HADOOP on LUBUNTU – *By Edgar Ruiz*

1. Install VM Ware Player
2. Download Lubuntu Linux distro
3. Create new VM Ware instance named Hadoop
   * Use the Lubuntu distro ISO as the source
   * Switch the OS type from Ubuntu to Ubuntu-64
4. Run the installation process of Lubuntu in VMWare
   * Use “hduser” as the default user (helps with instructions and other online help articles)
   * Choose “newuser” as password
   * After the installation is completed, accept the Software Updates
5. *Optional* – Change the resolution to 1024x768
6. In the new Linux desktop run System Tools/Synaptic Package Manager
   * Install “default-jdk” (Find it via a Search for “jdk”)
   * Install “gedit”
7. Download the latest Hadoop package
   * <http://www.apache.org/dyn/closer.cgi/hadoop/core>
   * Select the link to the recommended mirror
   * Select the “current” folder
   * Download Hadoop-#.#.#.tar.gz file. It will be the largest in the folder (200MB+)
8. Extract the file to the Desktop and rename the folder to Hadoop
9. Open the terminal emulator (Accessories/LXTerminal)
10. Copy the hadoop folder to usr/local
    * sudo mv Desktop/Hadoop /usr/local/hadoop
11. Add the following sub-folders in the usr/local/hadoop folder, using the command: sudo mkdir
    * app/
    * app/data
    * app/name
    * app/tmp
12. Configure the SSH server (back in the Terminal Emulator)
    * sudo apt-get install openssh-server
    * ssh-keygen -t dsa -P ‘’
    * cat $HOME/.ssh/id\_dsa.pub >> $HOME/.ssh/authorized\_keys
    * ssh localhost
      + If it ask to continue type “yes”, it means localhost works!
    * Reboot system
13. Let Hadoop know where the Java SDK is located
    * cd /usr/local/hadoop/etc/hadoop
    * sudo gedit hadoop-env.sh
    * In the file
      + Comment out the current JAVA-HOME line
      + Add a new line: “export JAVA\_HOME=/usr/lib/jvm/default-java”
    * Save and then Close
14. Connect the Temp, DataNode and NameNode to their folders
    * Open the “hdfs-site.xml” file, found in /usr/local/hadoop/etc/hadoop
    * Add the following to the “configuration” markup
      + <property><name>dfs.datanode.data.dir</name><value>app/data</value></property>
      + <property><name>dfs.namenode.name.dir</name><value>app/name</value></property>
    * Open the “core-site.xml” file in the same folder
    * Add the following to the “configuration” markup
      + <property> <name> hadoop.tmp.dir </name><value>app/tmp</value></property>
      + <property> <name> fs.default.name </name><value>hdfs://localhost:9000</value></property>
15. Start Hadoop
    * /usr/local/hadoop/bin/hadoop namenode –format
    * /usr/local/hadoop/sbin/start-dfs.sh
    * Jps
      + The PIDs for NameNode, DataNode and Jps should be listed

EXCEL 2013 with Power Query

1. Start the Hadoop cluster
   * sbin/start-dfs.sh
2. Download or transfer file to a standard Lubuntu folder
3. Add a file to the Hadoop file system
   * Open Terminal server
   * Go to the Hadoop folder (cd)
   * Create a folder called “data” (can be any name)
     1. bin/hadoop fs –mkdir /data/
   * Copy the file to the HDF file system
     1. bin/hadoop fs –put $HOME/Desktop/[Name of the file] /data/
4. Switch to the Host PC (not the VM)
5. Open Excel and go to the Power Query tab
6. Select “From Other Sources” and then “From Hadoop File (HDFS)”
7. Enter the VM’s IP address and then click OK
8. In the grid, right-click on the file and select “Drill Down”
9. If the error of “No host found” displays, add the IP to the Windows Host file
   * Open Notepad – Run as Administrator
   * Click Open
   * Go to C:\Windows\System32\drivers\etc
   * Select Host
   * Add a new line of the file
     1. VM’s IP address
     2. [Space]
     3. Name of the host from the error (mine was ‘hduser-virtual-machine’)
   * Refresh the Power Query
   * Try the Drill Down again

HIVE

1. Go to the Apache.org website and download the latest Hive
   * Select the link to the recommended mirror
   * Select the “stable” folder
   * Select the “bin-tar-gz” not the “src-tar-gz”
2. Extract the file to the Desktop and rename the folder to “hive”
3. Open the terminal emulator (Accessories/LXTerminal)
4. Copy the “Hive” folder to opt/hive
   * sudo mv Desktop/hive /opt/hive
5. Open the /etc/environment file
   * Add “:opt/hive/bin” to the end of the string
6. Add the Hadoop default location
   * cd $HOME
   * sudo gedit .bashrc
   * Add to the bottom of the file “export HADOOP\_HOME=/usr/local/hadoop”
7. Reboot
8. /usr/local/hadoop/sbin/start-dfs.sh
9. /opt/hive/bin/hive

Simple Hadoop cluster

1. Create 2 VM’s following the Single Node instructions
2. Start and logon to both Lubuntu VMs
3. In the Terminal Emulator
   1. Write down the IP address using “ifconfig”
   2. Edit the host file
      1. cd /etc
      2. sudo gedit hosts
      3. Add 2 lines
         1. [master’s IP] master
         2. [slave’s IP] slave
      4. Save and close
   3. From the Master, copy the SSH certificate to the Slave
      1. ssh-copy-id -i $HOME/.ssh/id\_rsa.pub hduser@slave
      2. It will prompt for Slave’s password, it should be “newuser”
   4. Logon to Master from Master and then Slave from Master
      1. ssh master
      2. ssh slave