

Data Engineering Major



Global Health Data

ELT with Airflow & BigQuery

Hafid GARHOU

haf0g



Overview

- Introduction 03
- ETL vs ELT 04
- Project Architecture 07
- Terraform 08
- Airflow Dags 09
- Looker Dashboard 10

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.

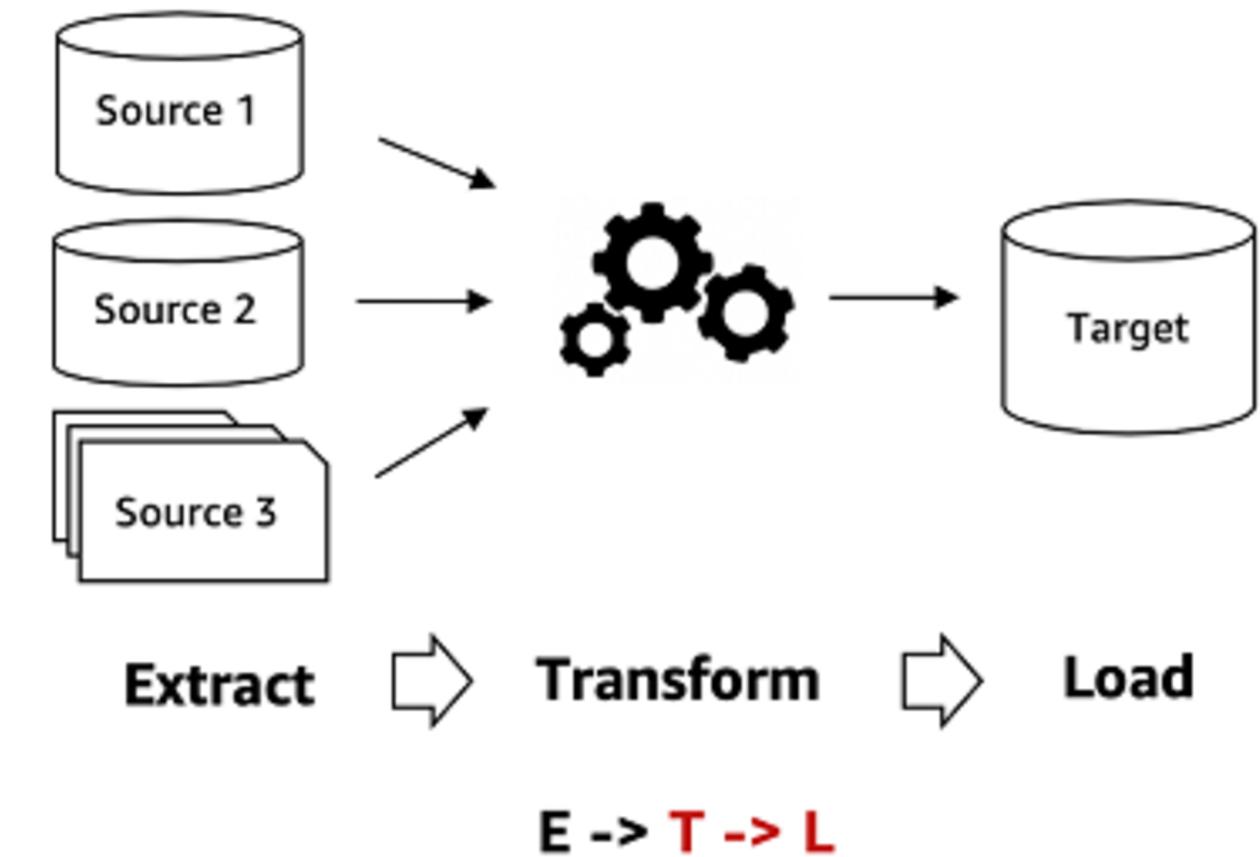


ETL vs ELT

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.

