Analysis of Consumer Purchase Patterns in a Large-Scale Multi-Category E-Commerce Platform

ABSTRACT

Analyzing Consumer Shopping Behavior from a large multi-category online store

The advent of technology and the internet has led to an increasing demand for Ecommerce industries. People of all ages prefer online shopping. This increasing trend for online purchasing was highly noticeable during the pandemic. The growth in online purchase has created a need to understand how the customer perceives online shopping and purchases. Large-scale data analytics in e-commerce is still in its nascent stage and there is plenty to learn about it in all facets**.** Analyzing such data is known as shopping behavior analysis which is predominantly used by many online retailers for identifying consumer's buying patterns and preferences. This information can be further used for making predictions about the market trend, setting up promotions, improving the overall user experience, recommending products which are likely to be bought based on their purchase history and in-turn also benefit the retailers by increasing their sales. This can be done by collecting the consumer's data such as their emotional experiences while buying a product, cost the consumer is willing to pay, the times at which most maximum shopping is observed, the product brand most preferred by consumer, etc. The purpose of this analysis is to understand what factors influence most in terms of consumer shopping experience and purchases. Ecommerce data being huge, scalability is an issue while analyzing the data. Hence the aim of this project is to analyze consumer behavior patterns using AWS cloud computing services (such as S3, AWS Glue, Athena) and visualize (using Tableau) the major factors affecting sales.

Table of Contents

Table of Figures

Table of Tables

Chapter 1. Introduction

1.1 Project goals and objectives

The main goal of our project research is to do an in-depth analysis of the dataset to gain insights which can help make strategic business decisions. Today we are in a data driven world whereby following a data driven approach, we can benefit from data, by analyzing it and drawing meaningful insights. The insights from data analysis can help us better understand our customers, target right audience for the relevant market, personalize customer experience, make better product recommendations, improve customer retention, help predict customer value, increase in the conversion rate, and have appropriate merchandise in stock. The insights from the analysis on the data available will help the business understand about its business transactions, consumer behavior and movement over the website, to further take appropriate decisions in improving the overall customer experience, sales, and revenue.

1.2 Problem and motivation

The advance in technology and the internet has led to an increasing demand for E- commerce industries which plays a vital role in retail. People of all ages prefer online shopping. This increasing trend for online purchasing was highly noticeable during the pandemic. In any business, understanding their customers behavior and preferences plays a major role in business improvement. For this purpose, they leverage the huge data available by customer activity online to derive insights about them and use it in making strategic business decisions. This led us to select the subject of analyzing consumer online shopping behavior for an ecommerce website and come with the insights which can be derived from big data available to help business understand the key aspects and make business decisions accordingly to improve the overall performance in terms of consumer experience, sales, and revenue.

1.3 Project application and impact

The consumer shopping behavior analysis is the process of discovering, interpreting, and communicating data patterns and insights related to E-commerce online business. It helps in measuring the user behavior, market and performance trends, and ROI. This analysis can be used by the business to show ROI for the campaigns and make better decisions to reduce costs, increase sales, and make business improvements accordingly.

1.4 Data Description

The dataset for this project of consumer shopping behavior analysis was obtained from a multi-category e-commerce platform available on Kaggle (Data Source). The dataset consisted of the shopping activity of around 285 million users from an e-commerce website for over a period of 7 months i.e., from October 2019 to April 2020. The total size of data available is 30 GB, but we have used October and November data (15 GB) for our analysis. The various attributes like product ID, product category, product subcategory, price, user session ID, brand, event_type, etc. were included in the dataset describing the user shopping activity.

Dataset Link:

https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

Chapter 2. Literature Survey

The research paper titled ‘The Impact of E-Commerce on Customer’s Purchasing Patterns in the Era of Big Data’ studies how the e-commerce can impact the purchasing habits of the customers. They analyzed that the human behavior can be predictable. They studied how cloud computing can be used to analyze the consumer’s behavior which can be influenced by various factors like the location of the seller, the price of the product, the aesthetics of the product, similar products bought along with the required product, etc. They have analyzed customer’s purchasing patterns with the help of a case study that was created by collecting information about the present state of the e-commerce industry, big data analytics and the latest advancements in technologies for studying this. They observed that the users tend to buy those products which the software will recommend them to buy. Also, on e-commerce websites, the quality of products cannot be determined. Hence, users tend to buy product based upon the comments of the users who have bought those products earlier and through the recommendation systems built using big data technologies. Such technological advancements can significantly improve the e-commerce market. They also analyzed that for the customer to keep using their e-commerce platform, the data privacy and security measures should be followed so that the customer’s data is not misused. Implementing these issues will optimize and innovate the existing e-commerce platform.

Another paper titled ‘Big data analytics: Impacting business in big way’ highlights the fact that the internet powered by Big Data is the third revolution that the world has seen after industrial and internet revolution. In today’s scenario, big data is mainly used to analyze customer’s buying patterns and behaviors, their likes, and dislikes on authenticated websites and social media platform. In this paper, authors discuss three Big Data related case studies – cell phone industry, e-commerce, and on-line insurance selling. For the success of any business, authors also propose Big Data analyzing engines to identify, collect, store, and analyze Big Data. Taking one of the case studies, E-commerce is simply the transaction of goods, services, funds or data between customer and business over an electronic network. According to them, the requirement of any major business to customer (B2C) ecommerce site is determine the promotion strategy, potential customers, and their purchase wish-list and to send promotional pricing to the targeted customers when available. To fulfill these requirements, authors have designed the E-commerce Monitoring System (EMS). First the customer data such as personal details, their likings and purchase patterns are recorded by tracking their web visits and preferences. Then the gathered data is processed and analyzed to obtain information like product popularity and customer’s product preferences. This information is used to enhance the business by sending product promotion and related product suggestion to target customers. Using this case study, author concludes that implementing Big Data analysis at business centers is utmost important for enhancing the business in every aspect.

The paper titled ‘The Establishment of Data Model about E-commerce’s Behavior Based on Hadoop platform’ states that the parallel computing capability on the Hadoop platform helps us in great ways to analyze and deal with e-commerce data so that the companies would get a clear understanding of the consumer's interests and needs and supply them with the products accordingly. It has become imperative to find a solution for the issue of how to effectively use transaction data to extract important information and serve as a guide for business marketing operations. Thus, we need a distributed high-performance network like Hadoop would help us to analyze a large amount of data. The data would be stored in the Hadoop platform and will undergo certain preprocessing techniques the MapReduce model would then calculate the data and arrange data appropriately. The calculated data would then be imported into the HDFS framework, and the data would be provided to the SQL. With the help of the test results, they observed that the transactions could be queried from different angles according to the period and price which is evident that Hadoop is effective in handling and processing large amounts of e-commerce data successfully without any data loss.

