Data Engineering Major

Google Cloud

Global Health Data

ELT with Airflow & BigQuery

Hafid GARHOUM

haf0g

Overview

Introduction

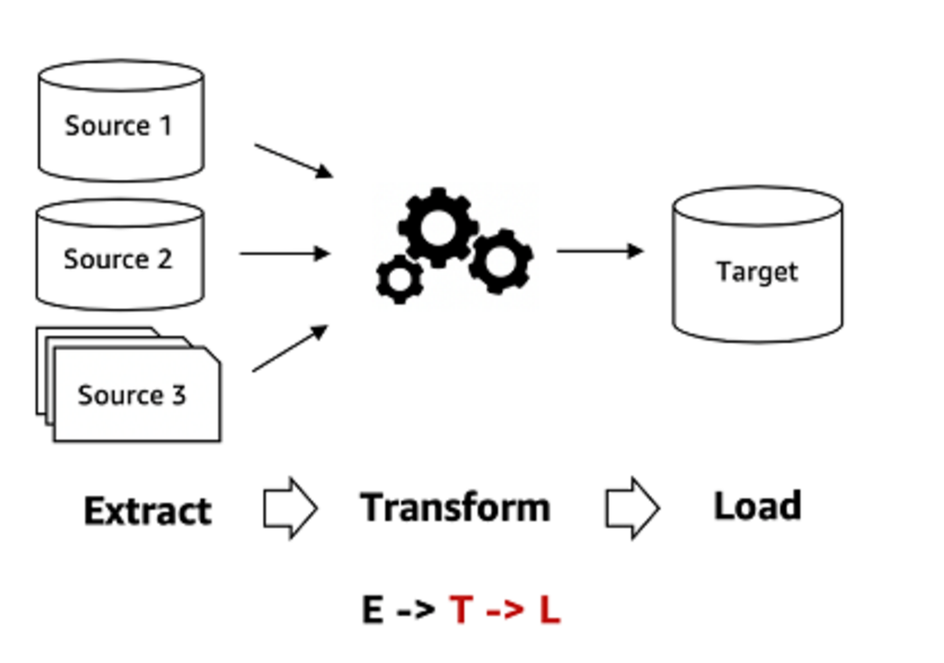
This presentation introduces a modern ELT (Extract, Load, Transform) data pipeline designed to run on Google Cloud Platform (GCP), orchestrated using Apache Airflow and fully deployed through Terraform. The project focuses on building a secure and scalable system capable of handling over one million records from a global medical dataset, ensuring efficient data processing, restricted access per country, and readiness for real-time analysis and reporting.

ETL vs ELT

a. ETL process

ETL consists of three distinct steps:

1. Extract raw data from various source systems.
2. Transform the data using an external processing server to clean, enrich, and reshape it.
3. Load the structured and ready-to-use data into a target database or data warehouse.



In ETL, data is transformed **before** loading to meet the target system's structure, ensuring only clean, ready-to-use data is moved.

