Instructions for Using the BigQuery-PySpark-MongoDB Data Engineering Project

This Data engineering project involves the use of Docker, Apache Airflow, Terraform, PySpark, BigQuery, and MongoDB. The project is designed for building a robust data pipeline with two microservices and ingest big amazon_review dataset from local host and save it into BigQuery and processes dataset with pyspark and loads data into MongoDB for using sentiment analysis ML app.

Project Overview

The project includes two microservices and components to handle data ingestion, preprocessing, and processing. This includes:

- **Docker Compose setup** for containerized services.
- **Apache Airflow DAGs** for managing workflows.
- **Terraform** configurations for infrastructure setup.
- **PySpark** scripts for data processing.

Credentials File

The project requires access to Google Cloud services for BigQuery. The credentials file contains sensitive information such as service account keys for Google Cloud. This file is ignored in the repository using .gitignore to prevent accidental commits of sensitive data.

Tip: If you want to create an account in BigQuery and MongoDB, generate your credentials JSON file for BigQuery and set the MONGODB_URI for MongoDB. Then, replace my credentials file with yours. Just don't forget to update the JSON file name in the ingestion and preprocessing Python code accordingly.

How to Use:

1- Firstly you should download this file Google Cloud Service Account (big-test-449715-2b0e9010365e.json) from (https://drive.google.com/drive/u/4/folders/16qyKNRDmMrKSCbH7vv9_51Ajmy5Byf_A) google drive address and paste it into these folders: microservices_data ingestion, microservices_data preprocessing, terraform.

2- Secondly you should download the dataset from (https://drive.google.com/drive/u/4/folders/16qyKNRDmMrKSCbH7vv9_51Ajmy5Byf_A) and then paste it into microservices_data ingestion folder.

3- And you should create two images from Ingestion DockerFile and Preprocessing DockerFile with the following commands , just don't forget firstly go to the path of each microservices and then run these command on the right path:

run this command in the path of Ingestion folder

3.1- docker build -t ingestion-microservice .

run this command in the path of preprocessing folder

3.2- docker build -t preprocess_micro .

4- And then run the docker desktop and build a docker-copmose and up it with the following command:

run these command in the path of Bigquery-Pyspark-MongoDB

4.1- docker compose build

4.2- docker-compose up

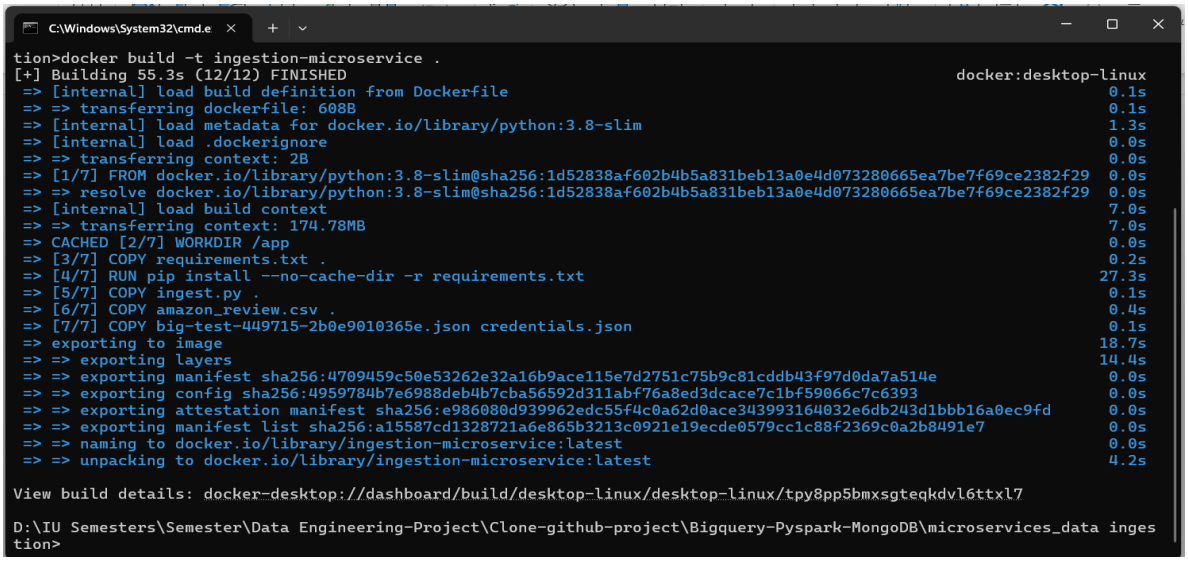
Or you can use this alternative command instead of two commands

