Optimizing Customer Journey through Advanced Analytics Techniques over Google Analytics 4 Data in Google BigQuery

BISERA CHAULEVA[1], DIJANA CAPESKA BOGATINOSKA[2], ALEKSANDAR KARADIMCE[1]
[1]Faculty of Information and Communication Sciences,
University of Information Science and Technology "St. Paul the Apostle",
Partizanska bb, 6000 Ohrid,
REPUBLIC OF NORTH MACEDONIA

[2]Faculty of Applied IT, Machine Intelligence and Robotics,
University of Information Science and Technology "St. Paul the Apostle",
Partizanska bb, 6000 Ohrid,
REPUBLIC OF NORTH MACEDONIA

Abstract: - In a highly competitive, data-driven marketplace, optimizing customer journeys is essential for businesses. This paper examines the combination of advanced analytics techniques with Google BigQuery's data warehousing capabilities, utilizing data from Google Analytics 4 (GA4). GA4 provides a comprehensive view of user interactions across platforms, but extracting actionable insights requires a robust data infrastructure. Google BigQuery's scalable architecture supports real-time analysis of massive datasets, offering valuable insights into customer behavior. This research explores methodologies such as sequence analysis, network analysis, and clustering to analyze customer journeys and enhance marketing strategies. Our technical contributions include the development of a scalable ELT pipeline using Dataform for processing GA4 data, the implementation of optimized star schema design for enhanced query performance in BigQuery, and the integration of advanced analytics techniques, such as sequence, cluster, and network analysis, to drive actionable insights and improve decision-making accuracy. Through practical implementations and real-world examples, the study demonstrates the effectiveness of this integration. Key findings show sequence analysis improves purchase flow, network analysis identifies product relationships, and clustering analysis enables customer segmentation for targeted marketing. The paper concludes with recommendations for businesses to fully leverage GA4 data, improving user experiences and fostering sustainable growth.

Key-Words: - Customer Journey Analysis, Advanced Analytics Techniques, Google Analytics 4, Google BigQuery, Sequence Analysis, Cluster Analysis, Network Analysis, Digital Analytics, Data Warehousing.

Received: April 14, 2024. Revised: November 9, 2024. Accepted: December 6, 2024. Published: December 31, 2024.

1 Introduction

Understanding and optimizing customer journeys has become essential for businesses that want to maintain a competitive edge in a data-driven environment. Advanced analytics and big data technologies provide businesses with new opportunities to gain deeper insights into customer behaviors and preferences. Specific tools need to be considered to collect and bring the user journey data to the cloud platform. In this research, we propose the use of the services Google Tag Manager (GTM) and Google Analytics 4 (GA4), which have the capability to collect all user interactions and save them as events such as page views, page scrolls, clicked CTA buttons, etc. The services are not limited only to web interaction but can also include interactions registered across mobile devices, tablets, SPAs, etc.

Google BigQuery (BQ) has been considered to preprocess, build the data model, and prepare data for analysis. BQ has been described as a fully managed serverless data warehouse mainly adapted to preprocess and analyze high volumes of data.

The study will go deeper and evaluate how processed and prepared data can be used to retrieve meaningful insights based on different techniques such as network analysis, sequence analysis, and clustering analysis, all done within the BQ environment. In the end, the results of all these techniques will be discussed, proving the capabilities of the platform and proposing end-to-end solutions for the most important issues for businesses. With the insights derived from analysis, the direction of investing in marketing strategies as well as updating content to retain the users and increase return on investment (ROI) can be determined.

The paper is organized into multiple sections: Section 2 elaborates on the related work, while Section 3 discusses the objectives and limitations. Section 4 introduces the research methodology, including data collection, data schema, data export, data preprocessing, and data analysis using advanced analytics techniques. Sections 5 and 6 present the results and findings based on insights derived from advanced analytics. Finally, Sections 7 and 8 conclude the paper and propose directions for future work.

2 Related Work

Customers today engage with companies through a variety of touchpoints across different channels and media, with their experiences becoming increasingly social. As a result, businesses must coordinate multiple internal functions and collaborate with external partners to effectively create and deliver positive customer experiences, [1].

Retail has transformed significantly over the past decade with the rise of online, mobile, and social media channels, altering business models, retail strategies, and shopper behavior. While multi-channel retailing was popular, there’s now a shift to omnichannel retailing, which takes a holistic view of how customers engage across different channels during their search and purchase journey, [2].

[3], presented the capability to infer consumers' psychological traits and states from their digital footprints, presenting exciting new possibilities for digital marketing.

Based on [4], the landscape of data-driven decision-making has shifted significantly, with top-performing companies now centering their competitive strategies on insights derived from data, leading to remarkable business outcomes. Their key advantage is analytics—advanced quantitative analysis, statistical techniques, and predictive modeling. Leading organizations use these tools to pinpoint their most valuable customers and offer tailored pricing, speed up product innovation, optimize supply chains, and uncover the real drivers behind financial success.

The marketing analytics field is undergoing significant transformation, especially in the advanced techniques and software tools used by professionals. Expertise in web analytics, particularly with Google Analytics, has become a crucial skill for marketing analysts worldwide, [5].

