CS 327E Milestone 4 due Thursday, 12/10 at **12pm**.

Part 1:

Convert your previously written Beam pipeline(s) to Dataflow. Run them on Dataflow over the entire input data; debug and fix as necessary.

General Coding Conventions:

- Create a new notebook `milestone4.ipynb` and call the pipelines from the notebook.
- The code should be commented sufficiently to follow the main logic of the transforms.

Dataflow Coding Conventions:

- A Beam pipeline should transform a single source table.
- All transforms applied to a source table should be placed in the same Beam pipeline.
- A pipeline script should be named `<table>_dataflow.py`.
- A table should be named `<table>_Dataflow` when produced by the Dataflow Runner.

Part 2:

Verify that each BigQuery output table (e.g. `<table>_Dataflow)` contains a valid primary key. Child tables must also have a valid foreign key. Run the appropriate SQL statements within your `milestone4` notebook to verify these constraints.

Update your ERD to reflect the schema of your transformed tables:

- Diagram should capture only the latest version of each table (e.g. `<table>_Dataflow).`
- Entity types should specify field names, data types, and keys for each table.
- Diagram should visually indicate the source of each entity (e.g. entities from `dataset1` can use one background color while entities from `dataset2` can use a different background color).
- Draw the relationships between the entities within `dataset2` as well as **across** the two datasets.
- Name your ERD file `final_unified_model.pdf.`

Part 3:

1. Implement your cross-dataset queries:
   - Develop and run three cross-dataset queries from your `milestone4` notebook
   - Queries should use the modeled tables from both refined datasets

- Each query should be wrapped into a view, created in your `insights` dataset
- A short comment should appear above each SQL statement to describe its function

2. Create visualizations in Data Studio:
   - Create a data visualization from each cross-dataset query
   - Data Sources query the SQL views from the previous section.
   - Charts should visualize the data in a compelling way.
   - Add the 3 charts to your existing Data Studio report (aka dashboard).
   - Download the report and save it as `final_dashboard.pdf`.

CS 327E Milestone 4 Rubric
Due Date: 12/10/20

Part 1 - Convert your Beam pipelines to Dataflow. Each Beam pipeline should have two Python scripts, `<table>_beam.py` and `<table>_dataflow.py` per source table. **-X** for each missing `<table>_dataflow.py` where X is dependent on the number of Beam pipelines. If you have 2, -**20** each. 3, -**13.3** each, and so on. **-10** Beam pipelines not using DataflowRunner **-10** Beam pipelines do not execute properly **-10** Beam pipelines not writing to output table `<table>_Dataflow` **-10** Beam pipeline run calls missing from `milestone4.ipynb` *(points will be broken based on number of pipelines)*	40
Part 2 - Verify primary key constraints on tables transformed by Beam. Verify foreign key constraints if those tables are also child tables. Add this logic to your notebook. **-10** missing or incorrect primary key verification on final output tables **-10** missing or incorrect foreign key verification on final child output tables Create an updated ERD that finalizes your table schema after Beam transforms have been applied. -**10** `./final_unified_model.pdf` not found in repository **-5** ERD is missing one or more entities **-5** ERD is missing one or more primary keys **-5** ERD is missing one or more foreign keys **-5** ERD is missing or incorrect relationship between entities	20
Part 3 - Implement and run your three cross-dataset queries. Comment each query with the function it performs. **-5** each missing or erroneous query, up to **-15** **-5** each missing or incorrect comment, up to **-15** **-5** each query not on a transformed table, up to **-15** Create 3 data visualizations and add them to your existing Data Studio report. The visualizations should represent the results from the three BQ views. The Data Studio report should contain a total of **5 charts**, 2 from Milestone 2 and 3 from the current milestone. Each chart should have a relevant title describing the dataset. **-20** `./final_dashboard.pdf` not found in repository **-10** each missing chart, up to **-20** -**10** each chart created from a BQ table instead of a BQ view, up to **-20** **-5** each missing title, up to **-15**	40
`submission.json` submitted into Canvas. Your project **will not** be graded without this	**Required**

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"`
`}`	

Total Credit:	**100**