Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

`

Department of Information Technology A.Y. 2024-2025

Class: TY BTech-IT, Semester: VI   Subject: Big Data Lab

NAME: Anish Sharma   **SAP:60003220045**

Experiment – 11

1. **Aim:** To implement Big Data Technologies for real world applications.

Procedure:

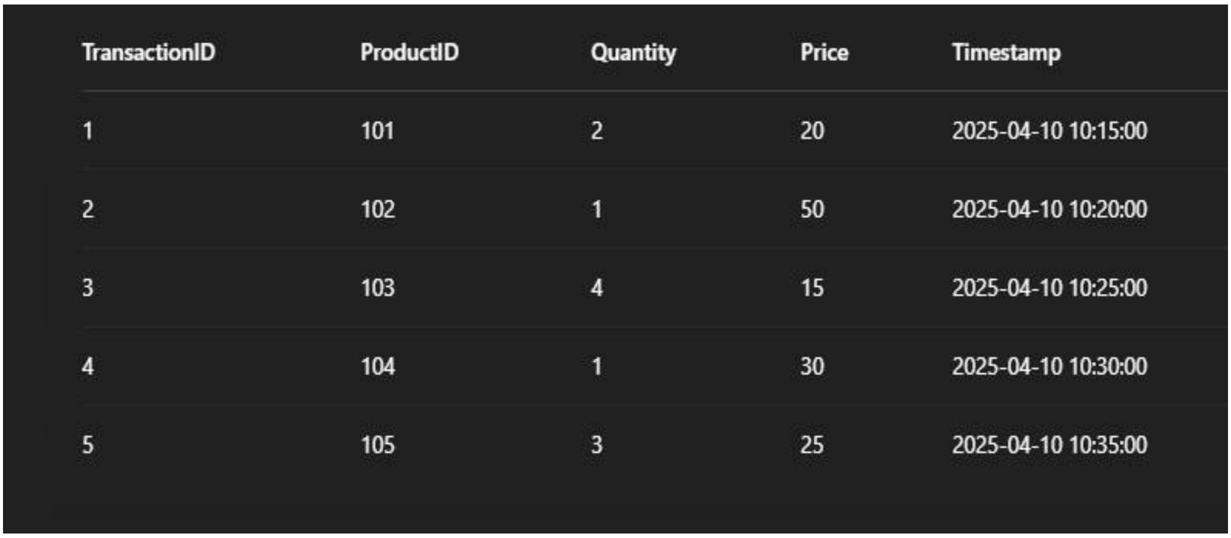
CODE:

```
from pyspark.sql import SparkSession from
pyspark.sql.functions import col

# Initialize the Spark session spark =
SparkSession.builder.appName("ECommerceAnalysis").getOrCreate()

# Sample data (assuming this is a CSV file with transactional data) data
= [
   (1, 101, 2, 20, "2025-04-10 10:15:00"),
   (2, 102, 1, 50, "2025-04-10 10:20:00"),
   (3, 103, 4, 15, "2025-04-10 10:25:00"),
   (4, 104, 1, 30, "2025-04-10 10:30:00"),
   (5, 105, 3, 25, "2025-04-10 10:35:00")
]

# Define schema
columns = ["TransactionID", "ProductID", "Quantity", "Price", "Timestamp"]

# Create DataFrame
df = spark.createDataFrame(data, columns)

# Show the loaded data df.show()

# Filter transactions where quantity is greater than 2
filtered_df = df.filter(col("Quantity") > 2)

# Add a new column for the total value (Quantity * Price)
```