Data warehousing solutions play a crucial role in enabling organizations to store, manage, and analyze large volumes of data for business intelligence and analytics purposes. The emergence of cloud-based data warehousing platforms has revolutionized the way companies approach data management and analytics, [6].

According to [7], analytic warehouses are powered by precise, integrated, and accessible customer data. It allows businesses to evaluate past decisions to enhance future interactions. Customer data is used to identify up-sell and cross-sell opportunities, address inefficiencies, stimulate demand, and improve customer retention. Historical data can also be utilized to develop models or scores that feedback into operational processes.

Considering analytics techniques for customer journey analysis, [8], pointed out that Graphs, or networks, present an innovative method for modeling the customer journey and converting this information into valuable insights. Additionally, they possess the adaptability and capability to facilitate this process across various segments and levels of detail.

A study [9], shows that a sequence analysis can be conducted to pinpoint touchpoints and channel interactions, followed by a sequential pattern analysis. The findings can reveal an increasing number of variable touchpoints and activities associated with high-involvement products.

Finally, study [10], shows that Customer journey clustering as a method can employ algorithms to analyze journeys and identify groups of consumers with similar purchasing behaviors. These clusters will allow for enhancements and personalization of the customer experience.

Building upon this body of research, our work integrates and extends previous findings in several key ways. While, [1], established the importance of coordinating multiple touchpoints, and [2], highlighted the shift towards omni-channel retailing, our research leverages these insights within the specific context of GA4's enhanced tracking capabilities. We extend [3], work on consumer psychology by applying it to real-time digital footprints through BQ's processing capabilities. Our approach combines [4], analytics-driven competitive strategy with [5], modern web analytics framework, specifically for GA4. Furthermore [6], provided the foundational data warehousing concepts, our implementation adapts these principles for cloud-based architectures through BQ. We build upon [7], customer segmentation methods by incorporating [8], graph-based analysis and [9], sequence analysis techniques.

3 Objectives and Limitations

This research proposes an end-to-end comprehensive solution that addresses the gaps in existing research regarding scalable customer journey analysis in cloud environments. BQ as a cloud platform will be used as preprocessing and Data warehousing solution, focusing on GA4 data. The main goal is to use BQ capabilities for deriving meaningful results from different types of analysis, detect issues in the user journey behavior, and finally, drive from the results to propose future solutions. This type of decision based on customer journey data can create a new perspective on the issues and guide the businesses to invest in the right content, marketing, and strategy.

Our approach will consider a dataset that is publicly available for the purpose of scenario replication in the future. The dataset can be found under Google's public datasets named as - "ga4_ecommerce", [11].

This dataset consists of the most often seen user interactions on the site, such as page views, view items, checkout, and purchases. Some custom events will not be considered in the study.

While BQ optimizes GA4 data analysis, alternative cloud platforms such as Azure or AWS present challenges like setup complexity and additional costs, furthermore, due to the sensitivity of GA4 data, strict compliance with data protection regulations, especially around user privacy and personally identifiable information, is required.

4 Research Methodology

Given that there is limited research on the use of advanced analytics in customer journey using BQ cloud on GA4 data, this study aims to explore different advanced analytics options to detect issues and enhance performance in the customer's journey.

4.1 Data Collection

GA4 provides an advanced framework for tracking and analyzing user behavior across both web and mobile platforms, [12]. To implement GA4, a property must be created via a Google account, followed by the configuration of data streams to monitor website or app traffic, [13]. GA4's built-in features automatically track common interactions, such as page views and clicks, via an enhanced measurement option; however, for more specific data collection, custom events can be established through Google Tag Manager (GTM), [14]. GTM facilitates the deployment of tracking tags without requiring direct modifications to the code, thus streamlining the process, [15], given in Figure 1.

Fig. 1: Collecting data from the Web and app with GTM and sending to GA4

GA4's event-based data model allows for more detailed and adaptable tracking of user behavior compared to the session-based model of Universal Analytics (UA), [16]. Additionally, GA4 has overcome issues that UA faced with compliancy and PII processing by giving an opportunity to completely remove or redact the values, [17].

Additionally, GA4 provides the capability to track cross-platform user journeys as well as compose customized reports called explorations. Due to their limitations and capabilities, the option of exporting to the cloud platform is considered to extend the preprocessing and ease of the analysis, [18].

4.2 Data Export

To begin with the export of GA4 data to BQ, there are certain steps to consider.

First and foremost, a new project must be created within the Google Cloud Platform (GCP), [19]. For the purpose of tracking the costs for preprocessing and modeling, a billing account is created and associated with the corresponding BQ project [19]. The major use of the billing project is for cost planning and possible limits that can be applied on the project level in order to fall within financial planning.

After the project is established, a new link between the GA4 property and the BQ project is enabled. This is done through the GA4 "Admin" panel and "BQ Linking" section, [20].

When choosing the connection options in GA4, the location should be aligned with the project location in BQ to avoid higher costs for processing as well as issues related to latency.

