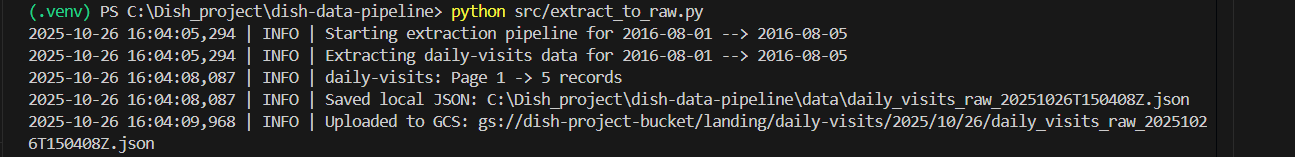
Dish Data Pipeline Project – Documentation and Testing Report

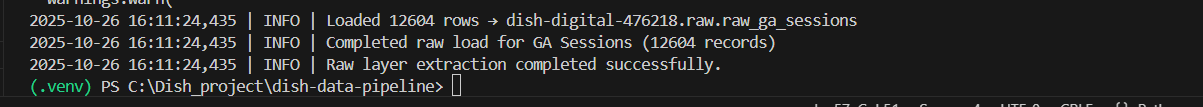
-- Created by: Dhanush H

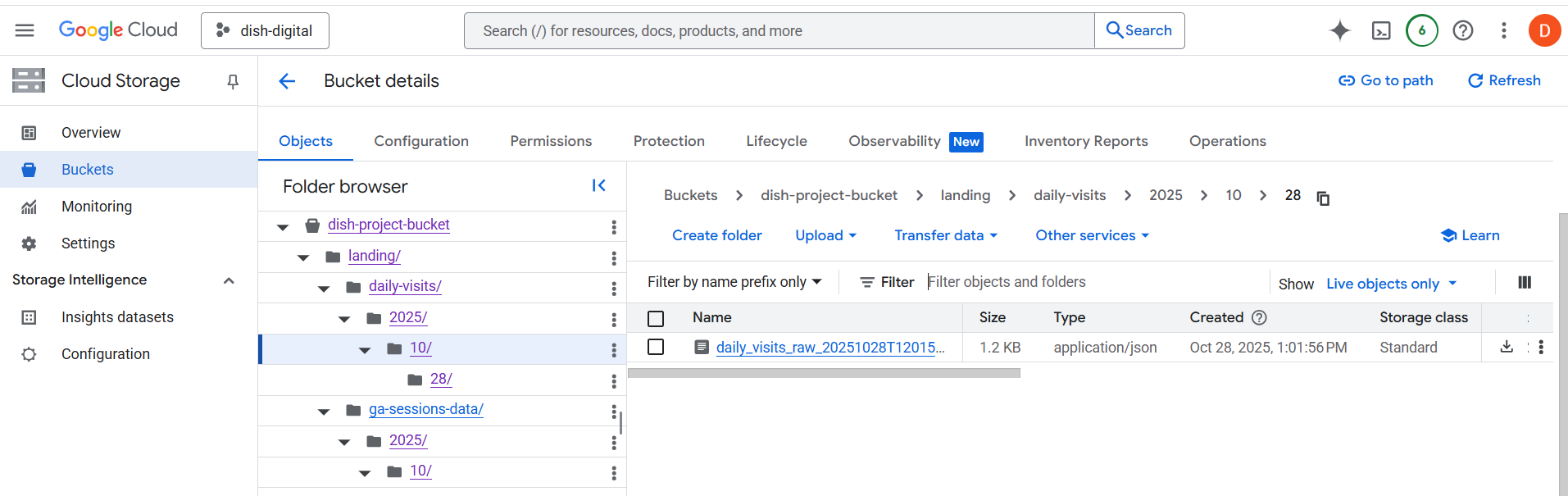
1. **Development and Testing data\_pipeline.py**

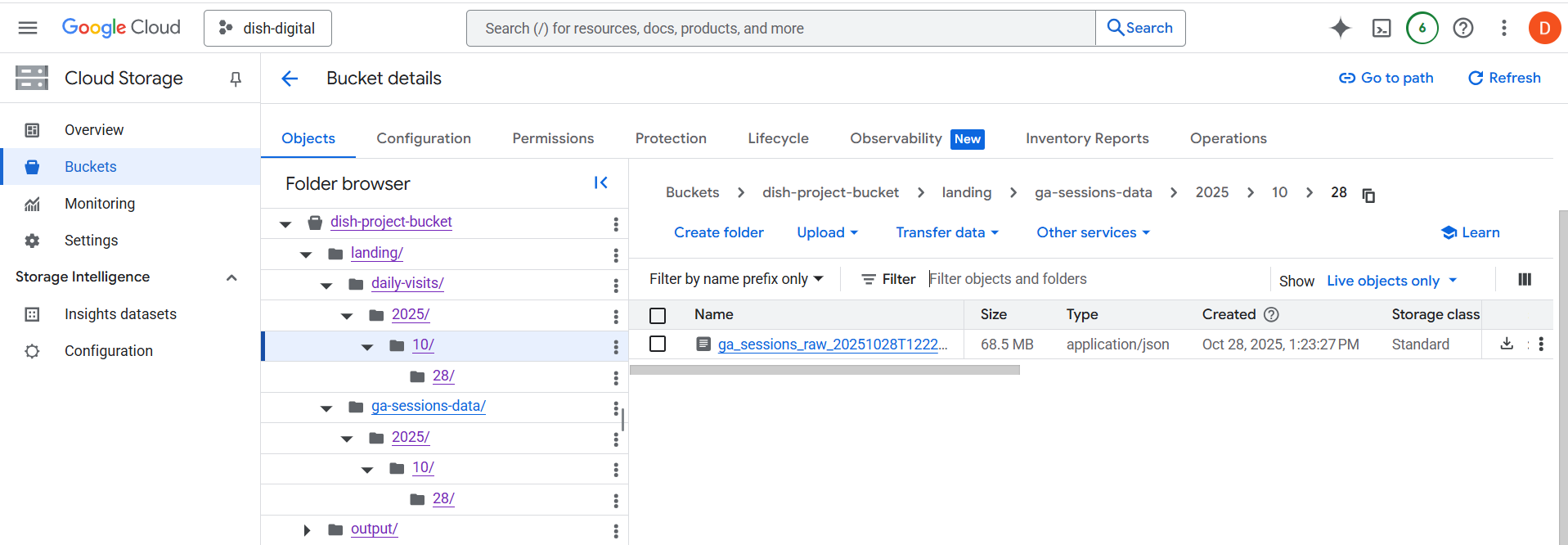
Note: Initially, I named data\_pipeline.py as extract\_to\_raw.py and tested.

This script will extract data from Dish API and stores raw json files locally and uploads it to GCS bucket. Then Loads those raw JSONs into BigQuery raw tables, preserving the source of truth





