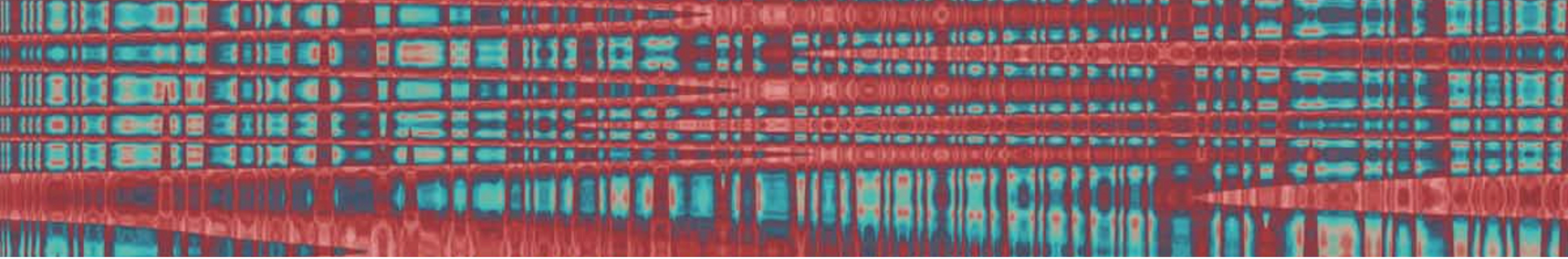
Data Warehouse & BI Dashboard for Healthcare Patient Outcome

Team 2

Michael Angelo Agua
Glenn Patrick Darriguez
Melvin Ignacio
Neil Jay Lacandazo
Alvin Tinaan

Data Warehousing / Denver Jhon Calantoc

November 15, 2025



Introduction & Objectives

Our project establishes a Healthcare Patient Outcome Data Warehouse to integrate and analyze medical data (patients, diagnoses, doctors, departments).

Business Problem

Healthcare administrators need to monitor admissions, costs, and outcomes.

Project Goal

Enable data-driven decisions for resource allocation and disease management.

Data Sources

Raw data from various hospital systems is extracted from CSV files, representing OLTP systems.

1

Diagnoses

Contains diagnosis codes and descriptions. (diagnoses.csv)

2

Patients

Includes patient demographic information. (patients.csv)

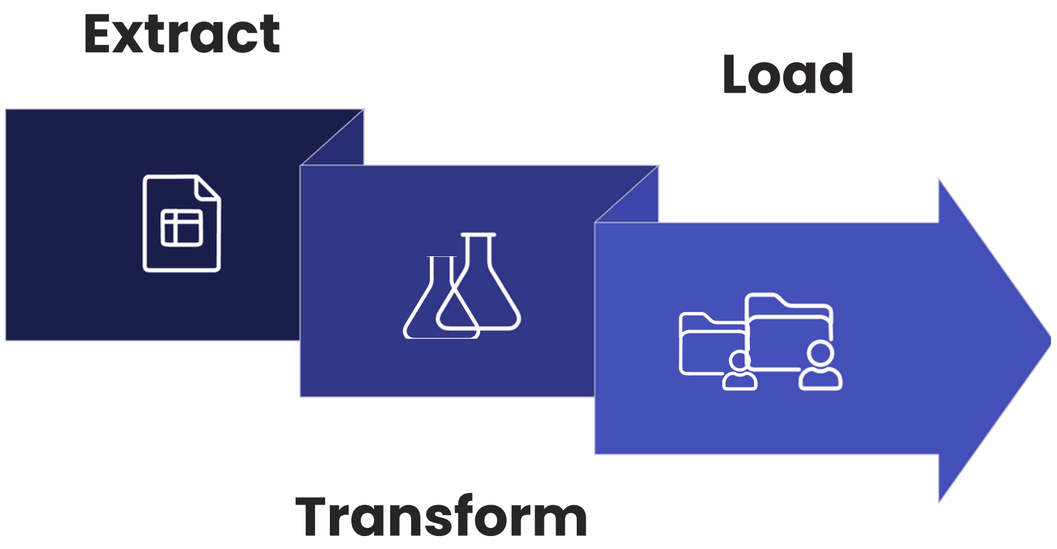
3

Treatments

Details treatment costs, doctors, and departments. (treatments.csv)

ETL Process Overview

Our ETL process ensures clean, consistent, and analyzable data using Python (pandas) and MySQL.



