

**Project Report On**

# Real Time E-commerce Dashboard

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA) From Know-IT(Pune)

**Guided by:**

**Anay Tamhankar Sir**

**Prasad Deshmukh Sir**

## Submitted By:

Aman Shrivastava (230943025005)

Pratik Reddy (230943025039)

Rushikesh Phapale (230943025041)

Saurabh Jambale (230943025044)



**CERTIFICATE**

**TO WHOMSOEVER IT MAY CONCERN**

**This is to certify that**

Aman Shrivastava (230943025005)

Pratik Reddy (230943025039)

Rushikesh Phapale (230943025041)

Saurabh Jambale (230943025044)

**Have successfully completed their project on.**

# Real Time E-commerce Dashboard

Under the guidance of Anay Tamhankar Sir and Prasad Deshmukh Sir



# ACKNOWLEDGEMENT

# This project “Real Time E-commerce Dashboard” was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of Anay Tamhankar Sir and Prasad Deshmukh Sir for his valuable guidance at work on this project. His guidance and support helped us to overcome various obstacles and intricacies during project work.

We are highly grateful to Mr. Vaibhav Inamdar Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks go to Mr. Prasad Deshmukh Sir (Course Coordinator-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

**From:**

Aman Shrivastava (230943025005)

Pratik Reddy (230943025039)

Rushikesh Phapale (230943025041)

Saurabh Jambale (230943025044)



# TABLE OF CONTENTS

**ABSTRACT**

1. **INTRODUCTION**
2. **DATA COLLECTION AND FEATURES**
3. **SYSTEM REQUIREMENTS**
   1. Software Requirements
   2. Hardware Requirements
4. **FUNCTIONAL REQUIREMENTS**
5. **SYSTEM ARCHITECTURE**
6. **WORKFLOW**
7. **DATA VISUALIZATION AND REPRESENTATION**
8. **FUTURE SCOPE**
9. **CONCLUSION**

**References**



# Abstract

Our Project Real Time E-commerce Dashboard focuses on developing a robust E-commerce. Analytics Dashboard by leveraging a sophisticated tech stack, including Kafka for real-time data streaming, Spark for efficient data processing, Tableau for advanced visualization, and MongoDB as an intermediary for seamless data flow between Tableau and Spark.

The project workflow is initiated by Python Faker, which generates realistic e-commerce data, setting the foundation for a diverse dataset.

The project architecture comprises two Kafka broker clusters (Broker 1 and Broker 2), providing a scalable and fault-tolerant environment for handling streaming data.

The generated e-commerce data is ingested into Kafka topics, serving as a resilient bridge for data communication between producers and consumers.

Apache Spark is employed to consume data from Kafka topics, ensuring efficient and parallelized processing of the diverse e-commerce dataset.

The computed insights are then pushed to MongoDB, functioning as a flexible and scalable data store. MongoDB acts as a crucial interface between Tableau and Spark, facilitating a seamless transfer of processed data for storage and retrieval.

Tableau, a powerful business intelligence tool, is utilized for real-time visualization of key performance indicators (KPIs) extracted from the e-commerce data.

The dynamic and interactive dashboards created in Tableau empower stakeholders to gain actionable insights into sales trends, customer behavior, and product popularity.

The project's unique workflow, incorporating Python Faker for data generation, Kafka for streaming, Spark for processing, MongoDB for storage, and Tableau for visualization, addresses the intricacies of real-time analytics in the dynamic e-commerce landscape.

The resulting E-commerce Analytics Dashboard offers a comprehensive solution for businesses aiming to enhance decision-making processes and gain a competitive edge in the market.



# INTRODUCTION

Real Time E-commerce Dashboard focuses on the development of a robust E-commerce Analytics Dashboard, leveraging a sophisticated tech stack to enable real-time visualization of key performance indicators (KPIs) extracted from e-commerce data. By integrating Python Faker for data generation, Kafka for real-time data streaming, Spark for efficient data processing, MongoDB for seamless data flow, and Tableau for advanced visualization, the project aims to address the complexities of real-time analytics in the dynamic e-commerce landscape. This unique workflow offers businesses a comprehensive solution to enhance decision-making processes and gain a competitive edge in the market.



