University of Ottawa

School of Electrical Engineering and Computer Science

CSI4142 Fundamentals of Data Science

Project Phase 2: Physical Design and Data Staging

Due Date: March 22, 2024, 11:59pm

Instructions:

A. This is a team assignment.

B. Follow the data staging steps (ETL) discussed during the practical class to populate the data mart, as follows:
   - Extraction: data can be extracted in various formats, such as CSV, XML, or JSON.
   - Transformation: this can include data cleaning (handling missing values, typos, and outliers, removing duplicates, converting data types etc.), transforming the data into a format that can be used for analysis (i.e., normalizing or scaling the data), data integration, data discretization (i.e., converting continuous data into discrete data by grouping it into bins or categories), Feature engineering (i.e., creating new features from existing data that may be more relevant or useful for analysis). This step may also involve aggregating or summarizing data.
   - Loading: generating the surrogate keys and loading your integrated/final dataset.

C. Create your data mart using a Database Management System (DBMS) of your choice.

Deliverables:

1. Submit your source code (well documented code) in a zipped folder on the Brightspace, or link to a GitHub repository (GitHub Preferred to get started using it if still not).

2. Submit a PDF file with the following details.

   A. A one-page schematic with your high-level data staging plan.

   B. Any other details you want to add.

   C. A list of data quality issues you encountered and how you handled them (i.e., how did you detect and handle missing or noisy data (if any). How did you integrate the data from different sources etc.?

D. Fill out the attached excel sheet (Team Planning) and include it in the PDF.