Assignment #1

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

Date Given: Jan 17, 2022
Due Date: Jan 30, 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

The Halifax city universities are trying to merge their databases under one city university initiative. In this regard, the council has decided to revisit the data and information that are already available, and rebuild a data model. Once the data model is obtained in the conceptual model, physical design of the database will be performed.

You are hired as a Strategist to create a conceptual model for this scenario. The council will not accept a poorly designed data model or data model with design flaws. Therefore, you are expected to follow some steps to create the conceptual model. This model is independent of a specific database or hardware, and therefore, while creating the model you do not have to consider any database or tables etc. This is a higher level conceptual model, which must be designed using tools, such as Erwin, Visio, Dia, Draw.io etc. or similar tools.

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

1. Install/Access or explore appropriate software tool (ErWin/ Visio/ Draw.io etc.) for ERD/EERD creation
2. Explore these three university websites (including sub-pages/domains) given below and write down names of possible entity sets required for the city university initiative project.
   (1) https://www.dal.ca/
   (2) https://www.msvu.ca/
   (3) https://www.smu.ca/

3. You must identify at least 25 entity sets, and for each entity write a reason (in single sentence/ one line) for your selection in the tabular format given below:

Table: Entity Identification

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

4. Once you got all the entities/ entity sets of your choice, now start building the ERD. You can start creating your ERD (Chen Model) using the software tool. At this point try to focus on the relationships, and cardinality. – {Label it as **ERD_Initial_P1**}
5. Identify and highlight if your designed ERD has any design issues. 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 (at least ½ page) on what problems you found and how are you going to solve in your final ERD.
6. Modify your current ERD and solve the design issues – {Label it as **ERD_Final_P1**}

Problem #1 Submission Requirements:
(1) A single Problem #1 PDF with list of entities,
(2) Image of ERD_Initial_P1,
(3) Image of ERD_Final_P1

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

Hypothetical Scenario

You are hired as a Strategist to create a conceptual model for a company working on WHO (World Health Organization). You are expected to create the conceptual model without flaws, and with a possibility of future enhancements. This model is independent of a specific database or hardware, and therefore, while creating the model you do not have to consider any database or tables or data points etc. This is a higher level conceptual model, which must be designed using tools, such as Erwin, Visio, Dia, Draw.io etc. or similar tools.

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

1. To create an Extended ERD (EERD), use the tool you already installed or accessed in the previous problem. If your EERD is large, you can always break it in multiple parts (smaller EERDs or ERDs).
2. You need to visit https://www.who.int/ and check the different sub-pages or sub-domains that are publicly available.
3. Without writing any script or program, just by performing some manual observations, you need to identify at least **20 entities** from the pages you have visited.
4. For writing the entities, create a same tabular format that you have done for the previous problem. In addition, add possible attributes (separated by commas) for the Entities in an additional column.

Entity	Reasons for considering	source	possible attributes

5. Once you got all the entities/ entity sets of your choice, now start building the EERD. You can start creating your EERD (Chen Model) using the software tool. At this point try to focus on the relationships, and cardinality. – {Label it as **EERD_Initial_P2**}
6. Identify and highlight if your designed EERD has any design issues. 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 (at least ½ page) on what problems you found and how are you going to solve in your final EERD.
7. Modify your current EERD and solve the design issues – {Label it as **EERD_Final_P2**}

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**