Project Report: Healthcare Costs - Stop Guessing, Start Saving

Course: BUAN 6390.S01 S25 (Friday- Group 6)
Project Sponsor: Vijay Koneru
Professor: Mandar Samant

GitHub link: https://github.com/NeelDevX/health-insurance-data-pipeline

1. Executive Summary:

Patients often face unpredictable medical costs due to the lack of pricing transparency in healthcare. This project aims to resolve this issue by designing a robust, open-source, end-to-end data pipeline that aggregates, cleans, transforms, and visualizes healthcare pricing data from multiple hospitals and insurance providers. Using Medallion Architecture and tools like Apache Spark, Apache Airflow, and PostgreSQL, the solution delivers actionable insights through dashboards, enabling users to make informed healthcare decisions.

2. Problem Statement:

In the current healthcare system, patients struggle with unexpected costs because there is no effortless way to compare treatment prices across different hospitals and insurers. This project addresses the gap by creating a scalable and automated solution that enables cost comparison and transparency through data analytics.

3. Project Objectives:

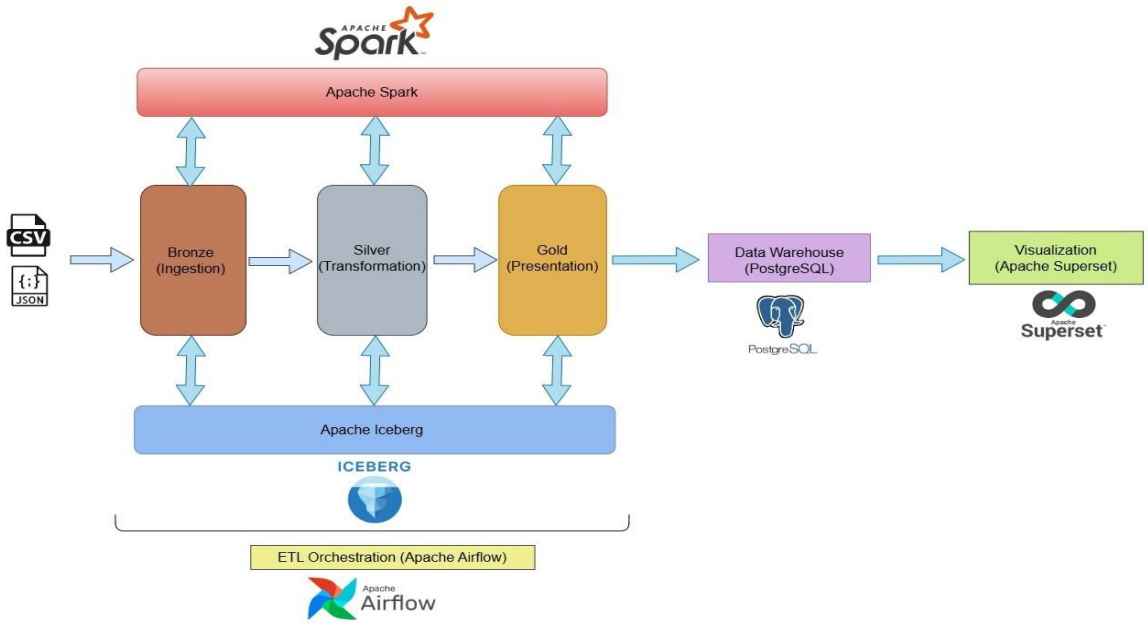
- Collect healthcare pricing data from diverse sources.
- Clean and standardize the data using Apache Spark
- Store data efficiently using Apache Iceberg
- Automate the data pipeline using Apache Airflow
- Use PostgreSQL to manage structured, relational data for warehouse querying and analytics.
- Visualize structured insights using Apache Superset

4. Project Scope:

In-Scope:

- CSV and JSON data ingestion
- Data transformation (cleaning, deduplication, masking PII)
- Aggregation for analytics and reporting
- Pipeline orchestration using Apache Airflow
- Management of structured data in PostgreSQL for relational querying and data warehousing.
- Dashboarding for healthcare cost comparison

5. Architecture Overview:



We used the **Medallion Architecture** comprising three layers:

Bronze Layer:

- Stores raw ingested files (CSV, JSON)
- Enables reprocessing and auditing.

Silver Layer:

- Cleans and standardizes data.
- Manages missing data, duplicates, PII masking.
- Convert data into Parquet format.

Gold Layer:

- Aggregates data for advanced analytics
- Includes KPIs like claim settlements, billing comparisons, and fraud detection.
- Stores structured data in PostgreSQL for downstream analytics.

Workflow Automation:

- Apache Airflow orchestrates DAGs for ingestion, cleaning, and reporting.

Data Warehouse Management:

- PostgreSQL serves as the central RDBMS for managing clean and aggregated data to support analytical queries and reporting needs.

Visualization:

- Dashboards built using Apache Superset

6. Tools & Technologies:

Layer	***Tool***	***Purpose***
Ingestion	Apache Airflow	Automate ingestion and pipeline management.
Storage	Apache Iceberg	ACID-compliant, scalable data storage
Processing	Apache Spark	Data cleaning, transformation, and aggregation
Orchestration	Apache Airflow	DAG-based automation of workflows
Storage PostgreSQL	PostgreSQL	RDBMS for data warehouse and relational data analysis
Visualization	Superset	Create cost comparison dashboards.

7. Milestone Timeline:

Milestone	**Timeline**
Data Collection & Schema Design	Feb 2025
Data Cleaning and Transformation	March 2025
Data Aggregation & Analytics	March 2025
Workflow Automation	April 2025

Milestone	Timeline
Dashboard Development	April 2025

8. Key Challenges:

- Managing large, unstructured healthcare datasets
- Configuration across systems.
- Performance issues during transformation and loading
- Handling bad or inconsistent data that impacts analysis.
- Optimizing Spark jobs for performance
- Scheduling and managing complex DAGs in Airflow

9. Project Impact:

- Delivered automated, transparent, and scalable data pipeline.
- Empowered users to view, compare, and make informed healthcare decisions.
- Set foundation for integrating ML models for risk scoring and cost prediction in the future.

10. Future Enhancements:

- Integrate NLP to auto-classify treatment descriptions.
- Introduce ML models for benchmarking and fraud detection.
- Optimize Airflow pipelines for production scaling.
- Expand data sources across more hospitals and states.

11. Conclusion:

This project equips stakeholders with critical cost transparency tools through modern data engineering practices. By automating the healthcare data pipeline, utilizing PostgreSQL for warehouse management, and presenting insights through interactive dashboards, we support smarter, cost-effective decision-making in the healthcare domain.

12. References:

1. Apache Airflow Documentation – https://airflow.apache.org
2. Apache Spark Documentation – https://spark.apache.org/docs
3. PostgreSQL Documentation – https://www.postgresql.org/docs
4. Apache Superset – https://superset.apache.org

Group 6 Members (Friday class):

- Monish Jayaprakash Seelam – MJS230000
- Mohitha Suresh Singh – MXS220125
- Vinoth Premkumar – VXP230018
- Panjam Puneeth Kumar Reddy – PXP230043
- Rupak Manikonda – RXM230010
- Viraja Nelagi – PXN230007
- Neel Kumar Kalavadiya – NXK230004
- Guru Kiranjha Bojja – GXB230013