# Dataset Collection and features

**Data Sources**

For our project, we generated a synthetic dataset simulating ecommerce data. The data was generated using Python libraries and tools, such as the Faker library for creating realistic data and MongoDB for storage. The decision to use simulated data was made to ensure data privacy and to have full control over the dataset's structure and content.

**Data Structure**

The dataset comprises several collections, each representing different aspects of the ecommerce ecosystem, such as customers, products, orders, and more. These collections are stored within a MongoDB database, providing a flexible and scalable storage solution.

**Dataset Size**

The generated dataset consists of approximately 100 records and can vary according to our requirement as we are generating our own data for each collection, resulting in a dataset of moderate size. Each collection contains specific information related to its domain, resulting in a rich and diverse dataset.

**Dataset with Features/Attributes**

Here is an overview of the key features (attributes) within our datasets:

**1.Customer Dataset**:

Attributes:

Customer ID, Name, Email, Age, Gender, Last Login, Account Created, Location

**2.Product Dataset:**

Attributes:

Product ID, Rating, Category, Name, Price, Supplier

**3.Transaction Dataset:**  
Attributes:

Transaction ID, Customer ID, Product ID, Category, Country, Payment Method, Pname, Price, Quantity, Region, Status, Supplier



# SYSTEM REQUIREMENTS

**Hardware Requirements**

* Computer: A computer with sufficient processing power and memory to run data processing and analysis tasks. A modern multicore processor and at least 8 GB of RAM are recommended.
* Storage: Adequate storage space to store the generated dataset and any additional datasets if required. An SSD (Solid State Drive) is recommended for faster data access.
* Internet Connection: A stable internet connection for downloading and installing software packages and libraries, as well as for any online resources needed during the project.

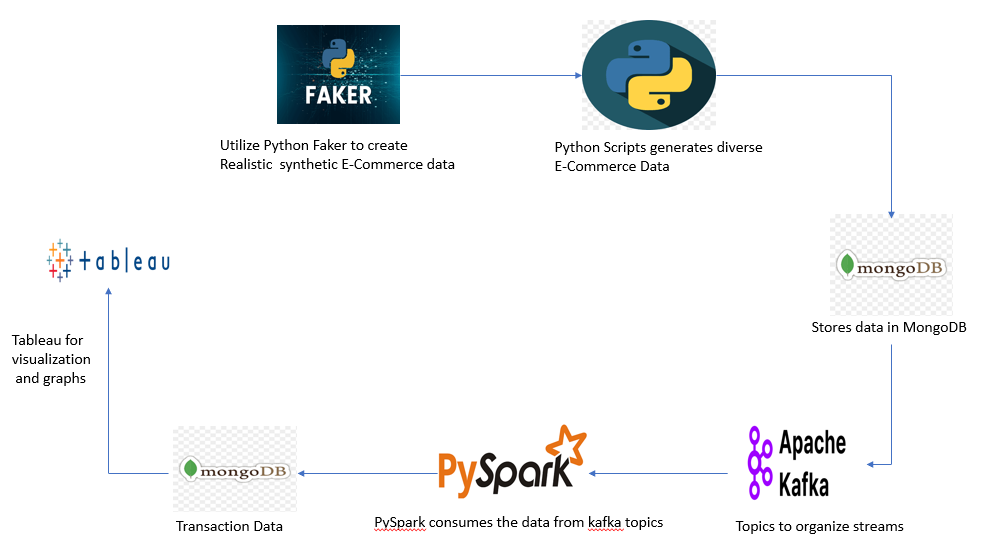
**Software Requirements**

* Operating System: Windows 10 or higher, Ubuntu
* Python: The project heavily relies on Python for data generation, analysis, and machine learning. Ensure Python is installed on your system.
* Python Libraries: Install the following Python libraries and dependencies using package managers like pip or conda:
  + NumPy: For numerical computing.
  + pandas: For data manipulation and analysis. Scikit learn: For machine learning tasks.
  + Matplotlib and Seaborn: For data visualization. Faker: For generating synthetic data.
* PyMongo: For interacting with MongoDB. Other libraries specific to your project's needs.
* MongoDB: Install MongoDB to store and manage the synthetic dataset. Ensure the MongoDB server is running.
* Apache Spark: If your project involves big data processing, consider installing Apache Spark. You can use PySpark to interact with Spark using Python.
* Kafka: If your project uses Kafka for real time data streaming, install and configure Kafka on your system.
* Integrated Development Environment (IDE): Choose a Python friendly IDE, such as PyCharm



