Retail Data Pipeline: Automated ETL & Business Intelligence

Smart data processing from raw CSV to actionable insights

Business challeines

Business Challenge

Addressing critical pain points in retail data management.

Manual Data Handling

Time-consuming, error-prone processes.

Inconsistent Data

Quality issues across 9 datasets.
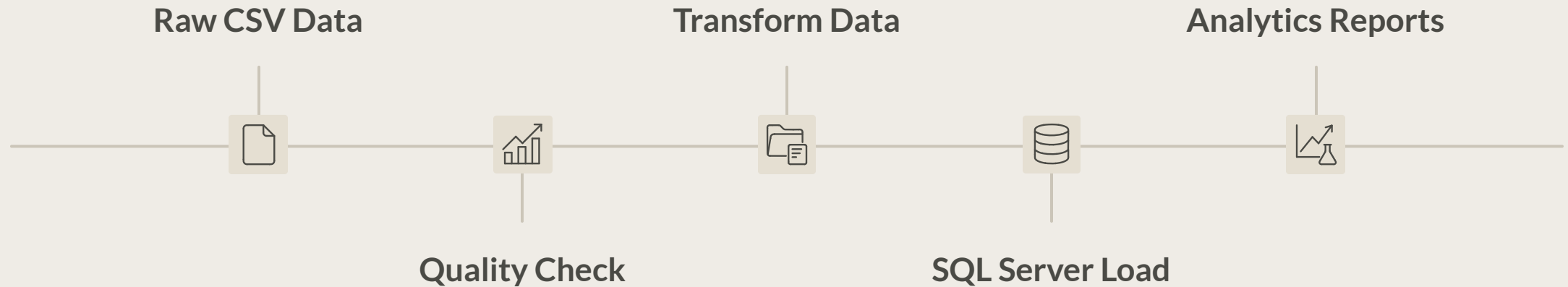
No Automated SQL

Lack of direct database loading.

Limited BI Reports

Absence of real-time insights.

Pipeline Architecture

A robust, automated solution for data integrity and insight generation.

Raw CSV Data

Transform Data

Analytics Reports

Quality Check

SQL Server Load

Stage 1: Data Quality Assessment

Ensuring data reliability from the very beginning.

Automatic Analysis
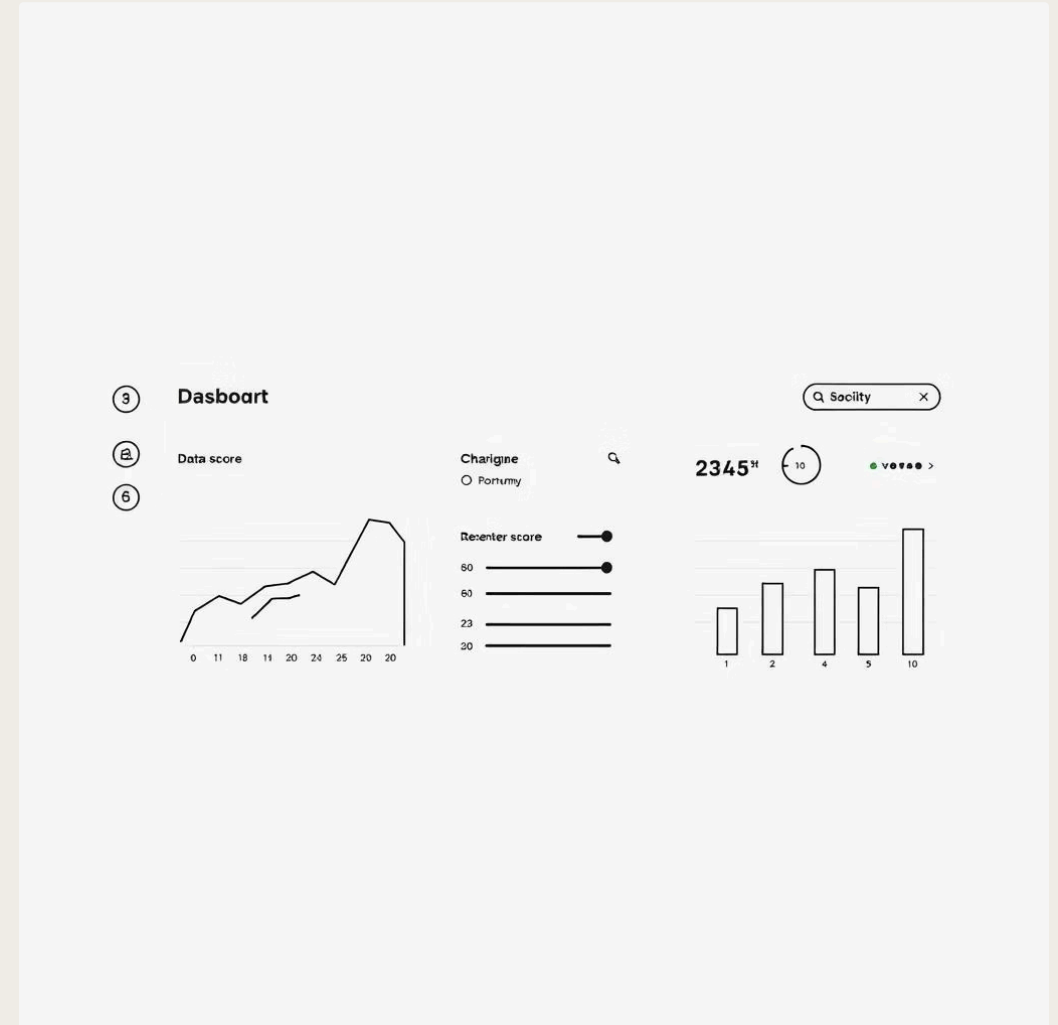
Scans 9 diverse datasets.

