Analysis of Ecommerce Behavior Using Apache Hadoop

Authors: Jonathan B. Reyes, Jose Ramirez, David Gomez Tagle, Kelvin Odii & Jing Zhu
Department of Information Systems, California State University - Los Angeles
CIS5200 System Analysis and Design
jreyes175@calstatela.edu, jramire3@calstatela.edu, dgomezt@calstatela.edu, kodii@calstatela.edu

Abstract

Global e-commerce sales are projected to cross $4 trillion by 2020. That represents the overall sales in the retail industry. The underlying fact is that more shoppers are now ready to pursue online retail shopping for a range of products or services rather than in person shopping. This change of behavior is presented within this analysis that demonstrated what category of products are most purchased between the months of October 2019 and April 2020. Still considering that large retail brands such as Walmart have reported a 29% increase in U.S. eCommerce sales. In recent years, consumer behavior analysis in the e-commerce industry has emerged as an effective analytical tool for knowing how any online shopper interacts with an e-commerce website.

For any online retailer, data containing online consumer behavior is one of the most important business assets that a retailer can leverage. Through behavior data analysis, e-commerce stores can extract deeper consumer information and present a more personalized website experience to these customers and further target consumers with specific types of products at specific times.

Consumer behavior analysis is a data-powered observation of online consumers and how they interact with companies. Behavioral analysis in e-commerce can be used to categorize consumers of the following behaviors:

1. How is the online consumer attracted to a particular e-commerce product?
2. What products do consumers buy?
3. When are purchases made?

Purchase behavior provides detailed insights into consumer needs and interest and more accurate indicators of consumer behavior. Data sets related to purchasing behavior can be deployed to determine buying patterns, how consumers respond to promotional activities, like product discounting or special offers and email advertisement announcements.

ECommerce retailers can derive multiple benefits from the insights taken from consumer behavior analysis tools and dashboards. These valuable insights can lead to a more personalized approach to customer needs that can increase their lifetime value to the business. Further, behavioral analysis in eCommerce can reduce customer acquisition costs, improve brand recommendations, and improve the lead generation process for sales and marking.

1. Introduction

Utilizing Apache Hadoop IOP 4.2, an analysis is presented to understand ecommerce behavior. The purpose of this analysis is to identify categories of items that are most frequently purchased, items that are most frequently purchased, and the time the items are most frequently purchased. Analysis of this consumer behavior can provide valuable insight in marketing and sales within the ecommerce industry.

This analysis is built from data provided by Michael Kechninov via Kaggle and Google Drive [1, 2]. The total size of the analyzed data is 43.10 GB and contains 310,190,105 unique records over the course of 7 months: October 2019 through April 2020. Additionally, these records are organized by 9 fields. For the purpose of this analysis, the data will be cleaned, and 3 of these unique fields will be used in creating our analysis.

By analyzing the occurrence of these records, a more in depth understanding in ecommerce purchase behavior will be recognized. Compiling these datasets from raw data into insightful meaning, the following will be provided:

1. What are the top 10 most purchased categories?
2. What are the top 10 most purchased items?
3. At what hours are most purchases made?
4. What month has the most purchases?

2. Related Work

According to one particular study, "An Investigation Into Facebook "Liking" Behavior," illustrates a first attempt to explore the different motives behind the use of the "Like" feature on Facebook [3]. Specifically, it takes into account both the gratified usage motives and the underlying motives associated with the "Liking" behavior. For example, when a Facebook user watches a video embedded in a post and finds it entertaining, this person may like the post or marketing material in order to secure the satisfaction obtained.

Another new research "3 Finding that Prove Instagram Drives Shopping Behavior" according to Iris Hearn, "reports that consumer perception of brands on the platform has produced some interesting findings" [4]. She explains that Instagram can drive different levels of authenticity from the initial discovery of the brand to even influencing purchase behavior. Data gathered from this study is based on 21,000 Instagram users between the ages of 13-64, all that have reported engaging with the platform at least once a week, and users were selected from 13 different countries, giving us an idea of how Instagram influences users across different cultures. The study found that just using sharing your brand on Instagram can have a positive impact on how the audience perceives it. They found that consumers view brands on Instagram as 78% more popular, 74% more relevant, 77 % more creative, 76% more entertaining, and 72 % more likely to build a community that those are not on Instagram.

