Star Schema Design Assignment

Accreditation Assignment

MIS 633 Business Intelligence and Data Integration
Professor Joseph Morabito

Shaowen (Carol) Weng
2025.4.3

Business Process Identification

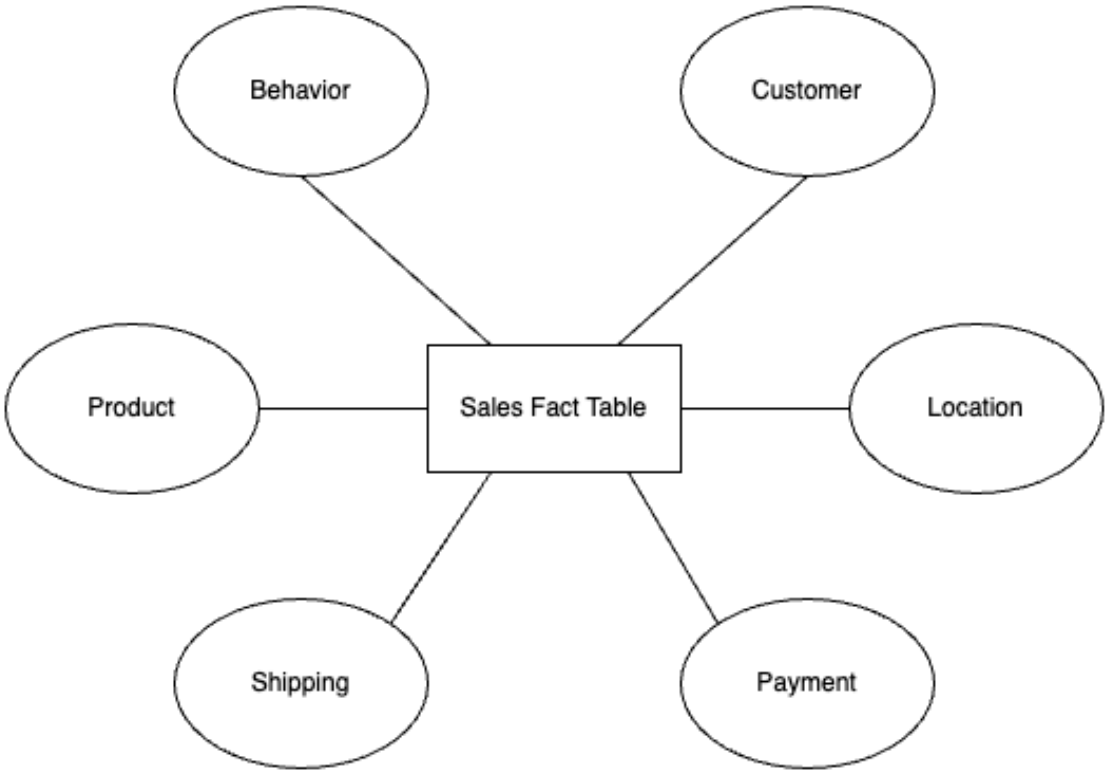
This is a clothing company and we need a multi-dimensional model developed to help effectively manage customers' data and see what more information is needed to better support the model and future analysis.

