Big Data Systems

Lab 1: Big Data Applications Study for eCommerce Platforms

Student Name: Soham Dave

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.

1. Introduction

Online shopping has turned e-commerce into one of the most data-heavy industries today. Every click, search, purchase, review, and shipment generates massive amounts of information, giving businesses valuable insights to improve user experience, optimize supply chains, and boost sales.

This study looks at big data applications in e-commerce using public datasets from platforms like Kaggle and Hugging Face. These datasets cover product catalogs, user activity, transaction histories, reviews, payments, shipping, inventory, and marketing results, with sources ranging from Brazilian retailers like Olist to global giants like Amazon, Flipkart, and Shein.

Through these datasets, we explore how big data powers personalized recommendations, sentiment analysis, churn prediction, dynamic pricing, fraud detection, and logistics optimization. Machine learning further enhances this by enabling customer segmentation, lifetime value prediction, and demand forecasting.

Overall, the report highlights how data analytics is reshaping e-commerce strategies, driving efficiency, and enhancing customer satisfaction.

2. E-commerce Data: Description and Analysis

E 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.
- **How does it justify the 5 V of Big Data?**:

Volume	Millions of SKUs and users on platforms like Amazon/Flipkart. Each item contributes product details (names, descriptions, images, specs, attributes), while users generate purchase history, searches, reviews, and interactions—together creating massive data volumes.
Velocity	Continuous real-time updates: transactions, stock levels, pricing adjustments, clickstream data, new launches, promotions, and recommendations—all processed instantly to avoid overselling, enable dynamic pricing, and optimize user experience.
Variety	**Structured:** product IDs/SKUs, names, categories, prices, stock levels, sales logs, customer demographics, transaction records.
	Unstructured: product descriptions, customer reviews, survey responses, chat logs, images, videos.
	Semi-structured: API/JSON/XML feeds from suppliers, competitor pricing data, clickstream/session logs, event tracking.
Veracity	Data errors lead to critical issues: wrong inventory counts cause overselling/stockouts; incorrect prices or descriptions trigger returns and complaints. Data deduplication, validation, integrity checks across IMS, POS, and e-commerce systems are vital.
Value	Big data analysis enables keyword mining for SEO, personalized recommendations, targeted marketing, demand forecasting, fraud detection, logistics optimization, and improved customer segmentation—directly boosting efficiency, engagement, and profitability.

Table 1

- **References**:
  - 7 Ways Big Data is Changing E-commerce (Case Study by Talend)
  - IDC: Expect 175 zettabytes of data worldwide by 2025 (Case Study by NetworkWorld)

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.
- **How does it justify the 5 V of Big Data?**:

Volume	Millions of users generate massive data volumes: personal details, browsing history, purchase history, searches, reviews, and support interactions. Each profile adds up over time, and multiplied by millions of active users, the scale becomes huge.
Velocity	User data is produced continuously and often processed in real-time. Clicks, views, and searches enable instant recommendations, dynamic content, and immediate follow-ups (e.g., abandoned cart emails).
Variety	**Structured:** Personal details (name, email, address), demographic info, purchase history (items, dates, prices).
	Unstructured: Customer reviews, survey responses, support chat logs, feedback text.
	Semi-structured: Web analytics logs, clickstream data, session details (time spent, navigation paths).
Veracity	Accuracy of user data is critical. Incorrect addresses cause delivery failures; wrong logs produce irrelevant recommendations. The verification of duplication, validation, and authenticity of feedback is essential for reliability.
Value	User data drives recommendations, targeted marketing, churn prediction, CLV estimation, and customer segmentation—boosting engagement, sales, and ROI while guiding strategy.

Table 2

- **References**:
  - The Role of Big Data in E-commerce: Opportunities and Challenges (Article by Explore S.M.M)
  - Big Data: Powering the Digital Revolution (Article by Lucent Innovation)

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.
- **How does it justify the 5 V of Big Data?**:

Volume	Each order generates large amounts of data: status updates (processing, packed, shipped, delivered), tracking numbers, shipping costs, and delivery dates. At scale—millions of daily orders—this quickly becomes enormous.
Velocity	Order and shipment data change constantly and must be updated in real-time or near real-time for accurate tracking, customer service, and logistical adjustments.
Variety	**Structured:** Order IDs, shipment IDs, tracking numbers, shipping costs, delivery dates, delivery addresses, carrier details, order status codes (e.g., processing, shipped, delivered).
	Unstructured: Customer inquiries about order/shipment status, internal logistics team notes on delivery issues or exceptions.
	Semi-structured: API responses from logistics and courier partners (JSON/XML formats), real-time tracking updates.
Veracity	Accuracy is crucial. Wrong addresses or tracking numbers cause failed deliveries; inaccurate dates reduce trust. Data reconciliation across logistics partners is a major challenge.
Value	Order and shipment analytics enable supply chain optimization, better customer service, cost reduction, and demand forecasting—driving efficiency and satisfaction.