The related work to this project is titled *Behavioral Issues in B2C E-commerce: The-state-of-the-art,* jointly authored by Farid Huseynov and Sevgi Ozkan Yildirim, both from Middle East Technical University. Similar to this project, the related work talks about business to consumer (B2C) e-commerce, online shopping, internet shopping, and consumer behavior. The related work refers to B2C as any type of commercial transaction conducted over computer networks such as the Internet. Business-to-Consumer (B2C) is one of the most common types of e-commerce and it has penetrated businesses in many ways. In a B2C e-commerce model, businesses provide goods and services to individual consumers over the Internet [5].

Business to consumer is one of the most widely recognized types of e-commerce and it has infiltrated businesses in numerous ways, and it has changed the way business is done. Firms make goods and services available to different consumers over the internet in a B2C e-commerce model. Many examples of B2C e-commerce models include online shopping, internet banking, online travels, etc. Furthermore, the behavior of consumers in e-commerce platforms try to establish many crucial factors that influence the behavioral intentions and attitudes of consumers toward online shopping. An example of factors identified are "consumers' Internet usage, previous online shopping experiences, shopping motivation, personal traits, risk and benefit perceptions, trust perception, subjective norms and perceived behavioral control " [5].

Finally, as this project explores the growth of business to consumer e-commerce and the behavior of online consumers, the related paper, *E-commerce and consumer's purchasing behaviour,* authored by Cuneyt Koyuncu And Donald Lien, equally examines the growth of B2C e-commerce. More particularly, this study analyzed the potential determinants of an individual's online shopping behavior in terms of demographic, economic, and the other human characteristics (such as sexual inclinations and essential place of online access). Moreover, studies are the impacts of an individual's assessments of certain critical internet issues (such as taxation of services, privacy, and censorship) on his/her online orders [6].

Two types of online consumer shopping motivation are said to be utilitarian and hedonic. It defines utilitarian shopping motivation as "goal orientated and mission critical" and refers to hedonic shopping motivation as "consumers' shopping behaviors that focus on enjoyment, satisfaction, happiness, and sensuality". These two types of consumer online shopping motivation have been discovered to be particularly connected with consumer satisfaction and buying intention.

Comparing the related studies that reference social media behavior, identification of the attitudes and motivation of consumers, and demographic, our research and analysis focuses on the behavior of online purchases to identify what specific categories of items are being purchased and at what time these purchases are being made. The referenced studies and related works that focus on identifying the demographics of users and what their underlying motives are, rather than what they are buying and when.

3. Workflow and Implementation

The data for this paper were retrieved from a publicly available Kaggle repository and Google Drive. These data were stored in seven files, each at a file size of at least 4GB when uncompressed.

The files were downloaded to a desktop before being migrated to Hadoop Distributed File System (HDFS). The process of migrating the data to HDFS involved three steps for most files. First, the files were compressed and uploaded from a desktop to Oracle Big Data Compute Edition (OBDCE). The files were uploaded compressed to reduce upload speeds. In Oracle, the files were then uncompressed before moving them to HDFS under a predetermined directory. These steps were repeated for each file.

Due to file size limitations of the provisioned Oracle server, some files required additional steps before the mentioned three steps could be followed. These files were split on a local desktop into several smaller files using a program called CSVSplitter. The program allows users to select a CSV file and specify the total records each file should contain. The number of records specified determines the number of output files and file size.

Once the data was in HDFS, it was analyzed using HiveQL. A master table was created to query all data stored in HDFS. Using this table, four additional tables were created to answer the research questions presented in this paper.

Three fields out of the nine provided in the dataset were utilized to produce output for the additional tables: 1) event_time, 2) event_type, 3) and category_code. The event_time field contained the date and time an event took place. The event_type field specified the type of event that took place, i.e. view, cart, and purchased. For this analysis, the purchased event type was utilized to filter the dataset. Lastly, the category_code field contained a string describing the type of item viewed, placed in cart, or purchased. The string contained the category, sub-category, and item type separated by a dot, e.g. "construction.tools.pump". Regular expressions were employed to extract the category, e.g. construction, and item type, e.g. pump. The aformentioned process is graphically displayed on Figure 1.

