## HU Extension Assignment 08 E63 Big Data Analytics

### Handed out: 03/26/2016 Due by 11:30PM EST, 04/01/2016

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.

Use most excellent and very detailed notes created by Marina Popova for Section 08 on Kafka and Streaming as you main guide for this assignment.

**Problem 1)** On your Cloudera VM or any other VM you might be using install Kafka. Just in case, install one of the recent Kafka 0.8 versions. Demonstrate that you can create a topic, publish messages to that topic and consume messages sent to that topic. Use Kafka command line interface.

**Problem 2)** Using Java or Python or any other (even scripting) language of your choice construct a producer and a consumer object. Let producer generate one random number between 0 or 1 and 10 every second. Let both producer and consumer run indefinitely or until you kill them. Demonstrate that your consumer is receiving messages by printing both the stream if numbers generated on the producer and the stream of numbers fetched by the consumer. You might find it easier to print to files and examine files afterwards. Once you terminate the exchange, examine Kafka’s log.

Instructions on how to write Java producer and consumer you can find on this URLs:

<https://cwiki.apache.org/confluence/display/KAFKA/0.8.0+Producer+Example>

<https://cwiki.apache.org/confluence/display/KAFKA/0.8.0+SimpleConsumer+Example>

Instructions on how to write Python clients for Kafka you could find on this URL:

<https://cwiki.apache.org/confluence/display/KAFKA/Clients#Clients-Python>

Instructions for Scala could be found here:

<https://cwiki.apache.org/confluence/display/KAFKA/Clients#Clients-ScalaDSL>

You are welcome to follow any other instructions and use any other programming or scripting language.

**Problem 3)** Starting from one of the attached Spark Streaming clients DirectKafkaWordCount in Java, Scala or Python write a consumer client that will replace the consumer from the previous problem. However, rather than simply printing every message it receives from the producer, let it print for us every 5 seconds the rolling count of numbers between 1 and 10 it received in the last 30 seconds. You might find it simpler to print to files and then examine those files afterwards. For Java build simple Maven Project with a single Java class and pom.xml file similar to the one provided. Build your projects following the process we used in Assignment 4.

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