
```

Test Your Hadoop Installation

We will now test the Hadoop installation by uploading some data.

13. Change directories to the assets directory, and unzip the shakespeare.txt.gz file. Check the file to ensure it has been extracted correctly.

```
$ cd ~/assets
$ gunzip shakespeare.txt.gz
$ head shakespeare.txt
```

14. Create a directory in HDFS and upload the file to that directory. Check to make sure the file is in place.

```
$ hadoop fs -mkdir input
$ hadoop fs -put shakespeare.txt input
$ hadoop fs -ls input
$ hadoop fs -tail input/shakespeare.txt
```

This is the end of the exercise.

Copyright © 2010-2013 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.