Big Data Systems

Lab 1: Big Data Applications Study

Student Name: Soham Dave

Guided By : Dr. Jaiprakash Verma

Abstract—Study and explore various applications of big data in different domains. Choose one of them and study in detail. Also, write down the report on different types of digital data generated in selected applications. For example: Big Data in Retail, Big Data in Healthcare, Big Data in Education, Big Data in E-commerce, Big Data in Media and Entertainment, Big Data in Finance, Big Data in Travel Industry, Big Data in Telecom.

Keywords—*Big Data Applications, Domain Applications, Retail Analytics, Healthcare Analytics, Education Analytics, E-commerce Analytics, Media Analytics, Finance Analytics, Travel Analytics, Telecom Analytics, Digital Data Types, Data Generation*

1. Introduction

The rapid growth of online shopping has transformed e-commerce into one of the most data-intensive industries today. Every click, search, purchase, review, and shipment generates vast volumes of structured and unstructured data, making e-commerce a prime domain for big data applications. From customer Browse patterns to real-time inventory updates, the digital footprint left by users and systems provides invaluable insights for businesses aiming to enhance user experience, optimize supply chains, and drive sales through data-driven decision-making.

This study focuses on big data applications in the e-commerce sector, leveraging publicly available datasets that capture various aspects of online retail operations. These include product catalogs, user profiles, behavioral logs, transaction histories, customer reviews, ratings, shipping details, payment methods, inventory levels, and marketing campaign outcomes. Platforms such as Kaggle and Hugging Face host rich datasets from global retailers—ranging from Brazilian e-marketplaces like Olist to international giants like Amazon, Flipkart, and Shein—enabling comprehensive analysis of consumer behavior and business performance.

By examining these datasets, we explore how big data technologies facilitate key e-commerce functionalities such as personalized recommendation engines, sentiment analysis of customer feedback, churn prediction, dynamic pricing, fraud detection in payments, and logistics optimization. Furthermore, the integration of machine learning models with large-scale transactional and behavioral data allows for advanced analytics, including customer segmentation, lifetime value prediction, and demand forecasting.

This report presents an overview of selected e-commerce datasets, discusses the types of digital data they contain, and illustrates how big data tools and frameworks can be applied to extract meaningful insights. The goal is to demonstrate the transformative impact of data analytics in shaping modern e-commerce strategies and improving both operational efficiency and customer satisfaction.

2. E-commerce Data: Description and Analysis

E-commerce platforms generate a diverse range of digital data, which can be broadly categorized based on the business function it supports. The following sections provide a detailed breakdown of these data types, their sources, and their applications.

2.1. Product and Site Data

This category encompasses all data related to the products themselves and the structure of the e-commerce site.

- **Description**: This data includes product names, descriptions, images, specifications, features,

benefits, materials, and nutritional information. It also includes information about the site itself, such as FAQs, policies, and base webpage content.
- **Source**: This information is typically sourced from internal databases, content management systems (CMS), and product information management (PIM) systems.
- **Datasets**:
  - E-commerce Data (Kaggle): This dataset contains a list of products with descriptions and other details.
  - Flipkart Products (Kaggle): A rich dataset with product names, descriptions, and categories from one of India's largest e-commerce platforms.
  - Products E-commerce Embeddings (Hugging Face): This dataset provides product descriptions alongside their vector embeddings, useful for semantic search and recommendation systems.
- **Advantages**: Provides a comprehensive understanding of the product catalog. Essential for building search functionality, product recommendation engines, and dynamic advertising content.
- **Disadvantages**: Data can be static and requires frequent updates to reflect changes. Descriptions can be inconsistent or incomplete, requiring significant data cleaning.
- **Insights and Improvements**: Analyzing product data can reveal popular keywords, helping optimize search engine results (SEO) and product discovery. Improvements can include using natural language processing (NLP) to standardize product descriptions and automatically generate tags for better categorization.

2.2. User Data

This data is centered around customer interactions and profiles.

- **Description**: It includes personal details (e.g., name, email, address), behavioral data (e.g., clicks, views, time spent on pages), and engagement data (e.g., sign-ups, subscriptions). It also includes customer feedback data from surveys or support interactions.
- **Source**: Internal user databases, web analytics tools (e.g., Google Analytics), and CRM (Customer Relationship Management) systems.
- **Datasets**:
  - Users E-commerce (Hugging Face): Provides user-specific data that can be used for segmentation and personalization.
  - E-commerce Customer Service (Hugging Face): Contains interactions with customer service, useful for sentiment analysis and identifying pain points.
  - Customer Analytics (Kaggle): Includes demographic and behavioral data for customer segmentation.
- **Advantages**: Enables personalized experiences, targeted marketing, and customer segmentation. This data is critical for predicting churn and calculating customer lifetime value (CLV).
- **Disadvantages**: Requires careful handling due to privacy concerns and regulatory compliance (e.g., GDPR). Data can be messy, with duplicate or incomplete user profiles.
- **Insights and Improvements**: User data can be used to build recommendation engines and predict user behavior. To improve the dataset, integrating real-time user activity streams could provide more immediate insights.