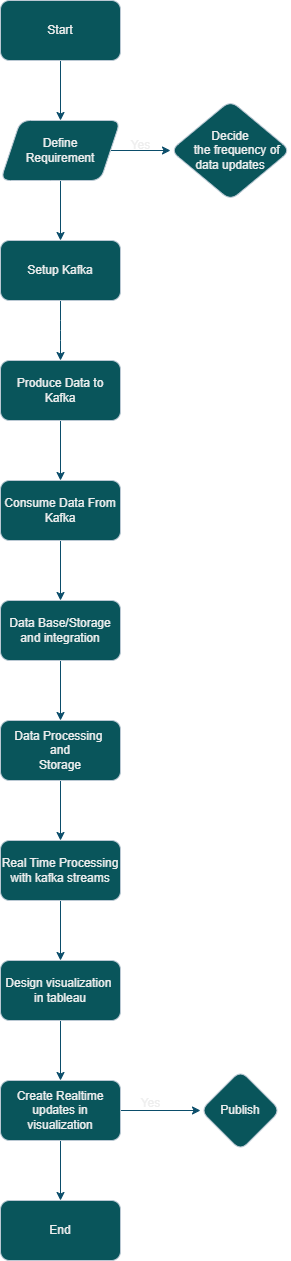
# FUNCTIONAL REQUIREMENTS

* **Python 3:**
* Python is a general purpose and high-level programming language.
* It is used for developing desktop GUI applications, websites, and web applications.
* Python allows focus on core functionality of the application by taking care of common programming tasks.
* Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, Small Talk, and Unix shell and other scripting languages.
* **PySpark:**
* PySpark is the Python API for Apache Spark.
* Used to perform real-time, large-scale data processing in a distributed environment using Python.
* PySpark provides a PySpark shell for interactively analyzing the data.
* PySpark supports all of Spark’s features such as Spark SQL, DataFrames, Structured Streaming, Machine Learning (MLlib) and Spark Core.
  + **Tableau:**
* Data visualization is the graphical representation of information and data.
* It helps create interactive elements like charts, graphs, and maps. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
* Tableau is widely used for Business Intelligence but is not limited to it.
* It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.
* All of this is made possible with gestures as simple as drag and drop.
* **Apache Kafka:**
* What is Kafka: Apache Kafka is an opensource stream processing platform and distributed event streaming platform developed by the Apache Software Foundation
* Key Features: Kafka is designed to handle real time data streams, making it a powerful tool for building and managing real time data pipelines, event driven architectures, and applications that require high throughput, fault tolerance, and scalability



