### \*\*Exercise: Analyzing a Sample Sales Dataset Using PySpark\*\*

### \*\*Part 1: Dataset Preparation\*\*

#### \*\*Step 1: Generate the Sample Sales Dataset\*\*

import pandas as pd

from datetime import datetime

# Sample sales data

data = {

"TransactionID": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],

"CustomerID": [101, 102, 103, 101, 104, 102, 103, 104, 101, 105],

"ProductID": [501, 502, 501, 503, 504, 502, 503, 504, 501, 505],

"Quantity": [2, 1, 4, 3, 1, 2, 5, 1, 2, 1],

"Price": [150.0, 250.0, 150.0, 300.0, 450.0, 250.0, 300.0, 450.0, 150.0, 550.0],

"Date": [

datetime(2024, 9, 1),

datetime(2024, 9, 1),

datetime(2024, 9, 2),

datetime(2024, 9, 2),

datetime(2024, 9, 3),

datetime(2024, 9, 3),

datetime(2024, 9, 4),

datetime(2024, 9, 4),

datetime(2024, 9, 5),

datetime(2024, 9, 5)

]

}

# Create a DataFrame

pandas\_df = pd.DataFrame(data)

# Save the DataFrame to a CSV file

pandas\_df.to\_csv('sales\_data.csv', index=False)

print("Sample sales dataset has been created and saved as 'sales\_data.csv'.")

#2. \*\*Verify the Dataset:\*\*

pandas\_df = pd.read\_csv('sales\_data.csv')

print(pandas\_df)

#### \*\*Step 2: Load the Dataset into PySpark\*\*

#1. \*\*Initialize the SparkSession:\*\*

spark = SparkSession.builder \

.appName("Sales Dataset Analysis") \

.getOrCreate()

#2. \*\*Load the CSV File into a PySpark DataFrame:\*\*

df = spark.read.csv('sales\_data.csv', header=True, inferSchema=True)

#Display the first few rows

df.show(5)

#### \*\*Step 3: Explore the Data\*\*

#1. \*\*Print the Schema:\*\*

df.printSchema()

#2. \*\*Show the First Few Rows:\*\*

df.show(5)

#3. \*\*Get Summary Statistics:\*\*

df.describe("Quantity", "Price").show()

#### \*\*Step 4: Perform Data Transformations and Analysis\*\*

#1. \*\*Calculate the Total Sales Value for Each Transaction:\*\*

df = df.withColumn("TotalSales", col("Quantity") \* col("Price"))

df.show()

#2. \*\*Group By ProductID and Calculate Total Sales Per Product:\*\*

df.groupBy("ProductID").sum("TotalSales").alias("TotalProductSales").show()

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

from pyspark.sql.functions import desc

df.groupBy("ProductID").sum("TotalSales").alias("TotalProductSales").orderBy(desc("sum(TotalSales)")).show(1)

#4. \*\*Calculate the Total Sales by Date:\*\*

df.groupBy("Date").sum("TotalSales").alias("TotalSalesByDate").orderBy("Date").show()

#5. \*\*Filter High-Value Transactions:\*\*

df.filter(col("TotalSales") > 500).show()

### \*\*Additional Challenge (Optional):\*\*

#1. \*\*Identify Repeat Customers:\*\*

df.groupBy("CustomerID").count().filter(col("count") > 1).show()

#2. \*\*Calculate the Average Sale Price Per Product:\*\*

df.groupBy("ProductID").avg("Price").alias("AvgPricePerProduct").show()