CS 329E Project 8, due Thursday, 04/04.

Change Data Capture (CDC) is a process by which changes to a database table are captured in the form of individual insert, update, and delete operations. In this project, we implement a CDC pipeline for one target table in the consumption layer.

Objectives

- Simulate a stream of changes to the CSV file(s) which source your target table of choice
- Ingest each affected CSV file into its own table into a loading area
- Detect the deltas between the loading and raw tables
- Refresh the raw table(s) with the changed records such that each raw table represents the latest snapshot of data
- Re-generate the tables in staging which are affected by the updated raw table(s)
- Merge the changes from the final staging table into the existing target table, while preserving the history of the changed records in the target table

Implementation Guidelines

- Implement your CDC pipeline in a Colab notebook named **cdc.ipynb**
- Make manual changes to your CSV file(s): add a few new rows, update a few existing rows, and delete a few rows from those file(s)
- Upload the changed CSV files into a new folder in your existing bucket. Name this folder named **incrementals**.
- Create a new dataset in BQ for the loading area and ingest each new CSV file into this dataset. Name this dataset **[domain]_ldg**.
- When applying changes to the raw table(s), ignore any records that are unchanged between loading and raw (i.e. don't update the load_time of any records in raw which are unchanged).
- Apply the normal logic in staging to each affected table. You don't need to re-create all the staging tables, only those that are affected by the changes in raw.
- Merge the final staging table into target, setting the **discontinue_time** and **status_flag** of each inactive record. The discontinue timestamp should be equal to the effective timestamp minus 1 second. The way that we are modeling the changes in the target table is known as a Slowly Changing Dimension of Type 2.
- Use the provided code samples, **p8-cdc.ipynb,** as a starting point for your work
- Publish to your repo: **cdc.ipynb**.

CS 329E Project 8 Rubric
Due Date: 04/04/24

Necessary tables exist in ldg dataset in BigQuery **-5** incorrect load time **-10** table not in ldg dataset **-15** table missing entirely (not in any dataset)	15
Correctly identifies deltas between tables **-5** does not account for nulls **-10** does not use full join **-15** missing output	15
Correctly creates changes within the table inside raw dataset **-5** for each operation not verified (select statement showing proof) **-10** missing insert operation **-10** missing update operation **-10** missing delete operation	30
Staging and target tables are correctly created and merged **-5** load time does not match with current time **-10** does not create staging table **-10** does not create target table **-15** merged table missing discontinue_time and status_flag **-15** does not verify merged table creation **-25** lack of populating newly merged table with create, insert, update	40

`submission.json` submitted into Canvas. Your project **will not** be graded without this submission. The file should have the following schema:
 	
`{`
`    "commit-id": "your most recent commit ID from Github",`
`    "project-id": "your project ID from GCP"`
`}`
 	
Example:
 	
`{`
`    "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",`
`    "project-id": "some-project-id"`
`}`
	Required

Total Credit:	**100**