IE 6750 – Data Warehousing & Integration

E-commerce Buyer Behaviour Analytics

Members:

Vinyas Naidu Karri, Param Madan

1. Executive Summary

To overcome fragmented e-commerce data, this project established an integrated E-Commerce Buyer Behavior Analytics Data Warehouse. Utilizing Talend for ETL, PostgreSQL for storage, and Tableau for visualization, our solution implements a dimensional star schema, featuring Fact_Sales, Fact_Review, and Fact_Payment fact tables, alongside a Type-2 SCD for customer historical tracking. This provides a unified foundation, enabling deep analysis to transform raw data into strategic insights for optimizing the e-commerce buyer journey.

2. Business Problem & Project Definition

- **The Challenge**

  E-commerce data is scattered across disparate operational systems, making comprehensive buyer behavior insights difficult.

- **Key Unanswered Questions**
  - Who are our most valuable customers?
  - Which products perform best?
  - What are the payment reliability trends?
  - How can we enhance the overall customer experience?

- **The Gap**

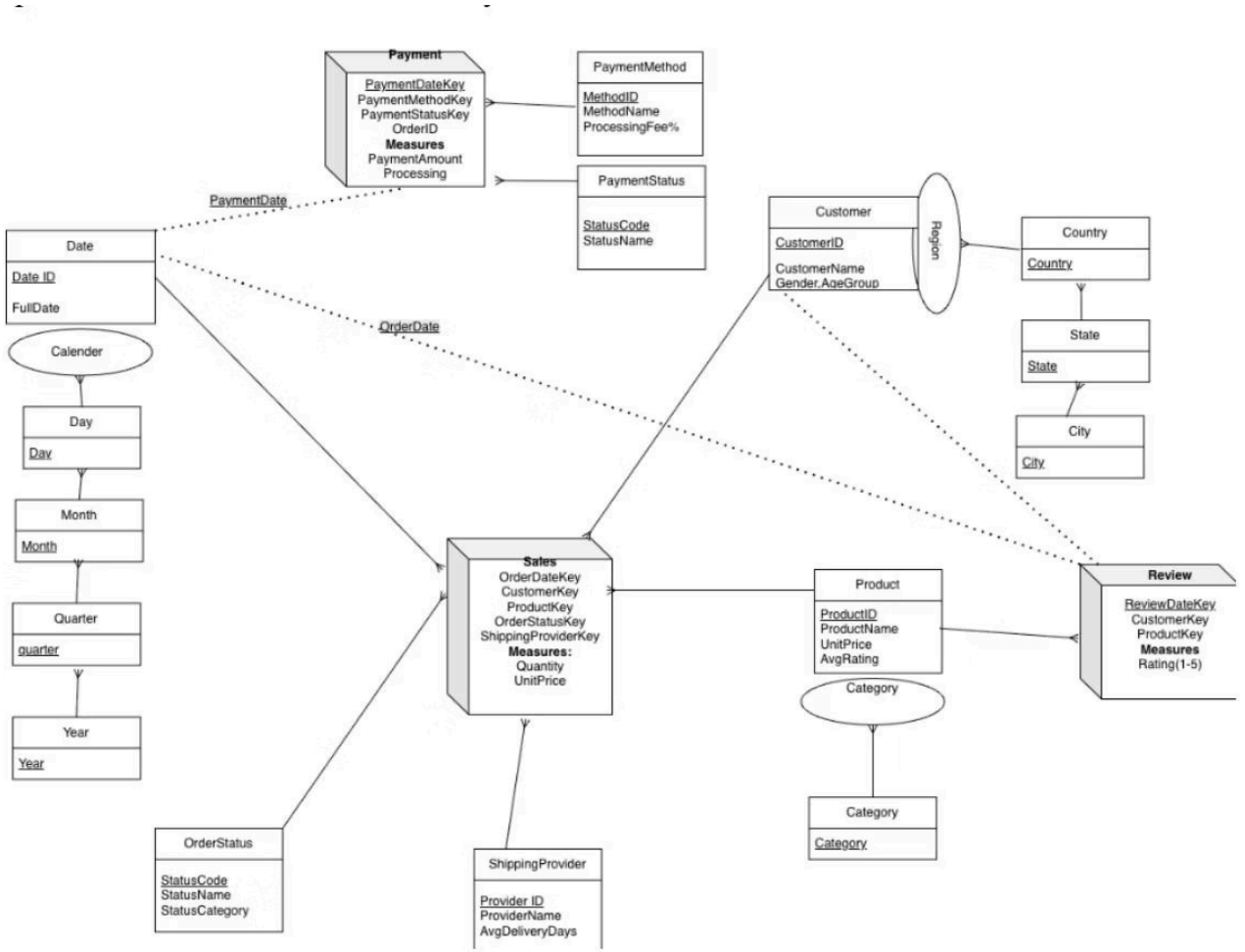
  Operational databases provide raw data, but lack the intelligence needed for strategic decision-making.

