BUAN 6390.S01 S25 – Analytics Practicum

Group 6 - Healthcare Costs: Stop Guessing, Start Saving

Project Sponsor: Vijay Koneru

Professor: Mandar Samant

Monish Jayaprakash Seelam – MJS230000

Mohitha Suresh Singh – MXS220125

Vinoth Premkumar – VXP230018

Panjam Puneeth Kumar Reddy – PXP230043

Rupak Manikonda – RXM230010

Viraja Nelagi – PXN230007

Neelkumar Kalavadiya – NXK230004

Guru Kiranjha Bojja - GXB230013

Problem Overview and Strategic Approach

Problem Statement
Patients lack transparent, accessible data to compare treatment costs across hospitals and insurance providers, leading to uninformed healthcare decisions and unexpected expenses.

Proposed Solution
We collect, clean, and analyze hospital and insurance pricing data to present clear cost comparisons, helping users choose affordable, suitable healthcare options through data-driven insights.

Discovery

Source:
We downloaded price transparency data from 3–4 hospitals in each of the 8 cities.

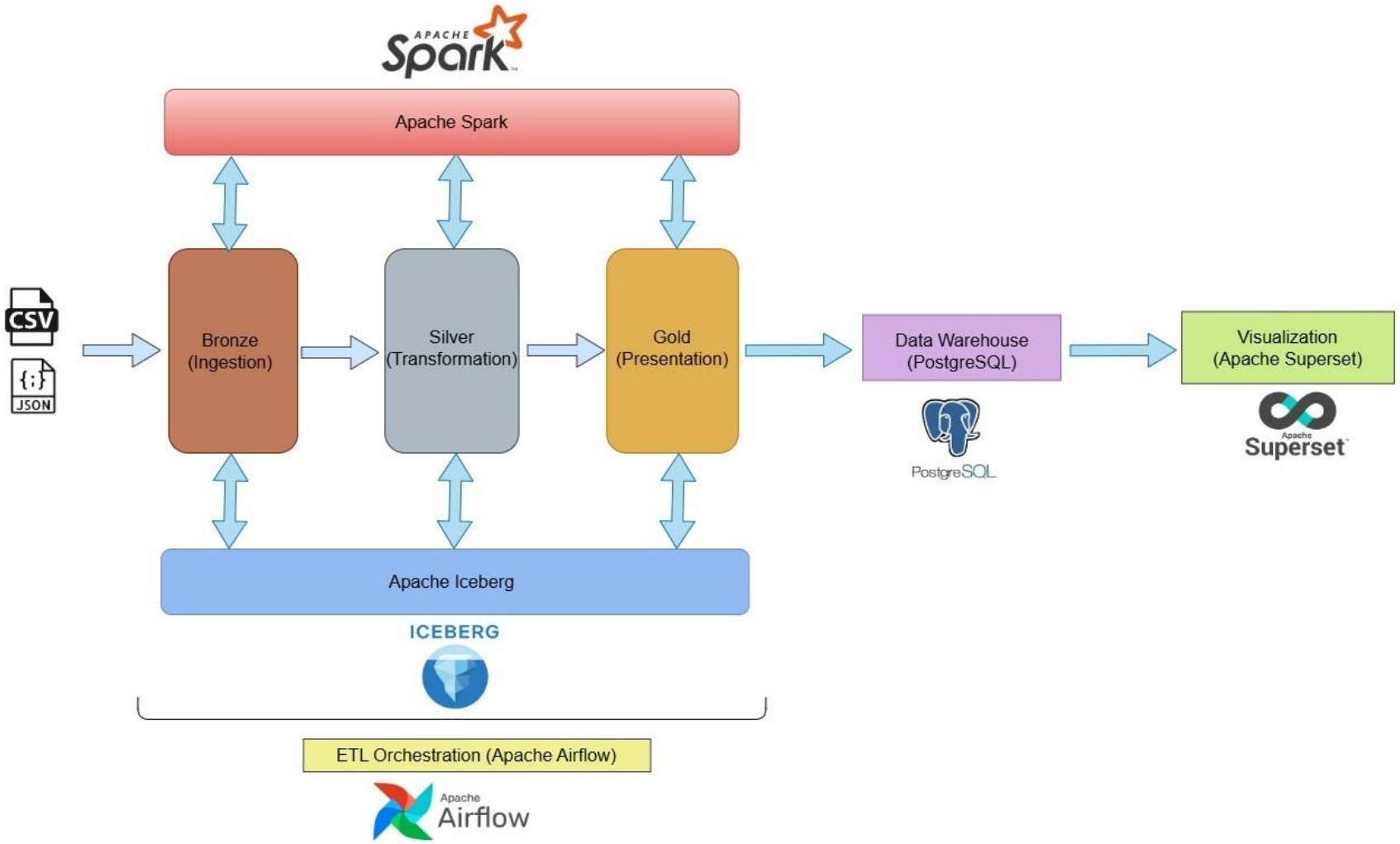
Dataset Details:
Each file is approximately 400 MB in size.

