E-Commerce Behavior Data Analysis Using Hadoop

Authors: Prathima Sarvani Alla, Vamsi Yarramreddy, Jeremy Esquivias,
Abhiram Natakarani

Department of Information Systems, California State University Los Angeles
CIS5200 System Analysis and Design
palla3@calstatela.edu, vyarram@calstatela.edu,jesqui35@calstatela.edu,
anataka@calstatela.edu

Abstract

The ecommerce dataset is a valuable resource for data analysts seeking to understand consumer behavior and trends in online shopping. This dataset contains a vast array of information, including product information, purchase history, and website engagement metrics.

By analyzing this data, businesses can gain insights into customer preferences, identify opportunities for growth, and optimize their online sales strategies.

This term paper aims to analyze behavior data for one day in the month of October 2019 from a large multi-category online store. Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users.

1. Introduction

The aim of this project is to analyze e-commerce dataset using Hadoop and Hive, concentrating on consumer behavior, tastes, and purchasing trends. We used the Kaggle-hosted e-commerce Dataset. The dataset was utilized because it offers information on customer behavior, including information on whether they examine products, add them to their carts, or buy them. By analyzing this dataset, we can learn important things about customer behavior and spot trends, patterns, and linkages by examining this collection. Determine which products clients frequently purchase in combination. Identify the products that the same consumer regularly buys at the same time. To uncover pricing trends and to make your prices more profitable, analyze the price and purchase history. The dataset is also ideal for use in big data analysis projects due to its size of 5 GB. Our analysis looks at corporate organizations and their major issues, the increasing rate of complaints year over year, and the overall attitude of the consumer. A typical ecommerce dataset contains data on the goods sold by an online shop as well as details on the clients who have made purchases there. This dataset may contain a variety of data items, such as usernames, user ids, and customer information, in addition to product information like names, descriptions, pricing, and categories.

Online retailers frequently use ecommerce datasets to learn more about consumer behavior, preferences, and buying habits. Retailers may improve their pricing tactics, expand their product lines, and boost consumer satisfaction by studying the data.

Additionally, researchers and data scientists use ecommerce analytics to create machine learning models and algorithms for forecasting consumer behavior, identifying fraud, and improving supply chain management. Online marketplaces like Amazon and eBay, as well as individual online stores and systems like Shopify and WooCommerce, are among common sources of ecommerce datasets. These datasets could be collected through collaborations with retailers or e-commerce platform providers, or they could be made publicly available.

To protect data privacy and security, it's crucial to keep in mind that ecommerce datasets may contain sensitive information, such as consumer payment information. Working with ecommerce datasets requires adherence to data protection laws like GDPR and CCPA to preserve client privacy. By offering insights into consumer behavior, customer happiness, customer lifetime value, sales success, and product performance, data analysis in ecommerce experimentation aids in business growth.

Customer segmentation: Use customer data analysis to divide customers into groups according to their characteristics, actions, and patterns of consumption. You can use this to target client groups with tailored marketing messages and promotions.

Product recommendations: Consider each customer's preferences by using their purchasing and browsing patterns as well as past purchases.

Pricing optimization: Examine sales data to spot pricing trends, then adjust prices to raise revenue. Utilize sales data to estimate demand and manage inventory levels so that you have adequate inventory to satisfy consumer demand while also reducing inventory costs.

We visualized the data using Tableau Software and Power BI, for sentiment analysis. Overall, our analysis provides a unique perspective on the E-commerce dataset and offers valuable insights into consumer behavior and trends in the industry.

1.1 Related Work

- Customer behavior patterns over two months, providing valuable insights for manufacturers and retailers to adjust product prices and gain customer trust. The analysis reveals that customers predominantly view products rather than make purchases. Smartphones emerge as the most popular category, with purchases primarily occurring around 9 o'clock. Furniture Bench and Jackets are the least purchased categories in both October and November. Samsung stands out as the most popular and purchased brand during this period, while Besafe and Ava are the least purchased brands. Casio is the most viewed but not purchased brand. User ID - 564068124 has made the most purchases. These insights can be utilized by business owners to enhance revenue and develop new strategies for increased profitability. Future work can involve analyzing customer behavior over an extended period. We visualized the data using Tableau Software and Power BI, for sentiment analysis. Overall, our analysis provides a unique perspective on the E-commerce dataset and offers valuable insights into consumer behavior and trends in the industry.

