Project Report

On

Real Time Data Analysis on Ecommerce Simulated Data

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PGDBDA)

From KnowIT(Pune)

Guided by:

Anay Tamhankar Sir & Prasad Deshmukh Sir

Submitted By:

Vaqash Khan (230343025053)

Vivek Kumar Sahu (230343025054)

Rishikesh More (230343025033)

Rakesh Kumar Bharti(230343025058)

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Vaqash Khan (230343025053)

Vivek Kumar Sahu (230343025054)

Rishikesh More (230343025033)

Rakesh Kumar Bharti (230343025058)

Have successfully completed their project on

Real Time Data Analysis on Ecommerce Simulated Data

Under the guidance of Anay Tamhankar Sir and Prasad Deshmukh sir

ACKNOWLEDGEMENT

This project Real time data analysis on ecommerce simulated data was a great learning experience for us and we are submitting this work to CDAC KnowIT (Pune).

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

We are highly grateful to Mr. Vaibhav Inamdar Manager (KnowIT), CDAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PGDBDA) through CDAC ACTS, Pune.

Our most heartfelt thanks goes to Mrs. Bakul Joshi (Course Coordinator, PGDBDA) 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 CDAC KnowIT, Pune.

TABLE OF CONTENTS

Abstract

In the age of data-driven decision-making, our project dives into the world of real-time data analysis, focusing on e-commerce. By employing simulated data, we replicate the complexities of online retail, enabling us to develop and evaluate advanced analytics solutions. Through the integration of Apache Kafka, Apache Spark, MongoDB, and machine learning techniques, we construct an end-to-end data processing pipeline. This project explores data generation, storage, analysis, and visualization, providing valuable insights into customer behavior, sales patterns, and marketing strategies. Additionally, we discuss the potential applications, scalability, and the future of real-time data analysis in the ever-evolving e-commerce landscape.

1. INTRODUCTION

The digital era has transformed the way commerce operates, with ecommerce emerging as a driving force behind global business. The rapid growth of online retail has created an unprecedented volume of data that, when harnessed effectively, can provide valuable insights for businesses. Realtime data analysis has become a critical tool in this context, allowing enterprises to make datadriven decisions, enhance customer experiences, optimize operations, and gain a competitive edge.

This project, titled RealTime Data Analysis on Ecommerce Simulated Data, delves into the realm of ecommerce data analytics, focusing on the processing and analysis of simulated realtime data. While realtime analytics has become essential for ecommerce, obtaining real data for analysis can be challenging due to privacy concerns, data access limitations, and the need for extensive historical data. Simulated data offers a pragmatic solution, allowing us to replicate the complexities of realworld ecommerce data while maintaining control and flexibility.

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.

Features/Attributes

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

1**.Customers**:

Attributes:

Customer ID, First Name,Last Name, Email,Phone Number, Registration Date, Age,Gender,Location ,State, Postal Code, Country, Purchase Frequency ,Average Order Value, Recency of Purchase, Average Items per Order, Loyalty Program Membership, Online Shopping Open Rate of Emails ,ClickThrough Rate,Total Spending,Repeat Purchases,Social Media Engagement,Customer, Feedback

2. **Products:**

Attributes:

Product ID,Product Name,Description,Price,Category,Brand,SKU (Stock Keeping Unit),Stock Quantity

3. **Orders:**

Attributes:

Order ID,Customer ID (Reference to Customers collection),Order Date,Total Amount,Payment Method,Shipping Address (Embedded Address),Order Status

4. **Order Items:**

Attributes:

Order Item ID,Order ID (Reference to Orders collection),Product ID (Reference to Products collection),Quantity,Subtotal

5. **Reviews:**

Attributes:

Review ID,Customer ID (Reference to Customers collection),Product ID (Reference to Products collection),Rating,Text,Timestamp

6. **Refunds:**

Attributes:

Refund ID,Order ID (Reference to Orders collection),Refund Date,Refunded Amount,Reason

7**. Web Traffic**:

Attributes:

Timestamp,Page Visited,Referral Source

8. **Marketing Campaigns:**

Attributes:

Campaign ID,Campaign Name,Start Date,End Date

9. **Campaign Metrics:**

Attributes:

Metric ID,Campaign ID (Reference to Marketing Campaigns collection),Clicks,Impressions,Conversions

10**. Locations:**

Attributes:

Location ID,Country,State,City,Address,Street,Postal Code ,Coordinates (Latitude, Longitude)

11**. Inventory**:

Attributes:

Product ID (Reference to Products collection),Location ID (Reference to Locations collection),Quantity Available

2. SYSTEM REQUIREMENTS

Hardware Requirements

1. 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.