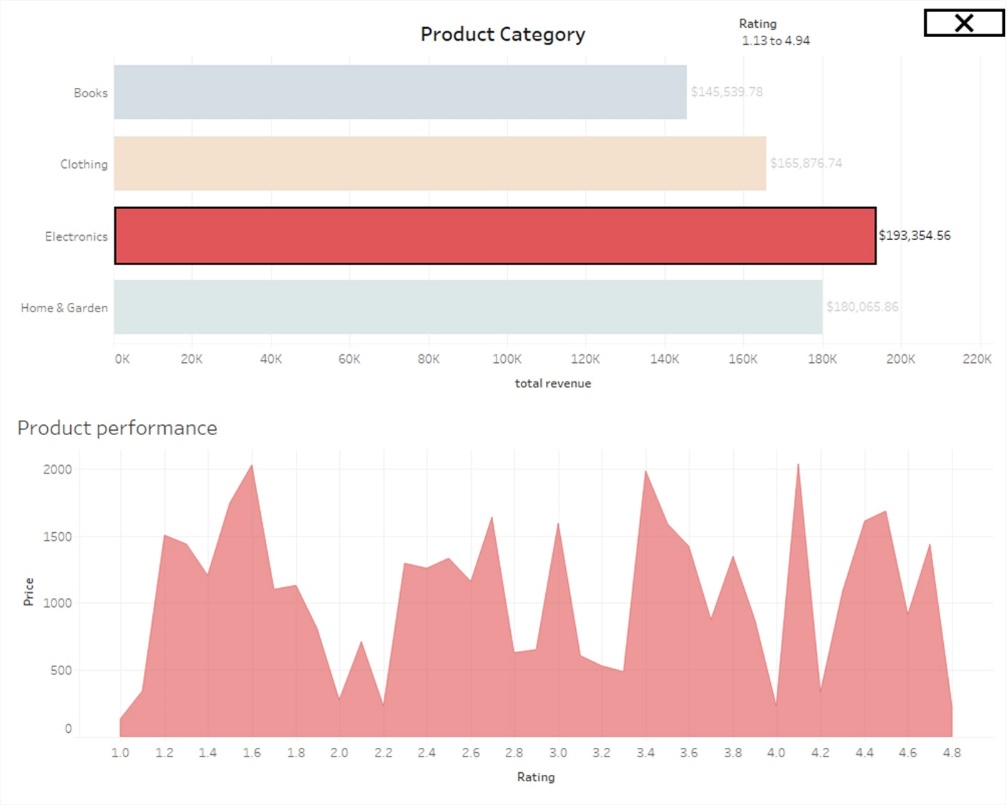
# SYSTEM ARCHITECTURE

**Fig: System Architecture Of “Real Time E-commerce Dashboard”**



# WORKFLOW

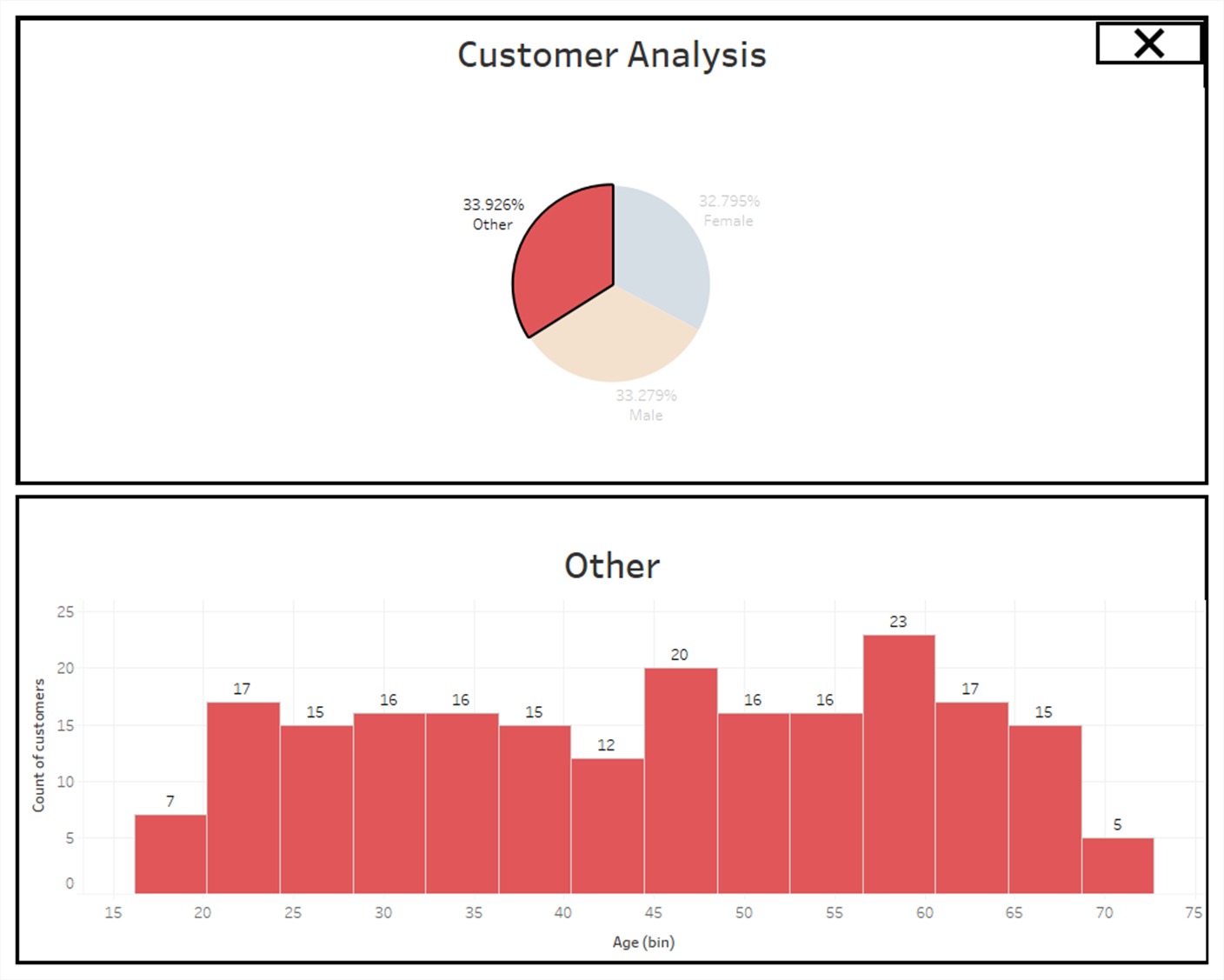
**Workflow of “Real Time E-commerce Dashboard”**