docker-compose up --build

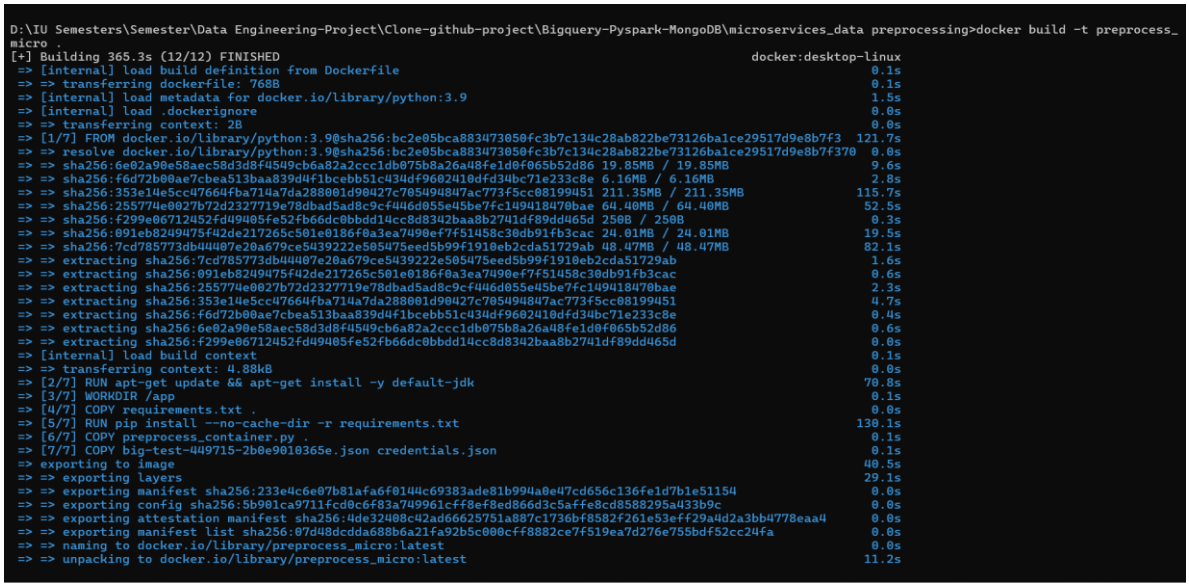
5- And then you can see all logs that shows ingest data from local server and save it into BigQuery and then with another microservices get a part of dataset into Pyspark and preprocess it and then save the cleaned-data into MongoDB.

