Assignment #1

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

Date Given: Sep 20, 2021
Due Date: Oct 3, 2021 at 11:59 pm

Late Submissions are not accepted. 10% deduction per day will be applied for late submissions.

Disclaimer: This assignment requires students to work on various research and open Datasets with appropriate citation. Submissions related to this assignment will not be considered for commercial purposes.

Objective:

- The objective of this assignment is to understand industry problems related to data capture, and database design. Create entity relationship model and perform normalization of the database.

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%)	Problem # where applied
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 #1
Correctness	All parts of the given tasks are correct	Most of the given tasks are correct However, some portions need minor modifications	Most of the given tasks are incorrect. The submission requires major modifications.	Incorrect and unacceptable	Problem #2
Novelty	The submission contains novel	The submission lacks novel contributions.	The submission does not contain	There is no novelty	Problem #2

	contribution in key segments, which is a clear indication of application knowledge	There are some evidences of novelty, however, it is not significant	novel contributions. However, there is an evidence of some effort		
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.

Hypothetical Scenario

- An established organization "HalifaxData5408" operates in Canada, and they have few clients overseas.
- Recently, "HalifaxData5408" signed a contract with Dalhousie University for processing, enhancing, and storing their data.
- You have joined "HalifaxData5408" as Information Specialist, and you will be in-charge of this entire operation, which includes two problems and sub tasks. Since you are reporting to the Manager, you need to document the entire operation and provide justification for the choices you make or decision you take.

Problem #1: Building a Data Model for Dalhousie University

Visit the website https://www.dal.ca/ and any pages within the website that you find appropriate to gather information on Dalhousie University. The university is trying to build an information system to capture all the key information related to the departments, infrastructure, services etc. Your initial task is to identify the key entities, attributes, and the relationships, so that at next phase of the task, the University can decide on how to create the database.

Therefore, at this stage of the project, the University is expecting you to provide a correct and flexible data modelling (EERD), which is free from any of the design flaws (e.g., absence of capturing historical data, chasm trap, and fan-trap etc.)

Conditions/Steps You must Follow (Do not skip any point):

1. This process does not require any web scrapping, therefore, do not perform such operations.

2. You must add the declaration in your submission:

"I ……., declare that in assignment 1 of CSCI 5408 course, data scrapping is not done programmatically or using any online or offline tools. However, the webpages within dal.ca domain are visited manually, and some useful information is gathered for education purpose only. Information, such as email, personal contact numbers, or names of people are not extracted. The course instructor or the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data"

3. You need to visit the webpages within dal.ca and document your findings in a systemic manner.
    - E.g., after visiting the website you find "***Libraries***" an entity, then in a single sentence in the PDF file, you need to mention, why did you consider "***Libraries***" as a valid entity. You should provide a tabular structure as mentioned in the 5th point.

4. Identify at least 18 valid entities, and that should not include sub-types. Therefore, with sub-types, it could be more than 18.
    - A valid entity means a proper strong or weak entity, which may have one or more attributes. E.g., "**CampusLife**" is not a valid entity, it can be a website menu item, but not an entity. However, "***CampusNews"*** can be a valid entity with attribute such as **ID**, **Headline**, **Date**, **Content**.

5. Create a table similar to the one given here with at least 18 valid entities and provide the reason of your selection.
   E.g.

Entity	Reasons for considering	Source
CampusNews	This entity represents news item or instances in the dal information system. It is a strong entity, because it exists without depending on other entities, and it has unique ID to identify each news item. This is a valid entity and capturing News will provide historical information about the system in future.	Information related to this entity is found in https://www.dal.ca/news.html

6. Create an initial data modelling (Chen model) with entities you identified with the possible attributes and try to establish the relationships between the entities. You should also add cardinality at this stage. **Perform this operation on a paper/ powerpoint/ word/paint etc**. At this stage you may get plenty of errors, design issues, and absence of attributes, or incorrect cardinalities, which are acceptable. This step will highlight your understanding of the problem, and the domain.
   You should not use Workbench or reverse engineering for Problem #1.

7. In the next step, you need to perform a systematic approach to find solution for the design issues, or attributes that were not considered, or entities that you discovered new, and document it with possible solution. You need to write (within ½ page) the problems that you found in your paper/initial design (6th point) and write your planning on how you are going to solve it.

8. Once you find the solution, it is the time to build the final correct data modelling (EERD) using a tool like ErWin/ Visio/ draw.io etc.

Submission Expectations:

(1) Report in PDF,
(2) image of Initial ERD/EERD, and
(3) image of final ERD/EERD.

Problem #2: Format Ocean Tracking Data and Report

Dalhousie Ocean Research wants you to explore the dataset they provided, and perform the following:

1. Read the document available at
   http://oceantrackingnetwork.org/about/#oceanmonitoring

2. Write a ½ page report (in your own words) on the different datasets, and attributes you discovered.

3. Clean and transform the dataset using spreadsheet formula/filtering. You do not need to write any code or use any other tools.
   a. remove NULL values.
   b. rearrange the columns if needed.
   c. transform the data in a column or attribute if required to fit a common format.
   d. Is there a possibility of combining some of the tables or attributes without losing information (de-normalization)? If yes, please perform the task and report your findings.
   e. Is there a possibility of decomposing some of the tables without losing information (normalization)? If yes, please perform the task and report your findings.

4. Based on the given dataset, create relational schema using MySQL DBMS

5. Using MySQL Workbench and reverse engineering create the possible ERD. Your report must contain the ERD produced by the reverse engineering. In addition, you need to add the cardinality.

6. Populate the database with clean and transformed dataset (if dataset is huge, then import at least 1000 random data points or rows).

Submission Expectations:

(1) Report in PDF file,
(2) ERD generated using MySQL Workbench,
(3) Normalization/Denormalization (Logic and reason in the PDF file),
(4) SQL Dump of Table structure and values (Before normalization or de-normalization)
(5) SQL Dump of Table structure and values (after normalization or de-normalization).