

Milestone 2, due Friday, 09/21.

1. Finalize your choice of datasets. Each dataset should have at least one CSV or JSON file or BigQuery table. **In addition, one dataset should have at least two related files or tables that are from the same source and are connected via a primary key to foreign key relationship.** If you've made any changes to your datasets selection since Milestone 1, update your `DATASETS` file so that it reflects the latest changes.
2. Create a bucket in GCS and a folder for each dataset. Upload the files for each dataset into their folder.
3. Create a dataset in BQ for each dataset.
4. Import the files from GCS into BQ as separate BQ tables:
  - Ensure each table is created in the appropriate dataset
  - Use schema auto detection for quicker import
  - Use consistent naming convention for all tables
5. In the BQ Console, write some simple queries to explore the data:
  - Keep track of the SQL from the most interesting queries.
  - Copy the SQL into a `simple-queries.sql` file.
  - Add a short comment above each SQL statement to describe the query. Comments should begin with a `--` (e.g. `--this is a legal comment in SQL`).
  - You should have at least 1 query per table. The queries should have a where clause and order by clause.

CS 327E Milestone 2 Rubric

**Due Date: 9/21/18**

<p>Import datasets into BigQuery</p> <ul style="list-style-type: none"> <li>-15/10 each missing table, depending on amount of datasets</li> <li>-30 no datasets present</li> <li>-10 inconsistent naming conventions across tables</li> </ul>	30
<p>Include SQL queries that explores the BigQuery data, and explain what the queries do.</p> <p>Each query should be preceded by a comment that explains what it does. All queries should be written and saved in <code>./simple-queries.sql</code> in your group's repository.</p> <ul style="list-style-type: none"> <li>-40 <code>./simple-queries.sql</code> not found in repository</li> <li>-20 no queries use the <code>WHERE</code> clause</li> <li>-20 no queries use the <code>ORDER BY</code> clause</li> <li>-20/15 each un-queried dataset, depending on amount of datasets</li> <li>-10 each missing comment</li> <li>-5 each incorrect comment</li> </ul>	40
<p>Finalize datasets from Milestone 1.</p> <p>Datasets should be described in a file named <code>DATASETS</code> (named exactly like so, no extensions) and each dataset should meet the following criteria:</p> <ul style="list-style-type: none"> <li>- Be available to you as a CSV/JSON file, or a BigQuery dataset</li> <li>- At least one dataset should be made up of multiple related files connected via a primary to foreign key</li> </ul> <ul style="list-style-type: none"> <li>-30 no <code>DATASETS</code> file found</li> <li>-20 all datasets are made up of only one file, or no files connect via a primary to foreign key</li> </ul>	30
<p><code>submission.json</code> submitted into Canvas. Your project <b>will not</b> be graded without this submission. The file should have the following schema:</p> <pre>{   "commit-id": "your most recent commit ID from Github",   "project-id": "your project ID from GCP" }</pre> <p>Example:</p> <pre>{   "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",   "project-id": "some-project-id" }</pre>	Required
<p><b>Total Credit:</b></p>	<b>100</b>