Finally, there is an option to choose between Daily Export or Streaming Export, [21]. The major difference between the two is in the quantity of times the data is coming to BQ. Daily Export updates only once per day at a non-specific time, comparingly, Streaming Export updates the data multiple times per minute and is a recommended option for real-time analysis, [22].

This step in our research will be skipped due to the use of a public dataset in BQ.

4.3 Export Schema

The GA4 schema considers one dataset with a specific definition, "analytics_random.number". If there are multiple links between different GA4 properties and the BQ project, multiple different datasets will be present with different "random.number" in the naming.

The dataset contains a table with the format "event_date" example: (events_YYYYMMDD), [23]. This table is updated, and once the data for the day arrives, the "date" part will change.

The table consists of normal fields as well as nested fields. In the normal fields fall the "event_date", "event_timestamp", "event_name", "event_previous_timestamp". Besides the date, there is also user-related data represented by fields such as "user_id", "session_id" and "user_pseudo_id", [24].

Additionally, there are nested fields "event_params" that hold key-value pairs with values for event parameters associated with events recorded under "event_name".

The final part of the schema is the geographic and traffic fields. With the help of these, we can analyze the source, medium, and campaign from where the user came from, as well as the country, city, and continent that was used for browsing activity.

The most important aspect is that all of this data doesn't include PII, and it is impossible to identify the actual user. The only exception is if the business intentionally decides to collect that data, for which purpose an anonymization approach for handling the data is needed.

The uniqueness of the GA4 schema is in the nested fields. They can be of different types, arrays, or structs (Figure 2.). Data within these fields is encapsulated based on related information. Arrays can hold more entries for one field, which can create repeated events or attributes. Compared to this, structures are grouping related fields together; this technique has an effect on the query performance and organization.

Before using the data, it needs to be addressed with UNNEST to flatten the structure and be able to query it.

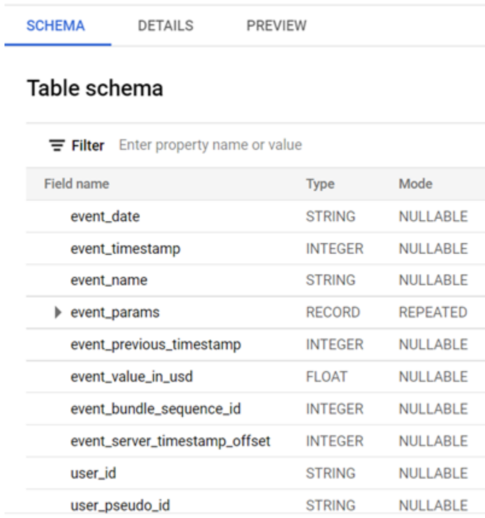

SCHEMA DETAILS PREVIEW

Table schema

Filter Enter property name or value

Field name	Type	Mode
event_date	STRING	NULLABLE
event_timestamp	INTEGER	NULLABLE
event_name	STRING	NULLABLE
event_params	RECORD	REPEATED
event_previous_timestamp	INTEGER	NULLABLE
event_value_in_usd	FLOAT	NULLABLE
event_bundle_sequence_id	INTEGER	NULLABLE
event_server_timestamp_offset	INTEGER	NULLABLE
user_id	STRING	NULLABLE
user_pseudo_id	STRING	NULLABLE

Fig. 2: GA4 Schema

4.4 ELT Process with Google Dataform

Part of the BQ is a service called Dataform (now Pipelines), which has a great variety of uses, such as version control and orchestrating workflows [25], but it can also be used for preprocessing and data modeling.

This study evaluates the process of ELT (Extract, Load, Transform), [26], which considers three major steps:

- extracting raw data from source systems (Creating Data Lake),
- cleaning the data and forming Dim and Fact tables, establishing keys (Staging environment)
- complete data and incremental load (Data Warehouse environment).

The major decision to incorporate Dataform stands in the capability to enhance data processing efficiency. Additionally, Dataform can help with cost reduction through the implementation of incremental load, [27].

The model proposed and used in this research is a star schema. For easier analysis and clear organization of the event parameters, Dim tables are used. The connection between Dim tables is established through keys available in the Fact table.

All the phases and the organization of the tables at each phase are visualized in Figure 3.

To begin with the extraction and load of the data, the Data Lake step is established. For this purpose, a specific directory is created, which holds the SQLX file for the event table where all GA4 data that is needed is preprocessed, mainly to unnest it or to combine and rename fields. In this case, there is only one dataset and, therefore one event table, however, if there are multiple datasets holding data for multiple GA4 properties, multiple event tables will be needed, each with specific definitions.

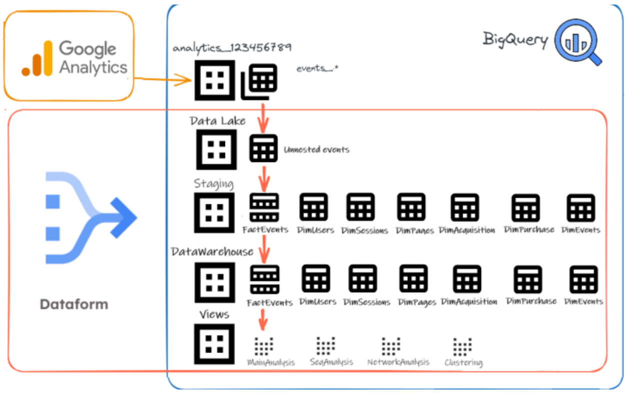


Fig. 3: Star Schema in Dataform

The next step considers creating a new directory for the Staging environment. This step will be composed of SQLX Dim tables and Fact tables that will consider validation, cleaning, combining, and definitions for processing null and non-valid data. The directory organization and format are given in Figure 4.