- **The Solution**

  Build a unified data warehouse for integrated buyer behavior analytics.

4. Conceptual Data Warehouse Model

The conceptual model follows a **star schema** architecture with three fact tables at the center surrounded by seven dimension tables. This design optimizes the data warehouse for OLAP queries and multidimensional analysis.



4. Logical Data Warehouse Model

The logical data warehouse model builds upon the conceptual design, detailing the implementation of the star schema architecture to optimize for analytical queries while ensuring data integrity and traceability.

Schema Architecture

This model features a **star schema** with 7 distinct dimension tables orbiting 3 central fact tables, designed for efficient multidimensional analysis.

Key Management

All primary keys are implemented as **Surrogate Keys (SK)** for performance and flexibility. Original **Natural Keys (NK)** are preserved within dimension tables to maintain traceability to source systems.

Key Dimensions

- Customer (SCD Type-2)
- Product
- Date
- Payment Method
- Order Status
- Shipping Method
- Review

Fact Tables

- Fact_Sales
- Fact_Payment
- Fact_Review

8. ETL Execution

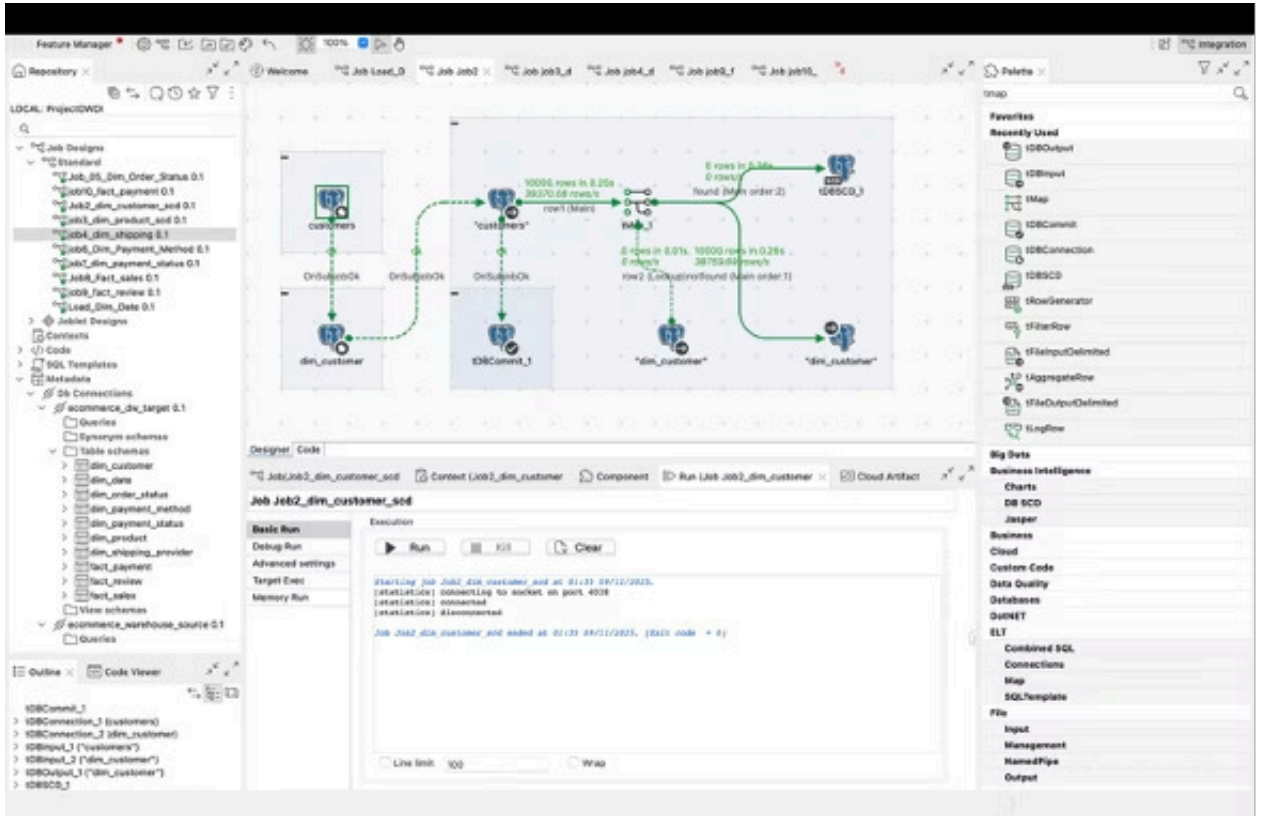
All data warehouse tables both dimension and fact were successfully designed, populated, and validated.

Surrogate keys were implemented correctly for all dimension tables. Both insert and update (incremental) load flows were executed successfully using tMap, tDBInput, and tDBOutput components.

Historical tracking for **SCD Type 2** was implemented in the **Customer Dimension**, ensuring versioning with effective_start_dt, effective_end_dt, and is_current flags. Control flow and commit logic were properly configured using tDBCommit and OnSubjobOK links to ensure atomic and consistent data loads.