2. 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.

3. 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

1. Operating System: Windows 10 or higher

2. Python: The project heavily relies on Python for data generation, analysis, and machine learning. Ensure Python is installed on your system.

3. 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.

scikitlearn: 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.

4. MongoDB: Install MongoDB to store and manage the synthetic dataset. Ensure the MongoDB server is running

5. Apache Spark: If your project involves big data processing, consider installing Apache Spark. You can use PySpark to interact with Spark using Python.

6. Kafka: If your project uses Kafka for realtime data streaming, install and configure Kafka on your system.

7. Integrated Development Environment (IDE): Choose a Pythonfriendly IDE, such as PyCharm, Jupyter Notebook, Visual Studio Code, or your preferred text editor.

Visualization Software

1. Tableau: If you plan to visualize and analyze data with Tableau, install Tableau Desktop.

3. **FUNCTIONAL REQUIREMENTS**

(1) Python 3:

- Python is a general purpose and high level programming language.
- It is use for developing desktop GUI applications, websites and web applications.
- Python allows to focus on core functionality of the application by taking care of common programming tasks.
- Python is derived from many other languages, including ABC, Modula3, C, C++, Algol68, Small Talk, and Unix shell and other scripting languages.

(2) Apache Spark:

- What is Spark: Apache Spark is an opensource distributed computing system designed for processing large volumes of data.
- Key Features: Spark provides a number of key features that make it wellsuited for processing big data, including inmemory processing, support for various data sources and formats, faulttolerance, and scalability.
- Spark also provides a range of APIs, including SQL, streaming, machine learning, and graph processing, making it a versatile platform for a wide range of use cases.

(3) 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 realtime data streams, making it a powerful tool for building and managing realtime data pipelines, eventdriven architectures, and applications that require highthroughput, fault tolerance, and scalability.

Overall, Apache Kafka serves as a backbone for realtime data processing, providing the infrastructure needed to handle large volumes of data, support eventdriven architectures, and enable applications with lowlatency requirements.

(4) 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.

ARCHITECTURE

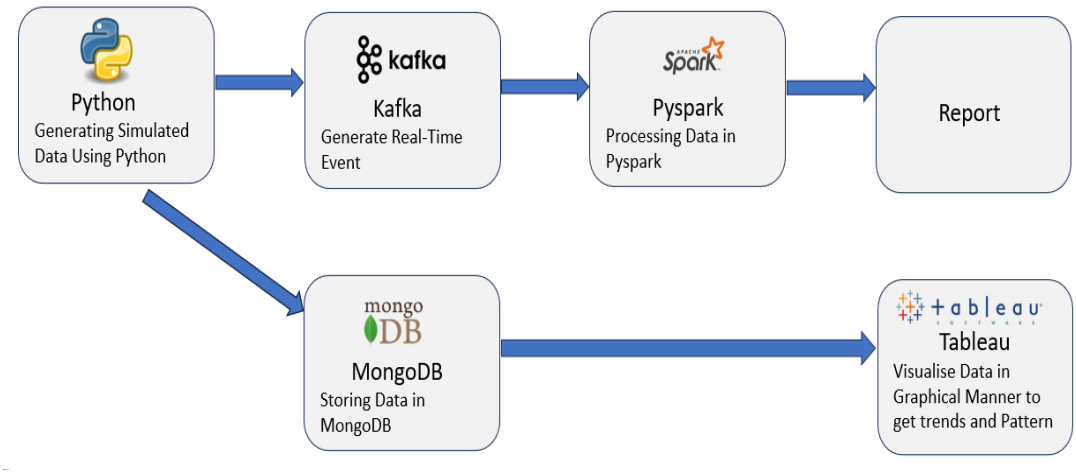


Fig: System Architecture Of Real time Data Analysis on ecommerce simulated data

MACHINE LEARNING ALGORITHMS

KMeans Clustering

Algorithm Overview:

Customer segmentation is a vital task in marketing and ecommerce analytics. It involves dividing customers into distinct groups based on shared characteristics, enabling businesses to tailor marketing strategies and services more effectively. In our project, we employ the KMeans Clustering algorithm for customer segmentation.