- By utilizing data to produce insights and address business issues, big data analytics (BDA) is fostering innovation and competitiveness in e-commerce. Leading e-commerce companies have embraced BDA and seen tremendous development, including Google, Amazon, and Facebook. This study offers a thorough analysis of the main facets of BDA and serves as a solid base for future studies in the field of growing e-commerce. It emphasizes how crucial data, sources, abilities, and systems are in forging a competitive edge. The study also underlines the necessity of defining BDA's scope, comprehending the many kinds of big data, and addressing difficulties in order to maximize its commercial worth. BDA helps businesses to use data to its fullest extent and to produce insights quickly, allowing them to succeed.

-

2. Specifications

This file contains behavior data for the month of October 2019 from a large multi-category online store. Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users. Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users. The size of the dataset is 5GB.

Below Table 1 shows files and size of the files from dataset.

Table 1 Data Specification

Data Set Size	5 GB
Number for files	1
Content Format	CSV

The Table 2 below shows the specification for Oracle cluster we are using and Hadoop specification for our project.

Table 2 H/W Specification

Number of nodes	5 (2 master nodes, 3 worker nodes)
CPU speed	1995.312 MHz
Storage	390 GB

3. Implementation Flowchart

The raw dataset, which contains the specifics of consumer activity data, was initially downloaded from Kaggle. This example illustrates the entire data manipulation process. The supplied dataset is in CSV format. The dataset was obtained in CSV format, and it was then uploaded to the Hadoop File System. The tables' schema is then created, the data is cleaned, a summary table is created, and the results are exported using the querying language HiveQL. After downloading the output file in CSV/xlsx, we utilized Power BI and Tableau to create the visuals.

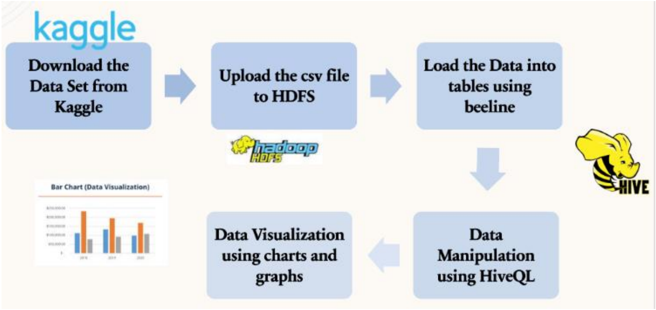
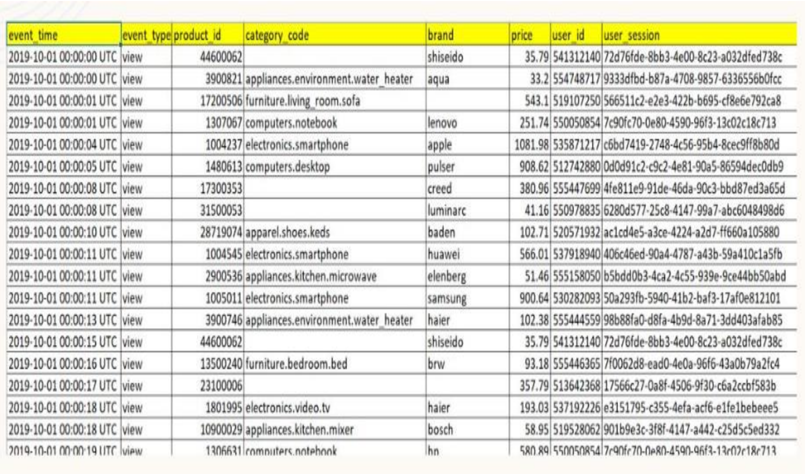

kaggle
Download the Data Set from Kaggle
Upload the csv file to HDFS
hadoop HDFS
Load the Data into tables using beeline
HIVE
Data Manipulation using HiveQL
Data Visualization using charts and graphs
Bar Chart (Data Visualization)