# DATA VISUALIZATION AND REPRESENTATION

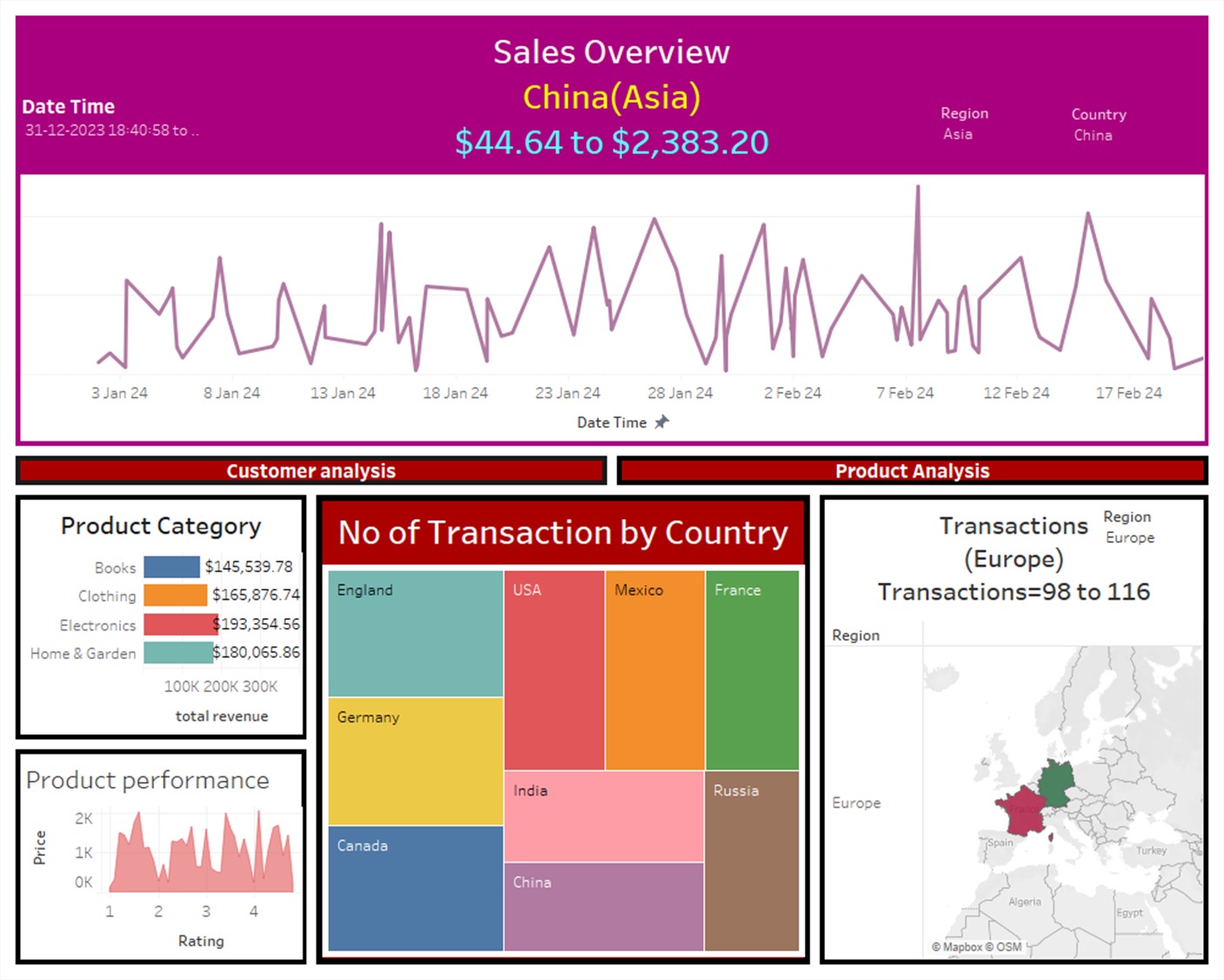
**FIG- PRODUCT ANALYSIS**

# DATA VISUALIZATION AND REPRESENTATION





**FIG:- CUSTOMER ANALYSIS DASHBOARD ACCORDING TO GENDER**



# DATA VISUALIZATION AND REPRESENTATION

FIG:- ANALYSIS OF TRANSACTIONS ON DIFFERENT PARAMETERS

# FUTURE SCOPE



The project, Real-Time E-commerce Dashboard, has laid a solid foundation for exploring various avenues of enhancement and expansion. While the current implementation has demonstrated the power of real-time data analysis in an e-commerce context, there are several exciting possibilities for future development and improvement:

**Integration of Real Data Sources:**

- The project primarily utilized simulated data for analysis. To make it more applicable to real-world scenarios, integrating real data sources from e-commerce platforms would be invaluable. This could involve web scraping, API integrations, or partnerships with e-commerce platforms.

**Interactive Dashboards:**

 Develop dashboards that dynamically adapt to user interactions and offer deeper insights through drill-down functionalities.

**Customer Segmentation:**

it involves dividing a customer base into groups that share similar characteristics, behaviors, or preferences. By analyzing customer segments, businesses can tailor their marketing strategies, product offerings, and customer experiences to better meet the needs of specific groups, ultimately improving customer satisfaction and increasing sales.

The future scope of this project is not limited to these suggestions; it extends as far as your imagination and business objectives. Continual adaptation and innovation are key to thriving in the dynamic e- commerce landscape, and leveraging data insights is central to this endeavor.

By exploring these avenues, businesses can remain competitive, adapt to changing customer behaviors, and make data-driven decisions that lead to sustainable growth and success.