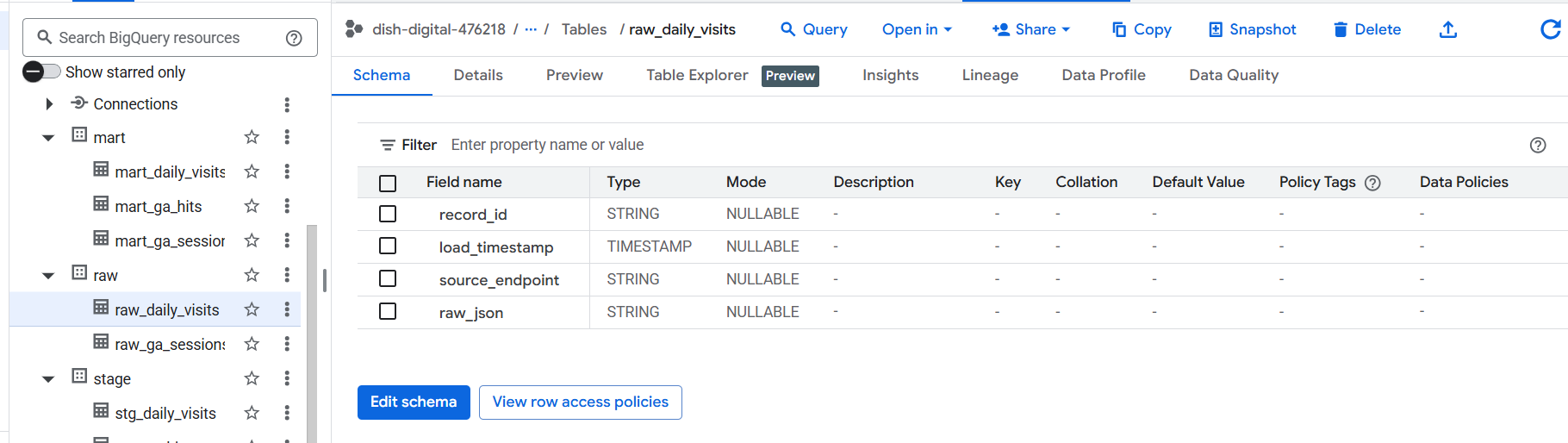




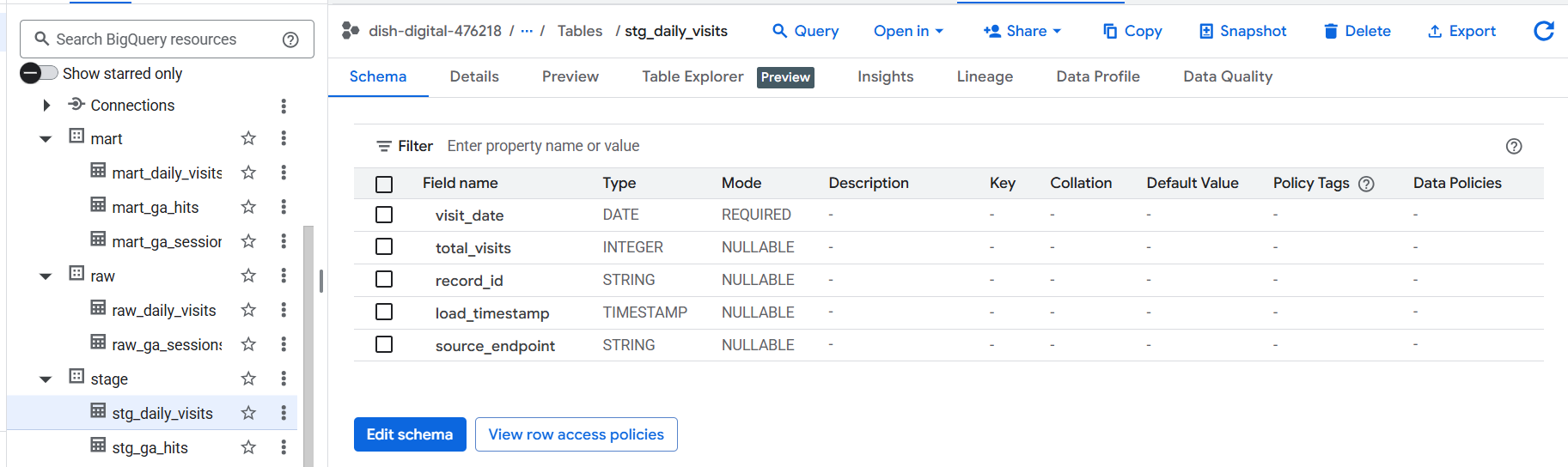
1. **BigQuery Tables and count validations**