Chapter 3. System Design

3.1 ELT Architecture

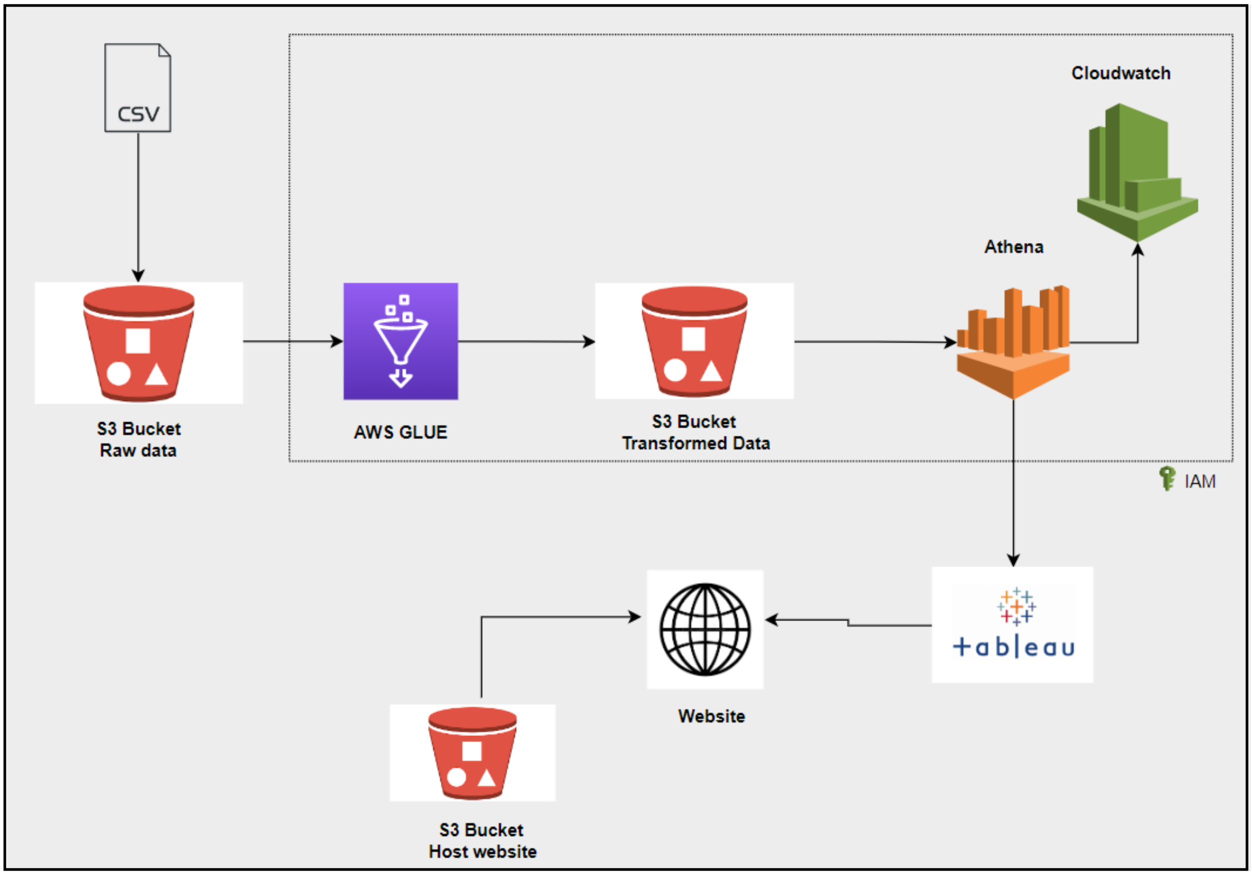

Figure 1. ELT Architecture

3.2 Process outline and description

The raw data which consisted of two files, one with data for the month of October and other for the month of November (in csv format) extracted from Kaggle was first loaded into the S3 bucket. The raw data had to be transformed. We used AWS Glue for ELT. Firstly, a crawler was created which crawls through all our data and creates a schema in the Data catalog. Once the schema is ready, we created a job in AWS Glue to do some transformations like changing the datatype of certain features, splitting the features into different columns, etc. Once the transformations are performed, we again crawl through the transformed data and the transformed data is stored in the S3. We then use Athena to query the data and to look at the schema of our data. The CloudWatch was used for monitoring the query-related metrics. The next step was analyzing the data through visualizations. We used Tableau Desktop for this purpose as it is free and user friendly and creates good interactive dashboards. We connected Athena to Tableau and extracted the data source to create visualizations and interactive dashboards. We then published dashboard to Tableau Public, generated an embed code which was then hosted on S3 bucket. We used IAM policies throughout to securely access to AWS resources used.

3.3 Tools Used

We have used AWS cloud services for most part of our project on consumer shopping behavior analysis. Uploaded csv files into Amazon S3 bucket. Performed ETL jobs in AWS glue for transforming the data uploaded in AmazonS3 bucket. The transformed data was then queried using Amazon Athena which queries data ins S3 using SQL, providing meaningful data extracts from the huge dataset via querying operations.

Finally, Visualizations were generated in Tableau by connecting it to Amazon Athena as a data source. Tableau visualizations are more appealing when compared to other tools and we could connect easily to Amazon Athena. Also had a free access to Tableau desktop and Tableau public for publishing our work for public access.

Amazon S3 – Scalable Storage in the cloud

AWS Glue – Data Catalog (Databases, Tables), Data Integration & ETL (AWS Glue Studio, Jobs)

Athena – Query Data in S3 Using SQL

Tableau – Amazon Athena connected to Tableau for Business Analytics and Visualizations

Tools	Usage/Description
AWS	Cloud Service for Data Storage, ETL, Querying
Amazon S3	Data Loading and Storage
AWS Glue	Data Transformation
Athena	Analytical Queries
Tableau	Data Visualization

Table 1. Tools used

Chapter 4. System Implementation

4.1 Data Loading and Transformation

As we are working on 15GB of data which is quite large we performed multipart upload of data using AWS Command Line Interface (CLI). Initially, AWS CLI was configured from our local machine and a user was created. The credentials were then created for the user using AWS Secret and Secret key and the credentials were downloaded in Comma Separated Values (CSV) format. We then imported the credentials to AWS CLI. To set the profile for the user we created and configured the named profile and set environmental variables for the profile user. The config file was verified if all the configurations were set accordingly. Regarding the multipart upload, AWS cp command along with recursive was used as two files had to be uploaded. We customized certain configurations and enabled amazon S3 transfer accelerations to increase the upload performance.

Exploratory Data Analysis and Extract Load Transform (ELT) was performed on the dataset which included removing duplicate data, handling missing values, changing the datatype to correct format, etc. After a thorough analysis, we performed the following cleaning and transformation on the dataset:

- All the columns were modified with appropriate datatypes.
- The column event_time was originally of the string datatype and consisted of 'UTC' along with the timestamp. Thus, we removed the 'UTC' and changed the event_time to timestamp datatype.
- The column category_code was splitted into three columns as it consisted of the product category, their subcategory and special category in one column itself. We splitted them with columns namely category_1, category_2 and category_3. The delimiter '.' was used to split the category_code column.

The figure 2 shows the script that was generated for transforming the data in Glue.

```python
# Script generated for node Custom Transform
def MyTransform(glueContext, dfc) -> DynamicFrameCollection:
    import pyspark.sql.functions as f

    df = dfc.select(list(dfc.keys())[0]).toDF()

    new_df = (
        df.withColumn("category_1", f.split(df["category_code"], "\.").getItem(0))
        .withColumn("category_2", f.split(df["category_code"], "\.").getItem(1))
        .withColumn(
            "category_3",
            f.when(
                (f.size(f.split(df["category_code"], "\.")) > 2),
                f.split(df["category_code"], "\.").getItem(2),
            ).otherwise(""),
        )
    )

    newcustomerdyc = DynamicFrame.fromDF(new_df, glueContext, "newcustomerdata")

    return DynamicFrameCollection({"CustomTransform0": newcustomerdyc}, glueContext)
```

Figure 2. Script for data transformations

Initially the data from the S3 bucket is crawled by creating a crawler under AWS Glue. After which a schema is generated into the Glue data catalog. Once the table is created, we then use it to perform ETL job. The transformed files are stored in the S3 bucket after which it is again crawled, and a final schema is created in the Glue catalog which will be used for querying via Athena. Figure 3 demonstrates the visual of the Glue Job.

