Business Use Case: Sales Analytics in Hive

# Problem Statement

A retail company wants to analyze its sales data using Hive. The business needs to understand the following metrics:  
1. Total sales by product category.  
2. Average sales per region.  
3. Top 3 products by total sales.  
4. Identify which regions have a higher-than-average sales amount.  
  
The company has the following datasets:  
1. Sales data containing product sales information.  
2. Product data containing product categories and other details.  
3. Region data containing regional sales details.

# Sample Data

## Sales Table (sales\_data)

Columns:  
- sale\_id: INT  
- product\_id: INT  
- region\_id: INT  
- sale\_amount: FLOAT  
- sale\_date: DATE  
  
Sample Data:  
1, 101, 1, 500.00, 2023-08-01  
2, 102, 1, 600.00, 2023-08-02  
3, 103, 2, 700.00, 2023-08-03  
4, 104, 2, 800.00, 2023-08-04  
5, 101, 3, 400.00, 2023-08-05

## Product Table (product\_data)

Columns:  
- product\_id: INT  
- product\_name: STRING  
- category: STRING  
  
Sample Data:  
101, Laptop, Electronics  
102, Mobile, Electronics  
103, Chair, Furniture  
104, Table, Furniture

## Region Table (region\_data)

Columns:  
- region\_id: INT  
- region\_name: STRING  
  
Sample Data:  
1, North  
2, South  
3, East

# Solution

## 1. Total Sales by Product Category

Query:  
SELECT p.category, SUM(s.sale\_amount) AS total\_sales  
FROM sales\_data s  
JOIN product\_data p ON s.product\_id = p.product\_id  
GROUP BY p.category;  
  
Expected Output:  
| category | total\_sales |  
|-------------|-------------|  
| Electronics | 1500.00 |  
| Furniture | 1500.00 |

## 2. Average Sales Per Region

Query:  
SELECT r.region\_name, AVG(s.sale\_amount) AS avg\_sales  
FROM sales\_data s  
JOIN region\_data r ON s.region\_id = r.region\_id  
GROUP BY r.region\_name;  
  
Expected Output:  
| region\_name | avg\_sales |  
|-------------|-------------|  
| North | 550.00 |  
| South | 750.00 |  
| East | 400.00 |

## 3. Top 3 Products by Total Sales

Query:  
SELECT p.product\_name, SUM(s.sale\_amount) AS total\_sales  
FROM sales\_data s  
JOIN product\_data p ON s.product\_id = p.product\_id  
GROUP BY p.product\_name  
ORDER BY total\_sales DESC  
LIMIT 3;  
  
Expected Output:  
| product\_name | total\_sales |  
|--------------|-------------|  
| Chair | 700.00 |  
| Table | 800.00 |  
| Laptop | 900.00 |

## 4. Regions with Higher-than-Average Sales

Query:  
WITH avg\_sales\_all AS (  
 SELECT AVG(sale\_amount) AS avg\_sales FROM sales\_data  
)  
SELECT r.region\_name, SUM(s.sale\_amount) AS total\_sales  
FROM sales\_data s  
JOIN region\_data r ON s.region\_id = r.region\_id  
GROUP BY r.region\_name  
HAVING SUM(s.sale\_amount) > (SELECT avg\_sales FROM avg\_sales\_all);  
  
Expected Output:  
| region\_name | total\_sales |  
|-------------|-------------|  
| South | 1500.00 |