------------------------------ReadMe---------------------------

AIM

The aim of the project is to work on end to end extraction ,transforming and Loading pipeline that can extract data from a CSV files and an API, could clean and enrich it with many transformations, and loads the results into a Google Cloude Storage Bucket.

About the Datasets Used

To explore more about the structure, duplicates, and missing values, the raw datasets were examined.

Transactions Dataset Schema

transaction\_id → int64

user\_id → int64

timestamp → object

amount → float64

currency → object

Null values

currency: \_\_\_ rows

amount: \_\_\_ rows

Clickstream Dataset Schema

user\_id → int64

session\_id → object

timestamp → object

url → object

Null values

url: \_\_\_ rows

timestamp: \_\_\_ rows

CSV files Extractions

The CSV files are read in chunks of 50,000 rows to handle the large datasets efficiently.

The live conversion rates are taken with the help of ExchangeRate API which is stored in rates.json file.

storage path:- data/raw/api\_currency/YYYY-MM-DD/rates.json

Basic Transformations

The datasets were made perfect by converting some of the factors in datasets. Timestamps are converted to UTC, conversion of column names into a combination of lowercases letters and underscores. For transactions datasets, a new column was introduced using the conversion rates from the API.

Example row : amount = 1000, currency = INR, USD rate = 83.1 → amount\_in\_usd = 12.03

Loading to Google Cloud

The processed datasets after transformations are saved as a partitioned Parquet files locally by ingestion date. Later they are Uploaded for Structural Storage to a Google Cloud Storage.

Processed files saved locally as Parquet:

data/processed/transactions/ingest\_date=YYYY-MM-DD/transactions\_part0.parquet

data/processed/clickstream/ingest\_date=YYYY-MM-DD/clickstream\_part0.parquet

Uploaded to Google Cloud Storage bucket:

gs://data-engineering-intern/processed/transactions/ingest\_date=2025-09-06/

gs://data-engineering-intern/processed/clickstream/ingest\_date=2025-09-06/

Alerts and Logs

The steps of pipeline were closely monitored usings Pythons Logging module, with proper messages if the process is working well and warnings if there is any issues.

Example log outputs:

INFO:root:Successfully fetched exchange rates

INFO:root:Saved cleaned chunk 0 → data/processed/transactions/...

Uploaded ... → gs://data-engineering-intern/processed/...

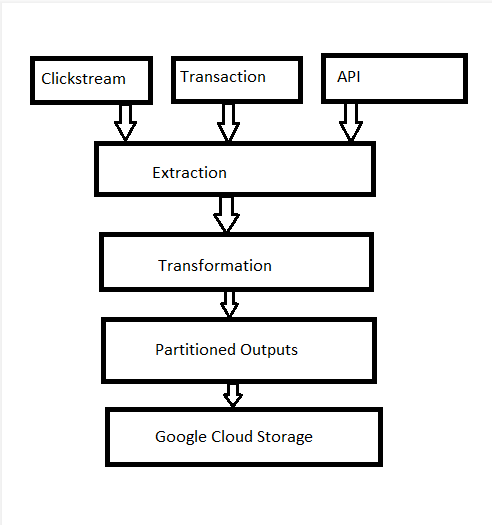
Pipeline finished successfully!

Warnings displayed if:

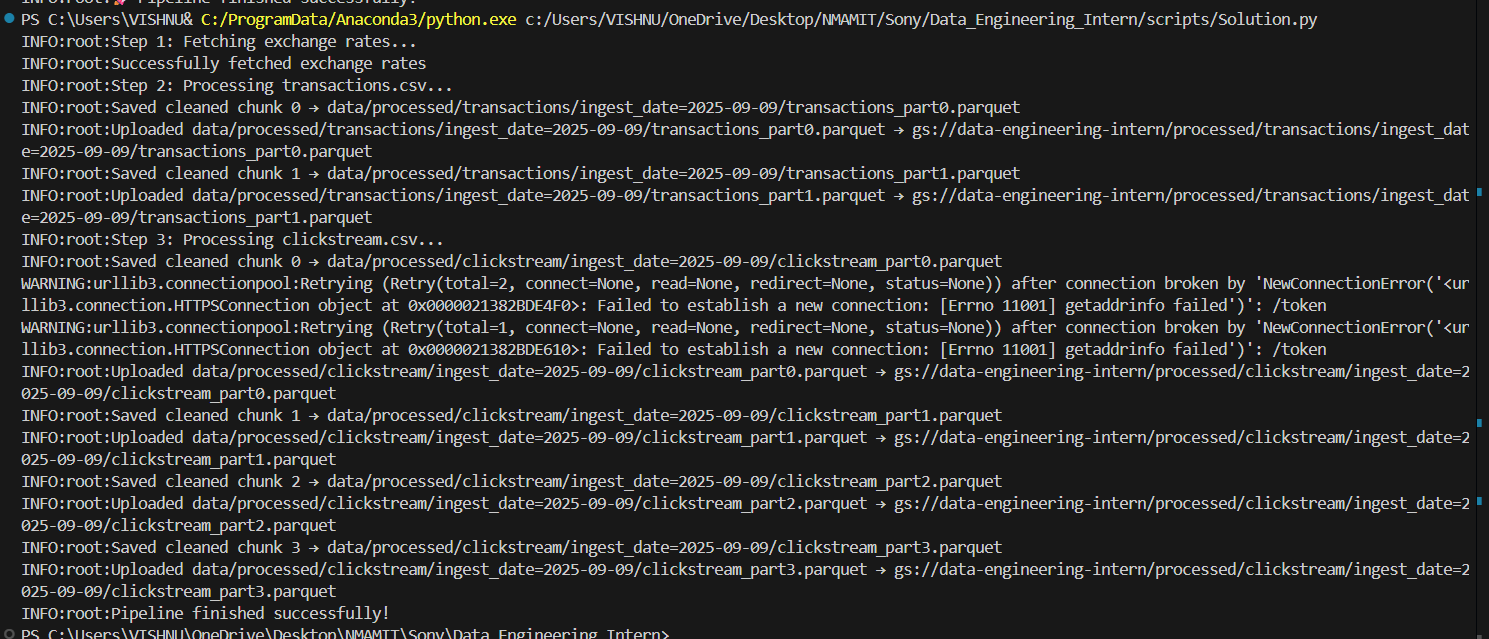
API fails to return results.