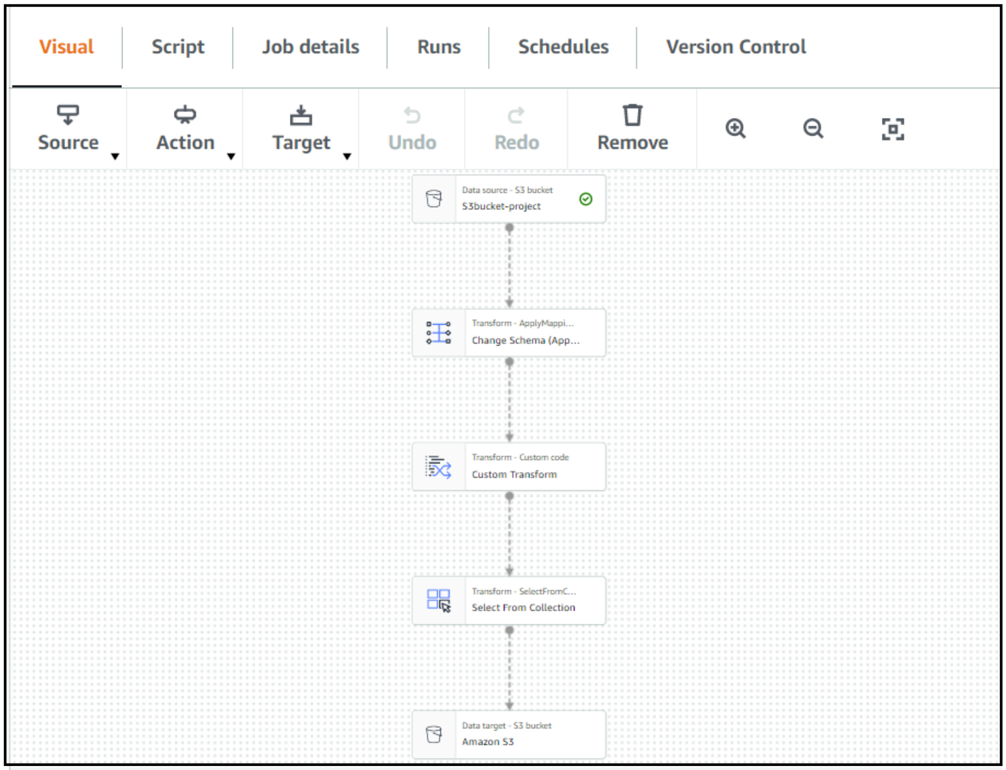

Figure 3. ETL job

Figure 4 illustrates the script that was generated after the datatype of the columns has been transformed.

```
# Script generated for node Change Schema (Apply Mapping)
ChangeSchemaApplyMapping_node1668023128365 = ApplyMapping.apply(
    frame=S3bucketproject_node1668022964314,
    mappings=[
        ("event_time", "string", "event_time", "timestamp"),
        ("event_type", "string", "event_type", "string"),
        ("product_id", "long", "product_id", "long"),
        ("category_id", "long", "category_id", "long"),
        ("category_code", "string", "category_code", "string"),
        ("brand", "string", "brand", "string"),
        ("price", "double", "price", "double"),
        ("user_id", "long", "user_id", "long"),
        ("user_session", "string", "user_session", "string"),
        ("partition_0", "string", "partition_0", "string"),
    ],
    transformation_ctx="ChangeSchemaApplyMapping_node1668023128365",
)
```

Figure 4. Datatype Transformation

Figure 5 shows the S3 buckets where the input and the transformed files are stored.

