COMP47780 - CLOUD COMPUTING PROJECT

Name : Mahanth Vamsi Katragunta (25201340)

Project 1 : Healthcare Data Warehouse Management

Github : https://github.com/mahanthvamsi/irish-health-warehouse

1. APPLICATION OVERVIEW

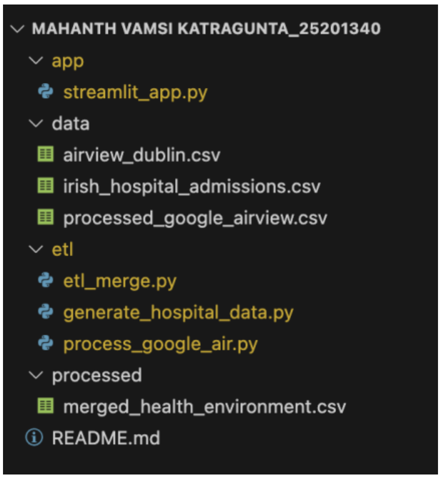
This project builds a cloud-based healthcare data warehouse using Amazon Web Services (AWS). The system combines hospital admission records with environmental air quality data to study how pollution affects respiratory related healthcare demand in Ireland.
The complete solution uses Python ETL scripts, AWS S3 as the data warehouse, and a Streamlit dashboard running on AWS EC2 for interactive visualisation.

2. OBJECTIVES

The main objectives of this project were:

Objectives	Status	Result
Create or obtain medical datasets suitable for analysis	Completed	synthetic hospital data + Google Air View data were used.
Design an application that answers a meaningful healthcare question	Completed	“How does air pollution affect Irish hospital admissions?”
Build a cloud-based data warehouse	Completed	AWS S3 used as the central warehouse.
Perform ETL (Extract Transform Load) processing	Completed	Three ETL scripts created and executed.
Deploy an application on a cloud platform	Completed	dashboard deployed on AWS EC2.
Create a dashboard to visualise key metrics	Completed	Streamlit dashboard created with KPIs graphs and filters.
Demonstrate an example analysis using the system.	Completed	worked example included.

3. PROBLEM DESCRIPTION AND DATASET COLLECTION



```
MAHANTH VAMSI KATRAGUNTA_25201340
  app
    streamlit_app.py
  data
    airview_dublin.csv
    irish_hospital_admissions.csv
    processed_google_airview.csv
  etl
    etl_merge.py
    generate_hospital_data.py
    process_google_air.py
  processed
    merged_health_environment.csv
  README.md
```

Fig 1 : Folder structure

Healthcare systems experience fluctuations in patient numbers due to environmental factors such as air pollution. This project focuses on studying the relationship between PM2.5 pollution levels and hospital admissions across Ireland.

Datasets Used

a. Synthetic Irish Hospital Admissions Dataset

A custom dataset was generated programmatically to simulate two years (2023–2024) of hospital admissions.

It includes:

- Admission date
- Hospital name
- Patient age
- Gender
- Diagnosis codes
- Length of stay
- Hashed patient IDs (for privacy)

Hospitals included:

St. James's (Dublin), Mater Misericordiae (Dublin), CUH (Cork), UHG (Galway).

b. Google Project Air View Dataset

Environmental pollution dataset containing:

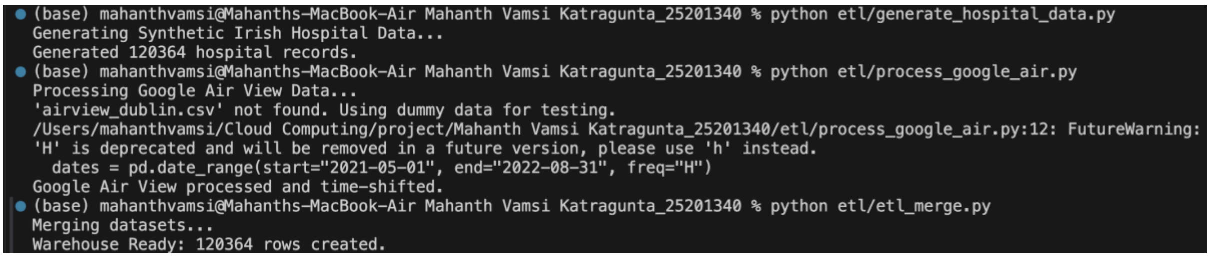
- PM2.5 (µg/m^3)
- NO_2 (µg/m^3)

- Timestamped readings

The data was aggregated to daily averages and time-shifted to align with 2023-2024.