DIM Customer ETL Process Details

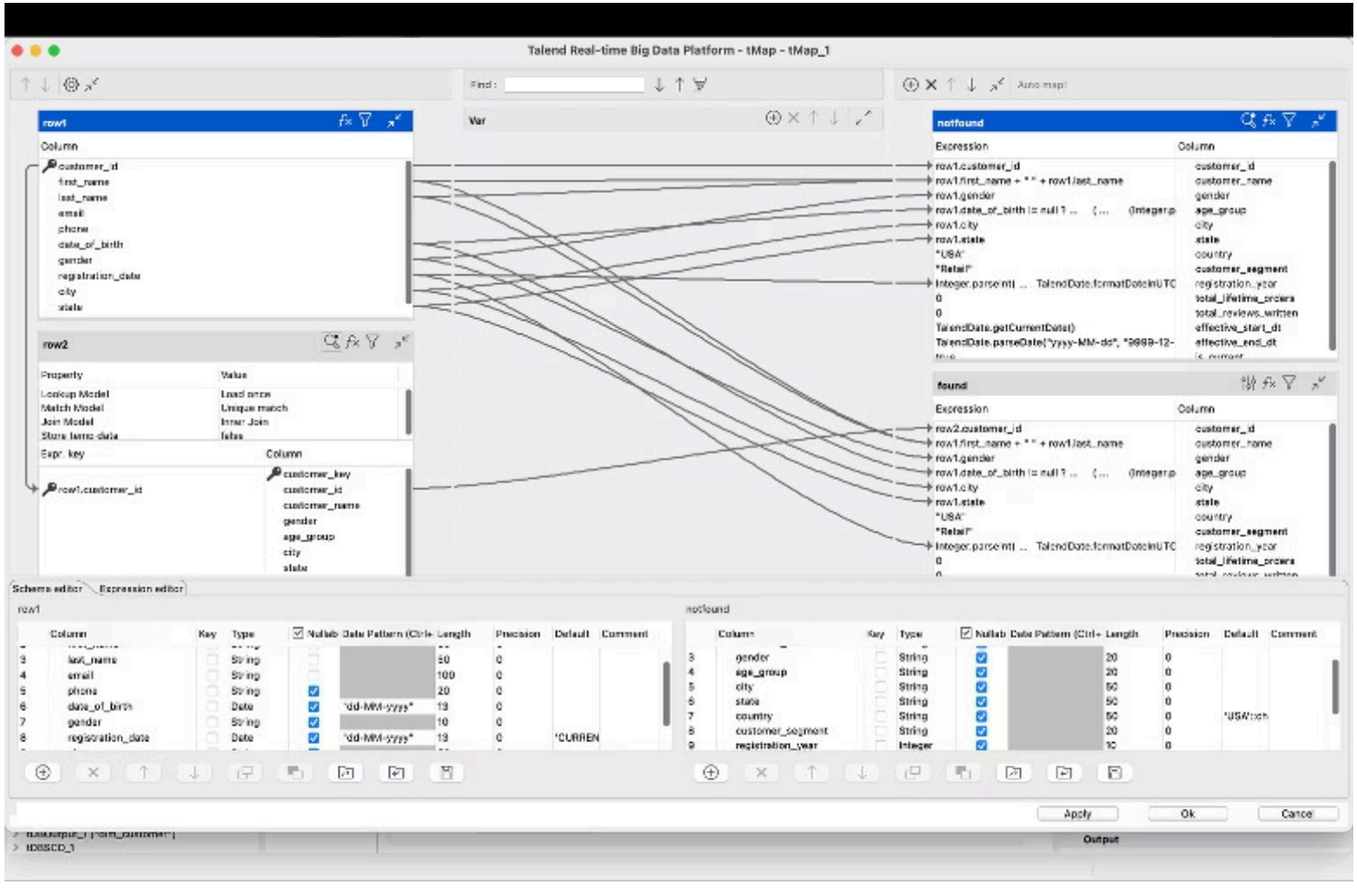
Talend ETL



Data Mapping and Transformation (tMap)

