**Big Data Engineer**

**Student name:** Amit Ronen

**Student ID:** 205733975

**Student name:** Naim Ghanim

**Student ID:** 212687537

Mid-semester project:

**warehouse solution for a business**

# 

# **Data Warehouse Solution for E-commerce Platform**

**Business Domain: E-commerce Retail Platform**

We focused on an online retail platform that sells products across multiple channels (website, mobile app, and marketplace partnerships).

**Data Sources**

**1. Streaming Data Source: User Activity Events**

* Real-time clickstream data from website and mobile app
* User interactions (page views, product clicks, add-to-cart, etc.)
* Session tracking and user journey analysis

Implemented using event-driven architecture where user interactions trigger events.

Events include page views, product detail views, searches, add-to-cart actions, wishlist additions, etc.

**2. Late Arrival Source: Third-party Marketplace Sales**

* Sales data from partner marketplaces (Amazon, eBay, etc.)
* Data arrives up to 48 hours after transaction time
* Includes order details, shipping information, and payment status

Data arrives via API integrations or batch file imports from marketplace partners.

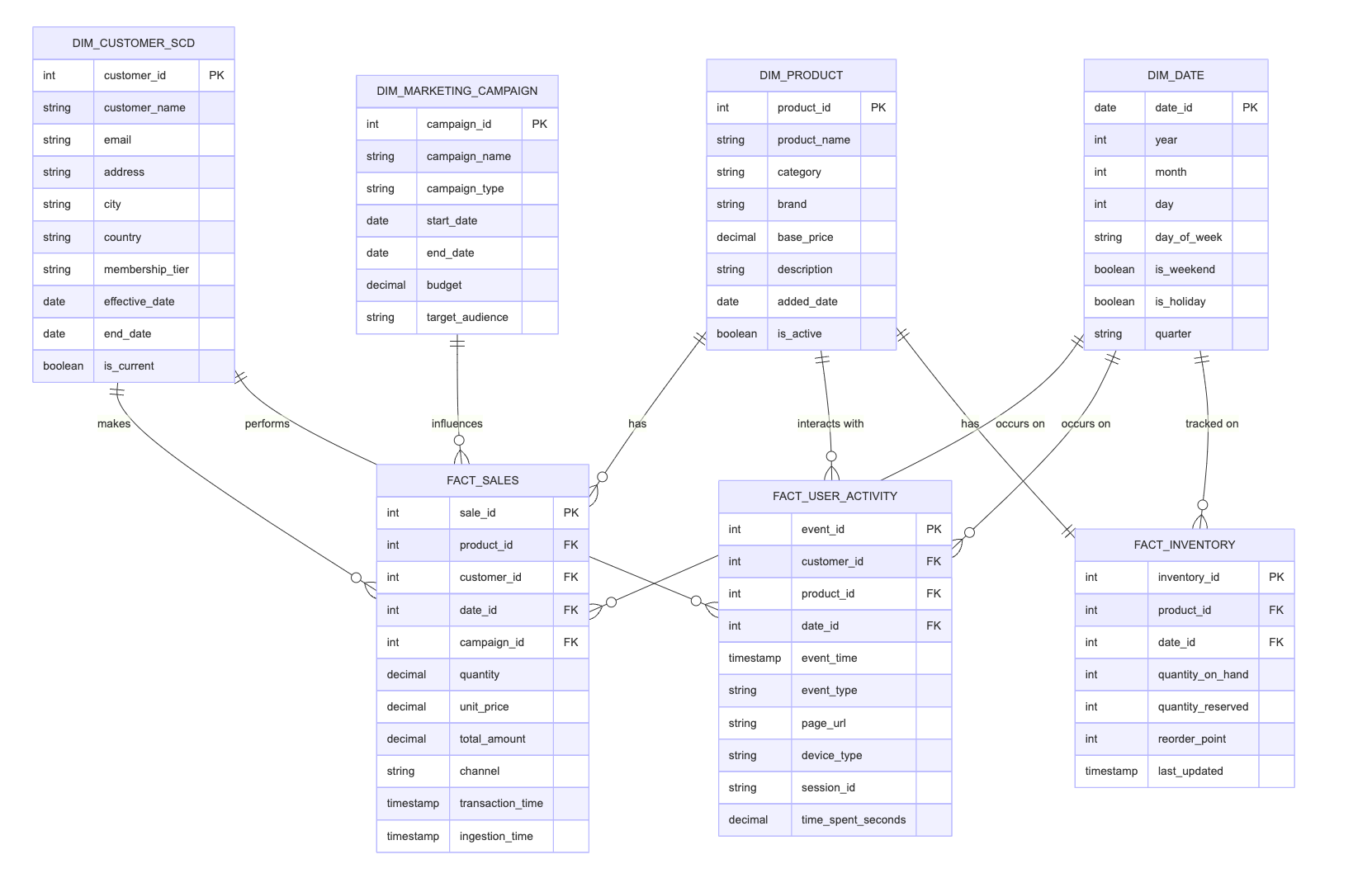
Late arrivals occur due to marketplace settlement cycles, manual reconciliation, and third-party processing delays.

