1. Find a Dataset

Dataset Choice:

- Title: "eCommerce Behavior Data"
- Kaggle Link: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store/data?select=2019-Nov.csv

Why This Dataset:

- The dataset contains rich information that can be separated into different tables such as Products, Customers, Orders, Reviews, and Categories.
- It's a real-world, practical dataset that allows for a variety of interesting queries and database operations.

2. Define a Scenario

Scenario Context:

- The database will be used for an online eCommerce platform. It will help in managing product listings, customer information, order tracking, and product reviews.

Entities and Use Cases:

1. **Product**

- Attributes: Product ID (PK), Name, Category ID (FK), Price, Description, Stock Quantity.
- Use Case: Each product listed on the platform will have detailed information including its category, pricing, and availability.

2. **Customer**

- Attributes: Customer ID (PK), Name, Email, Address, Registration Date.
- Use Case: Each customer's personal and contact information is stored for order processing and marketing purposes.

3. **Order**

- Attributes: Order ID (PK), Customer ID (FK), Order Date, Total Amount, Shipping Address.

- Use Case: Tracks each order placed by customers, including the purchase details and shipping information.

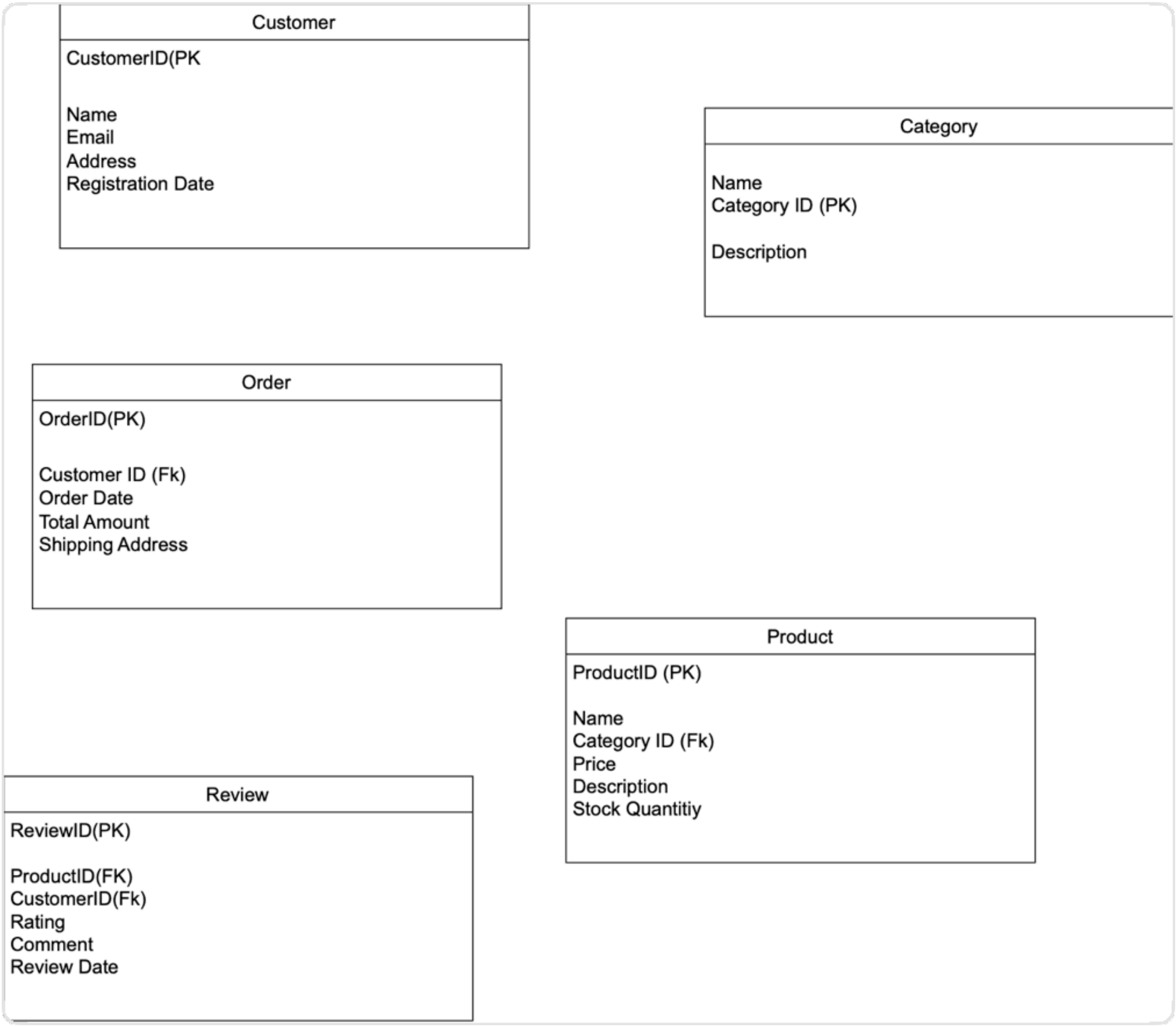
4. **Review**

- Attributes: Review ID (PK), Product ID (FK), Customer ID (FK) — ??, Rating, Comment, Review Date.
- Use Case: Customers can leave reviews on products they have purchased, which includes a rating and textual feedback.

5. **Category**

- Attributes: Category ID (PK), Name, Description.
- Use Case: Products are categorized for easier management and navigation on the platform.

3. Create an ER Diagram



Customer

CustomerID(PK

Name
Email
Address
Registration Date

Category

Name
Category ID (PK)

Description

Order

OrderID(PK)

Customer ID (Fk)
Order Date
Total Amount
Shipping Address

Product

ProductID (PK)

Name
Category ID (Fk)
Price
Description
Stock Quantitiy

Review

ReviewID(PK)

ProductID(FK)
CustomerID(Fk)
Rating
Comment
Review Date

Follow Up Summary

Differences from ER Diagram

