Group Project Task 2 - Uploading Dataset to GCP and Processing the data through ETL

Due Date: Friday November 21, 2025, 11:59 p.m.

- Upload the selected data sets chosen from your Task-1 of the group project to GCP GS Bucket.
- Connect to GCP either via GCP Console or GCP CLI. Seek help from friends or internet to get help if you are stuck.
- Connect to GCP from your local machine either
    1. using the terminal, or
    2. using GCP console.
- Create a GS bucket “msds-694-cohort-14-<project_group_Number>
    1. E.g. if I am in group 6, my group’s GS bucket will be msds-694-cohort-14-group6
    2. Create a folder called “data”.
    3. Upload the data to this folder using either GCP console or terminal cli commands.
- Start understanding the dataset you have chosen
    1. Based on your understanding of the dataset you have chosen, start playing with a subset of that data on your local Jupyter notebook using PySpark code.
    2. Take first 1000-5000 lines from the original dataset and save it a separate file, creating a smaller subset of the data.
    3. Write a spark code that will create a meaningful dataset from this subset or smaller dataset.
    4. Later, we will take this code and run it via spark-submit command.

Deliverables:

- Take a screenshot of the data uploaded to GS and upload it to canvas.
- Upload the current working PySpark code or the Jupyter notebook indicating the job that creates the summarization or aggregation job. It does not have to be complete work but a work in progress.
- E.g. If I have Orders data from Amazon, I can create a summarization job that creates monthly sales report or number of orders placed by the customers.
- E.g. HomeDepot wants to find its unique customer base based on past visit or order history.