DATA ENGINEER HOME ASSIGNMENT

1. Data modeling

We have a MariaDB/MySQL database for each client's online commerce. Below are the key fields for three of the most important tables:

client_products:
- **id** (varchar) PRIMARY KEY
- **title** (varchar)
- **price** (decimal)
- **stock_availability** (boolean)
- **main_category** (varchar)
- **is_active** (boolean)
- **updated_at** (datetime)

matchings:
- **matching_id** (int) PRIMARY KEY autoincremental
- **product_id** (varchar) UNIQUE KEY `unique_matching`
- **market_shop_id** (int) UNIQUE KEY `unique_matching`
- **market_product_id** (varchar) UNIQUE KEY `unique_matching`
- **created_at** (datetime)

market_products
- **shop_id** (int) PRIMARY KEY
- **id** (varchar) PRIMARY KEY
- **title** (varchar)
- **price** (decimal)
- **stock_availability** (boolean)
- **is_active** (boolean)
- **updated_at** (datetime)
- **failed_to_update_at** (datetime)

The client_products table contains data on the client's product catalog, while the market_products table stores information about products from other online shops in the market.

Each client product can have multiple matchings with different products from various online shops.

We generally do not enforce foreign keys at the database level. Instead, we establish relationships at the application level and when transferring data to the data warehouse.

Assignment:

- Using any preferred format (Word, Excel, PDF, or any other application), create a simple diagram illustrating the relationships between these three tables. Additionally, provide any comments or explanations that might help us better understand your reasoning.

2. Dataflow

A data analyst on your team wants to create visualizations for one of our clients using our dashboarding tool, Looker Studio. This tool is directly connected to our data warehouse (BigQuery). Therefore, the analyst has requested your assistance in preparing the necessary data sources via a dataflow process that will extract data from the production database (which includes the 3 aforementioned tables in section 1). Additionally, they may require support in writing queries to retrieve data from the data warehouse.

The analyst mentioned the following data requirements for building the dashboards:

- A simple chart displaying all matchings for the client’s active products, including relevant product data (1 row per combination);
- Showing the customer's top 10 most competitive products in the market;
- Showing the customer's top 10 worst categories in terms of market competitiveness.

Assignment:

Using any preferred format (Word, Excel, PDF, or any other application), complete the following tasks:

1. Diagram the dataflow process: provide a simple representation of the dataflow process, including any relevant comments or explanations to clarify your approach;
2. Write SQL queries when needed for the related steps of the dataflow process;
3. Write SQL queries to retrieve the data based on the 3 specified afore-mentioned criteria to assist the data analyst in building the dashboards.

3. Data stack

A growing eCommerce startup wants to build a robust data infrastructure that can support the creation of client-facing data products, dashboards, and data-driven decision-making. You have been asked to propose a data stack that efficiently moves data from the production database to a data warehouse, ensuring that data is clean, reliable, well-structured, and ready for delivery to analysts and business users.

Assignment:

Using any preferred format (Word, Excel, PDF, or any other application), complete the following tasks:

- Propose a high-level data stack architecture, including key tools and technologies;
- Justify your choices by explaining how they align with scalability, cost-efficiency, and ease of use;
- If applicable, describe any trade-offs or alternatives that could be considered.

4. Data visualization

Assignment:

- Create a Looker Studio dashboard (or use a similar tool like Tableau, Power BI, etc.) using the data source named "data_visualization_home_assignment" and generate the necessary visualizations to answer the following questions:

    1. What is the average percentage difference between my product prices and those of my competitors?
    2. Which product has the largest price difference?
    3. Am I more competitive in certain brands?
    4. Which competitor offers the lowest prices?
    5. For which products should I lower my prices to be closer to the cheapest competitor (i.e., my current price is significantly higher than the lowest competitor's price)?
    6. For which products should I increase my prices to be closer to the cheapest competitor (i.e., I currently have the lowest prices among all competitors)?