README for the StrataConf 2012 Hive Tutorial

February 24, 2012

Dean Wampler dean.wampler@thinkbiganalytics.com @deanwampler.
Jason Rutherglen jason.rutherglen@thinkbiganalytics.com @JasonRutherglen.

Check List

Please come with the following items already installed on your laptop. We'll explain the purpose of each item as we go along, as well as how to obtain and install some of them.

- VMWare, VirtualBox, or KVM.
- The Cloudera Virtual Machine.
- *Secure shell* (ssh) software (strongly recommended, e.g., PuTTY for Windows).
- Everything in the polyglotprogramming.com/Strata2012.zip Zip file for the tutorial.

Advance Preparation for the Tutorial

Since you learn best by doing, this tutorial is hands-on. You'll run Hive commands and queries along with us, using various data sets. Because we'll have limited time during the tutorial, we ask that you setup your laptop beforehand. These instructions will tell you how.

If you have trouble, send Dean email at the above address. We'll also arrive early for the tutorial to offer last-minute assistance and we'll arrive at the conference Monday evening. If you are unable to setup these tools before the tutorial, don't worry; you'll still learn a lot and the materials in the tutorial Zip file will let you try the exercises on your own afterwards. **We will have no time available during the tutorial to debug setup problems.**

Since Hive and Hadoop are primarily designed for server use, it takes a little effort to set them up on a laptop. Currently, they aren't supported *at all* on Windows. Therefore, to make it as painless as possible, we will do the tutorial using a Linux *virtual machine* (VM) with everything already configured. We ask you to download and install it in advance.

You'll run the VM in the **VMWare Player** (Windows or Linux - free), **VMWare Fusion** (Mac OSX - inexpensive with a free, 30-day evaluation), the freeware **VirtualBox** (available for all platforms), or **KVM** (Linux).

NOTE: We recommend using VMWare for this tutorial instead of VirtualBox and KVM, because we have done more testing using VMWare.

If you already have Hive and Hadoop installed on your laptop or a virtual machine running Hive and Hadoop, make sure you're running recent versions of these tools and that you can complete the **Sanity Checks and Final Steps** below.

Here's what you need to do:

Download and Install VMWare, VirtualBox, or KVM

Pick the VM host system you want to use, then go to one of the following links to download the version for your operating system:

- **VMWare Player:** (Windows and Linux) downloads.vmware.com/d/info/desktop_end_user_computing/vmware_player/4_0.
- **VMWare Fusion:** (Mac OSX) vmware.com/products/fusion/overview.html.
- **VirtualBox:** (All operating systems) virtualbox.org/wiki/Downloads. Also download and install the "extension pack" available.
- **KVM:** (Linux) www.linux-kvm.org/page/Main_Page.

Download the Cloudera Virtual Machine

Cloudera has a VM image of their CDH3u2 Hadoop release for VMWare, VirtualBox, and KVM. It uses the CentOS variant of Linux with a bare-bones GUI. It has many Hadoop ecosystem tools already installed, configured, and running when you start the OS.

- Go to ccp.cloudera.com/display/SUPPORT/Cloudera's+Hadoop+Demo+VM.
- Find the correct VM for your chosen VM host system and download it. All of them are big, so be sure you have a good Internet connection and plenty of time!

Expand the archive once you have downloaded it. Follow the instructions for your VM host system for loading the VM. For example, with VMWare, you can double-click the `cloudera-demo-vm.vmx` file. With VirtualBox, you create a new VM and specify the cloudera "hard drive" in the opening wizard dialog.

Tips:

- Allocate at least 512MB for the VM and preferably 1GB.
- When you're prompted to specify a "hard disk", use the Cloudera VM "hard disk".

Sanity Checks and Final Steps

Let's make sure you have the VM ready to go and then finish installing the tutorial software.

TIP: The VM will grab your mouse when you use it in the GUI. Press either the Control, Command (Mac OSX), or the Alt key (and sometimes a combination of the two) to release the mouse. The bottom right corner of the VM window indicates the key combination to use.

- Start VMWare, VirtualBox, or KVM.
- Start the Cloudera VM.
- Log in as user `cloudera`, with password `cloudera`. (You might get logged in automatically.)

After the VM starts, there are extra tools you should install:

- If using VMWare, select the *Virtual Machine -> Install VMWare Tools* menu item.
- If using VirtualBox, select the "Device -> install Guest Additions", then open a terminal window and enter the following command:

  sudo /media/VBOXADDITIONS_4.1.8_75467/VBoxLinuxAdditions.run

(Hit return until it finishes.)

Install the Tutorial Software

- Double click the Earth icon at the botton of the VM window to open a browser *inside* the VM.
- Enter the following URL, which downloads the tutorial presentation, the exercises, and the data to the VM:

http://polyglotprogramming.com/Strata2012.zip

- Unzip the archive.
- Open a terminal window in the VM (double click the black rectangle at the bottom).
- Change to the "root" directory of the expanded archive.
- Run the following command, watching for error messages:

  setup.sh

- To verify that it worked, run this command:

  hadoop fs -ls /data

You should see something like the following listed at the end:

```
drwxr-xr-x  3 cloudera  supergroup    0 2012-02-24 08:00 dividends/
drwxr-xr-x  3 cloudera  supergroup    0 2012-02-24 08:00 employees/
drwxr-xr-x  3 cloudera  supergroup    0 2012-02-24 08:00 shakespeare/
drwxr-xr-x  3 cloudera  supergroup    0 2012-02-24 08:00 stocks/
drwxr-xr-x  3 cloudera  supergroup    0 2012-02-24 08:00 twitter/
```

Final Checks

In the terminal window, we'll type a few commands using the `hive` command-line interface (CLI). The following is a transcript: the `[cloudera@localhost ~]$` is the Linux prompt, while the `hive>` is the Hive CLI prompt and everything else is output from commands:

```
[cloudera@localhost ~]$ hive
Hive history file=/tmp/cloudera/hive_job_log_cloudera_201202231101_1447274284.txt
hive> show tables;
OK
Time taken: 4.157 seconds
hive> exit;
[cloudera@localhost ~]$
```

If all these steps worked, you're ready to go!

About the Exercises

This tutorial is adapted from a longer Hive + Pig course taught by Think Big Academy. We have included all the Hive exercises from this course, so you'll have plenty to play with afterwards. We'll walk through parts of most of the exercises in the tutorial. Note that the exercise solutions are described in solutions/Hive Exercise Solutions.pdf.