Architecture Used:
Medallion Architecture (Bronze, Silver, Gold layers) to manage data lifecycle.

Goal:
Improve healthcare cost transparency by aggregating and standardizing pricing data.

Expected Outcomes:
Enable patients and stakeholders to make informed decisions through automated pipelines and insightful dashboards.

Solution Architecture



Tools used for Development

Apache Iceberg (Data Storage)
•Why: Provides a scalable, ACID-compliant table format ideal for large-scale analytics. It ensures data versioning, time travel, and schema evolution—all important for healthcare data integrity.

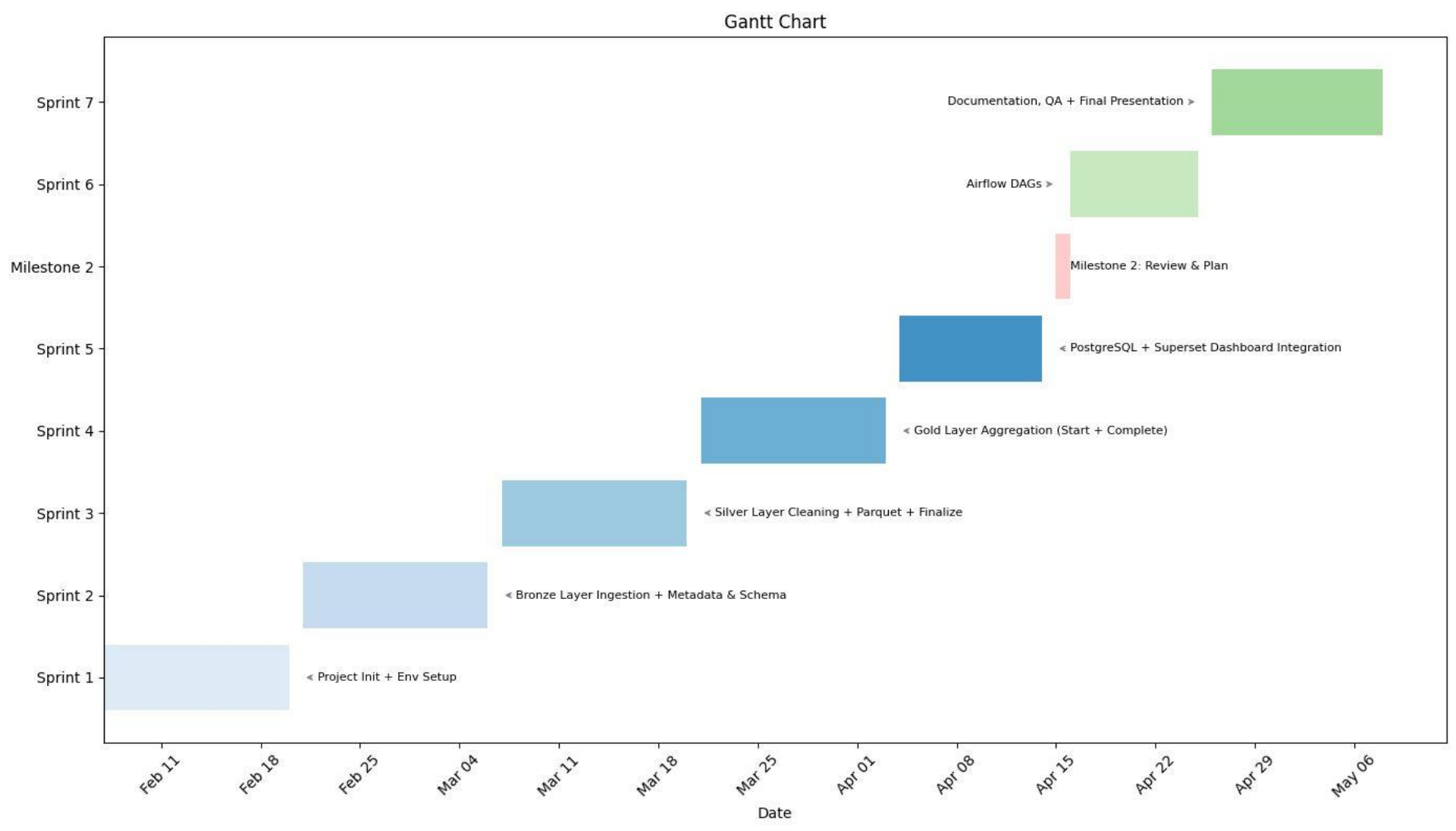
PostgreSQL (Data Warehouse)
Stores the final, processed (Gold Layer) data in a structured format that's easy to query for dashboard tools. It supports complex joins and analytics.

