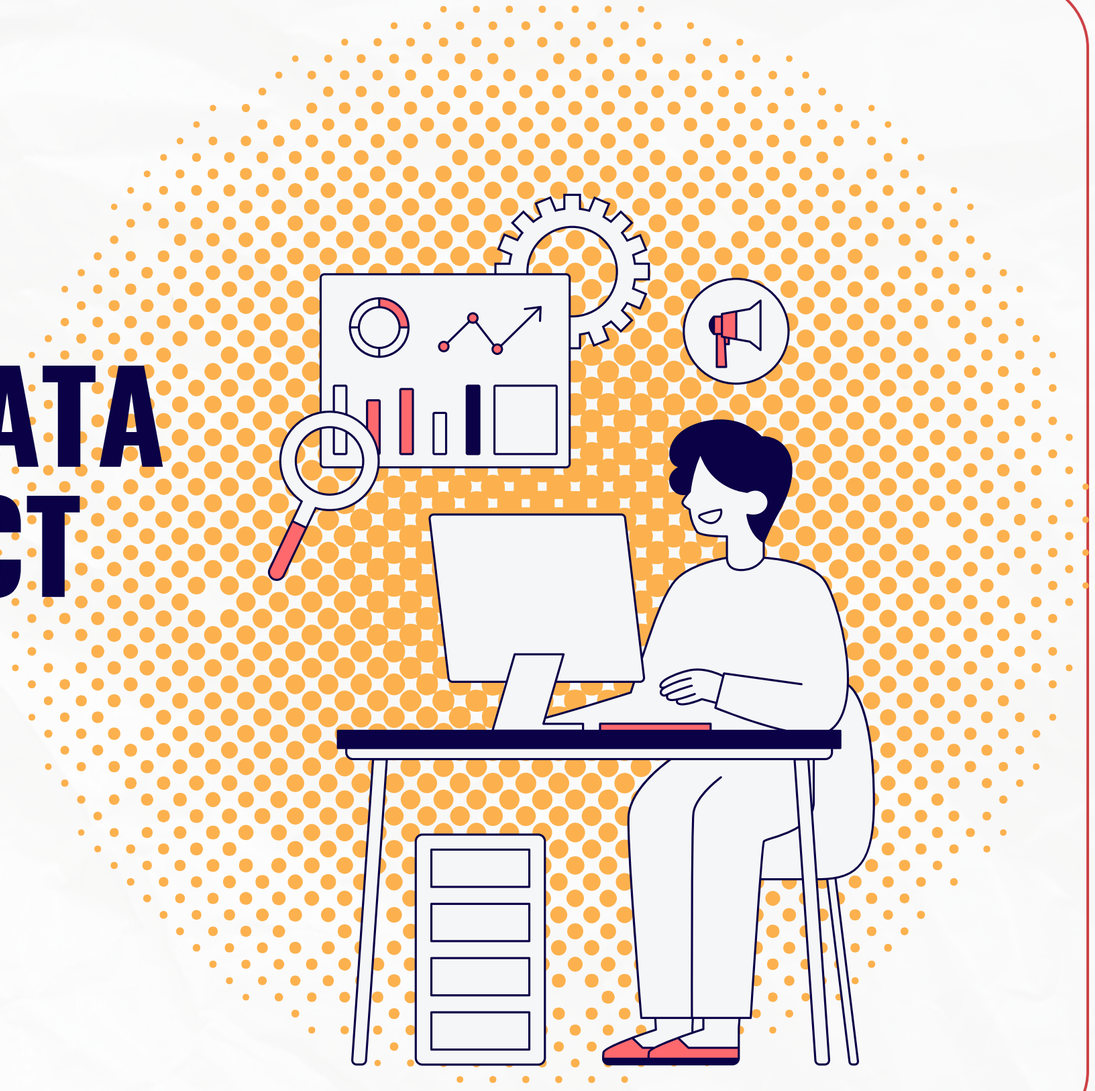
ETL PIPELINE FOR HOSPITAL DATA: A DATA ENGINEERING PROJECT WITH AIRFLOW

INTE 42232 - Data Engineering

Group 02



Introduction

Hospital data is fragmented & inconsistent

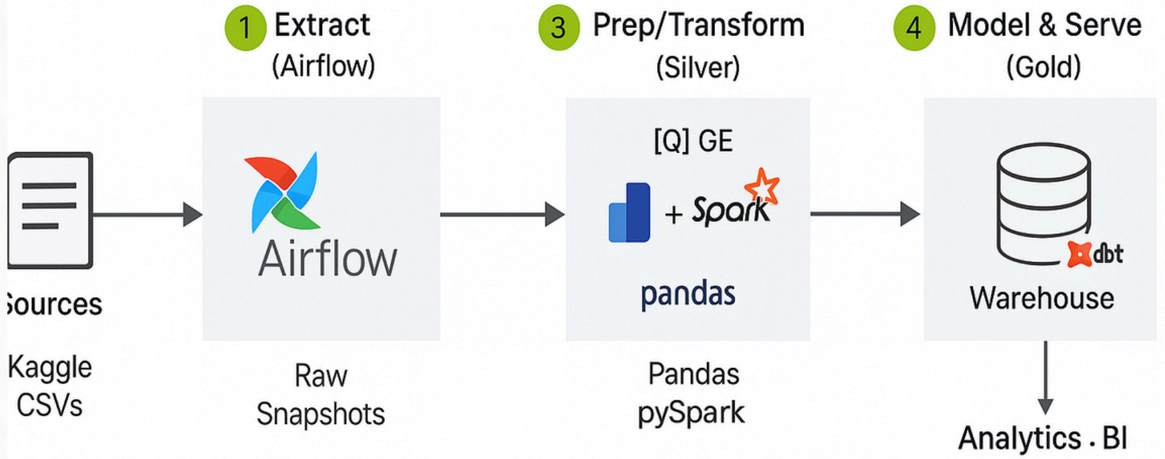
ETL pipeline: extract, clean, transform, load

Covers major data engineering tasks

Solution: automated pipeline with Airflow

Architecture

1. Extraction: hospital CSVs → Bronze (raw storage)
2. Transformation: cleaning, standardization → Silver
3. Loading: star-schema warehouse → Gold
4. Orchestration: Apache Airflow DAGs
5. Validation: Great Expectations & dbt tests



Dataset & Storage Strategy

Kaggle: CA Hospital Dataset Q1 2025

Bronze: raw immutable CSVs