Table 3

- **References**:

  - Big data analytics in E-commerce: a systematic review and agenda for future research (ResearchGate Article)
  - Recommender Systems in E-Commerce (ResearchGate Article)

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.
- **How does it justify the 5 V of Big Data?**:

Volume	Vast product catalogs with numerous SKUs generate huge data: inventory levels, sales figures across multiple channels, dynamic pricing info, and long-term sales history—all accumulating rapidly.
Velocity	Inventory and sales data update at high speed. Purchases instantly adjust stock counts; flash sales and dynamic pricing require real-time recalculations based on demand or competitor prices.
Variety	**Structured:** SKU numbers, product IDs, stock levels, sales transaction records, COGS, pricing rules.
	Semi-structured: Supplier stock feeds, competitor pricing data (JSON/XML).
	Unstructured: Customer reviews, market trend data impacting product performance.
Veracity	Errors in inventory counts cause overselling or stockouts. Inaccurate sales records distort demand forecasting. Synchronizing IMS, POS, and e-commerce platforms is a key integrity challenge.
Value	Inventory and sales analytics enable demand forecasting, smarter purchasing, optimized pricing, and efficient stock management—boosting profitability and customer satisfaction.

Table 4

- **References**:

  - The Role of Big Data in Personalizing the Customer Experience (TechnoRely Article)
  - Research on E-commerce Data Analysis Algorithms and Applications Based on Artificial Intelligence(ACM Digital Library Research-Article)
  - Dynamic Pricing in Ecommerce: A Guide for Businesses of All Sizes (By Nected AI)

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.
- **How does it justify the 5 V of Big Data?**:

Volume	Every purchase on an e-commerce platform generates a payment transaction. For large platforms processing millions of transactions daily, the sheer number of records (payment method, transaction details, amounts, dates, times, fraud scores) rapidly accumulates into a massive volume of data.
Velocity	When a customer attempts a purchase, the payment gateway processes the transaction instantly. Fraud detection systems analyze transaction details in milliseconds to approve or decline the payment, demonstrating the critical need for high-velocity data processing.
Variety	**Structured:** Transaction IDs, amounts, payment method types (credit card, PayPal, etc.), currency, timestamps, merchant IDs, authorization codes. This typically comes from payment gateways and internal financial systems.
	Semi-structured: API responses from credit card processors or fraud detection services, sometimes including risk scores or reason codes.
	Unstructured: While not directly payments data, linking payment details with user location, device type, IP address, and purchase history adds crucial context for robust fraud detection.
Veracity	Incorrect transaction amounts can lead to disputes. False positives in fraud detection can block legitimate customers, while false negatives can lead to financial fraud. Maintaining the accuracy of sensitive financial information and ensuring it's not tampered with is crucial.
Value	Analyzing payments data provides critical insights for financial reporting, optimizing payment processes, and, most importantly, detecting and preventing fraud, directly impacting profitability and platform security.

Table 5

- **References**:
  - Fraud Detection at Scale — Stripe Radar Case Study
  - Using Big Data Analytics to Combat Payment Fraud in E-commerce (Research Gate)

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.
- **How does it justify the 5 V of Big Data?**:

