# Tasks

# 1. \*\*Join the DataFrames:\*\*

# Join the `product\_df` and `sales\_df` DataFrames on `ProductID` to create a combined DataFrame with product and sales data

product\_sales\_df = product\_df.join(sales\_df, on="ProductID")

product\_sales\_df.show()

# Calculate Total Sales Value:\*\*

# For each product, calculate the total sales value by multiplying the price by the quantity sold.

total\_sale\_product\_df = product\_sales\_df.withColumn("TotalSalesValue", col("Price") \* col("Quantity"))

total\_sale\_product\_df.show()

# Find the Total Sales for Each Product Category:\*\*

# Group the data by the `Category` column and calculate the total sales value for each product category.

total\_sale\_category\_df = total\_sale\_product\_df.groupBy("Category").sum("TotalSalesValue").withColumnRenamed("sum(TotalSalesValue)","TotalSales")

total\_sale\_category\_df.show()

# Identify the Top-Selling Product:\*\*

# Find the product that generated the highest total sales value.

total\_sale\_productName = total\_sale\_product\_df.groupBy("ProductName").sum("TotalSalesValue").withColumnRenamed("sum(TotalSalesValue)","TotalSales")

total\_sale\_productName.show()

sorted\_product\_Sale = total\_sale\_productName.orderBy(col("TotalSales").desc())

top\_selling\_product = sorted\_product\_Sale.limit(1)

top\_selling\_product.show()

# Sort the Products by Total Sales Value:\*\*

# Sort the products by total sales value in descending order.

sorted\_product\_Sale.show()

# Count the Number of Sales for Each Product:\*\*

# Count the number of sales transactions for each product.

sales\_count\_df = product\_sales\_df.groupBy("ProductID").count().withColumnRenamed("count","TransactionCount")

sales\_count\_df.show()

# Filter the Products with Total Sales Value Greater Than ₹50,000:\*\*

# Filter out the products that have a total sales value greater than ₹50,000.

product\_big\_price\_df = sorted\_product\_Sale.filter(col("TotalSales") > 50000)

product\_big\_price\_df.show()