CS 327E Milestone 4 due **Thursday, 05/13**.
Hard deadline: no late submissions will be accepted.

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 the latest version of all tables in your datamart (e.g. `<table>_Dataflow).`
- Entity types should specify field names, data types, and keys for each table.
- Diagram should include all valid relationships between the entities.
- Name your new ERD `final_project_datamart.pdf.`

Part 3:

1. Implement three cross-dataset queries in your `milestone4` notebook:
   - Develop and run three queries that join across both sources of data (primary and secondary datasets)
   - Queries should use the datamart tables (not the staging tables)
   - Wrap the queries into views and create the views in your `reports` dataset
   - Add a short Markdown comment above each SQL statement to describe its function

2. Create visualizations in Data Studio:
    - Create a data visualization with the results from each cross-dataset query
    - Data Sources in Data Studio should query the views (not the tables directly).
    - Charts should visualize the data in a compelling way.
    - Charts should have a relevant title that describes the data.
    - Add the three charts to your existing Data Studio report (aka dashboard).
    - Download the report and save it as `final_project_dashboard.pdf`.

CS 327E Milestone 4 Rubric
Due Date: 05/13/21

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 schemas after Beam transforms have been applied. -**10** `./final_project_datamart.pdf` not found in repository **-4** ERD is missing one or more entity types **-2** ERD is missing one or more primary keys **-2** ERD is missing one or more foreign keys **-1** 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 which doesn't join the two sources of data, 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 data. **-20** `./final_project_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**