**Tables structures:**

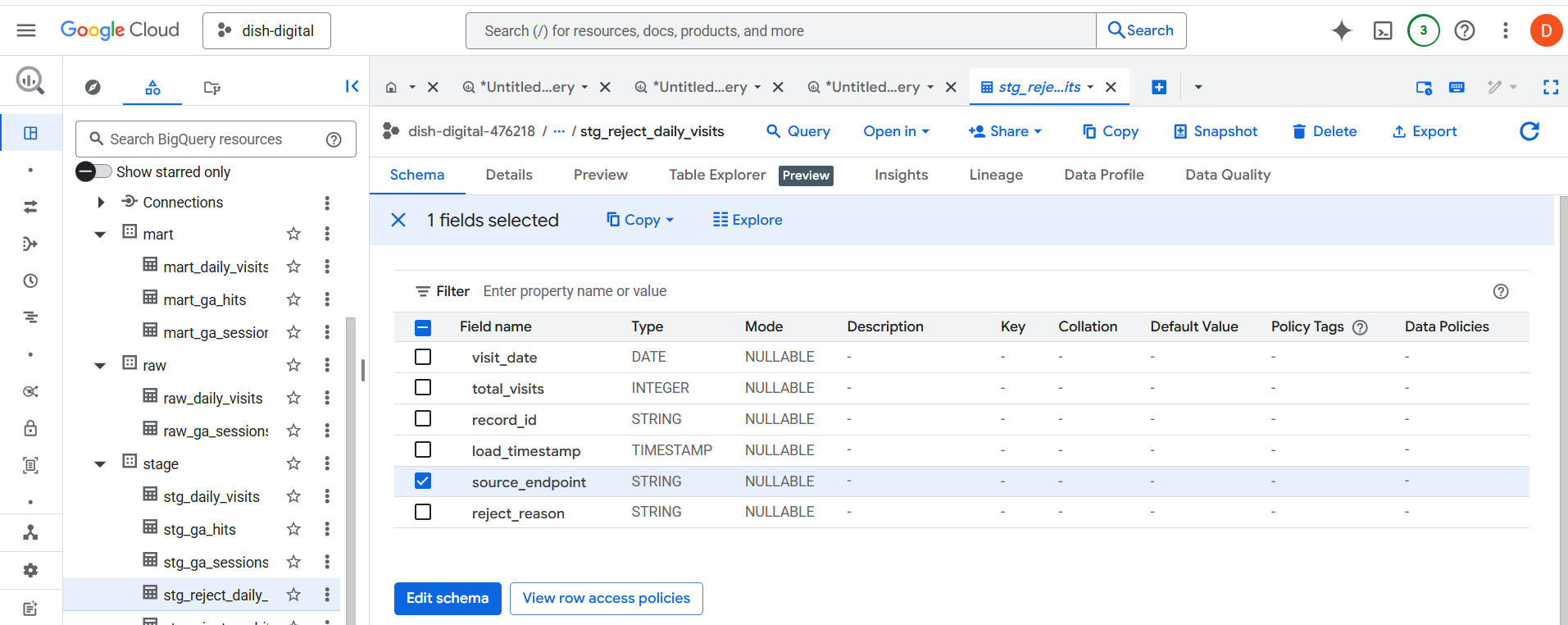
**raw\_daily\_visits**



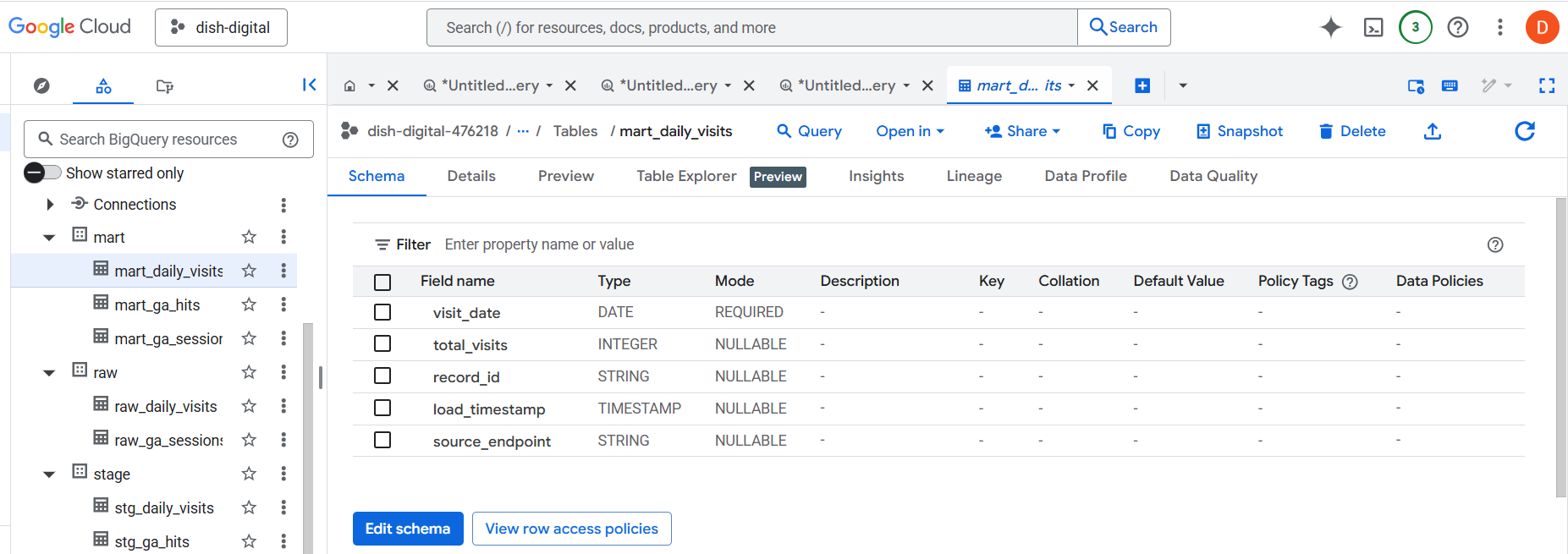
**Stg\_daily\_visits**



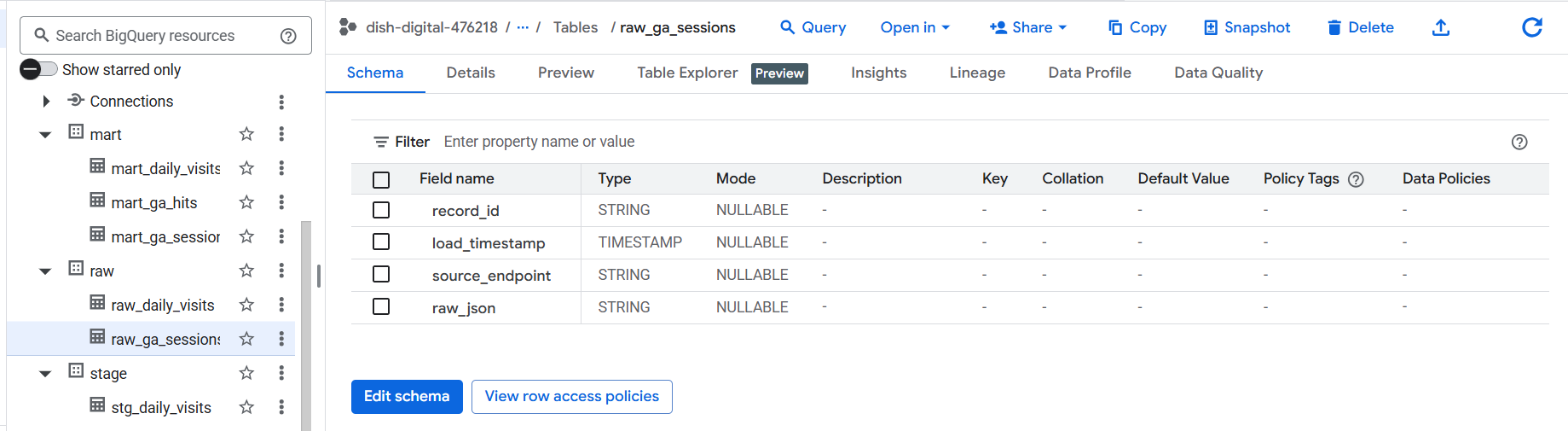
**Stg\_reject\_daily\_visits**



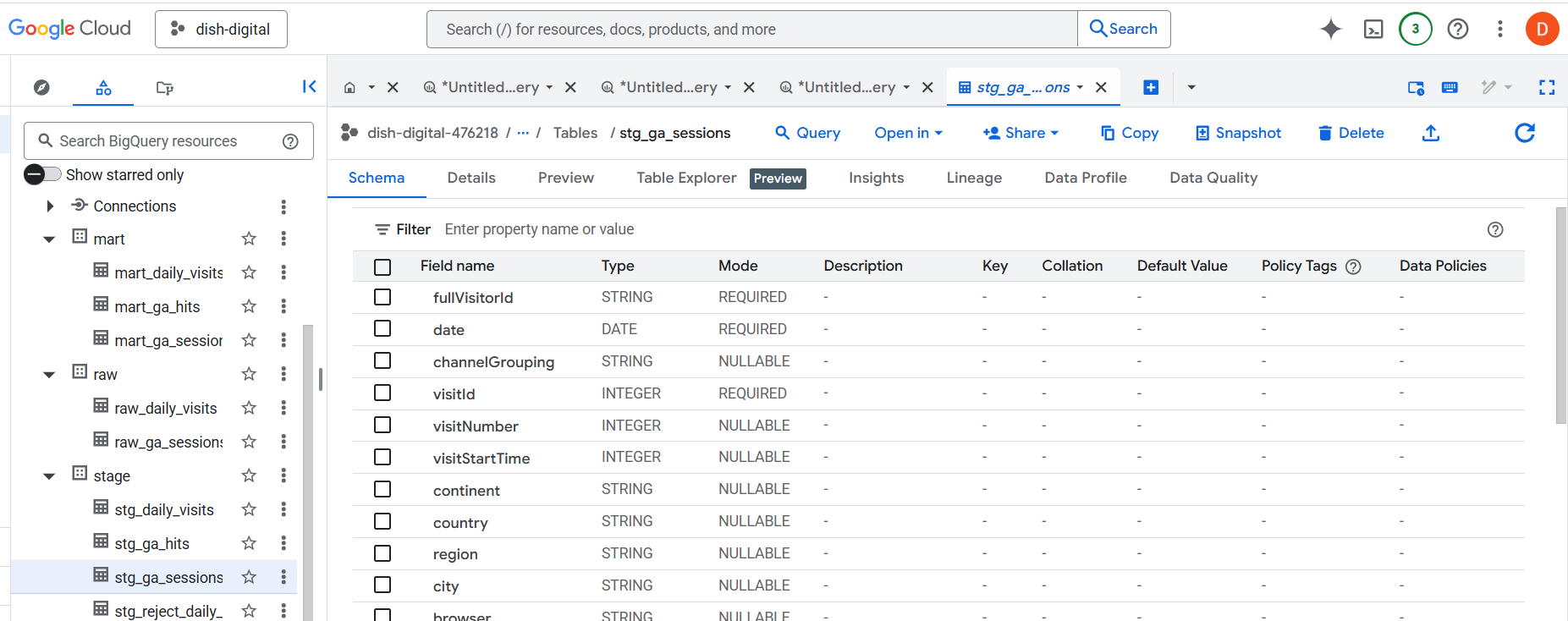