Volume	A single social media campaign can generate millions of impressions and thousands of clicks and engagements. Running multiple campaigns across platforms like Google Ads, Facebook Ads, and email marketing for a year generates petabytes of data that need to be collected and analyzed. According to Systems Plus, a logistics startup, with the help of data analytics, reduced fuel costs and optimized reloads.
Velocity	Campaigns generate impressions, clicks, and engagements in real-time across multiple platforms. Data arrives continuously from ads, emails, and customer interactions, requiring instant collection and processing to monitor performance and optimize decisions.
Variety	**Structured:** Campaign performance metrics (CTR, ROAS, cost per click) from advertising platforms, customer survey responses (ratings, scores), sales data linked to campaigns.
	Semi-structured: Web analytics logs, JSON data from APIs.
	Unstructured: Social media posts, comments, customer feedback from open-ended survey questions, transcripts of chatbot interactions.
Veracity	Accurately attributing a sale to the correct marketing touchpoint (e.g., the specific ad that led to the conversion) is complex. Data discrepancies between platforms, bot traffic impacting click data, and incomplete customer profiles can severely affect the veracity of insights, leading to misinformed decisions. A ReBid article highlights that data fragmentation, when customer data is scattered across various systems and databases, can hinder personalization and understanding of customer behavior.
Value	The primary goal of analyzing marketing data is to extract actionable insights that optimize marketing spend, improve campaign targeting, and ultimately drive revenue and customer loyalty.

Table 6

- **References**:
  - 8 case studies and real world examples of how Big Data has helped keep on top of competition
  - Social Media Analytics For Brands: The Secrets of Data-Driven
  - Addressing the Challenge of Data Fragmentation in Marketing (ReBid Article)

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. Figures

Fig. 4 shows the types of data set that we have covered:

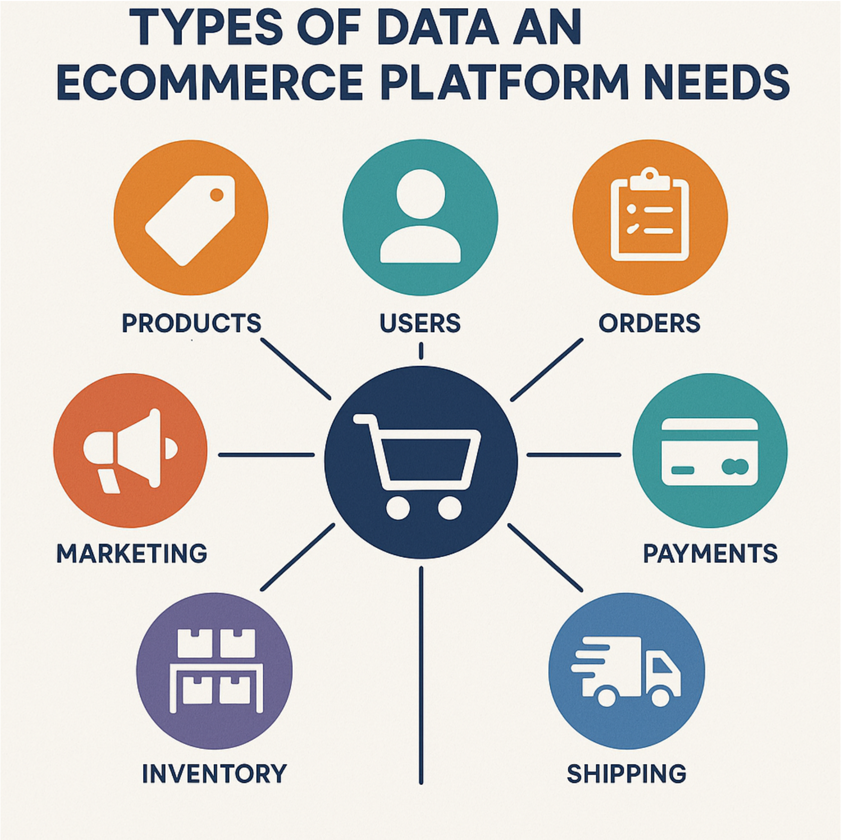
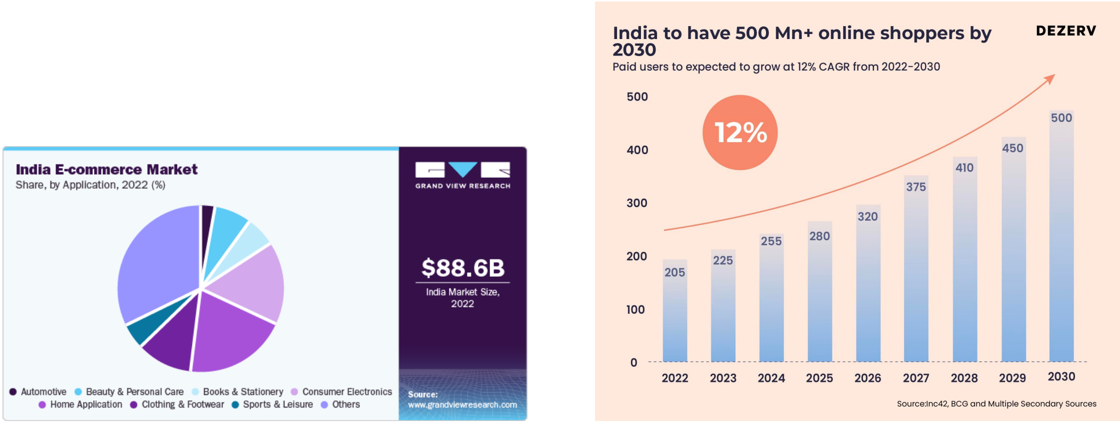