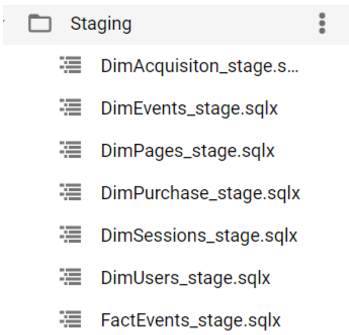


Fig. 4: Staging directory organization

In order to create surrogate keys, a hash key approach is used. This is possible with the FARM_FINGERPRINT function that considers all columns in a table and concatenates them, composing unique 64-bit hashed values. This function is very practical not just for the surrogate keys but also for mapping any length into fixed-length output due to the unique value generated, which guarantees uniqueness, [28]. Once the workflow for the Staging part is finished and the tables are filled with data, the next step is the Data Warehouse.

When working with large volumes of data, it is crucial to avoid loading and transformation on full data. Instead, the most practical approach is to consider chunks of data that are newly arrived. In Dataform, this is possible with the help of incremental load.

To decrease the amount of data even more, an additional step is to use partitioning, so instead of working with one big table, data is stored in several smaller tables. Due to the way of query execution, these declarations must be defined in the pre-operations part.

Finally, to fill tables with data, first, a full load is initialized, after which the incremental load is established, and only the new data coming in is considered for processing, with old data not being affected.

4.5 Data Analysis: Sequence, Clustering, and Network Analysis

In order to add an abstract layer between the original tables and the users, as well as to join the Fact and Dim tables and simplify further queries, Views are introduced, [29]. In our research, three specific views are created:

- Sequence analysis;
- Network analysis;
- Clustering analysis;

Sequence Analysis – In order to detect a critical point in the user journey, either related to behavior or to a present issue, sequence analysis can be considered. In this study, the analysis is implemented in the cart-checkout-purchase flow to detect any drop-offs caused by issues in the process, as well as some higher conversion points that might be updated further.

To detect which action has followed after the key pages, in this case, cart, checkout, and purchase correspondingly, the LEAD() function is considered, [30]. With the help of the function over the mentioned pages, there is the possibility to easily detect users dropping off from the cart page or from the checkout page before making an actual conversion.

By aggregating the sequences of page views and analyzing their frequency, insights into user behaviors are gained. If certain sequences are common among users who convert, these can be enhanced, while sequences leading to drop-offs can be considered for evaluation for possible technical issues in the journeys and further evaluation based on device type.

Additional reverse sequence analysis using the LAG() function provides a deeper understanding of user behavior immediately before conversion, [30]. This dual analysis—forward and reverse—enables businesses to capture a full picture of the user journey, identifying potential pain points in both directions.

Cluster Analysis - Clustering analysis goes beyond simple customer segmentation by allowing businesses to uncover actionable insights that drive customer engagement and retention. This analysis focuses on the use of RFM metrics, [31]:

- Recency: How recently a customer has made a purchase;
- Frequency: How often a customer makes purchases over a given period;
- Monetary Value: The total amount spent by the customer;

These metrics are crucial as they provide a quantitative foundation for understanding customer behavior and purchasing patterns.

The K-means algorithm is then applied to the consolidated dataset to group customers based on their RFM metrics. By specifying a number of clusters (in the testing case, three), the model segments users into different categories, such as 'Loyal', 'At-Risk', and 'Churned', [31].

The analysis pulls data from various tables as sources, including event details, session information, and user demographics, followed by the creation and exploration of k-means models using BQ ML, [32].

The resulting clusters are analyzed to gain insights into the characteristics of each group, including average recency, frequency, and monetary values.

This segmentation allows for targeted marketing strategies tailored to each cluster. For instance, loyalty programs will be developed for high-value customers, while re-engagement campaigns can target those at risk of churn.

Network Analysis – Network analysis is a key technique that any e-commerce business should consider if struggling with sales of certain products. This is the analysis that provides information on the relationship of the products, such as products bought together or related products, [33]. As described further in the study, this analysis can give information on the possible placement, marketing, and promotions regarding the products, as well as further personalization tools that can retarget customers and bring them back to the page to make conversion.

In order to relate products with specific events of interest, we used the “purchase” event in the study.

This event has been enriched with self-join operation [34], which helps to detect the co-occurrence of items within the session and the possibility to relate the items. The following step is used to count the frequency of occurrence in the item pairs within sessions.

As a final step, the results derived can be exported from the BQ and imported into specific platforms such as CDP/CRM platforms that can reposition and optimize the items on the page, as well as optimize strategies and personalize item targeting of the users.