**Mart\_daily\_visits**



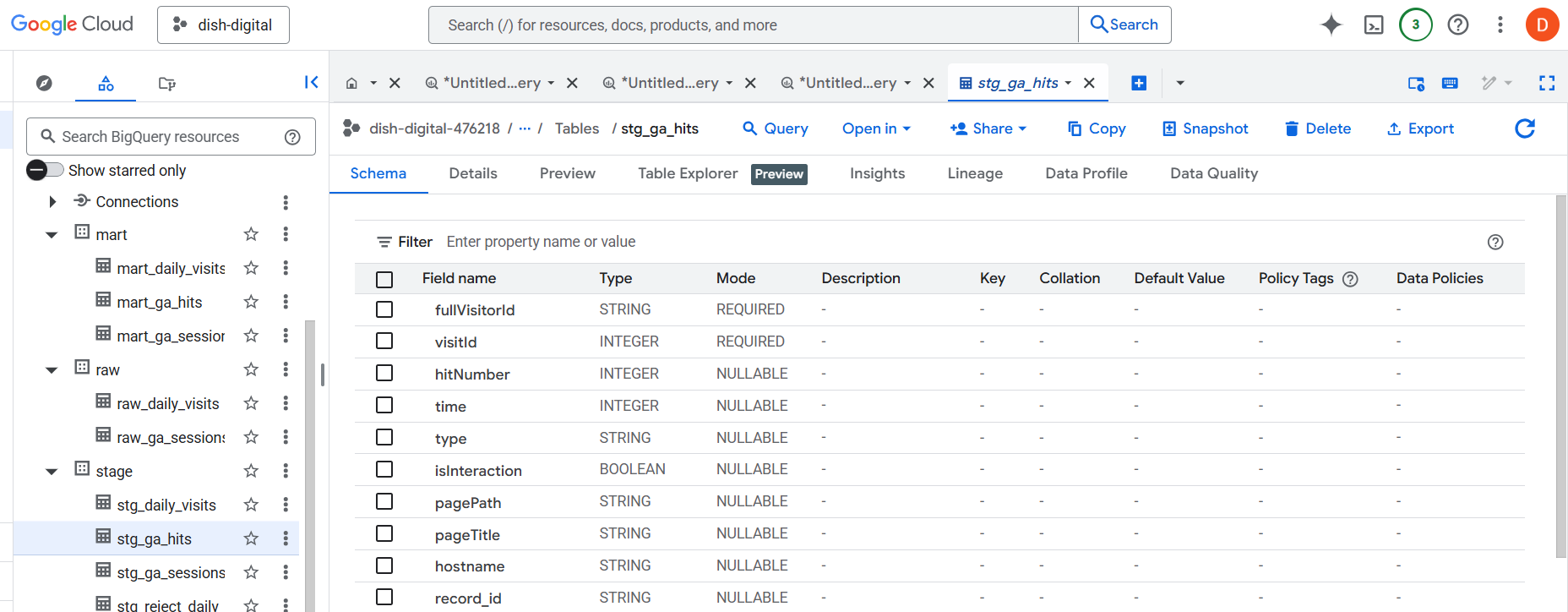
**raw\_ga\_sessions**



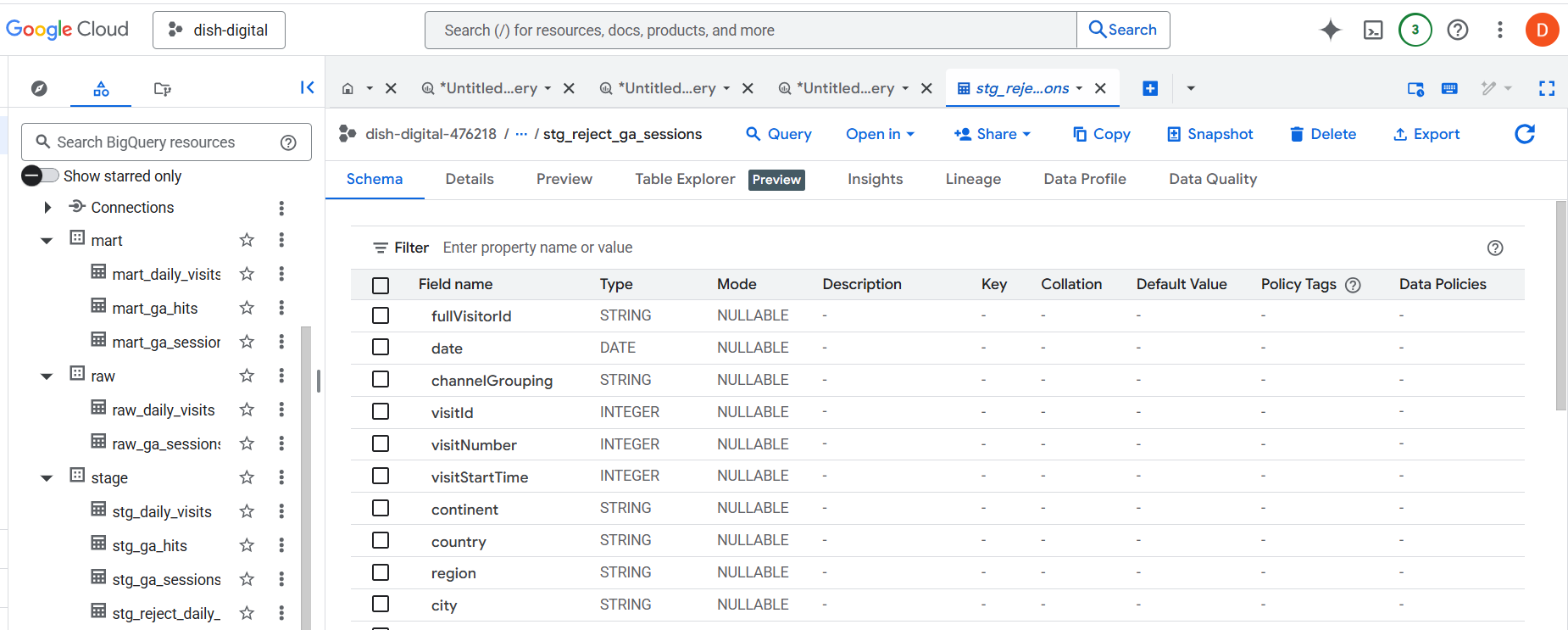
**Stg\_ga\_sessions**



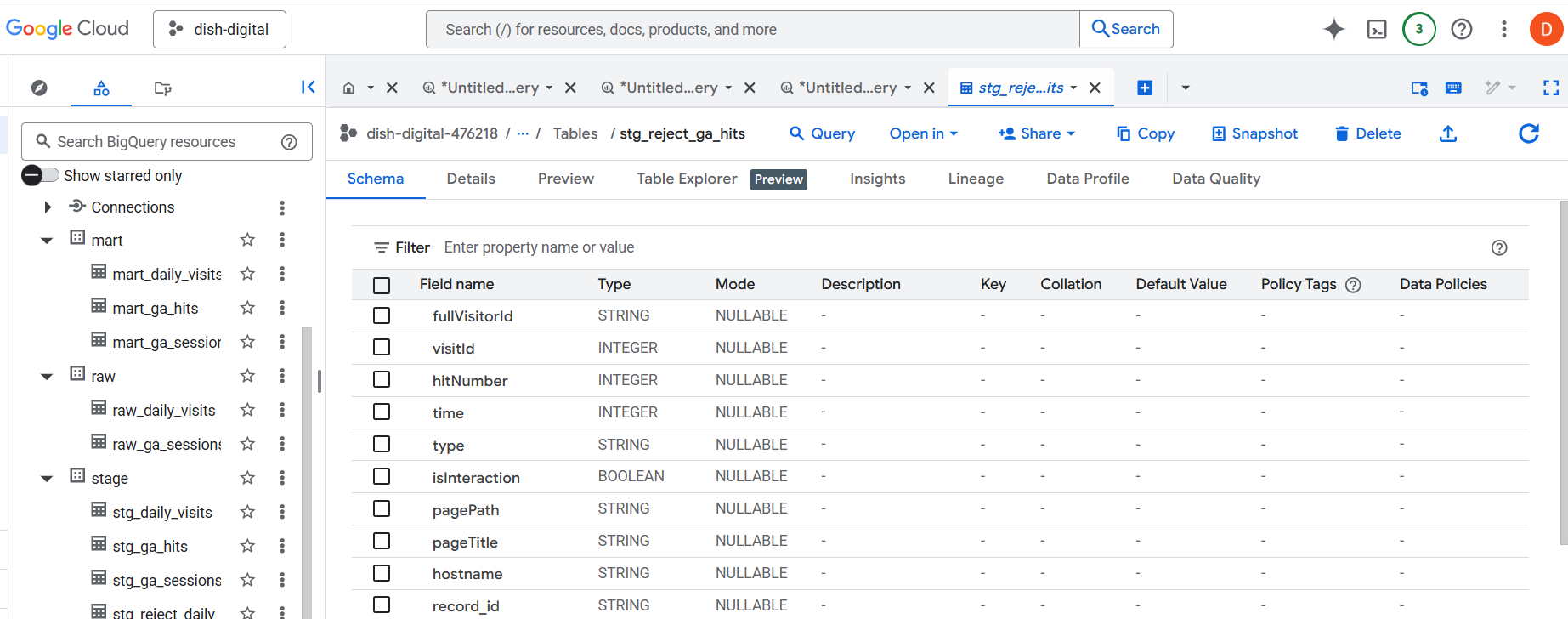
**Stg\_ga\_hits**



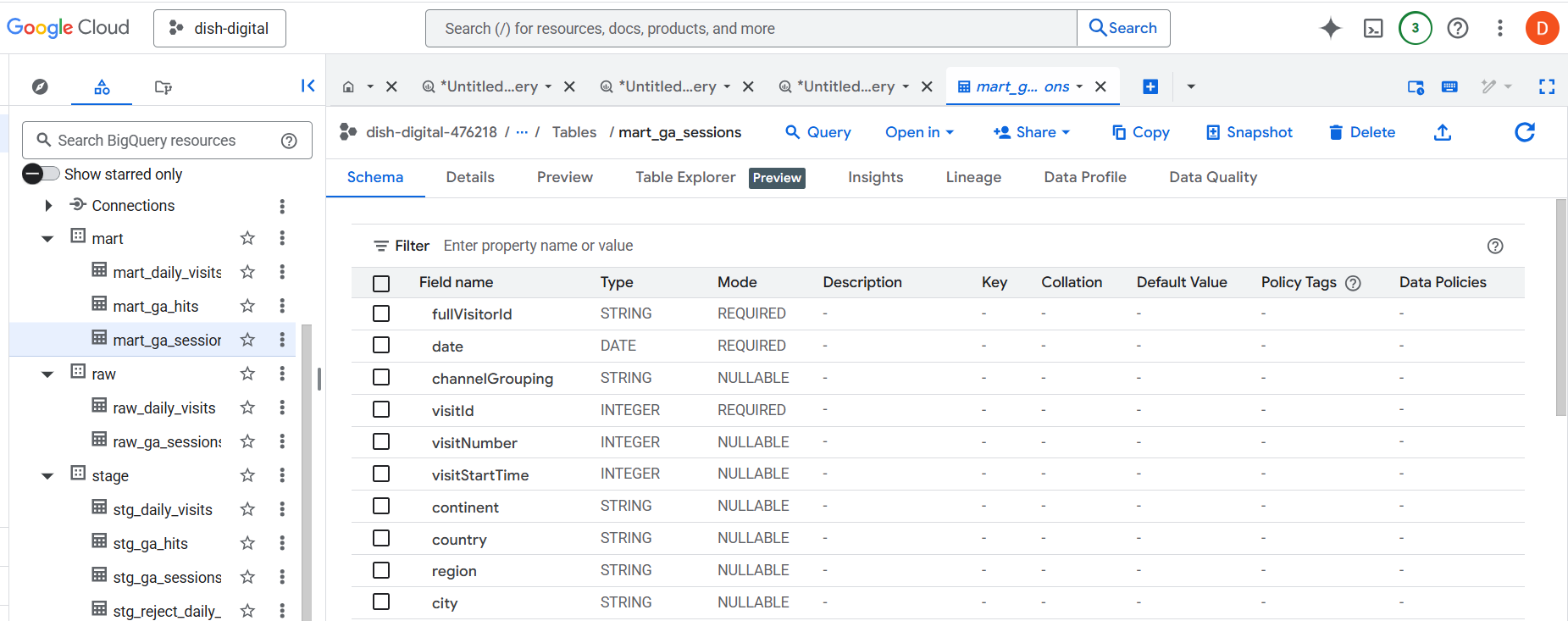
