CS 329E Project 9, due Thursday, 04/12.

Objectives

- Identity three properties/attributes to enrich within one or more existing entities
- Implement a data enrichment pipeline for each of your chosen properties/attributes
- Use the `ML.GENERATE_TEXT` function in BQ to predict the values of your chosen properties/attributes
- Experiment with prompting to boost the language model's performance
- Merge the changes into the affected tables in the staging and consumption layers
- Update your existing ERDs to reflect the latest schema of your data enriched entities

Implementation Guidelines

- Implement the data enrichment pipelines in a Colab notebook named **data-enrichment.ipynb**.
- Create a dataset in BigQuery for storing the raw output from the `ML.GENERATE_TEXT` function. The dataset should be named **[domain]_stg_ai**.
- Each property/attribute that you select for enrichment should fall into one of two categories: it exists in the staging table but contains empty values in a subset of records or it does not yet exist in the staging table and would be a brand new property/attribute.
- If your table has more than 1000 records, request a quota increase for gemini-pro from the Quota page. The current value is only 300 QPM, you can request 1000 QPM.
- If your table has more than 10,000 records, aim to enrich only a subset of its records, up to 10,000 records. You can select which records to enrich by creating a smaller table if needed. Keep in mind that a BigQuery query will timeout after 6 hours.
- When merging the predictions into the enriched staging table, be sure to also update the **data_source** value of the record to indicate that the record was partly AI generated (e.g. open_food_facts_ai).
- When merging the enriched staging table into the target table, apply the same logic as from Project 8 (i.e. discontinue the previous record and insert the enriched row).
- If you made changes to your ERDs, name them **erd-stg-v3.pdf** and **erd-csp-v2.pdf.**
- Publish to your repo: **data-enrichment.ipynb, erd-stg-v3.pdf,** and **erd-csp-v2.pdf.** Note: the ERDs are only needed if you made schema changes.

CS 329E Project 9 Rubric
Due Date: 04/12/24

Necessary tables exist in stg_ai dataset in BigQuery, and remote_models dataset exists **-5** data source field not correct in stg dataset **-10** table not in stg_ai dataset **-10** remote_models dataset does not exist **-15** table missing entirely (not in any dataset)	15
Gemini prompt is generated correctly, and output is stored properly - This part of the rubric applies to any part in the .ipynb file (applies to larger scale and anywhere there is a prompt generation) **-10** does not verify prompt outputs (with select statement) **-15** prompt is incorrect or incorrectly formatted **-15** output is not formatted into JSON correctly **-20** JSON outputs not converted into tables properly	50
Original table is not updated with LLM generated output - This part of the rubric applies to anywhere outputs were generated from the LLM (for any part in the .ipynb file) **-5** does not verify updated table with select statement **-10** incorrectly adds field and/or incorrectly updates original table	20
Incorrectly merges new table to old table **-10** incorrectly updates original table with new output **-15** does not merge changes into target table	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**