# Pharmacy Analytics Lakehouse

From Raw Data to Intelligent Healthcare Insights

Guru Saran Satsangi Peddinti

# Domain Context: Retail Pharmacy & Healthcare Supply Chain

## Non-Clinical Data Focus

This project exclusively deals with non-clinical, non-sensitive pharmacy operational data, ensuring patient privacy and data security.

## Operational Challenges

Pharmacies face constant battles with inventory management, stock-outs, and expiry losses, directly impacting patient care and profitability.

## Analytics for Medicine Availability

Leveraging data analytics is crucial to optimizing medicine availability, a critical factor in public health and patient satisfaction.

# The Problem: Transforming Pharmacy Data into Actionable Insights

How can raw pharmacy transaction and inventory data be transformed into a governed analytics platform to address critical operational challenges?

### Reduce Stock-Outs

Ensure essential medicines are always in stock, preventing missed sales and patient inconvenience.

### Minimize Expiry Losses

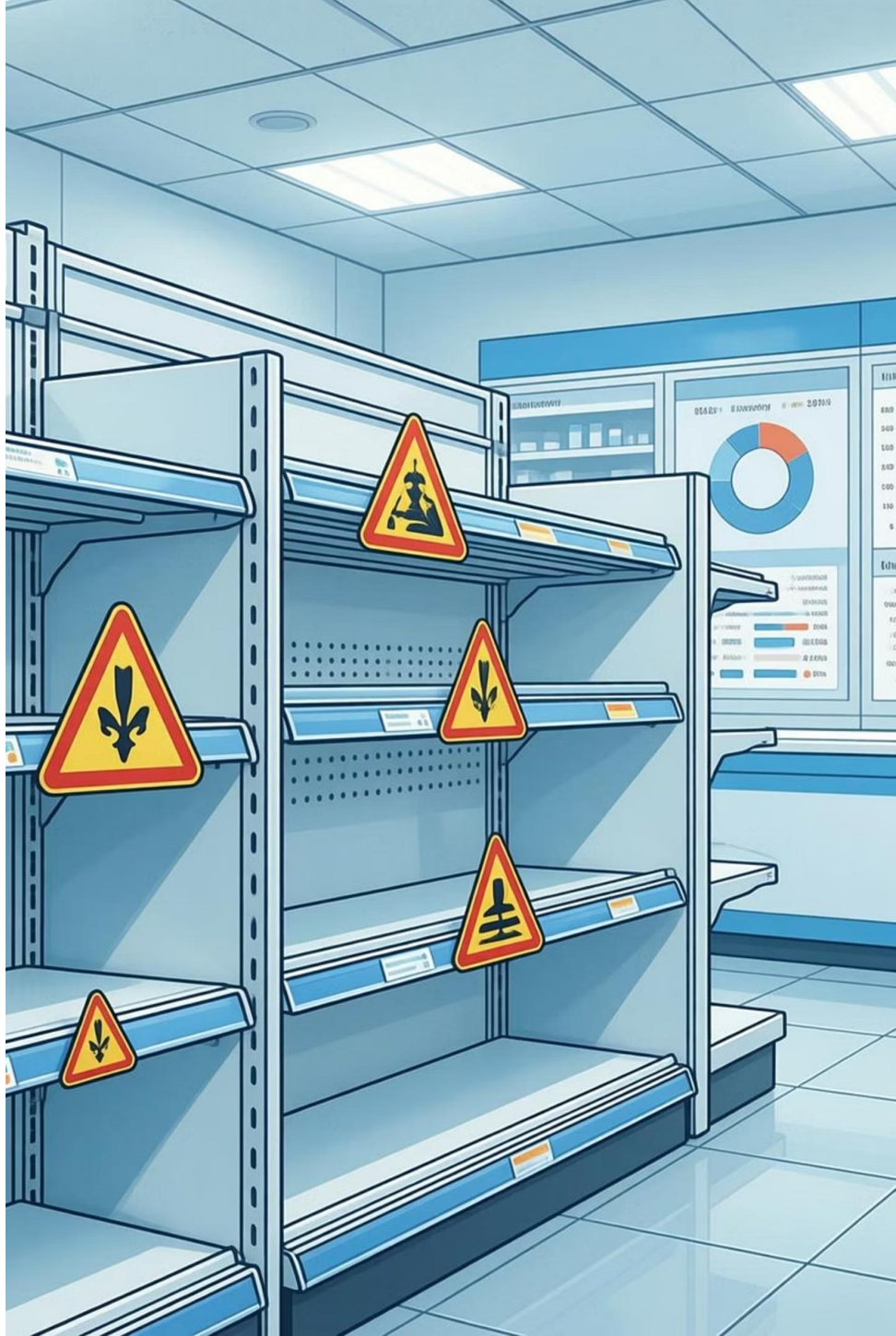
Implement proactive measures to reduce financial losses from expired inventory.

