MARKETPLACE PLATFORM

AMALIMART

Project Overview

The marketplace platform operates as a robust B2B/B2C multi-vendor ecosystem, supporting a broad range of users, including Administrators, Vendors, Delivery Partners, and Support Agents. To enable effective data-driven decision-making and maintain operational excellence, there is a critical need for a unified, scalable, and intelligent analytics solution.

This initiative focuses on developing a cloud-native analytics platform that delivers curated, role-specific insights through modern data pipelines, empowering stakeholders with timely and relevant information to inform their decisions.

This project will deliver the first production-ready version of the analytics platform, built using modern data engineering best practices, including data orchestration, observability, governance, and cloud scalability.

Business Objectives

The primary objectives of the analytics platform initiative are as follows:

- **Provide centralized and persona-specific analytics dashboards**
    - Deliver tailored dashboards and reports for different user roles to access key business metrics and operational KPIs relevant to their responsibilities.
- **Enable near real-time visibility for operational teams**
    - Support decision-making and rapid response by processing and surfacing critical data with low latency.
- **Ensure data governance, lineage, and monitoring**
    - Establish strong data quality controls, traceability, and platform observability to maintain trust, compliance, and operational transparency.
- **Establish reusable data models to support future business scaling**
    - Design modular, extensible data models that can be leveraged across departments and use cases as the platform grows in complexity and scale.

Insights and Use Cases

The analytics platform will deliver curated insights tailored to the unique needs of each user group within the marketplace ecosystem. These insights will be accessible via dashboards and APIs, supporting both strategic planning and real-time operational decision-making.

Persona	**Key Questions**
Admin	- Who are the top-performing vendors by GMV this month? - What is the breakdown of COD vs prepaid orders?
Vendor	- What are my best-selling products? - What are my return and cancellation trends?
Support Agent	- Which orders are delayed or failed? - What are common support request patterns by category?
Analyst	- What are our customer acquisition costs and LTV over the last 3 quarters? - Which states have the lowest fulfillment rates?

Cross-Cutting Metrics For Visualization.

The platform will also include comprehensive dashboards with drill-down capabilities across key business dimensions:

Sales & Product Performance

- Sales Overview (Total Sales, Orders, GMV) filterable by day, week, month
- Product Performance (Best sellers, Low performers)
- Return & Cancellation Rates
- Customer Ratings & Feedback Scores