Explanation:

KMeans Clustering is an unsupervised machine learning algorithm used for partitioning data into 'K' distinct, nonoverlapping clusters or groups. Each cluster represents a set of data points that are similar to each other based on selected features. Here's how KMeans Clustering works in our project:

1. **Feature Selection:**

We select relevant features from our customer dataset, such as purchase history, demographics, and behavioral data. These features help define the similarity between customers.

2. **Normalization:**

Before applying KMeans, we normalize the data to ensure that features with different scales do not dominate the clustering process.

3. **KMeans Algorithm:**

We initialize 'K' centroids (cluster centers) randomly within the feature space.

The algorithm iteratively assigns each customer to the nearest centroid based on a chosen distance metric (usually Euclidean distance).

After all customers have been assigned, the centroids are recalculated as the mean of all data points within their respective clusters.

Steps 2 and 3 are repeated until convergence, i.e., until the centroids no longer change significantly.

4. **Cluster Interpretation:**

Once clustering is complete, customers are grouped into 'K' clusters. Each cluster represents a segment of customers who exhibit similar behaviors or characteristics.

5. **Business Insights:**

These customer segments can provide valuable business insights. For example, one cluster might contain highvalue, frequent customers, while another might represent occasional shoppers. Marketing strategies, product recommendations, and promotions can then be customized for each group.

Benefits:

Personalized Marketing: Customer segmentation allows businesses to create targeted marketing campaigns. For instance, high value customers can be offered loyalty rewards, while price conscious shoppers can receive discounts.

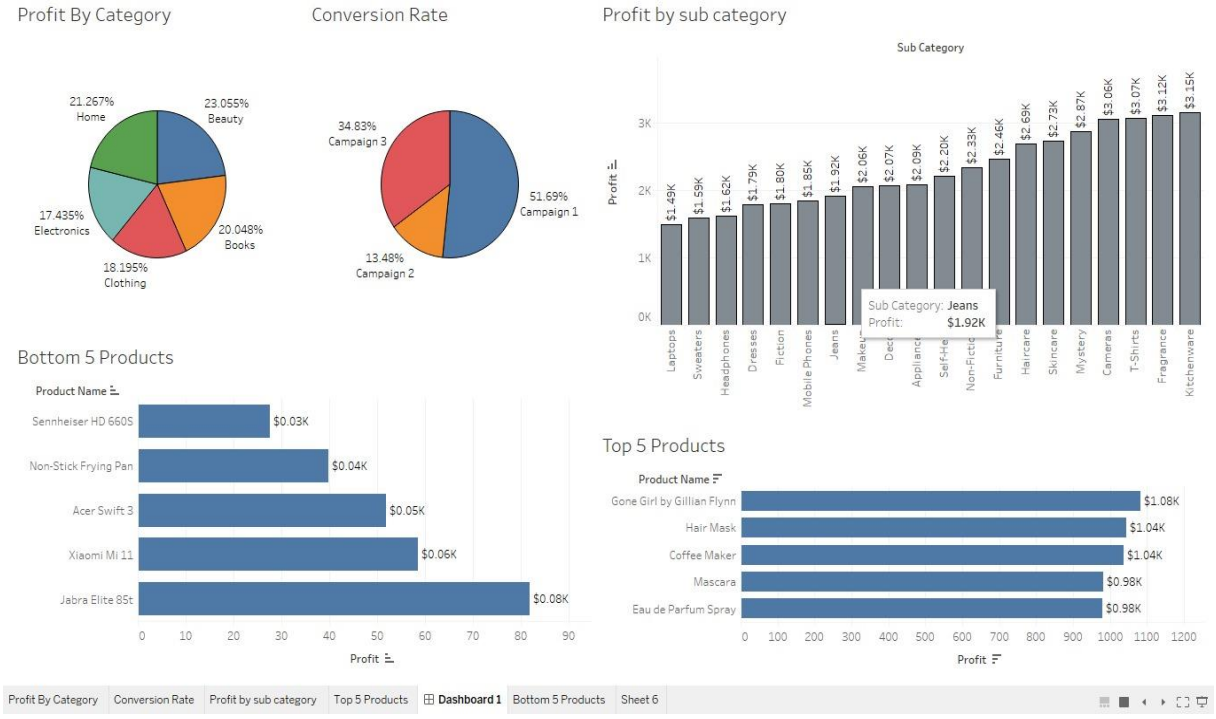
Improved Product Recommendations: Understanding customer segments helps enhance product recommendations. If a segment consists of tech enthusiasts, tech-related products can be suggested.