screenshots from different steps of project

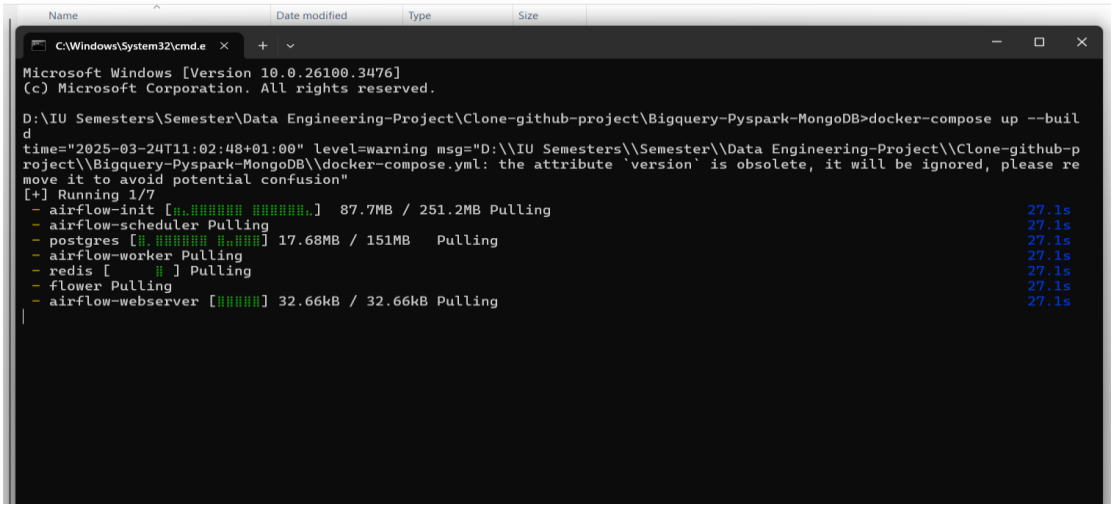
Creating Image docker for ingestion



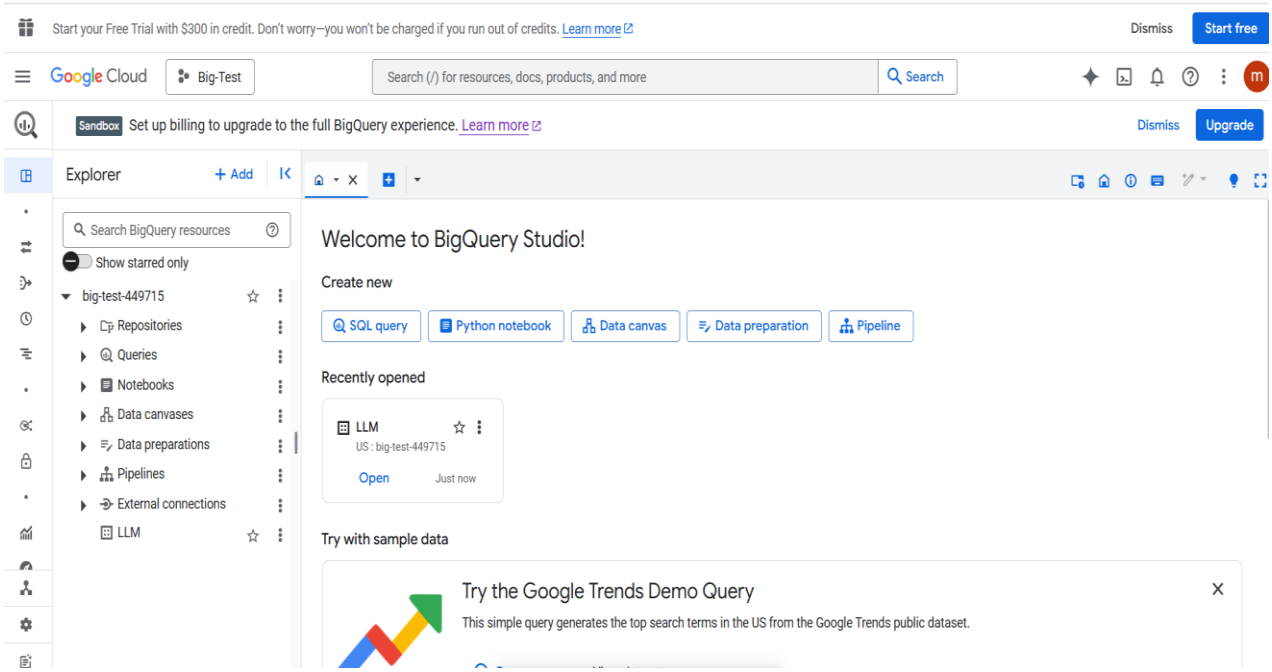
Creating Image docker for preprocessing



build docker container with docker compose file

