Chisom Nnamani

Data Engineering Project

Automating ELT Data Pipeline with Airflow: Load & Transform data to BigQuery

Extract, Load, & transform 1 Million Health Records

Date: 06-01-2025

Table of Contents

Chisom Nnamani

First of, a Recap

ASPECT	ETL	ELT
Transformation	Performed before loading	Performed after loading
Where it happens	External tools (e.g Dataflow, Dataprep)	Inside the data warehouse (e.g Bigquery, SQLServer)
Scalability	Limited by the ETL Tools Capabilities	Highly scalable with BigQuery processing power
Use Cases	Legcay systems, Data governance requiements	Cloud-native analytics, large datasets

Positioning the Project: An ELT Data Pipeline

Chisom Nnamani

Project Requirements

The Medical Research Team receives a global health statistics data file containing disease data for all countries.

Each country's Health Minister should have access only to their respective country's medical data. Additionally, they need the ability to analyze diseases for which no treatment or vaccination is currently available.

Challenges

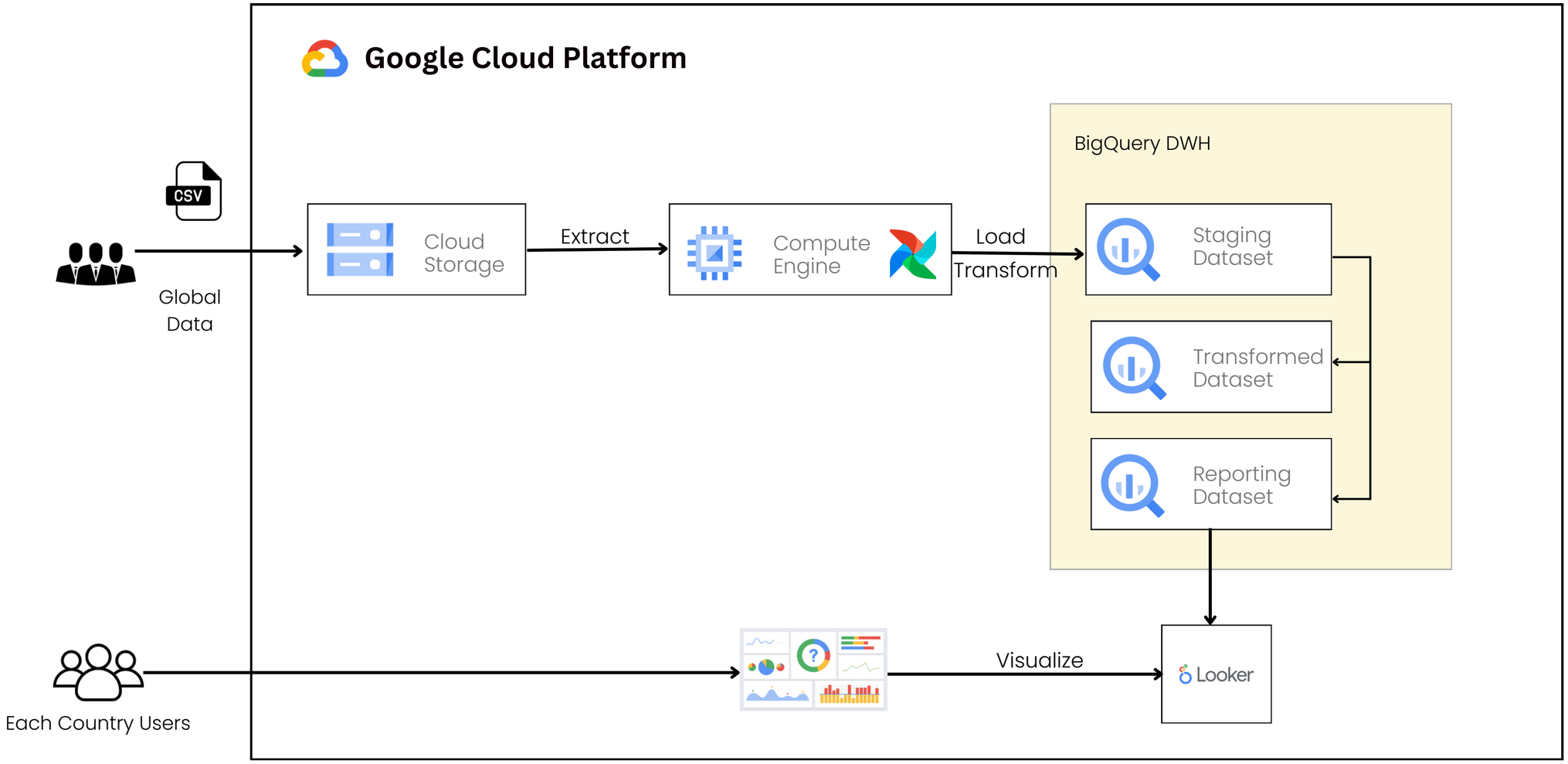
- The data is currently provided as a single file containing over 1 million records for all countries.
- due to the confidential nature of the data, sharing the entire file with everyone is not feasible.
- analyzing such a CSV file to extract meaningful insights is complex and inefficient.