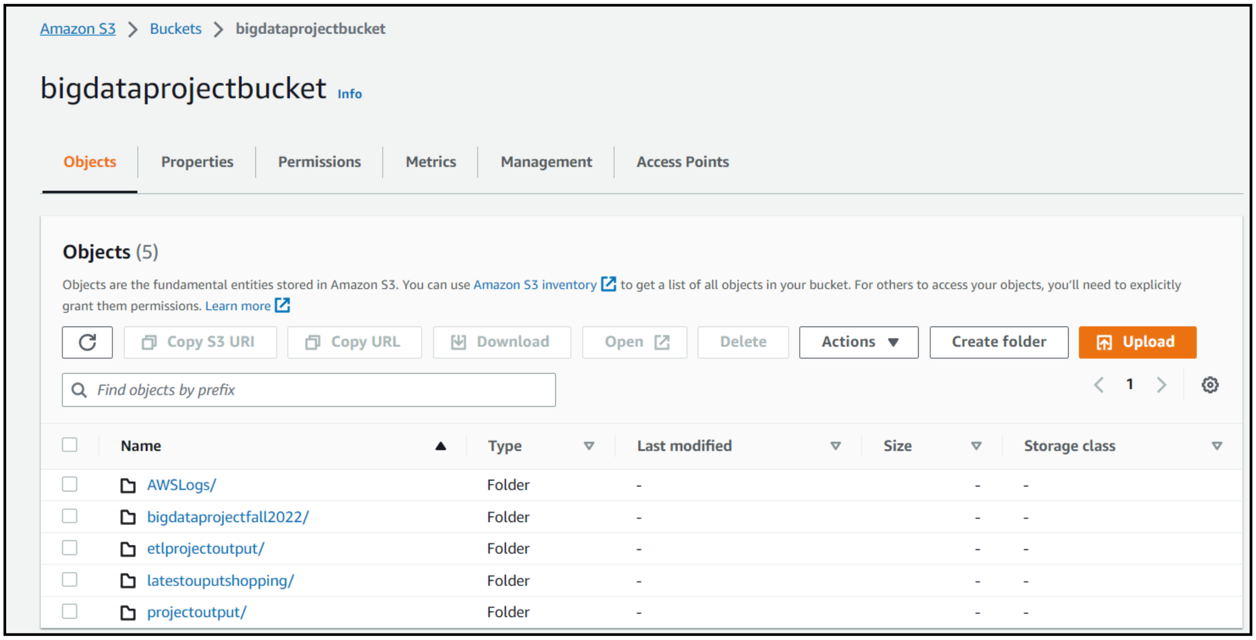

Figure 5. Storage of input and transformed files in s3 bucket

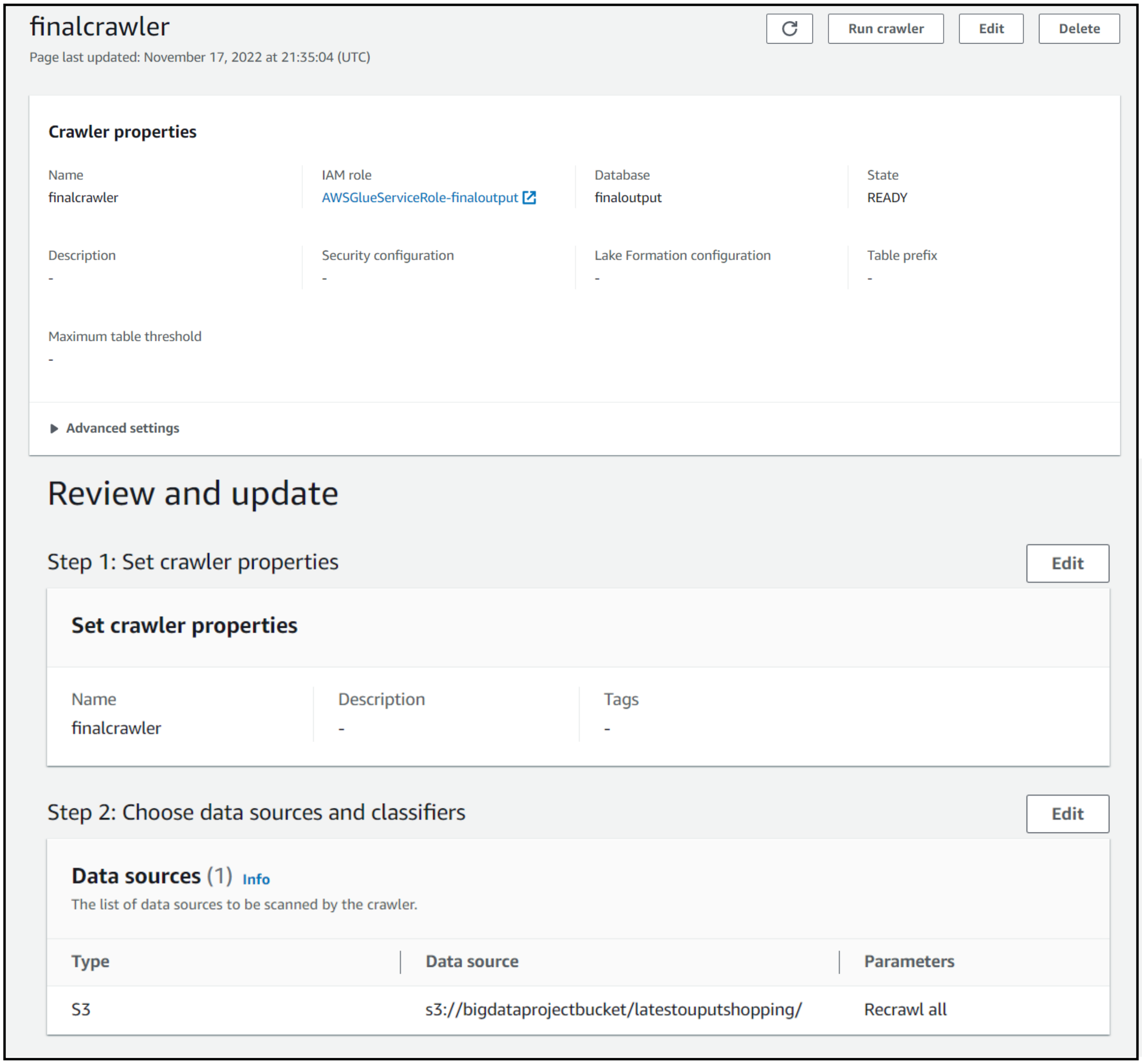

Figure 6. Crawler configurations

4.2 Data Analysis

The data in amazon S3 could be easily analyzed using Amazon Athena. Athena is serverless, data can be queried without setting up data warehouses or managing any servers. In Athena data is sited in amazon S3, definition of schema is done, and querying is achieved with the help of SQL. Although a significant quantity of data up to 15 G was utilized, most of the answers were supplied immediately, and we only pay for the queries that we actually conduct. We performed certain queries using Athena and the results were stored in latestoutputshopping1 table. One such query ran into Athena for querying the data from the table can be seen from figure 7.

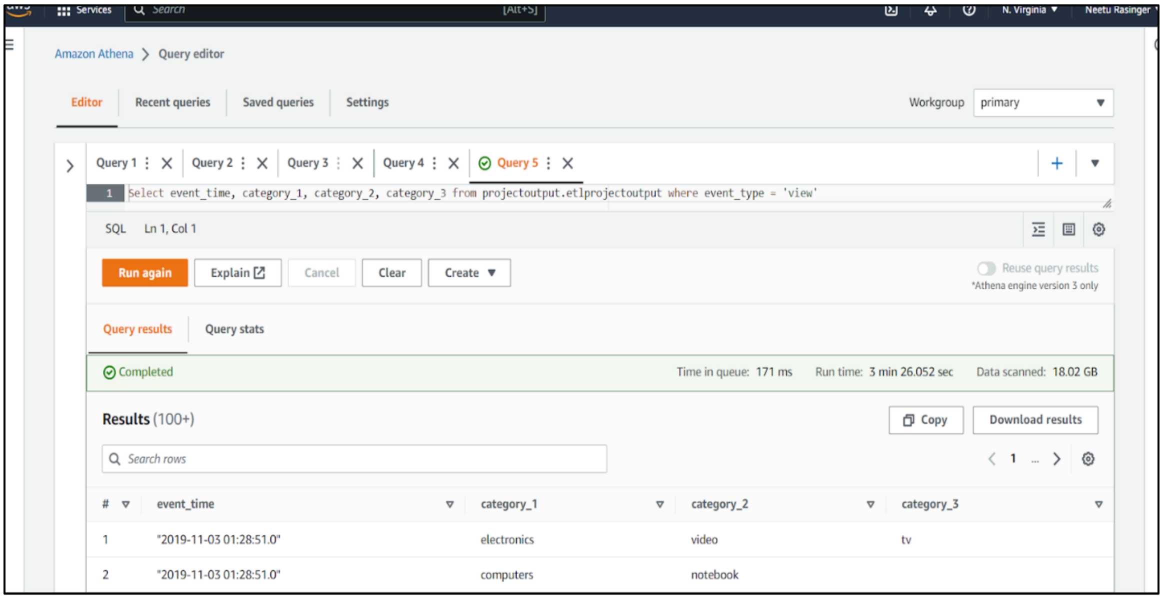

Figure 7. Data Querying in Athena

4.3 System implementation issues and resolutions

Firstly, we downloaded the large files onto our system and then tried uploading them directly into S3 bucket. However, the upload failed due to the network failure. Also, the time taken for the direct upload was larger and hence we resolved this issue by uploading the data to S3 through AWS CLI.

The second problem that we faced was extracting the data from Athena to Tableau using live connection. It was taking long hours to upload the entire data in Tableau. We resolved this issue by using extract connection where a snapshot of the data is created in Tableau and we filtered unnecessary columns and selected a random sample with one million users for our analysis.

4.4 Web Integration

Connection to Tableau

Tableau desktop is used to make connections with Athena for creating dashboards. We connect Tableau Desktop to AWS Athena by inserting the Server, Port, and the S3 Staging Directory (The S3 URL where the queried data is stored) along with the Access Key ID of our user which can be seen from Figure 8.

