Data Engineering Technical Assessment
Part III: Behavioral Segmentation with Elasticsearch & AI

Mohamed Amin Soumih

May 27, 2025

Objective

Segment users based on behavioral data stored in Elasticsearch. The goal is to extract user search activity, understand intent patterns, and cluster users into actionable marketing personas using machine learning.

Data Preparation

- 1,000 realistic user session documents were generated and inserted into Elasticsearch (index: `user_sessions`).
- Each document contains: `user_id`, `search_query`, `clicked_product_ids`, and `timestamp`.
- Search queries were designed to reflect real e-commerce behavior (e.g., `"gaming laptop"`, `"cheap phone"`).

Clustering Approach

- **TF-IDF** was used to vectorize search queries.
- **KMeans** clustering was applied to group users into 5 segments based on their search intent.
- Each user was assigned a cluster ID and then mapped to a descriptive label based on dominant keywords in the cluster.

Segment Labels

Users were categorized into meaningful groups such as:

- `tech_enthusiast`: frequent terms like ”gaming”, ”keyboard”, ”monitor”
- `budget_buyer`: queries containing ”cheap”, ”discount”, ”offer”
- `fashion_oriented`, `other`: detected from keywords like ”dress”, ”handbag”, or fallback logic

Final Storage

- The segmented users were written back into Elasticsearch under a new index: `user_segments`.
- Each document contains: `user_id`, `search_query`, and `segment`.
- Bulk insertion was used for efficiency (`helpers.bulk`).

Conclusion

This project aimed to deliver a comprehensive data-driven solution for understanding and optimizing user behavior on an e-commerce platform. Across the three phases, we conducted data exploration, cohort retention analysis, and user segmentation to extract actionable insights.

In particular, **Part 3 focused on behavioral segmentation** using the search queries of users. We applied TF-IDF vectorization to numerically represent user interests, followed by KMeans clustering to group users into meaningful segments. These segments, such as *tech_enthusiasts*, *budget_buyers*, and *fashion_oriented*, allow the platform to tailor its marketing, product recommendations, and user experience strategies.

The final segmented data was indexed into Elasticsearch, enabling fast querying and future integrations with dashboards or APIs. We also provided a Dockerized environment for easy setup and reproducibility, ensuring our solution can be deployed efficiently in real-world settings.

Overall, this end-to-end solution demonstrates the impact of combining data science, clustering algorithms, and engineering best practices to support personalized, scalable, and insightful decision-making in an e-commerce context.