5 Major Findings

The section will give information on the insights derived from the previously introduced analyses.

5.1 Sequence Analysis Results

Sequence analysis considering page_view events occurring on the cart, checkout, and purchase pages has been used to analyze the number of users and their sessions on these steps. With the help of LAG() and LEAD() functions, forward and backward analysis was possible in order to detect a stable flow of users on each of the steps.

The forward sequence analysis has shown that nearly 48% of users moved from the cart to checkout and purchased without any drop-offs. This is a high percent of users that are making conversions on the site. However, this percentage might grow further if coupon codes are delivered to the customers through email marketing campaigns to bring back the users and finish the journey with conversion (Figure 5).

The backward sequence analysis has shown that 94% of users ended the purchase in less than three steps, meaning there were no repeated movements between cart, checkout, and purchase, also proving the absence of technical issues in the purchase process (Figure 5).

As a conclusion from the sequence analysis, the website has a healthy checkout and purchase journey, with no pain points detected that also might consider better marketing strategies such as email retargeting to drive back users that didn’t finish the journey with conversion.

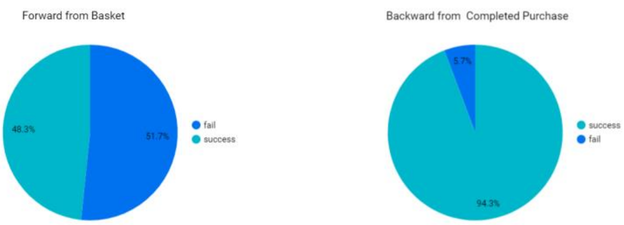


Fig. 5: Pie chart of data derived from sequence analysis

5.2 Network Analysis Results

Network analysis provided insights into complex interactions among products, providing a deeper understanding of user purchase patterns.

To better understand the connections between items, data derived from BQ analysis was exported to a tool that supports network visualizations by Flourish, [35]. The visualization used for this purpose is a radial network graph, which provides information on items' interconnections.

Radial network graphs display items as nodes, and the relationships between them are represented by lines. Each item has a relationship with a different weight (force) from the item edges, and the interconnected items are assigned a weighted factor. To provide a clearer image of the most popular and frequently bought products, the plotting is limited to the top 50 products to avoid cluttering the network.

As seen in Figure 6, the radial network graph illustrates key items such as “Google Light Pen Green” and “Google Camp Mug Ivory” as central nodes, indicating high co-purchase frequencies. This suggests potential for product bundling strategies or placing of recommendations.

Peripheral nodes/items like “YouTube Small Sticker Sheet” and “Google Red YoYo” show fewer connections, pointing to specific demands or items of low interest that could benefit from targeted marketing efforts.

Considering items such as “Google Crewneck Sweatshirt Grey” and “Google Camp Tee Red,” we can see the relationship that shows frequent co-occurrence, presenting opportunities for effective cross-selling marketing strategy.

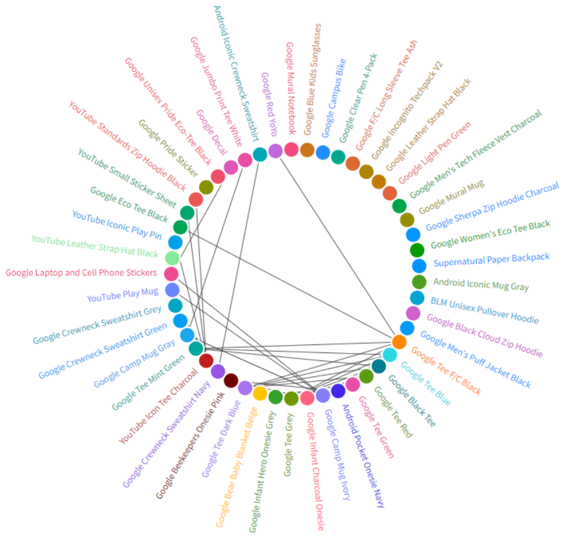


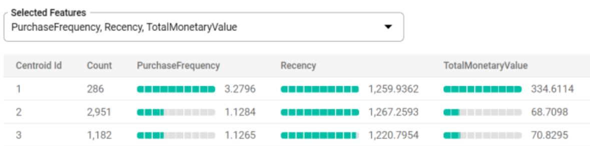
Fig. 6: Radial Network Graph for representing Network Analysis results, [35]

These findings highlight how network analysis can inform marketing strategies by leveraging item associations to influence customer purchase behavior, providing ideas on how to organize and place items, and adjusting the recommendation engines on the site.

5.3 Clustering Results

Clustering analysis provides insights into customer engagement levels, identifying segments of users based on their purchase behavior.

To prepare customer engagement analysis, the K-means clustering method is used. Users were classified into three distinct clusters based on frequency, recency, and monetary value of purchases. Cluster 1 contained the smallest number of users but represented the most valuable regular customers, while Clusters 2 and 3 included a larger portion of users with significantly lower purchase frequency, for example, the division given in Figure 7.


