## HU Extension Assignment 06 E63 Big Data Analytics

### Handed out: 10/06/2017 Due by 4:00 PM EST, 10/14/2017

In your solution, please leave the text of every problem as presented here. Add your solution below the problem statement. It is important for us and we will take points if you ignore this request. Please make sure that you provide numeric or textual results of your calculations. If there are no results, we will treat the problem as not addressed. Just providing some code without results will give you 0 point.

**Problem 1)** Lecture notes contain script network-count.py in both Spark Streaming APIand Spark Structured Streaming API. Use Linux nc (NetCat) utility to demonstrate that scripts work. Run both scripts on your own VM with Spark 2.2 installation. Cloudera VM with Spark 1.6 does not have Spark Structured Streaming API.

(20%)

**Problem 2)** Expand provide orders.tar.gz file. Also, download shell scrips splitAndSend.original.sh and splitAndSend.sh and the Python script count-buys.py. First run splitAndSend.original.sh and count-buys.py. Record the failure mode of count-buys.py. Simply read the error message produced and tell us what is happening. Then run script splitAndSend.sh and Python program count-buys.py and tell us what the results are. In both cases show use contents of your HDFS directories input, output and staging. The second script splitAndSend.sh is supposed to reduce or eliminate the race condition. You might want to rename HDFS directory output from the first run in order to preserve it’s content. In both cases, show the partial contents of your HDFS directories input, output and staging. In the second run, locate an output file named part-00000 that is not empty and show its content to us. Run these experiments on Cloudera VM. You need HDFS for these programs to run.

(30%)

**Problem 3).** In the second run of the previous problem you will notice that many of part-00000 files in your output directory are empty. Could you explain why.

(10%)

**Problem 4)** Could you rewrite count-buys.sh in Spark Structured Streaming API. If you do that change script splitAndSend.sh to move generated chunks from the local files system directory staging to local file system directory input. Run this experiment on your VM with Spark 2.2.

(20%)

**Problem 5)** Examine provided Python program stateful\_wordcount.py. Make it work as is. If there are errors on the code, fix them. Modify the code so that it outputs the number of words starting with letters a and b. Demonstrate that modified program work. You should provide several both positive and negative examples.

(20%)

You are welcome to implement your solution in any language of your choice.

You are welcome to follow any other instructions and use any other programming or scripting language to accomplish the above goals.

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used as separate files. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. Please, submit to the class drop box. For issues and comments visit the class Discussion Board. You are not obliged to use Java or Eclipse. You are welcome to use any language and any IDE of your choice.