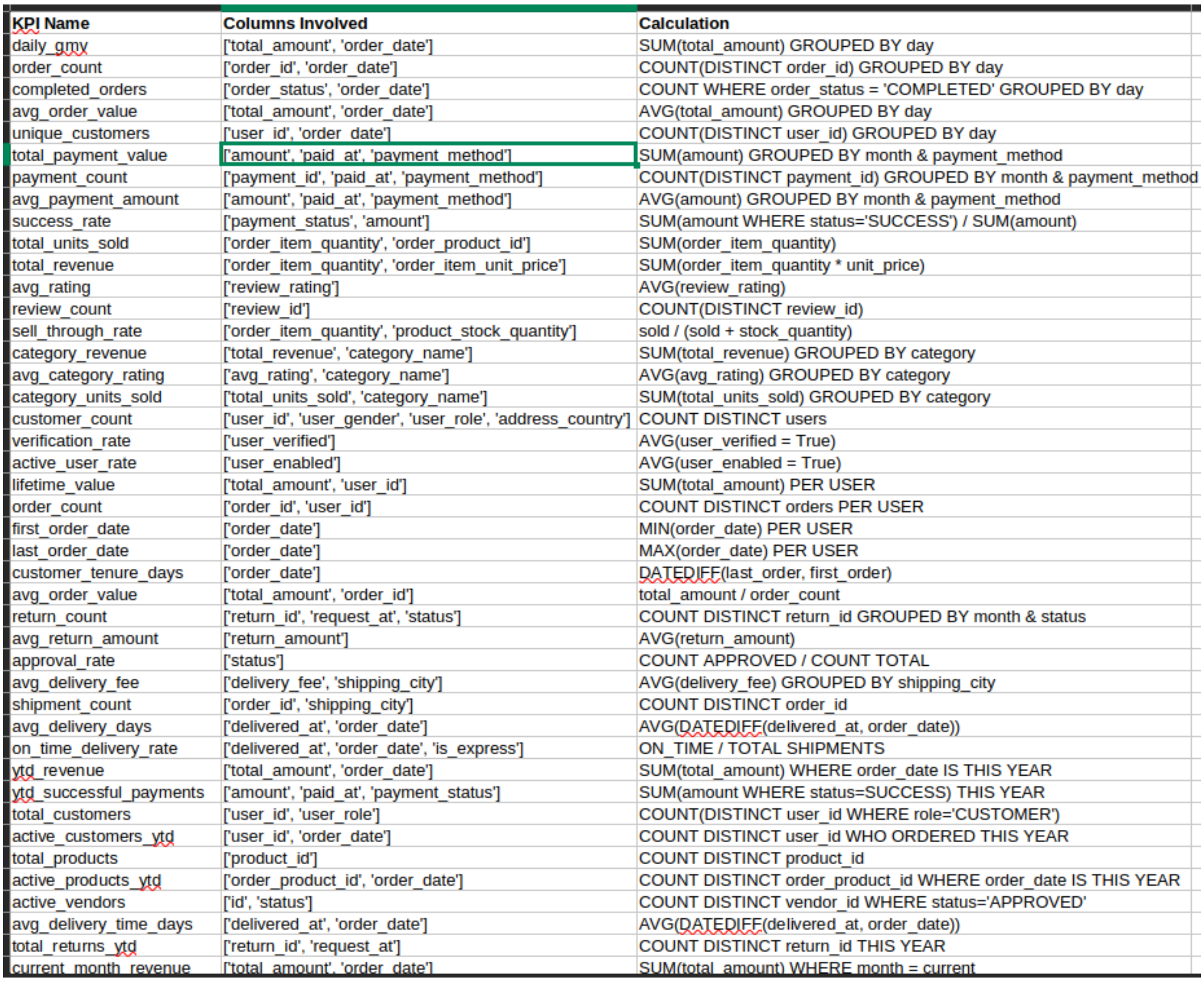
Inventory & Order Management

- Total Products Listed
- Active vs Inactive Products
- Category Distribution
- Out-of-Stock Products
- Order Frequency & Repeat Purchase Behavior

Customer Insights

- Customer Demographic Breakdown segmented by Gender, Age Group, Location, and User Type
- Average Purchase Value
- Customer Retention and Lifetime Value (LTV)

