

CS 329E Project 3, due Thursday, **02/05**.

### Objective

- Build the second half of the raw layer of lakehouse

### In-Scope

- The unstructured data sources (pdf, jpg, mp3, etc.)

### Out-of-Scope

- The structured data sources (csv, tsv, json, etc.)

### Work Items

- Extract attributes from unstructured datasets
- Convert results to parquet file(s)
- Create the Iceberg tables in BQ
- Load the tables with the parquet files
- Extend the tables with `_data_source` and `_load_time` fields
- Populate the `_data_source` and `_load_time`
- Check your work against our validation criteria

### Code Samples

See [snippets](#) repo for code samples

### Implementation Details

- Create a new folder in your repo and name it `project3`. Store all your artifacts for this project in the `project3` folder.
- Create a Colab notebook named `project3-attribute-extraction.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.
- Consult the Batch Inference Job [documentation](#) as needed.
- Create the `jsonl` folder in the root of the bucket to hold the input files.
- Each dataset should have its own `input_data.jsonl` file. Make a subfolder for this file under the `bucket/jsonl` path. For example,  
`gs://air-travel-open-access/jsonl/airport-maps/input-data.jsonl`.
- Before you run the Batch Inference Job, create a temp dataset in BQ for storing the response objects. Follow the naming convention: `[your-domain]_tmp` for your dataset (e.g. `air_travel_tmp`, etc.).
- If you run into any issues with the Batch Inference job, go to the [UI](#) to see the state of the job. If your job gets stuck, you can try to cancel it and run it in a different region (e.g.

us-east1) or with a different [model](#). When choosing a model, please keep in mind the [pricing](#).

- Each dataset should have its own parquet folder under the `lakehouse` area. For example, `gs://air-travel-open-access/jsonl/airport-maps/input-data.jsonl`.
- After loading your table, run some basic sanity checks by retrieving some sample rows and ensuring that the results look fine.
- Refine the prompt (and experiment with different models) until you get the results you're expecting. This is an iterative process.
- Update the data dictionary and ERD to reflect your new tables in the raw layer. 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.
- Review the [validation criteria](#) and cross-reference them with your tables. Determine which ones you have satisfied to-date and which ones are still missing. Update your Markdown file (`criteria-analysis.md`) with the results from your analysis. Please include a specific example for each criteria you are able to satisfy. If you are missing a criteria that is required by project 3 (as indicated by the Required By field), try to resolve it ASAP. If you are not able to do so on your own, make note of it in your analysis and discuss it with the Prof. To be clear, you are expected to meet the first 5 validation criteria, the ones that are marked as required by all projects and project 3.
- Commit and publish your work to your GitHub repo. Remember that all artifacts need to go into a `project3` 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: 09/18/25

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <p><code>project3-attribute-extraction.ipynb</code> correctly written and indicates successful execution</p> <ul style="list-style-type: none"><li>-10 for each missing or empty table in temp dataset in BQ</li><li>-10 for each missing or empty table in raw dataset in BQ</li><li>-7 for each output table in the raw dataset that has fewer rows than its source table in the temp dataset</li><li>-7 at least one data file (parquet) and Iceberg file (metadata) are not located in the right path</li><li>-7 at least one table in raw dataset is missing or has empty <code>data_source</code> or <code>load_time</code> fields</li><li>-10 ERD and/or data dictionary missing or out-of-date with tables in the raw layer</li><li>-90 missing notebook</li><li>-80 didn't run the code</li></ul> | 90         |
| <p><code>criteria-analysis.md</code> is thorough and has all information. All 10 criteria, including the 6 anomaly types should be included in the analysis. Include placeholders for any criteria that you are not able to satisfy.</p> <ul style="list-style-type: none"><li>-5 at least one of the required criteria for this project is not met (criteria 1-5)</li><li>-2 for each required criteria whose explanation is not thorough</li><li>-2 for crucial misunderstanding of a required criteria</li><li>-10 missing file</li></ul>                                                                                                                                                                                                                                                               | 10         |
| <p><code>submission.json</code> submitted into Canvas. Your project <b>will not</b> be graded without this submission. The file should have the following schema:</p> <pre>{   "commit-id": "your most recent commit ID from Github",   "project-id": "your project ID from GCP" }</pre> <p>Example:</p> <pre>{   "commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",   "project-id": "some-project-id" }</pre>                                                                                                                                                                                                                                                                                                                                                                                       | Required   |
| <b>Total Credit:</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | <b>100</b> |