Detects Issues

Finds duplicates, missing values, type errors, rule violations.

Generates Quality Score

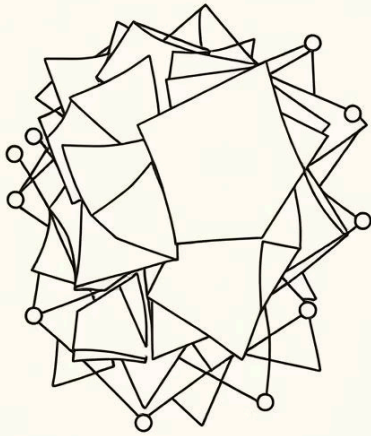Provides a percentage (0-100%).

Creates Config File

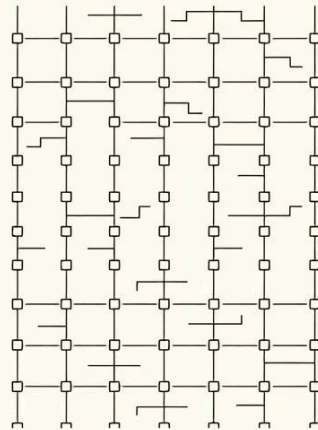Outputs smart `pipeline_config.yaml`.

Stage 2: Smart Data Transformation

Intelligent processing for clean, usable data.

Before

Ater Da

→ **Column Standardization**

Lowercase, consistent underscores.

→ **Advanced Data Cleaning**

Handles duplicates, nulls, invalid values.

→ **Type Conversions**

Accurate date parsing, format adjustments.

→ **Business Calculations**

Automates order totals, data enrichment.

→ **Cleaned CSV Output**

Ready for database loading.

Stage 3: Database Integration

Seamless loading into SQL Server with defined relationships.

Automated Schema
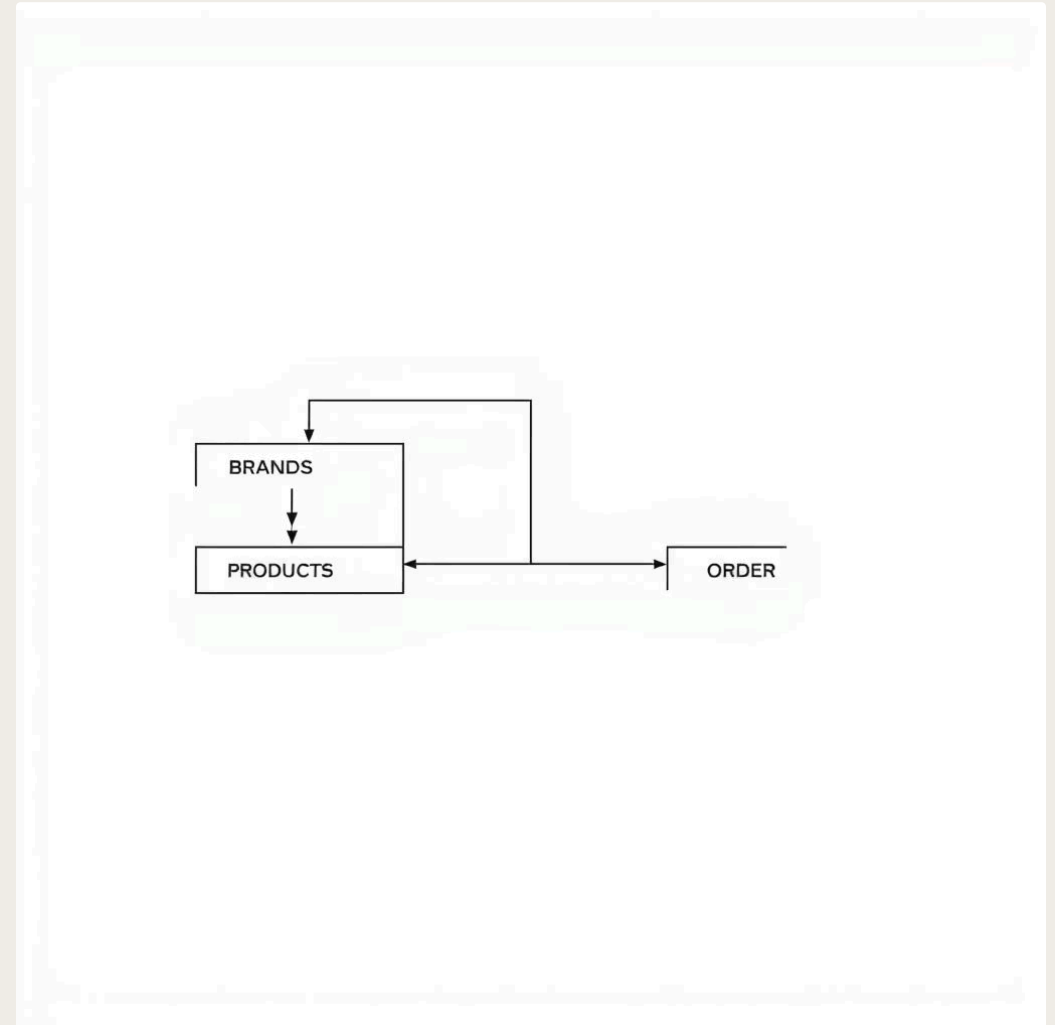
Creates 9 new tables.