METRICS



KPI Name	Columns Involved	Calculation
daily_gmv	['total_amount', 'order_date']	SUM(total_amount) GROUPED BY day
order_count	['order_id', 'order_date']	COUNT(DISTINCT order_id) GROUPED BY day
completed_orders	['order_status', 'order_date']	COUNT WHERE order_status = 'COMPLETED' GROUPED BY day
avg_order_value	['total_amount', 'order_date']	AVG(total_amount) GROUPED BY day
unique_customers	['user_id', 'order_date']	COUNT(DISTINCT user_id) GROUPED BY day
total_payment_value	['amount', 'paid_at', 'payment_method']	SUM(amount) GROUPED BY month & payment_method
payment_count	['payment_id', 'paid_at', 'payment_method']	COUNT(DISTINCT payment_id) GROUPED BY month & payment_method
avg_payment_amount	['amount', 'paid_at', 'payment_method']	AVG(amount) GROUPED BY month & payment_method
success_rate	['payment_status', 'amount']	SUM(amount WHERE status='SUCCESS') / SUM(amount)
total_units_sold	['order_item_quantity', 'order_product_id']	SUM(order_item_quantity)
total_revenue	['order_item_quantity', 'order_item_unit_price']	SUM(order_item_quantity * unit_price)
avg_rating	['review_rating']	AVG(review_rating)
review_count	['review_id']	COUNT(DISTINCT review_id)
sell_through_rate	['order_item_quantity', 'product_stock_quantity']	sold / (sold + stock_quantity)
category_revenue	['total_revenue', 'category_name']	SUM(total_revenue) GROUPED BY category
avg_category_rating	['avg_rating', 'category_name']	AVG(avg_rating) GROUPED BY category
category_units_sold	['total_units_sold', 'category_name']	SUM(total_units_sold) GROUPED BY category
customer_count	['user_id', 'user_gender', 'user_role', 'address_country']	COUNT DISTINCT users
verification_rate	['user_verified']	AVG(user_verified = True)
active_user_rate	['user_enabled']	AVG(user_enabled = True)
lifetime_value	['total_amount', 'user_id']	SUM(total_amount) PER USER
order_count	['order_id', 'user_id']	COUNT DISTINCT orders PER USER
first_order_date	['order_date']	MIN(order_date) PER USER
last_order_date	['order_date']	MAX(order_date) PER USER
customer_tenure_days	['order_date']	DATEDIFF(last_order, first_order)
avg_order_value	['total_amount', 'order_id']	total_amount / order_count
return_count	['return_id', 'request_at', 'status']	COUNT DISTINCT return_id GROUPED BY month & status
avg_return_amount	['return_amount']	AVG(return_amount)
approval_rate	['status']	COUNT APPROVED / COUNT TOTAL
avg_delivery_fee	['delivery_fee', 'shipping_city']	AVG(delivery_fee) GROUPED BY shipping_city
shipment_count	['order_id', 'shipping_city']	COUNT DISTINCT order_id
avg_delivery_days	['delivered_at', 'order_date']	AVG(DATEDIFF(delivered_at, order_date))
on_time_delivery_rate	['delivered_at', 'order_date', 'is_express']	ON_TIME / TOTAL SHIPMENTS
ytd_revenue	['total_amount', 'order_date']	SUM(total_amount) WHERE order_date IS THIS YEAR
ytd_successful_payments	['amount', 'paid_at', 'payment_status']	SUM(amount WHERE status=SUCCESS) THIS YEAR
total_customers	['user_id', 'user_role']	COUNT(DISTINCT user_id WHERE role='CUSTOMER')
active_customers_ytd	['user_id', 'order_date']	COUNT DISTINCT user_id WHO ORDERED THIS YEAR
total_products	['product_id']	COUNT DISTINCT product_id
active_products_ytd	['order_product_id', 'order_date']	COUNT DISTINCT order_product_id WHERE order_date IS THIS YEAR
active_vendors	['id', 'status']	COUNT DISTINCT vendor_id WHERE status='APPROVED'
avg_delivery_time_days	['delivered_at', 'order_date']	AVG(DATEDIFF(delivered_at, order_date))
total_returns_ytd	['return_id', 'request_at']	COUNT DISTINCT return_id THIS YEAR
current_month_revenue	['total_amount', 'order_date']	SUM(total_amount) WHERE month = current

DATA MODEL

The analytics platform is built on a normalized relational schema that captures the full range of operational and transactional data required to power the marketplace. The data model is designed to support extensibility, scalability, and analytical performance, ensuring consistent integration with both batch and real-time processing pipelines.

The schema comprises multiple domains, each representing a logical entity group within the system:

Core Domains and Entity Groups

- **User & Identity Management:** Captures user profiles, authentication providers, roles, and audit history for system access and behavioral analysis.

- **Product & Catalog Management:** Includes products, variants, categories, images, tags, and business/vendor information.

- **Order Lifecycle:** Represents carts, orders, order items, status history, payments, shipping, returns, and fulfillment metadata.

- **Customer Engagement:** Covers reviews, wishlists, messages, and support ticketing infrastructure for tracking customer interaction and satisfaction.

- **Business Operations:** Handles vendor applications, businesses, stores, representatives, and discount campaigns.

- **Geographic & Demographic Context:** Incorporates address details, regions, cities, and customer segmentation attributes.

Key Characteristics

All tables follow a timestamped, auditable structure with created_at and updated_at fields to support change tracking and temporal analysis.

Referential integrity is enforced through foreign keys and lookup relations to ensure consistency across joins and aggregations.

Enumerated types (enum) are used across the schema to maintain data integrity and facilitate efficient filtering during downstream analysis (e.g., order status, roles, product states).

The model supports **multi-tenancy** and vendor segmentation through clear ownership references (e.g., vendor_id, business_id) for secure data partitioning and access control.