**Stg\_reject\_ga\_sessions**



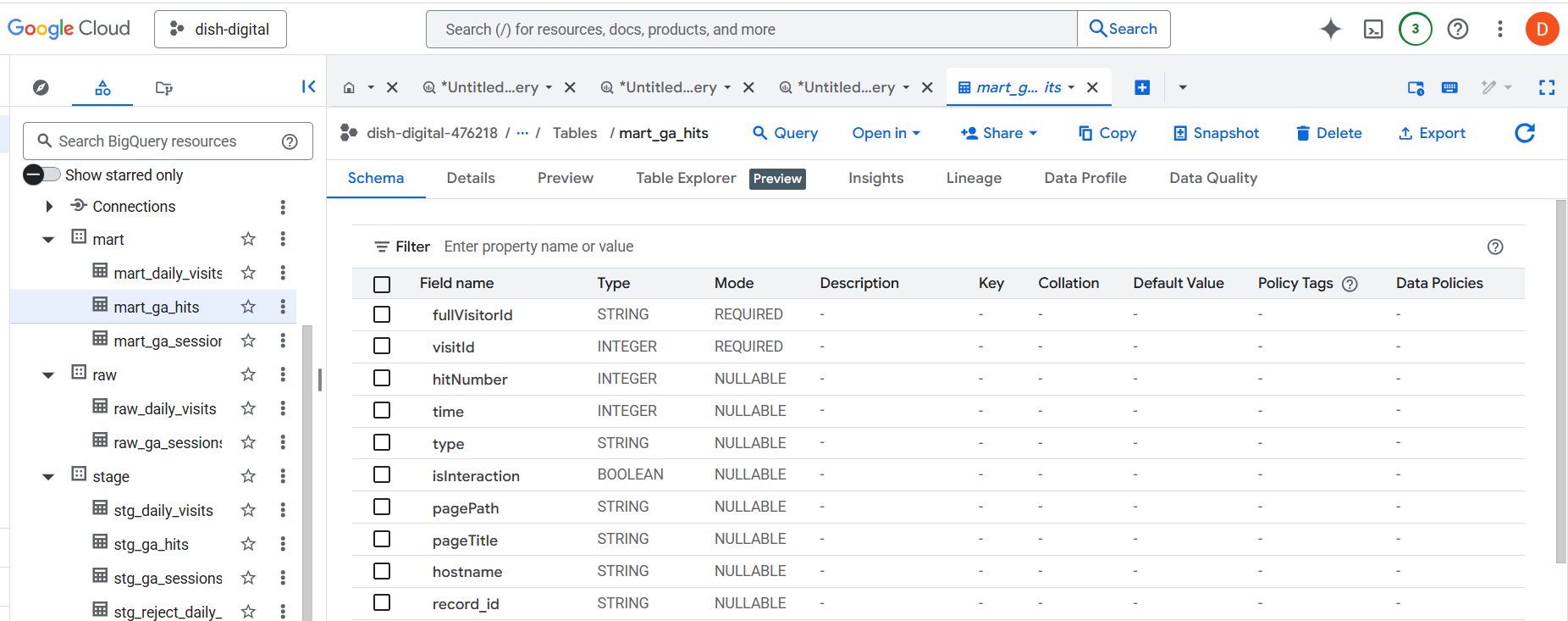
**Stg\_reject\_ga\_hits**



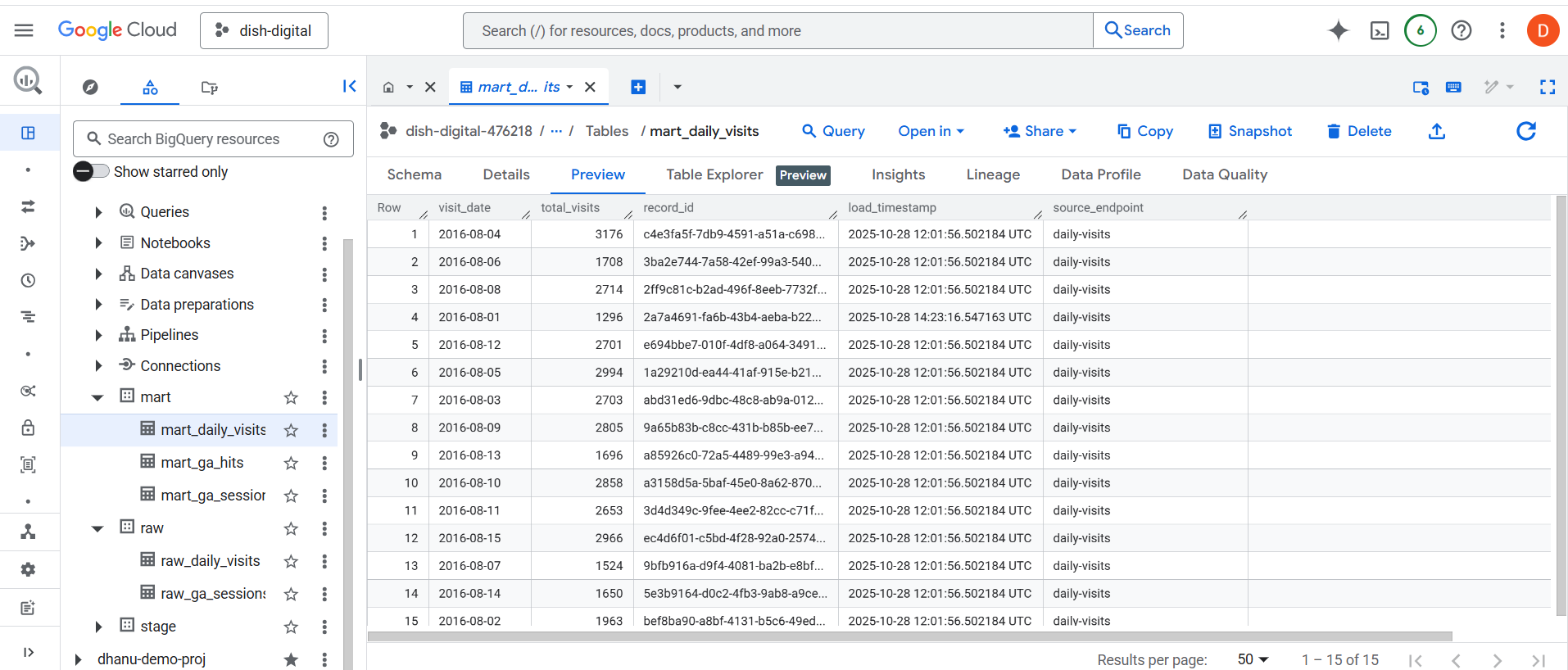
**Mart\_ga\_sessions**

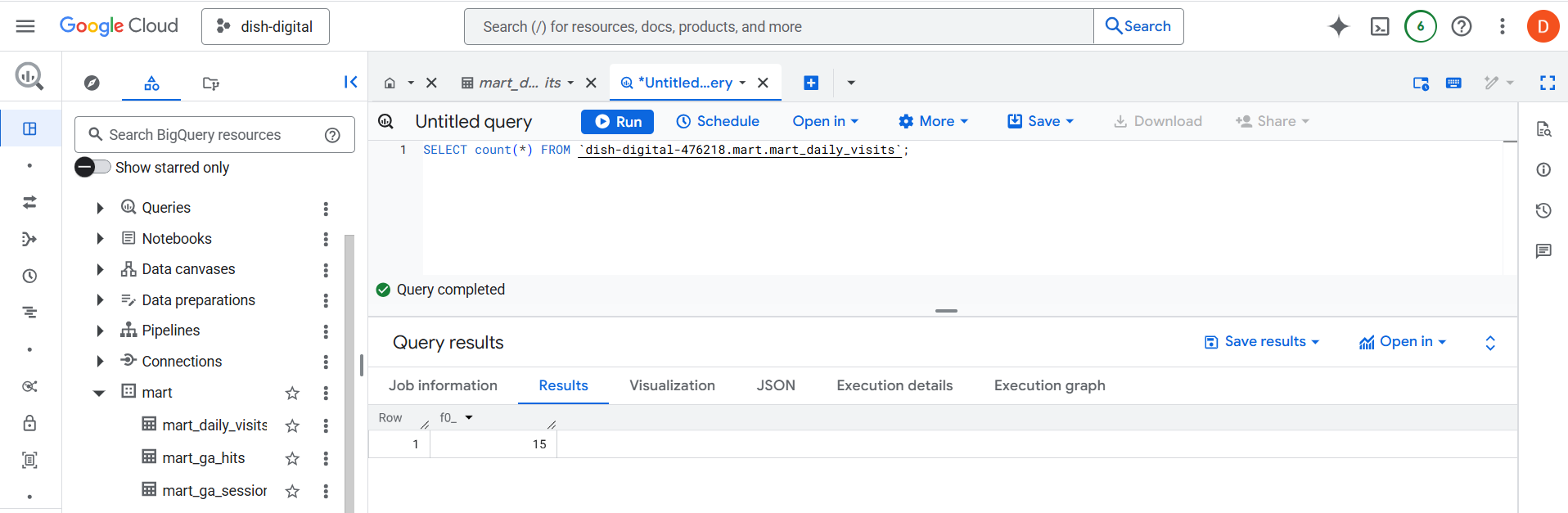


**Mart\_ga\_hits**

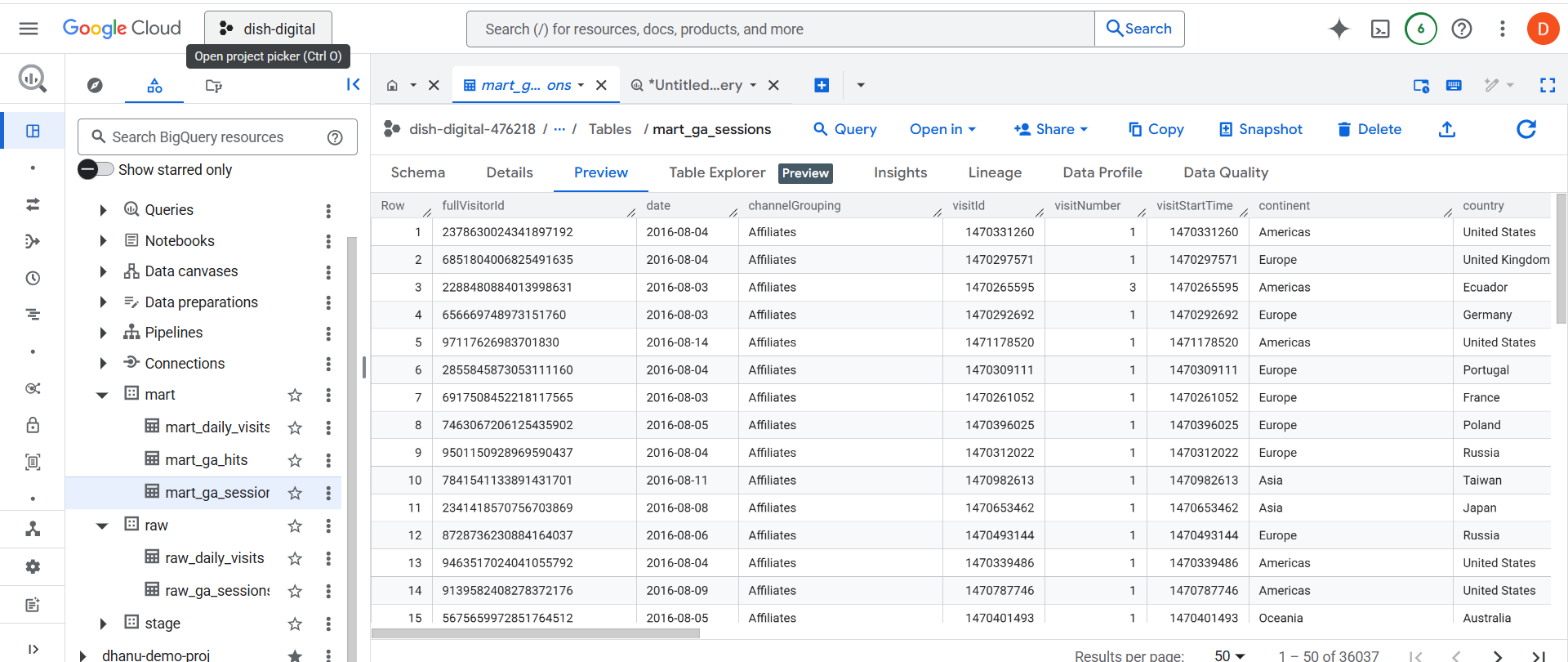


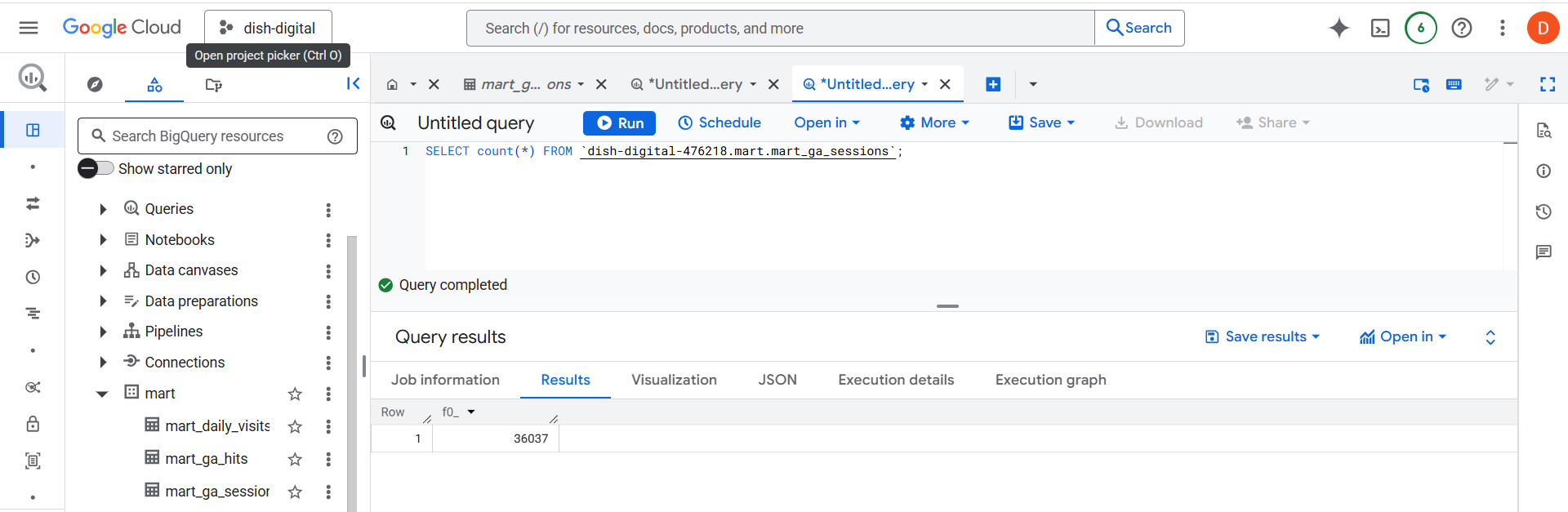