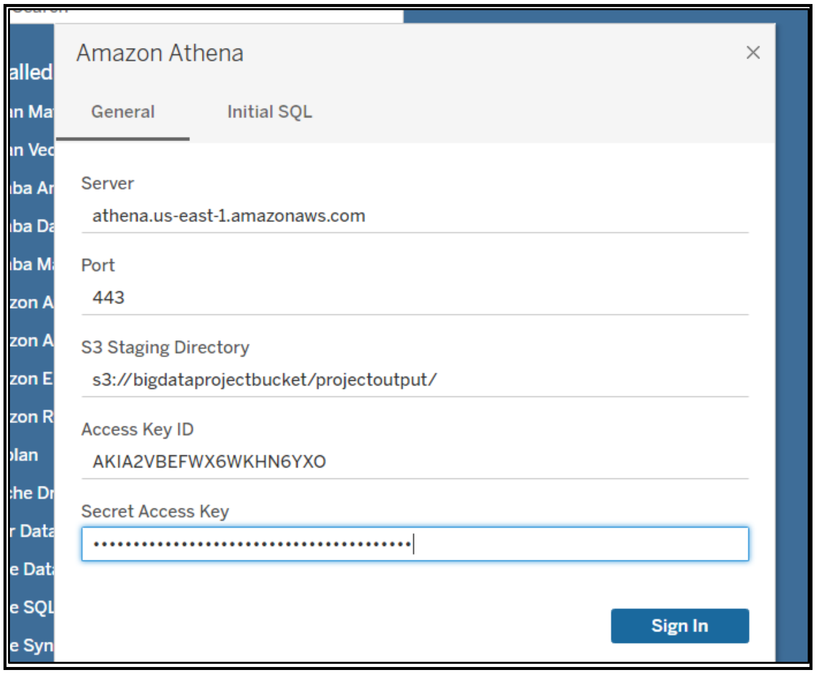

Figure 8 Connecting Tableau to Athena

Once, the connection is successful, we extract a subset of data (around 1 million records) as shown in Figure 9 from Athena to Tableau for our analysis. Once the data has been extracted, we can then start analyzing the data.

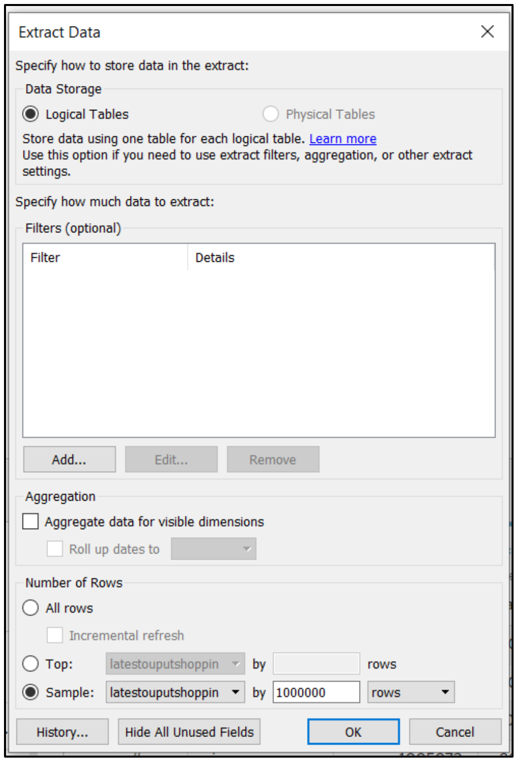

Figure 9. Data Extraction

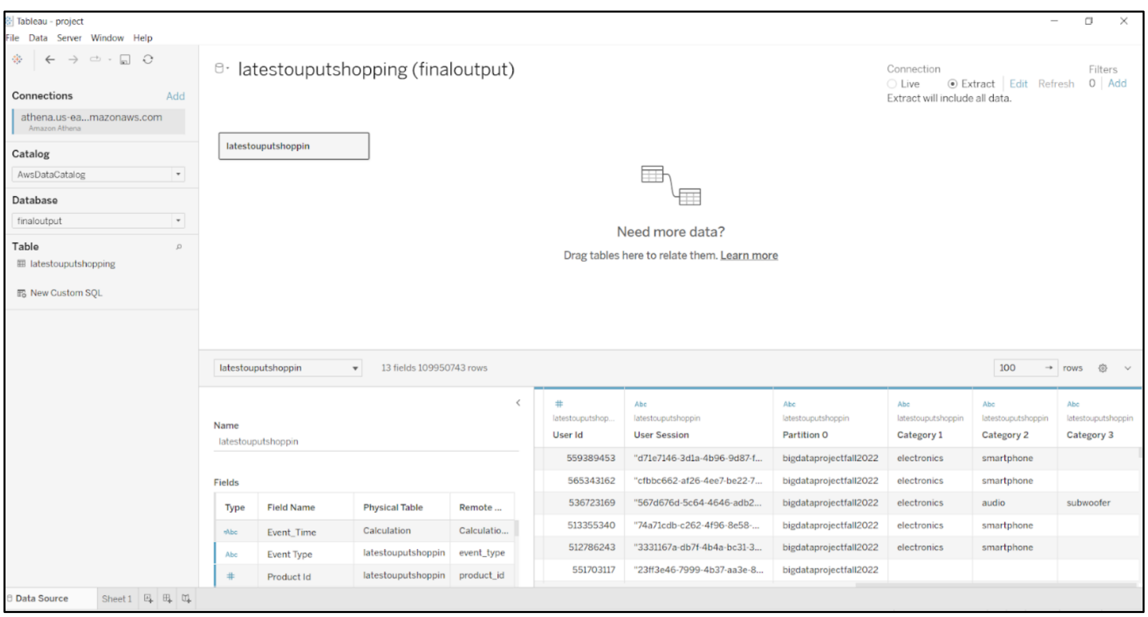

Figure 10. Data Loaded in Tableau Desktop

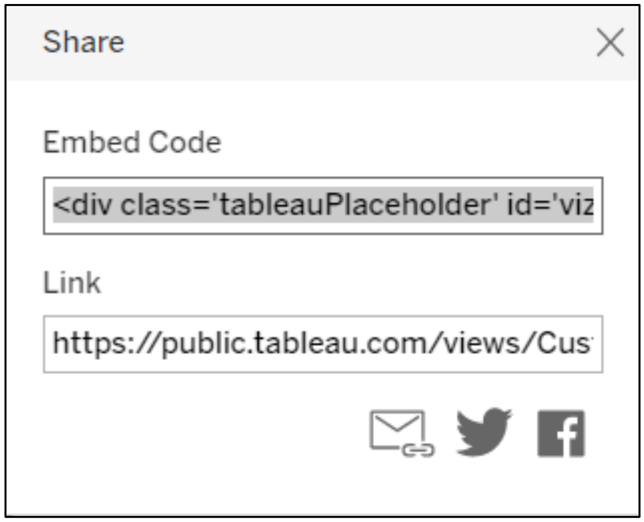

Figure 11. Embed code generated in tableau public

Once the dashboard is ready, we publish the dashboard in Tableau Public and copy the embed code (Figure 11) and upload it in the S3 bucket for hosting it as a static website for public access.