2.3. Orders and Shipments Data

This category covers the entire order fulfillment lifecycle.

- **Description**: This data includes order tracking information, order fulfillment details, shipping costs, delivery dates, and the current status of a shipment.

- **Source**: Order Management Systems (OMS) and logistics partners' tracking APIs.
- **Datasets**:
  - Amazon Seller Order Status Prediction (Kaggle): Contains data to predict the final status of an order.
  - Shipping (Hugging Face): Focuses on shipping details, including delivery times and costs.
  - Olist Brazilian E-commerce (Kaggle): This dataset is a rich source of order, shipment, and review data from a Brazilian e-commerce platform.
- **Advantages**: Optimizes logistics, supply chain management, and delivery routes. Enables proactive customer service by providing real-time tracking information.
- **Disadvantages**: Data can be highly dynamic and sensitive to external factors like weather or traffic. Integrating data from multiple logistics partners can be complex.
- **Insights and Improvements**: Analyzing this data can help identify bottlenecks in the supply chain. Predicting order delivery times can improve customer satisfaction. The dataset could be improved by including more granular data on warehouse operations and carrier performance.

2.4. Inventory and Sales Data

This involves the management of products and financial transactions.

- **Description**: This data includes inventory levels, product availability, stock-keeping units (SKUs), sales figures, and pricing information.
- **Source**: Inventory management systems (IMS) and point-of-sale (POS) systems.
- **Datasets**:
  - Grocery Inventory (Kaggle): Provides inventory and stock data for a grocery retailer.
  - Transactional E-commerce (Kaggle): Contains detailed transaction records, including sales and order data.
- **Advantages**: Facilitates demand forecasting and inventory optimization, minimizing stockouts and overstocking. Supports dynamic pricing strategies.
- **Disadvantages**: Requires real-time updates to be effective. Inaccurate data can lead to significant financial losses.
- **Insights and Improvements**: Analyzing inventory data alongside sales trends helps in forecasting demand for seasonal products. Enhancements could include integrating supplier data to improve supply chain visibility and reduce lead times.

2.5. Payments Data

This deals with all financial transactions on the platform.

- **Description**: Includes payment methods, transaction details, and fraud scores.
- **Source**: Payment gateways and internal financial systems.
- **Datasets**:
  - Transactional E-commerce (Kaggle): A good source for payment and transaction data.
- **Advantages**: Crucial for financial reporting and fraud detection. Ensures secure and smooth transactions.
- **Disadvantages**: Requires strict security protocols to protect sensitive financial information.
- **Insights and Improvements**: This data can be used to build machine learning models for real-time fraud detection. The dataset could be improved by including more features related to the transaction context, such as user location and device type, to enhance fraud models.

2.6. Marketing Data

This category tracks the performance of marketing activities.

- **Description**: This includes data from marketing campaigns, social media interactions, and customer surveys. It measures performance indicators such as click-through rates (CTR), conversion rates, and return on ad spend (ROAS).
- **Source**: Ad platforms (e.g., Google Ads, Facebook Ads), social media APIs, and survey tools.
- **Datasets**:
  - Retail Sales Data with Marketing (Kaggle): This dataset combines sales data with marketing spend.
  - Superstore Marketing Campaign Dataset (Kaggle): Focuses on marketing campaign effectiveness.
  - Social Media Marketing Data (Hugging Face): Provides social media campaign data.
- **Advantages**: Optimizes marketing spend and campaign targeting. Enables A/B testing and performance analysis.
- **Disadvantages**: Data from different platforms can be siloed, making unified analysis challenging. Requires careful attribution modeling to measure effectiveness.
- **Insights and Improvements**: Analyzing this data helps optimize campaigns and personalize marketing messages. Enhancements could include integrating customer feedback from social media to refine marketing strategies.

2.7. Analytics Data

This is the output of data processing and analysis.

- **Description**: This includes performance indicators (KPIs), dashboards, and insights derived from other data types.
- **Source**: Business intelligence (BI) tools and data warehouses.
- **Datasets**: This is typically an output, not a raw dataset.
- **Advantages**: Provides a holistic view of the business, enabling data-driven decision-making.
- **Disadvantages**: The quality of the insights depends entirely on the quality of the raw data.
- **Insights and Improvements**: Analytics can reveal hidden trends and correlations. To improve, we can incorporate machine learning models to generate predictive analytics, such as demand forecasting and churn probability, which adds a layer of proactivity to the business strategy.

