CS 329E Project 5, due Thursday, 02/29.

In this project, we automate the data pipelines that ingest our raw data into BigQuery. We use Apache Airflow, an open source tool for authoring and orchestrating pipelines. In the next project, we will continue to work with Airflow to automate the remaining stages of our workflow. Our goal is to run the entire process from the raw layer to the consumption layer without a human in the loop.

Objectives

- Create Cloud Composer environment, a managed Airflow service on Google Cloud
- Develop an Airflow DAG that creates the raw dataset in BQ
- Develop an Airflow DAG that creates and populates the raw tables in BQ
- Delete and re-create your Cloud Composer environment to reduce billing charges

Implementation Guidelines

Please follow these guidelines when developing your Airflow DAGs:

- Store the tables generated from Airflow in a new dataset in BigQuery. The name of the dataset should follow the convention **[domain]_*raw_af*** where [domain] is the name of your data domain and ***af*** is short for Airflow. For example, airline_raw_af.
- Use the provided code samples **p5-ingest-controller.py** and **p5-ingest-table.py** as a starting point. Note: you shouldn't need to modify **p5-ingest-table.py** much, if at all.
- Ensure that the resulting tables generated through Airflow match the raw tables created from your notebook, in terms of the field names, types, and number of records per table. However, you can also make improvements in this iteration. For example, you can define a field as not null in the table if you know that to be a valid constraint.
- When not actively using Composer, delete the environment to avoid burning through all of your GCP credits. Note: there is no way to stop a running Composer instance. Follow this guide to spin up a new instance.
- Publish to your repo: **ingest-controller.py** and **ingest-table.py**.

CS 329E Project 5 Rubric

Due Date: 02/29/24

`ingest-controller.py` has all required info and correctly populates schemas **-10** did not update global variables **-10** for each schema missing **-10** if no schema_full present (with load_time) **-10** for each TriggerDagRunOperator object missing **-15** if upload .ipynb instead of .py **-60** missing file	60
Google Cloud BigQuery bucket has properly loaded all tables **-5** for each missing table **-10** if tables not under “raw_af” dataset **-30** missing file	30
`ingest-table.py` has all required methods **-5** missing create_table method **-5** missing load_table method **-10** missing file	10

`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**