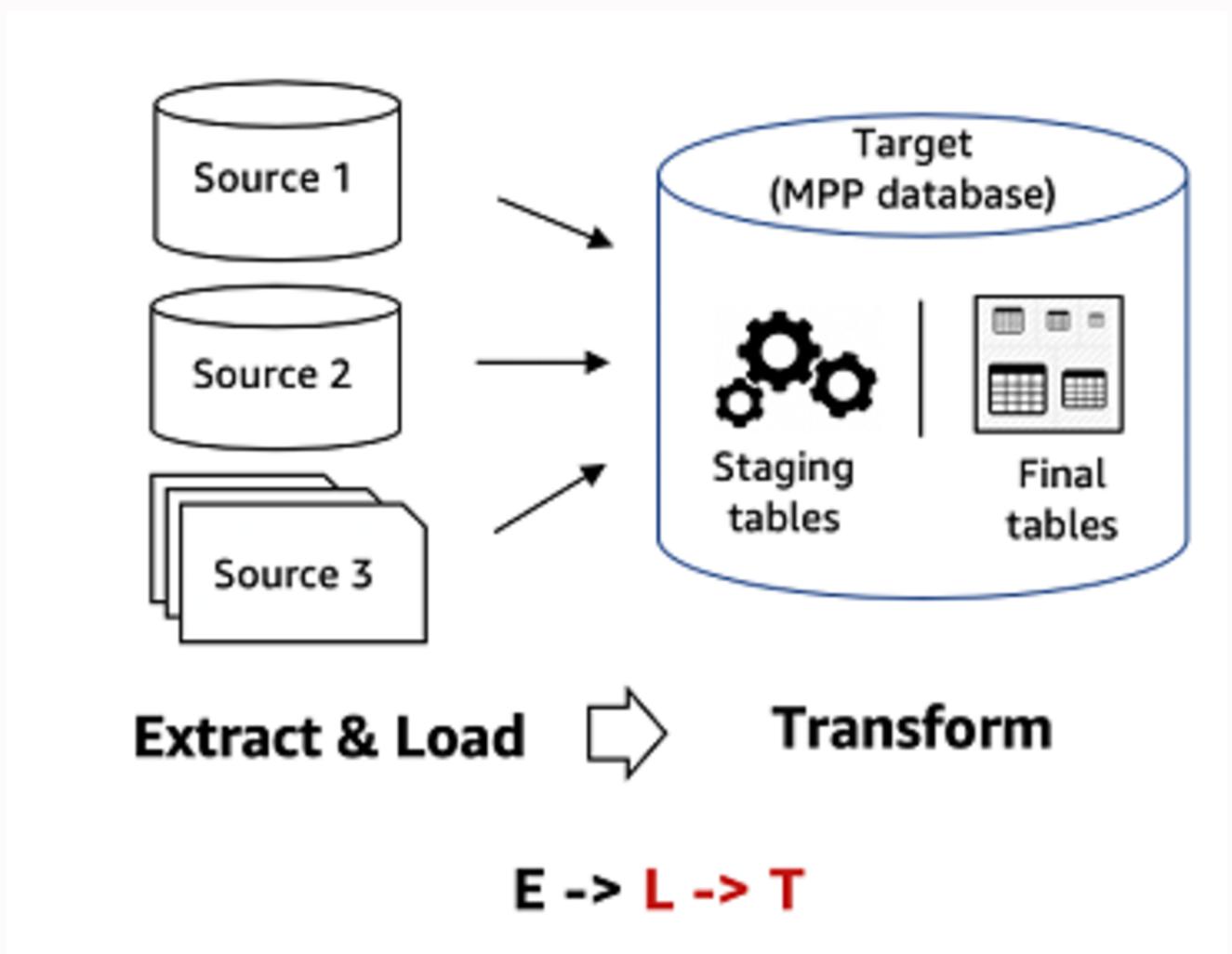
ETL vs ELT

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.