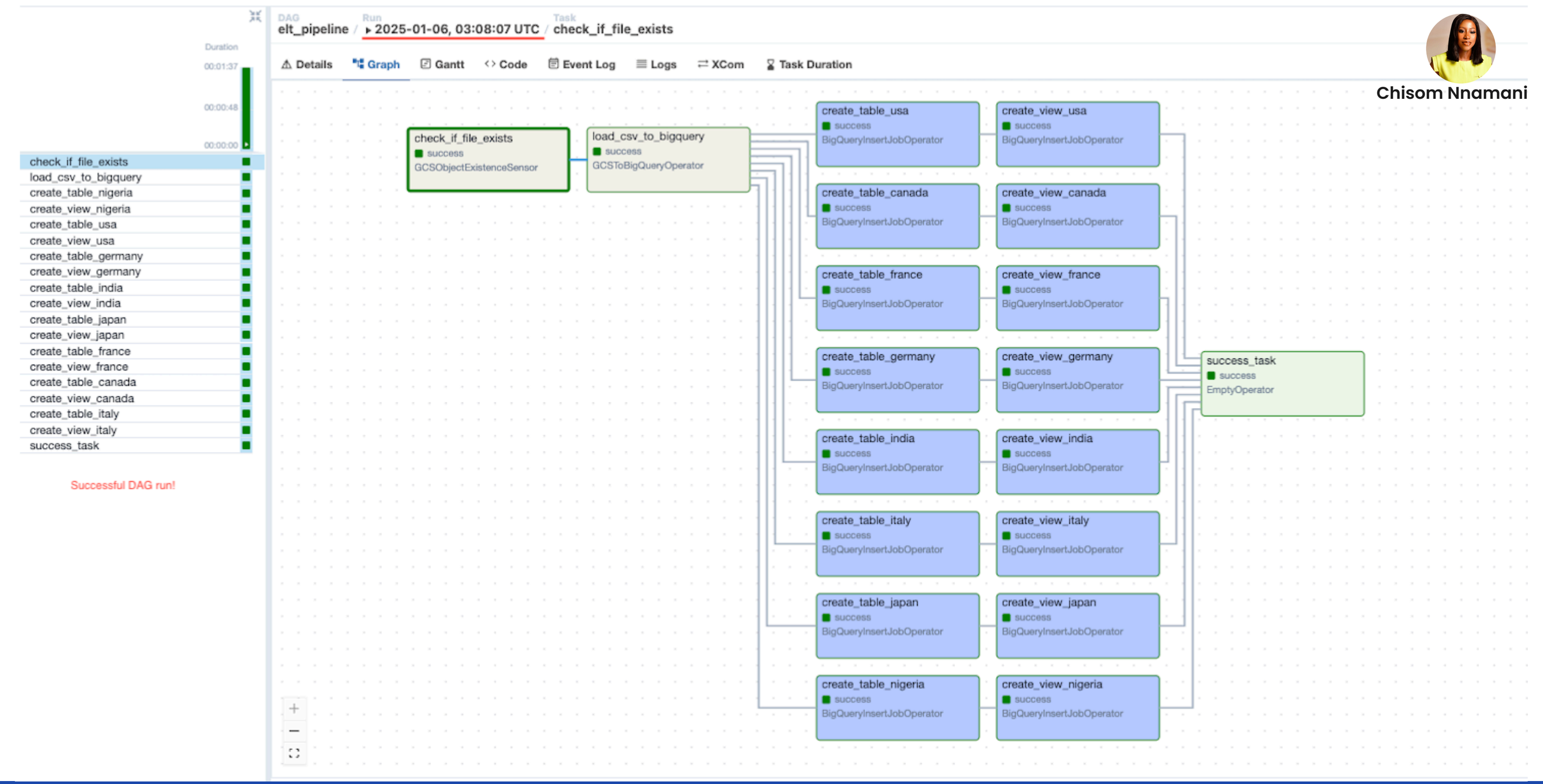
Objective

Develop a robust data analytics solution to securely manage and filter this data securely securely, ensuring restricted and enabling efficient analysis of diseases without available treatment or vaccination.

