VirtualBox and Cloudera Quick Start

Cecilia Earls

7/5/2018

Cloudera Quick Start

Cloudera Quick Start runs on a virtual machine and is really just a learning tool. It is a single-node cluster. So, there is no replication and everything is on the same node. The following includes the steps for setting up Cloudera Quick Start.

Virtual Box

Note: Use 7-zip to extract the zipped files.

1) Download and install virtual box from https://www.virtualbox.org/. Virtual box is functional for many different operating systems.
2) You also are going to want vmware tools. Use commands in *InstallingVMWare* in /Documents/hadoop/SetUp folder on this Mac (also on GitHub) to get these installed easily.

Cloudera Quick Start

1) Navigate to https://www.cloudera.com/downloads/quickstart_vms/5-13.html. *Note: the version may have changed.* Choose virtual box as the platform.
2) Register using the pop-up.
3) Download cloudera-quickstart-vm-5.13.zip. Again, extract with 7-zip.
4) Look at the document *install-cloudera-vm.pdf* in the /Documents/hadoop/SetUp folder on this Mac (also on GitHub). Or navigate to https://vgc.poly.edu/~juliana/courses/BigData2015/cloudera-vm.pdf. Follow the instructions in this document to get Cloudera set up and working properly.
5) Cloudera username and password are both *cloudera*

Terminal

1) Pretty much anything you want to do in the terminal, you want to start as logging in as a superuser. Commands:
    a) su
    b) type in *cloudera* as the password
2) Install useful utilities:

```
sudo yum -y install wget git
```

3) Install EPEL repository (to get R) *Note: of course the repository link might have changed. So, check that this is still accurate first.*

```
sudo rpm -ivh https://mirror.colorado.edu/fedora/epel/7/x86_64/Packages/e/epel-release-7-11.noarch.rpm
```

Setting up a Shared Folder Between the Host and Guest

Follow the directions in *SharingFilesBetweenHostandVM* on this Mac or on Github.

Using Sqoop

1. Basically I looked at the directions found on the Cloudera tutorial. The following document contains some information from that tutorial, *Sqoop.docx*. The idea is that SQL is accessed through the command line and Sqoop will allow files to be transferred to the hdfs from SQL.
2. Also look at the site: https://dzone.com/articles/why-fast-data-is-hard-top-9-challenges-ranked-by-2

R Hadoop

1) I did roughly start by following the directions in this slide presentation by Jeffrey Breen. https://www.slideshare.net/jeffreybreen/big-data-stepbystep-part-1-local-vm
2) However, I deviated somewhat from these steps. In particular, I found that I couldn't use bridged (I wasn't connected to the internet on the VM which caused a ton of problems!). So, I used NAT. This may just be something I need to figure out.
3) After systems were set up as above, install R with:

sudo yum -y install R

4) Hadoop Environment variables need to be set. I needed to use different paths than Mr. Breen used. Look at the docment *What I did to install R.* on this Mac or Github
5) The following page will get you started installing RStudio: *InstallingRStudioCentos*.
6) Need to install `libcurl` by using the terminal command:

install yum libcurl.devel

7) Quite a few packages need to be installed in R:
   a) Use command "R" to start R in the terminal.
   b) install.packages("devtools"), install.packages("httr"), install.packages("curl")
8) Install rhdfs, rmr2, and plyrmr (which took quite a bit of time) using the `install_github` function (in devtools - so library(devtools) first):

For example: install_github('RevolutionAnalytics/rhdfs/pkg')

9) Next start by looking at the following website on installing RHadoop on RHEL (red hat) https://github.com/RevolutionAnalytics/RHadoop/wiki/Installing-RHadoop-on-RHEL

a) Determine what has already been done.

b) Set the environment variables in the terminal
   i. sudo nano ~/.bashrc
   ii. edit this file by appending the following:

   export PKG_CONFIG_PATH = $PKG_CONFIG_PATH:/usr/local/lib/pkgconfig/

   export HADOOP_HOME = /usr/lib/hadoop

   export HADOOP_CMD = /usr/bin/hadoop

export HADOOP_STREAMING = /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.15.0.jar

Of course these paths and versions may change. A picture of this configuration is in SettingEnvironmentVars on github.

iii. Source the environmental variables:

source ~/.bashrc (This may need to be done every time you want to use RHadoop - I'm not sure yet.)

c) The package rJava needs to be installed in R, but it requires a 64bit version of Java. Try: https://www.java.com/en/download/faq/java_win64bit.xml (but see the next comment first)

Now, I recall downloading this on my host computer (maybe to do a test on my version of R?), but I think it would also need to be downloaded on the virtual machine's operating system. It looks like to me that Java came with cloudera so this may be a non-issue on the virtual machine.

d) To start with thrift I used a combination of 2 websites: https://thrift.apache.org/docs/install/centos and http://diggdata.in/post/67561846971/fetch-data-from-hbase-database-from-r-using-rhbase. I did all of the updates recommended by the first site; I am not sure it was necessary?

e) Install the *Apache Thrift* dependancies:

yum -y install automake libtool flex bison pkgconfig gcc-c++ boost-devel libevent-devel zlib-devel python-devel ruby-devel openssl-devel

f) *Apache Thrift* needs to be installed. I could not get it to work using the most up to date version of Thrift. The following website suggested installing version 9.0, http://diggdata.in/post/67561846971/fetch-data-from-hbase-database-from-r-using-rhbase . I basically followed his instructions. So, assuming version 9.0 is already downloaded:

i. sudo tar xvzf /home/Downloads/thrift-0.9.0.tar.gz

ii. cd thrift-0.9.0/

iii. sudo ./configure

iv. sudo make

v. sudo make install

g) Make sure hbase is running on cloudera manager. I am not sure this is necessary to check the installation, but it might be.

h) I followed most of the steps from http://diggdata.in/post/67561846971/fetch-data-from-hbase-database-from-r-using-rhba including putting *hbase thrift start* in the command line. But after using this command, I went to step 5, but I installed version 1.2.1 of `rhbase`. Just use whatever version is the current version.

i) Install `ravro`:

i. wget https://raw.github.com/RevolutionAnalytics/rhbase/master/build/ravro_1.0.4.tar.gz (note: the link said 3 but it was really 4)

ii. R CMD INSTALL ravro_1.0.4.tar.gz