**Data Sources Documentation**

**1. Introduction:**

**a. Brief Overview of the Datasets**

There are three main datasets in the data warehouse project: Customers\_FlipKart, Products\_FlipKart, and Sales\_FlipKart. The Customers collection keeps track of information about each customer's demographics and behavior, such as what they buy. The Products dataset has important details about products, like their category, price, and availability. The Sales Transactions dataset is the main fact table. It stores information about every purchase and uses unique identifiers to connect customers and goods.

**How I created my Dataset?**

Using the Faker tool in Jupyter Notebook, I made the datasets for this project. I found the most important fields and relationships for a full retail data warehouse by studying e-commerce data structures on Kaggle. I used a star schema design with Sales as the fact table to make three datasets that are all linked to each other: Customers\_FlipKart, Products\_FlipKart, and Sales\_FlipKart. I made realistic fake data using Faker, such as customer information, product details, and transaction records. Each dataset has more than 1000 records to make sure there is enough information for research.

**b. Purpose of the Datasets in the Data Warehouse**

These datasets are meant to help people make decisions based on data by making it easy to store, find, and analyze sales and customer contacts. Businesses can learn about the demographics and behavior of their customers with the Customers dataset. The Products dataset, on the other hand, helps them keep track of their supplies, prices, and how well their products are doing. The Sales Transactions dataset is an important part of business intelligence and reporting because it shows trends in income, buying patterns, and operational efficiency.

**c. How Datasets are Connected**

The three sets of data are linked together using the primary and foreign keys, creating a star schema in the data warehouse. In this case, the Sales Transactions dataset is the fact table, and the customer\_id and product\_id in the Customers dataset connect them. This structure lets you run quick and effective analytical questions, which lets you keep an eye on what customers buy, how much demand there is for products, and how much money is being made at different levels of detail.

**2. Dataset Descriptions:**

**2.1 Customers\_FlipKart Dataset**

The Customers dataset has information about customers' demographics and behaviors, which helps businesses look at buying patterns and divide customers into groups. A primary key (customer\_id) is used to identify each customer individually. There are fields in the dataset for things like birth year, level of schooling, marital status, income, number of children, purchase history, and loyalty score. For customer segmentation, personalized marketing, and behavior research, this dataset is a must-have.

**Key Attributes:**

* product\_id (Primary Key) – Unique identifier for each product.
* product\_name – The name of the product.
* category – The category the product belongs to (e.g., Electronics, Furniture).
* price – The listed price of the product.
* stock\_quantity – Available stock levels to monitor inventory.
* rating – Customer review rating for the product.

A screenshot of a computer

AI-generated content may be incorrect.

**2.2 Sales\_FlipKart Dataset**

The Sales\_FlipKart dataset is the Fact Table in the data center. It keeps track of all the purchases that customers make. It has information about the transaction, like the order details, buy date, customer ID, product ID, quantity, total amount, and payment method. This dataset is very important for tracking sales success, analyzing revenue, and predicting demand because it has foreign keys (customer\_id and product\_id) that connect customers and products.

**Key Attributes:**

* transaction\_id (Primary Key) – Unique identifier for each transaction.
* customer\_id (Foreign Key) – Links to the Customers dataset.
* product\_id (Foreign Key) – Links to the Products dataset.
* order\_date – Date of the purchase transaction.
* quantity – Number of units purchased in the transaction.
* total\_amount – The final purchase amount after discounts.
* payment\_method – Payment type used (Credit Card, PayPal, etc.).

A screenshot of a computer

AI-generated content may be incorrect.

**2.3 Products\_FlipKart Dataset**

The Product\_FlipKart dataset has information about items that are for sale, such as their categories, brands, prices, availability, and discounts. A primary key (product\_id) is used to identify each product individually. This set of data helps companies keep track of their supplies, study how well their products are doing, and find the best ways to set prices.

**Key Attributes:**

* product\_id (Primary Key) – Unique identifier for each product.
* product\_name – The name of the product.
* category – The category the product belongs to (e.g., Electronics, Furniture).
* price – The listed price of the product.
* stock\_quantity – Available stock levels to monitor inventory.
* rating – Customer review rating for the product.

A screenshot of a computer

AI-generated content may be incorrect.

**3. Relationships between Datasets:**

The Customers, Products, and Sales Transactions datasets are interconnected using primary and foreign keys, forming a star schema model that enables efficient analytical querying.

**3.1 Customers\_FlipKart Dataset (customer\_id)**

* The **primary key** customer\_id uniquely identifies each customer.
* The **Sales\_FlipKart dataset** contains customer\_id as a **foreign key**, linking each transaction to a specific customer.

**3.2 Products\_FlipKart Dataset (product\_id)**

* The **primary key** product\_id uniquely identifies each product.
* The **Sales\_FlipKart dataset** contains product\_id as a **foreign key**, linking each transaction to the product being purchased.

**3.3 Sales\_FlipKart Dataset (Fact Table)**

* Acts as the **Fact Table**, linking **Customers and Products** datasets.
* Contains both customer\_id and product\_id as **foreign keys**, ensuring that each sale references a valid customer and product.

**Relationship Summary**

* **One-to-Many Relationship** (Customers → Sales Transactions):
* Each customer can make more than one purchase, but each purchase is only linked to one customer.
* **One-to-Many Relationship** (Products → Sales Transactions):
* Every product can be part of more than one transaction, but every transaction is only linked to one product.