Assignment #3

CSCI 5408 (Data Management, Warehousing, Analytics)
Faculty of Computer Science, Dalhousie University

Date Given: Oct 27, 2021
Due Date: Nov 9, 2020 at 11:59 pm

Late penalty: 10% deduction/day

Disclaimer: This assignment requires students to work on Spark framework for unstructured data processing, MongoDb for data storing, and Neo4j graph database for visualization. Submissions related to this assignment will not be used for commercial purposes.

Objective:

- The objective of this assignment is to understand Big Data processing problems, and NoSQL database (document, and graph).

Plagiarism Policy:

- This assignment is an individual task. Collaboration of any type amounts to a violation of the academic integrity policy and will be reported to the AIO.
- Content cannot be copied verbatim from any source(s). Please understand the concept and write in your own words. In addition, cite the actual source. Failing to do so will be considered as plagiarism and/or cheating.
- The Dalhousie Academic Integrity policy applies to all material submitted as part of this course. Please understand the policy, which is available at: https://www.dal.ca/dept/university_secretariat/academic-integrity.html

Assignment Rubric

	Excellent (25%)	Proficient (15%)	Marginal (5%)	Unacceptable (0%)	
Completeness including Citation	All required tasks are completed	Submission highlights tasks completion. However, missed some tasks in between, which created a disconnection	Some tasks are completed, which are disjoint in nature.	Incorrect and irrelevant	Problem #4 (Neo4j)
Correctness	All parts of the given tasks are correct	Most of the given tasks are correct However, some portions need	Most of the given tasks are incorrect. The submission	Incorrect and unacceptable	Problem #2 (

		minor modifications	requires major modifications.		
Novelty	The submission contains novel contribution in key segments, which is a clear indication of application knowledge	The submission lacks novel contributions. There are some evidences of novelty, however, it is not significant	The submission does not contain novel contributions. However, there is an evidence of some effort	There is no novelty	Problem #1
Clarity	The written or graphical materials, and developed applications provide a clear picture of the concept, and highlights the clarity	The written or graphical materials and developed applications do not show clear picture of the concept. There is room for improvement	The written or graphical materials, and developed applications fail to prove the clarity. Background knowledge is needed	Failed to prove the clarity. Need proper background knowledge to perform the tasks	Problem #3

Citation:
McKinney, B. (2018). The impact of program-wide discussion board grading rubrics on students' and faculty satisfaction. Online Learning, 22(2), 289-299.

Tasks

This assignment requires you to submit programming codes on gitlab, and a single PDF file on Brightspace with all the required details mentioned in the steps below.

Problem #1

Step 1: Using your GCP cloud account, configure and initialize Apache Spark cluster.
(Follow the tutorials provided in Lab session).
Note: If for some reason, you fail to work on GCP cloud account (valid reasons required), you need to create local standalone Hadoop/Spark cluster to perform the next set of operations.
Apache Spark will be used for MapReduce at Step 10

Step 2: Create a Twitter developer account
Step 3: Explore the Twitter search API and data format it returns as results
Step 4: Write a Java program to extract data from Twitter. Execute/Run the program on GCP
(Do not use any online program codes. You can only use API specification codes given by Twitter only)
- The search keywords are (not case sensitive) "**weather**", "**hockey**", "**Canada**", "**Temperature**", **"Education"**

You need to include a flowchart/algorithm of your tweet extraction program in the PDF file.

Step 5: You need to extract the tweets and metadata related to the given keywords.

- For some keywords, you may get less number of tweets, which is not a problem. Including all keywords - you should get approximately 2000 to 3000 tweets.
- You should extract tweets, and retweets along with provided meta data, such as location, time etc.

Step 6: The captured **raw** data should be kept (automated process must be done through your program) in MongoDB

- To store raw data, use Database Name: **RawDb**. You can create different collections within that RawDb

Step 7: Your program should automatically clean and transform the data stored in RawDb, and then upload to new MongodB database **ProcessedDb**

- For cleaning and transformation -Remove special characters, URLs, emoticons etc.
- Write your own regular expression logic. **You cannot use libraries such as, beautifulsoup, jsoup, JTidy etc.**

Step 9: You need to include a flowchart/algorithm of your tweet cleaning/transformation program on the PDF file.

Problem #2

Step 10: From the two given Reuter news files (reut2-009.sgm, and reut2-014.sgm), create MongoDb Database – **ReuterDb**, where each **Document** contains a news article. The task must be done using a Java program.

- To perform this operation, you need to write a code to scan the required texts between two **<REUTERS></ REUTERS >** tags, **<TEXT></ TEXT>** tags, and **<TITLE></ TITLE >** tags.
- In addition, you need to include a flowchart/algorithm of your News Artricle extraction/transformation program in the PDF file.

Problem #3

Step 11: Write a MapReduce program to count (frequency count) the following substrings or words. Your MapReduce should perform the frequency count on the stored clean tweets (**ProcessedDb**) and the stored news articles (**ReuterDb**). You can export the ProcessedDb, and ReuterDb content to files and perform the Spark MapReduce job on those files.

- "**Education**", "**Canada**", "**hot**", **"cold",** "**Flu**", "**Snow**", "**Indoor**", **"rain", "ice"**
- You need to include a flowchart/algorithm of your MapReduce program on the PDF file.

Step 12: In your PDF file, report the words that have highest and lowest frequencies.

Problem #4

Step 13: Explore Neo4j graph database, understand the concept, and learn cypher query language
Step 14: Using Cypher, create graph nodes with names of cities, provinces, and territories of Canada. You can visit https://www.worldatlas.com/geography/capital-cities-of-canada-s-provinces-territories.html as a source of information.

- If there are relationships between nodes, then connect them accordingly. E.g. **Halifax** is neighbour of **Fredricton** and **Charlottetown.** Therefore, these will be connected using edges.
- Include your Cypher and generated graph in the PDF file.