Extract

Load CSVs into Python pandas DataFrames.

Transform

Cleanse, standardize, and enrich data using Python.

Load

Populate MySQL data warehouse via Python script.

Data Warehouse Design: Star Schema

The data warehouse uses a Star Schema, optimizing for analytical queries.

dim_patient

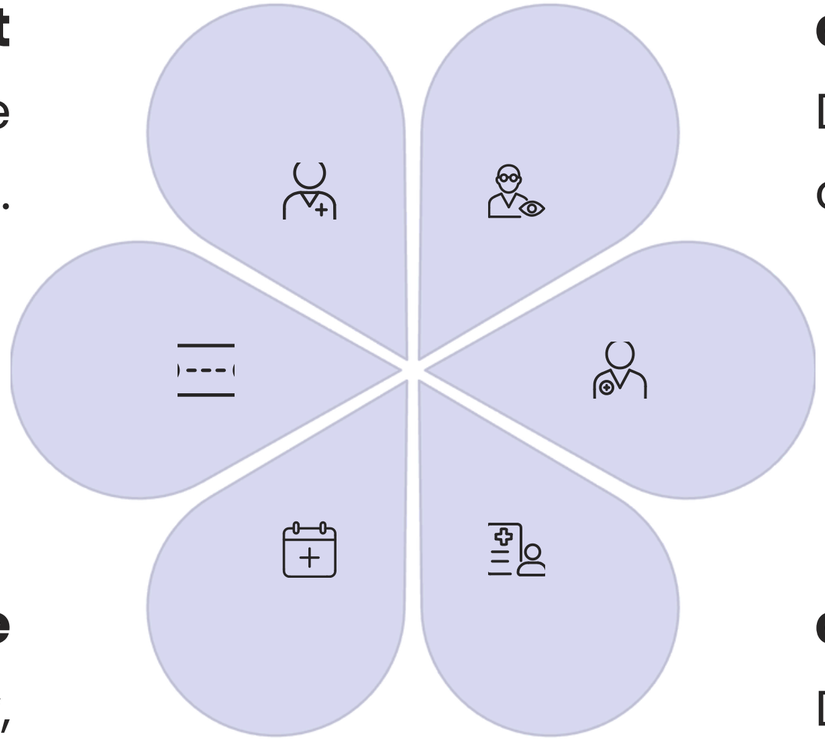
Patient demographics, age groups.

fact_admissions

Central fact table: treatment costs, linked to dimensions.

dim_date

Time-based data (day, month, year).



dim_diagnosis

Diagnosis codes and categories.

dim_doctor

Doctor names and specializations.

dim_department

Department names and locations.

Schema & Table Design

Our schema ensures data consistency and referential integrity with surrogate keys and foreign key relationships.

dim_patient	patient_id	INT (PK)	Surrogate key
dim_diagnosis	diagnosis_id	INT (PK)	Surrogate key
fact_admissions	patient_id	INT (FK)	Links to dim_patient
fact_admissions	treatment_cost	DECIMAL(10,2)	Measure field for analysis

Data Loading Process

Cleaned datasets are loaded into the MySQL data warehouse via an automated Python script.

01

Connect to DB

Python script connects to MySQL (healthcare_dw).

02

Execute SQL

Reads script.sql for schema creation and data loading.

03

Load Data

LOAD DATA LOCAL INFILE imports cleaned CSVs into staging tables.

04

Populate DW

INSERT INTO ... SELECT populates dimension and fact tables.