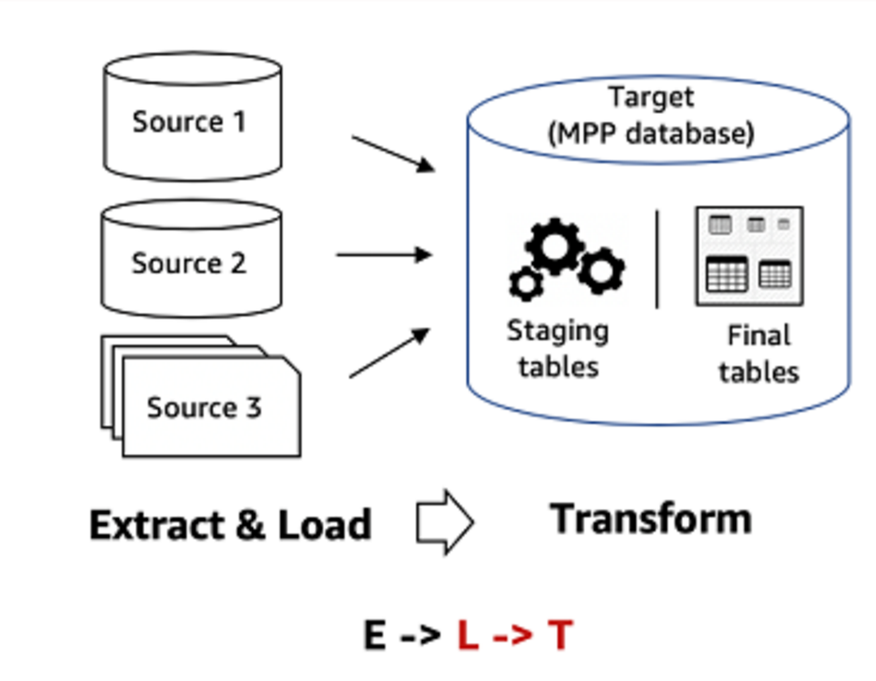
ETL vs ELT

b. ELT process

ELT, on the other hand, follows a slightly different sequence:

1. Extract raw data from various sources.
2. Load it in its raw state into a data warehouse or data lake.
3. Transform the data directly within the destination system as needed.

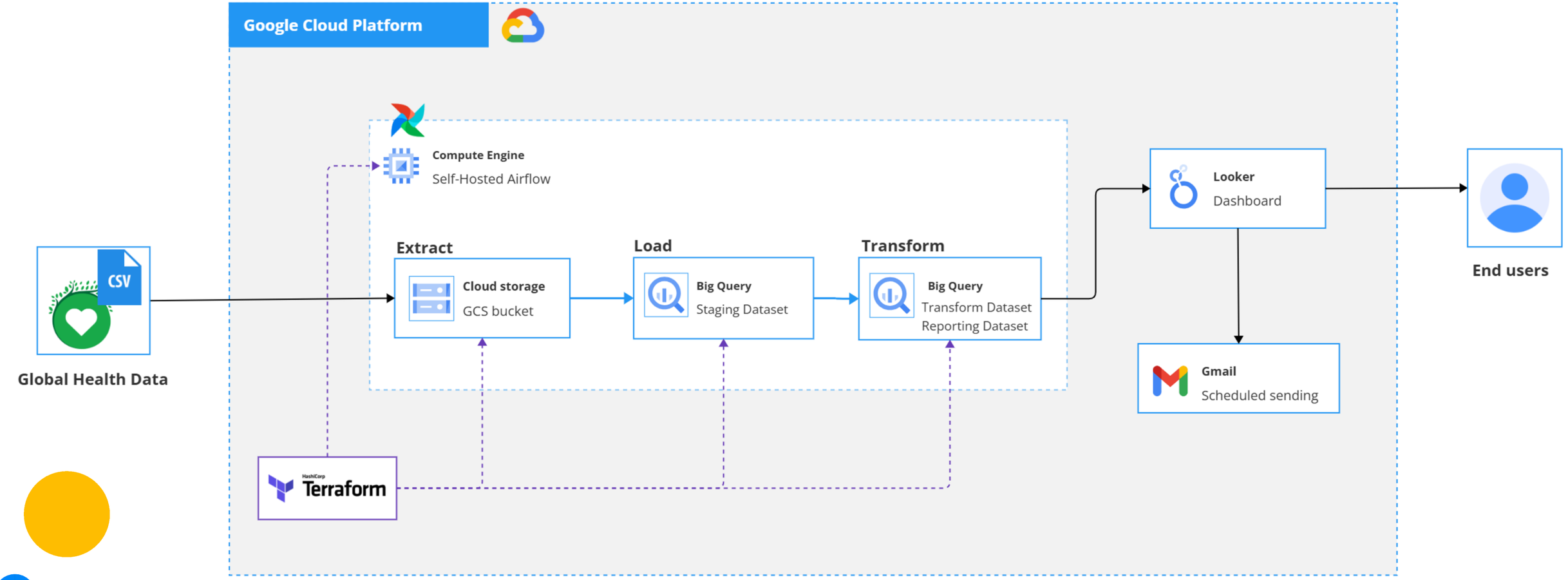
ELT is faster and better suited for modern cloud platforms like BigQuery, as it loads raw data directly into the warehouse, making it ideal for processing large data volumes efficiently.



ETL vs ELT

In our project, we adopted the **ELT** approach because it leverages **BigQuery**'s power to handle transformations efficiently **within the warehouse**, simplifies the pipeline, and scales better with large datasets, especially important when processing over 1 million medical records.