Static website link: https://shoppingdashboard.s3.us-west-1.amazonaws.com/index.html

Chapter 5. Data Visualization and analysis

The next step is to analyze the cleaned and the transformed data through visualizations. Visualizations are created to identify the activity trends of customers throughout the day, the customer's preference of products for online shopping, the percentage of customers viewing the products versus buying the products, etc. We have used Tableau Desktop for our analysis. After the data has been transformed and queried using Athena, we connect Tableau Desktop to AWS Athena by inserting the Server, Port and the S3 Staging Directory (The S3 URL where the queried data is stored) along with the Access Key ID of our user.

5.1 Visualizations

5.1.1 Total Sales Per week

It can be seen from the Figure 12 that the total sales increases after first week of each month. For the month of October, the highest sales were observed on October 13th, 2019, which was Sunday. Similarly, for the month of November, the highest sales recorded (which was the overall highest sales) was on November 17th which was also third Sunday. Overall sales for November were higher. One of the reasons could be the approach of Thanksgiving which falls on the last week of November.

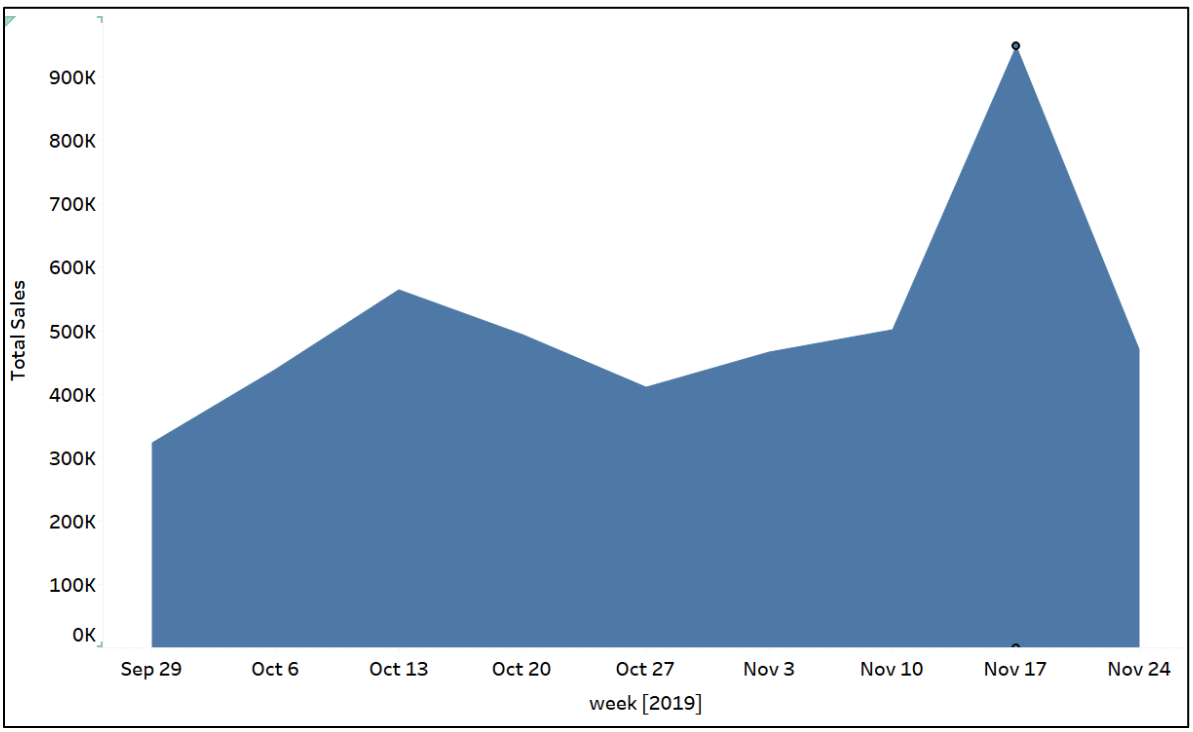

Figure 12. Total Sales per week

5.2.1 Top Sold Category of products

Figure 13 shows a bubble chart showing the most sold products on the e- commerce website. Electronics tops the rank with around 8248 items sold. The second top sold category is appliances (1538 items), computers and so on.

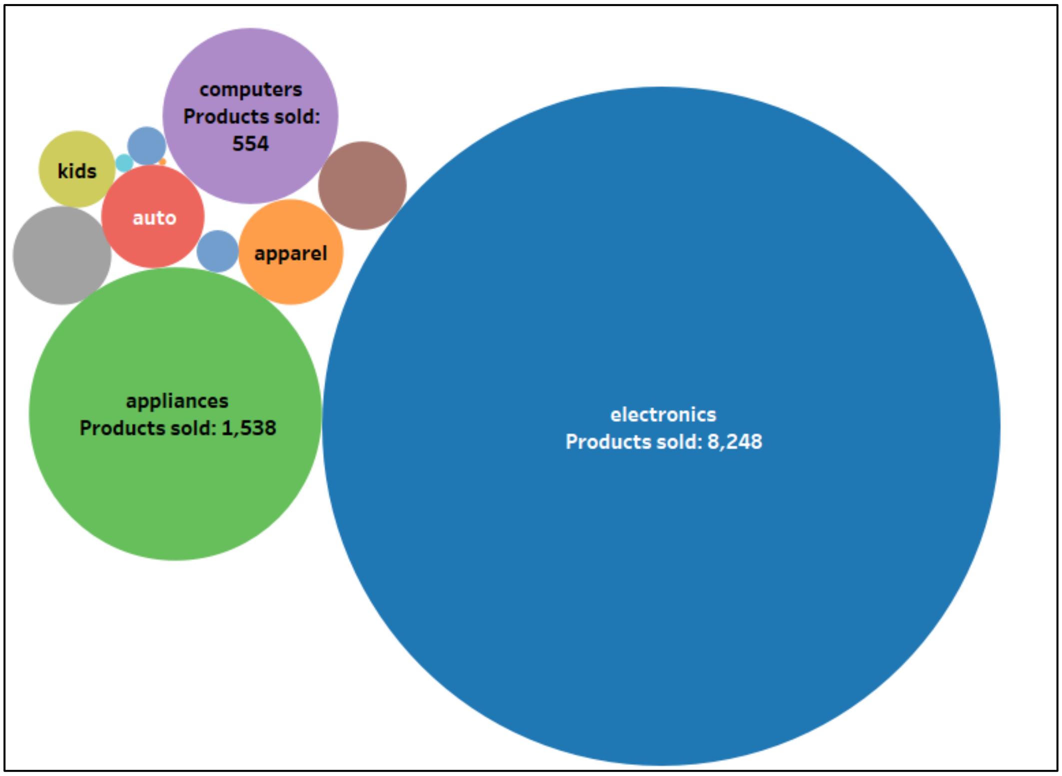

Figure 13. Top Sold Product Category

5.3.1 Top Sold sub-category of products

It can be seen from the figure below that from electronics category, smartphones are the most purchased items by customers which is followed by audio and video related items. It was observed that camera is the least purchased product for this e-commerce website. We can change the category from electronics to another category to see the most sold items sold for that respective category using the drop-down menu.

