Big Data Analytics Mini Project

# Analyzing Retail Customer Purchase Behavior using MapReduce Simulation

Submitted by:  
Swastika Das  
Registration No: 2023pgdm1286

## 🎯 Objective:

To implement a basic simulation of the MapReduce framework for analyzing customer purchase behavior using a small-scale transactional dataset. This project highlights how MapReduce concepts can be translated into business insights, particularly for retail/e-commerce operations.

## 🔹 Project Overview:

In this project, we simulate a simplified MapReduce model using spreadsheet tools such as Microsoft Excel or Google Sheets. The goal is to process transaction data to generate key insights like:  
- Most purchased products  
- Revenue contribution by product categories  
- Most loyal/repeated customers

## 🔹 Scenario:

Imagine you are hired as a junior data analyst in a mid-sized online retail company. Your first task is to assist in identifying key sales and customer behavior patterns using a dataset of historical transactions.

## Step 1: Dataset Creation

Prepare a transactional dataset with the following columns:  
- Transaction\_ID  
- Customer\_ID  
- Product\_Name  
- Product\_Category  
- Quantity\_Purchased  
- Unit\_Price  
- Purchase\_Date  
  
You may use:  
- UCI Machine Learning Repository  
- Kaggle Retail Datasets  
- Or simulate at least 50 rows manually

## Step 2: Map Phase (Logic Breakdown)

In this phase, you extract key-value pairs as:  
- Product\_Name → Quantity\_Purchased  
- Product\_Category → Revenue (Quantity × Price)  
- Customer\_ID → Count of Transactions  
  
Example Output:  
(ProductX, 2), (ProductY, 5), (ProductX, 3)  
(CategoryA, 200), (CategoryB, 150)  
(Cust\_101, 3), (Cust\_102, 1)

## Step 3: Reduce Phase

Aggregate the data from the mapping stage:  
- Sum of quantities for each product  
- Total revenue for each product category  
- Purchase frequency for each customer

## Step 4: Visualization (Optional)

Represent your findings using visual tools such as:  
- Bar chart: Top 5 Products Sold  
- Pie chart: Revenue Share by Category  
- Column chart: Frequent Buyers (Top 5)  
  
Tools: Excel, Tableau, Google Sheets, or Power BI

## Step 5: Final Output & Observations

From the above simulation, we can derive:  
- Products contributing most to sales volume  
- Categories generating the highest revenue  
- High-value customers based on purchase frequency  
  
These findings will help the marketing and inventory team to focus promotions and restocking strategies accordingly.

## 🔸 Tools Used:

- Microsoft Excel / Google Sheets  
- Simulated dataset  
- Chart/Graph visualizations  
- MapReduce logic

## 🔸 Conclusion:

This mini-project demonstrates how Big Data principles like MapReduce can be practically applied in a spreadsheet-based environment to deliver actionable business insights. Although the scale is small, the methodology scales well for big data tools like Hadoop.