CSE4334/5334 Data Mining

Spring 2015, Prof. Chengkai Li

Department of Computer Science and Engineering
University of Texas at Arlington

Setting Up the Environment for Programming Assignment 3 (P3)

In this assignment, you will write MapReduce programs in Python. You will use Apache Hadoop, an open-source implementation of Google's proprietary MapReduce system. Since we don't have a cluster to use for this course, you will set up a single-node Hadoop environment on your own personal computer. The programs you write will work in a real cluster. It is just that you won't be able to observe performance advantage against solving the problems in a centralized system. In fact, you will observe worse execution efficiency, since the overhead of Hadoop environment cannot pay off in a single-node setup.

Setting up your own Hadoop environment can be non-trivial, even if it has only one node. To save the hassle, we will use a virtual machine made ready by Hortonworks. The virtualization software we will use is VMware. The Hadoop virtual machine we will use is Hortonworks Sandbox. There are other virtualization software and Hadoop virtual machines. Our following discussion is based on the setup of VMware Player + Hortonworks Sandbox.

You need at least 15GB free space on your hard drive.

1 Setting Up the Environment

Step 1.1: Enable BIOS Support for Virtualization

Here is an example of how to do it: `https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/7_0`.

This probably is only necessary if you have a PC. It appears to be irrelevant to Mac, but I couldn't verify. If you encounter troubles using an Mac, this page might have some useful information for you: `http://bit.ly/1BZoaKe`.

Step 1.2: Download and Install VMware Player 7 from `https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/7_0`

VMware Player 7 is what I used and tested for this assignment. Other versions may work as well. For instance, if you have an Mac, it seems that you need to use VMware Fusion.

Step 2.3: Download Hortonworks Sandbox HDP 2.2 from `http://hortonworks.com/products/hortonworks-sandbox/#install`

According to this page, this virtual machine should work on 32-bit and 64-bit OS (Windows XP, Windows 7, Windows 8 and Mac OSX).

Step 2.4: Install Hortonworks Sandbox HDP 2.2 in VMware

From the above page, you can see that, for VMware, they only have "Install Guide" for Mac, at `http://bit.ly/19tJfpK`. But the steps for Windows are pretty similar. You don't really need a guide. Just "Open" the file Sandbox_HDP_2.2_VMware.ova in VMware. After a while, you should be able to see the screenshot in Step 9 of the Install Guide. The virtual machine with Hadoop environment is ready.