## HU Extension School E-63 Big Data Analytics

## Assignment 02

### Handed out: 02/06/2016 Due by 11:30 PM on Friday, 02/12/2016

Capture all steps of your implementation with comments indicating what are accomplishing with every step. Place those in this MS Word document bellow the problem statement. Please send comments and questions to the Discussion Forum on the class site.

### 

**Problem 1)** Please, download and install VMware Workstation 11 on your 64 bit Windows PC or VMWare Fusion 7, if you are on a MAC. Please download 64 bit CentOS6.7 and create a 64 bit VM. If you know what you are doing and you work with another flavor of Linux supported by CDH5.5.1, please be free to create a virtual machine based on your favorite Linux flavor. Provide your virtual machine with some 40GB of disk space, if you can spare it. For whatever reasons, Hadoop installation appears to prefer to have more than 20 GB of available space. Name the main user of your VM cloudera. Do not use name hadoop. “hadoop” is a bad name for a user, since Hadoop framework has an executable called hadoop and it creates many directories with that same name and those would not necessarily be owned by the VM user called hadoop. On that VM create yet another user called joe. Make both users sudo users. Once your CentOS is fully installed, please shut the VM down and make a copy of the entire directory containing that VM. Name the folder containing that copy differently. Two VMs are identical and you could even run them simultaneously if your machine has enough memory. In the folder of each VM add a text file describing OS on your VM, usernames and passwords of important users. This little file will make your VMs useful long into the future. The reason you are creating the backup VM is to save time, if you damage the one VM on which you are installing your software.

Please do not capture installation of Workstation 12 or Fusion. Please do not capture every step in creation of VM. Show addition of the second network adapter. Show steps in creation of user joe. Demonstrate that joe is a sudo user. Show results of your ifconfig command.

**Problem 2)** Use one of above VMs and follow closely steps in the CDH5.5.1 Quick Start Guide, or my notes. PDF and PPT formats and characters on PC do not always map well into Unix (Linux) characters. If you want to copy commands from the guide you are better off doing it from the HTML version of the CDH Quick Start Guide, which you could find at:

[http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cm\_qs\_quick\_start.html](http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cm_qs_quick_start.html%20) and open from with your VM.

Both my notes and the Quick Start Guide will lead you through a “semi-automated” process of installing Hadoop. Please install YARN version of Hadoop. My notes add a few explanations beyond what you can see in the Cloudera’s guide. Read the notes and the guide very carefully. Do not execute commands for flavors of Linux other than RedHat (CentOS) unless you are working with another flavor purposefully. You will know that you have successfully installed Hadoop if all of tests described in the guide work properly.

**Problem 3)** As your new Linux user joe fetch the .txt version of James Joyce's Ulysses by issuing the following command on the command prompt:

wget <http://www.gutenberg.org/files/4300/4300.zip>

Unzip the file. Open the resulting txt file with Vi and convince yourself that the life of Buck Mulligan is in front of you. Create a HDFS directory called ulysses and copy the .txt file into that HDFS directory. Do not create another HDFS directory called counted. The Map Reduce job you will run will create that directory for its output. Actually, if the directory preexists the job will raise an error. That same hadoop-mapreduce-examples.jar file mentioned in class notes and you used as the final proof that MapReduce works contains another program called wordcount. wordcount will tell you how many times a word appears in a provided text. Invoke wordcount by the following command:

$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount ulysses counted

Once the job is finished visit site <http://localhost:19888>. You will see some statistics on MapReduce jobs executed on your cluster. There will not be much for your short job. In general that is a very useful site.

Copy results of word count analysis to the local file system. Write a small program in any language (or scripting tool) of your choice and order the counting results by the decreasing count. Present the portion of your final result which does not contain so called stop words (the, a, and, or, …) in your report. Submit top 200 words in separate .txt file with your report.

**Problem 4).** Consider a symmetric matrix

A =

Using R demonstrate that all three eigenvectors of that matrix are mutually orthogonal. Let be the matrix of eigenvectors of matrix A. Calculate product of tree matrices:

A Λ

Symbol T indicates the transpose matrix. Google around for properties of eigenvectors and eigenvalues of real symmetric matrices. What is the general statement you can make about the observation on the value of the above product. Include copies of your R commands in your MS Word report.