# CONCLUSION

In the rapidly evolving landscape of e-commerce, real-time data analysis has emerged as a critical driver of success. This project, Real-Time Data Analysis on E-commerce Simulated Data, has explored the intricacies of data generation, collection, analysis, and visualization to simulate an e- commerce environment and glean actionable insights.

Our project has revolved around several key aspects:

**Data Generation and Simulation:**

We developed a robust data generation system capable of creating synthetic e-commerce data, encompassing customers, products, orders, reviews, and more. This simulated data served as a foundation for our analysis.

**Data Storage and Management:**

MongoDB, a NoSQL database, was employed to efficiently store and organize our data collections. We discussed various strategies for data modeling and how to establish relationships between different data entities.

**Real-Time Data Processing:**

- Apache Kafka played a pivotal role in our project, enabling the real-time streaming of data. It served as a conduit for data produced by the simulated e-commerce environment, mimicking real- world scenarios.

**Real-Time Data Processing:**

- Apache Kafka played a pivotal role in our project, enabling the real-time streaming of data. It served as a conduit for data produced by the simulated e-commerce environment, mimicking real- world scenarios.

**Data Visualization and Insights:**

- Visualizations created using Tableau showcased real-time data trends, providing businesses with valuable insights for decision-making.

.



# REFERENCES

* <https://spark.apache.org/docs/latest/api/python/index.html>
* <https://medium.com/@simardeep.oberoi/building-a-robust-real-time-e-commerce-analytics-pipeline-a-data-engineering-project-b57db9e9bfc4>
* <https://chat.openai.com/c/b6c94468-a142-4faf-9097-77c6f57d1b7e>
* <https://kafka.apache.org/>
* <https://www.tableau.com/support/help>