CS 329E Project 1, due Thursday, 02/01.

Objectives

This project has 6 key objectives:

1. Find additional data sources that are related to your BIRD dataset. These can be public APIs, Faker library, public datasets or all of the above.
2. Ensure that your data meets the acceptance criteria. See details below.
3. For each new data source, write a Python script that outputs the data as a CSV file with a header row on the first line of the file.
4. Copy the CSV files to your landing zone (GCS bucket).
5. Ingest all the CSV files from your landing zone into a single dataset in BQ. This dataset will be known as the raw area.
6. Update your existing ERD and data dictionary to reflect your newly added data sources.

Acceptance Criteria

The raw data that you load into BQ needs to meet certain criteria for subsequent projects to make sense and have the right scope of work. These are detailed in the table below. The criteria apply to your dataset as a whole, not to individual tables or columns, unless otherwise noted.

Criteria	Description	Min No.	Examples
1	Dataset must have multiple unique entities. These are logical entities as opposed to how the raw tables are layed out.	5	Air Carrier, Airport, Flight, Flight History, Snack, and Meal
2	Dataset must come from multiple sources of data. You are free to come up with your own sources, you do not need to use the same ones I did.	4	BIRD, Faker, Open Food Facts, and The Meal DB.
3	Functional dependencies must hold on all tables, which means that the values are consistent across each record. For example, if a record has (city, state, country), we want the values of city and state to determine the value of country.	Applies to all tables	meals.meal_name -> meals.cat_name bird_airports.code -> bird_airports.description
4	There exists a column among the raw tables that stores more than one property in a given cell. Description and comment columns are usually a good place to look for such embeddings.	2	bird_airports.description contains these components (city, state, airport name)
5	There exists two raw tables coming from two	1	bird_airport and

	different sources that represent the same entity. However, the entity may have slightly different properties in one table from another.		faker_airport both represent an Airport entity
6	There exists a raw table that represents more than one entity. You can usually spot those tables by looking for repeated values among their records.	1	The airlines table represents two different entities: Flight and Flight History.
7	There exists at least two disjoint entities coming from different sources that could be connected through a third entity. The third entity is not present in the dataset.	1	Flights, Snacks, and Meals can be joined through an In Flight Shopping or an In Flight Meal Service entity. Neither one is present in the raw area.

What should you do if your raw data does not meet **all** of these criteria?

- Keep searching for additional sources of data until the missing criteria is met.
- If that doesn't work, change your domain by going back to BIRD and choosing a different starting database.
- Get help from the instructors if you have questions about your scenario or are struggling with one of the criteria. We may be able to offer exceptions if your data exceeds some of the criteria.

Provided Code Samples

- Faker script (uses the Faker library to generate airport records)
- Meal script (pulls data from the public API, themealdb.com, which I found by doing a google search)
- data ingest notebook (how to load the CSV files from GCS into BQ)
- ERD (represents the entities in the landing zone / raw area)
- Data Dictionary (represents the entities and sample data that are in the landing zone / raw area)

You do not need to run the Faker script or the Meal script, these are available for your reference. However, you should run through the data ingest notebook. This will allow you to load the airline data into your project and follow along in future projects.

Implementation Guidelines

- In a Markdown file, document how your data meets the 7 acceptance criteria with some specific examples. If your data doesn't satisfy a criteria or if you're unsure, make note of that as well. Name your document **acceptance_criteria.md**.

- Copy the CSV files into a folder that's in your bucket and name the folder **initial_load**.
- Create a dataset in BigQuery for storing the raw tables. The dataset name should follow the naming convention **[domain]_raw** where [domain] is the name of your data domain (e.g. airline, basketball, disney, etc.).
- When you create the tables, be sure to include a **load_time** field as the last field in the schema.
- Choose table names and field names that are meaningful.
- Table names should be in lowercase in the raw area.
- Column names should be in lowercase across the board.
- Annotate your notebook with Markdown comments to improve code readability.
- Name your updated ERD and data dictionary, **erd-raw-v2.pdf** and **data-dict-raw-v2.xlsx**, respectively.
- Publish to your repo: **acceptance_criteria.md**, **erd-raw-v2.pdf**, **data-dict-raw-v2.xlsx**, and **data-ingest.ipynb**.
- Create a submission.json file and upload it to Canvas by the deadline. Only one person per group needs to do this step.

CS 329E Project 1 Rubric
Due Date: 02/01/24

`data-ingest.ipynb` correctly written and indicates successful table creation **-5** for each missing output **-5** column names and/or table names not lowercase **-10** did not include load_time field in tables **-10** did not update variable values (project_id, bucket_name, etc) **-15** unable to verify loads through bigquery commands **-30** missing file **-20** Didn't run the code	
`acceptance_criteria.md` is thorough and has all information, criteria should be met **-5** for each criteria explanation not thorough **-10** for each missing criteria **-10** if crucial misunderstanding for each criteria **-50** missing file	50
ERD and Data Dictionary **-5** for each missing important link in ERD **-10** data dictionary does not contain essential information of a table **-10** ERD not aligned with data dictionary columns **-20** missing ERD and/or data dictionary	20

`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**