**INTERNATIONAL INSTITUTE OF BUSINESS STUDY**

**IV TRIMESTER REPORT OF INTERNSHIP**

**BATCH OF 2023 – 2025**

**REPORT**

**ON**

# Big data Analytics Mini Project

Kashish Bhatia

2023PGM1121

**V TRIMESTER PGDM**

**Faculty:**

Akriti Gupta

Assistant Professor

IIBS, Bengaluru

**Course:**

Big Data Analytics

PGDM

2023-2025

Report of Internship submitted to the IIBS in partial fulfilment of the requirements of IV Trimester PGDM

**International Institute of Business Study, Bangalore – 562 157**

# Big data Analytics Mini Project

**Simulating MapReduce for Customer Purchase Pattern Analysis:**

**A Management Perspective**

**Objective**: The goal of this mini project is to simulate the functionality of the MapReduce model using a sample e-commerce transaction dataset to get the key business insights. This simulation was carried out using Microsoft Excel to understand customer purchase behaviour and product sales trends. The MapReduce framework was used to break down data processing into two main phases: Map (data transformation into key-value pairs) and Reduce (aggregation based on keys).

**Business Scenario:** As a data analyst in a growing e-commerce business, management requires insights into product performance and customer engagement. The role involves:

Track the product sale to get the idea that which product is in high demand.

Analyze the category-wise revenue to determine which product segments generate the most income.