Data Dictionary & Metadata

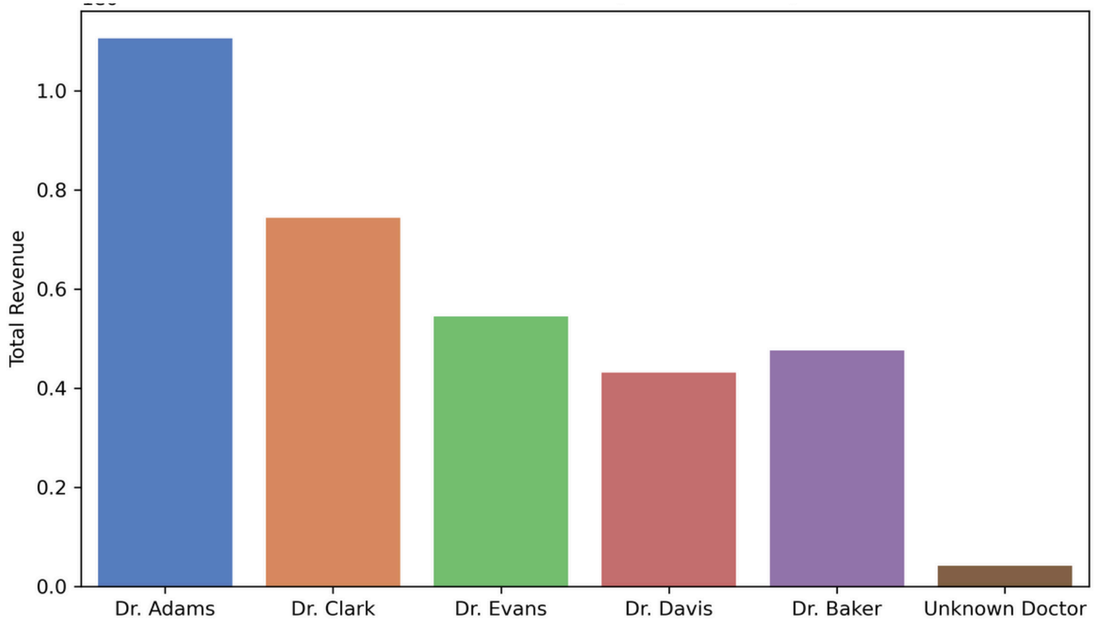
Our data dictionary provides clear definitions and naming conventions for all tables and columns.

diagnoses	diagnosis_code	VARCHAR(50)	Standard medical code
patients	age	INT	Patient’s age at admission
treatments	treatment_cost	DECIMAL(10,2)	Cost of the treatment

Reports & Insights

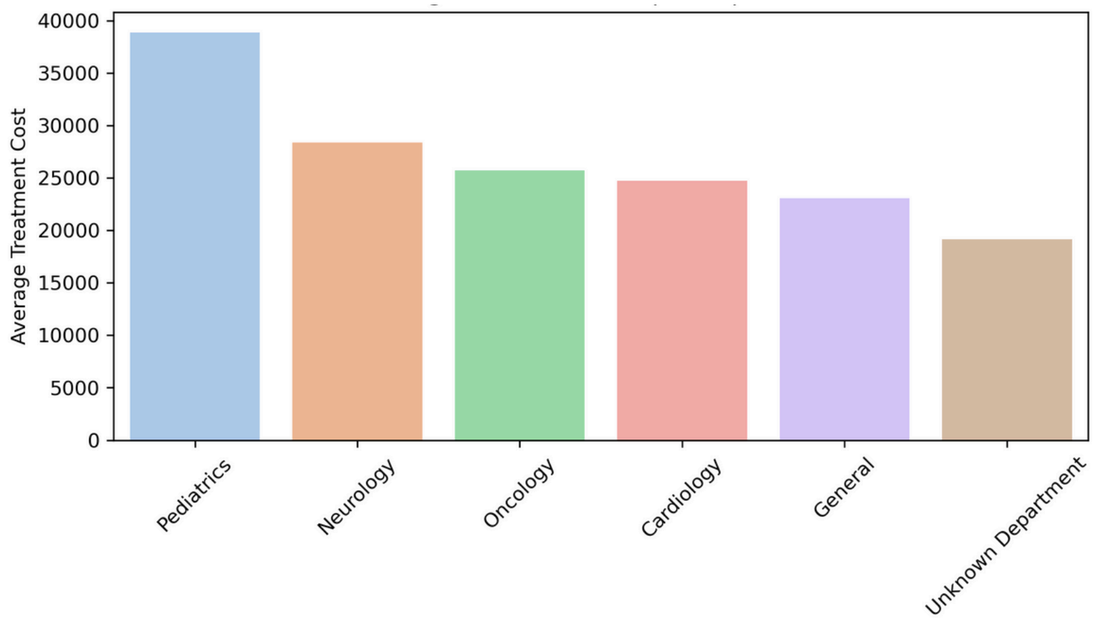
The data warehouse enables critical analytical results and visualizations.

