The Dataset

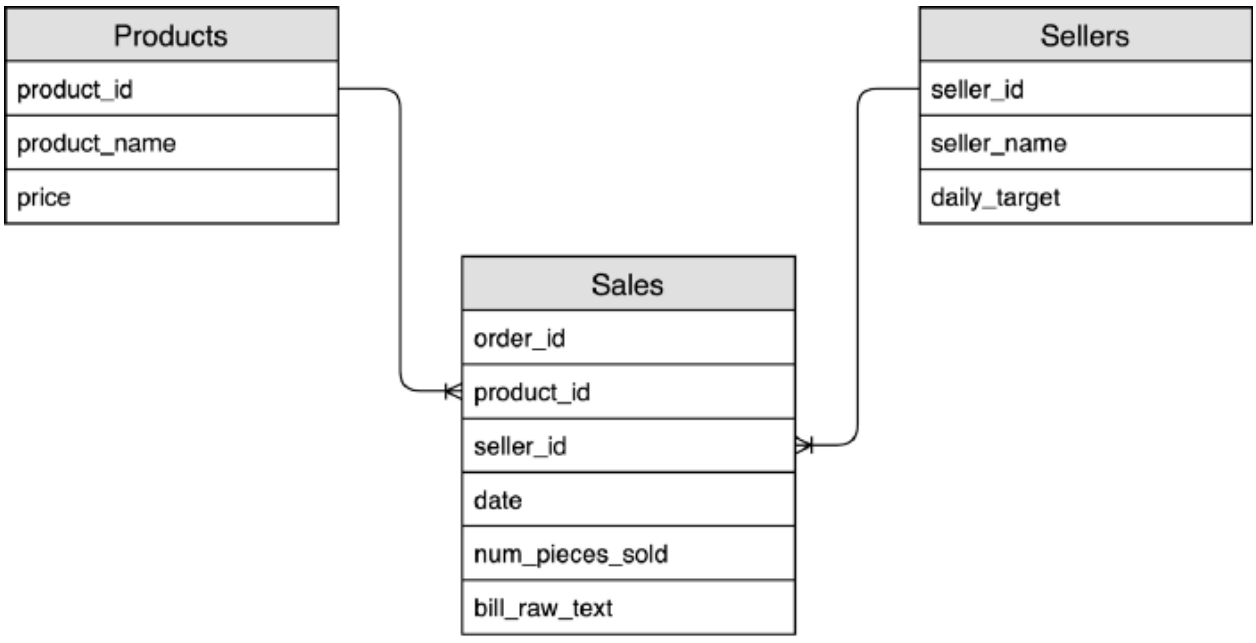
Let's describe the dataset briefly that we will use: it consists of three tables coming from the database of a shop, with products, sales, and sellers. The complete materials needed for this exercise can be downloaded from our public cloud here:

https://cloudshellsatiplytics.blob.core.windows.net/public-share/spark-exercise-materials.zip

You will find 3 CSV files describing the data and the format of this data so that it is easier for you to see it; these CSV's are of no use for the exercise. They are simply an example of the data. The actual data you will need to parse is in PARQUET, stored in 3 folders (Product_parquet, sales_parquet, sellers_parquet).

The Data Model:

While you explore the below Data schema, feel free to open the CSV files and study the format and the data in general to get an understanding of the model.



Sales Table:

Each row in this table is an order, and every order can contain only one product. Each row stores the following fields:

- order_id: The order ID
- product_id: The single product sold in the order. All orders have exactly one product)
- seller_id: The selling employee ID that sold the product
- num_pieces_sold: The number of units sold for the specific product in the order
- bill_raw_text: A string that represents the raw text of the bill associated with the order
- date: The date of the order.

Products Table:

Each row represents a distinct product. The fields are:

- product_id: The product ID
- product_name: The product name
- price: The product price

Sellers Table:

This table contains the list of all the sellers:

- seller_id: The seller ID
- seller_name: The seller name
- daily_target: The number of items (regardless of the product type) that the seller needs to hit his/her quota. For example, if the daily target is 100,000, the employee needs to sell 100,000 products; He can hit the quota by selling 100,000 units of product_0 or by selling 30,000 units of product_1 and 70,000 of product_2.

The Exercises:

Warm-up #1:

- Find out how many orders, how many products and how many sellers are in the data.
- How many products have been sold at least once? Which is the product contained in more orders?

Warm-up #2:

- How many distinct products have been sold each day?

Exercise #1:

- What is the average revenue of the orders?

Exercise #2:

- For each seller, what is the average % contribution of an order to the seller's daily quota?

Exercise #3:

- Who are the second most selling and the least selling persons (sellers) for each product? Who are those for the product with `product_id = 0`

Exercise #4:

- **if the order_id is even**:
  - apply MD5 hashing iteratively to the bill_raw_text field, once for each 'A' (capital 'A') present in the text. E.g. if the bill text is 'nbAAnllA', you would apply hashing three times iteratively **(only if the order number is even)**
- **if the order_id is odd:**
  - Apply SHA256 hashing to the bill text
- **Finally**, check if there are any duplicate on the new column

Thank you for taking the time and spending the effort with us!