Identify the frequent buyers to target the loyal customers for promotions. These insights help to take the strategic decisions which are related to inventory management, marketing campaigns, and sale forecasting.

**Dataset Preparation:** A dataset was created manually in Excel to simulate 50+ customer transactions. The dataset includes:

**Transaction ID:** Unique identifier for each transaction (TXN001, TXN002…).

**Customer ID:** Identifies each customer (C001, C003...).

**Product ID**: Product codes (P006, P007...).

**Product Name:** Descriptive name of the item sold (Blender, Laptop…).

**Category:** Product category (Kitchen, Apparel, Electronics…).

**Quantity:** Number of units purchased in each transaction.

**Price**: Price per unit.

**Date:** Date of purchase.

**Revenue:** Computed as Quantity × Price for each row.

I. **Map Phase Simulation:** The dataset was processed to simulate the Map step of the MapReduce paradigm. Each record in the dataset was converted into key-value pairs for different analysis goals:

a) **Product → Quantity:** This shows the quantity of products sold in individual transactions. It helps in calculating total units sold per product. It helps in calculating total units sold per product. Example: (Blender, 5), (Laptop, 5), (Headphones, 5).

b) **Category → Revenue:** It tracks the revenue generated by each product category in individual transaction. Example: (Kitchen, 23305), (Electronics, 14525).

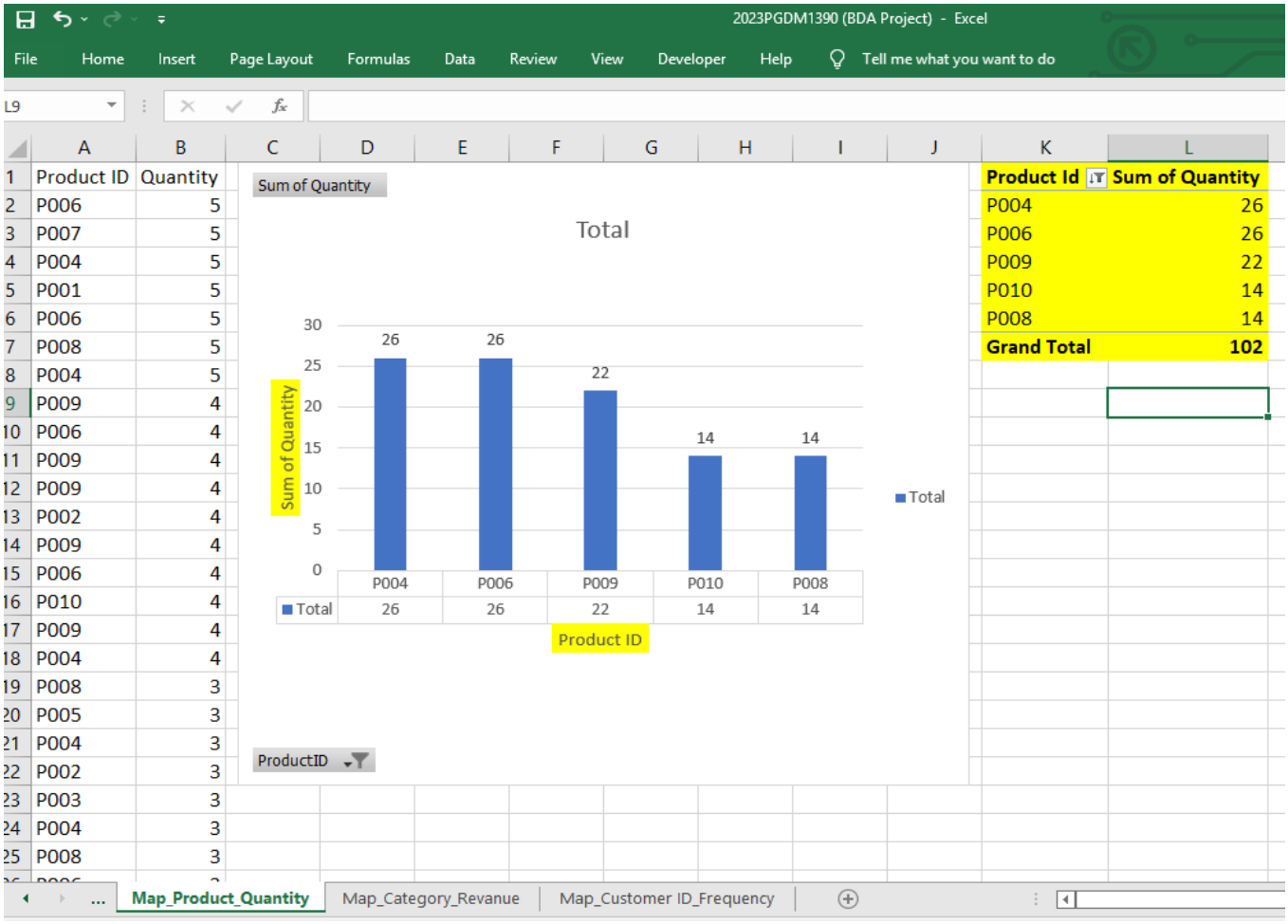
c) **Customer ID → Frequency:** This tracks the customers ID on the basis of how frequent they are purchasing the product. Example: (C003, 1), (C004, 1), (C001, 1).