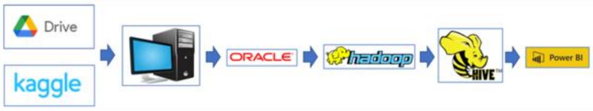

Figure 1. Workflow

4. Hardware Specifications

The composition of HDFS allows for the execution of schemas, tables, and queries. A unique cluster environment hosted by California State University - Los Angeles was used in the creation of this analysis. This cluster was accessed remotely and the entirety of this analysis was performed using HiveQL. The components of this system are as follows:

1. Cluster Version 20.3.2-2
2. Number of Nodes: 3
3. # of CPUs: 12
4. CPU Speed: 2.20 GHZ
5. Memory Size: 180 GB

5. Analysis

Applying HiveQL data warehousing to datasets stored within Apache HDFS, data querying, summarization, and analyses was created to supply a broad understanding of existing purchase records. Data structure was defined, schema was created, tables were created, and queries were run to extract organized data from HDFS database. Following HiveQL data warehousing and exporting this data, Microsoft PowerBI was used in creating visuals analyses of the exported data. The following illustrates ecommerce purchase behavior.

5. 1 Top 10 Most Purchased Categories

Analysing the relationship between category and event types, allow for an understanding of what category resolves in the event type of "purchase." Purchase behavior is important to understand when assessing the likes and dislikes of consumers. The below figure presents the top 10 most purchased categories. It is identified that "construction" is the most purchased category at45.96%. That is followed by "appliances" at 14.49%.

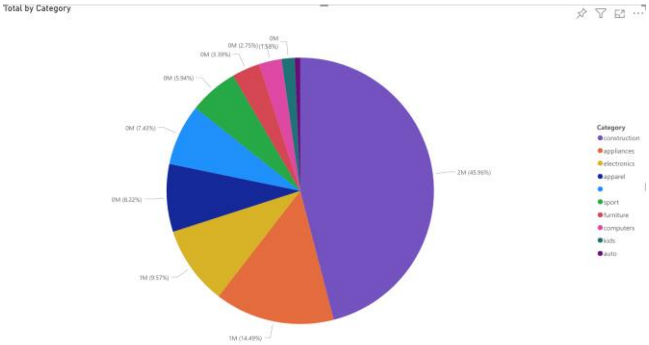

Figure 2. Top 10 Most Purchased Categories October 2019 - April 2020

5. 2 Top 10 Most Purchased Items

The "construction" category has been identified as receiving the most purchases. To create a more in depth analysis in identifying what of construction is most purchases, the following pie chart provides a visual representation of the proportional data. The proportional data is organized by percentages representing each slice. From this visual representation, it is discovered that "lights" are the most purchased item of the "construction" category by 95.3%. Second, is "faucets" by 1.86%. The distinguishing of what categories and what items are most popularly purchased presents a potential marketing and sales focus within the e-commerce industry.

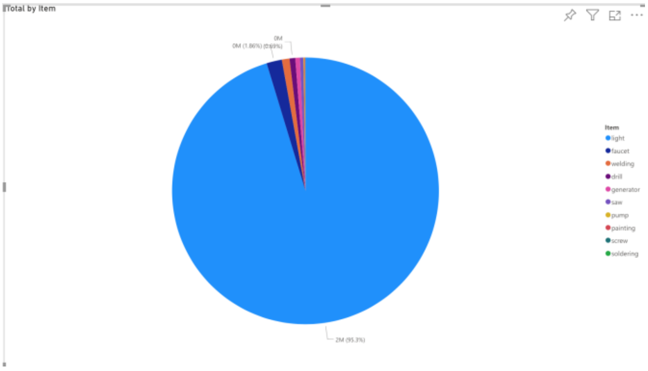

Figure 3. Top 10 MOst Purchased Items October 2019 - April 2020

5. 3 Hours with Most Purchases

Following the distinguishing of most purchased categories and most purchased items of a specific category, below is presented an analysis of what hour of the day these purchases are being made. This valuable insight allows for marketing and sales professionals to understand their customer and further understand when they are likely to make purchases. From this analysis, it is viewed that at 8:00 AM and 9AM most purchases are made online. 8:00 AM received 175,000 purchases and 9:00 AM received 173,000 purchases. Later hours such as 12:00AM and 11:00PM received the least amount of purchases.