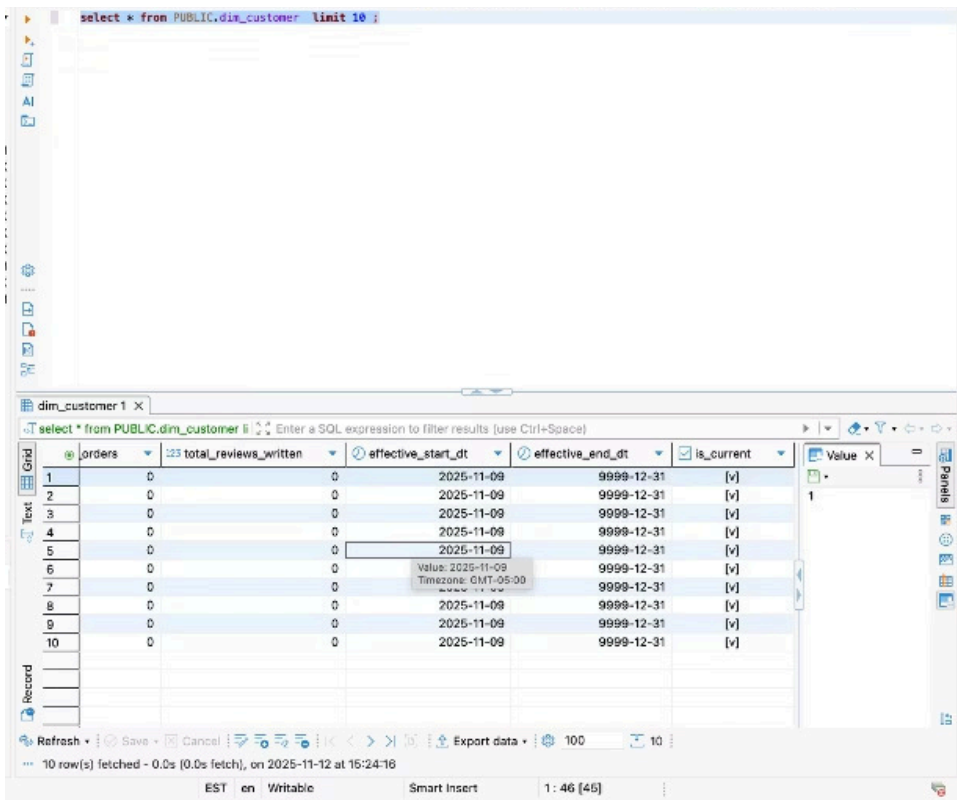
The tMap component was critically used for transforming source data into the required structure for the DIM Customer table.