I. **Reduce Phase Simulation:** In the next step we need to aggregate the key-value pairs to simulate the reduce phase. This provided the following insights:

**A. Top 5 products by Quantity:** These products represent the most frequently purchased items across all customers. P004 (Headphones) and P006 (Blender) are the joint top selling products with 26 units each. Followed by P009 (Mobile Charger) with 22 units etc.

| Product ID | Total Quantity Sold |
|---|---|
| P004 | 26 |
| P006 | 26 |
| P009 | 22 |
| P010 | 14 |
| P008 | 14 |

**Bar Chart**: Top 5 Products by Quantity Sold

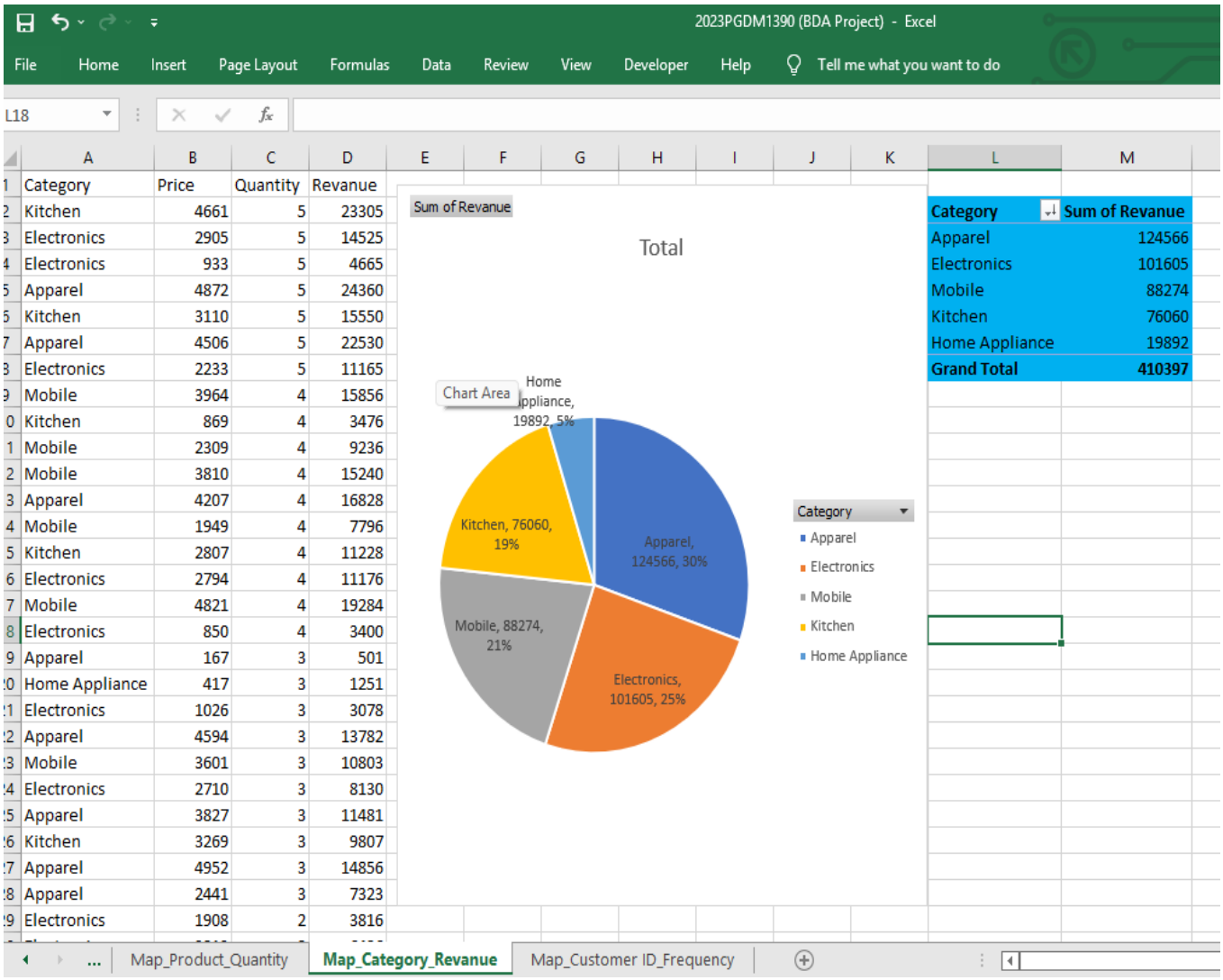


## A. Revenue by Category (Sum of Quantity × Price):

Revenue by Category (with name, revenue. The Apparel category contributed the highest revenue, indicating strong performance in fashion and clothing items. Apparel is the highest revenue-generating category. Home Appliance contributes the least and may need business attention or improvement.

| Category | Total Revenue (Rs.) |
|---|---|
| Apparel | Rs.1,24,566 |
| Electronics | Rs.1,01,605 |
| Mobile | Rs.88,274 |
| Kitchen | Rs.76,000 |
| Home Appliance | RS,19,892 |

**Pie Chart**: Revenue by top 5 Categories (Category, Sum of Revanue)




| Category | Price | Quantity | Revanue |
|---|---|---|---|
| Kitchen | 4661 | 5 | 23305 |
| Electronics | 2905 | 5 | 14525 |
| Electronics | 933 | 5 | 4665 |
| Apparel | 4872 | 5 | 24360 |
| Kitchen | 3110 | 5 | 15550 |
| Apparel | 4506 | 5 | 22530 |
| Electronics | 2233 | 5 | 11165 |
| Mobile | 3964 | 4 | 15856 |
| Kitchen | 869 | 4 | 3476 |
| Mobile | 2309 | 4 | 9236 |
| Mobile | 3810 | 4 | 15240 |
| Apparel | 4207 | 4 | 16828 |
| Mobile | 1949 | 4 | 7796 |
| Kitchen | 2807 | 4 | 11228 |
| Electronics | 2794 | 4 | 11176 |
| Mobile | 4821 | 4 | 19284 |
| Electronics | 850 | 4 | 3400 |
| Apparel | 167 | 3 | 501 |
| Home Appliance | 417 | 3 | 1251 |
| Electronics | 1026 | 3 | 3078 |
| Apparel | 4594 | 3 | 13782 |
| Mobile | 3601 | 3 | 10803 |
| Electronics | 2710 | 3 | 8130 |
| Apparel | 3827 | 3 | 11481 |
| Kitchen | 3269 | 3 | 9807 |
| Apparel | 4952 | 3 | 14856 |
| Apparel | 2441 | 3 | 7323 |
| Electronics | 1908 | 2 | 3816 |

