CS 329E Project 6, due Thursday, 03/07.

In this project, we continue our automation effort and focus on orchestrating the pipelines in the staging layer.

Objectives

- Create Cloud Composer environment, a managed Airflow service on Google Cloud
- Develop an Airflow pipeline that creates and populates the tables in staging
- Execute the Airflow pipeline
- 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 staging tables in a new dataset in BigQuery. The name of the staging dataset should follow the convention of **[domain]_*stg_af***.
- Use the provided code samples **p6-model-controller.py**, **p6-create-bird-airports.py, and p6-create-meal-snack-service.py** as a starting point for your own DAGs.
- Ensure that the resulting tables generated through Airflow match the staging tables created from your notebook. Both the record count and contents should match although in this iteration, you may use a more precise schema (with not null constraints and non-string types).
- Take a screenshot of your model controller execution showing that all tasks completed successfully and name it **model-controller-run.png**.
- 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: **model-controller.ipy, model-controller-run.png,** and any additional sub-DAGs triggered from your main DAG.

CS 329E Project 6 Rubric
Due Date: 03/07/24

`model-controller.py` has all required info and correctly populates tables **-10** did not update global variables **-10** for each table missing **-10** for each BigQueryInsertJobOperator object missing **-15** if upload .ipynb instead of .py **-40** missing file	40
At least one file for separate sub-DAGs is present and has all required information - If multiple files, divide points by amount of DAG files **-10** does not have serialize_datetime or remove_none_values methods **-20** does not load records properly **-20** missing record insertion method (will use @task decorator) **-40** missing file	35
`model-controller-run.png` shows proper proof of Airflow controller execution **-10** missing file	10
Google Cloud BigQuery bucket has properly loaded all tables **-5** for each missing table **-10** if tables not under “stg_af” dataset **-20** missing file	15

`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**