### Improve Medicine Availability

Optimize inventory levels to meet demand consistently and efficiently.

### Optimize Revenue

Maximize category and product-level revenue without relying on sensitive patient or clinical data.



# Data Overview: Synthetic Pharmacy Transactions

This project utilizes a synthetic dataset of 1 million pharmacy records, meticulously designed to mimic real-world pharmacy operations while maintaining data privacy.

## Key Attributes:

- Medicine Name, Combination
- Manufacturer
- Category (e.g., Pain Relief, Antibiotics)
- Manufacturing & Expiry Dates
- Price
- Batch Number
- Form (Tablet, Syrup, Injection, etc.)
- Quantity, Amount

# Tools & Technologies Used

This project was built entirely on Databricks, leveraging its unified Lakehouse capabilities end to end.

## Core Tools & Tech used

### Databricks Workspace

Central environment for notebooks, jobs, dashboards, and versioned development

### Databricks Notebooks

Used for data generation, Bronze/Silver/Gold transformations, exploratory analysis, and ML preparation

### Apache Spark (Databricks Runtime)

Distributed data processing, PySpark DataFrames & Spark SQL, large-scale transformations on 1M+ records

### Delta Lake

ACID-compliant storage, Bronze–Silver–Gold Medallion Architecture, schema enforcement & evolution

### Databricks SQL

Analytical queries, KPI computation, powering dashboards and visualizations

### Databricks Dashboards

Interactive analytics, category-wise and expiry-risk insights, business-facing visual layer

### Databricks Jobs (Workflows)

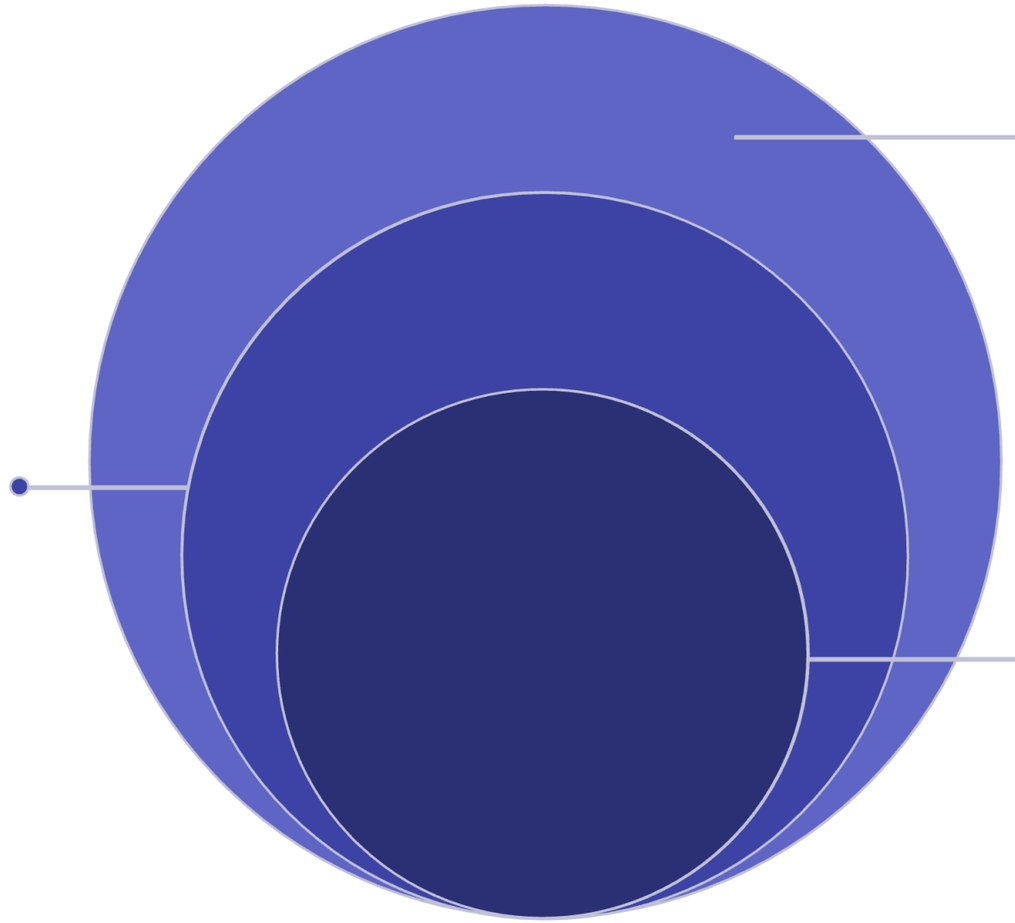
Orchestrated pipelines, Bronze → Silver → Gold execution flow, production-style scheduling