Data Pipeline Architecture

Chisom Nnamani





Chisom Nnamani

ELT Data Pipeline

Looker Studio Dashboard

Chisom Nnamani

Global Health Analytics Dashboard

Total Years	Total Record Count
2	2,095

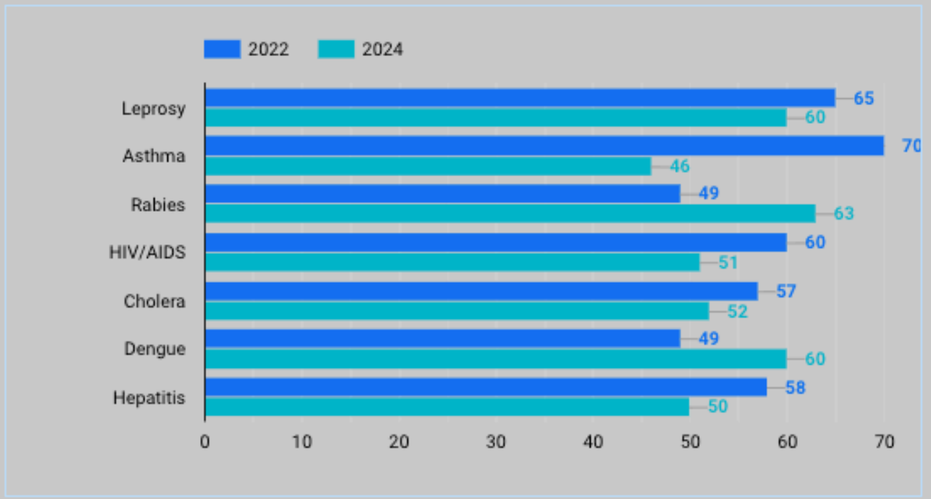
Year: 2022, 2024 (2)

Record Count by Disease Name and Year

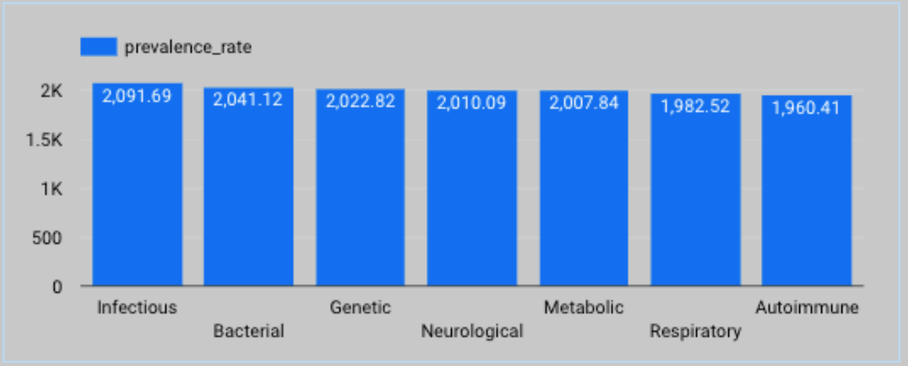
	Year	disease_name	Record Count
1.	2022	Asthma	70
2.	2022	Malaria	65
3.	2022	Leprosy	65
4.	2024	Rabies	63
5.	2024	Ebola	61
6.	2022	HIV/AIDS	60
7.	2024	Leprosy	60
8.	2024	Dengue	60
9.	2022	Zika	58
10.	2022	Polio	58
11.	2024	COVID-19	58
12.	2022	Hepatitis	58
13.	2022	Alzheimer's Disease	57
14.	2022	Cholera	57
15.	2024	Parkinson's Disease	57
16.	2024	Measles	55
17.	2022	Diabetes	54

1 - 40 / 40

Record Count by Disease Name



Prevalence Rate by Disease Category



Data Last Updated: 1/7/2025 10:57:10 AM

Chisom Nnamani

Conclusion

This end-to-end ELT data pipeline project on a global scale will help in:

- Healthcare Policy Analysis: Understanding which diseases are most prevalent and which countries require more investment in healthcare infrastructure.
- Epidemiological Studies: Studying the correlation between disease prevalence and socio-economic factors like income, education, and urbanization.
- Machine Learning Models: Training predictive models to forecast disease trends, mortality rates, and treatment effectiveness based on historical data.
- Global Health Research: Identifying regions that need targeted interventions or public health campaigns.

Building Scalable Data Architectures