

BDA Mini Project Report

Title:

Simulating MapReduce for Customer Purchase Pattern Analysis: A Management Perspective

Submitted By:

Name: *Daisy Rabha*

Registration Number: 2023PGDM1068

Course: *Big Data Analytics*

Trimester: V

Institution: International Institute of Business Study

Objective:

This project aims to simulate the core logic of **MapReduce**, a fundamental concept in Big Data analytics, using **Microsoft Excel** as the primary tool. The simulation was conducted on a **retail transaction dataset**, with the goal of extracting valuable business insights relevant to **sales performance, customer behavior, and product categorization**. Specifically, we focused on:

- Identifying the **Top 5 Best-Selling Products** by quantity sold
- Calculating **Revenue Generated per Product Category**
- Determining the **Most Frequent Buyers** based on transaction counts

By doing this, we bridge the gap between big data theory and real-world retail business scenarios, demonstrating how scalable logic can be emulated on a smaller scale.

Dataset Description:

A synthetic dataset of **100 customer transactions** was created to represent a retail environment. The data includes a mix of product categories, customer IDs, purchase dates, and transactional details. The columns in the dataset were:

- **TransactionID** – A unique identifier for each transaction
- **CustomerID** – Unique ID for each customer
- **Product** – Name of the product purchased
- **Category** – Type of product (e.g., Electronics, Clothing, Groceries, Home & Kitchen)
- **Quantity** – Number of units purchased
- **Price** – Price per unit of the product
- **Date** – Date on which the transaction occurred

The dataset reflects a variety of customer behaviors and purchase patterns across different time periods, mimicking a typical retail store or e-commerce platform.

Methodology: Simulating MapReduce

The simulation followed the two primary phases of the MapReduce framework:

1. Map Phase – Key-Value Pair Creation

The **Map Phase** involves generating intermediate key-value pairs from the raw data. In this simulation, three mappings were established:

- **Product → Quantity:** Each product paired with the number of units purchased
- **Category → Revenue:** Each product category paired with the revenue generated from that transaction (calculated as $\text{Quantity} \times \text{Price}$)
- **CustomerID → Frequency:** Each transaction assigned a value of 1 to be counted later for customer frequency

These mappings were created using Excel formulas in separate sheets (`Map_Product_Quantity`, `Map_Category_Revenue`, `Map_Customer_Frequency`). This step mirrors how a mapper extracts useful components from raw input data in distributed computing.

2. Reduce Phase – Aggregation and Analysis

In the **Reduce Phase**, we consolidated the key-value pairs to derive meaningful results:

- **Reduce_Product_Sales:** Aggregated total quantity sold for each product using a Pivot Table.
- **Reduce_Category_Revenue:** Summed total revenue generated per product category.
- **Reduce_Customer_Activity:** Counted the number of transactions per customer to measure activity levels.

These pivot tables enabled us to identify top-selling products, high-revenue categories, and loyal or frequent buyers—all of which are crucial metrics in retail business management.

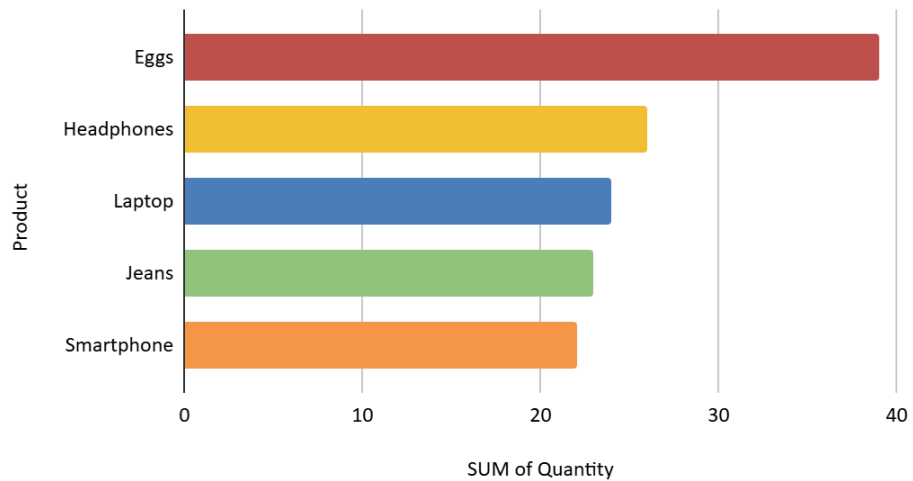
Visualizations:

To enhance clarity and present insights in a digestible format, the following charts were created:

1. Bar Chart – Top 5 Products by Quantity Sold

Displayed the most in-demand products. This helps in inventory planning and promotional focus.

