CS 327E Milestone 6 due Sunday, 03/08.

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

In the previous milestone, you transformed one of the tables you identified in `TRANSFORMS.txt` using a simple `ParDo`. In this milestone, you will expand this work as follows:

- Transform every table listed in `TRANSFORMS.txt`.
- Apply the appropriate Beam transforms to cleanse the data (e.g. `ParDo`, `GroupByKey`, `CoGroupByKey`, `Flatten`).
- Create two versions of each pipeline, one which uses the Direct Runner and processes a small subset of the source data (with the `LIMIT` clause) and another that processes all of the source data using the Dataflow Runner.
- Execute the pipelines from your `<source>_modeled.ipynb` notebook.
- Verify that the BigQuery result tables contain a valid primary key. Child tables must also have a valid foreign key. Run the appropriate SQL statements from your `<source>_modeled.ipynb` notebook to verify these constraints.

Coding Conventions:

- A 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>_beam.py` if it runs the pipeline with the Direct Runner.
- A pipeline script should be named `<table>_beam_dataflow.py` if it runs the pipeline with the Dataflow Runner.
- A table should be named `<table>_Beam` if it was produced by a Direct Runner execution.
- A table should be named `<table>_Beam_DF` if it was produced by a Dataflow Runner execution.
- The code should be commented sufficiently to understand the main logic of the transforms.

CS 327E Milestone 6 Rubric
Due Date: 03/08/20

Create a number of Python scripts, `<table>_beam.py` and `<table>_beam_dataflow.py` based on the transforms specified in `TRANSFORMS.txt`. The above two files should exist for each transform you make. **-X** for each missing `<table>_beam.py/<table>_beam_dataflow.py` where X is dependent on the number of transforms you have. If you have 2, -**50** each. 3, -**33** each, and so on. **-10** pipeline does not work as intended **-10** each pipeline not using both DirectRunner and DataflowRunner **-10** each pipeline not writing to output tables `<table>_Beam` and `<table>_Beam_DF` **-10** missing pipeline run calls from `<source>_modeled.ipynb` **-10** missing or incorrect primary key verification on output tables in `<source>_modeled.ipynb` **-10** missing or incorrect foreign key verification on child output tables in `<source>_modeled.ipynb`	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**