ETL Task for Programmatic Inventory Support Candidate

Objective

You are given raw CSV files representing sales transactions and marketing campaign data. Your goal is to design and implement an ETL process that transforms the data into a clean, analytics-ready format and populates a reporting table.

📂 Files

1. `sales_data.csv`

   - `transaction_id` (string)

   - `product_id` (string)

   - `customer_id` (string)

   - `quantity` (integer)

   - `price_per_unit` (decimal)

   - `transaction_date` (YYYY-MM-DD)

   - `region` (string)

2. `marketing_campaigns.csv`

   - `campaign_id` (string)

   - `product_id` (string)

   - `campaign_name` (string)

   - `start_date` (YYYY-MM-DD)

   - `end_date` (YYYY-MM-DD)

- channel (e.g. Email, Social Media, Display)

📑 Requirements

1. Extract

Load both CSV files into staging tables or memory (depending on the environment: SQL, Python, or dbt).

2. Transform

Create a clean `fact_sales` table [or CSV file] with the following transformations:

- Ensure `price_per_unit` is non-negative. If not, discard the row.

- Remove rows with missing `product_id` or `customer_id`.

- Calculate `total_sales` = `quantity` × `price_per_unit`.

- Add a derived column: `campaign_name` (string) by joining with `marketing_campaigns.csv` using:

  - `product_id`

  - and checking whether the `transaction_date` falls between `start_date` and `end_date`.

    Edge Case: If multiple campaigns overlap, pick the one with the **most recent start date**.

3. Load

Populate a new table [or CSV file] called `bi_sales_summary` with:

- `region`

- `campaign_name`

- `total_sales`

- `transaction_month` (YYYY-MM format)

- `sales_count` (number of transactions)

4. Bonus

- A script in Python, Node.js or SQL (with DDL/DML) or dbt models implementing the ETL pipeline.

- A document (Markdown or PDF) explaining your assumptions, logic, and any caveats. [with any files or database dumps produced]

- Answers to the bonus questions.

🛠️ Expected Deliverables

- A script in the language of choice: Python (pandas) or SQL (with DDL/DML) or dbt models implementing the ETL pipeline.

- A document (Markdown or PDF) explaining your assumptions, logic, and any caveats.

- Answers to the bonus questions.