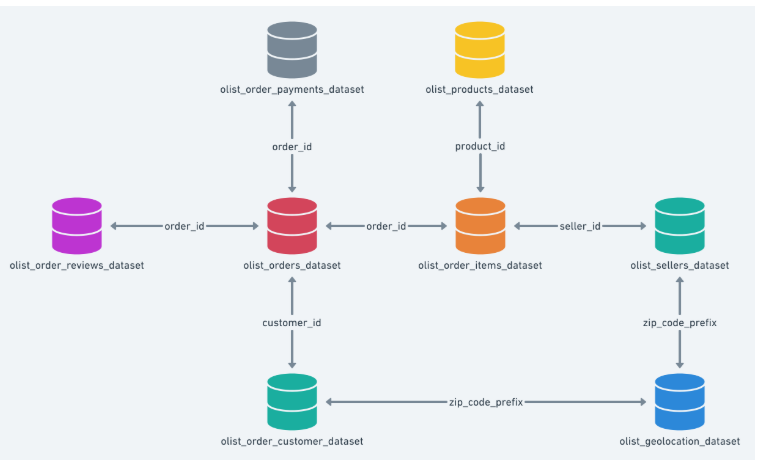
**Assignment # 3: Data Analysis using SQL**

In this assignment we have provided you with the ***Brazilian E-Commerce Public Dataset by Olist*** This dataset provides public access to Brazilian e-commerce order data from the Olist Store. It contains detailed information on approximately 100,000 orders placed between 2016 and 2018 across various marketplaces in Brazil. The dataset includes multiple features that allow for an in-depth analysis of each order from different perspectives. Users can explore details such as order status, pricing, payment methods, and freight performance. Additionally, it provides insights into customer locations, product attributes, and customer reviews, enabling a comprehensive understanding of the e-commerce landscape. To further enhance geographical analysis, a geolocation dataset is also available, mapping Brazilian zip codes to their respective latitude and longitude coordinates. However, you will not be working with that for now.

This dataset represents real commercial transactions, though it has been carefully anonymized to protect sensitive information. To maintain privacy, references to specific companies and business partners within the review texts have been removed. Instead, these names have been replaced with some other names.

**Data Schema**



As a data analyst, your role is crucial in extracting valuable insights from this dataset. With access to detailed e-commerce transaction data, you are tasked with performing various types of analyses to uncover trends, patterns, and key business insights. Using SQL queries, you will analyze customer behavior, order performance, and sales trends to provide data-driven recommendations. Your work will involve predicting future sales, identifying factors influencing customer satisfaction, and optimizing logistics based on freight and payment data. By leveraging SQL’s powerful querying capabilities, you will transform raw data into actionable intelligence, helping businesses make informed decisions and improve their marketplace strategies.

**Task # 1 Data Uploading and Cleaning**

Load the data from the CSV files to tables that you have created according to the schema provided. The primary key to the geolocation table is not unique, you have to make changes in the key to make it unique so each location is uniquely identified. You must note that this primary key is also linked to other tables through the foreign key. Refer to the dataset and perform your own data analysis to confirm that your chosen key(s) enforce uniqueness and maintain referential integrity across related tables. This approach will help you decide when to use a single-column primary key or a composite key for your tables.

**Task # 2 Data Retrieval**

You are required to make the following data analysis using SQL.

**ORDER ANALYSIS**

We often lose our customers as the orders are delayed with the estimated delivery date. You need to find the following information

1. Percentage of orders that got delayed beyond the estimated date
2. What are the peak months of order delays?
3. Which state experiences the highest order delays?
4. See how many orders are still in **“pending”** status for each year
5. What is the average delay duration per seller?
6. How do shipping costs impact order delays? Find the average shipping cost for delayed and on-time orders.
7. Which product category experience the most order delays.
8. How do number of items per order affect the delays? Find the average number of items for delayed and on time orders.

**CUSTOMER ANALYSIS**

**We want to find the pattern and buying behavior of our customers, hence a through analysis is required for the customers as well. You need to find the following information**

1. What percentage of customers have made only one order?
2. Find the top five cities with the most repeat customers (customers who have placed orders more than once)
3. **Calculate the average order price of customers for each state**
4. **Find the top ten customers with the highest number of orders placed.**
5. **Which customers have the longest average delivery time**
6. How does customer order frequency change over time? Find the average number of orders placed per customer per year.
7. Which top 5 customers have spent the most money in year 2017
8. Which customers have the highest order cancellations

**PRODUCT ANALYSIS**

1. What is the most profitable product category per state? Find the most profitable product category in each state based on total sales.
2. What are the peak hours for order placements per product category? Find **which hours of the day** have the **highest number of order placements** for each **product category**.
3. Which product categories experience the most delays? Find the **top 5 product categories with the highest number of delayed orders.**
4. What is the impact of product price on sales volume? Find whether **higher-priced products sell less** by comparing **average price vs. total sales**.
5. Which products are most frequently bought together? Find the **most frequently purchased product pairs**.
6. Calculate the total revenue per product category by summing order item prices.
7. Compute the average review score for each product category.
8. Retrieve the top 5 products based on total sales revenue.

**SELLER AND SHIPMENT ANALYSIS**

1. What is the return rate per seller?



1. Which sellers sell the most expensive products on average? Find sellers whose **products have the highest average price**.
2. What is the profit margin per seller? Find the **profit margin per seller**, assuming:



1. How do shipping costs impact order delays? Find the **average shipping cost** (freight\_value) for delayed vs. non-delayed orders.
2. What is the number of delayed shipments in 2017
3. What is the correlation between shipping cost and delivery speed? Find the **average shipping cost** (freight\_value) for delayed vs. on-time shipments.
4. Sum the total freight cost for each seller

**Task # 3 Presentation**

All the queries mentioned above that return more than 10 rows in the result table should be saved in a CSV file. Once the data is stored, a graphical representation should be created using Excel or Google Sheets. A screenshot of the generated graph should then be included in the report.

**Submission Guidelines**

You should submit a report that includes queries and the respective graphs you have made. (Only those queries that give a result table that has more than 10 queries. In the report write the query question and the screenshot of the graph. No SQL code is required. ***No Submission after deadline will be considered.***

1. **Report Structure**

* The report should be submitted as a **PDF file**.
* The report should be structured as follows:
* **Title Page**: Assignment title, student name, roll number, and course name.
* **Table of Contents** (optional but recommended for better organization).
* **Task 1: Data Uploading & Cleaning**
* Brief explanation of the steps taken to upload and clean the data.
* Mention the modifications made to ensure the uniqueness of the geolocation primary key.
* Any changes made for unique identification should be mentioned
* **Task 2: Data Retrieval (SQL Analysis)**
* Each question must be included as a heading, followed by:
* **Query result summary** (e.g., key insights from the data).
* **Graphical representation**: Only include a **screenshot** of the chart/graph (no SQL code required).
* **Task 3: Presentation**
* Explanation of how data was exported and visualized.
* Screenshot of the graphical represntation
* Summary of key findings from the graphs.

1. **SQL Query Execution & Exporting**

* Only queries that return more than **10 rows** should be visualized.
* Students should **export these query results to CSV files** and then generate relevant graphs using **Excel or Google Sheets**.
* All **graphs should be labeled properly**, with appropriate titles, axes, and legends.

1. **Submission Format**

* Submit a **single compressed ZIP folder** containing:
* **PDF report** (as described above).
* **CSV files** for queries with more than 10 rows.
* **Sql file**
* Name your ZIP file as: ***Assignment2\_Section\_StudentID1\_StudentID2.zip***