The original ER diagram planned for tables like `Products`, `Customers`, `Orders`, `Reviews`, and `Categories`. In practice, due to the unavailability of the customer dataset, I had to adapt and work with the available "eCommerce Behavior Data". This led to a focus on `Products`, `Users`, and `Events` tables primarily, reflecting the dataset's structure of user interactions and product details.

Schema Changes

The actual database schema is significantly different then the initial ER diagram. I removed of multiple interconnected entities like `Customers`, `Orders`, and `Reviews`, and simplified it. The `Products` table was retained but adjusted to the dataset specifics, and `Users` and `Events` tables were introduced to capture the user activity and product interactions.

Data Cleaning and Preparation

For data insertion, the raw dataset required cleaning and formatting to fit the SQL schema. This included converting timestamps to a MySQL-friendly format. I also had to download the 2019 csv file from original Kaggle link since the 2020 was too large.

Challenges in Data Cleaning

One challenge was the date-time format conversion, where the original 'UTC' timestamp needed to be reformatted to match MySQL's datetime standards. Additionally, handling missing data, especially for categories and brands in the `Products` table, required careful consideration to maintain data integrity.

View and Trigger Implementation

The `ProductSummary` view was created to provide a quick summary of product interactions, offering insights into views and purchases per product. This aligns with the database's purpose of tracking user-product interactions and aids in decision-making for inventory and marketing strategies.

The `BeforePriceUpdate` trigger was implemented to log changes in product prices, enhancing the database's ability to audit and track significant data changes. This addition serves to maintain a history of price adjustments, crucial for financial analysis and strategic planning.

Conclusion and Future Directions

The shift from the proposed comprehensive e-commerce platform to a more focused user-product interaction database was a significant pivot in the project's direction.

The challenges faced, particularly in data preparation and schema adjustment, underscored the need for robust data cleaning and validation processes. Going forward, improving data validation, expanding the database to include more purchasing data from customers. Also, if I had to choose a new database, I would try to understand a dataset by actually finishing the ERD diagram.