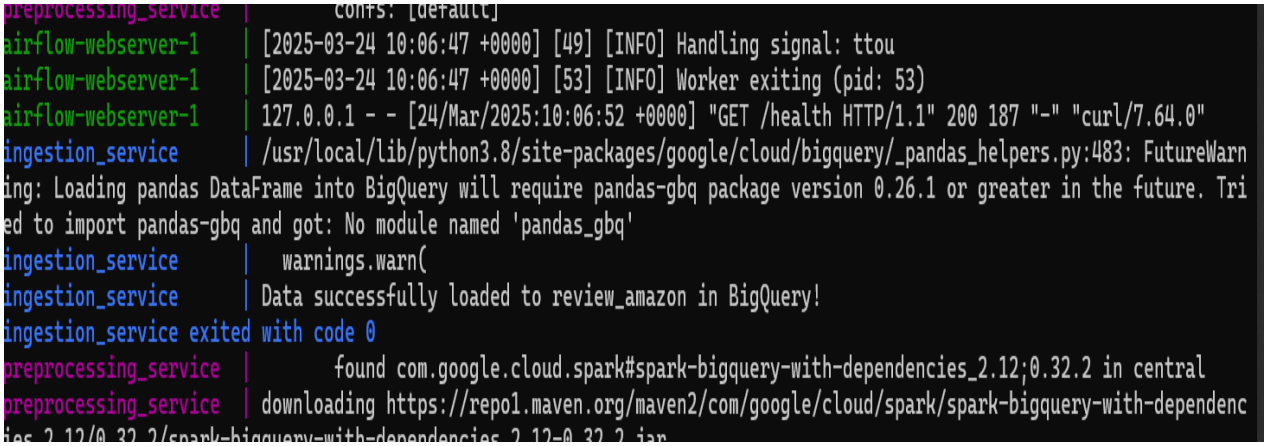


Bigquery account before ingesting data into LLM table

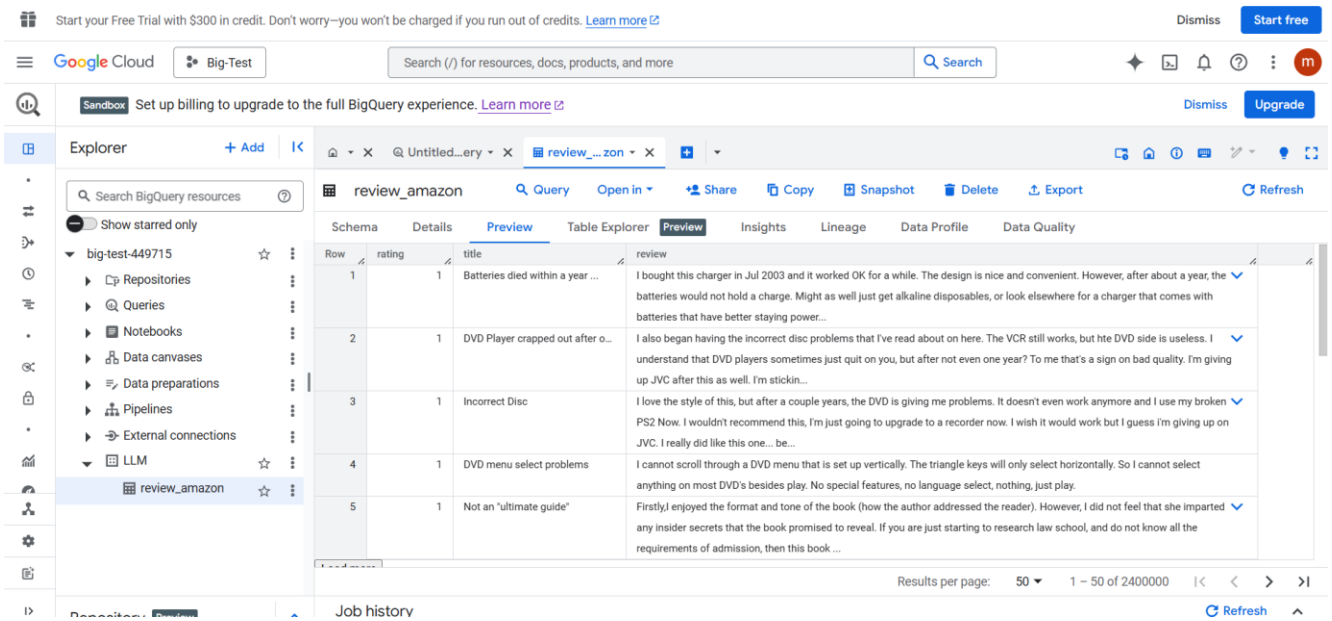


Log for ingesting big dataset from local server and save it into BigQuery account

