Assignment 1 (10%)
Date Given: Jan 27, 2020
Submission Due: Feb 10, 2020 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 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 assignments.

Grading Scheme:

- Identification of datasets and Data collection: 10%
- Answer the feasibility analysis questions: 20%
- Initial Data Model Design: 25% (Conceptual Model, 1st paper sketch, 2nd using ERD tools like draw.io
- Design Issues identification and solution: 10% (3rd and Final Data Model)
- Database implementation based on the designed data modelling: 10%
- Normalization: 10%
- DML for the questions given in section C: 10%
- 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:

HalifaxInfo is a startup in Halifax, which is planning to build a data management portal for the Halifax region. The system can be conceptualized as a content management system. The project has three components,
(1) Data management,
(2) Visualization-Analytics, and
(3) Front-end design.

HalifaxInfo is trying to identify key performance indicators (KPIs) in the Halifax region to improve the *business*, *education*, *lifestyle*, and *safety*. In the first phase of the project, the company is trying to gather relevant structured data and information that are collected by various sources. Once the relevant datasets are identified and data are collected, the company will find key entity sets, their attributes, and relationships that will be critical to the future system. This is an important stage for the project development.

*** Your Tasks for this Assignment ***

CSCI-5408 assignment series covers the (1) Data management, and (2) Visualization-Analytics component. The first two assignments will focus on the data management component, and the next two assignments will focus on the visualization-analytics component.

As an *information specialist,* you are expected to perform a series of tasks that are mentioned in section A to section D. To start building the project, *HalifaxInfo* has identified some URLs, which provide various datasets.
http://catalogue-hrm.opendata.arcgis.com/
https://dal.ca.libguides.com/data/novascotia
https://data.novascotia.ca/

Specific Tasks (section A to D) – complete them as specified and submit on Brightspace based on the submission instructions on page #3

A. Feasibility Analysis: (Answer these questions)

1. Do you have enough datasets to identify the key performance indicators to improve *business*, *education*, *lifestyle*, and *safety* of Halifax region?

2. List all the datasets, you find critical for the project. Note: You need to provide URLs of the actual datasets, and write reason of selection in single sentence.

> **Answer:**
> e.g. Name: Business Establishments 2010-2011 dataset
> URL: https://data.novascotia.ca/Business-and-Industry/Business-Establishments-2010-2011/wa8g-ji9a
> Reason: This dataset will help to understand types, and sizes of industries established in the region during a specific period, which can provide information on business opportunities.

3. What are the programming languages and scripts or tools you used to extract or collect the required datasets?

4. From your selected datasets, identify entity sets. You can create your own entity set by omitting certain fields/attributes from a dataset, or by merging more than one datasets. Consider fields/attributes that you find crucial to identify KPIs.
   E.g. I identify dataset <name of dataset D1>, <name of dataset D2>, and <name of dataset D3> useful.
   Now, after exploring the data in D1, I found D1 has 5 fields, but I just need 2 fields.
   I consider D1 with 2 fields as my entity set that has 2 attributes.
   After exploring D2, I found 10 fields, and all 10 fields are important. I create one entity set from the data set. I found dataset D3 and dataset D2 has many common elements, I merged these two datasets, and finally extracted the useful attributes

5. List the strong/weak entity sets in your system that are important to identify the KPIs.

B. Data Modelling:

Initial Design

1. Identify the relationships, and cardinality between the entity sets you created.
2. Make the initial sketch on paper and take picture/scan (This is 1st and rough design)

3. Construct an extended ERD (2nd design). Your ERD should highlight if any overlap/disjoint subtype exists.
4. The extended ERD should be created using a tool, such as ErWin/ Visio/ draw.io etc.

Final Design

5. Is your initial design free from any design issues?
    a. If No, solve the design issue and construct a new extended ERD (3rd design).
    b. Submit all diagrams and your findings.

C. DDL (Data definition language) & DML (data manipulation language):

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.

❖ **Test if your newly created database provides answers to the following questions**

- Which business organization or type of business organization has highest employees?
- Which area in Halifax region has more schools?
- Which street in Halifax region has more number of reported crimes?

D. Normalization:

Once you create the tables, identify the functional dependencies, and normalize the tables up to 3 NF (no partial and 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.

e.g. Database1 contains one table, which is in 1 NF

PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CLASS	CHG_HOUR	HOURS
15	Evergreen	103	June E. Arbough	Elect. Engineer	84.50	23.8
15	Evergreen	101	John G. News	Database Designer	105.00	19.4
15	Evergreen	105	Alice K. Johnson *	Database Designer	105.00	35.7
15	Evergreen	106	William Smithfield	Programmer	35.75	12.6
15	Evergreen	102	David H. Senior	Systems Analyst	96.75	23.8

Database2 contains tables after converting to 2 NF

PROJ_NUM	PROJ_NAME
15	Evergreen
15	Evergreen
15	Evergreen
15	Evergreen

EMP_NUM	EMP_NAME	JOB_CLASS	CHG_HOUR
103	June E. Arbough	Elect. Engineer	84.50
101	John G. News	Database Designer	105.00
105	Alice K. Johnson *	Database Designer	105.00
106	William Smithfield	Programmer	35.75
102	David H. Senior	Systems Analyst	96.75
114	Annelise Jones	Applications Designer	48.10

PROJ_NUM	EMP_NUM	HOURS
15	103	23.8
15	101	19.4
15	105	35.7
15	106	12.6
15	102	23.8

Submission Instruction:

- Create a Folder with your name and B00 number, and store all your files – PDF file with answers, SQL script file, images (if any) 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: Feb 10, 2020 at 11:59 pm (midnight)