
Spark SQL Exercises – Tasks Only

Database & Table Tasks

1. Create a new database named `sales_db`.
2. Set the current database to `sales_db`.
3. Create a table `product_sales` with columns:
    - `ProductID` (INT)
    - `ProductName` (STRING)
    - `Category` (STRING)
    - `Price` (DOUBLE)
    - `Quantity` (INT)
    - `SaleDate` (DATE)
4. Insert at least 5 rows into `product_sales`.

Query Tasks

5. Select all records from `product_sales`.
6. Retrieve products where price is above 500.
7. Calculate total sale amount (`Price * Quantity`) for each product.
8. Find the number of products sold in each `Category`.
9. Sort products by total sales in descending order.

Temporary View Tasks

10. Create a PySpark DataFrame with dummy product data.
11. Register it as a temporary view called `temp_orders`.
12. Run a SQL query to filter `temp_orders` where quantity > 1.

Global View Tasks

13. Create a global temp view from a PySpark DataFrame named `global_orders`.
14. Run a SQL query on the global view from another notebook cell/session.

Join Tasks

15. Create a second table `customer_details` with:
    - `CustomerID`, `Name`, `Gender`, `City`, `SignupDate`
16. Insert at least 3 records into `customer_details`.
17. Write a SQL join between `product_sales` and `customer_details` based on `ProductID = CustomerID` (simulate a match).
18. List customers who bought more than 2 products.

View & Summary Tasks

19. Create a SQL view `sales_summary` that includes:
    - `ProductName`, `Price`, `Quantity`, `Total = Price * Quantity`

20. Query the view for records with `Total > 1000`.

🧹 Cleanup Tasks

21. Drop the view `sales_summary`.
22. Drop the tables `product_sales` and `customer_details`.
23. Drop the database `sales_db`.