**Final data Available in Mart\_daily\_visits**



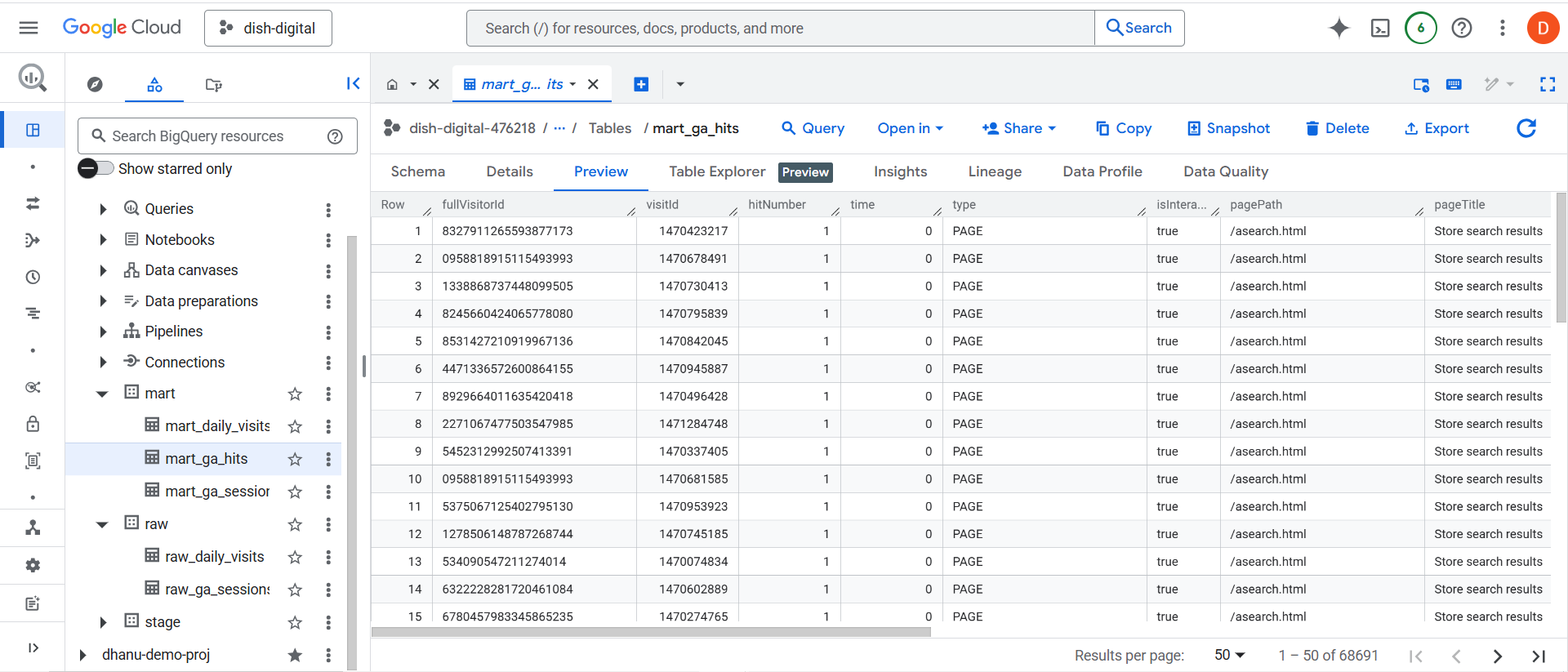


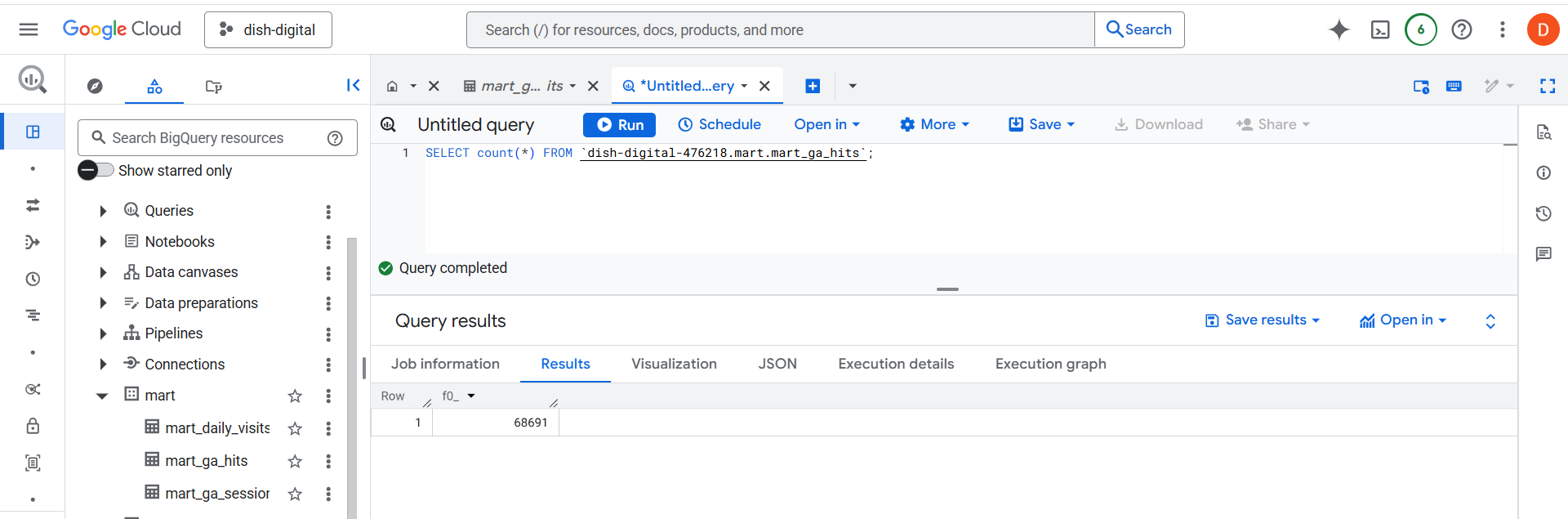
**Final data Available in Mart\_ga\_sessions**





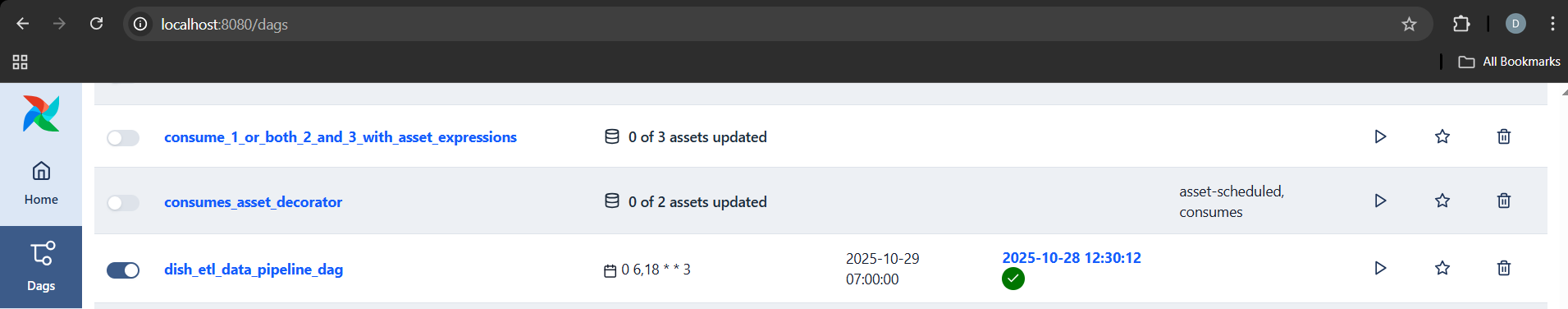
**Final data Available in Mart\_ga\_hits**