Selected Features
PurchaseFrequency, Recency, TotalMonetaryValue

Centroid Id	Count	PurchaseFrequency	Recency	TotalMonetaryValue
1	286	3.2796	1,259.9362	334.6114
2	2,951	1.1284	1,267.2593	68.7098
3	1,182	1.1265	1,220.7954	70.8295

Fig. 7: Cluster division of users based on their FRM

The findings indicate that more than 90% of users fall into Clusters 2 and 3 (Figure 8), categorizing them as potentially "endangered" customers who may require targeted promotions or engagement strategies to reestablish their purchasing interest.

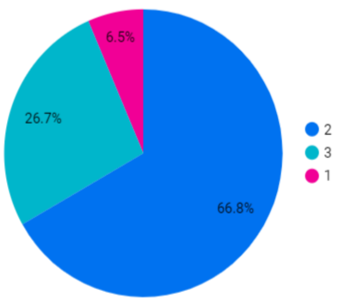


Fig. 8: Pie chart visualization of clusters results

Most customers originated from the USA, India, and Canada, with referral traffic from other Google domains being the primary source of new customers. Considering that most of the users come from referral traffic, consideration about enlarging the budget and focusing on increasing further referrals must be done.

These results underscore the need for a refined marketing strategy to retain new customers, focusing on engagement activities and promotions that can reestablish customer loyalty (Figure 9).

Fig. 9: Categorical features, Source, and Location of customers and their separation per cluster

6 Discussion

For applicability, the research utilized "ga4_ecommerce," the only publicly available dataset from BQ. This dataset captures the most frequently used eCommerce events across websites and apps, offering a good foundation for comprehensive analysis of user behavior through sequence analysis, network analysis, and clustering analysis, which has given significant insights into the efficiency of the e-commerce platform and the dynamics of customer interactions.

The sequence analysis, employing both forward and backward methods, demonstrated a high level of efficiency in the cart-to-purchase flow. Nearly 50% of customers progressed to the purchase stage within five steps of visiting the cart, indicating a streamlined funnel with intuitive navigation. Additionally, backward analysis confirmed that most customers followed the expected path to the purchase page without encountering major technical issues. These results align with industry benchmarks for optimized purchase funnels [36], emphasizing the importance of maintaining a user-friendly interface. However, the limitations of sequence analysis should be acknowledged, particularly in capturing user behaviors that need further investigation. Those limitations can be addressed by examining various steps of the user journey beyond just the basket and purchase relationship, as well as taking into consideration other events beyond just simple page views.

Network analysis, using data displayed as a radial network graph, revealed insights into the connections between items. These insights propose possible placing adjustments in the content, such as creating personalized recommendations that can be positioned on item details pages or cart pages to stimulate spending. Additionally, less related items that are isolated from others can stand out with retargeting and promotional campaigns or vouchers delivered through email. Besides the deep understanding of item relationships, there is a limitation from BQ in the visualization of them. This can be solved by exporting and importing the data into specific visualization software or tools.

With cluster analysis, we could identify engagement patterns in the user journey, helping to understand which customers to classify as “regular” and which to consider being “at risk” of losing. The clustering considered several conditions such as frequency and recency of purchases as well as monetary value (spending). Results concluded that creating targeted marketing strategies and promotions is essential for re-engaging "at-risk" customers and retaining "regular" customers.

The major limitation of the analysis is the limited number and type of clusters. The solution needs to consider extending the analysis to include more key performance indicators (KPIs), such as engagement time and the number of custom events, such as clicks on promotional banners.

7 Conclusion

Customer journey analysis is the foundation of targeted and personalized marketing strategy. The deeper the analysis is, the better, and more related products are offered to the corresponding audience. With the help of BQ’s capabilities to process, model, and analyze GA4 data, we could understand the behavior of customers through their journey on e-commerce sites, as well as how to organize and place items and which audience to consider with more campaigns.

The study introduced three different analyses. Sequence analysis was established through forward and backward evaluation of page view events considering steps needed for the user to pass from cart to purchase and vice versa. With the help of data, we could conclude that the purchase flow was normal and there were no anomalies from a technical perspective. Additionally, the analysis provided information for the people who drop off at the cart stage, leading to an audience definition that can be used for retargeting in later email campaigns.

Network analysis was used for the purpose of detecting relationships between items, leading to the conclusion which items to position on site together, as well as to create recommendations.

Finally, the cluster analysis provided an overall evaluation of engagement and organized the customers into different clusters, such as those who are at risk since they haven't purchased in a while and need retargeting or aggressive marketing and those who are regular customers who could be preserved with some promotional offers.

In conclusion, the duo GA4 data and BQ platform can provide a thorough customer journey analysis. An e-commerce site is a major consideration, where evaluating the journey is crucial to success. By deriving the right information from the journey data, the business can create a detailed and specific marketing strategy that can decrease costs and increase interactions and conversions on the site.

8 Future Work

Introduced work provided a deep understanding of the concepts and their capability using ecommerce GA4 data. However, there are more details that could enrich the analyses further.

From a sequence analysis perspective, evaluating more customer interactions at each step, not just relying on page views, could be used.