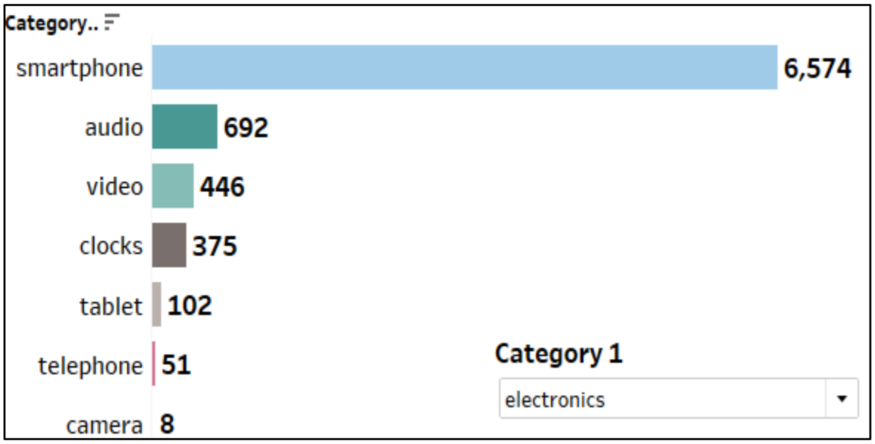

Figure 14.Top Sold Product Sub-category

5.4.1 Popular brands

It can be seen from the bar chart below that in the electronics category, Samsung is the most popular brand which is followed by Apple and Xiaomi. The category can be changed to other category to look for the popular brands for the category using the drop-down list.

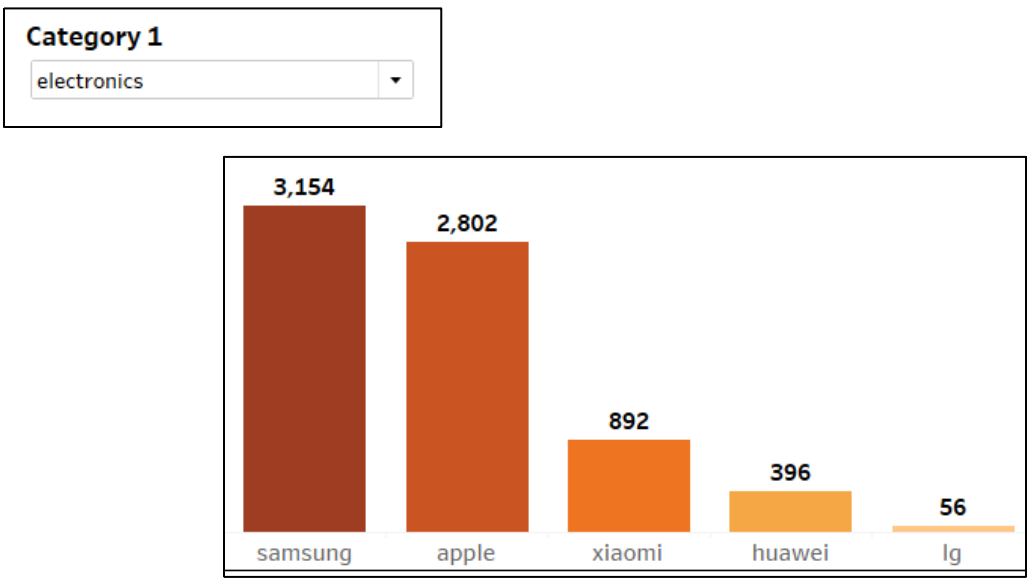

Figure 15. Popular brands

5.5.1 Distribution for each event

The donut chart shown in Figure 16 shows that the percentage of people viewing the products on e-commerce website is the largest (94.94%). The percentage of people adding these items into their cart is around 3.58 % whereas only 1.49% of the total people will purchase the products.

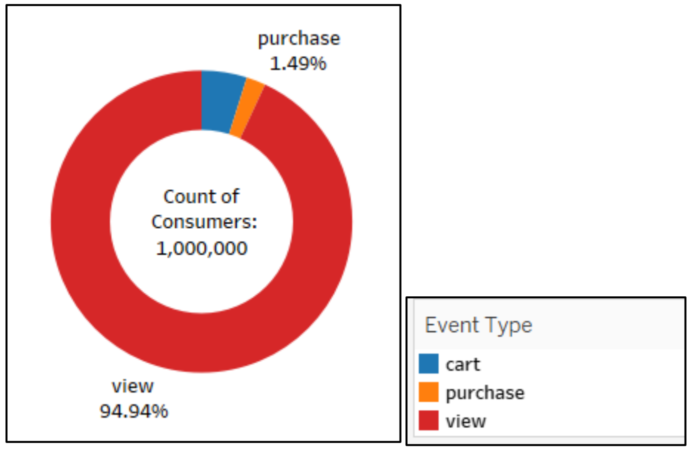

Figure 16. Event distribution for viewing, adding in cart and purchasing.

5.6.1 Activity trend throughout the day

The line chart below shows the activity trend of online customers on e-commerce platform for each hour of the day. The trend shows an increasing trend till 8 am and then a slight dip is observed at around 12 pm. Most of the people are active around 4 pm.

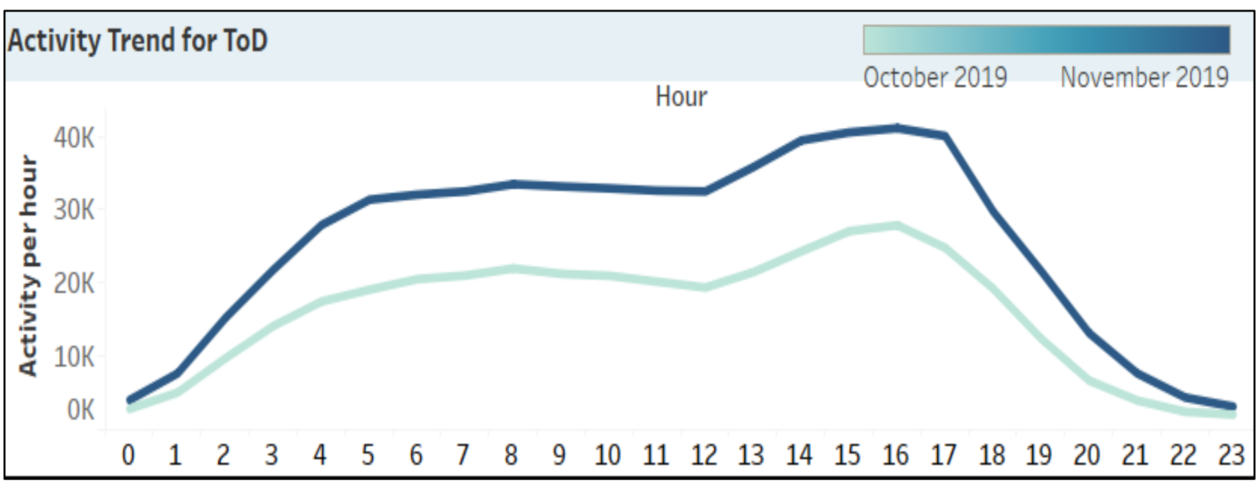

Figure 17. Activity Trend throughout the day

5.2.6 Month-Over-Month Growth Percentage

The table below shows the count of products bought by each consumer, the total earnings by the e-commerce website for the month of October and November respectively. Apparel lead to the highest increase in sales for November (around 72% increase).