### Unity Catalog (Governance)

Logical data organization, schema-level separation, foundation for secure analytics

### MLflow (Databricks Native)

Model experimentation, metric tracking, regression & classification workflows



**Gold, ML & Dashboards**

Serve analytics, models, and visual insights

**Bronze & Silver**

Clean, validate, and structure curated data

**Raw CSV**

Ingest unprocessed files from sources

The Lakehouse architecture provides a robust and scalable framework for data processing, from raw ingestion to advanced analytics and machine learning applications.

# Lakehouse Architecture Layers

Our Lakehouse architecture organizes data into distinct layers, each serving a specific purpose in transforming raw data into actionable intelligence.

## Bronze Layer (Raw)

The initial ingestion point for raw, unprocessed data from various sources, maintained in its original format with minimal transformation (append-only).

## Silver Layer (Cleaned)

Data undergoes cleaning, validation, and enrichment here, creating a single source of truth that is structured and ready for business use cases.

## Gold Layer (Curated)

Highly aggregated and refined data, optimized for dashboards, reporting, and advanced analytics, providing key performance indicators (KPIs) and business insights.

# Bronze Layer: Raw Data Ingestion

The Bronze layer serves as the initial landing zone for all incoming raw data, preserving its original format and ensuring immutability.

### Append-Only Ingestion

Raw CSV files are ingested as append-only events, capturing every data point without alteration.

### Minimal Transformation

Data undergoes minimal schema inference and conversion to Delta Lake format, maintaining fidelity to the source.

### Ingestion Timestamp

Each record is stamped with an ingestion timestamp for auditing and traceability, crucial for data governance.

### Delta Lake Format

Leveraging Delta Lake provides ACID transactions, schema enforcement, and time travel capabilities from the very first layer.

# Silver Layer: Clean & Enriched Data

The Silver layer refines the raw data, applying data quality checks and enriching it for business readiness.

### Data Quality Checks

Validation rules are applied to ensure data integrity and consistency across all attributes.

### Duplicate Removal

Identification and elimination of duplicate records to ensure accurate analysis.