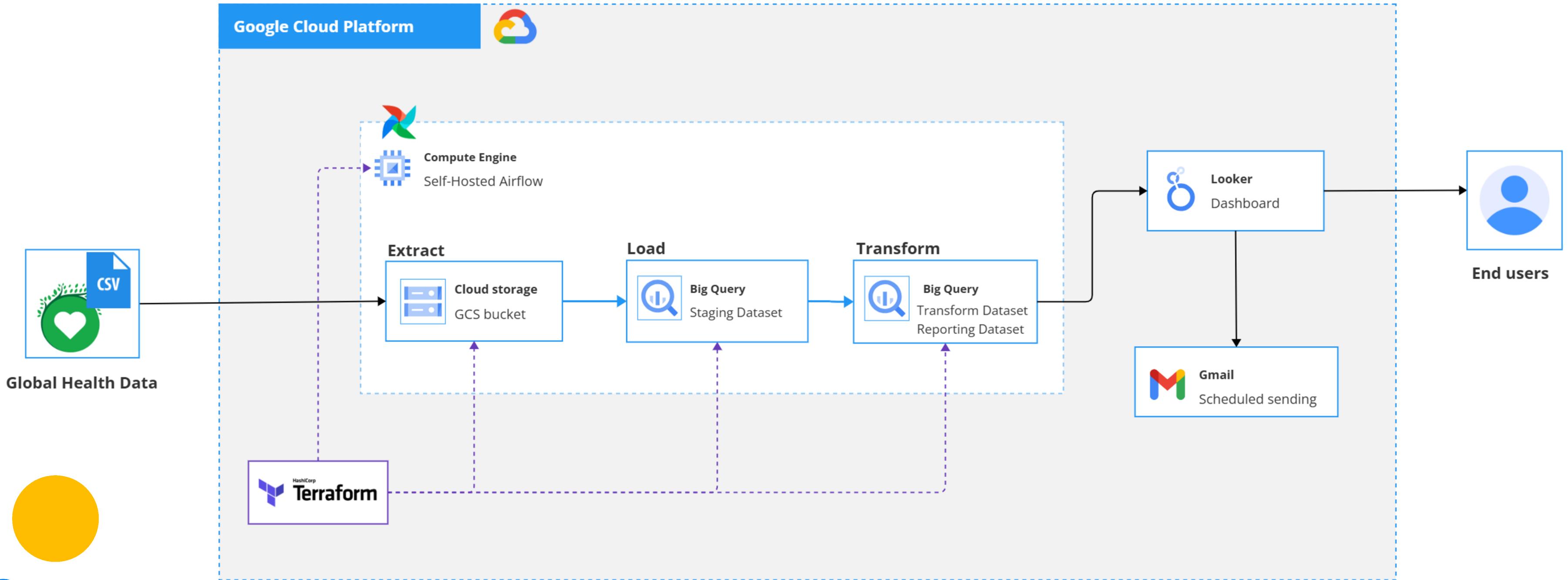
ELT is faster and better suited for modern cloud platforms like BigQuery, as it loads raw data directly into the warehouse, making it ideal for processing large data volumes efficiently.



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.

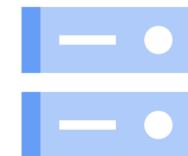
Project Architecture



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.