It handled data type conversions, derived new attributes, and performed lookups against existing dimensions to ensure data integrity and consistency before loading.



SCD Output & Data Loading

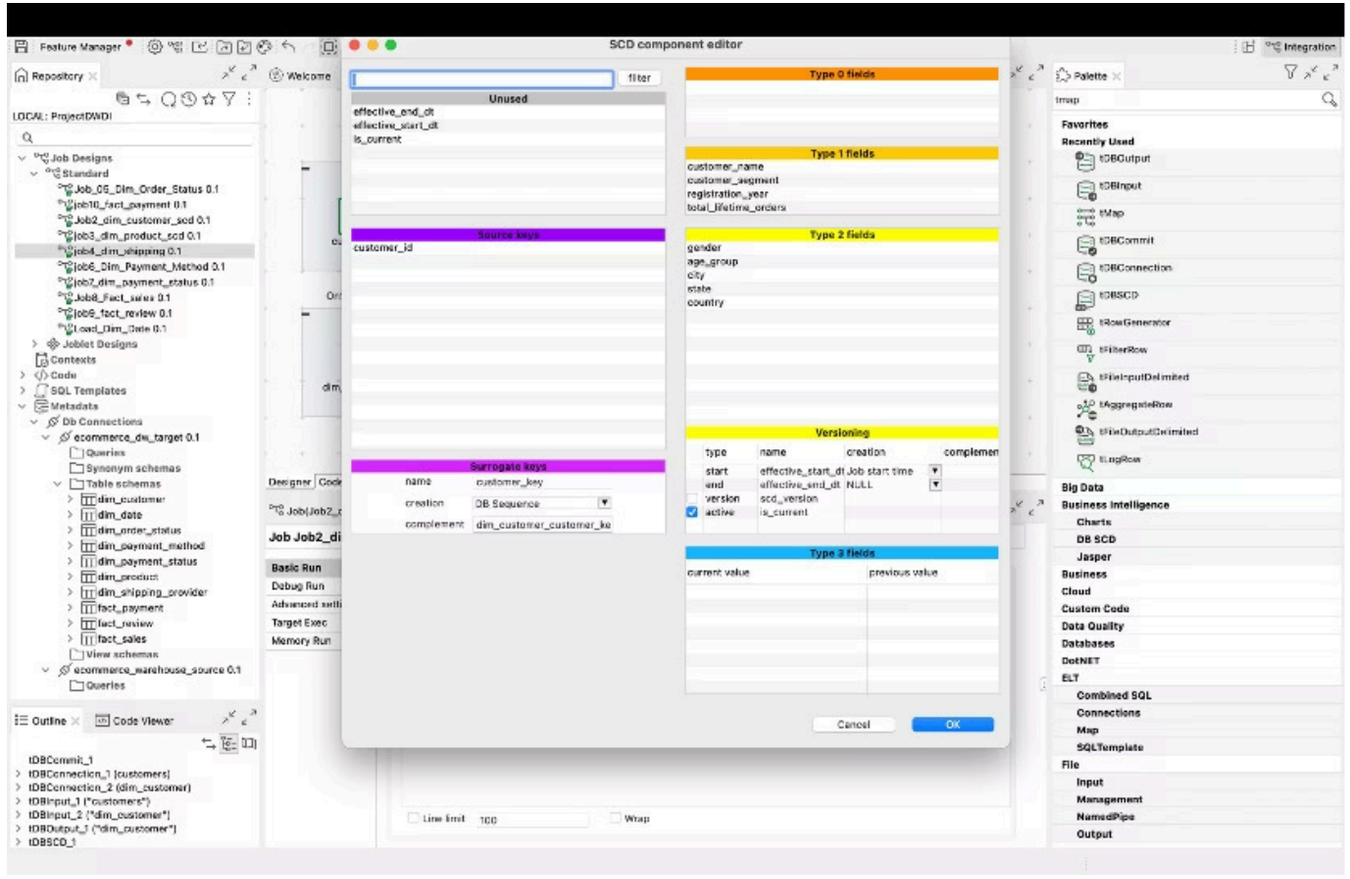
This output stream directly fed the transformed customer data into the DIM Customer table. It managed both initial loads and incremental updates, ensuring that new customer records were inserted and existing ones were updated efficiently. The combination of tDBInput, tMap, and tDBOutput components with commit logic ensures data accuracy, consistency, and traceability for all customer data within the data warehouse.



