CS 327E Milestone 5 due Sunday, 03/01.

This is the first of two milestones that makes use of Apache Beam for cleansing your main dataset.

1. Review your latest tables in your `<source>_modeled` dataset. Make a list of the remaining standardization and normalization problems present in your data. For example, duplicate records in table `XYZ` caused by non-conforming dates in column `abc`. Add those issues to the `TRANSFORMS.txt` file.

2. Choose one of the tables you identified in `TRANSFORMS.txt`. Write a short Beam pipeline that cleanses the data in this table. The pipeline should satisfy the following requirements:
   - use the Direct Runner to execute the pipeline
   - run a BigQuery query that contains a `limit` clause over the table(s) in your `<source_modeled>` dataset
   - make an input `PCollection` from the BigQuery results
   - write the input `PCollection` to a local file named `input.txt`
   - apply one or more custom `DoFns` through a ParDo
   - write the output `PCollection` to a local file named `output.txt`
   - write the output `PCollection` to a new BigQuery table in your `<source>_modeled` dataset
   - execute the pipeline from your `<source>_modeled.ipynb` notebook.

3. Verify that the BigQuery result table from the previous step contains a primary key. If it's a child table, it must also have a foreign key. Run the SQL statements to verify these constraints from your `<source>_modeled.ipynb` notebook.

Coding Conventions:

- The Beam pipeline should be in a file named `<table>_beam.py` where `<table>` is the name of the table that is being transformed.
- The BigQuery result table should be named `<table>_Beam` and reside in your `<source_modeled>` dataset.
- The DoFn code should be commented sufficiently to understand the main logic of the transform(s).

CS 327E Milestone 5 Rubric
Due Date: 03/01/20

Create a file `<table>_beam.py` that takes in data from your `<source>_modeled` dataset, performs a DoFn transform on the data, and writes it back out into another table. Sufficiently comment the code to show understanding of the Apache Beam pipeline. In addition, a `TRANSFORMS.txt` file should now be present for all groups. If a transformation could not be found, please refer to the TAs for assistance. **-100** missing `<table>_beam.py` from repository **-50** code does not implement the DoFn transform **-50** code does not pull from or write back to your dataset **-40** code does not write to two output files `input.txt` and `output.txt` (these text files need not be pushed to your repo) **-50** code does not write to output table `<table>_Beam` **-30** code missing comments **-40** missing pipeline run call from `<source>_modeled.ipynb` **-40** missing `TRANSFORMS.txt` **-20** missing or incorrect primary key verification on output table `<table>_Beam` in `<source>_modeled.ipynb` **-20** missing or incorrect foreign key verification on output table `<table>_Beam` in `<source>_modeled.ipynb` if output table is a child table	100

`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**