CSCI 5408, Summer 2018

Assignment 5: Using Data Lake Analytics to capture and analyze real time traffic data set and summarize findings

(Issue: July 03, Due: July 17, 11:59PM)

- TA: Dijana Kosmajac (dijana.kosmajac@dal.ca)
- Tutorial: July 04, 11:05-12:25, Room: 127
- Help Hours: Fri, 11:05-12:25, Room 127; Wed, 18:00-20:00, Room: 127

---

1. **Objectives:**
    1) To learn data analytics using MS Azure Data Lake
    2) To learn use of data analytics dashboard
    3) To learn real time data analysis with pattern detection
    4) To learn big data analysis using big data tools available in cloud

2. **Tasks:**
    1) Create a free account on Microsoft Azure Portal (Data Lake Analytics) https://azure.microsoft.com/en-ca/services/data-lake-analytics/
    2) Download a visual analytical tool of your choice (for example Power BI or Tableau) and install it on your local system.
    3) Download **Traffic Signal Vehicle and Pedestrian Volumes Data** dataset (CSV) from following location: https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#7c8e7c62-7630-8b0f-43ed-a2dfe24aadc9
    4) Load the provided data into MS Azure Data Lake using framework of your choice. You can use any given SDK (.Net, Java, Node.js or Python).
    5) Read the documentation provided for the given data set (This dataset contains the most recent 8 peak hour vehicle and pedestrian volume counts collected at intersections where there are traffic signals. The data is typically collected between the hours of 7:30 a.m. and 6:00 p.m.)
    6) You need to perform following queries on the given data. ***For all below queries export the output in CSV format and plot the data using visual analytical tool of your choice.***
        1. Aggregate results based on “Main Street Name” and calculate average volume of vehicles for all available years provided in the dataset. You need to calculate one average value for each individual “Main Street Name”.
        2. Based on past 5 years of data, identify which 10 traffic locations are busiest during peak hours (consider both vehicle traffic and pedestrian traffic).
        3. Aggregate results based on individual year (2017, 2016, 2015, etc.) and calculate sum of vehicles and pedestrians traffic count for all available

locations. You need to do sum on all available locations and group them based on individual years.

4. Considering all historic years of data and all available locations, identify which day of the week (out of 7 days in a week) has been the busiest with vehicle and pedestrian traffic. Export sum of final counts for all 7 days of week for plotting.
5. Aggregate results based on "Main Street Name", identify which day of the week (out of 7 days in a week) has been the busiest with vehicle and pedestrian traffic for each individual location. Include all historic data in observation. [HINT: Group By Main Street Name, Day of Week and calculate SUM(Vehicle Traffic + Pedestrian Traffic)]

7) Write a report including the following sections:
   a. Tool Selection: Describe which visual analytic tool you selected and why. Also briefly explain the capabilities of Azure Data Lake.
   b. Data Loading: Provide detailed steps that you performed in data loading and data cleaning before performing 5 analytical queries in Azure Data Lake.
   c. Dashboard: Provide screenshots of output of all 5 queries after plotting them in visual analytic tool of your choice.
   d. Output: Write a brief summary of different patterns identified from the output of given queries. Should not be more than one page.
   e. Code submission and output data: Provide code (U-SQL) scripts and output result set of all 5 queries in a compressed (zipped) format. Upload them on BrightSpace or add them to your GitHub profile and share the link.
   f. Note: You are allowed to take help and guidance from any code source available online (research papers or GitHub repos). However, you are not allowed to copy-paste code into your program. Please provide references for all external sources/libraries including data source and GitHub repositories that you used for performing this assignment.

3. Submit your Ass5 report electronically:

1) Please use Bright Space to submit your assignment.
2) Please also provide any additional scripts you used in this assignment (all uploaded to GitHub in a single repository).

*** Plagiarism and Intellectual Honesty:** (http://plagiarism.dal.ca)

Dalhousie University defines "plagiarism as the presentation of the work of another author in such a way as to give one's reader reason to think it to be one's own." Plagiarism is considered a serious academic offense which may lead to loss of credit, suspension or expulsion from the University, or even the revocation of a degree.