Input CSV files are missing.

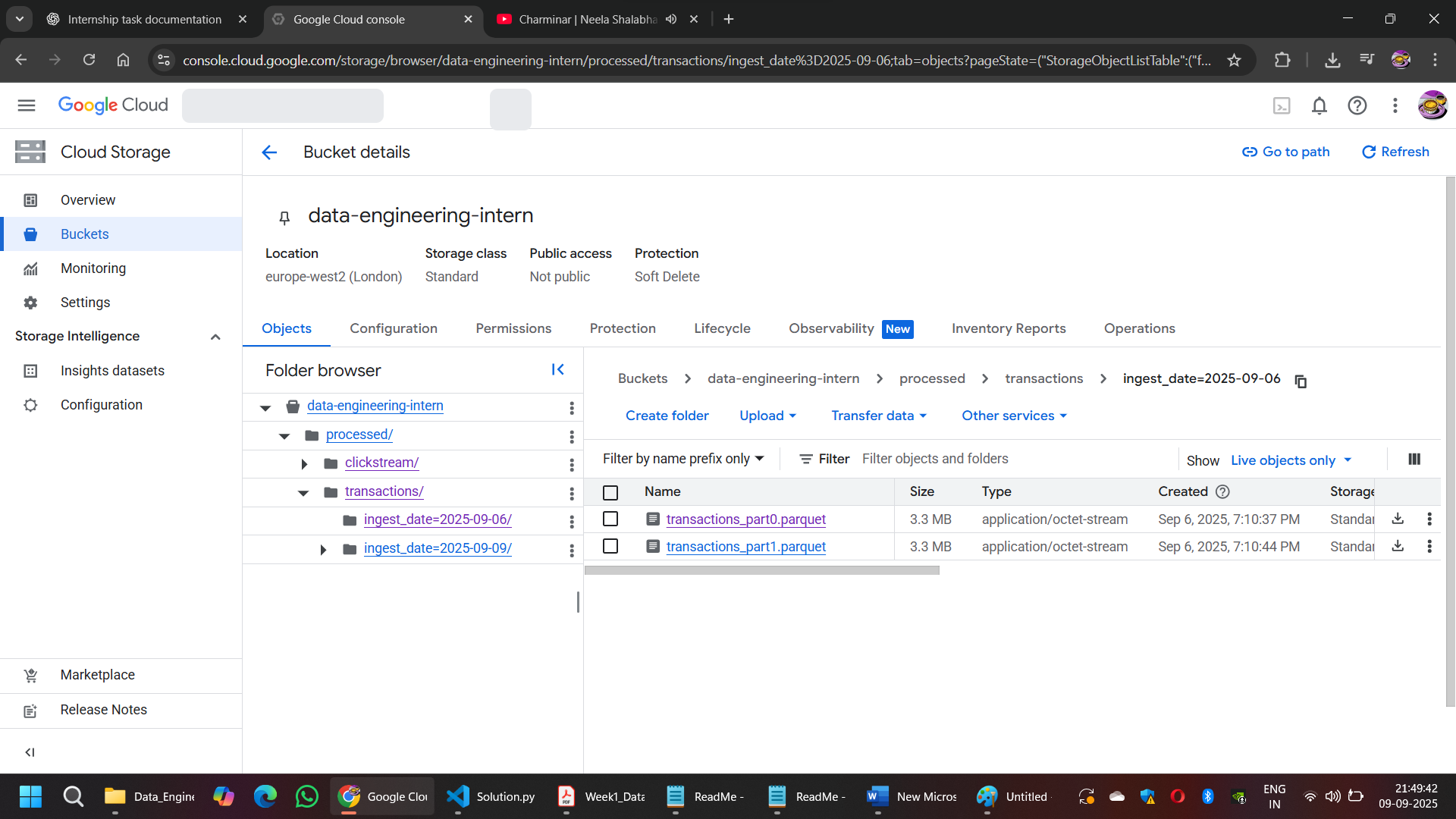
Architecture Diagram and Relevant Screenshots



Terminal Logs



Google Cloud Storage



Verified Outputs