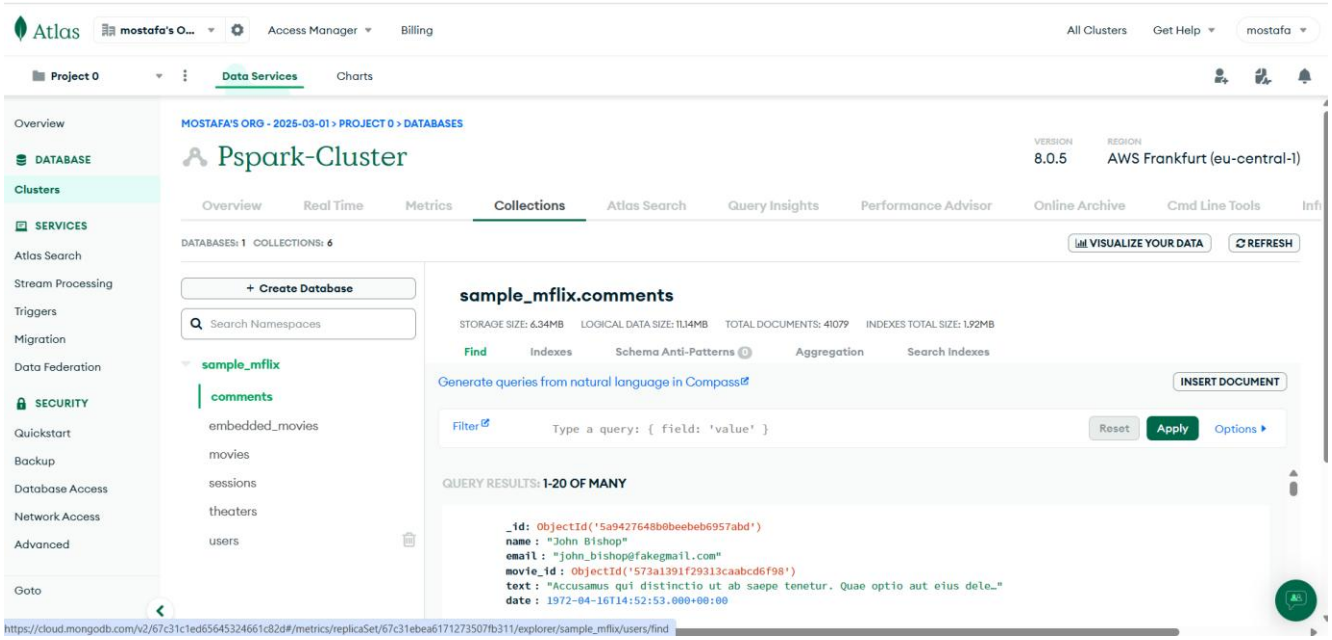
Data Successfully loaded to review_amazon in Bigquery!



Dataset is loaded in Bigquery



mongoDB account before saving cleaned-data into preprocessed-review table



Data loaded from Bigquery into pyspark and cleand and save it into mongoDB

```
w.providers.google.common.hooks.leveldb.LevelDBHook' from 'apache-airflow-providers-google' package: No module named 'ai
rflow.providers.google.common.hooks.leveldb'
airflow-webserver-1     | [2025-03-24 10:07:18 +0000] [49] [INFO] Handling signal: ttou
airflow-webserver-1     | [2025-03-24 10:07:18 +0000] [54] [INFO] Worker exiting (pid: 54)
airflow-webserver-1     | 127.0.0.1 - - [24/Mar/2025:10:07:22 +0000] "GET /health HTTP/1.1" 200 187 "-" "curl/7.64.0"
preprocessing_service   | Loading data from BigQuery...
preprocessing_service   | Data loaded! Cleaning now...
preprocessing_service   | Connecting to MongoDB...
preprocessing_service   | Inserted 100 records into MongoDB
preprocessing_service   | PySpark session stopped
ingestion_service       | /usr/local/lib/python3.8/site-packages/google/cloud/bigquery/_pandas_helpers.py:483: FutureWarn
ing: Loading pandas DataFrame into BigQuery will require pandas-gbq package version 0.26.1 or greater in the future. Tri
ed to import pandas-gbq and got: No module named 'pandas_gbq'
ingestion_service       |   warnings.warn(
ingestion_service       | Data successfully loaded to review_amazon in BigQuery!

ingestion_service exited with code 0
preprocessing_service exited with code 0
airflow-webserver-1     | 127.0.0.1 - - [24/Mar/2025:10:07:32 +0000] "GET /health HTTP/1.1" 200 187 "-" "curl/7.64.0"
airflow-webserver-1     | [2025-03-24 10:07:41 +0000] [49] [INFO] Handling signal: ttin
```