Doctor Performance



Dr. Adams is the top performer, generating over $1.1 million.

Average Treatment Cost per Department



Pediatrics has the highest average treatment cost at $39,000.

Key Findings

Doctor Performance & Data Integrity

Dr. Adams leads in revenue generation with $1.1M, highlighting top individual contributions. However, a significant number of cases are linked to "Unknown Doctor," indicating a need for improved data entry to accurately track physician performance and resource utilization.

Disease Trend Analysis

Seasonal patterns for Asthma and Flu are evident, allowing for proactive resource planning. An unusual dip in all disease diagnoses in August warrants further investigation into potential data anomalies or operational shifts during that period.

Key Findings

Departmental Cost Efficiency

Pediatrics records the highest average treatment cost at $39,000, prompting a review of cost drivers in this department. Additionally, substantial costs associated with "Unknown Department" point to critical data gaps impacting accurate cost allocation and budgeting.

These detailed insights support administrators in optimizing resource allocation, enhancing operational efficiency, and refining strategic planning for better patient outcomes and financial management.

Challenges and Solutions

1

Data Inconsistencies & Code Mismatches

Implemented robust data validation rules and standardization processes, addressing inconsistencies and resolving code mismatches for enhanced data accuracy.

2

Missing Values

Utilized advanced imputation techniques and systematic cleansing routines to effectively fill data gaps and ensure the completeness and reliability of datasets.

Challenges and Solutions

3

ETL Complexity

Designed a comprehensive and modular ETL workflow, incorporating incremental loading strategies and robust error handling to streamline data integration processes.

4

Performance Optimization

Applied indexing, partitioning, and pre-aggregated tables to dramatically reduce query execution times and enhance overall reporting speed, ensuring timely insights.

Group Roles & Contributions

Neil Jay Lacandazo	ETL Engineer	Designed data transformation scripts
Glenn Patrick Darriguez	Database Architect	Created schema and database objects
Glenn Patrick Darriguez	BI Analyst	Built reports and dashboards
Neil Jay Lacandazo	Data Modeler	Documented ERD and metadata
Neil Jay Lacandazo	Project Manager	Coordinate tasks, compile final report, lead presentation

Conclusion

Reflection on Learning

Throughout this project, our group gained a deeper understanding of the entire ETL (Extract, Transform, Load) lifecycle. This included practical experience in:

- Designing and implementing efficient data cleaning processes to ensure data integrity.
- Developing robust schema designs and applying dimensional modeling principles for optimal data warehousing.
- Enhancing our technical skills in Python for data manipulation and SQL for database querying and management.

Conclusion

Importance of Data Warehousing for Healthcare

Data warehousing plays a critical role in modern healthcare by:

- Consolidating disparate data sources into a single, unified source of truth, eliminating data silos.
- Enabling comprehensive, multi-angle analysis of patient data, operational efficiency, and treatment outcomes.
- Empowering decision-makers with actionable insights to improve patient care, streamline operations, and reduce costs.

Conclusion

Possible Improvements & Next Steps

To further enhance our data warehousing solution, we propose the following improvements and next steps:

- Integrating additional data sources, such as electronic health records (EHRs), medical imaging, and claims data, for a more holistic view.
- Automating ETL processes using tools like Apache Airflow to improve efficiency and reduce manual effort.
- Implementing advanced analytics and predictive modeling techniques to forecast trends and identify potential health risks.

Thank You