Silver: cleaned Parquet files

Gold: star-schema warehouse (dim/fact)

Scheduling: Airflow tasks

Methodology

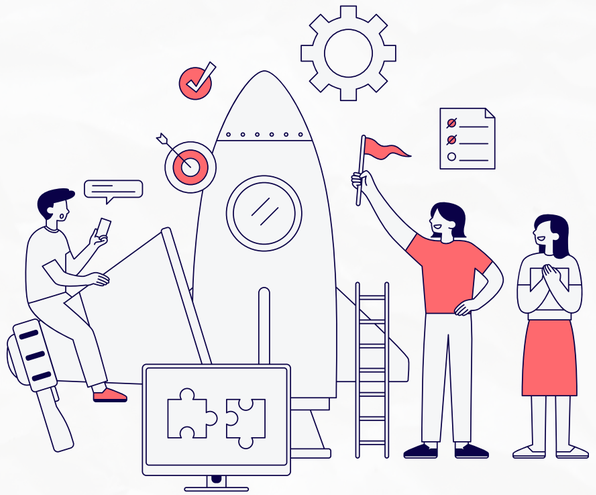
1. Extract → Ingest CSVs into Bronze
2. Validate → Schema & quality checks
3. Transform → Cleaning, deduplication, enrichment
4. Load → Warehouse with star schema
5. Validate again with dbt tests
6. Analytics → Dashboards

Tools Required

- Apache Airflow – orchestration
- Pandas / PySpark – transformations
- PostgreSQL / BigQuery – warehouse
- dbt – modeling & testing
- Great Expectations – data quality
- Tableau / Power BI – visualization

Expected Outcomes

1. Automated hospital ETL pipeline
2. Clean, integrated warehouse
3. Star schema for analytics



Thank You