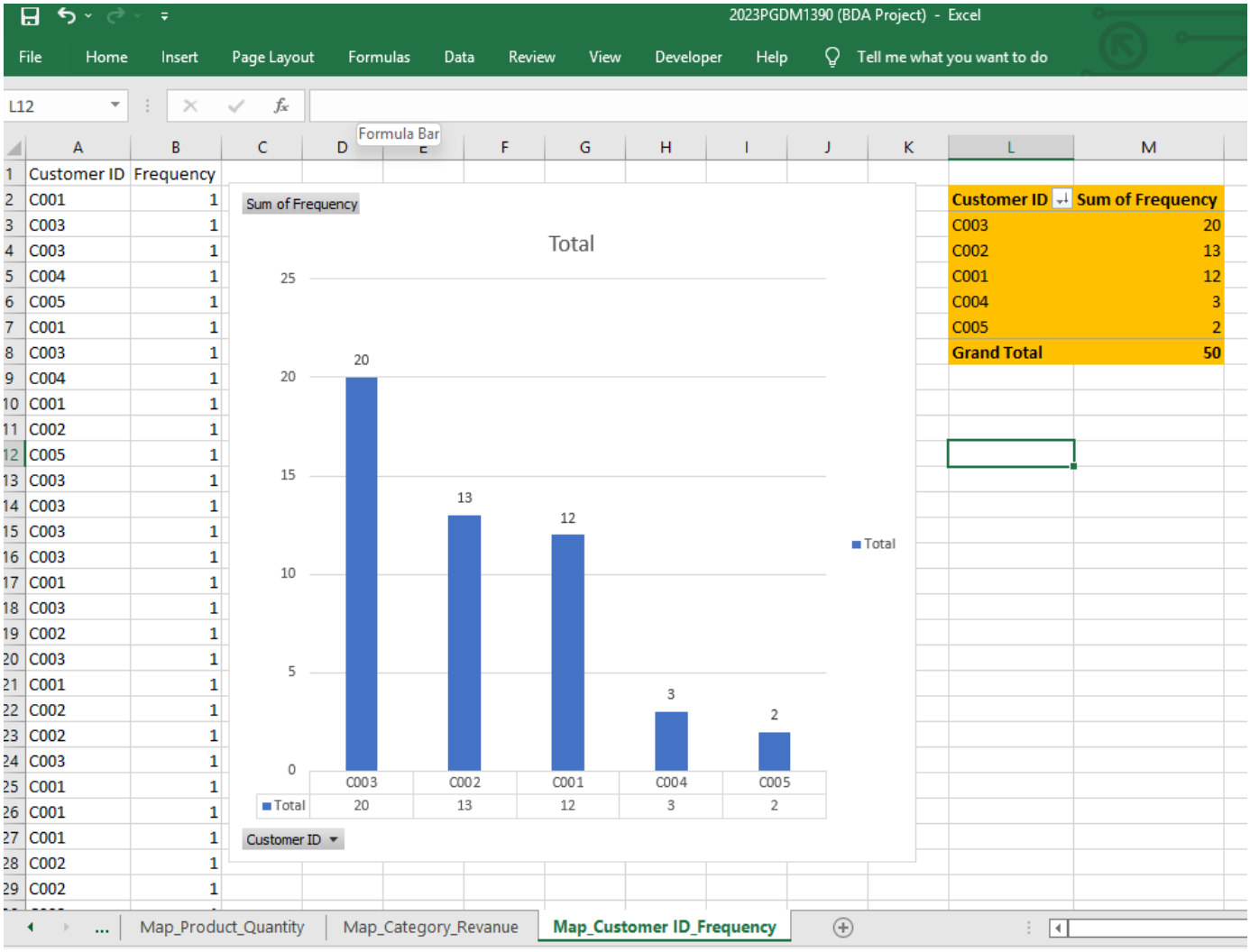
| Category | Sum of Revanue |
|---|---|
| Apparel | 124566 |
| Electronics | 101605 |
| Mobile | 88274 |
| Kitchen | 76060 |
| Home Appliance | 19892 |
| **Grand Total** | **410397** |

## B. Top 5 Customers by Frequency:

The pie chart was created to visualize the share of total revenue generated by each product category in the retail dataset. It helps management quickly understand which categories are contributing the most to overall sales and category 3 purchase the most to overall.

| Customer ID | Frequency |
|---|---|
| C003 | 20 purchases |
| C002 | 13 purchases |
| C001 | 12 purchases |
| C004 | 3 purchases |
| C005 | 2 purchases |

**Bar Chart**: Top 5 Products by Quantity Sold:



## Conclusion

This mini project showed how the MapReduce framework, a fundamental concept in big data analytics, can be simulated using Excel to analyze customer purchasing patterns. MapReduce is typically associated with complexion but this simulation used its logic and practical setting which is suitable for business users and we were able to understand and apply the same idea using simple tools.

### Data transformation (Map phase):

In this step, we changed the original transaction data into smaller, useful parts called **key-value pairs**. For example:

**I.** We looked at each product and how many times it was sold.
**II.** We looked at each category and how much money it earned.
**III.** We looked at each customer and how often they made purchases.

### Reduce Phase – Adding it All Together

In this step, we added all the useful information:

I. Total quantity sold for each product
II. Total number of purchases for each customer
III. Total revenue for each category

### Business insight extraction through simple tools

From this simulation, we were able to extract actionable insights such as:

**I.** Which products are most popular?
**II.** Which categories generate the most revenue?
**III.** Who the most loyal or frequent customers are?

These insights are crucial for business decision-making, helps managers to focus on inventory management, targeted promotions, and customer engagement strategies. Even without using a programming language or distributed infrastructure, Excel it allowed us to know valuable trends and patterns those are in the dataset.