mongoDB account after saved preprocessed-data into preprocessed-review table

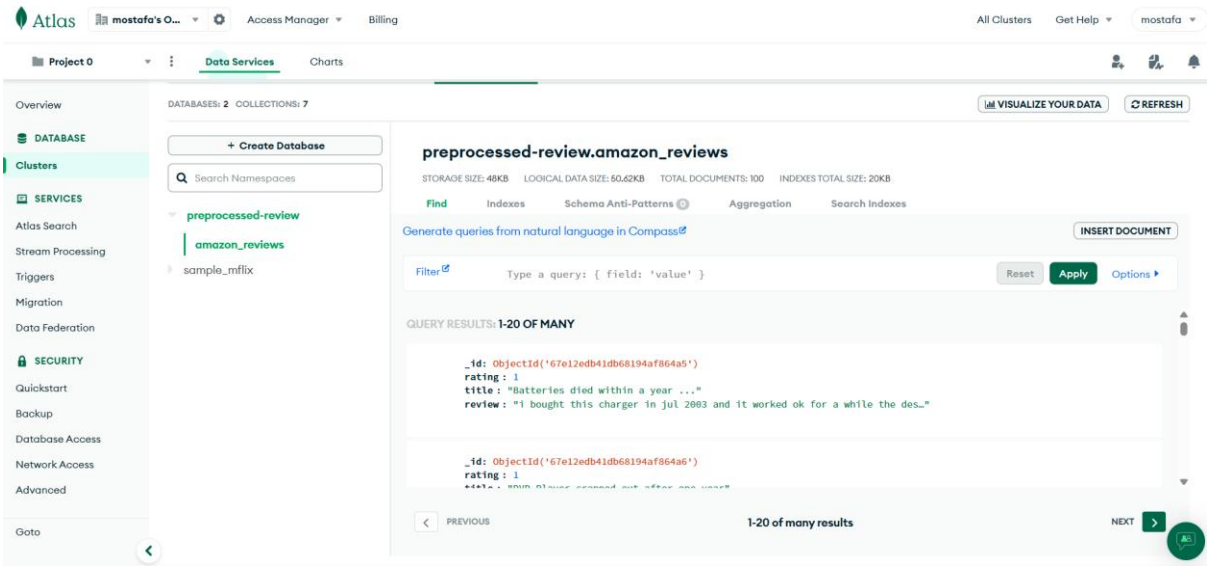


