Assignment #1

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

Date Given: May 20, 2022
Due Date: Jun 3, 2022 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 various websites and open Datasets with appropriate citation. Submissions related to this assignment will not be used 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 for 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%)	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 #1 Problem #2 Problem #3
Correctness	All parts of the given tasks are correct	Most of the given tasks are correct However, some	Most of the given tasks are incorrect. The	Incorrect and unacceptable	Problem #1 Problem #2 Problem #3

		portions need minor modifications	submission 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 Problem #2 Problem #3
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 Problem #2 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.

Explanation of the Rubric:

Suppose you received different grades in ***Clarity*** for the 3 given problems.
Problem #1: 25% in Clarity,
Problem #2: 15% in Clarity,
Problem #3: 15% in Clarity
Then your overall grade for Clarity will be avg. of {25+15+15} % = approx. 20%

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 or the domain mentioned in this document 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"

Problem #1: This item will be graded based on aforesaid rubric.

Hypothetical Scenario

Establishing relationships and identifying key entities, attributes are the main requirements of this project. The problem is not well defined; however, the client trusts your expertise, and believe that you will provide an enhanced data modelling for the conceptual phase, and sample table structure for the physical design. Apart from the design, and the table structure, the client wants your assumptions in the form of one page written report.

Note:

To obtain information on various collected data, you can visit: https://resources-covid19canada.hub.arcgis.com/search

Problem #1 Submission Expectations:

(1) Report, ERD/EERD, Normalization/Denormalization, SQL Dump of Table structure and values.

Problem #2: This item will be graded based on aforesaid rubric.

Hypothetical Scenario

Visit the website https://parks.novascotia.ca/ and any other related websites that you find appropriate to gather information on Nova Scotia Parks. The province is trying to build an information system to capture all the key information related to the parks that are operating in the province. Your initial task is to identify the key entities and the relationships, so that at next phase of the project Nova Scotia can decide on how to create the database.

Therefore, at this stage of the project, the province is expecting you to provide a correct and flexible data modelling, 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 need to visit the website(s) and document your findings in a systemic manner.
   - E.g., after visiting the website you find "Parks" an entity, then in a single sentence in the PDF file mention, why did you consider "Parks" as a valid entity. You should provide a tabular structure as mentioned in the 5th point

3. Identify at least 12 valid entities, and that does not include sub-types if you are considering an EERD.

4. A valid entity means a proper strong or weak entity, which may have one or more attributes. E.g., "ParkName" is not a valid entity, it can be an attribute of entity "Parks".

5. Create a table of entities and provide the reason of your selection.

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.

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 (6th point) design 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 (ERD or EERD) using a tool like ErWin/ Visio/ draw.io etc.
   - ○ If you include EERD, then highlight the part in your ERD (e.g., drawing a circle around the entity sets) that you want to extend.

Problem #2 Submission Requirements:

(1) A single Problem #2 pdf with list of entities, reasons, sources, attributes
(2) Image of EERD_Initial_P2,
(3) Image of EERD_Final_P2

Problem #3: This item will be graded based on aforesaid rubric.

Format, Clean, Store ***Ocean Tracking Data*** and Report your findings

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

- Read the document available at http://oceantrackingnetwork.org/about/#oceanmonitoring
- Write a report on what are the different datasets, and attributes you discovered.
- Clean and transform the dataset using combination of manual work/ spreadsheet filtration/ code written in Java. Include your cleaning/transformation steps in the ***Problem #3 pdf file***. (**Note**: If you write any programming script, it must be added to gitlab as part of the submission)

  Transformation and Cleaning requirements.
  - remove NULL values
  - rearrange the columns, if columns are shifted and not matching the flow.
  - transform the data in a column or attribute if required to fit a common format (e.g. date format)

  - In the clean spreadsheets/CSVs you created, is there a possibility of combining some of the files, or columns in the files (without losing information)? If yes, please perform the task and add your findings in the ***Problem #3 pdf file***.

    OR
  - In the clean spreadsheets/CSVs you created, is there a possibility of further decomposing of the files, or columns in the files (without losing information)? If yes, please perform the task and report your findings in ***Problem #3 pdf file***.

- Based on your final CSVs or spreadsheet files, create relational schema using MySQL DBMS.
- Populate the database with your transformed dataset. If the dataset is large you can consider uploading a random subset of maximum 3000 data points on the database.
- 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.

Problem #3 Submission Requirements:

(1) Any findings, logic etc. must be included in the Problem #3 pdf file.
(2) SQL Dump of Table structure and values must be submitted.
(3 - optional) If you write a script/program to clean/format data, then upload your program code to gitlab (https://git.cs.dal.ca).
(4) You can export the ERD from workbench and include it as an image file

Assignment Submission Instructions:

All files must be added to a single .zip file before uploading to Brightspace.
Do not use any other compression format.
Rename the .zip file as **Your_FirstNameB00xxxxx.zip**