Figure 1 Architecture Flow Chart



event_time	event_type	product_id	category_code	brand	price	user_id	user_session
2019-10-01 00:00:00 UTC	view	44600062		shiseido	35.79	541312140	72d76fde-8bb3-4e00-8c23-a032dfed738c
2019-10-01 00:00:00 UTC	view	3900821	appliances.environment.water_heater	aqua	33.2	554748717	9333dfbd-b87a-4708-9857-6336556b0fcc
2019-10-01 00:00:01 UTC	view	17200506	furniture.living_room.sofa		543.1	519107250	566511c2-e2e3-422b-b695-cf8e6e792ca8
2019-10-01 00:00:01 UTC	view	1307067	computers.notebook	lenovo	251.74	550050854	7c90fc70-0e80-4590-96f3-13c02c18c713
2019-10-01 00:00:04 UTC	view	1004237	electronics.smartphone	apple	1081.98	535871217	c6bd7419-2748-4c56-95b4-8cec9ff8b80d
2019-10-01 00:00:05 UTC	view	1480613	computers.desktop	pulser	908.62	512742880	0d0d91c2-c9c2-4e81-90a5-86594dec0db9
2019-10-01 00:00:08 UTC	view	17300353		creed	380.96	555447699	4fe811e9-91de-46da-90c3-bbd87ed3a65d
2019-10-01 00:00:08 UTC	view	31500053		luminarc	41.16	550978835	6280d577-25c8-4147-99a7-abc6048498d6
2019-10-01 00:00:10 UTC	view	28719074	apparel.shoes.keds	baden	102.71	520571932	ac1cd4e5-a3ce-4224-a2d7-ff660a105880
2019-10-01 00:00:11 UTC	view	1004545	electronics.smartphone	huawei	566.01	537918940	406c46ed-90a4-4787-a43b-59a410c1a5fb
2019-10-01 00:00:11 UTC	view	2900536	appliances.kitchen.microwave	elenberg	51.46	555158050	b5bdd0b3-4ca2-4c55-939e-9ce44bb50abd
2019-10-01 00:00:11 UTC	view	1005011	electronics.smartphone	samsung	900.64	530282093	50a293fb-5940-41b2-baf3-17af0e812101
2019-10-01 00:00:13 UTC	view	3900746	appliances.environment.water_heater	haier	102.38	555444559	98b88fa0-d8fa-4b9d-8a71-3dd403afab85
2019-10-01 00:00:15 UTC	view	44600062		shiseido	35.79	541312140	72d76fde-8bb3-4e00-8c23-a032dfed738c
2019-10-01 00:00:16 UTC	view	13500240	furniture.bedroom.bed	brw	93.18	555446365	7f0062d8-ead0-4e0a-96f6-43a0b79a2fc4
2019-10-01 00:00:17 UTC	view	23100006			357.79	513642368	17566c27-0a8f-4506-9f30-c6a2ccbf583b
2019-10-01 00:00:18 UTC	view	1801995	electronics.video.tv	haier	193.03	537192226	e3151795-c355-4efa-acf6-e1fe1bebeee5
2019-10-01 00:00:18 UTC	view	10900029	appliances.kitchen.mixer	bosch	58.95	519528062	901b9e3c-3f8f-4147-a442-c25d5c5ed332
2019-10-01 00:00:19 UTC	view	1306631	computers.notebook	hp	580.89	550050854	7c90fc70-0e80-4590-96f3-13c02c18c713

Figure 2 Sample Dataset

4. Data Cleaning

Beeline Client was used to load raw files into tables once they had been uploaded to HDFS and stored there. The dataset doesn't need to be completely clean because there aren't many NULL or missing values. It does, however, have some cleaned-up

duplicate values. The Category Code has been divided into three categories, subcategories, and products.
Example:
Category Code: appliances.environment.water_heater
Category: Appliances
Subcategory: Environment
Product: Water Heater

Water_Heater is the category code. Environment is a subcategory of the category of appliances. The item is a water heater. And we filled the hive table with the entire file. Deleted all records with null values for Category and Brand to create a new table.

5. Analysis and Visualization

Data was cleaned up, and files were extracted into Excel and Power BI in order to prepare them for more analysis. To display information based on views, carts, and purchases using a pie chart, we employed several interactive maps. A pie chart is used to show the categories that are frequently purchased. A 3D map is used to display popular brands in highly purchased categories. Using Power BI, the type of event is visually analyzed together with its location and timing.

5.1 3D Map in Excel

The first visualization Figure 2 Category Analysis, a 3D map, was made in Excel and it is an animated map with a timeline for one month, December 2020. This visualization uses a bubble chart to represent each state on a map, with the size of the bubble indicating the sentiment count (i.e., number of complaints) in that state. The layers with different colors of the bubbles indicate the product category such as apparel, appliances, auto, computers, and electronics. The map is arranged by state and the time element is represented by one month of date received, allowing us to analyze cart, view, and purchase trends over time.

