# Assignment #1

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

**Date Given: May 20, 2021**
**Due Date: Jun 1, 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 #3 |

| | | | | | |
|---|---|---|---|---|---|
| | 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 the province "Nova Scotia", and other organizations 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 three problems or 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 Nova Scotia on its Provincial Parks

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.

**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).

**Problem #3:** Opportunities in Halifax (*Flexible project with abstract problem, and dataset*)

A real-estate client wants to explore various factors that can create opportunities for future business investments in Halifax. The client wants to know if Halifax is good for Education, and if it has good shopping areas. They asked you to capture data from various public domains and create an enhanced model. In addition, they want a database populated with few hundred records just to analyze the data pattern. They are expecting an EERD that shows the possible relationships between various factors, such as Education, and Shopping/Lifestyle.
To obtain information on various collected data, you can visit: https://data.novascotia.ca

**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 given and document your findings in a systemic manner.

3. Identify at least 5 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., "**SchoolName**" is not a valid entity, it can be an attribute of entity "***School***".

5. 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**. 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

do not need to solve design issues for this problem. However, it will be appreciated if you can solve the design issues)

6. After downloading the required dataset, perform possible normalization/cleaning/de-normalization using spreadsheet formula/filtering and include the process in the PDF file with images.

7. Import the clean dataset on MySQL Workbench and create the ERD using reverse engineering.

**Submission Expectations:**

(1) Report in PDF,
(2) Image of ERD/EERD you created on paper.
(3) Image of ERD obtained from MySQL Workbench.
(4) SQL Dump of Table structure and values.
(5) ** ERD if you solve the design issues.

** Not a mandatory requirement.