3. Conclusion

The e-commerce domain is a rich source of big data, offering vast opportunities for analytics and business optimization. By leveraging a wide array of datasets—from product descriptions to user behavior logs and marketing campaign results—businesses can gain a competitive edge. The provided datasets from platforms like Kaggle and Hugging Face serve as excellent resources for studying these applications. By analyzing and enhancing these datasets, businesses can improve everything from personalized recommendations and supply chain efficiency to fraud detection and marketing effectiveness. This study underscores the critical role of big data in modern e-commerce, highlighting how a data-driven approach is essential for growth, customer satisfaction, and operational excellence.

4. Tables and figures

4.1. Tables

Check the following table for the datasets :

Table 1. E-commerce Datasets from Kaggle and Hugging Face

Sr. No.	Dataset Link	Description
1.	https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce	Comprehensive data from a Brazilian e-commerce platform.
2.	https://www.kaggle.com/datasets/carrie1/ecommerce-data	E-commerce data containing transactional details.
3.	https://www.kaggle.com/datasets/surajjha101/bigbasket-entire-product-list-28k-datapoints	Product list of a large grocery retailer.
4.	https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews	Customer reviews and ratings for Amazon products.
5.	https://www.kaggle.com/datasets/prachi13/customer-analytics	Data for customer segmentation and analytics.
6.	https://www.kaggle.com/datasets/palvinder2006/zepto-inventory-dataset	Inventory data for a quick commerce grocery service.
7.	https://www.kaggle.com/datasets/willianoliveiragibin/grocery-inventory	Inventory and stock data for a grocery retailer.
8.	https://www.kaggle.com/datasets/pranalibose/amazon-seller-order-status-prediction	Data to predict the final status of an Amazon order.
9.	https://www.kaggle.com/datasets/PromptCloudHQ/flipkart-products	Product descriptions and categories from Flipkart.
10.	https://www.kaggle.com/datasets/bytadit/transactional-ecommerce	Detailed transactional and sales records.
11.	https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset	Marketing campaign performance data.
12.	https://www.kaggle.com/datasets/abdullah0a/retail-sales-data-with-seasonal-trends-and-marketing	Sales data combined with marketing campaign information.
13.	https://www.kaggle.com/datasets/trainingdatapro/shein-e-commerce-dataset	Product and user data from the fashion retailer Shein.
14.	https://huggingface.co/datasets/manumartinm/users_ecommerce	User-specific data for e-commerce platforms.
15.	https://huggingface.co/datasets/saattrupdan/womens-clothing-ecommerce-reviews	Reviews of women's clothing from an e-commerce site.
16.	https://huggingface.co/datasets/LukeSajkowski/products_ecommerce_embeddings	Product descriptions with vector embeddings.
17.	https://huggingface.co/datasets/TrainingDataPro/asos-e-commerce-dataset	E-commerce dataset from the fashion retailer ASOS.
18.	https://huggingface.co/datasets/qgyd2021/e_commerce_customer_service	Customer service interactions and feedback.
19.	https://huggingface.co/datasets/withpi/social-media-marketing-data-v01-formatted_alt_questions	Social media marketing campaign data.
20.	https://huggingface.co/datasets/aeroplayer/shipping	Data focused on shipping costs and delivery times.

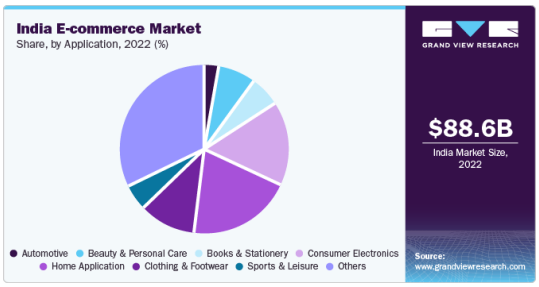
Note: This table lists a selection of publicly available datasets for e-commerce big data analysis.

4.2. Figures

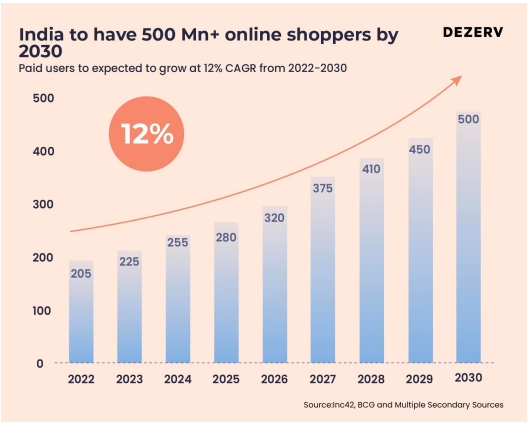
Fig. 1 shows the types of data set that we have covered:

Figure 1. Various types of Data in eCommerce Platforms.

Fig. 2 shows two graphs about category-wise bestselling online products in India, and projection of the growth of eCommerce platforms' revenue in India till 2030.



(a) eCommerce products categories.



(b) Growth of eCommerce in India.

Figure 2. Statistical Graphs about Data revolving around eCommerce business in India