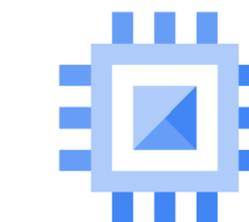
02



Deploy BigQuery Datasets

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

03



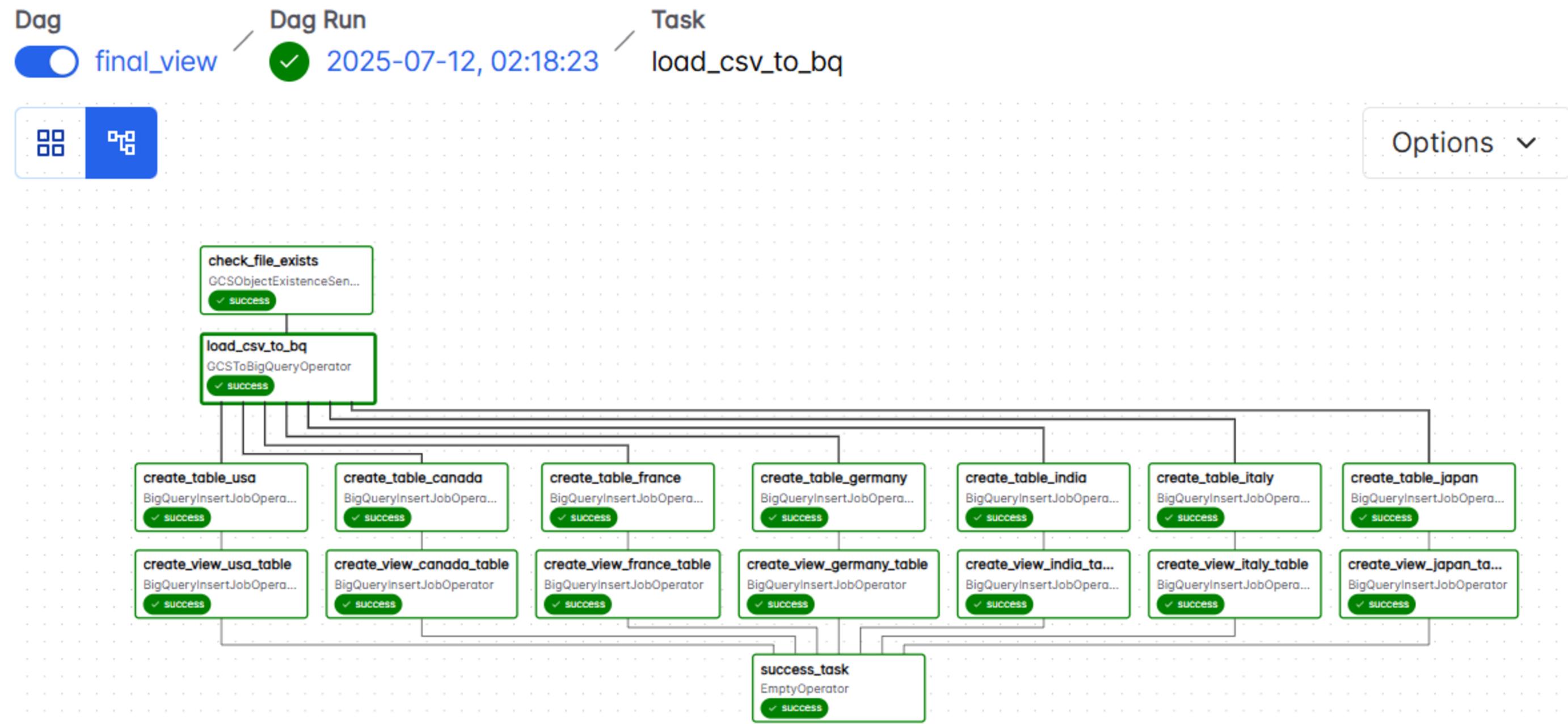
Launch Airflow VM

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

Global Health Data Analysis

Looker Studio
Author: Hafid GARHOUM

Record Count

6973

disease_category: Viral, Respir...(3)

prevalence_rate par country et year

country	2023	2024
USA	3,1 k	2,8 k
Germany	2,8 k	3 k
Italy	2,6 k	2,9 k
France	2,8 k	2,6 k
Japan	2,7 k	2,7 k
Canada	2,6 k	2,7 k
India	2,7 k	2,6 k

Record Count au fil du temps par disease_category

Legend: Neurological (blue), Viral (orange), Respiratory (purple)

country par Record Count

1 mai 2001 - 1 mai 2024

disease_na...	year	Record ...
Zika	2014	26
Alzheimer's Di...	2020	25
Diabetes	2014	24
Leprosy	2014	24
Diabetes	2019	23
Alzheimer's Di...	2010	23
Ebola	2004	23
Asthma	2015	23
HIV/AIDS	2014	23
Cancer	2008	22
Polio	2013	22
Diabetes	2006	22
Asthma	2003	21
Tuberculosis	2018	21
HIV/AIDS	2004	21
Alzheimer's Di...	2014	21
HIV/AIDS	2001	21

10

Looker Dashboard

Daily reports sent to predefined recipients via Gmail API

Health_data_analysis - Jul 12, 2025 Inbox ×

 Hafid GARHOUIM (via Looker Studio)  <looker-studio-noreply@google.com>
to me ▾

⌚ 2:31PM (1 hour ago)  

  Looker Studio

[View the interactive report: Health_data_analysis](#)

© 2025 Google LLC 1600 Amphitheatre Parkway, Mountain View, CA 94043

You received this email because you scheduled it to be sent to you regularly. You can [edit this scheduled email](#) by logging into Looker Studio.
This email and its content are subject to the [Looker Studio Terms of Service](#) you have agreed to. If you have not agreed to the Looker Studio Terms of Service, the [Google Terms of Service](#) shall apply. This email and its content are also subject to the [Google Privacy Policy](#).

One attachment • Scanned by Gmail ⓘ


PDF Health_data_anal...



Feel free to connect and explore more!

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



Hafid GARHOUM



haf0g