Customer Retention & Cohort Analysis with the Modern Data Stack

July 9, 2025

Project Overview

In this project, we analyzed customer retention and repeat purchases using modern tools:

- **Fivetran** to ingest data from GCP Cloud SQL
- **Databricks** to transform and model the data (Delta Lake, SQL)
- **Dashboards** to visualize cohort trends

We focused on:

- When customers return to buy again
- How retention differs across monthly cohorts
- What insights we can use to improve marketing

Data Pipeline

How the Data Flows

- **Data Ingestion**:
  Fivetran pulls e-commerce sales data from GCP Cloud SQL into Delta Lake in Databricks

- **Data Transformation**:
  SQL in Databricks is used to:

  - Find each customer's first and second purchase dates

  - Group customers into cohorts by first purchase month

- **Visualization**:
  Dashboards in Databricks show retention trends and repeat purchase behavior

What Is a Cohort?

A **cohort** is a group of customers who made their **first purchase in the same month**.

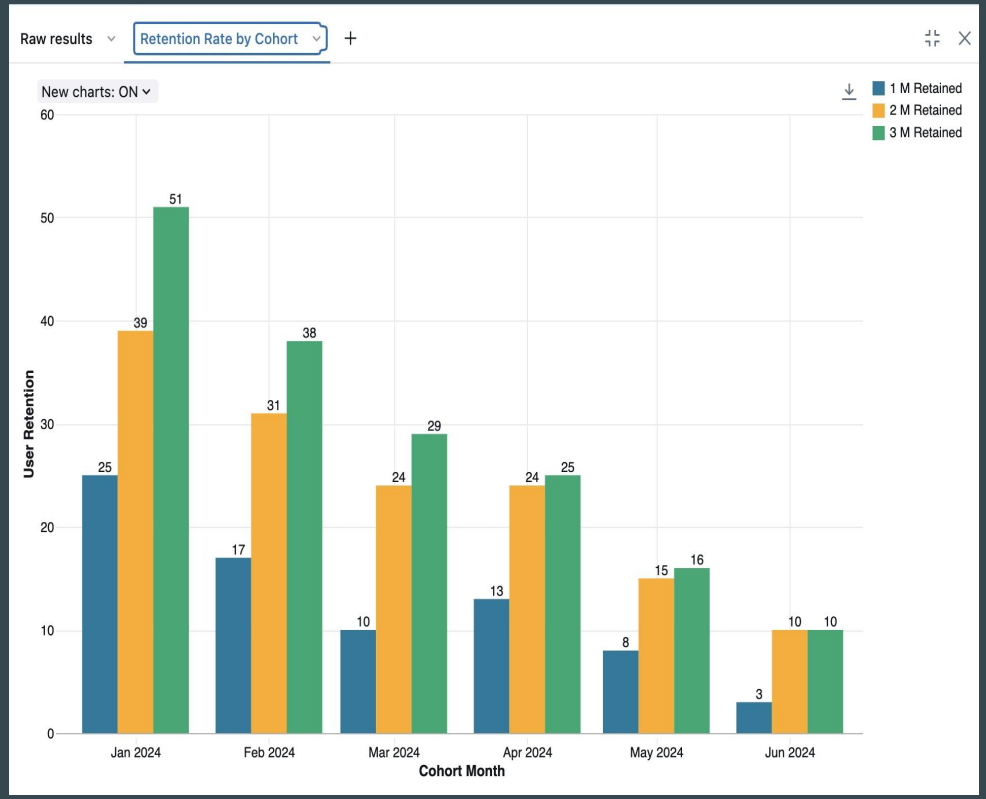
We then track:

- How many of them made a **second**, **third**, or **fourth purchase**
- How quickly they returned

This helps us understand how **customer behavior changes over time**.

Visualization 1 –

Retention Rate by Cohort



1-Month Retention:

- Highest: **May 2024 (50%)**

2-Month Retention:

- Highest: **June 2024 (100%)** *(only 10 customers – small sample)*

3-Month Retention:

