eCommerce behaviour data from multi-category store

Lekha Ajit Kumar, Sushmitha Dandu, Dauren Omarov, Navyasree Sriramoju
Department of Information Systems, California State University Los Angeles
Los Angeles, USA
lajitku@calstatela.edu, sdandu3@calstatela.edu, domarov@calstatela.edu, nsriram@calstatela.edu

Abstract This paper explains about customer behavior patterns for an online eCommerce muti-category store. We can interpret data by seeing the customer buying and viewing behavior for a particular brand and category. The paper explains the method and process used for data manipulation and further analysis. The major goal of the project is to give a clear flow of handling big data files and data cleaning processes using Hadoop and Beeline. The data interpretation and analysis of this data was done using Excel, Tableau and Power BI. Charts and visuals like as bar diagrams, timeline and charts on customer behavior patterns for an online eCommerce muti-category store.

I. Introduction

This dataset contains customer behavior data for October and November 2019. The size of the dataset is 5GB. There are 10 columns in this set, which are, as follows: time, type, product id, category id, category code, brand, price, user id, and user session. In this data set our goal is to analyze customer buying behavior. We have Data from 2 different months, hence we also compare the buying and viewing patterns of customers. We have chosen this dataset because it will help companies understand requirements of the users and help maximize their sales and production of demanding goods.

II. Related Work

There are plenty of studies regarding e-commerce shops behavior. One study is based on the data set from Kaggle.com[1]. In this research, the authors study the behavior of an online e-commerce store and its habits. In their methodology, the authors removed the irrelevant features of the study. The authors used bar charts and pie charts. For example, the authors disregarded null values, and used only Event type, product_id, category_id, brand, price columns from the data set. Also, in the study duplicate entries are ignored. Moreover, one column 'purchased' was added to the dataset to indicate if there was a purchase action. In the study results, more people are viewing the merchandise than buying it. Hence, most customers simply looked at the goods and left it there. The authors concluded that so few people purchased the item because users are knowledgeable enough to recognize its significance in their lives, and as a result, they must have looked it up elsewhere before adding it to their basket. In our study we are using dataset from the same source but implementing different approach as we are using Big Data analysis using Hive. Another study made in 2020 focuses on eCommerce behavior data from multi category store [2]. The dataset was derived from [4][5][6]. Obviously, traffic department of Louisville currently have a sustained flow of data and use it on a daily basis, whereas in our work, even if the dataset has the same structure, we were limited to a small portion of data. However our work, apart from analysis of traffics, also Kaggle.com. The author of the study established several primary goals. One of them is to calculate the daily traffic in November. The second is to find out which product categories and brands are most popular. The third is to find out whether the buyer will buy goods by adding them to the cart. In the research the author is using Python 3 with squarify and xgboost libraries, and basic prediction model through XGBoost. The author calculated that in November there were 3,696,117 website visitors. Electronics.smartphone, Electronics.video.tv, Computers.Notebook, Electronics.Clocks, and Apparel.Shoes are the leading 5 categories that people view, as the author says. The most popular brands include Samsung, Apple, Xiaomi, and other brands. The author comes to the conclusion that only a tiny percentage of users actually add the item to the basket, which results in a purchase conversion of about 30% (1.36% divided by 4.49%) once the item was added to the cart by buyers. In contrast to our study, which uses a Big Data model using Hive, the author of this study used Python 3 libraries and a simple prediction model through XGBoost.

III. Specifications

The dataset consists of behavior data for 2 months (from October and November 2019) from a large multi-category online store. This dataset is publicly open and stored in a CSV file format. The dataset is of the size 14.68 GB and covers several days in October and November, the file consists of irregularities majority in the month of November. Although the data is of a big size, we assume that the same data processes can be applied to a similar dataset (as large as 100 GB+ or as small as 25 MB). Table 1 shows the files and size of the files from the dataset.

IV. Implementation Flowchart

Initially, the first version of the dataset known as raw dataset, which comprises all detail customer behavior patterns for an online eCommerce muti-category store, was downloaded from Kaggle. The whole process of data manipulation

Data Set	Size (Total 14.68BG)
2019-Nov	9.01 GB
2019-Oct	5.67 GB

TABLE I
DATA SPECIFICATION

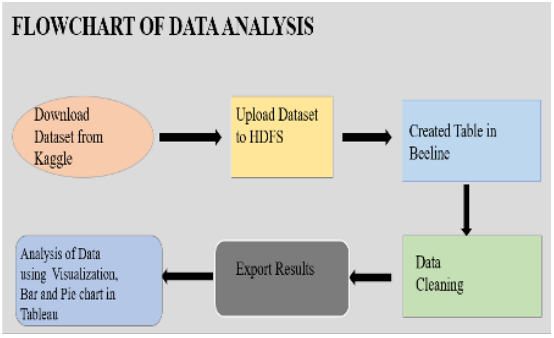

Fig. 1. Figure 1 - Implementation Flowchart

and analysis is shown in the below flowchart (Figure 1).There are two data logs in csv format were uploaded to the Hadoop File System. After that, we created the tables' schema, cleaned the data, create a cleaned data table and summary table, then exported the results. Once the output file has been downloaded we opened in Excel format, we use Excel's in built charts, Power BI and tableau to obtain the visualizations.

V. DATA CLEANING

Firstly the uncleaned or raw files were uploaded and stored in HDFS in the tmp directory. Then we shifted it to our directores ecommerce1 and ecommerce2 which we created in our login. Then created tables and loaded the raw data into tables using Beeline. Since we had two files, each was separately cleaned and then exported for further analysis. Data cleaning was conducted using different techniques such as regular expressions, select statements, like "NULL" and exporting tables. Below is one peace of the code used for cleaning data. SELECT * from octuncleaned where category_code not like "NULL" AND brand not like "NULL" AND user_session not like "NULL";

In addition to that we created a cleaned table to portray all cleaned data in a seprate table so we can permorm qeuire on it freely without always not removing null values.

VI. ANALYSIS AND VISUALIZATION

After data cleaning and preparation for further analysis, files were extracted into Excel and Power BI. We used different interactive maps in order to show jams and alerts clearly on the map as well as time line and different charts.

A. 3D Map in Excel

The first visualization (Figure 2), a 3D map, was made in Excel and it is an animated map with a timeline for one day, since the dataset had full cover of one day in January 2018. We used the heat map to show the count of jams, and clustered columns to show the count of accidents (green bar), road closure (blue bar) and weather hazard (yellow cones). By using the time column, we were able to come up with a map with a time flow. Initially, this visualization was in a video format, by playing the video, it is clear to see that the bars grow faster after 5:00 pm of the day, which means that mostly traffic reports are made during that time. Note that weather hazards alerts are almost not seen, since in LA weather is mostly calm.

B. Power BI

The bubble map from Power BI has been used in the next map (Figure 3), and the location field "City" is being used for geo-coding. The map clearly states that the city of Los Angeles2 is a top city by reported traffic problems. From the time-bar chart shown in Figure 4, we see that the peak-hour of alerts is from 6:00 pm to 7:00 pm and generally, rush hour starts from 4:00 pm and getting better after 8:00 pm.

VII. CONCLUSION