### Date Normalization

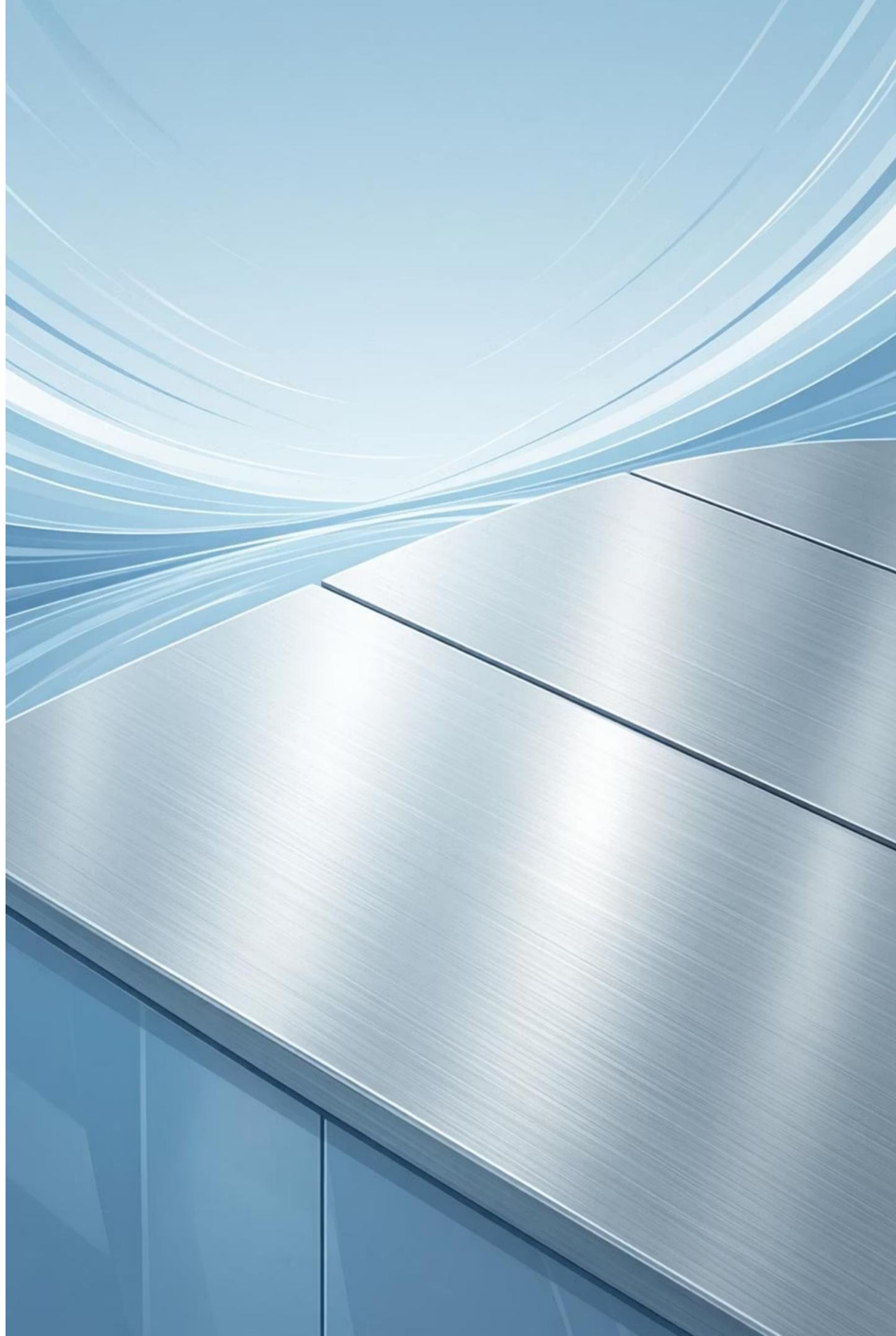
Standardization of date formats for consistent time-series analysis.