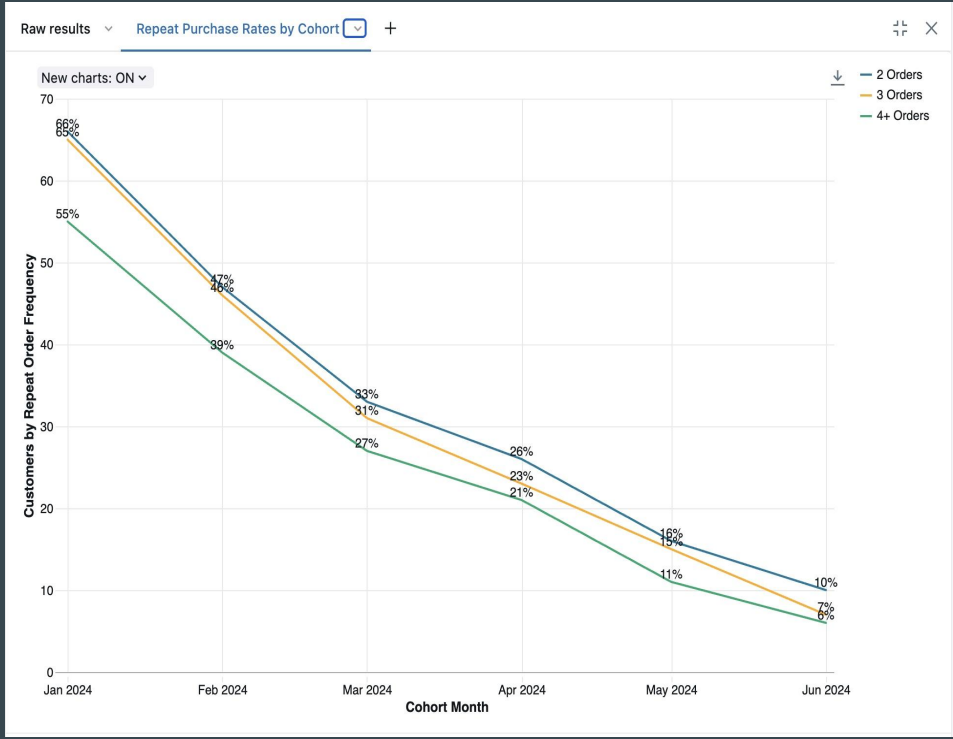
- Highest: **May & June 2024 (100%)**

Observation:

- **Newer cohorts** (Apr–Jun) show **better short-term retention**
- **Older cohorts** (Jan–Mar) had lower early retention (~30–38%)

Visualization 2 –

Repeat Purchase by Cohort



2nd Purchase Rate:

- Very high (96–100%) across all cohorts

3rd Purchase Rate:

- Declines over time: Jan (98.5%) → Jun (70%)

4th Purchase Rate:

- Clear drop: Jan (83.3%) → Jun (60%)

Interpretation:

- **Older cohorts** (Jan–Mar) had stronger long-term loyalty
- **Newer cohorts** make fewer follow-up purchases

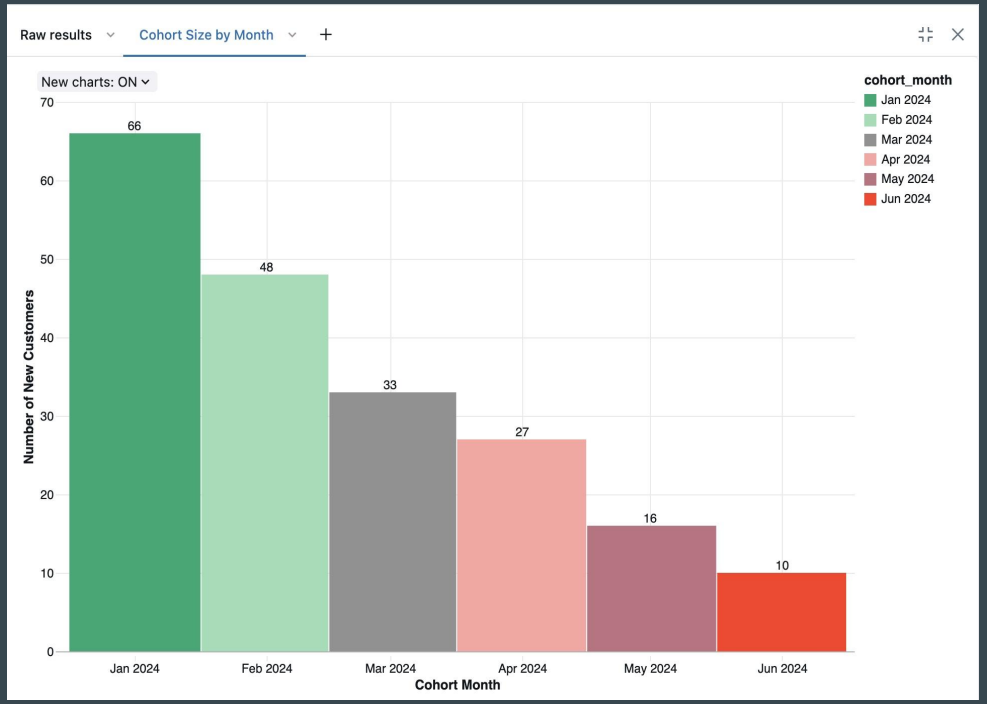
Visualization 3 –

Cohort Size by Month

New customer numbers are going **down each month**:

- **January 2024**: 66 customers *(peak month)*
- June 2024: only 10 customers

This shows a **steady drop in new acquisitions**, month over month.



Summary of Insights

- Retention is **improving in the short term** for newer cohorts
- **Repeat purchase rates** are **declining** after the 2nd order
- **Older cohorts** had stronger long-term engagement
- New customer **acquisition is slowing down**

These insights help us improve:

- **Customer retention strategies**
- **Marketing campaigns**
- **User experience**

Conclusion

- This project showed how tools like Fivetran and Databricks can work together to build a **scalable, automated cohort analysis solution**. It provided key insights into retention patterns and helped identify opportunities for **marketing and product strategy improvements**.

- By understanding customer behavior over time, the company can take **data-driven actions** to improve retention, engagement, and long-term value.