

Milestone 5 due Friday, 10/11.

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

1. Review your latest schema in your `<source_modeled>` dataset. Make a list of all the remaining formatting issues present in this data. For example, duplicate records in table `XYZ` caused by inconsistent naming of field `abc`, non-conforming dates, etc. Add those issues to the file `TRANSFORMS.txt`.

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 contain the following logic:

- run a BigQuery query on your `<source_modeled>` dataset
- make an input `PCollection` from the BigQuery results
- write the input `PCollection` to a local file named `input.txt`
- apply your custom `DoFn` 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

### Coding Conventions:

- The Beam pipeline should be in a file named `pardo_<table>.py` where `<table>` is the name of the table that is being transformed.
- The BigQuery input and output tables should reside in your `<source_modeled>` dataset.
- The `ParDo` code should be commented sufficiently to understand the main logic of the transform.

CS 327E Milestone 5 Rubric

**Due Date: 10/11/19**

<p>Create a file <code>pardo_&lt;table&gt;.py</code> that takes in data from your <code>&lt;source_modeled&gt;</code> dataset, performs a <code>ParDo</code> transform on the data, and writes it back out into another table. Sufficiently comment the code to show understanding of the Apache Beam pipeline.</p> <p>In addition, a <code>TRANSFORMS.txt</code> file should now be present for all groups. If a transformation could not be found, please refer to the TAs for assistance.</p> <ul style="list-style-type: none"> <li>-100 missing <code>pardo_&lt;table&gt;.py</code> from repository</li> <li>-50 code does not implement the <code>ParDo</code> transform</li> <li>-50 code does not pull from or write back to your dataset</li> <li>-40 code does not write to two output files <code>input.txt</code> and <code>output.txt</code> (these text files need not be pushed to your repo)</li> <li>-30 code missing comments</li> <li>-40 missing <code>TRANSFORMS.txt</code></li> </ul>	<p>100</p>
<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>	<p>Required</p>
<p><b>Total Credit:</b></p>	<p><b>100</b></p>