Percentage change in sales

	Event_Time					
	Count of Products		Earnings		Percentage Change	
Category 1	October	November	October	November	October	November
apparel	71	128	$6K	$10K	0.00%	72.82%
appliances	678	860	$119K	$167K	0.00%	40.60%
auto	95	96	$13K	$13K	0.00%	-3.25%
computers	243	311	$97K	$134K	0.00%	38.39%
construction	72	68	$10K	$7K	0.00%	-32.33%
electronics	3,839	4,409	$1,643K	$1,856K	0.00%	12.92%
furniture	74	100	$16K	$20K	0.00%	25.26%
kids	53	53	$9K	$8K	0.00%	-7.67%

Figure 18. Percentage change in Sales

Once the visualizations are created, we created an interactive dashboard applying month and category as a filter as shown in figure 19.

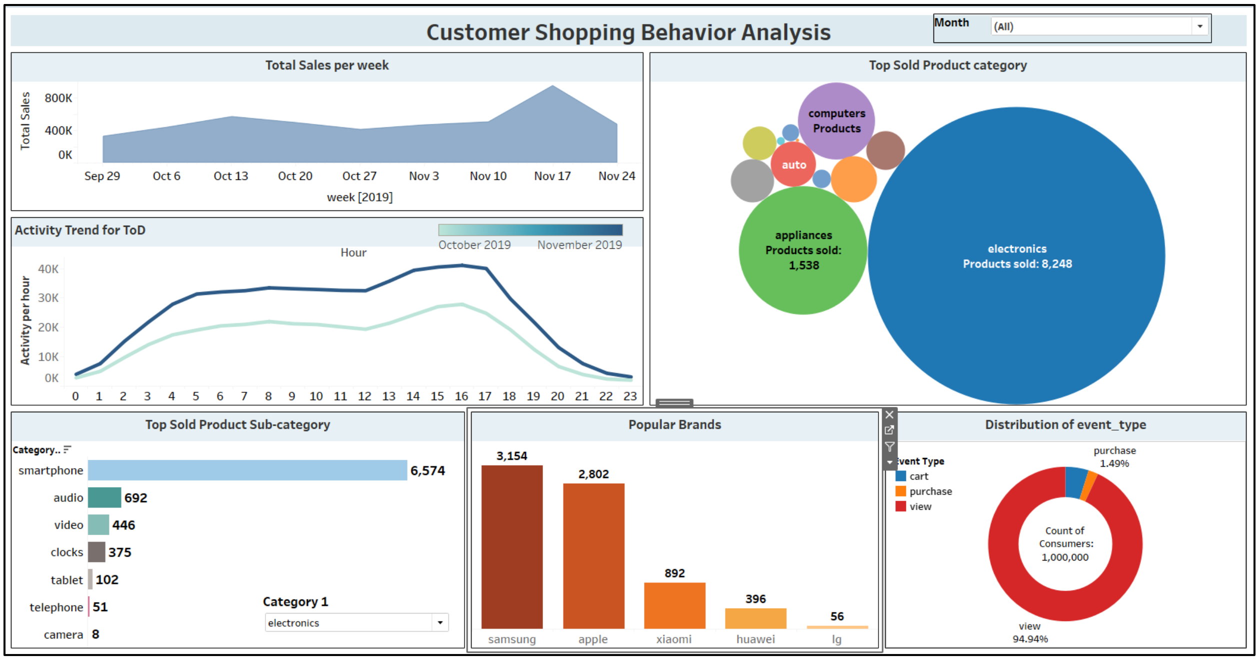

Figure 19. Customer Shopping Behavior Analysis dashboard

Chapter 6. Conclusion and Future Work

6.1 Project summary

We analyzed the consumer shopping behavior on e-commerce platform using AWS services for our project. We observed the user activity for a period of two months. The analysis helped us understand the time for which the consumer is most active online, the products that most users are interested in shopping online, the preferred brands for the product categories. This analysis can help business to make sure that there is no server crash during the most active period thereby improving the consumer experience. Also, offering discounts on products during weekdays can attract the users for shopping through the week further contributing to increase in sales.

6.2 Future work

Our analysis was focused for a period of two months. We can increase the data and improve our analysis by including the data for long period of time. We can also use machine learning algorithms for predicting the demands for all products based on the consumer shopping behavior so that the inventory can be maintained beforehand by the e-commerce platform. We can build recommendation systems which will help recommend the products to the users based on their past purchases and predicting their future purchases.

References

Wu, G., Cheng, L., & Dong, L. (2019). The Impact of E-Commerce on Customers' Purchasing Patterns in the Era of Big Data. *Journal of Advances in Information Technology*, *10*(3), 109–113. https://doi.org/10.12720/jait.10.3.109-113

Sharma, N., Sawai, D., & Surve, G. (2017). *Big data analytics: Impacting business in big way. 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI).* doi:10.1109/icdmai.2017.8073494

Shao, F., & Yao, J. (2018). *The Establishment of Data Analysis Model about E-Commerce's Behavior Based on Hadoop Platform. 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS).* doi:10.1109/icitbs.2018.00116

https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

https://www.shopify.com/blog/marketing-analytics

https://www.sciencedirect.com/science/article/pii/S2666518221000383

https://www.researchgate.net/publication/298739144_Big_data_analytics_in_E-commerce_a_systematic_review_and_agenda_for_future_research

https://d1wqtxts1xzle7.cloudfront.net/50409172/10.1.1.655.7217-with-cover-page-v2.pdf?Expires=1663392470&Signature=IZRvJYmHZ-hYlnQNB7L9s1wsj~P~5bZ3VnPa2Ow17EVdnnygeA218SM6DRsSa31jxPlNWvU5gFWQ7RankUa31-pR0O5ZCFAKaNjW5GpJmtFOpTIv25cn5xdrBfrs~8dsA0O6nLY-

https://help.tableau.com/current/pro/desktop/en-us/examples_amazonathena.htm

https://docs.aws.amazon.com/cli/latest/topic/s3-config.html

https://docs.aws.amazon.com/cli/latest/userguide/cli-configure-profiles.html

https://www.google.com/search?q=website+icon+images&tbm=isch&ved=2ahUKEwij5divzrb7AhXtmo4IHZIEDe0Q2-cCegQIABAA&oq=website+icon+images&gs_lcp=CgNpbWcQAzIECCMQJzIFCAAQgAQyBggAEAcQHjoECAAQQzoICAAQCBAHEB5QwwdYng1goBJoAHAAeACAAV2IAcwDkgEBNpgBAKABAaoBC2d3cy13aXotaW1nwAEB&sclient=img&ei=W-R2Y-PLIu21uvQPkom06A4&bih=714&biw=1536&rlz=1C1CHBF_enUS912US912#imgrc=27rmHmjzHIpI3M

https://www.google.com/search?q=s3+bucket+icon&rlz=1C1CHBF_enUS912US912&sxsrf=ALiCzsYRsacjmLplqXwBYMKc_bG0q9gjIw:1668736032502&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiGr

MOTzrb7AhXUI30KHV_uAzQQ_AUoAXoECAMQAw&biw=1536&bih=714&dpr=1.25#imgrc=HTL--CMhT7OOWM

https://www.google.com/search?q=tableau+image&rlz=1C1CHBF_enUS912US912&sxsrf=ALiCzsaTsl5T47u_3kEAhagdxKpFR2M5vg:1668736170109&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiznJLVzrb7AhVzPH0KHc27A0oQ_AUoAXoECAMQAw&biw=1536&bih=714&dpr=1.25#imgrc=IC2ieWfwkth2fM