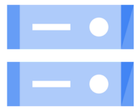
Project Architecture



Terraform

Terraform is an open-source Infrastructure as Code (IaC) tool that enables you to automate and version the provisioning of cloud resources in a reliable and repeatable manner.

01



Provision Cloud Storage

Created a secure GCS bucket to store the global health dataset before loading into BigQuery.

02

Deploy BigQuery Datasets

Automatically provisioned BigQuery datasets for each layer: staging, transform, and reporting, allowing structured ELT processing.

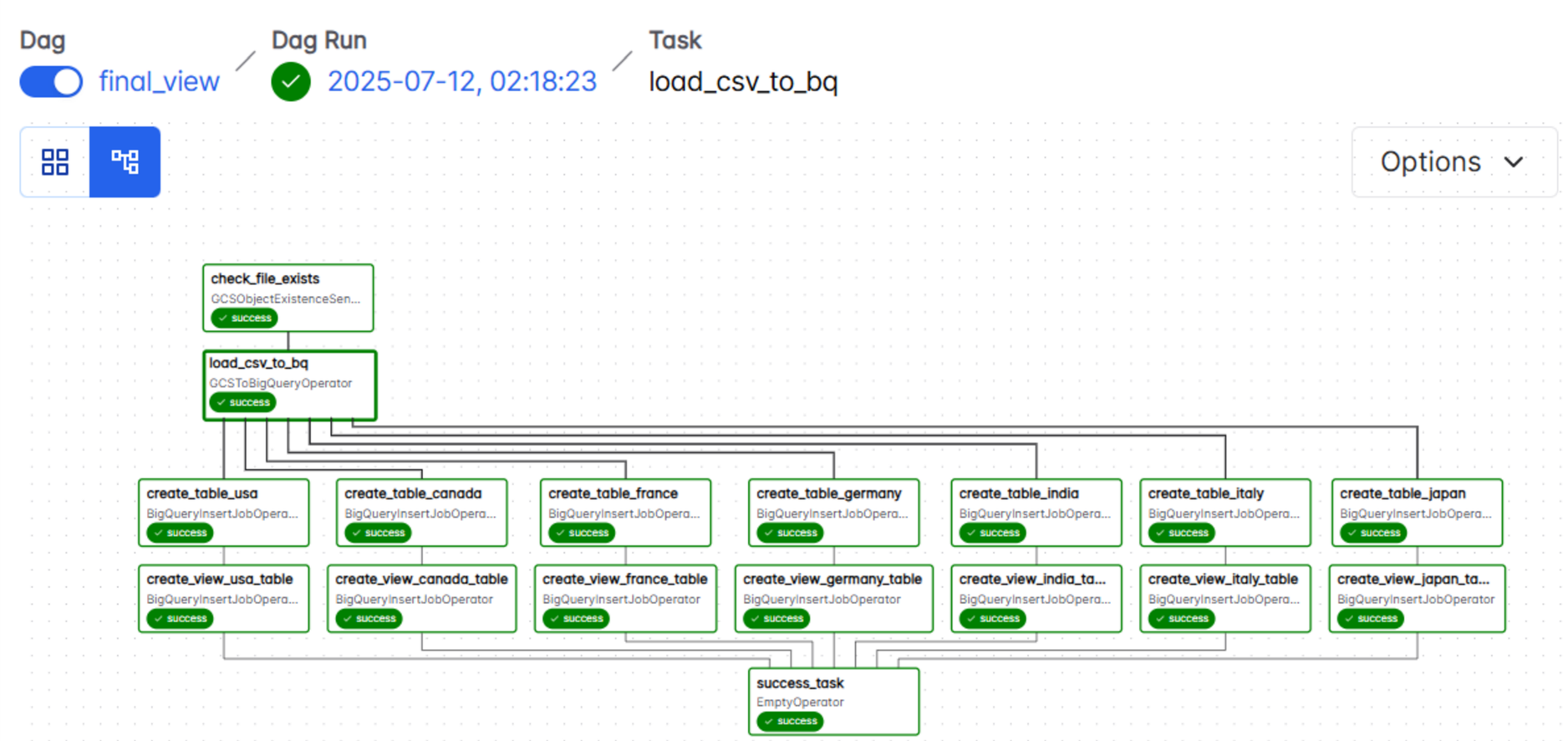
03

Launch Airflow VM

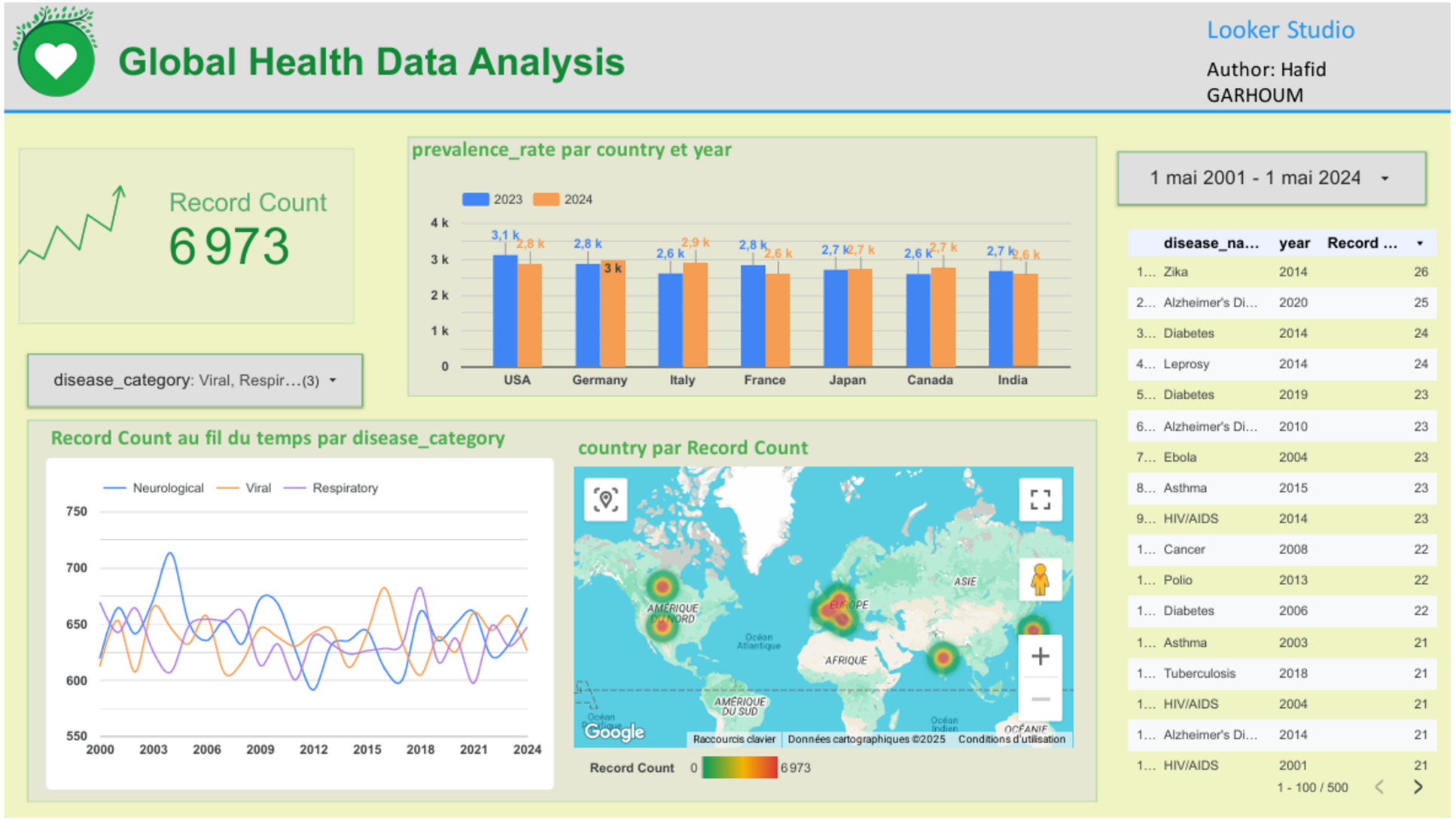
Provisioned a Compute Engine VM to host the self-managed Airflow instance, which orchestrates the ELT pipeline reliably.

Airflow Dags

Apache Airflow is a platform to programmatically author, schedule, and monitor workflows. In this project, it orchestrates the ELT pipeline tasks in a reliable and modular way.