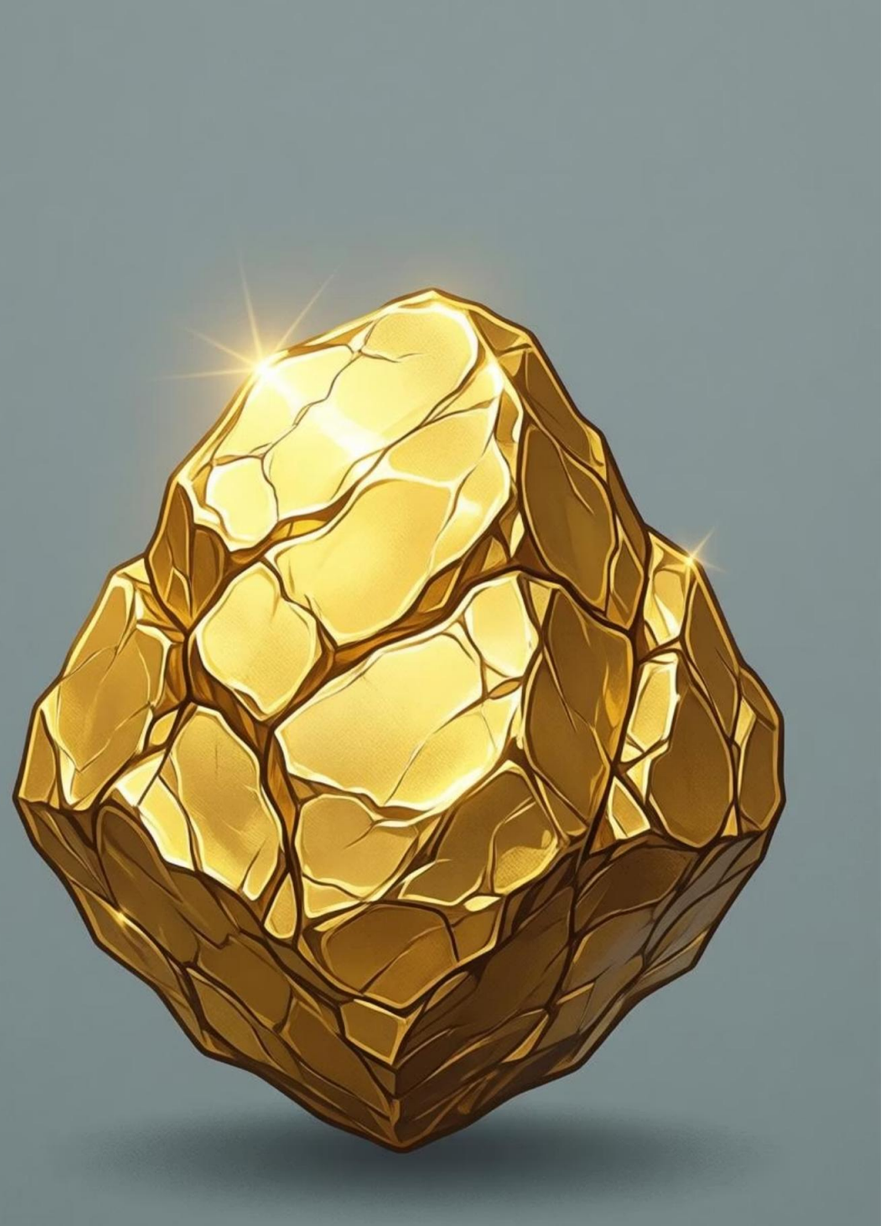
### Expiry Risk Flags

Introduction of flags to identify medicines approaching their expiry date, enabling proactive management.

### Business-Ready Schema

Transformation into a user-friendly schema, optimizing it for downstream analytics and reporting.





# Gold Layer: Business Aggregates

The Gold layer provides highly refined, aggregated data tailored for specific business use cases, dashboards, and advanced analytics.

## Category-Level KPIs

Aggregated metrics like total sales, inventory value, and profit margins by medicine category.

## Product-Level Metrics

Detailed performance indicators for individual medicines, including sales velocity and stock turnover.