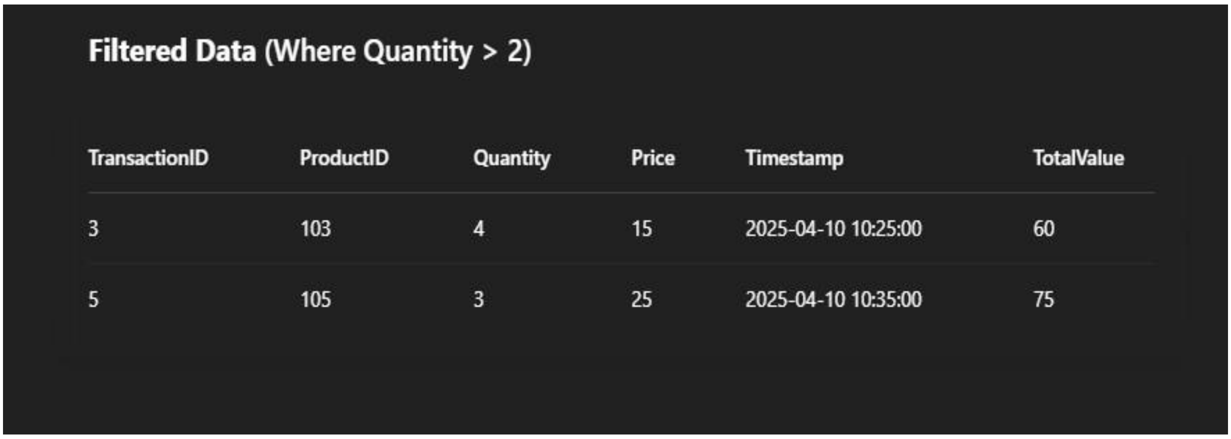
```
transformed_df = filtered_df.withColumn("TotalValue", col("Quantity") * col("Price"))

# Show the transformed data transformed_df.show()

# Aggregate total sales per ProductID sales_per_product_df
=
transformed_df.groupBy("ProductID").sum("TotalValue").withColumnRenamed("sum(TotalValue)"
, "TotalSales")

# Show the aggregated results sales_per_product_df.show()

# Optional: Write the result to a CSV file
# sales_per_product_df.write.csv("total_sales_per_product.csv", header=True)

# Stop the Spark session spark.stop()
```

2. **Requirements:** PC, Internet

OUTPUT:

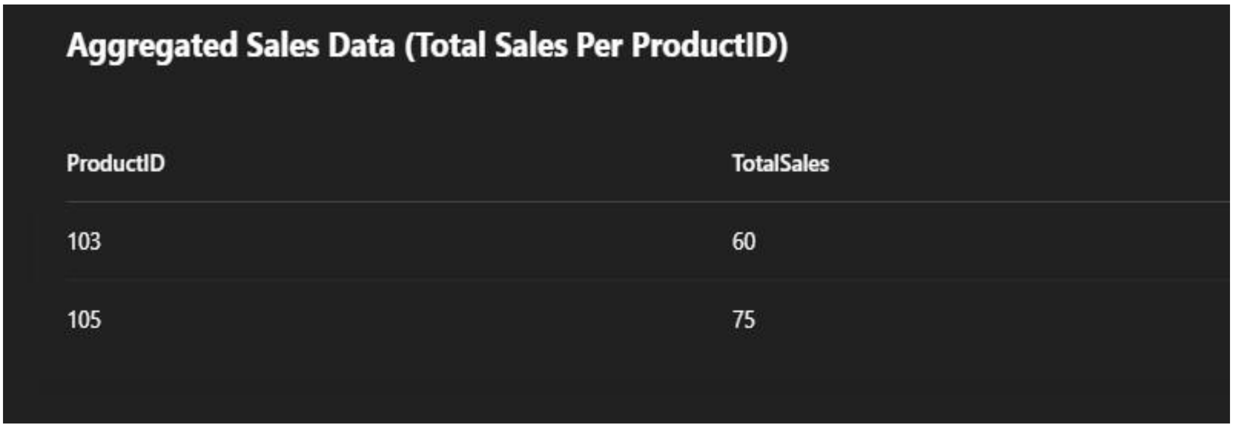


TransactionID	ProductID	Quantity	Price	Timestamp
1	101	2	20	2025-04-10 10:15:00
2	102	1	50	2025-04-10 10:20:00
3	103	4	15	2025-04-10 10:25:00
4	104	1	30	2025-04-10 10:30:00
5	105	3	25	2025-04-10 10:35:00



Filtered Data (Where Quantity > 2)

TransactionID	ProductID	Quantity	Price	Timestamp	TotalValue
3	103	4	15	2025-04-10 10:25:00	60
5	105	3	25	2025-04-10 10:35:00	75



Aggregated Sales Data (Total Sales Per ProductID)

ProductID	TotalSales
103	60
105	75

3. **Conclusion:** Thus, in this experiment, we implemented Big Data technologies for real world applications.