**Exam 2 - Part 2, Applying Big Data with Cloudera QuickStart VM**

**Objective:** The objective of this tutorial is for you to gain exposure to Hadoop, its framework, and its associated tools like Hive, Spark, and HDFS through Cloudera’s QuickStart virtual machine. The exposure, although prescriptive, will give you the opportunity to see how Hadoop works in action with a sample data set.

**Points: 35**

**How you will be assessed:**

Completing the steps for the tutorial is not enough to earn full points for this assignment. Perfect work not only includes well-documented screenshots and notes of what was complete, but also connects the tutorial to your learning in this class related to Big Data. You are encouraged to refer back to the textbook, readings, and outside research to inform your work.

Submit your assignment in a separate word document from Exam 2, Part 1 and upload to Exam 2 in Moodle.

**Note:** This must be done on a SCO 312 lab computer. Unless your computer has 16gb or more of RAM, do not attempt this on your personal machine. If you do this on your own machine, it will require that you also have Oracle VirtualBox installed.

**Pre-work: Getting the VM downloaded and configured in VirtualBox (5 pts)**

**Step 1:** [Download Cloudera QuickStart VM](https://www.cloudera.com/downloads/quickstart_vms/5-13.html) - It’s a 5.5gb file, so it will take some time. Select VirtualBox as the platform and complete the information form.

**Step 2:** Go to your Downloads folder, unzip coudera-quickstart-vm-5.13.0-0-virtualbox.zip. Then double click the cloudera-quickstart-vm-5.13.0-0-virtualbox ovf file. The VirtualBox appliance dialog box should appear as below.



**Step 3:** Change the RAM value to 8192 MB. Then click Import.

**Step 4:** Launch the Cloudera virtual machine

[Documentation:](https://www.cloudera.com/documentation/enterprise/5-13-x/topics/cloudera_quickstart_vm.html) The following contains documentation on user account information and other administrative information regarding your Cloudera VM. You may need to reference this later in the tutorial for things like usernames and passwords.

**You are ready to begin the Cloudera tutorial**

**Exercise 1 (5 pts) and 2 (5 pts)**

**Documentation and Connection to Big Data (10 pts).**

Follow the steps contained in the [Cloudera QuickStart VM Tutorial](https://www.cloudera.com/developers/get-started-with-hadoop-tutorial/setup.html) (Exercises 1 and 2) . For each step, take screenshots along the way to show your progress, and place them into a Word document. Document successful messages and failure messages. It is likely you may run into errors and that you will need to troubleshoot this process along the way. Document the sources you reference and the steps you complete in order to resolve the issue. Although this is an exam, it is OK to discuss the lab with your classmates and solicit assistance as needed. However, all of the work needs to be done individually.

At the end of each step, connect the step back to the course material, specifically, the text. For instance, If you use Spark in a specific step, discuss how Spark is used (and when), and how it fits into the Hadoop ecosystem. What tasks is Spark particularly good at accomplishing? Your job is not to just go through the tutorial, but to connect the dots between the course material and the lab to give you greater understanding for the material. Simply completing the lab and providing screenshots will not earn you full points.

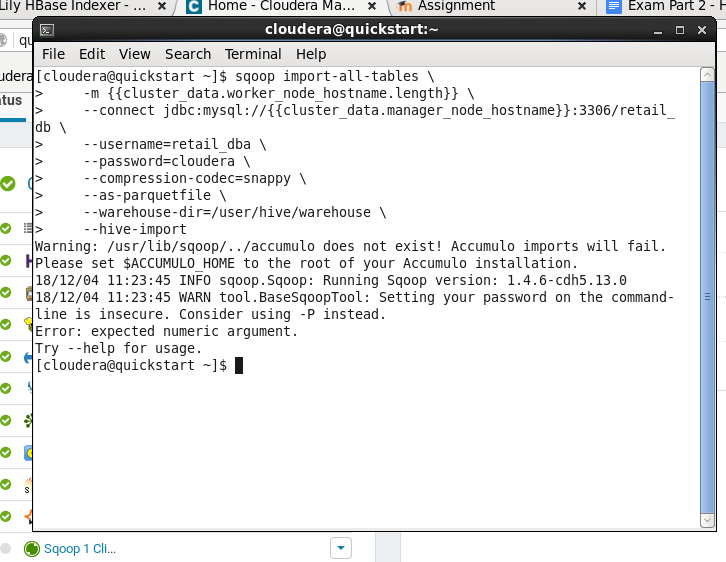
**Last Step (5 points)**

Answer the following questions at the end of your Word document.

1. What did you like the most from these 2 exercises? The least?
2. After completing the first 2 exercises, do you have a desire to continue and complete the remaining exercises in the tutorial? Why or why not?
3. How did this tutorial help you better understand Hadoop and its associated tools?

**Notes / Troubleshooting**

1.



If you get the following error message found above when executing the sqoop command on exercise 1, complete the following through a command window.

sudo mkdir /var/lib/accumulo

ACCUMULO\_HOME=’/var/lib/accumulo’

export ACCUMULO\_HOME

Then, re-run the sqoop command.

--------------------------------------------------------------------------------------------------------

2.

In exercise 2, during the CREATE EXTERNAL TABLE command, the job may be executing into infinity (i.e. it never stops). If so, check out the following Cloudera thread:

<http://community.cloudera.com/t5/Hadoop-101-Training-Quickstart/Exercise-2-Hive-data-load-not-completing/td-p/34673>

A poster makes the assertion that shutting down the VM, increasing the RAM to 10gb, and restarting the VM may resolve the issue. You may need to experiment with this to see if it gets you the results you want. I wouldn’t recommend allocating more than 12gb RAM to the Cloudera VM

--------------------------------------------------------------------------------------------------------