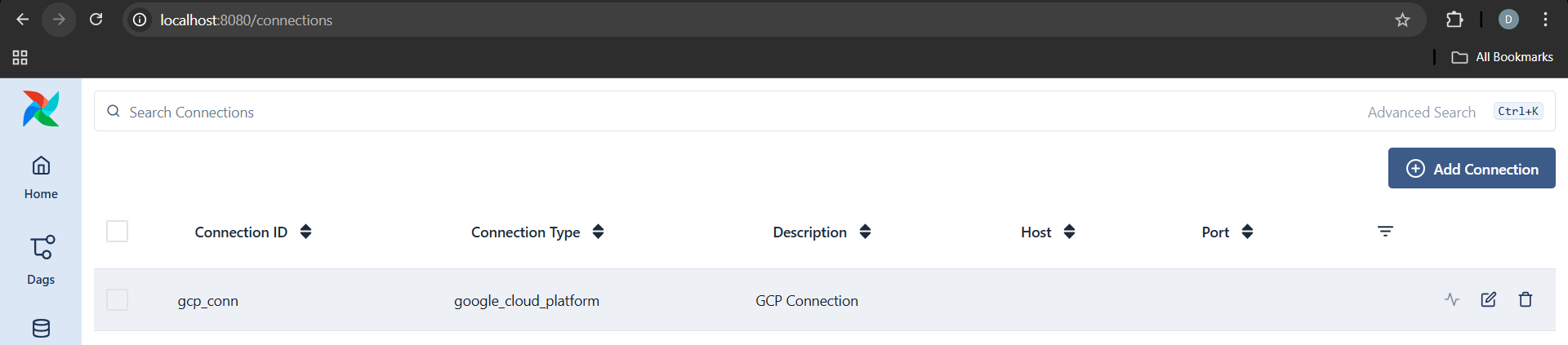


1. **Airflow testing**

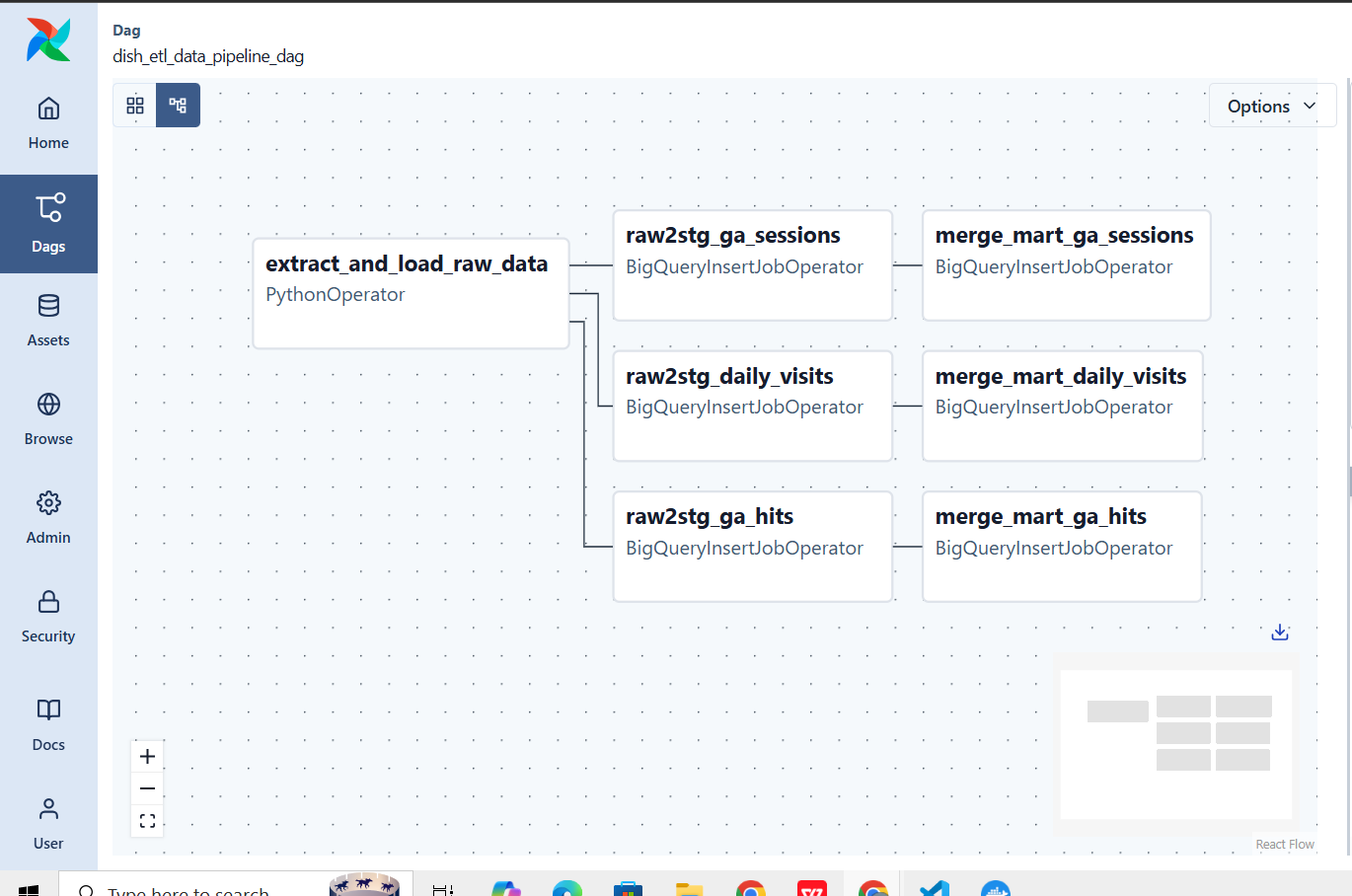
**Airflow dag script: etl\_google\_analytics\_dag.py**



