Multi-source Retail Data Integration Hub

ABSTRACT

Retail businesses generate a large amount of data from multiple sales channels such as offline stores and online platforms. To extract meaningful insights, this project applies a complete ETL (Extract–Transform–Load) pipeline using PySpark. Two datasets—offline sales data and online sales data—were cleaned, standardized, transformed, analyzed, and visualized.
The project provides insights on monthly sales trends, category performance, regional revenue distribution, payment behavior, online vs offline comparison, and the relationship between units sold and revenue.

The results help the business make data-driven decisions related to marketing, inventory management, and revenue enhancement.

INTRODUCTION

Retail analytics helps organizations understand sales performance, customer behavior, and revenue drivers. In modern retail environments, data comes from multiple channels, and merging these datasets requires systematic processing.

This project focuses on analyzing two retail datasets using PySpark, a distributed computing engine suitable for large-scale data processing. Through ETL, cleaning, transformation, and visualization, the project extracts clear business insights from raw data.

The goal is to provide actionable insights that help improve decision-making across product categories, regions, and sales channels.

PROBLEM STATEMENT

Retail companies collect sales data from both offline stores and online platforms, but this data is often unorganized, inconsistent, and difficult to analyze. To support better decision-making, there is a need to integrate these datasets, clean and standardize the information, and extract meaningful insights. This project aims to build an ETL pipeline to combine online and offline sales data and generate clear, actionable business insights.

DATASET DESCRIPTION

Two datasets were used:

- Sales Dataset - https://www.kaggle.com/datasets/vinothkannaece/sales-dataset
- Online Sales Dataset - https://www.kaggle.com/datasets/shreyanshverma27/online-sales-dataset-popular-marketplace-data

Sales Dataset - 1:

Contains transactional data from physical retail stores.
Important columns include:

- Product ID
- Sale Date
- Region

- Sales Amount (converted to revenue)
- Quantity Sold
- Payment Method
- Product Category
- Unit Price
- Sales Representative

This dataset helps analyze in-store customer purchasing patterns

Online Sales Dataset - 2:

Represents sales made through e-commerce platforms.
Key columns include:

- Transaction ID
- Date
- Product Category
- Product Name
- Units Sold
- Unit Price
- Total Revenue
- Region
- Payment Method

This dataset helps analyze online consumer behavior and compare it with offline performance.

OBJECTIVES OF THE PROJECT

✓ Integrate and analyze offline and online sales data

✓ Perform ETL and data cleaning using PySpark

✓ Standardize inconsistent column names and formats

✓ Identify category-wise, region-wise, and month-wise sales trends

✓ Compare online vs offline revenue performance

✓ Understand customer payment preferences

✓ Visualize insights for better interpretation

✓ Support business decision-making through data-driven insights

TOOLS & TECHNOLOGIES USED

- Python
- PySpark (Spark DataFrames)
- Matplotlib
- Pandas
- PyCharm IDE
- CSV files (as input and cleaned output)

ETL PROCESS

Extract

- Loaded raw CSV files (offline and online datasets) into PySpark.
- Automatically inferred schema and converted data into DataFrames.

Transform

Major cleaning and transformation steps:

- **Removed duplicates**

  To ensure accuracy and avoid counting the same transaction twice.

- **Identified missing values**

  Each column was checked for null values to evaluate data completeness.

- **Standardized column names**

  Different naming styles were converted to a uniform format.
  Example:

  "Sales_Amount" → "revenue"

  "Quantity_Sold" → "units_sold"

  "product category" → "product_category"

- **Cleaned payment methods**

  Converted variations like "Card Payment", "card", "credit card" → **card**
  Converted "UPI Payment", "upi" → **upi**

- **Region grouping**

  Mapped regions into global groups:

- North → North America
- South & East → Asia
- West → Europe

Extracted month from date: Used for monthly trend analysis.

Saved cleaned datasets: Converted both DataFrames into:

- **cleaned_sales.csv**
- **cleaned_online_sales.csv**

Load: Loaded cleaned CSVs for transformation, insights generation, and visualization.

DATA TRANSFORMATION SUMMARY

After cleaning, additional transformations were done:

- Month-wise grouping of revenue and units sold
- Category-wise revenue calculations
- Region-wise comparison of online & offline sales
- Payment method grouping
- Total revenue aggregation for channel comparison
- Units sold vs revenue correlation preparation

These transformations made the datasets analysis-ready.

INSIGHTS & INTERPRETATION

Based on the transformed data, the following insights were generated:

Month-wise Category Revenue:
line chart Shows how each product category performs across different months.

Region-wise Revenue : A bar chart was used

A bar chart showing the total revenue contributed by each region group.

Category-wise Revenue:

Simple bar chart comparing revenue from:

- Food
- Electronics
- Clothing
- Furniture

Payment Method Preference:

Pie chart showing the usage percentage of UPI, Card, Cash

Online vs Offline Revenue:

A bar chart comparing the combined revenue.

Units Sold vs Revenue :

Scatter plot showing how units sold relate to total revenue.

CONCLUSION

This project successfully demonstrated how PySpark can handle multi-source retail data to produce valuable insights. A complete ETL pipeline was developed to extract, clean, transform, and load the datasets.

KEY TAKEAWAYS INCLUDE:

- Clothing and Electronics are strong product categories.
- North America is the leading revenue-generating region.
- UPI is the most preferred payment method.
- Offline sales outperform online sales significantly.
- Units sold directly influence revenue, indicating healthy sales behavior.

These insights can help businesses optimize marketing, pricing, inventory, and channel strategies.

PREPARED BY

MAHASWETHA A S

DATA ENGINEERING - DATABRICKS