**3. Additional Data Sources:**

* Product Catalog API (updated daily)
  + Contains complete product attributes: names, descriptions, images, pricing, categories, etc.
* Customer Account Database (batch extraction)
  + Contains customer profiles, account settings, preferences, and contact information
* Marketing Campaign Management System (periodic exports)
  + Exports campaign definitions, targeting criteria, budget allocations, and creative assets

**A diagram of a company

Description automatically generatedData Model Design- Bronze**

**Data Model Design- Silver**

### Key Elements of the Data Model:

1. **Static Dimension Table:**

* **DIM\_MARKETING\_CAMPAIGN**:
  + Reference data for marketing initiatives
  + Contains campaign metadata, budget allocations, and targeting parameters

1. **Type 2 Slowly Changing Dimension:**

* **DIM\_CUSTOMER\_SCD**:
  + Implements Type 2 SCD methodology to track historical changes to customer data
  + Every change creates a new record with updated values while preserving the old record

1. **Fact Tables:**

* **FACT\_SALES**:
  + Records all transaction details across all sales channels
  + Contains measures like quantity, unit price, and total amount
* **FACT\_USER\_ACTIVITY**:
  + Captures user behavior data from web and mobile applications
  + Records event types, timestamps, session information, and contextual metadata
* **FACT\_INVENTORY**:
  + Tracks product inventory levels over time
  + Contains point-in-time snapshots of quantity on hand, reserved stock, and reorder points

## Medallion Architecture

### Bronze Layer (Raw Data)

1. **Streaming User Events Pipeline:**
   * Schema Registry for event validation
   * Raw storage with minimal transformations
   * Event-time watermarking for processing
2. **Marketplace Sales with Late Arrivals:** 
   * Batch import with 48-hour window
   * Event timestamp preservation
   * Deduplication logic
3. **Additional Data Sources:** 
   * API connectors for product catalog
   * Database CDC for customer information
   * File imports for marketing campaigns

