Milestone 10 due Friday, 04/19.

Part 1:

Develop the Beam pipelines in support of your cross-dataset queries. The pipelines should read the records from your secondary dataset tables, perform one or more Beam transforms on the input data, and write the cleansed and normalized records back to BigQuery as a new table.

Develop and test the pipelines on a small subset of the data using the `DirectRunner`. Once tested, convert the pipelines to use the `DataflowRunner` and run them over the entire input data.

Coding Conventions:

For consistency and readability, follow these programming conventions:

- Each pipeline transforms a different input table.
- All the transforms related to a given input table are contained in the same Beam pipeline.
- Each Beam pipeline has two flavors, one that runs on a single VM using `DirectRunner` and the other version that runs on a distributed cluster using `DataflowRunner`.
- The file name for a pipeline is `transform_<table>_single.py` or `transform_<table>_cluster.py` where `<table>` is the actual table name being transformed and `single` versus `cluster` indicates the compute environment used by the pipeline.
- Push both flavors of each pipeline to your GitHub repo.

Part 2:

Update your ERD to reflect the newest version of all your tables:

- Diagram should represent the latest tables in `dataset1` and `dataset2`.
- Entities should specify the field names, data types, and keys of each table.
- Diagram should visually identify the dataset that each table belongs to (e.g. use a different background color for each dataset).
- Name your updated ERD file `ERD-v5.pdf`.

CS 327E Milestone 10 Rubric
Due Date: 04/19/19

Part 1 - Create files `transform_<table>_single.py` and `transform_<table>_cluster.py`, to transform tables in your new dataset. The file marked "single" should use the DirectRunner, while "cluster" should use the DataflowRunner. **-60** transform files missing, but transform listed in `TRANSFORMS.txt` **-60** transforms do not use DirectRunner *and* DataflowRunner **-60** transforms do not execute properly, or are unrelated to the table *(points will be broken based on number of transforms)*	60
Part 2 - Create file `ERD-v5.pdf` outlining the schema of your datasets after transforming the tables. The ERD should follow the same guidelines set in previous milestones. **-40** no `ERD-v5.pdf` in repository **-10** each missing table, up to **-40** **-10** each missing key, up to **-40** **-10** each missing data type, up to **-40** **-10** each incorrect relationship, up to **-40**	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**