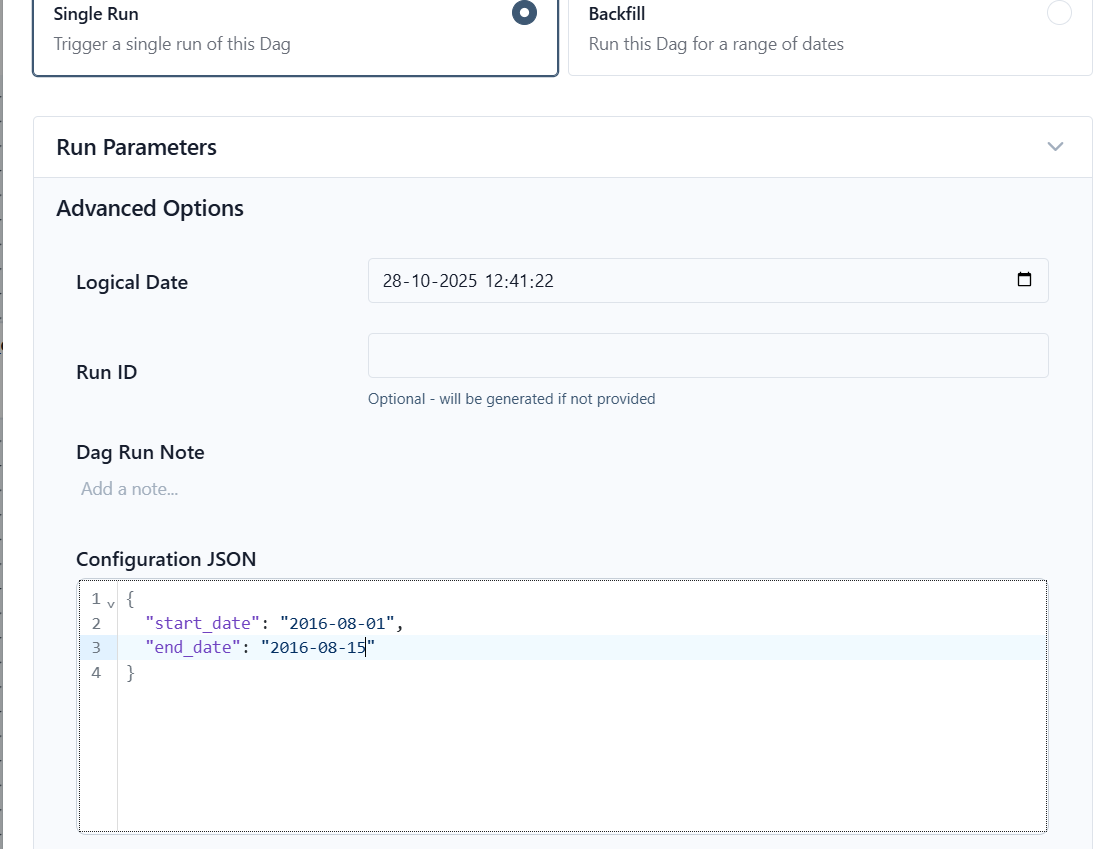
**Connection with GCP:**



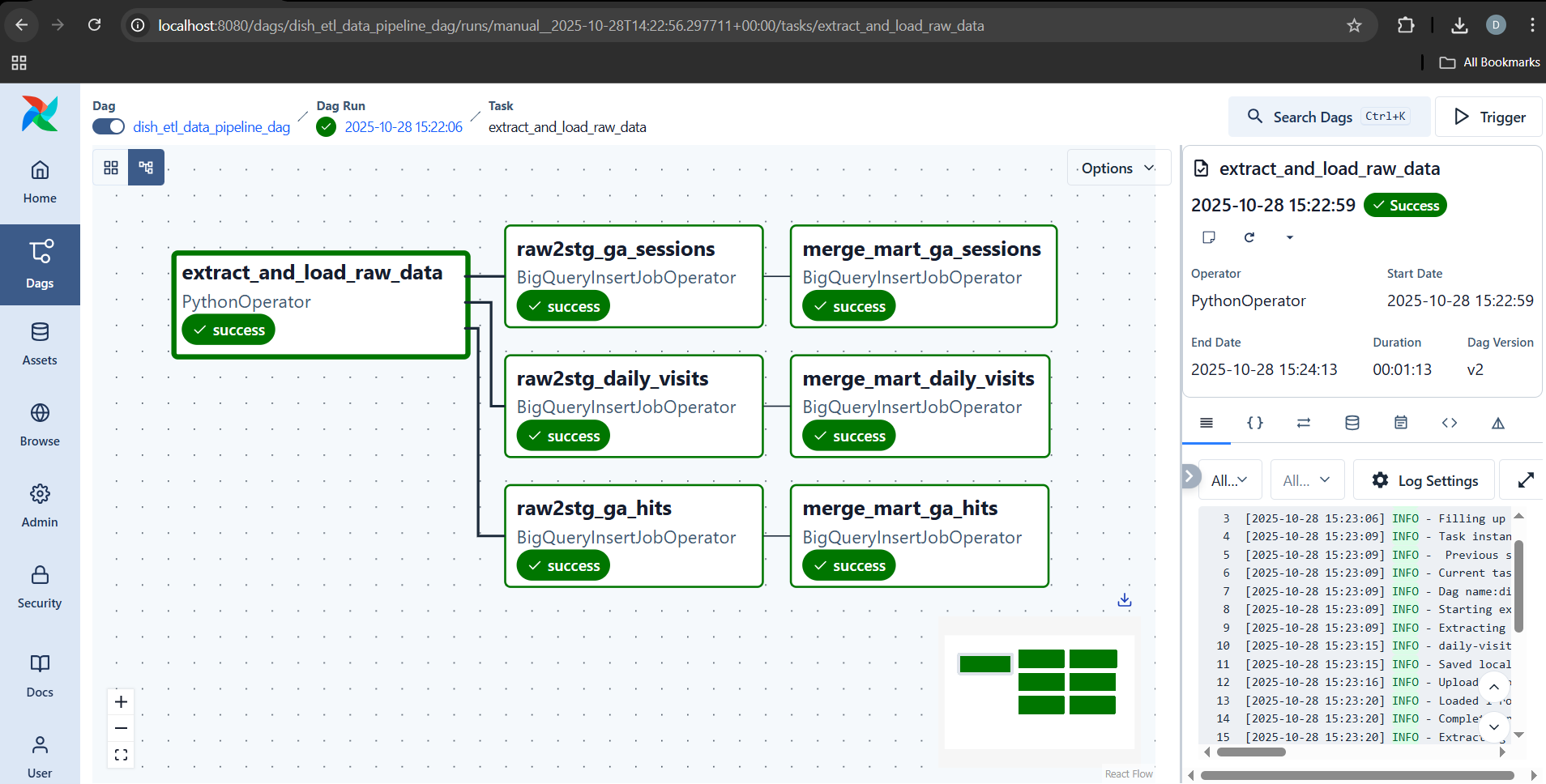
**Dag Lineage:**



**Triggering Dag by passing Custom date as parameters**

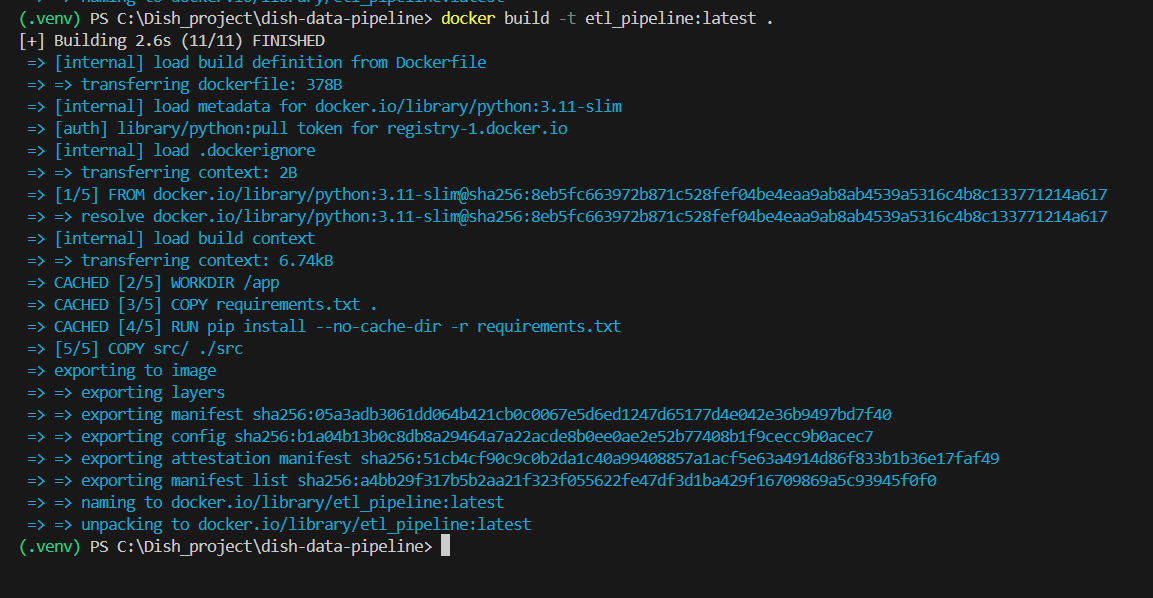


**Successful airflow run**

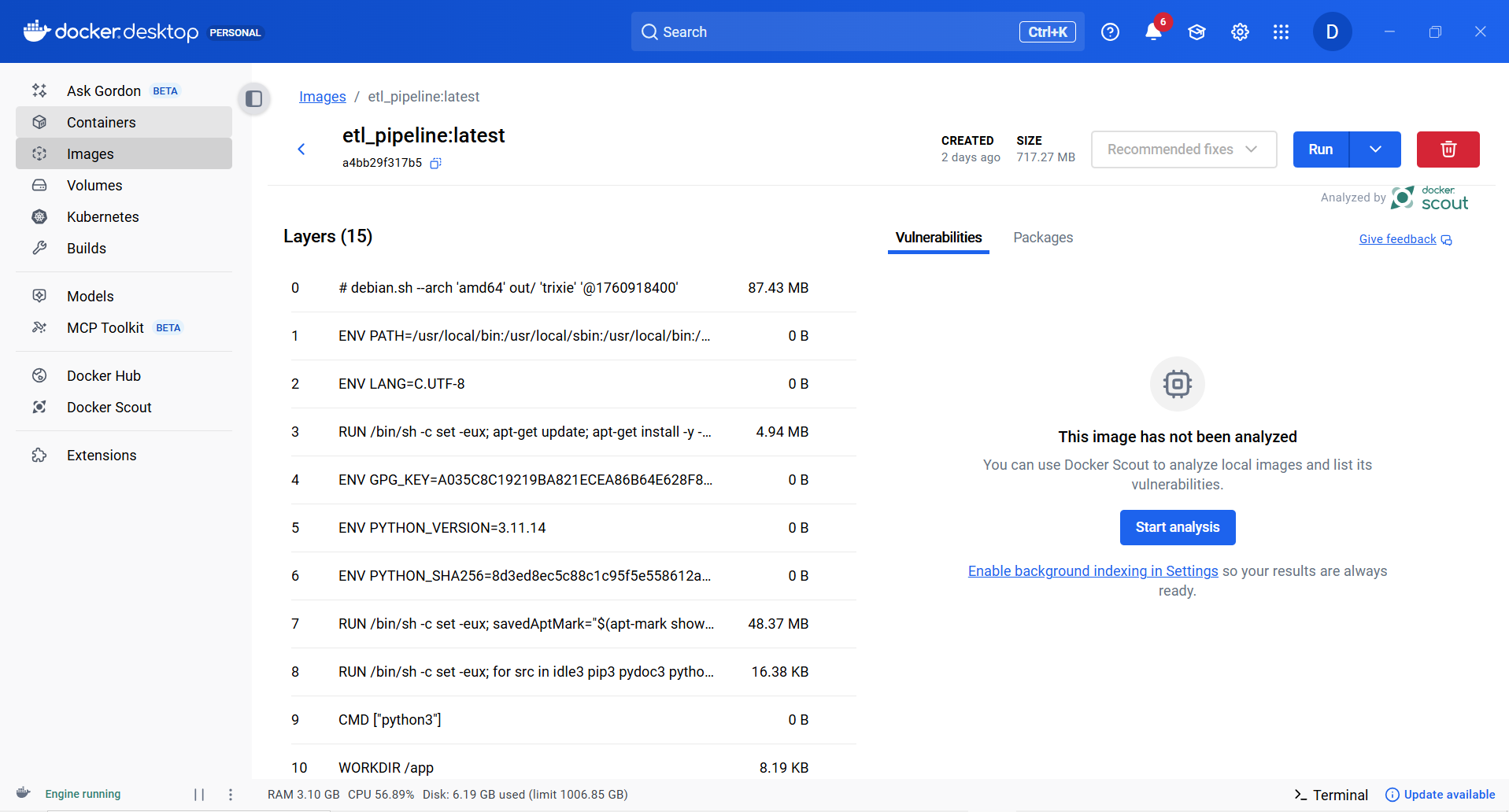


1. **Docker Containerization**

**Creating Docker container**



**Docker Container in Docker Desktop**



**Running the pipline script by setting appropriate entry point and by passing command line arguments:**