### Silver Layer (Standardized & Enriched)

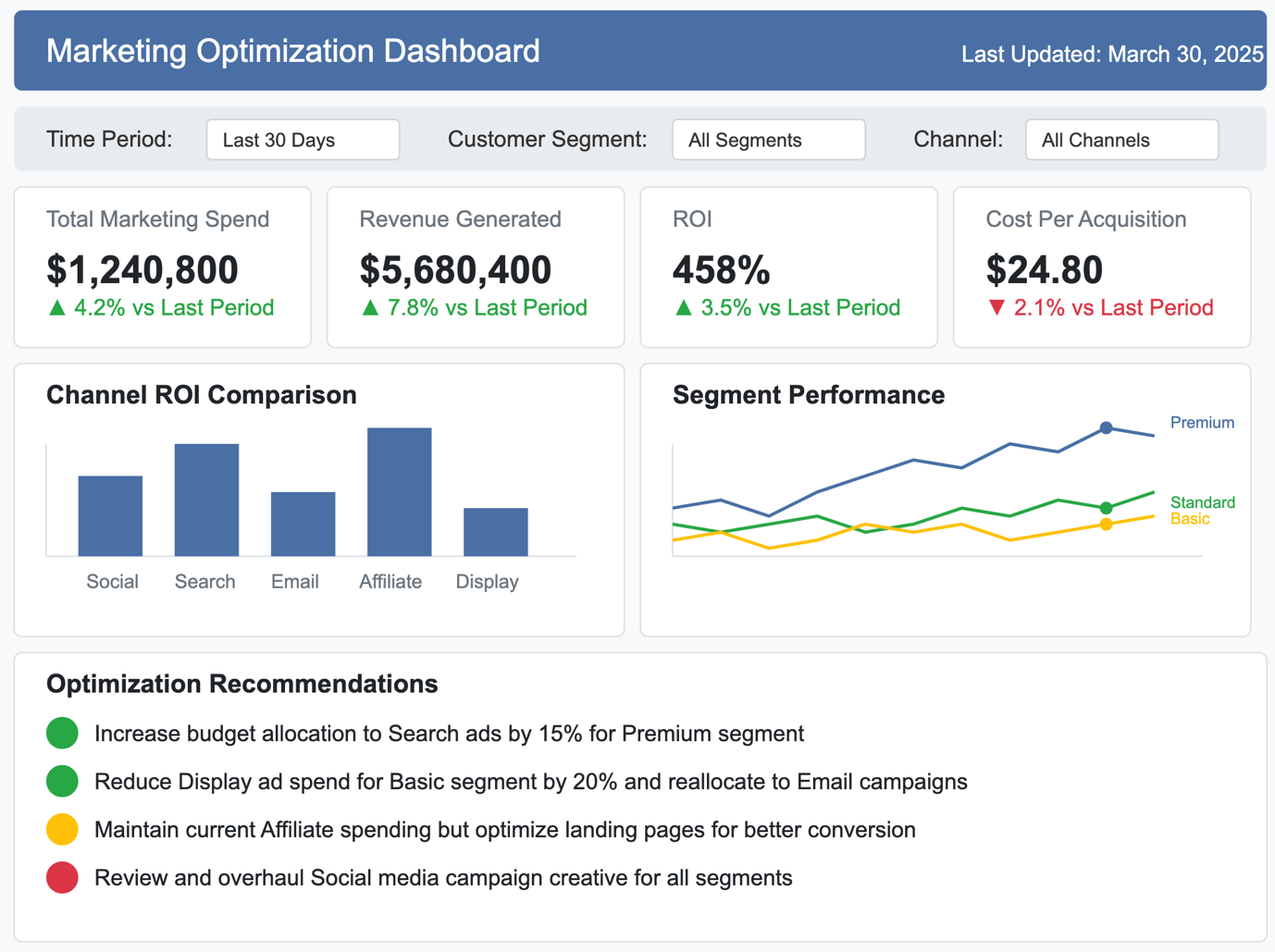
1. **Standardization Processes:** 
   * Data type normalization
   * Null handling and validation
   * Deduplication
2. **Join & Enrichment Logic:** 
   * Lookup enrichment for product details
   * Customer profile enhancement
   * Geographic enrichment
   * Currency standardization
   * Market segment tagging

### Gold Layer (Business Ready)

1. **Business Metrics:**
   * Sales performance by channel
   * Product category performance
   * Customer segmentation metrics
   * Marketing campaign effectiveness
2. **Aggregation Strategies:** 
   * Daily/weekly/monthly rollups
   * Customer lifetime value calculations
   * Product performance scoring
   * User journey funnel metrics
3. **End User Views:** 
   * Executive dashboard
   * Marketing analytics
   * Inventory management
   * Customer insights

## Business Analytics Requirements

### Key Business Question

"How can we optimize our marketing spend across channels based on customer segments and their purchase behavior?"