Configured FKs

Foreign key relationships set.

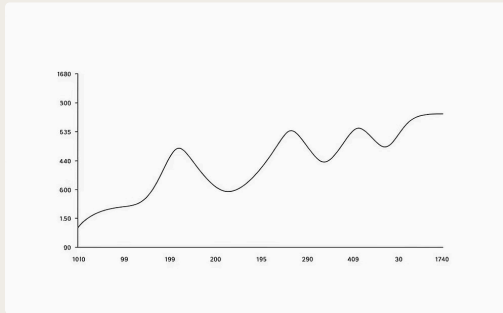## Intelligent Loading

Respects data dependencies.

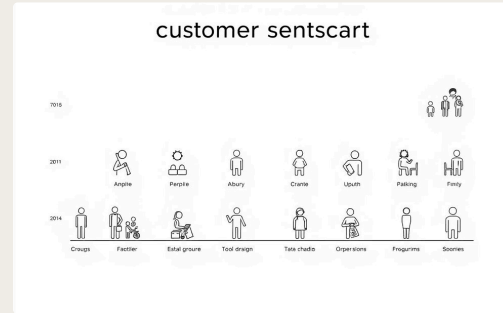## Data Verification

Ensures accuracy, integrity.

Stage 4: Business Intelligence Reports

Actionable insights delivered through pre-built analytics.

Sales Analysis

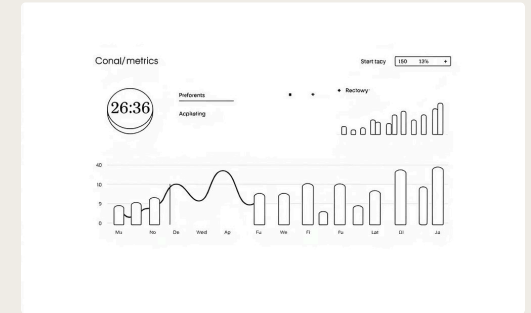Top products, revenue trends.

Customer Insights

Segmentation, spending patterns.

Inventory Management

Stock levels, automated alerts.

Staff Performance

Key operational metrics.

Over 15 pre-built analytical queries ensure comprehensive reporting.

Made with GAMMA

Why This Pipeline?

Unlocking efficiency and accuracy for retail data.

Key Features

Fully Automated

End-to-end workflow.

Intelligent Config

YAML-driven settings.

Modular Design

Flexible, scalable architecture.

Comprehensive Testing

Ensures robust functionality.

Benefits

Save 10+ Hours

Per week on data tasks.

98.999% Accuracy

Reliable data, fewer errors.

Real-Time Insights

Faster, informed decisions.

Easy to Maintain

Streamlined operations.

Technology & Architecture

Leveraging modern tools for a robust and scalable solution.

Python 3.7+ Pandas, NumPy, SQLAlchemy.	**SQL Server 2016+** Efficient ODBC connectivity.	**Matplotlib/Seaborn** Rich data visualizations.

PyTest 85% code coverage.	**YAML Config** Declarative setup.

Real-World Impact

Tangible results showcasing the pipeline's effectiveness.

10,000+

Records Processed

Automatically handled daily.

65% to 95%

Quality Score

Significant improvement achieved.

< 5 Min

Report Generation

Rapid insight delivery.

9

Interactive Visuals

Key metrics at a glance.

1

HTML Dashboard

Centralized data view.