Apache Spark (Data Processing)
Handles distributed data processing efficiently. It's ideal for cleaning, transforming, and aggregating large CSV/JSON healthcare datasets into Parquet format.

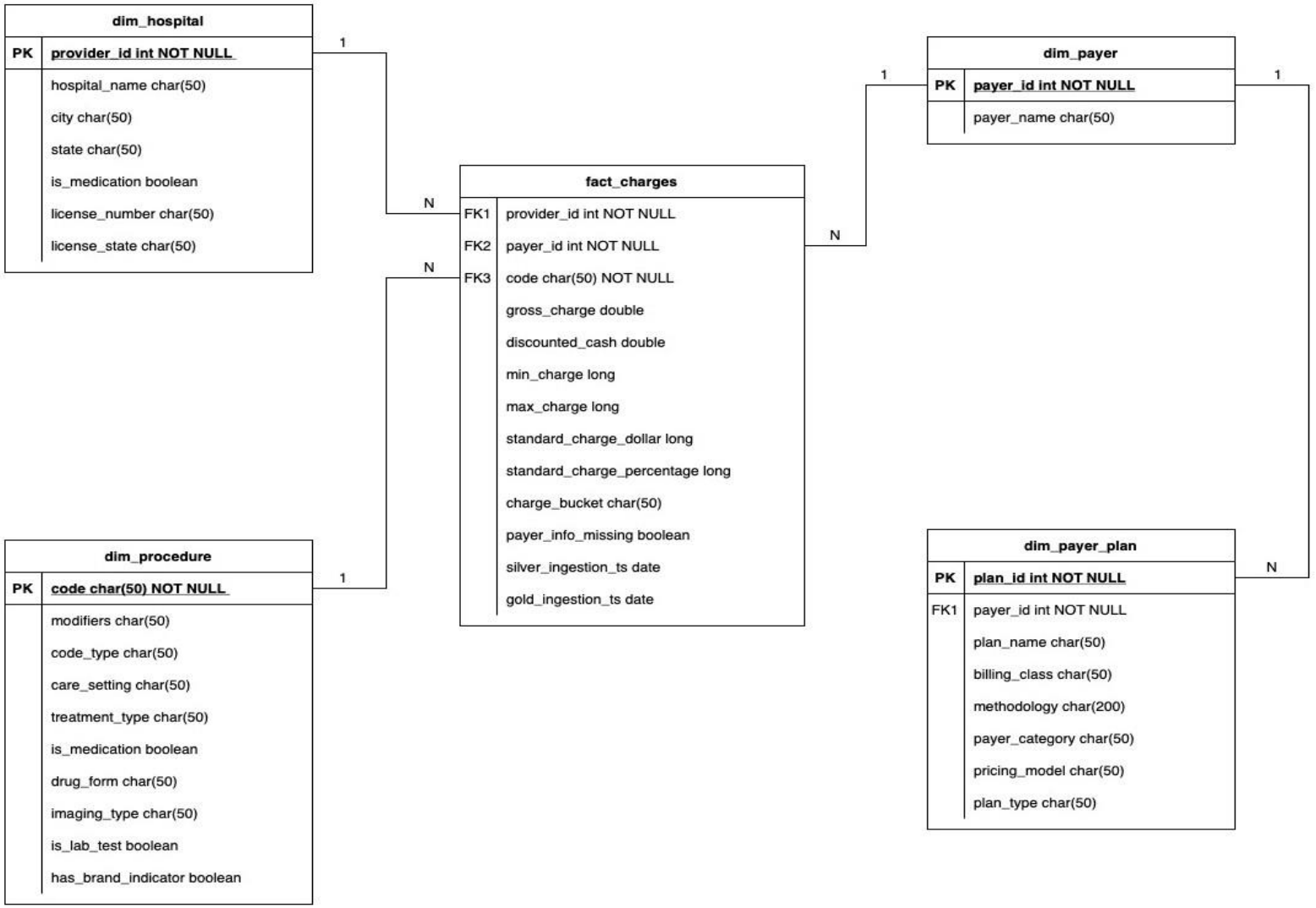
Apache Superset (Visualization)
Creates interactive dashboards that provide cost insights, risk analysis, and trends, helping stakeholders and patients make data-driven decisions.

Apache Airflow (Workflow Orchestration)
Automates the end-to-end workflow—from ingestion to transformation to reporting—through DAGs (Directed Acyclic Graphs), reducing manual intervention and improving reliability.

Roadmap


Gantt Chart
Sprint 7
Sprint 6
Milestone 2
Sprint 5
Sprint 4
Sprint 3
Sprint 2
Sprint 1
Documentation, QA + Final Presentation
Airflow DAGs
Milestone 2: Review & Plan
PostgreSQL + Superset Dashboard Integration
Gold Layer Aggregation (Start + Complete)
Silver Layer Cleaning + Parquet + Finalize
Bronze Layer Ingestion + Metadata & Schema
Project Init + Env Setup
Feb 11
Feb 18
Feb 25
Mar 04
Mar 11
Mar 18
Mar 25
Apr 01
Apr 08
Apr 15
Apr 22
Apr 29
May 06
Date

Entity Relationship Diagram (ERD)



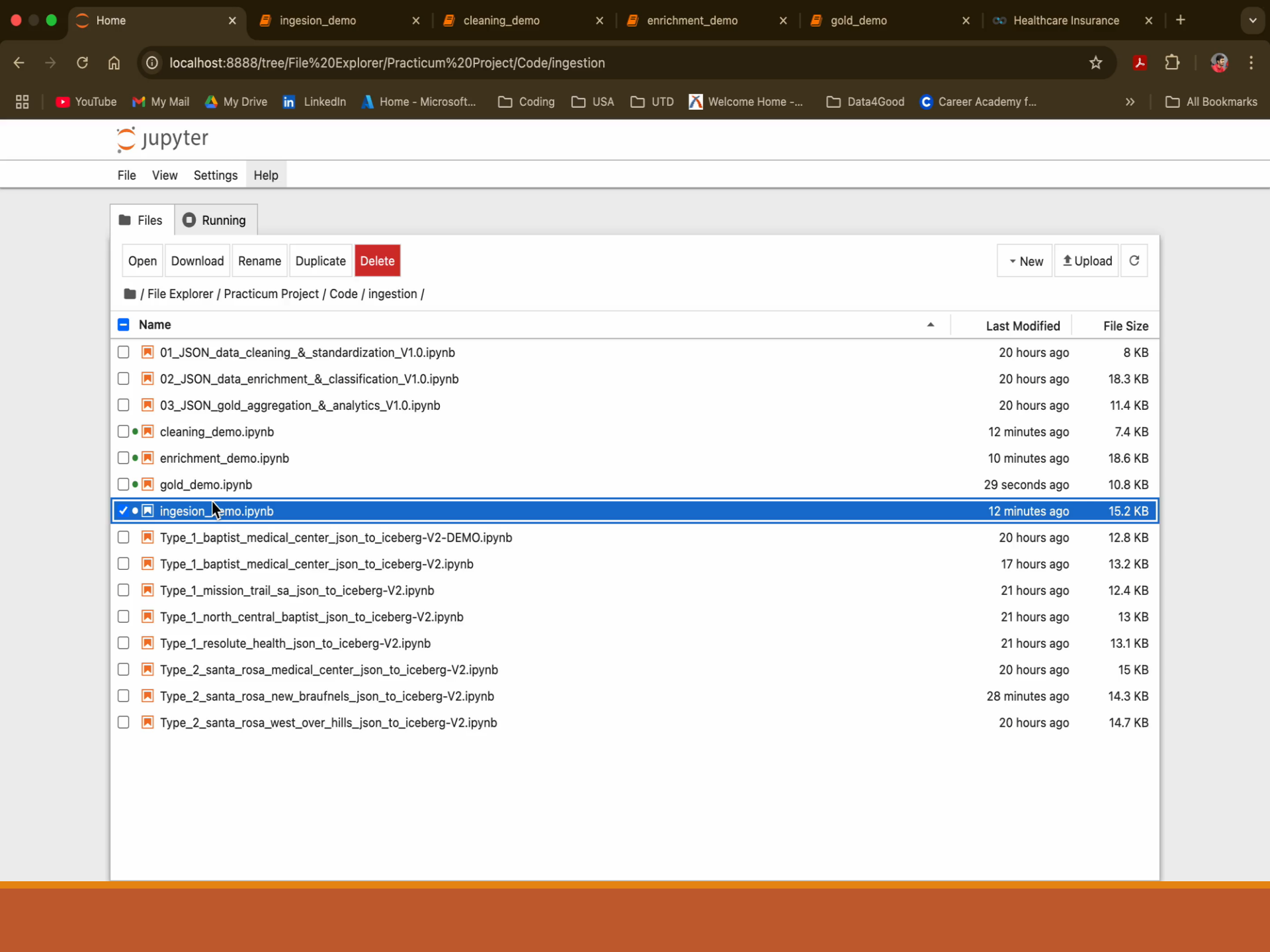
Future Extension

1. **Workflow Automation**: Implement Apache Airflow to orchestrate and automate data pipeline tasks for enhanced efficiency and reliability.

2. **Standardized Framework:** Establish a unified workflow and standardization process to ensure consistency across the entire pipeline.

3. **Documentation Creation:** Develop clear and concise project documentation to outline the architecture, workflow, and implementation details.

Demonstration of the project
Video


Home
ingesion_demo
cleaning_demo
enrichment_demo
gold_demo
Healthcare Insurance
