DWBI Problem Statement: SQL Generation with LLM

# Problem Statement

You are part of the Data Warehouse & Business Intelligence (DWBI) team at a retail company. The business team frequently asks ad-hoc questions across two data sources:  
  
1. Sales Data Warehouse (sales\_dw)  
2. Marketing Data Warehouse (marketing\_dw)  
  
Both sources are in Table.  
  
The challenge is to use a Large Language Model (LLM) to generate SQL queries automatically. You will provide the schema information of both data sources, feed 20 business questions, and capture the SQL queries generated. The output should be stored in a CSV with columns:

* question\_id
* question (as given)
* target\_source (sales\_dw or marketing\_dw)
* sql (the query generated)
* assumptions (if any made by the LLM)
* confidence (0.0–1.0 scale)

# Data Source 1: Sales Data Warehouse (sales\_dw)

## Table: sales

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| sale\_id | INT | Unique identifier for each sale |
| product\_id | INT | Foreign key → products.product\_id |
| region | VARCHAR | Sales region |
| sale\_date | DATE | Date of transaction |
| sales\_amount | DECIMAL | Revenue from transaction |
| quantity | INT | Number of units sold |

## Table: products

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| product\_id | INT | Unique product ID |
| product\_name | VARCHAR | Name of the product |
| category | VARCHAR | Product category |
| subcategory | VARCHAR | Subcategory |
| brand | VARCHAR | Product brand |

# Data Source 2: Marketing Data Warehouse (marketing\_dw)

## Table: campaigns

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| campaign\_id | INT | Unique campaign ID |
| channel | VARCHAR | Marketing channel |
| start\_date | DATE | Campaign start date |
| end\_date | DATE | Campaign end date |
| budget | DECIMAL | Campaign budget |

## Table: impressions

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| campaign\_id | INT | Foreign key → campaigns.campaign\_id |
| day | DATE | Date of impressions |
| impressions | INT | Number of impressions shown |
| clicks | INT | Number of clicks received |

# Business Questions

1. What are the top 5 products by sales amount in the last 90 days?

2. Show the month-over-month sales growth by region for the past 6 months.

3. Which categories contributed the most to total revenue in the last year?

4. Find the average order value (AOV) per region in the current quarter.

5. Identify the top 3 brands with highest quantity sold in the last 30 days.

6. Which subcategory had the sharpest decline in sales compared to the previous quarter?

7. What is the percentage contribution of each region to total sales this year?

8. Show the trend of sales\_amount vs quantity sold for Electronics products.

9. Find the product with the highest sales per unit (sales\_amount ÷ quantity) in the last 60 days.

10. List the top 10 customers by revenue (if customer table exists).

11. Which channel had the highest total impressions in the last quarter?

12. Calculate the average click-through rate (CTR) per channel last month.

13. Which campaign delivered the lowest cost per click (CPC) in the last 6 months?

14. Find the total budget spent per channel in the last year.

15. Identify the top 3 campaigns by impressions during their active periods.

16. What is the daily average impressions vs clicks trend for Social Media campaigns?

17. Which channel shows the highest conversion ratio (clicks ÷ impressions) overall?

18. List campaigns that ran for more than 60 days and their total spend.

19. Compare campaign budgets vs actual clicks to highlight underperforming campaigns.

20. Find the month with the highest total impressions across all campaigns.