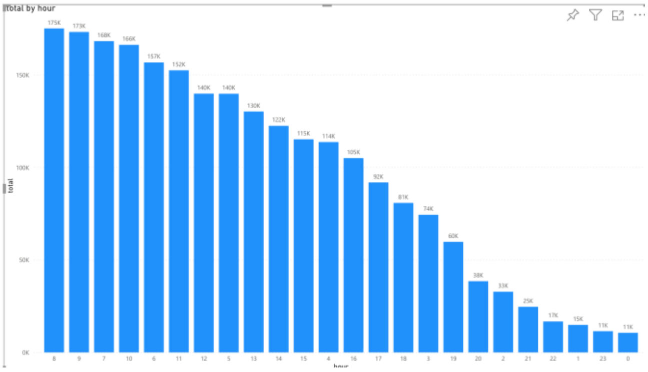

Figure 4. Count of Total Purchases by the Hour (24 Hours) October 2019 - April 2020

5. 4 Top Months of Purchases

Time of ecommerce sales is critical to understanding e-commerce purchase behavior. Following the identification of count of purchases by the hours, the below analysis provides insight to the month that holds the most sales. Identification of seasonality can provide a further understanding of the predictability of consumer traffic. The seasonality is categorized by month. From the analysis, February and December receive the highest number of purchases of the "construction" category. It is recognized that February received .62 million purchases and December at received .52 million purchases.

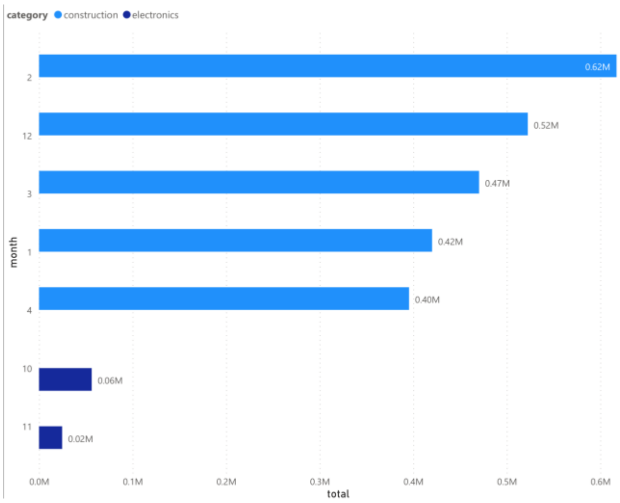

Figure 5. Count of Total Purchases by the Month October 2019 - April 2020

6. Conclusion

Analyzing data such as when purchases were made, if items were purchased, and type of items purchased provides an industry-wide understanding of e-commerce customer behavior. Utilizing extensive expertise in customizing cloud-based data analytics for e-commerce customers, this report provides insight to this behavior.

The role and impact of e-Commerce and customer behavior drive the industry. Behavior analysis is essential in attracting more shoppers and improving "how businesses conduct business." From this report, marketing, advertising, and sales can monitor performance. The provided insight can aid in the identification of what opportunities exist, what forecasting can be prepared, and what activity can lead to greater sales revenue.

References

[1] Kechinov, M., ECommerce Behavior Data from Multi-Category Store. Available: https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Oct.csv [Accessed December 10, 2020].

[2] Kechinov,M., Datasets. Available: https://drive.google.com/drive/folders/1Nan8X33H8xrXS5XhCKZmSpClFTCJsSpE [Accessed December 10, 2020].

[3] Ozannam M., Navas, A. C., Mattila, A. S., Van Hoof, H. B. May 10, 2017. An Investigation into Facebook "Liking" Behavior An Exploratory Study. Available: https://journals.sagepub.com/doi/full/10.1177/2056305117706785 [Accessed December 10, 2020]

[4] Hearn, I. 2019, February 11). 3 Findings That Prove Instagram Drives Shopping Behavior [NEW DATA]. Available: https://www.impactplus.com/blog/instagram-drives-shopping-behavior-new-data [Accessed December 10, 2020]

[5] Huseynov, F., & Yıldırım, S. Ö. 2016. Behavioral Issues in B2C E-commerce. *Information Development, 32*(5), 1343-1358. doi:10.1177/0266666915599586 [Accessed December 10, 2020]

[6] Koyuncu, Cuneyt, and Donald Lien. "E-Commerce and Consumer's Purchasing Behaviour." *Applied Economics*,. Available: web.a.ebscohost.com.mimas.calstatela.edu/ehost/pdfviewer/pdfviewer?vid=1&sid=7588d8f7-39b0-415f-9ee3-ded895977dbd%40sdc-v-sessmgr01.[Accessed December 10, 2020]