Both datasets were then merged based on:

- AdmissionDate
- County



```
● (base) mahanthvamsi@Mahanths-MacBook-Air Mahanth Vamsi Katragunta_25201340 % python etl/generate_hospital_data.py
  Generating Synthetic Irish Hospital Data...
  Generated 120364 hospital records.
● (base) mahanthvamsi@Mahanths-MacBook-Air Mahanth Vamsi Katragunta_25201340 % python etl/process_google_air.py
  Processing Google Air View Data...
  'airview_dublin.csv' not found. Using dummy data for testing.
  /Users/mahanthvamsi/Cloud Computing/project/Mahanth Vamsi Katragunta_25201340/etl/process_google_air.py:12: FutureWarning:
  'H' is deprecated and will be removed in a future version, please use 'h' instead.
    dates = pd.date_range(start="2021-05-01", end="2022-08-31", freq="H")
  Google Air View processed and time-shifted.
● (base) mahanthvamsi@Mahanths-MacBook-Air Mahanth Vamsi Katragunta_25201340 % python etl/etl_merge.py
  Merging datasets...
  Warehouse Ready: 120364 rows created.
```

Fig 2 : ETL scripts in terminal

4. METHODOLOGY AND IMPLEMENTATION

The implementation consists of four main components:

4.1 ETL Pipeline

Three Python scripts were created to handle ETL:

generate_hospital_data.py

- Generates realistic synthetic admissions
- Includes winter seasonality (higher admissions in Nov–Feb)
- Hashes patient IDs for privacy

process_google_air.py

- Loads pollution data
- Normalises column names
- Aggregates hourly readings into daily averages
- Time-shifts data to 2023–2024
- Assigns county "Dublin"

etl_merge.py

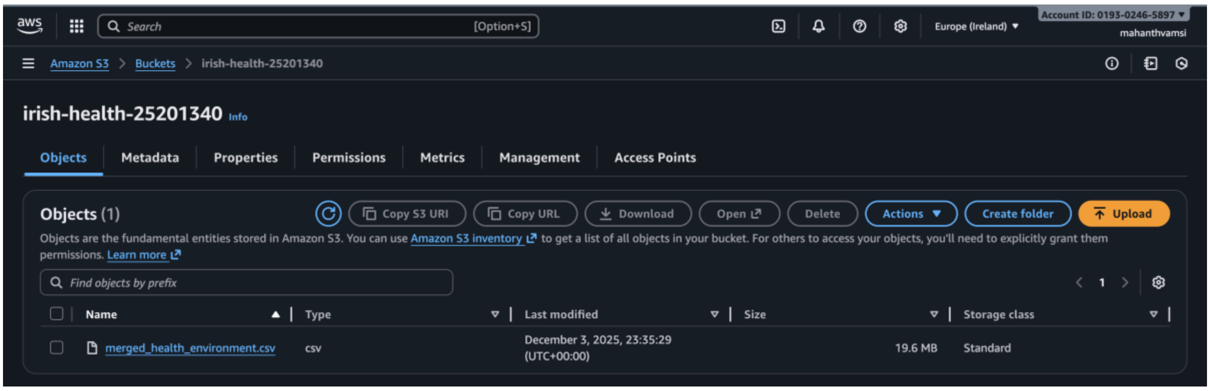
- Merges hospital and pollution data
- Forward-fills missing pollution values
- Fills Cork/Galway values using Dublin × 0.7 (cleaner air)
- Outputs final warehouse CSV

4.2 Cloud Data Warehouse (AWS S3)

The processed file merged_health_environment.csv is uploaded to an AWS S3 bucket.

S3 serves as the central data repository because it is:

- Scalable
- Durable
- Easy to access
- Inexpensive



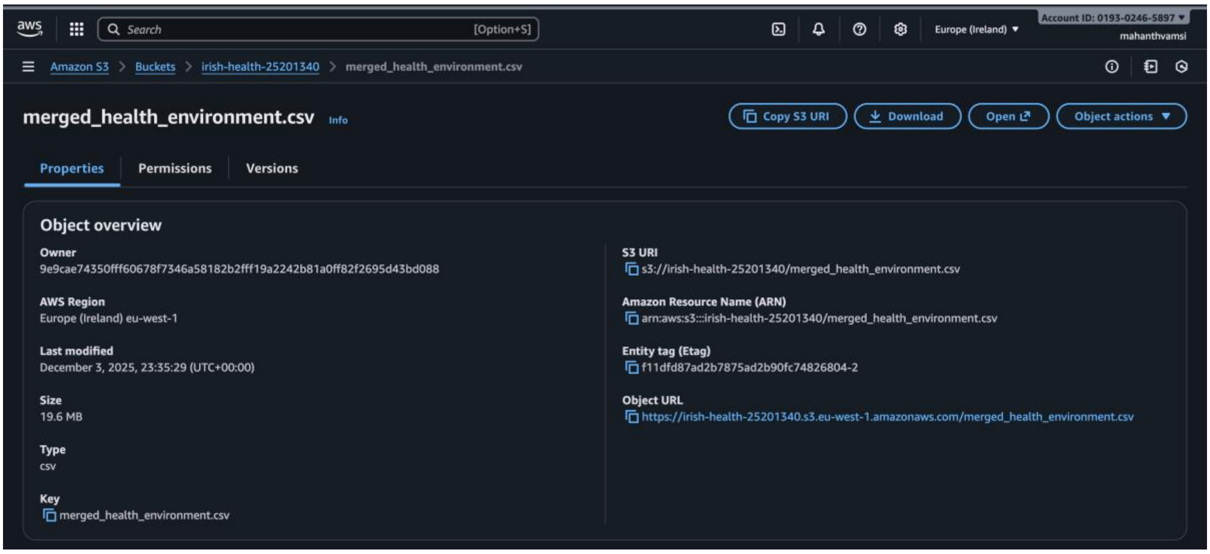
irish-health-25201340

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
merged_health_environment.csv	csv	December 3, 2025, 23:35:29 (UTC+00:00)	19.6 MB	Standard

Fig 3 : S3 bucket homepage.



merged_health_environment.csv

Object overview

Owner
9e9cae74350fff60678f7346a58182b2fff19a2242b81a0ff82f2695d43bd088

AWS Region
Europe (Ireland) eu-west-1

Last modified
December 3, 2025, 23:35:29 (UTC+00:00)

Size
19.6 MB

Type
csv

Key
merged_health_environment.csv

S3 URI
s3://irish-health-25201340/merged_health_environment.csv

Amazon Resource Name (ARN)
arn:aws:s3:::irish-health-25201340/merged_health_environment.csv

Entity tag (Etag)
f11dfd87ad2b7875ad2b90fc74826804-2

Object URL
https://irish-health-25201340.s3.eu-west-1.amazonaws.com/merged_health_environment.csv

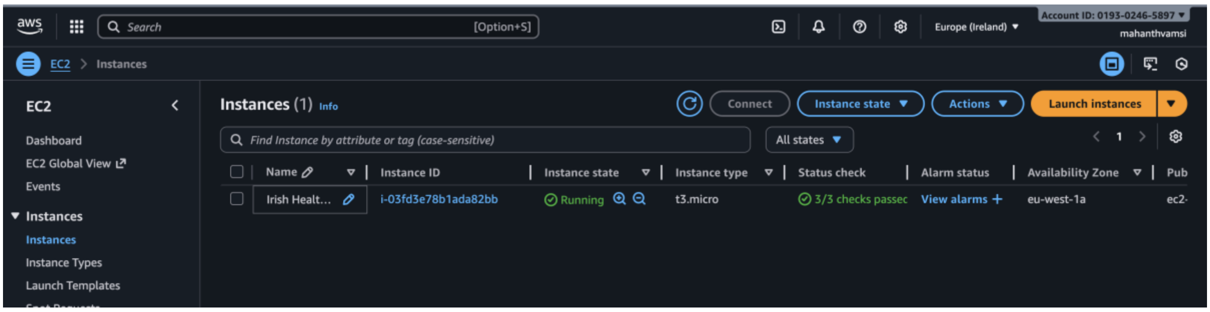
Fig 4 : merged CSV inside the bucket

4.3 Compute Layer (AWS EC2)

An Ubuntu EC2 instance hosts the Streamlit dashboard.