Slowly Changing Dimensions (SCD Type 2)

SCD Type 2 logic was meticulously applied to the Customer Dimension to track historical changes. This implementation involved:

This approach allows for comprehensive historical analysis of customer data over time.



5.2 ETL Process: Other Dimensions

The ETL process for our remaining dimensions—Product, Date, Payment Method, Order Status, Shipping Method, and Review—follows robust procedures designed for efficiency, integrity, and analytical readiness.

1

Incremental Loading

An incremental load strategy is implemented across these dimensions, utilizing change data capture (CDC) techniques to process only new or modified records. This significantly optimizes ETL job runtimes and resource consumption.

2

Surrogate Key Generation

For each dimension, a unique surrogate key is generated and assigned. These keys are crucial for establishing efficient relationships with fact tables and ensuring data warehouse query performance, distinct from source natural keys.

3

Lookup & Transformation Logic

`tMap` components are extensively used to perform lookups, join data, and apply necessary transformations. This includes standardizing values, converting data types, and handling nulls to prepare data for its dimensional structure.

4

Data Quality Checks

Integrated data quality checks are performed during the loading process, focusing on critical attributes to ensure completeness and accuracy. This includes validating data formats and enforcing business rules to prevent inconsistencies.

9. Analytical Dashboards, KPIs and OLAP Operations

To evaluate buyer behaviour across the e-commerce lifecycle, three Tableau dashboards were developed on top of the dimensional warehouse:

- **Sales & Buyer Behaviour Overview**
- **Payment & Experience Insights**
- **Fulfilment & Logistics**