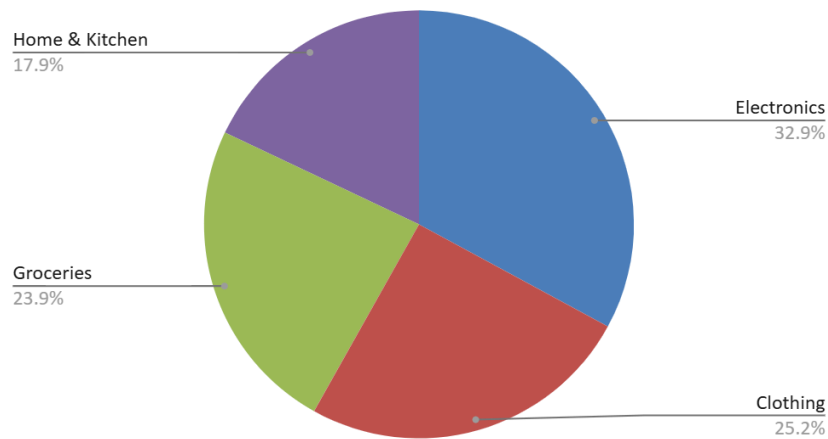
Top 5 Selling Products by Quantity



2. Pie Chart – Revenue by Category

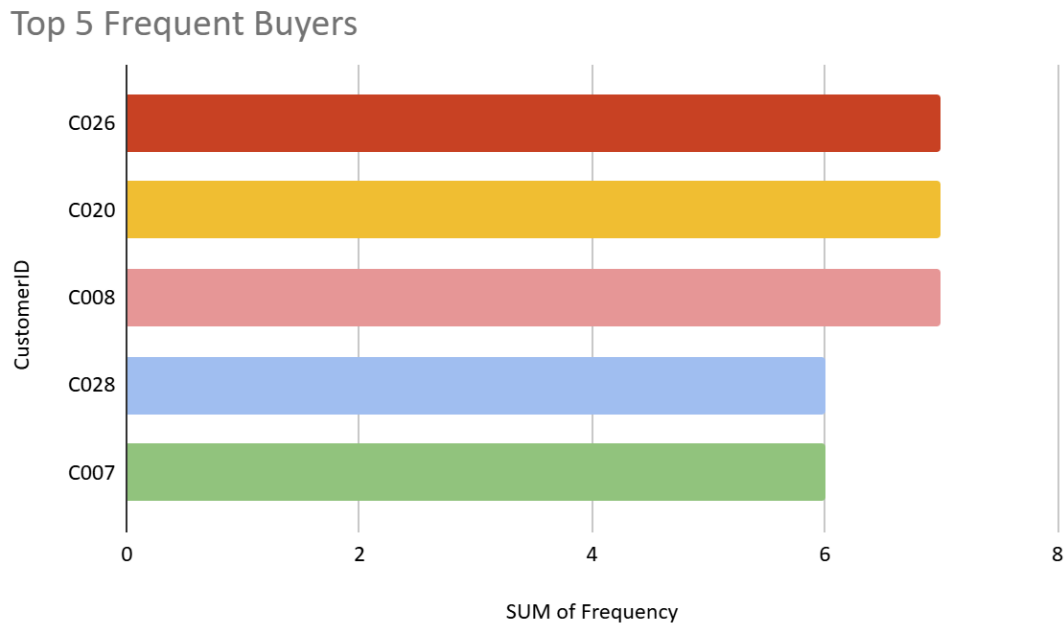
Offered a clear view of which product categories contributed most to overall revenue. Useful for category-level business strategy.

Revenue Distribution by Category



3. Bar Chart – Top 5 Frequent Buyers

Highlighted customers who made the most purchases. These insights are critical for loyalty programs and targeted marketing.



Each chart was created from the respective pivot tables and placed in a separate **Charts** sheet for better presentation.

Key Business Insights:

Based on the analysis, we observed the following patterns:

- **Top-Selling Products** tended to be lower-cost, everyday-use items, indicating frequent consumer demand.
- **Revenue Leaders** were from premium product categories (e.g., Electronics), despite selling fewer units.

- A small number of **high-frequency buyers** accounted for a significant portion of total transactions, validating the 80/20 rule in retail (Pareto Principle).

These insights provide actionable direction for:

- Restocking strategies
 - Pricing decisions
 - Personalized marketing
 - Customer segmentation
-

Conclusion:

This simulation effectively demonstrated how **MapReduce principles** can be applied in Excel to perform scalable logic on a manageable dataset. While tools like **Hadoop and Spark** are used in enterprise settings for real-time processing of massive datasets, this project reflects how similar thinking can guide decision-making even in small businesses or startups.

By integrating **data structuring, aggregation, and visual storytelling**, this project highlights how even non-programmatic tools like Excel can empower data analysts and managers to uncover valuable trends, customer behaviors, and operational opportunities.