CS 378 Project 4, due Thursday, **09/25**.

Objectives

- Enrich the tables in the lakehouse with the language model
- Create any missing data anomalies (criteria 6 - 10) in the raw layer

In-Scope

- Finding the anomalies in the raw dataset and showing that they exist

Out-of-Scope

- Fixing the anomalies in the raw layer

Work Items

- Query the tables in the raw dataset to show the presence of anomalies in the data
- For each missing anomaly, choose a strategy for introducing the anomaly:
  - data collection
  - data generation
  - hybrid (collection and generation)
- If you are collecting data, select the new source and bring it into the lakehouse, following the same methodology as past projects
- If you are generating data to create the missing anomaly, try to combine it with the data enrichment task to keep the number of batch prediction jobs to a minimum. Although not a hard requirement, this could save you time
- For data enrichment, choose one or more tables in the raw dataset to enrich with 3 additional attributes. Use the language model to create them. The enriched attributes should be derived from existing attributes present in the Iceberg tables
- If you are missing one or more anomalies, you have the option to create them while enriching the data to save time. Follow these guidelines:
  - If missing criteria 6, inject “\n” or similar character for null values
  - If missing criteria 7, inject a list of values for different attributes into the same cell and across multiple rows
  - If missing criteria 8, inject a list of values for the same attribute in the same cell and across multiple rows
  - If missing criteria 9, identify an entity that is represented in more than one physical table of your lakehouse, usually in a parent-child relationship. Inject a new identifier in the child table as its foreign key
  - If missing criteria 10, inject a derived attribute into a child table such that the attribute describes the parent relation
- Replace the existing Iceberg tables with the changed tables in the raw layer
- Extend the new tables with the `_data_source` and `_load_time` fields and populate them as normal

Code Samples

See snippets repo for code samples

Implementation Details

- Create a new folder in your repo and name it `project4`. Store all your artifacts for this project in the `project4` folder.
- Create a Colab notebook named `project4-data-enrichment.ipynb`.
- Annotate your notebook with section headers and short Markdown comments throughout your work.
- Import the provided notebook into your Colab and follow the same steps.
- Once you have extracted the predictions from the response object, store the results into a separate table in the tmp dataset (`[your-domain]_tmp`). Name this table `[source_table]_enriched`. For example, if your prediction table is named `airport_reviews_data`, the intermediate table should be named `airport_reviews_data_enriched`.
- Join the enriched table with the iceberg table from the raw dataset and materialize the results into a new iceberg table, replacing the old one
- Update your data dictionary and ERD to reflect the new and changed tables in the raw dataset. The dictionary should include all the fields for every table that you have loaded into the lakehouse. The ERD should contain the most important fields for every table or entity in your lakehouse. This data types and relationships. You do not need to list the `_data_source` and `_load_time` fields in the ERD or any other field that is unimportant. Think of the ERD as a bird's eye view of the raw layer and the data dictionary as a granular view of the raw layer.
- Commit and publish your work to your GitHub repo. Remember that all artifacts need to go into a `project4` folder. This includes your notebook, data dictionary, ERD, and criteria analysis.
- Create a submission.json file and upload it to Canvas by the deadline. Only one person per group needs to do this step.

Grading Rubric

Due Date: 09/25/25

Data enrichment sections of `project4-data-enrichment.ipynb` are correctly written and they indicate successful execution **-10** for each missing or empty enrichment attribute **-8** for each enrichment attribute that was not derived from existing attributes **-7** for each output table in the raw dataset that has fewer rows than its source tables in the raw and temp datasets **-7** at least one data file (parquet) and Iceberg file (metadata) are not located in the expected path **-7** at least one final table is missing or has empty `data_source` or `load_time` fields **-7** at least one final table is not located in the `[your-domain]_raw` dataset **-10** ERD and/or data dictionary missing or out-of-date with tables in the raw layer **-75** missing notebook **-70** didn't run the code	75
`Data anomaly` sections of `project4-data-enrichment.ipynb` are thorough and have all the information showing that the anomalies exist in the raw layer (whether they were sourced from a new dataset or generated by the language model or simply queried). All 5 anomaly types (criteria 6-10) should be covered in the notebook. **-5** for each missing or misunderstood anomaly **-5** for each anomaly which was not obtained via approved methods	25

`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**