### Dashboard Elements

1. **Marketing Channel ROI**
   * Spend vs. Revenue by channel
   * Cost per acquisition trends
   * Attribution modeling insights
2. **Customer Segment Performance** 
   * Segment-specific response rates
   * Channel preference by segment
   * Customer lifetime value by acquisition source
3. **Campaign Effectiveness** 
   * Campaign ROI comparison
   * Creative performance metrics
   * A/B test results visualization

### Required Tables for Dashboard

* FACT\_SALES joined with DIM\_CUSTOMER\_SCD and DIM\_MARKETING\_CAMPAIGN
* Aggregated view of FACT\_USER\_ACTIVITY for funnel analysis
* Historical trends from both fact tables with date dimension

## Machine Learning Requirements

### Business Problem: Product Recommendation Engine

Improve customer engagement and average order value by implementing a personalized product recommendation system based on browsing behavior and purchase history:

* **Problem Statement**: Improve customer engagement and average order value through personalized product recommendations
* **Business Impact**:
  + Increases cross-sell and up-sell opportunities
  + Enhances customer experience through personalization
* **ML Approach**:
  + Collaborative filtering to identify patterns across user behaviors
  + Content-based recommendations using product attributes

### Feature Table Design

### Feature Table Design Explanation:

* **Primary Keys**:
  + customer\_id and product\_id form a composite key
  + Represents the interaction between a specific customer and product
* **Frequency Features**:
  + Quantifies interaction intensity through count metrics
  + Multiple time windows (7d, 30d, 90d) to capture both recent and established patterns
  + Distinguishes between passive (views) and active (purchases) interactions
  + Cart abandonment tracking identifies products of interest not purchased
  + Provides strong signals for collaborative filtering algorithms
* **Product Interaction Features**:
  + Measures engagement depth through time spent and interaction types
  + Image view ratio indicates visual interest in the product
* **Category Affinity Features**:
  + Extends product-specific interactions to category-level preferences
  + Category purchase count establishes demonstrated preference
* **Customer Features**:
  + Provides customer context for better recommendation relevance
  + Tenure indicates familiarity with the platform
* **Product Features**:
  + Adds product context independent of specific customer
  + Popularity score balances personal relevance with general appeal
* **Target Variables**:
  + Defines the prediction objectives for supervised learning
  + Binary purchase flags for different time horizons (7d, 30d)

### Feature Data Flow

### A diagram of a diagram Description automatically generated with medium confidenceFeature Data By Groups

**Why We Structured the Future Data This Way:**

1. **Enhanced Clarity and Usability**  
   Categorizing future data (Events, Symptoms, Locations) ensures that insights are organized, easy to interpret, and actionable.
2. **Modular and Scalable Design**  
   This structure allows us to seamlessly add or modify categories as the system evolves, without disrupting the overall framework.
3. **Deeper Insight and Analysis**Grouping related future elements helps uncover meaningful patterns and relationships, supporting more accurate predictions and smarter decisions.

### SQL Logic

**Our SQL Logic – How It Works**

1. **Clean and Modular**  
   We broke the logic into smaller SQL blocks like recency, frequency, and intensity, which makes the whole system easier to understand, maintain, and scale.
2. **Smart Data Transformation**  
   Raw data from multiple input tables gets processed using SQL to create meaningful features, like how often something happens or when it last occurred.
3. **One Unified Output**  
   All the features are pulled together into a single, clean output table that’s ready to be used for training models or making predictions.

### ML Pipeline Design

### Required Tables for ML Model

* FACT\_USER\_ACTIVITY for interaction data
* FACT\_SALES for purchase history
* DIM\_PRODUCT for product attributes
* DIM\_CUSTOMER\_SCD for customer profiles
* Derived feature tables aggregating behavioral patterns