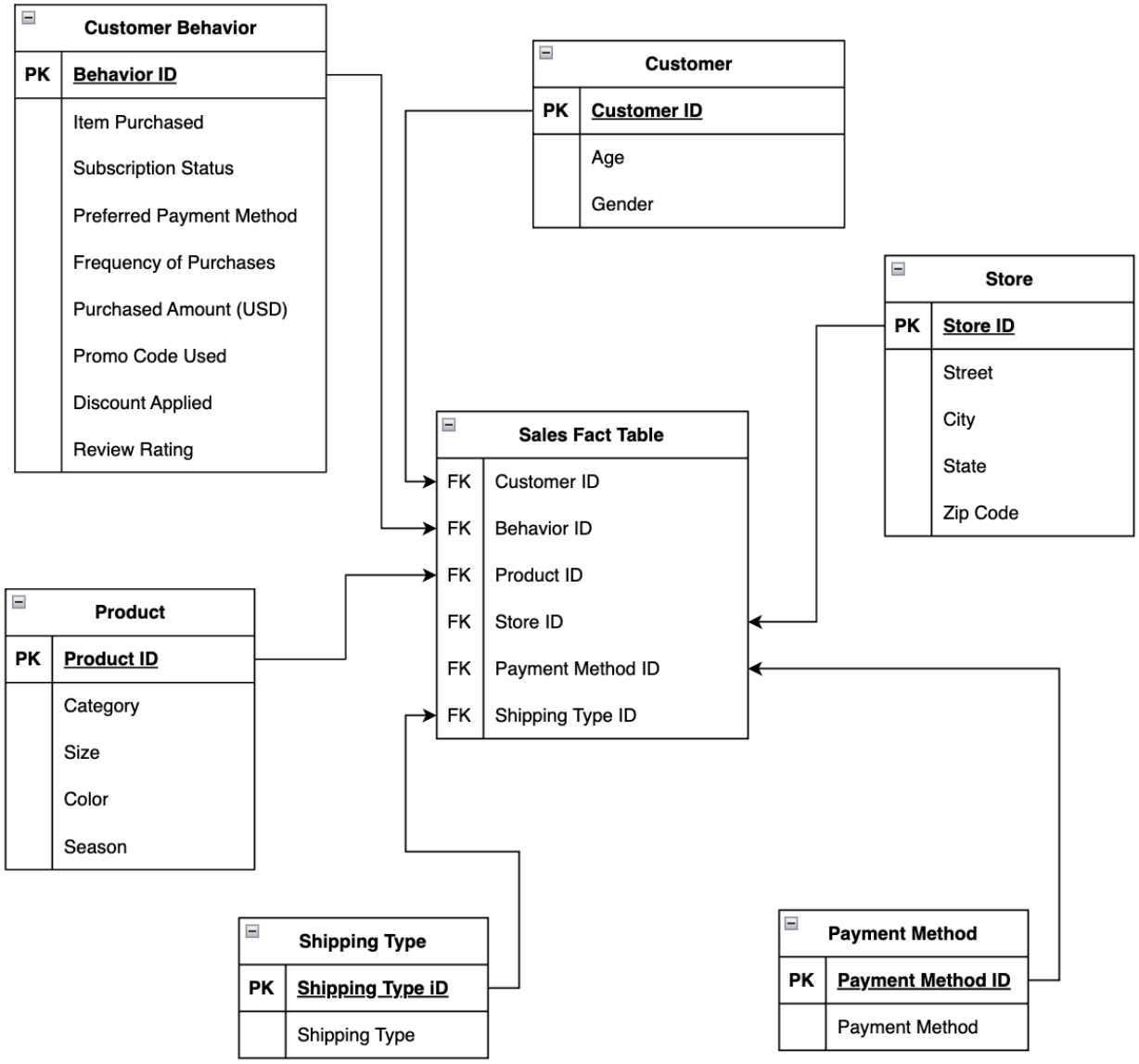
Access of Data

The data is accessed through Kaggle https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset. However, the columns are limited and do not specify which company the data is from.

Dimensional Model Design

High-Level Model Diagram




Customer Behavior
PK
Behavior ID
Item Purchased
Subscription Status
Preferred Payment Method
Frequency of Purchases
Purchased Amount (USD)
Promo Code Used
Discount Applied
Review Rating
Customer
PK
Customer ID
Age
Gender
Store
PK
Store ID
Street
City
State
Zip Code
Sales Fact Table
FK
Customer ID
FK
Behavior ID
FK
Product ID
FK
Store ID
FK
Payment Method ID
FK
Shipping Type ID
Product
PK
Product ID
Category
Size
Color
Season
Shipping Type
PK
Shipping Type iD
Shipping Type
Payment Method
PK
Payment Method ID
Payment Method

Fact Table

- The Sales Fact Table is at the center of the dimensional model which captures individual purchase transactions.

Keys

- In the sales fact table, I have foreign keys as composite primary keys, which have the information of Customer, Customer Behavior, Product, Store, Payment Method, and Shipping Type that is linked to its own dimension table.
- This is designed where a fact table can be joined with any dimension to analyze from different perspectives.

Selected Attributes

- Other than the foreign keys in the fact table which are the primary keys from the dimension tables, are included in the fact table. These are important attributes when doing analysis and business intelligence reporting to find patterns and trends.

Grain

- The grain of this fact table is at the individual purchase level, each row represents a single item purchased by a customer.
- This provides flexibility for future analysis while still being able to support the aggregation at a higher level.

Dimension Tables

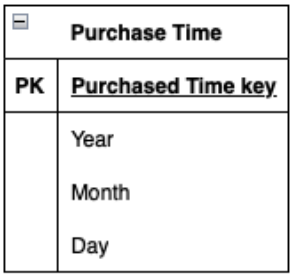
Keys

- The Primary Key (PK) for all dimensions are surrogate keys rather than business keys.