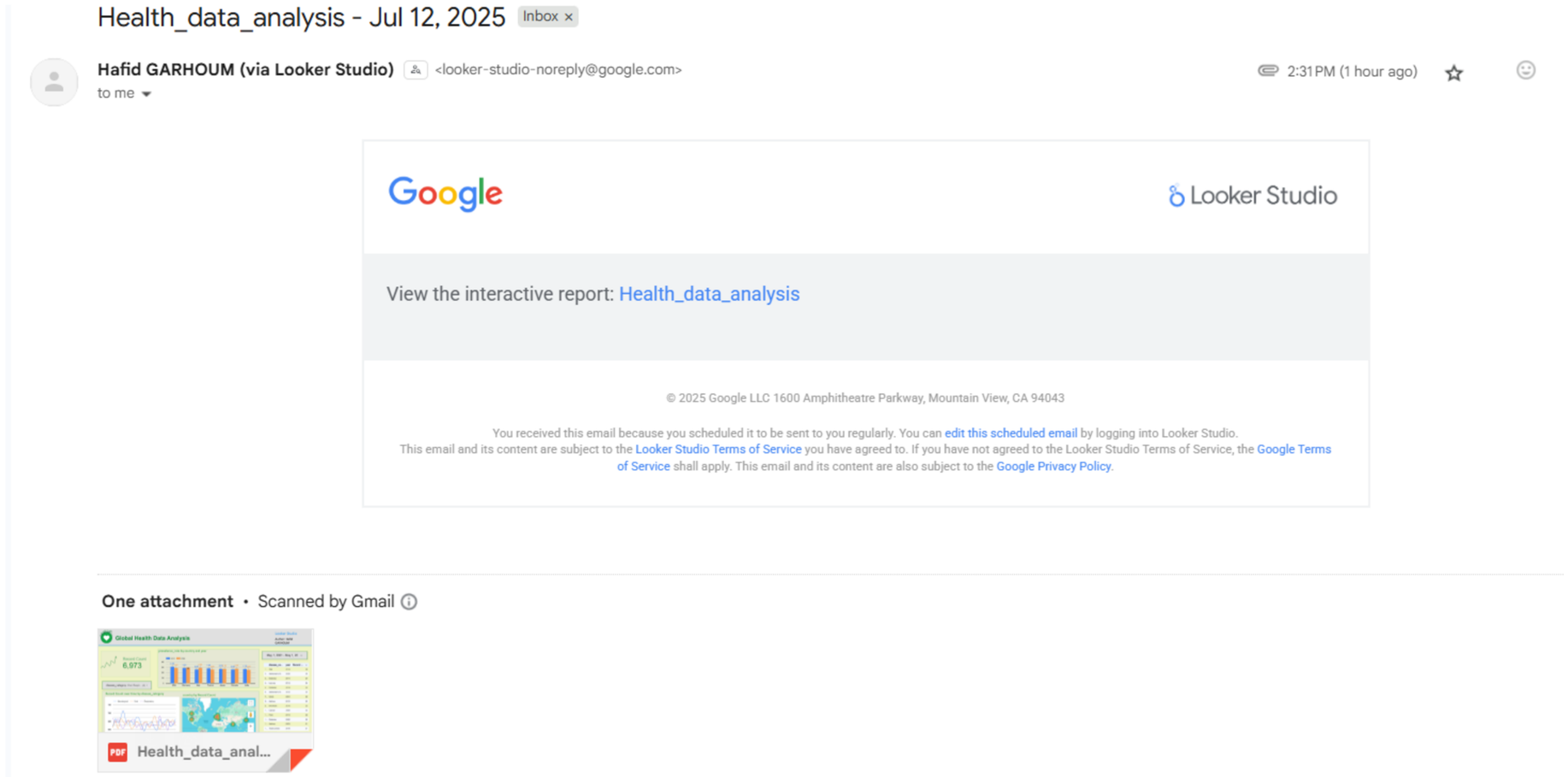


Looker Dashboard



Looker Dashboard

Daily reports sent to predefined recipients via Gmail API



Feel free to connect and explore more!

Dive into the full project code, DAGs, Terraform modules, and dashboards on my GitHub

Hafid GARHOUM

haf0g