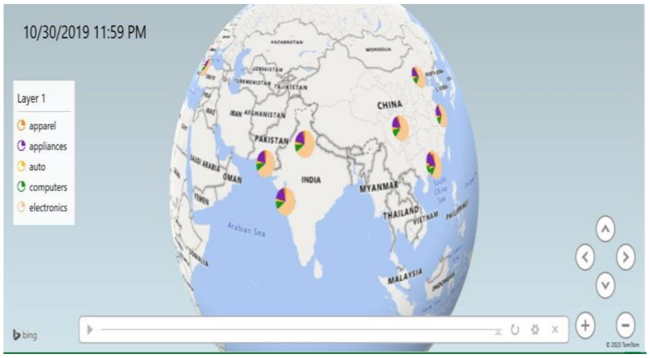


Figure 2 Event Type, Location, Time and Top Categories (Excel 3D Map)

5.2 Power BI

For this research paper utilized Power BI, a robust data visualization tool, to analyze ecommerce data. By executing a query that extracted eventtype, location, and eventtime from the dataset, we gained valuable insights as shown below in *Figure 3*. The visual analysis confirmed a significant finding: a consistent level of product viewing activity throughout the day. This insight suggests the importance of optimizing product information and browsing features to enhance customer experiences and drive sales. By leveraging Power BI's capabilities, businesses can make data-driven decisions, improve customer satisfaction, and guide strategic planning in the dynamic landscape of ecommerce.

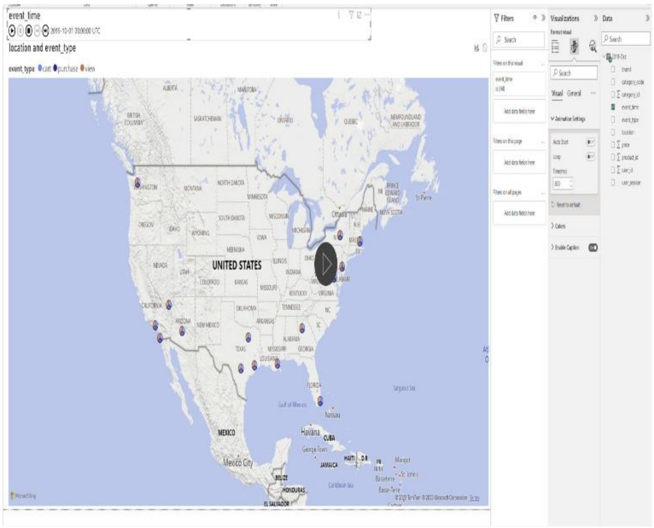


Figure 3 Tempo Spatial Analysis Using Power BI

5.3 Weekly Revenue Report Graphs

The 4-week line graphs in the Weekly Report for October 2019's revenue breakdown visually represent the data from Purchases. The first week displayed a decrease in sales on October 4th, followed by an increase during the weekend. The subsequent week showed relatively balanced sales at the start, a slight decline, and a recovery towards the end. The purchase trend in the third week closely resembled that of the second week. The final week of the month exhibited a volatile purchase trend, with an initial upswing and a sharp decline in sales that persisted throughout the weekend, culminating in the lowest number of purchases overall. Overall, the month of October 2019 experienced significant fluctuations in sales revenue, reflecting a volatile purchase trend. This highlights the challenges faced by businesses seeking stable sales revenue and emphasizes the importance of proactive measures to mitigate the impact of unpredictable market conditions on consumer behavior.

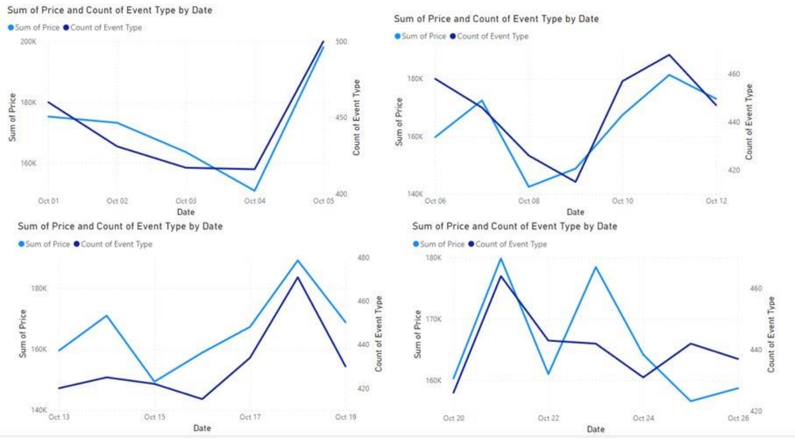


Figure 4 Weekly Revenue Report (Week 1-4)

5.4 Popular Brands in Highly Purchased Category Bar Graph

The visual analysis, in *Figure 5* below. revealed that Samsung and Apple are the most popular brands within the highly purchased category of electronics. This suggests that customers are drawn to these brands due to their desirable features, quality, and strong reputation. The accompanying bar chart provides valuable insights into customer preferences, enabling businesses to understand which brands are favored by their target audience. Furthermore, this information can guide businesses in identifying less popular brands, allowing them to introduce gift bundles and implement targeted marketing strategies to boost sales. By leveraging this visual analysis, businesses can align their offerings with customer preferences and make informed decisions to enhance their competitiveness in the market.