Considering the network analysis, alternative visualization tools from a non-paid nature can be evaluated, replacing the proposed service – Flourish.

Finally, cluster analysis can be extended to evaluate more metrics like engagement time or different custom interactions. There is an opportunity to enhance the cluster analysis with time series in order to evaluate time from add-to-carts to an actual purchase.

Leading research can be done in cross-channel attribution. There is an opportunity to derive insights from different marketing options used – social media, paid advertising, or email campaigns. The values for conversion rates and click-through rates can provide an idea of the engagement of customers on each of these channels, choosing the right strategy and investment in the campaigns, [37].

Finally, in the era of AI and machine learning, there is an opportunity to use BQ ML for predictive analysis and detection of customers' buying patterns. This research can provide an idea of how powerful ML can be in marketing decision-making, [38].

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used Grammarly and ChatGPT to correct grammatical and syntactical errors, improving the study's correctness and readability. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References:

[1] Lemon N. K. and Verhoef C. P. Understanding Customer Experience Throughout the Customer Journey, *Journal of Marketing,* vol. 80, no. 6, 2016, pp. 69-96. DOI: 10.1509/jm.15.0420.

[2] Verhoef C Peter, Kannan P.K. , Inman J. Jeffrey ,From Multi-Channel Retailing to Omni-Channel Retailing Introduction to the Special Issue on Multi-Channel Retailing, *Journal of Retailing,* 91(2), 2015, pp.174-181. DOI: 10.1016/j.jretai.2015.02.005.

[3] Matz, Sandra C, Netzer, Oded, Using Big Data as a window into consumers' psychology, *Current Opinion in Behavioral Sciences,* 18, 2017, pp.7-12. DOI: 10.1016/j.cobeha.2017.05.009.

[4] Davenport H. Thomas and Harris G. Jeanne, *Competing on Analytics: The New Science of Winning*, Harvard Business Press, 2006.

[5] McGuirk Mike, Performing web analytics with Google Analytics 4: a platform review, *Journal of Marketing Analytics,* 11, 2023, pp.854–868. DOI: 10.1057/s41270-023-00244-4.

[6] Kimball, Ralph, and Margy Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling,* Wiley, 2013. ISBN: 978-1-118-53080-1.

[7] Gomes, Miguel A., and Tobias Meisen, A review on customer segmentation methods for personalized customer targeting in e-commerce use cases, *Inf Syst E-Bus Manage,* 21(6), 2023, pp:527-570. DOI: 10.1007/s10257-023-00640-4.

[8] Marquez H. Jacob, Decoding the customer journey with graph node embeddings, Medium, 2023, [Online]. https://medium.com/data-science-at-microsoft/decoding-the-customer-journey-with-graph-node-embeddings-74eb983e9847 (Accessed Date: May 5, 2024).

[9] Santos Susana and Gonçalves M. Helena, Consumer decision journey: Mapping with

real-time longitudinal online and offline touchpoint data, *European Management Journal*, 42(3), 2024, pp.397-413. DOI: 10.1016/j.emj.2022.10.001.

[10] Heka.ai, Understanding the customer purchase journey with visualization and clustering, Medium, 2023, [Online]. https://heka-ai.medium.com/understanding-the-customer-purchase-journey-with-visualization-and-clustering-f0be391810ec (Accessed Date: May 6, 2024).

[11] Google Analytics, BigQuery sample dataset for Google Analytics ecommerce web implementation, Google, 2024, [Online]. https://developers.google.com/analytics/bigquery/web-ecommerce-demo-dataset (Accessed Date: May 10, 2024).

[12] Google Analytics, [GA4] Introducing the next generation of Analytics, Google Analytics 4, Google, 2024, [Online]. https://support.google.com/analytics/answer/10089681?hl=en (Accessed Date: May 10, 2024).

[13] Google Analytics, Set up Analytics for a website and/or app, 2024, [Online]. https://support.google.com/analytics/answer/9304153?hl=en (Accessed Date: May 10, 2024).

[14] Fedorovicius Julius, How to Track Events with Google Analytics 4 and Google Tag Manager, 2024, [Online]. https://www.analyticsmania.com/post/how-to-track-events-with-google-analytics-4-and-google-tag-manager/ (Accessed Date: May 10, 2024).

[15] Metriclabs, Google Tag Manager, 2024, [Online]. https://metriclabs.com.au/google-tag-manager/ (Accessed Date: May 6, 2024).

[16] Google Analytics, [UA→GA4] How differences between Universal Analytics and Google Analytics 4 impact key event counts, Google, 2024, [Online]. https://support.google.com/analytics/answer/13881440?hl=en (Accessed Date: May 6, 2024).

[17] Maneno Vincent, The Role of Google Analytics 4 in Data Privacy and Compliance: What You Need to Know, Metricslab, 2024, [Online]. https://metriclabs.com.au/the-role-of-google-analytics-4-in-data-privacy-and-compliance-what-you-need-to-know (Accessed Date: May 9, 2024).

[18] Ayesha Khan, How does Google Analytics 4 detect Trend Changes and what Insights does it provide?, Analytico, 2024, [Online]. https://www.analyticodigital.com/blog/google-analytics-4-detect-trend-changes-and-insights (Accessed Date: May 9, 2024).

