## McKesson Assignment 12 Deep Azure

## 

### Handed out: 01/13/2018 Due by 11:59 PM, midnight (CST) on Saturday, 01/20/2018

**Problem 01.** Create your own Hadoop cluster using HDInsight service. Demonstrate that you can establish an ssh session with the master node of that cluster. Tell us where is the hadoop executable by issuing command $ which Hadoop. Tell us which version of Python and Java are installed. Demonstrate that you have wiped out all components of that cluster.

(25%)

**Problem 02** Download and import Hortonworks HDP 2.6.3 sandbox VM. **Create a new** Linux user centos. Create HDFS home directory for that user. Move attached 4300-0.txt file to a subdirectory in centos’ HDFS home directory. Run Hadoop MapReduce grep program on that file. Copy the result of that analysis to your local file system and examine top 10 and bottom 10 lines.

(25%)

**Problem 03.** Use Spark 1.6.3 to find out the number of lines in 4300-0.txt which contain word “future”. Repeat the exercise with Spark 2.2.

(25%)

**Problem 04**. Use hive command line client to demonstrate that you can import bible word frequencies and Shakespeare word frequencies into respective tables bible and shakespeare. Convince yourself that you can run a select statement joining those two tables.

(25%)

SUBMISSION INSTRUCTIONS:

Your main submission should be a MS Word or PDF document containing descriptions of your action while configuring Azure services. If your MS Word document is larger than 1 MB, save it as a MINIMIZED PDF. Please be merciful and capture small JPGs. Describe the purpose of every action and the significance of the results. Start with the text of this homework assignment as the template. Please add the entire text of your JAVA, C# or Python programs to the end of your MS Word/PDF document. Please write your solution as if you are writing a tutorial for your colleagues. Please make your text readable. Make sure that your fonts, especially in captured images are not unreadable. Please do not provide ZIP or RAR or any other archives. Canvas cannot open those archives and they turn into a nuisance for us.