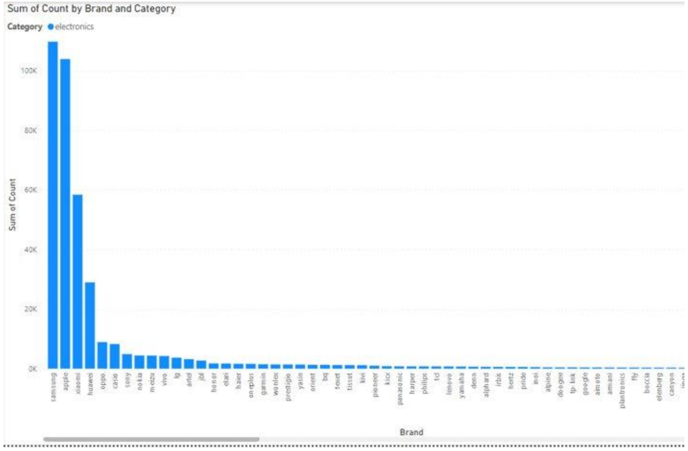


Figure 5 Popular Brands in Highly Purchased Category

5.4 Event Type of Each Category Bar Graph Analysis

The following analysis of the electronics category revealed that only a small percentage (2.5%) of users made a purchase. Additionally, 3.2% of users in this category added items to their cart. Most users engaged in product views without completing a purchase or adding items to their cart. These findings highlight the significance of understanding customer behavior and identifying potential areas for improvement in the conversion rate. Businesses in the electronics category can leverage these insights to optimize their strategies, enhance user experience, and increase the likelihood of conversion by addressing barriers or incentives that may be hindering purchases.

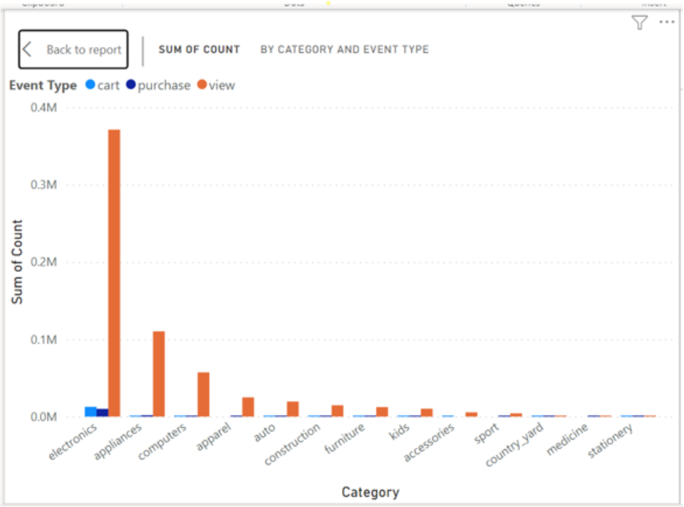


Figure 6 Views, Cart, and Purchases of Each Category

6. Conclusion

In summary, this research paper has provided an overview of ecommerce data, focusing on key aspects such as categories with high purchase activity, popular brands within those categories, and the tempo-spatial distribution of customer engagement. Using visualizations, we have presented valuable insights that can guide decision-making processes, drive revenue growth, and enhance customer targeting and satisfaction. By understanding which categories experience the most sales, identifying popular brands, and analyzing customer engagement patterns across different locations and timeframes, businesses can make data-driven decisions that optimize their online sales strategies. These visualizations serve as powerful tools that enable companies to stay competitive in the dynamic landscape of ecommerce, adapt to evolving consumer preferences, and ultimately achieve success in the digital marketplace.

For more information, dashboards, and code visit the project's GitHub link[2].

References

[1] Lab Tutorial Documents

[2] https://chat.openai.com/

[3] eCommerce behavior data from multi category store | Kaggle: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store?resource=download

[4] YouTube.com