Assignment #3

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

Date Given: Jun 23, 2021
Due Date: Jul 5, 2021 at 11:59 pm

Late Submissions are not accepted and will result in a late penalty of 10% deductions / day in the assignment.

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%)	This Rubric Applied to
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 #2
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 #1

		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 #1

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

This assignment requires you to submit programming codes on gitlab, and a single PDF file on Brightspace.

Problem #1: This problem contains three tasks.

Task 1: Cluster Setup - Apache Spark Framework on GCP
(if no GCP credit available – Hadoop or Spark setup in personal Linux machine)

Using your GCP cloud account, configure and initialize Apache Spark cluster.

Create a flowchart or write **½ page explanation on how you completed the task**, include this part in your PDF file.

Task 2: Data Extraction and Preprocessing Engine: Sources – NewsAPI

Steps for NewsAPI Operation

Step 1: Visit the news API https://newsapi.org/
Step 2: Create a developer account
Step 3: Search keywords – "Canada", "University", "Dalhousie", "Halifax", "Canada Education", "Moncton", "Toronto", "

Step 3: Write a well-formed script/program using Java to extract data (**Extraction Engine**) from NewsAPI.

(Do not use any online program codes or scripts, which is not part of the official API documentation and specification.)

Step 4: You need to **include an appropriate pseudocode of your data extraction program in the PDF file**.

Step 5: **The captured raw data should be kept (programmatically) in files. Each file should not contain more than 5 news articles. These files will be needed for "Problem #1-Task 3"**

Step 6: Your program (**Filtration Engine**) should automatically clean and transform the data stored in the files, and then upload each record to new MongodB database **myMongoNews**

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

Step 7: You need to **include a flowchart of Step 6 in the PDF file.**

Task 3: Data Processing using Spark – MapReduce (written in Java) to perform count

Step 1: Write a MapReduce program **(WordCounter Engine)** to count (frequency count) the following substrings or words. Your MapReduce should perform the frequency count on the stored raw news files (titles and contents of the news articles)

- "**Canada**", "**Nova Scotia**", "**education", "higher", "learning", "city", "accommodation", "price" -** (case sensitive)
- You need to include a flowchart/algorithm of your MapReduce program on the PDF file.

Step 2: In your PDF file, report the words that have highest and lowest frequencies (it must be computed programmatically).

Problem #2: This problem contains one task

Task 1: Data Visualization using Graph Database – Neo4j for graph generation

Step 1: Explore Neo4j graph database, understand the concept, and learn cypher query language
Step 2: Visit NovaScotia parks website that you used in Assignment 1.
Step 3: Using Cypher, create graph nodes with

- names of each region (e.g. Cape Breton Island Parks) as node, and
- names of parks as nodes.

You should add properties to the nodes. For adding properties, you should check the dataset that you used in Assignment 1. E.g. location, street name, size etc. could be added as properties

- All regions are parts of Nova Scotia, so all regions should be connected using edges.
- Each region has multiple parks, and therefore, there should be edges between parks and the region.
- Once the graph is constructed on Neo4j - using cypher language, find which region has more number of parks. Provide the screenshot on the PDF file.
- **Include all your Cyphers (graph construction, find query etc.) and generated graph image in the PDF file.**

Assignment 3 Submission Format:
1) Compress all your reports/files into a single .zip file and give it a meaningful name.
You are free to choose any meaningful file name, preferably - **BannerId_Lastname_firstname_5408_A3** but avoid generic names like assignment-3.
2) Submit your reports only in PDF format.
Please avoid submitting .doc/.docx and submit only the PDF version. You can merge all the reports into a single PDF. **You should also include output (if any) and test cases (if any) in the PDF file.**
3) Your executable code/script needs to be submitted on https://git.cs.dal.ca/