## Expiry Risk Summaries

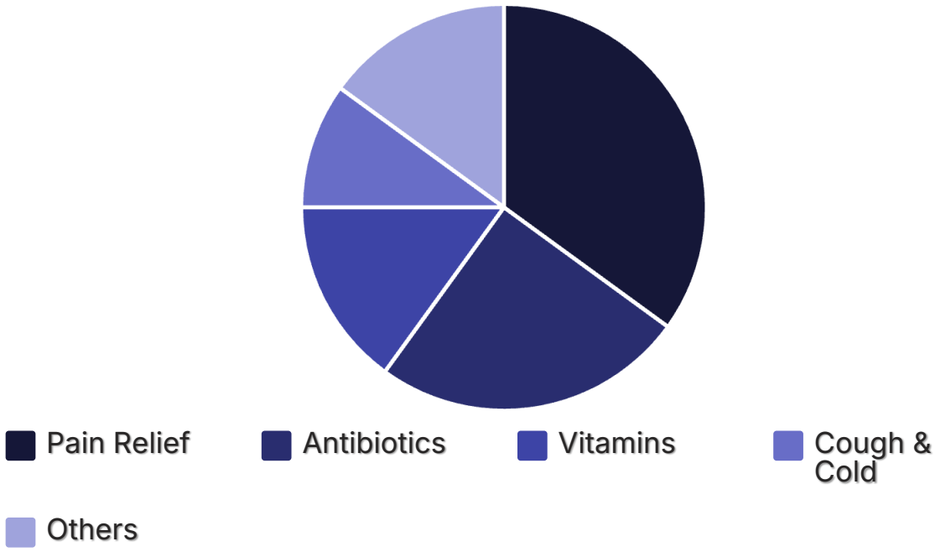
Summaries of near-expiry inventory, highlighting categories and products most at risk.

## Dashboard-Ready Tables

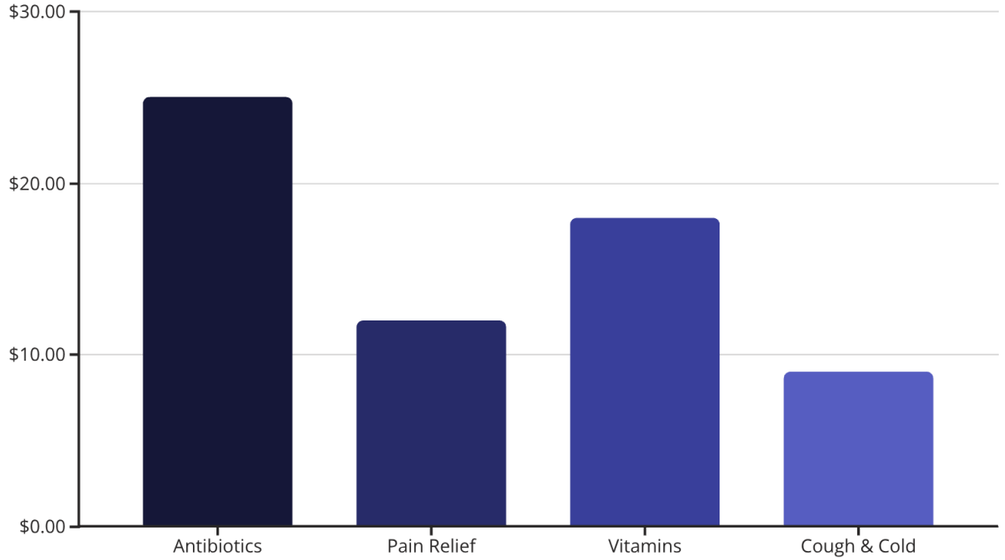
Optimized tables designed for fast query performance in business intelligence tools.

# Analytics & Dashboards: Driving Business Insights

Visual dashboards provide a comprehensive view of pharmacy operations, enabling data-driven decision-making.




Category distribution: Understanding the proportion of different medicine categories in inventory.




Average price by category: Identifying high-value and high-volume product segments.

- Interactive filters enable deep dives into specific products or manufacturers.
- Key Performance Indicators (KPIs) track inventory health and sales trends.

# About the Author

An AI/ML enthusiast with a strong foundation in data engineering and analytics, building real-world healthcare data platforms inspired by hands-on pharmacy operations. This project reflects a passion for transforming complex data into actionable insights, driving efficiency and better healthcare outcomes.

Connect with me: