Assignment 1 (10%)
Date Given: Sep 16, 2019
Submission Due: Sep 29, 2019 at 11:59 pm (midnight)
**** Late submissions are not accepted and will result in a 0 on the assignment**

Objective:

This assignment covers concepts related to data modelling (conceptual design), and planning of a hypothetical data management project. Consider this assignment as the first phase of an industry project. The designed model and data gathered in this assignment will be used in the next phase of the project.

Grading Scheme:

- Initial data model (drawn using tools) with description: 10% (initial sketch – conceptual model based on business requirement)
- Data Extraction and Data collection: 30% (writing program/script to gather useful data)
- Data insert: 10% (Actual table creation, insert/upload data from XML file)
  [Disclaimer (available on page 3) must be included in the report]
- Final Data Model Design: 20% (Actual Data model after addressing design issues if any – Include report).
- Normalization up to 3NF (if possible): 15% (Present SQL scripts, and ERDs before and after normalization)
- Answer the given question: 10% (Present SQL scripts)
- Adding citation in IEEE/ACM Format only. Use reliable information source: 5%

Academic Integrity:

- This assignment does not require group work. Therefore, each student is expected to complete their work by themselves. Collaboration of any type amounts to a violation of the academic integrity policy and will be reported to the AIO.
- Do not copy texts verbatim from online or printed materials
- Do not copy texts from other's work
- Do not submit other's work
- If you obtain help from Tutor(s), please acknowledge
- Provide citation for texts, images, tables, data etc.
- 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

Hypothetical Scenario:

A Company, *"Analytics-5408"* is trying to establish its position in the business domain in Canada. Recently it hired you as an information specialist for one of its clients (Dalhousie University). Your company wants to create an app to provide people higher education related information. This is not a NEWS app; therefore, as the initial requirement it does not provide current NEWS items. The app will analyze data and guide people. The project has two components,

(1) Data management, and
(2) Business Intelligence

As the Information specialist, you should gather all required information during the first phase. A rough design must be produced from the initial conversation before finalizing the data model. Once the requirements are gathered, you will collect data from different sources, or sources provided by your company. This data must be cleaned, formatted, decomposed etc. before uploading to actual database. A final report is due at end of each phase.

In addition, your goal is to ensure that the designed app receives a good customer rating; Therefore, you will perform an initial study (market research) of Google Apps' ratings/ reviews to identify user behavior. In this phase of the project, you need to identify all possible domains (e.g. Faculties, Maintenance, Security, Management, Benefits, Health etc.), entity sets, attributes, and relationships. Once it is established, you need to implement and populate the database using relational database management system, and run some queries to answer the given questions (section F)

Dataset:
Google App Rating: https://www.kaggle.com/lava18/google-play-store-apps
Education: visit https://www.dal.ca/ and all the important web pages within the website domain

***** Your Tasks for this Assignment *****

A. Data Modelling (Initial Design or the Conceptual Model)

You need to create two types of data models. One for the Google App market research, and other for the university related information system.

1. **Data Model from Existing Data:** Study the Google App ratings data that are available on Kaggle.com. Identify suitable entities, attributes, and their relationships. For the initial design, you can use a model like Chen's model.

2. **Data Model from Business Problem:** The objective of the project is to design an information system for people, who want to join Dalhousie University as a student, faculty member, or staff. Further, the system will help analysts to perform comparative analysis of university academic programs, courses, number of faculty members, benefits, etc.

   You can build this initial data model after visiting the DAL website. This section does not require any data extraction. However, you need to envision the overall information system to identify the entities, attributes, and relationships.

B. Data Extraction and Data Collection

1. Collect Google App ratings data from Kaggle.com, and store it in an RDBMS of your choice. This step does not require any programming.

2. Write a program /script to extract data from DAL website, and store it in an XML format. The XML file must be created programmatically.
   Note: Do not extract personal information, such as employees' email, contact number etc. You must add the following Disclaimer in your report.

Disclaimer (Data Scraping)

Name: <Your Name>
#ID: <B00xxxxxx>

In assignment 1 of CSCI 5408 course, data scraping is done manually or programmatically from Dalhousie University's website, and it is used only for educational purpose. Sensitive information, such as personal Email, personal contact numbers are not extracted. However, names of instructors, professors, or other staff members available on the Dalhousie University websites are extracted for course (CSCI 5408) related analysis, such as "find how many employees have similar first name etc." The scope of the extracted data usage is limited to the course CSCI 5408 only. The course instructor and the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data.

C. Data Insertion:

Create database and tables using the extended ERD you created, and populate the tables with the data you obtained from the selected datasets. You can use MySQL or MSSQL or Oracle DBMS systems to create your database. If you do not want to install the DBMS on your system, you are free to use cloud based database applications.

D. Final Data Modelling:

1. Is your initial sketch/design free from any design issues? (Yes/No)
   Provide justification to support your answer.

Final Design

1. Identify the relationships, and cardinality between the entity sets you created.
2. Construct an extended ERD. Your ERD should highlight if any overlap/disjoint subtype exists.
3. The extended ERD should be created using a tool, such as MySQL workbench/ ErWin/ Visio etc.

E. Normalization:

Once you create the tables, identify the functional dependencies, and normalize the tables up to 3 NF (no transitive dependency). You should create 2 copies of your database – one copy should contain table(s) before normalization process, and other should contain tables after normalization. In addition, you need to generate ERDs before and after the normalization.

F. Write SQL Query:

1. Find the name of the department or faculty that has the highest number of employees having last name starting with an "A"
2. Find the name of the department or faculty that has the highest number of undergraduate programs

Submission Instruction:

- Create a Folder with your name and B00 number, and store following files in the folder–
  - PDF file with answers, and disclaimer
  - Program source code, and sample run screen capture.
  - SQL script file(s) for schema, (before and after normalization)
  - SQL script file(s) for data,
  - images (ERDs, data models etc.) in the folder.
- Compress the folder and create a .ZIP file (do not use other compression formats)
- Upload the .ZIP file on Brightspace.
- Submission Due: Sep 29, 2019 at 11:59 pm (midnight)