CSCI 5408, Winter 2017

Assignment 3: Apache Spark, Real Time data pipelines and analysis

(Issue: Feb 07, Due: Feb 21, 11:59PM)

- **TA:** Abhinav Kalra (Abhinav.Kalra@Dal.Ca)
- **Ass3 Tutorial:** Feb 08, 1:00-2:30 PM, Room: 1020 Rowe Mgmt.
- **Ass3 Help Hours:** Wed, 1:00-2:30 PM, CS 134; Fri, 1:00-2:30 PM,CS 233

---

1. Objectives:

1) To learn concepts of Big Data Systems running on Clouds
2) To learn ad-hoc and in-memory data analysis
3) To learn using Apache Spark
4) To learn Cloud technology supporting Big Data applications

2. Tasks:

1. Download Apache Spark (http://spark.apache.org/downloads.html) and unzip the tar file. No further installation is required. The easiest way to run Spark is on any Linux distro such as Ubuntu. Alternately Spark can also be run on a VM Linux instance.(Contact Help desk in CS for any technical help)

2. Select a data set to work with:
   a) You can either use the given datasets with this assignment, or datasets of your own, or other data sources, which are suitable for this exercise, e.g.
      1. http://open.canada.ca/data/en/dataset
      2. https://www.kaggle.com/datasets
      3. https://aws.amazon.com/public-data-sets/
   b) Your task would be to pick one dataset which can support features of word count. Import the data into Apache Spark and run your program. You need to perform data cleaning (removal of quotations, comma, case conversion etc. within your application).
   c) You need to write an application which can count distinct words and number of occurrences of each word in the dataset.
   d) You will also create another spark application to run queries and find the below mentioned information from the datasets in step 3 and 4.
   e) Document your application and final outputs in your report.
3. Use Baby Names dataset and write a program (using Python Spark SQL) to capture the following information:
   - Total number of birth registered in a year
   - Total number of births registered in a year by gender
   - Input a year and populate top 5 most popular names registered that year
   - Input a child name and populate total number of birth registrations throughout the dataset for that name

4. Use NYPD Motor Vehicles Collision dataset and perform the following tasks (using Spark SQL):
    - Preprocess and clean dataset if required
    - Capture total injuries and fatalities associated with each motor collision record(identified by a unique incident key)
    - Capture total incident counts in a year (grouped by year)
    - Capture total injuries(can be sum of injuries and fatalities) grouped by year and quarter
    - Capture total injuries(sum of injuries and fatalities) and incident count grouped by Borough, year and month

5. Write a report including the following sections:
    a) Task Description: Introduction and brief application scenario.
    b) Spark Design: Provide an overview of what configuration steps are taken to setup Apache Spark on local system.
    c) Application Queries: Provide description, syntax and run time for each application query executed on Spark
    d) Outputs: Outputs generated in response to word count calculation and Spark SQL queries ran on Spark
    e) Summary: Provide a summary of your work & observations on the application, the developed program and the experience of using the software tools (i.e. your comments & recommendations, etc.).

3. Submit your Ass3 report electronically:

1. Please use Bright Space to submit your assignment
2. In addition to the report submit source dataset(if using your own dataset),python code and output files generated from the application
3. Include a README file
4. Submit all material within one Zip file

*** 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.