Figure 1. Various types of Data in eCommerce Platforms.

Fig. 2 shows two graphs about category-wise bestselling online products in India, and a projection of the growth of the revenue of e-Commerce platforms in India until 2030.



(a) eCommerce products categories.

(b) Growth of eCommerce in India.

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

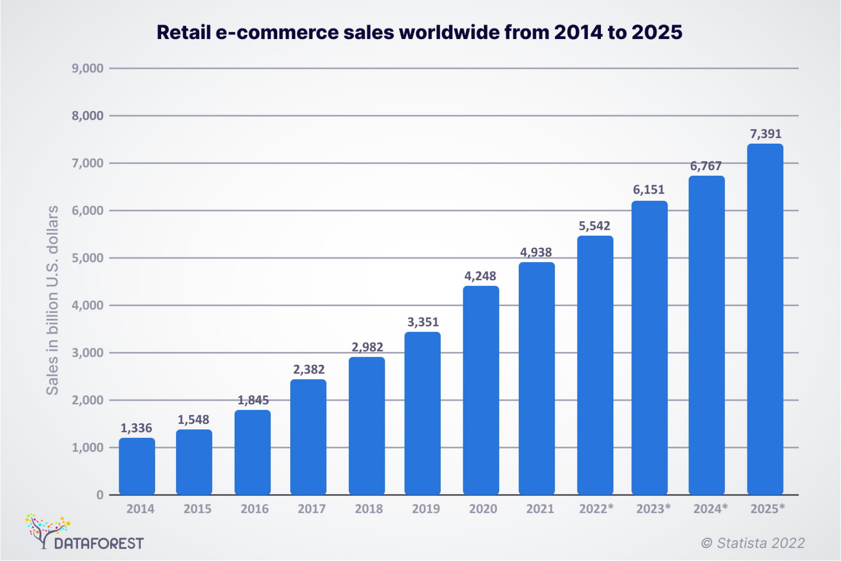


Figure 3. worldwide retail ecommerce sales

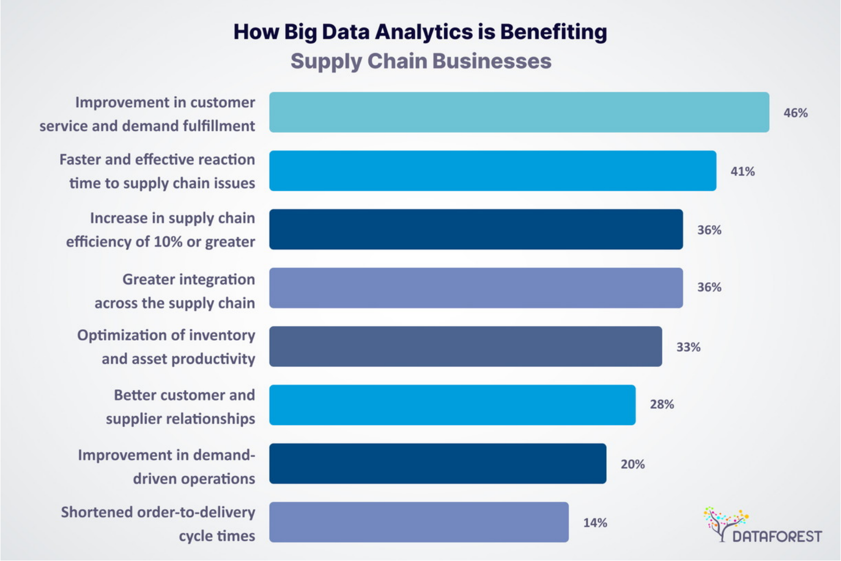


Figure 4. How big data analytics is helping supply chain businesses

4.2. Tables

Check the following table for the datasets :

Table 7. 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.