METROPOLITAN STATE UNIVERSITY

ICS 613: Introduction to Big Data Computing Systems
Fall 2018

Course Project

This project is to be completed in groups of two students.

The final of your project report is to be submitted to D2L by **11:59pm Saturday, December 8.**

Objective

The goal of this project is for you to demonstrate a good understanding of the big data computing tools and how they are used to analyze data. When evaluating your project, I will consider:

- The creation of a data set (by downloading/aggregating/cleaning, ETL).
- Data analysis and difficulty of computation.
- Presentation, summarization, and visualization of your results.
- Report writing.

Project Steps

- **Step 0 (due 10/13/2018):** Find your partner and send me one email with the names of your group's members.

- **Step 1 (due 10/20/2018): Introduction, problem motivation, and data (at least 300 words)**
  - There are tons of public datasets available online you have to find a dataset that is of interest to you. Hence, instead of starting your search with a dataset, start by thinking about what kinds of problems you think would be interesting. Once you have a topic of interest, it will be much easier to find possible datasets. You can use Google Dataset Search (https://toolbox.google.com/datasetsearch) to look for data sets in the area of your interest.
  - Describe the dataset(s) you choose and include the data source (i.e., from where did you download your data), data size, and schema information. You can also include few lines of your data set for demonstration.
  - Think of a list of ~6-8 interesting questions you can ask on your chosen data set.

- **Step 2 (due 11/3/2018): Computations (at least 500 words)**
  - Choose at least 4 computations to implement over your data, highlight them, and walk through them in depth.
  - For each computation, include high-level diagram and/or explanation of the steps you are going to follow to complete the computation.
  - At least one computation must be implemented in map-reduce with two map reduce jobs where the output of the first job is used as input to the second job.
  - At least one computation must be implemented as a Spark application.
  - The other 2 computations can be implemented either in map-reduce or spark. If you choose Spark, then each computation must have at least a sequence of 5 functions.
  -
- **Step 3 (due 11/17/2018): Implementation**
  - Implement the computations you explained in step 2 above.
  - You must use **Oozie** to run the map-reduce computation that consists of two jobs.
  - Include the source code of your implementation as an Appendix in your report.

- **Step 4 (due 12/1/2018): Results & Presentation**
  - Give a detailed explanation of the results of your analysis tasks, broken down into sections where there is one section for each task you implemented.
  - Include visualizations (e.g., charts) of your results and/or screenshots of your output.
  - Each group will have a 10-minutes presentation to the class on 12/1 and 12/8.