localhost:8888/tree/File%20Explorer/Practicum%20Project/Code/ingestion
YouTube
My Mail
My Drive
LinkedIn
Home - Microsoft...
Coding
USA
UTD
Welcome Home -...
Data4Good
Career Academy f...
All Bookmarks
jupyter
File
View
Settings
Help
Files
Running
Open
Download
Rename
Duplicate
Delete
New
Upload
/ File Explorer / Practicum Project / Code / ingestion /
Name
Last Modified
File Size
01_JSON_data_cleaning_&_standardization_V1.0.ipynb 20 hours ago 8 KB
02_JSON_data_enrichment_&_classification_V1.0.ipynb 20 hours ago 18.3 KB
03_JSON_gold_aggregation_&_analytics_V1.0.ipynb 20 hours ago 11.4 KB
cleaning_demo.ipynb 12 minutes ago 7.4 KB
enrichment_demo.ipynb 10 minutes ago 18.6 KB
gold_demo.ipynb 29 seconds ago 10.8 KB
ingesion_demo.ipynb 12 minutes ago 15.2 KB
Type_1_baptist_medical_center_json_to_iceberg-V2-DEMO.ipynb 20 hours ago 12.8 KB
Type_1_baptist_medical_center_json_to_iceberg-V2.ipynb 17 hours ago 13.2 KB
Type_1_mission_trail_sa_json_to_iceberg-V2.ipynb 21 hours ago 12.4 KB
Type_1_north_central_baptist_json_to_iceberg-V2.ipynb 21 hours ago 13 KB
Type_1_resolute_health_json_to_iceberg-V2.ipynb 21 hours ago 13.1 KB
Type_2_santa_rosa_medical_center_json_to_iceberg-V2.ipynb 20 hours ago 15 KB
Type_2_santa_rosa_new_braufnels_json_to_iceberg-V2.ipynb 28 minutes ago 14.3 KB
Type_2_santa_rosa_west_over_hills_json_to_iceberg-V2.ipynb 20 hours ago 14.7 KB

THANK YOU