Data Engineering Blueprint

Author: Abhilash Mishra, Senior Data Analyst

Purpose:

This document aims to outline the essential requirements for crafting and integrating a segment of the data architecture that addresses core business challenges. These guidelines are derived from a comprehensive analysis of customer interactions and transaction data, aiming to design a system that enables swift data retrieval and insightful analysis.

Integration Approach:

The redesigned data architecture will unify information from several sources to form a holistic view of customer activities. The integration will cover:

Customer Identification: This includes tracking unique customer IDs along with personal details like names, ages, and geographical locations.

Transaction Tracking: Incorporates unique identifiers for transactions, detailed item descriptions, and total spending per transaction.

Logistics Monitoring: Captures unique shipping identifiers along with the delivery status, indicating whether items are in transit or have reached their destination.

Ensuring Data Quality:

To uphold data quality standards, the following protocols must be enforced:

Consistency in Identifiers: The Customer_ID must be uniformly applied across all datasets, acting as the key link between customer, order, and shipping information.

Correct Relationship Mapping: Ensure that Order_ID and Shipping_ID accurately correspond to their respective Customer_ID entries, establishing reliable links between transactions and customers.

Resolution of Discrepancies: Proactively manage issues such as missing transaction details or misalignments between order and delivery data.

Operational Guidelines:

Product Details in Orders: Maintain product details within the transaction records to support detailed insights into purchasing behavior, even though separating this data has been considered. This approach minimizes disruptions to current reporting workflows.

Accurate Transaction Amounts: Ensure the recorded total spending per transaction includes all costs, accounting for taxes and any discounts applied.

Optimization Strategies:

The data architecture should be structured to facilitate high-speed data retrieval, especially for commonly accessed reports such as those analyzing customer profiles and buying patterns.

Data Verification Measures:

To maintain the integrity of the data, adhere to these validation steps:

Avoid Null Entries: All key fields must be complete, with no missing values in essential linking columns.

Value Range Checks: Confirm that numerical fields, like spending amounts and ages, fall within logical and acceptable ranges.

Cross-Table Consistency: Validate that all data remains consistent across different tables before merging into the final comprehensive dataset.

Urgency: Critical

This data architecture will serve as the backbone for all analytical outputs, driving business insights including customer spending behaviours, delivery status evaluations, and identifying popular products across different customer segments. As a Data Engineer, your responsibility is to implement this structure while adhering strictly to the outlined quality and performance standards.

Note: This document serves as a directive for developing and refining our data models to align with the organization's strategic and operational needs.