[19] Dealer insights, Setting Up a BigQuery Project: A Step-by-Step Guide – Part 1, DealerInsights, 2023, [Online]. https://dealerinsights.com/blog/setting-up-a-bigquery-project-a-step-by-step-guide (Accessed Date: May 10, 2024).

[20] GA4Bigquery, How to set up BigQuery linking in your Google Analytics 4 property, ga4bigquery, 2024, [Online]. https://www.ga4bigquery.com/tutorial-how-to-set-up-bigquery-linking-in-your-google-analytics-4-property-ga4/ (Accessed Date: May 10, 2024).

[21] Google Analytics, [GA4] Set up BigQuery Export, Google, 2024, [Online]. https://support.google.com/analytics/answer/9823238 (Accessed Date: May 10, 2024).

[22] Ievgen Krasovytskyi, [GA4] BigQuery Export: How to Connect Step-by-Step, OWOX, 2024, [Online]. https://www.owox.com/blog/articles/google-analytics-4-to-bigquery-guide/ (Accessed Date: May 10, 2024).

[23] Alyona Samovar, Anna Panchenko, Ievgen Krasovytskyi, [GA4] BigQuery Export: Events Table Schema and Managing Dates, OWOX, 2024, [Online]. https://www.owox.com/blog/articles/ga4-bigquery-export-event-table-schema-and-dates/ (Accessed Date: May 10, 2024).

[24] Google Analytics, [GA4] BigQuery Export schema, Google, [Online]. https://support.google.com/analytics/answer/7029846?hl=en (Accessed Date: May 10, 2024).

[25] Dolan Lee, Dataform, what's the story?, Medium, 2023, [Online]. https://medium.com/cts-technologies/dataform-whats-the-story-ffa66d83d484 (Accessed Date: May 12, 2024).

[26] IBM, What is extract, load, transform (ELT)?, IBM, 2024, [Online]. https://www.ibm.com/topics/elt (Accessed Date: May 12, 2024).

[27] GA4bigquery, How to create and maintain a Dataform powered GA4 export SQL workflow in BigQuery, ga4bigquery, 2024, [Online]. https://www.ga4bigquery.com/how-to-create-and-maintain-a-ga4-export-sql-workflow-in-bigquery-using-dataform/ (Accessed Date: May 12, 2024).

[28] Artem Korneev, Dataform: prepare GA4 events and keep them shiny, gtm-gear, 2024, [Online]. https://gtm-gear.com/posts/dataform-ga4-events/ (Accessed Date: May 16, 2024).
[29] Google Cloud, Introduction to views, Google, 2024, [Online]. https://cloud.google.com/bigquery/docs/views-intro (Accessed Date: May 17, 2024).
[30] Griese Gunnar, How to replicate the GA4 Path Exploration report with BigQuery SQL?, GG, 2022, [Online]. https://gunnargriese.com/posts/bigquery-path-exploration/ (Accessed Date: May 17, 2024).
[31] Anjum Munaz, Understanding 'At-Risk' and 'Loyal' Customers with GA4's Public Data Using Clustering, Medium, 2023, [Online]. https://munaz-anjum.medium.com/understanding-at-risk-and-loyal-customers-with-ga4-s-public-data-using-clustering-3dbaad45411a (Accessed Date: May 18, 2024).
[32] Google Cloud, The CREATE MODEL statement for K-means models, Google, 2024, [Online]. https://cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-syntax-create-kmeans (Accessed Date: May 20, 2024).
[33] Ma, Jian & Sun, Yong-Hong & Hao, Jinxing & Sun, Yongqiang & Zhao, Yongqiang, Using Social Network Analysis as a Strategy for E-Commerce Recommendation, *PACIS 2009,* 106, 2009
[34] Samuel Langenfeld, Understanding SQL Self Joins, Medium, 2017, [Online]. https://medium.com/@samlangenfeld/understanding-sql-self-joins-b39426404aa . (Accessed Date: May 22, 2024).
[35] Flourish, Data visualization with Radial Network Graph, Flourish, 2024, [Online]. https://app.flourish.studio/@flourish/network-graph (Accessed Date: May 22, 2024).
[36] Lauren O'Bryan, B2B SaaS Funnel Conversion Benchmarks, poweredbysearch, 2024, [Online]. https://www.poweredbysearch.com/learn/b2b-saas-funnel-conversion-benchmarks/ (Accessed Date: May 22, 2024).
[37] Luka Cempre, GA4 Bigquery Tips: Guide to Attribution, Adswerve, 2023, [Online]. https://adswerve.com/technical-insights/ga4-bigquery-tips-guide-to-attribution (Accessed Date: May 23, 2024).
[38] Google Cloud, Create machine learning models in BigQuery ML , Google, 2024, [Online]. https://cloud.google.com/bigquery/docs/create-machine-learning-model (Accessed Date: May 24, 2024).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
Each author made an equal contribution to the current study, from the conceptualization of the problem to the final conclusions and resolution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The authors have no conflicts of interest to declare.