Selected Attributes

- The Customer dimension table contains demographics of Age and Gender.
- Behavioral attributes include Behavior ID, Item Purchase, Subscription Status, Preferred Payment Method, Frequency of Purchases, Purchased Amount, Promo Code Used, Discount Applied, and Review Rating.
- The Product dimension table includes Product ID, Category, Size, and Season. I placed season in this dimension table instead of creating a separate time dimension because it appears to be a product attribute rather than a transaction date in this dataset.

- The data only included state for Store but street, city, and zip code are added in the dimension table to make the locations more specific which is better for analysis.
- Payment Method and Shipping Type are small dimensions that contain lookup values for these attributes. These are separated into their own dimension rather than in the fact table because the model and be more flexible, for example, if we want to analyze review ratings for Express Shipping.
- Time is not included in this data but it would be recommended to include the Purchase Date and Time information for each transaction.

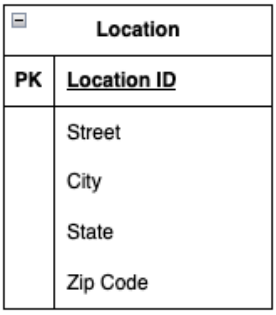


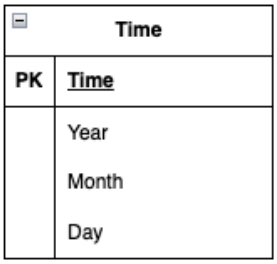
Embedded hierarchies

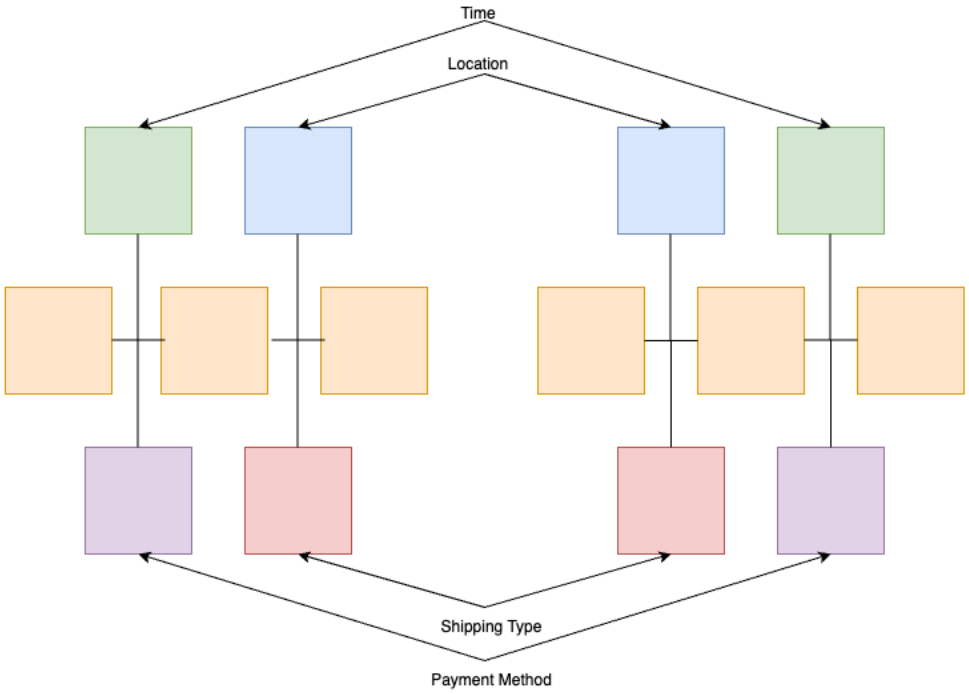
- There are no embedded hierarchies in this model.

Conformed dimensions

- The Location, Time, Shipping Type, and Payment Method are the conformed dimensions because they could be shared across other fact tables such as Return or Exchange and Warehouse Inventory.
- The Location dimension can be shared with other fact tables that need location such as store, warehouse inventory, and shipping.
- Time can be used to time Return or Exchange and Sales as in this design.







Bridge Tables

- Current data only had one category for products, but when we get a more complete dataset, one product might belong to multiple categories. A bridge table can be created when a product has multiple categories. For example, a product can be both “Summer” and “Casual”.
- When we get more data regarding customer preferences, a bridge table can be created when the customers have multiple preferences. For example, if a customer prefers size S and color teal.