Image dockers

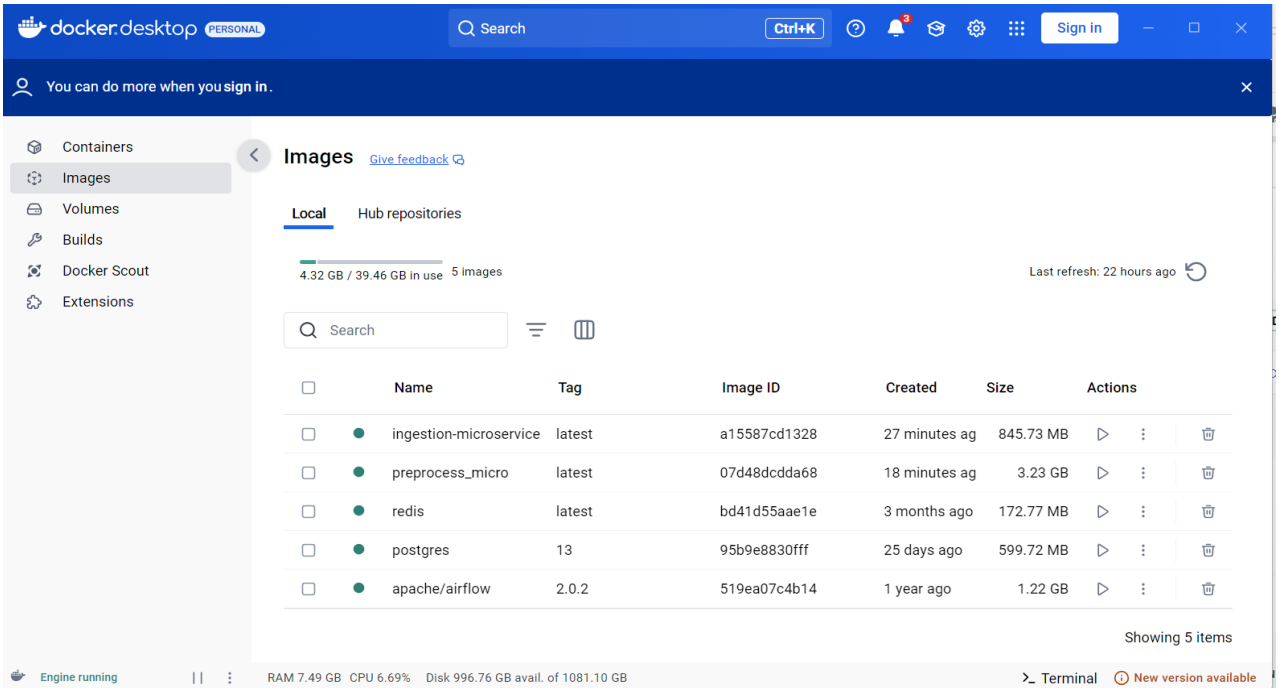
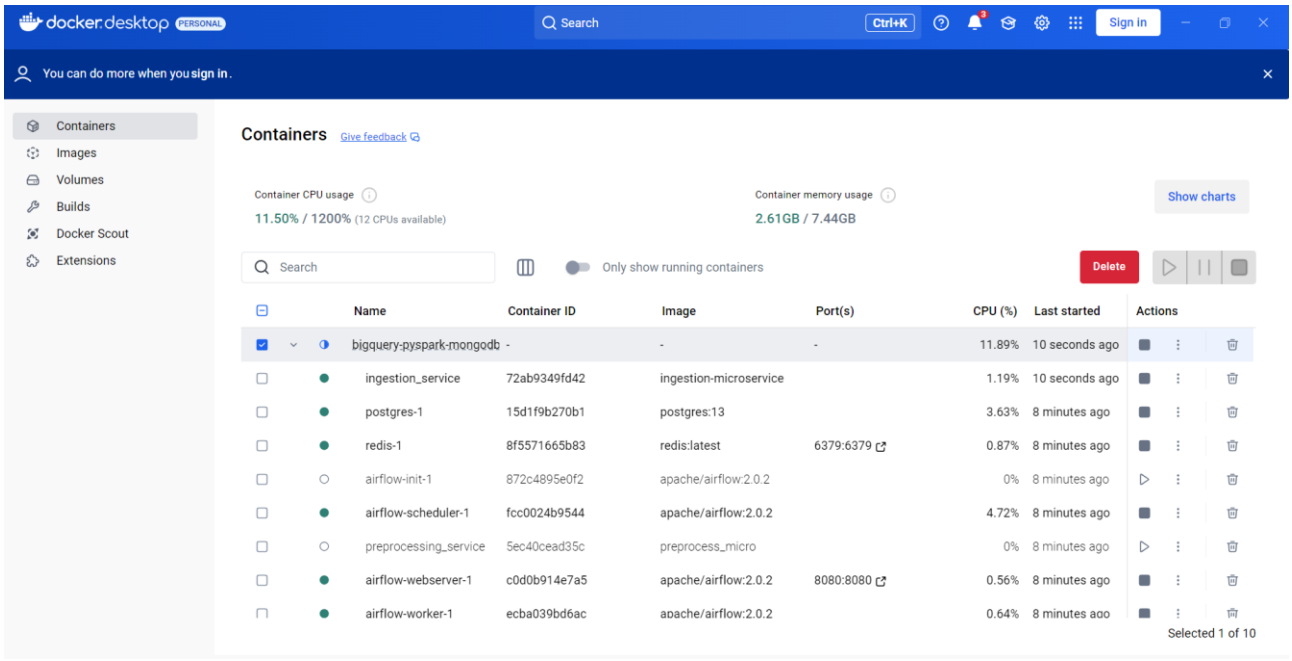


Image container



Ruuning Airflow