9.1 Sales & Buyer Behaviour Overview

Recommendations from Dashboard 1

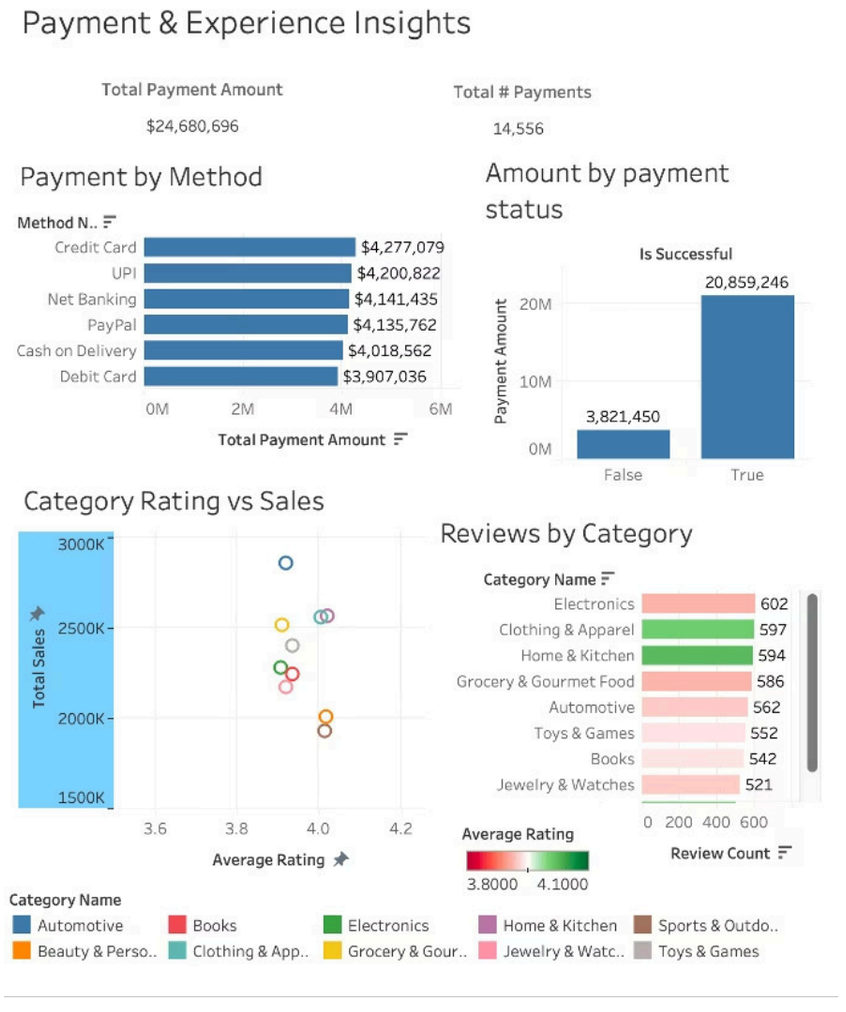
- Prioritize **retention strategies** (loyalty points, personalised recommendations) since returning customers drive a large portion of revenue.
- Use **seasonality patterns** to schedule marketing campaigns and stock replenishment around peak months.

Focus **regional marketing and logistics investments** in the top-performing states, while testing growth strategies in under-penetrated regions

9.2 Payment & Experience Insights

Recommendations from Dashboard 2

- **Stabilize high-value payment methods** (e.g., credit cards, UPI) by monitoring their failure rates and working with gateways to reduce declines.
- For methods with lower usage but high success rates, consider **incentives (cashback/discounts)** to diversify payment mix and reduce dependence on a single provider.
- Use the **Rating vs Sales** view to identify categories that are important but under-performing in experience; prioritize these for product QA, better images/descriptions, or improved after-sales support.
- Encourage more customers to leave reviews (e.g., post-purchase nudges), especially in high-revenue categories where review counts are low, to strengthen social proof.