Enhanced Customer Experience: Tailored services and support can be provided based on customer segments, leading to higher satisfaction and retention.

Conclusion:

K-Means Clustering is a powerful tool for customer segmentation in our project. It helps businesses gain insights into customer behavior, preferences, and needs, ultimately driving more effective marketing and sales strategies.

DATA VISUALIZATION AND REPRESENTATION



CONCLUSION AND FUTURE SCOPE

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:

1. 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.

2. 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.

3. 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.

4. Data Analysis with Apache Spark:
   - Apache Spark, a powerful data processing framework, facilitated data analysis. We demonstrated how to extract, transform, and load data from MongoDB into Spark DataFrames, allowing for in-depth analysis.

5. Machine Learning and Customer Segmentation:
   - We utilized the K-Means Clustering algorithm to perform customer segmentation. This ML technique partitioned customers into distinct groups, enabling businesses to personalize marketing strategies.

6. Data Visualization and Insights:
   - Visualizations created using Tableau showcased real-time data trends, providing businesses with valuable insights for decision-making.

7. Future Scope and Enhancements:

- Our project has vast potential for expansion. Future enhancements could include incorporating real data sources, implementing more advanced machine learning models, and deploying predictive analytics for sales forecasting.

8. Business Relevance:
- The project underscored the immense value of real-time data analysis in e-commerce. From personalized marketing to tailored product recommendations, the insights derived from this analysis can significantly impact business success.

In conclusion, Real-Time Data Analysis on E-commerce Simulated Data exemplifies the power of data-driven decision-making in the e-commerce sector. Through simulated data, innovative technologies, and advanced analytics, this project offers a glimpse into the future of data-driven e-commerce, where businesses can thrive by understanding and responding to customer needs in real time.

As the e-commerce landscape continues to evolve, the insights gained from this project serve as a solid foundation for businesses seeking to leverage data to gain a competitive edge.

Future Scope:

The project, Real-Time Data Analysis on E-commerce Simulated Data, 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:

1. 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.

2. Advanced Machine Learning Models:

- Enhance the machine learning component by incorporating more advanced algorithms such as Random Forests, Gradient Boosting, or Neural Networks. These models can provide deeper insights into customer behavior, product recommendations, and sales forecasting.

3. Predictive Analytics:

- Implement predictive analytics to forecast future sales trends and customer behavior. Predictive models can help businesses anticipate demand, optimize inventory management, and plan marketing campaigns effectively.

4. Real-Time Monitoring and Alerts:

- Develop a real-time monitoring system that alerts businesses to significant events or anomalies in their e-commerce data. For instance, sudden spikes in website traffic, unusually high shopping cart abandonment rates, or a surge in product reviews.

5. A/B Testing and Personalization:

- Incorporate A/B testing frameworks to evaluate the effectiveness of different marketing strategies, website designs, or product placements. Additionally, invest in personalized recommendation systems to enhance the customer shopping experience.

6. Customer Sentiment Analysis:

- Extend the analysis to include sentiment analysis of customer reviews and social media interactions. Understanding customer sentiment can help businesses gauge public opinion, identify areas for improvement, and manage their brand reputation effectively.

7. Mobile Application Integration:

- Develop a mobile application that provides customers with a personalized shopping experience, real-time updates on orders, and seamless integration with the e-commerce platform. Mobile apps can significantly enhance customer engagement.

8. International Expansion:

- If the e-commerce platform operates in multiple countries, consider expanding the analysis to accoun for global sales trends, regional preferences, and currency exchange rates.

9. Cloud-Based Scalability:

- Migrate the data processing and storage components to cloud platforms like AWS, Azure, or Google Cloud. This ensures scalability, high availability, and cost-efficiency as data volumes grow.

10. Regulatory Compliance:

- Stay up-to-date with evolving data privacy regulations such as GDPR and CCPA. Implement data protection measures and ensure compliance with these regulations to maintain trust with customers.

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.

References

1. https://towardsdatascience.com/top-3-python-packages-to-generate-synthetic-data-33a351a5de0c

2. Apache Spark. [https://spark.apache.org/]

3. MongoDB. [https://www.mongodb.com/]

4. Kafka. [https://kafka.apache.org/]

5. Python. [https://www.python.org/]

6. scikit-learn. [https://scikit-learn.org/]

7. Confluent Python Client. [https://docs.confluent.io/platform/current/clients/confluent-kafka-python/html/index.html]

8. Faker. [https://faker.readthedocs.io/en/master/]