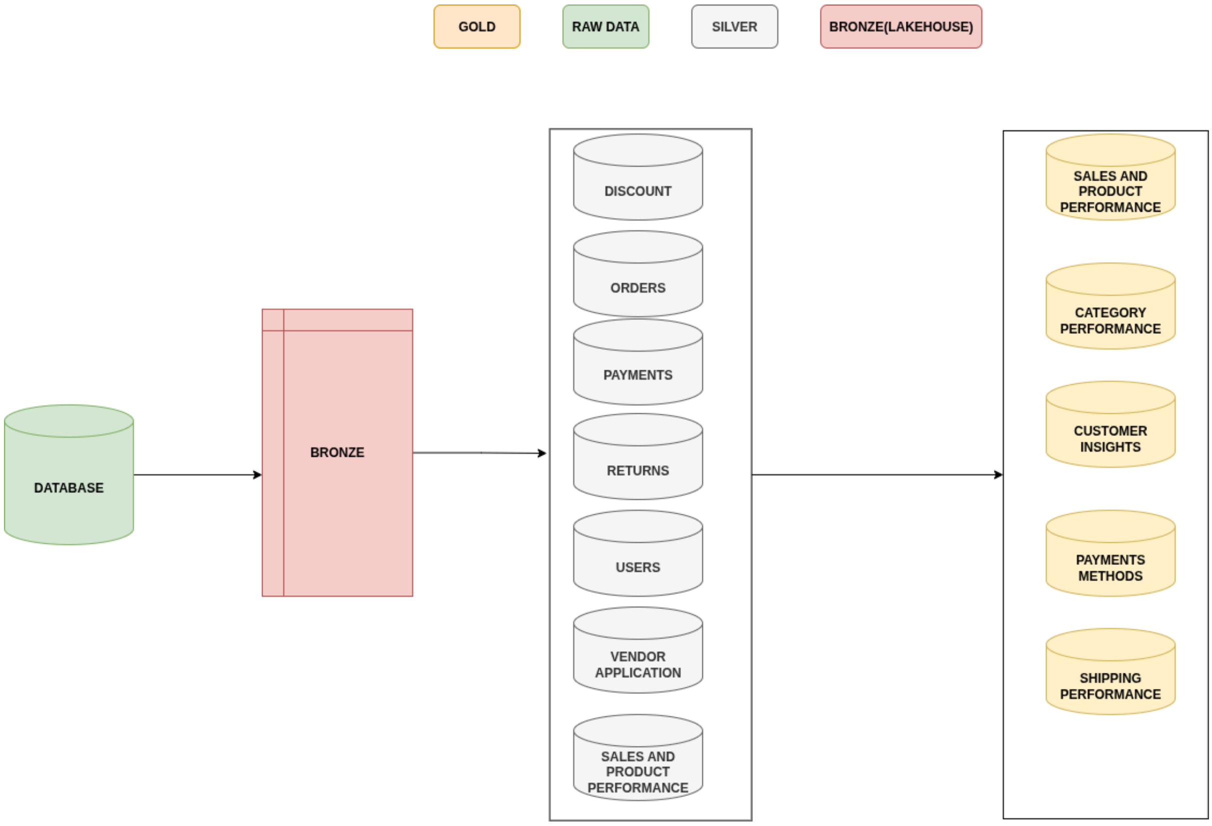
Each domain is optimized for **analytical queries**, making it suitable for dimensional modeling (such as star/snowflake schemas) and data warehouse loading.

This structured schema lays the foundation for high-fidelity data pipelines, enabling transformation into curated data models and analytical views for dashboards, reports, and machine learning use cases.

DATA QUALITY CHECKS

- Row count for every batch of data scheduled (weekly)
- Seasonality-adjusted row count. (Yearly, quarterly)

FLOW DIAGRAM



The data pipeline follows a medallion architecture composed of three core layers: Bronze, Silver, and Gold.

Raw Data Ingestion (Database to Bronze Layer)

The pipeline begins with raw data extracted from the source Database, which may include transactional records from various systems. This raw data is ingested directly into the Bronze layer of the data lakehouse. At this stage, the data is stored in its original form with minimal processing to ensure traceability and historical accuracy.

Data Cleaning and Structuring (Bronze to Silver Layer)

From the Bronze layer, the data is processed and cleaned before being loaded into the Silver layer. In this stage, data is refined, validated, deduplicated, and structured into domain-specific tables such as Discounts, Orders, Payments, Returns, Users, Vendor

Applications, and Sales and Product Performance. This layer serves as the foundation for analytical processing and ensures data quality.

Business-Level Aggregation (Silver to Gold Layer)

The curated data in the Silver layer is further aggregated and transformed into meaningful business insights within the Gold layer. This layer comprises analytical datasets, including **Sales and Product Performance**, **Category Performance**, **Customer Insights**, **Payment Methods**, and **Shipping Performance**. These aggregated outputs support business intelligence, reporting, and data-driven decision-making.

Orchestration