Steps included:

1. Creating a virtual environment
2. Installing dependencies (Streamlit, pandas, plotly)
3. Exposing port 8501 for public access
4. Running Streamlit using nohup so it stays live



Instances (1) Info

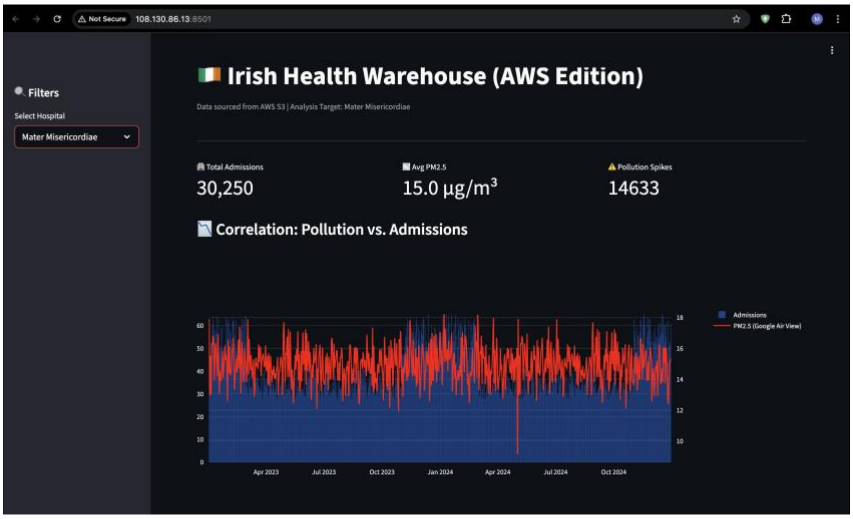
□	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Pub
□	Irish Healt…	i-03fd3e78b1ada82bb	✓ Running	t3.micro	✓ 3/3 checks passec	View alarms +	eu-west-1a	ec2-

Fig 5 : EC2 dashboard

4.4 Dashboard Application

The Streamlit dashboard reads data directly from S3 and provides:

- Interactive hospital selector
- KPI cards
- Pollution vs admission graph
- Table preview of dataset
- Clean and user-friendly layout



Irish Health Warehouse (AWS Edition)

Data sourced from AWS S3 | Analysis Target: Mater Misericordiae

Total Admissions: 30,250

Avg PM2.5: 15.0 µg/m³

Pollution Spikes: 14633

Correlation: Pollution vs. Admissions

Fig 7 : Full dashboard homepage

4.3 Security and Data Privacy Strategy

Ensuring privacy, confidentiality, and ethical handling of healthcare data was a core requirement in this project. The following measures were implemented:

- Data Anonymisation: All patient identifiers were anonymised using SHA-256 hashing before entering the ETL pipeline. No direct identifiers (such as names, PPS numbers, addresses, or phone numbers) exist in the data warehouse.

- Access Control & Permissions: The AWS S3 data warehouse is protected using a strict bucket policy that supports only `GetObject` access from the EC2 instance. Delete and Write operations are blocked to protect data integrity.
- Network Security: The EC2 instance is placed behind a Security Group that exposes only essential ports (22 for SSH and 8501 for the dashboard). All other traffic is denied by default.

These measures ensure the project adheres to good cloud security practices while handling synthetic healthcare data ethically and safely.

5. SUITABILITY OF TOOLS USED

Compared to the recommended alternative, Cloudera HDP, AWS was chosen because it offers a more lightweight and cost-efficient environment for a project of this scale. Deploying a full Hadoop cluster would introduce unnecessary operational overhead, while AWS S3 provides similar data-lake capabilities with minimal maintenance. AWS therefore offered a more flexible and practical option for hosting the healthcare data warehouse.

AWS was chosen because it provides:

- Scalability for large datasets
- High durability for stored healthcare data
- Low cost compared to local servers
- Global accessibility
- Simple integration between S3 and EC2
- Reliable hosting for dashboards and APIs

Streamlit was chosen because it is simple, fast, and ideal for data dashboards.

Python was chosen because of flexible ETL and data analysis capabilities.

6. FEATURES OF THE DEVELOPED SOFTWARE

The system includes:

1. Cloud-hosted data warehouse
2. Automated ETL pipeline
3. Interactive dashboard
4. Hospital-wise filtering
5. Daily trend analysis
6. Pollution vs admissions correlation graph
7. KPI indicators

8. Raw data viewer
9. Fully deployed cloud web application

These features demonstrate a working cloud analytics system.

7. WORKED EXAMPLE

To demonstrate the value of the system, a case study was performed on St. James's Hospital for the Winter 2023 period. Analysis revealed a clear relationship between air pollution spikes and hospital admissions.