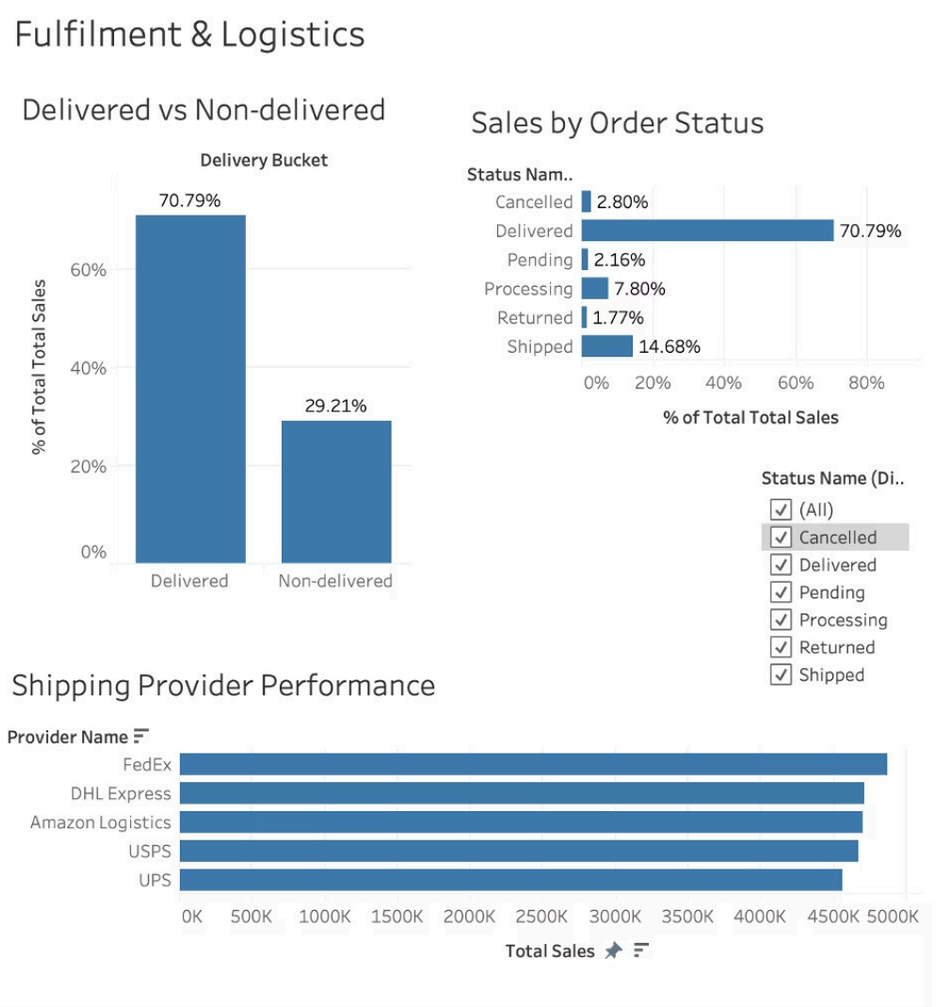


9.3 Fulfilment & Logistics

Recommendations from Dashboard 3

- Monitor the **ratio of Delivered vs Non-delivered revenue** as a key operational KPI. Set thresholds and alerts if non-delivered share grows beyond an acceptable level.
- Investigate categories, regions, or providers associated with **higher cancellation/return rates** to identify root causes (shipping delays, item quality, inaccurate descriptions).

Use **provider-level performance metrics** in contract negotiations and logistics planning—shifting volume toward providers that deliver reliably in key states can improve customer experience and reduce refunds/returns



9.4 Overall Impact of the Dashboards

9.4 Overall Impact of the Dashboards

The three dashboards collectively show that the **dimensional warehouse and ETL process successfully support OLAP-style analysis** across multiple business domains:

- **Sales and customer behaviour** (Fact_Sales + Dim_Customer + Dim_Product + Dim_Date)
- **Payment performance and customer satisfaction** (Fact_Payment + Fact_Review + shared dimensions)
- **Order fulfilment and logistics** (Fact_Sales + Dim_Order_Status + Dim_Shipping_Provider)

KPIs are defined as aggregations directly over fact tables, and each visualization is a concrete implementation of one or more OLAP operations specified in the design section. This confirms that the warehouse is **analytics-ready**, and that the star schema with conformed dimensions is effective for answering complex buyer-behaviour questions that the original transactional systems could not address.

Thank You

Thank you for your attention. We appreciate your time and engagement.