HOMEWORK 2 Lidiya Hanicheva

Github: https://github.com/lidiyaganicheva/uku_airflow/tree/main/homework_2

Provide a Docker Compose file within your repository. This file should:

Extend the default Docker Compose configuration as outlined here.

Include any additional libraries that need to be installed.

Override default configurations (e.g., set load_examples = False).

Define a separate database for storing information, distinct from the Airflow metastore.

1) Separate database postgres_storage is added
2) Load_examples is set to False
3) A Dockerfile with additional libraries is created (run docker compose build is required)
4) Init_connections.sh file is created – all necessary connections and variables are configured automatically
5) /dags, /logs, /config and other volumes are mounted

The pipeline must utilize Google OCR or Easy OCR to recognize data and store links to the marketing materials along with the recognized data.

The pipeline should extract company information using the domain names identified by OCR and enrich this data using one of the following methods: PeopleDataLabs, Web Scraping combined with ChatGPT, or BrandFetch.

Implement a deduplication process to ensure:

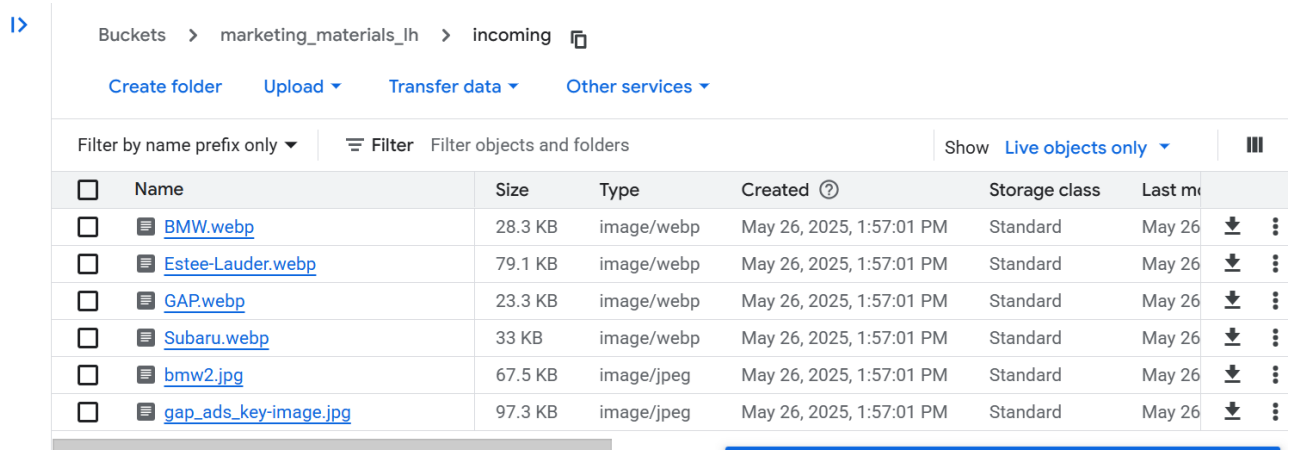
Only new companies are added to the database.

Offers from different runs but for the same company are consolidated under that company.

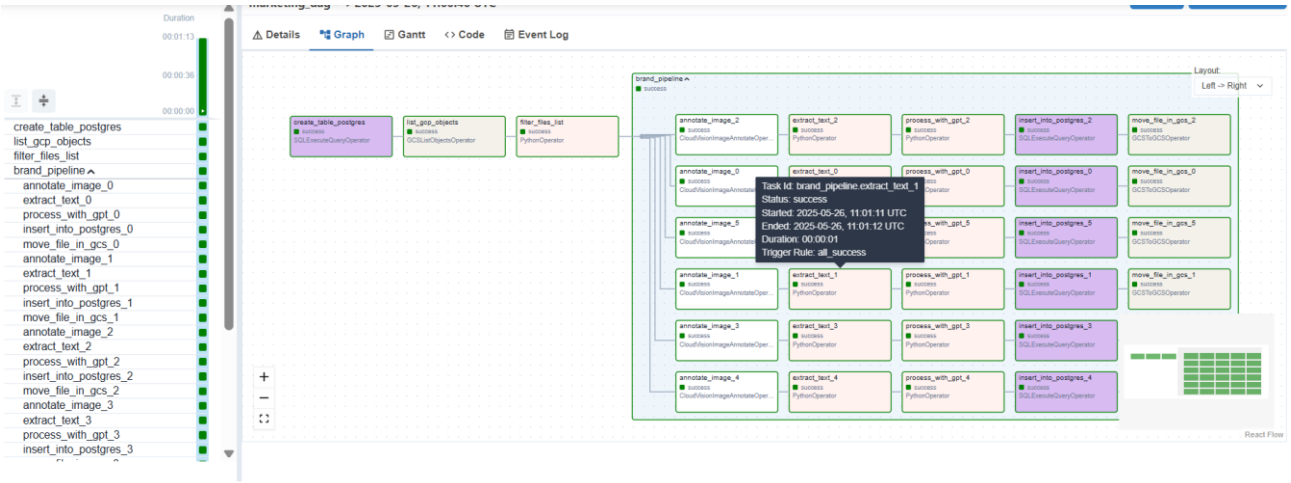
1) Marketing materials are placed in GCP bucket, in incoming/ folder
2) Table materials is created using SQLExecuteQueryOperator
3) Dag receives list of available marketing materials using GCSListObjectsOperator
4) Root folder incoming/ is removed from this list using PythonOperator
5) Materials are processed in parallel using TaskGroup:
6) Using CloudVisionImageAnnotateOperator (Google Vison AI) all text on an image is recognized
7) As Google Vision AI output contains a lot of information that is not necessary in our case, it is cleansed using PythonOperator
8) Using PythonOperator, cleansed text is sent to Chat Gpt and name of the brand and additional data are received
9) Output is written on the database table. To avoid duplicates, only new brands are inserted, others are updated. If there are several offers from the one company, they will be represented as a list. As processing is running concurrently brand record is locking for update.
10) After processing the materials, they are removed from folder incoming/ and placed to the folder processed/

Job run example:

1) Add materials to bucket. There are 2 offers from BMW and GAP to check deduplication



2) Dag run



3) Resulting record in the database. All offers from one brand are consolidated to a list.

brand_name	description	marketing_materials
ESTEE LAUDER	A renowned cosmetics and skincare brand known for its high-quality products.	{Estee-Lauder.webp}
GAP	American clothing and accessories retailer	{GAP.webp,gap_ads_key-image.jpg}
BMW	Bavarian Motor Works (BMW) is a well-known automobile manufacturer known for producing luxury vehicles.	{bmw2.jpg,BMW.webp}
SUBARU	Subaru is a well-known automotive manufacturer specializing in vehicles known for reputation, reliability, and safety.	{Subaru.webp}

(4 rows)