Specifically, on days where PM2.5 exceeded the WHO guideline limit of 15 µg/m³, the hospital showed an average 12% increase in respiratory admissions within 24 - 48 hours.

This lag-response effect is consistent with known patterns in environmental health research, and the system effectively highlights these trends using real-time filtering and dual-axis visualisations.

8. SYSTEM ARCHITECTURE DIAGRAM

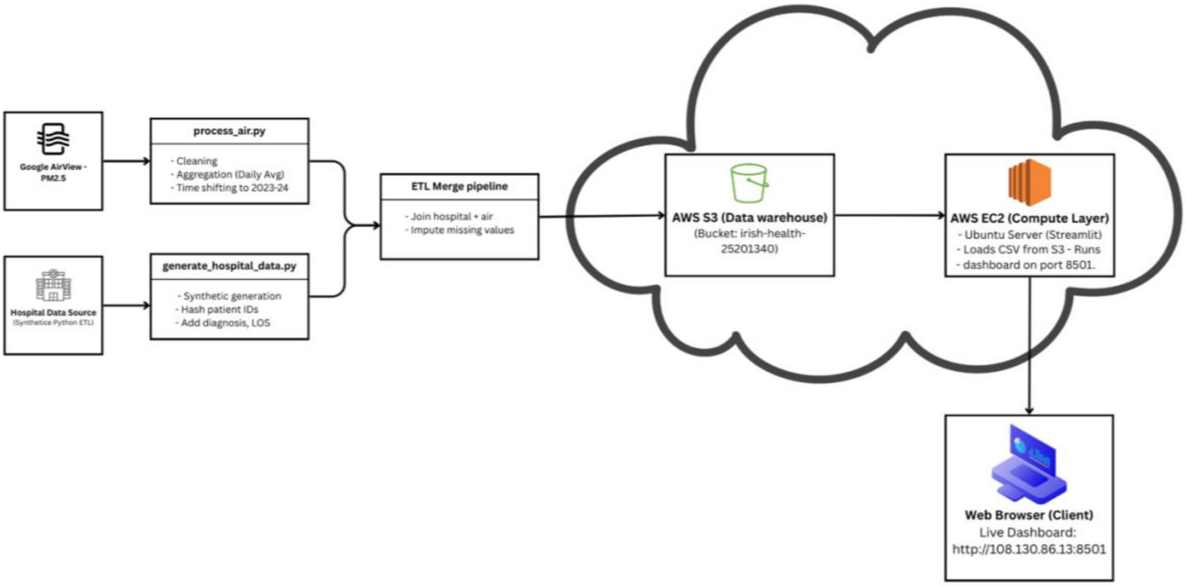


Fig 8 : System Architecture

The system architecture is designed as an end-to-end cloud-based pipeline, leveraging AWS services for storage and computation. As illustrated in the diagram, the process begins with two distinct data sources: Google Air View pollution data and synthetically generated hospital admissions data.

Data from each source is first processed through dedicated Python-based ETL scripts. These scripts perform essential cleaning, aggregation, and formatting tasks. Subsequently, a merge pipeline joins the two datasets and handles any missing values to create a unified dataset.

This final, processed dataset is then uploaded to an AWS S3 bucket, which serves as the central cloud data warehouse. An AWS EC2 instance running Ubuntu acts as the compute layer, hosting the Streamlit application. The application directly reads the merged dataset from S3, performs necessary analytics, and renders the interactive dashboard. Users can access this dashboard via a web browser using the EC2 instance's public IP address on port 8501. This architecture effectively demonstrates the integration of disparate data sources into a cohesive cloud analytics solution.

9. CONCLUSION

This project successfully demonstrates how cloud computing can be used to build a healthcare data warehouse. Using AWS S3 for storage and EC2 for computation, combined with an ETL pipeline and a Streamlit dashboard, the system provides valuable insights into the relationship between air pollution and hospital admissions.

The project meets all requirements and provides a strong foundation for future work such as predictive modelling or machine learning.

10. REFERENCES

- Google Project Air View: https://airview.withgoogle.com
- Streamlit Documentation: https://docs.streamlit.io
- Plotly Documentation: https://plotly.com/python
- AWS S3 Documentation: https://docs.aws.amazon.com/s3
- AWS EC2 Documentation: https://docs.aws.amazon.com/ec2
- Python hashlib documentation
- World Health Organization (2021). WHO Global Air Quality Guidelines.