**Creating two node Hadoop cluster (Master and Slave)**

Host OS: Windows 7

GuestOS: Fedora 16, 64 Bit

Technology: Oracle VirtualBox and Vagrant

1. Install VirtualBox (<http://www.virtualbox.org>) latest version in Windows 7.
2. Follow instructions in this URL to first install Fedora 16 VirtualBox : <http://pyfunc.blogspot.in/2011/11/creating-base-box-from-scratch-for.html>
   1. Follow till Step 18 in the URL.
   2. Use Fedora 16, 64 bit ISO file and install Fedora Linux on it.
   3. Do NOT use live version of Linux ISO file.
   4. While installing, Do NOT install OS GUI. This is NOT needed and will increase Vagrant’s BaseBox size when we create it.
   5. During OS installation system asks if you want to install Server, desktop or minimal versionof OS.
      1. Say “minimal” installation.
   6. Install Vagrant tool in Fedora VM.
   7. Put the Virtualbox installation path to environment variable (PATH)
   8. You can remove installation steps for Puppet and Chef if not needed.
   9. Last command will be “mv package.box Fedora-16.box”
3. Now you have a Fedora 16’s Vagrant BaseBox (Feoxdora-16.box)
4. Pick a folder for your work. Assuming D:\vagrant\hadoop for the purpose of this documentation.
   1. Work in this folder from now on.
5. Create VirtualMachines from BaseBox (Fedora-16.box)
   1. Create “master” base box
      1. Vagrant box add “master” “<Path to the Fedora-16.box>”
      2. Vagrant init
   2. Create “slave” base box
      1. Vagrant box add “slave” “<Path to the Fedora-16.box >”
      2. Vagrant init
6. Folder will now have a vagrantfile. Replace the vagrant file with our vagrant file.
   1. Our vagrant file will start the master as “Hadoop Master” and slave as “Hadoop Slave”
   2. Hadoop master will have the IP: 10.10.10.50 and slave will have 10.10.10.51
7. Do “vagrant up”
   1. Both master and slave machine should be up and running and should be visible in VirtualBox UI as “Hadoop Master” and “Hadoop Slave”
8. Logging into the machines
   1. On Windows you can use Putty-Gen to create a private key (.ppk) file from the Private key which is in .vagrant.d folder
   2. You can access the machines with the IP we have given.
   3. You will get the shell when you login.
9. Do the following on both Master and Slave machines to install JDK and Hadoop
   1. Go to the site: <http://www.oracle.com/technetwork/java/javase/downloads/jdk7u9-downloads-1859576.html>
   2. Accept lincense
   3. Download file [**jdk-7u9-linux-x64.rpm**](http://download.oracle.com/otn-pub/java/jdk/7u9-b05/jdk-7u9-linux-x64.rpm)
   4. Create a folder ~/bin and copy jdk\_fedora.sh and Hadoop.sh into this folder.
   5. Run command “sudo usermod -a -G root vagrant” – This will add vagrant user to root group.
   6. Run jdk\_fedora.sh
   7. Run Hadoop.sh
   8. To the file ~/hadoop/conf/Hadoop-env.sh file add the line
      1. “export JAVA\_HOME=/usr/lib/jvm/java-7-oracle”
10. Now you are ready to configure Hadoop on both machines.
11. Changes on Master
    1. Edit file ~hadoop/conf/masters
       1. Remove localhost
       2. Add line “master”
    2. Edit file ~hadoop/conf/slaves
       1. Remove localhost
       2. Add line “slave”
    3. Edit file .ssh/id\_rsa
       1. Add the data in private key that will be there in .vagrant.d folder
       2. This will allow auto login from master to slave.
    4. Edit /etc/hostname file and change “localhost” to “master”
    5. Edit /etc/hosts file
       1. Remove localhost lines.
       2. Add line “10.10.10.50 master”
       3. Add line “10.10.10.51 slave”
12. Changes on Slave
    1. Edit file ~hadoop/conf/masters
       1. Remove localhost
    2. Edit file ~hadoop/conf/slaves
       1. Remove localhost
       2. Add line “slave”
    3. Edit /etc/hostname file and change “localhost” to “slave”
    4. Edit /etc/hosts file
       1. Remove localhost lines.
       2. Add line “10.10.10.51 slave”
13. Do the following commands
    1. Vagrant halt
    2. Vagrant up
    3. This will bring up both master and slave machines with the changes we did.
14. Login to Master machine
    1. Create folder ~/HDFS
    2. Run the command “hadoop namenode –format”
       1. This will create HDFS under ~/HDFS folder
    3. Run command “start-dfs.sh”
       1. This will start daemons
          1. NameNode and SecondaryNameNode on master
          2. DataNode on slave
    4. Run command “start-mapred.sh”
       1. This will start daemons
          1. JobTracker on master
          2. TaskTracker on slave
15. Running command “jps” on master should give the three master daemons
16. Running command “jps” on slave should give the two slave daemons.
17. On master
    1. Run command “hadoop fs –put <filename> /” to put a file in HDFS
    2. Run command “hadoop fs –ls /” to check if you can see the uploaded file in HDFS.
18. Two node Hadoop cluster is configured now on two VMs
19. Using “vagrant package” command to create your own BaseBoxes “hmaster” and “hslave” to create another Hadoop cluster.
    1. Delete downloaded tar/gz files JDK and Hadoop from both master and slave.
    2. Run Command “apt-get clean”
       1. Tmp files and cached system files will be cleaned. You don’t need to export these to basebox. This will reduce the size of BaseBox.
    3. Run command “vagrant package master –vagrantfile Vagrantfile –output hmaster.box
    4. Run command “vagrant package slave –vagrantfile Vagrantfile –output hslave.box
    5. Now you can use these BaseBoxes and create two other VMs that will have Hadoop installed and configured on it.

**Sharing files between host and guest machines:**

1. Go to VirtualBox UI and open. Right click on the machine that you want to share files with
   1. Select setting and click on shared folder.
   2. Remove the shared folder listed. (If nothing listed then not a problem.)
   3. Add a new shared folder by clicking on “+” sign on the side.
   4. Give Folder Path as “D:\Vagrant\hadoop”
   5. Give Folder Name as “v-root”
   6. Uncheck read-only and auto-mount
   7. Login to the VM box that you created
      1. “cd /home/vagrant”
      2. “mkdir v-root”
      3. “sudo mount -t vboxsf v-root v-root”
      4. “cd v-root”
      5. “ls –la” – You will see files in this folder that are stored in D:\Vagrant\hadoop” folder on Host machine.

**Reference Sites:**

* Vagrant BaseBoxes for ready download
  + <http://www.vagrantbox.es/>
* Creating Vagrant BaseBox from scratch
  + <http://pyfunc.blogspot.in/2011/11/creating-base-box-from-scratch-for.html>
* Sharing files between Host and Guest OS in VirtualBox
  + <http://helpdeskgeek.com/virtualization/virtualbox-share-folder-host-guest/>