# Understanding Customer Behavior

## Cohort Insights Powered by a Modern ELT Pipeline

Roberto Pera

# Unlocking Customer Value:
# Cohort-Based Insights from 2024

01

**Data:**

- Jan-Dec 2024
- 200 customers
- 1.000 orders

02

**Research Questions:**

- How quickly do customers repurchase?
- How does retention differ across cohorts?
- How strong are repeat-purchase funnels?
- How have cohort sizes changed over time?

# Disclaimer
## Data Context & Limitations

01 **Datascope:**

- customer first purchases fall within Jan–Jun 2024 → only six cohorts
- sample too small to infer robust causal patterns

02 **Limitations:**

- we do **not** know the business model, industry, or seasonality patterns
- we do **not** know the marketing mix, budget, or campaign history

03 **Interpretation:**

- findings should be viewed as **directional**, not definitive
- recommendations are **hypotheses** for further investigation
- additional data would be needed for strategic decisions

# ELT Data Stack:
# From Cloud SQL Extraction to Databricks Transformation


**Extract**

raw data from Cloud SQL

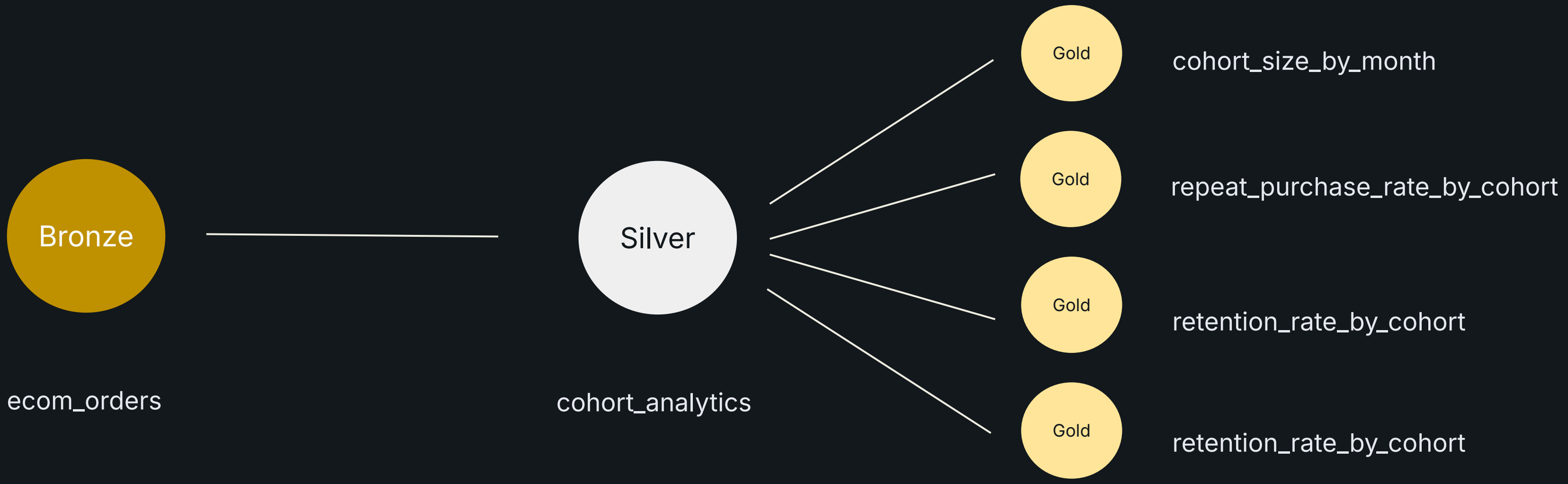
**Load**

directly into Databricks via Fivetran

**Transform**

entirely inside the Lakehouse (SQL + Delta)

# Medallion Architecture
## Structured Data Layers ensure Analysis Quality and Reproducibility



# Medallion Architecture

Silver

```sql
-- Step 1: Calculate the First Purchase Date
WITH first_purchases AS (
  SELECT
    customer_id,
    MIN(order_date) AS first_purchase_date,
    COUNT(*) AS total_orders
  FROM workspace.bigquery_db_cohort_db.bronze_ecom_orders
  GROUP BY customer_id
),

-- Step 2: Calculate the Second Purchase Date
second_purchases AS (
  SELECT
    eo.customer_id,
    MIN(eo.order_date) AS second_purchase_date
  FROM workspace.bigquery_db_cohort_db.bronze_ecom_orders AS eo
  JOIN first_purchases fp ON eo.customer_id = fp.customer_id
  WHERE eo.order_date > fp.first_purchase_date
  GROUP BY eo.customer_id
)

-- Step 3: Combine the Results and Calculate the Days Between Purchases
SELECT
  fp.customer_id,
  first_purchase_date,
  total_orders,
  second_purchase_date,
  DATEDIFF(second_purchase_date, first_purchase_date) AS days_between_first_and_second
FROM first_purchases AS fp
LEFT JOIN second_purchases AS sp ON fp.customer_id = sp.customer_id
```

- Derive 1st and 2nd purchase dates
- Calculate days between purchases
- Count total customer orders

# Medallion Architecture

Gold

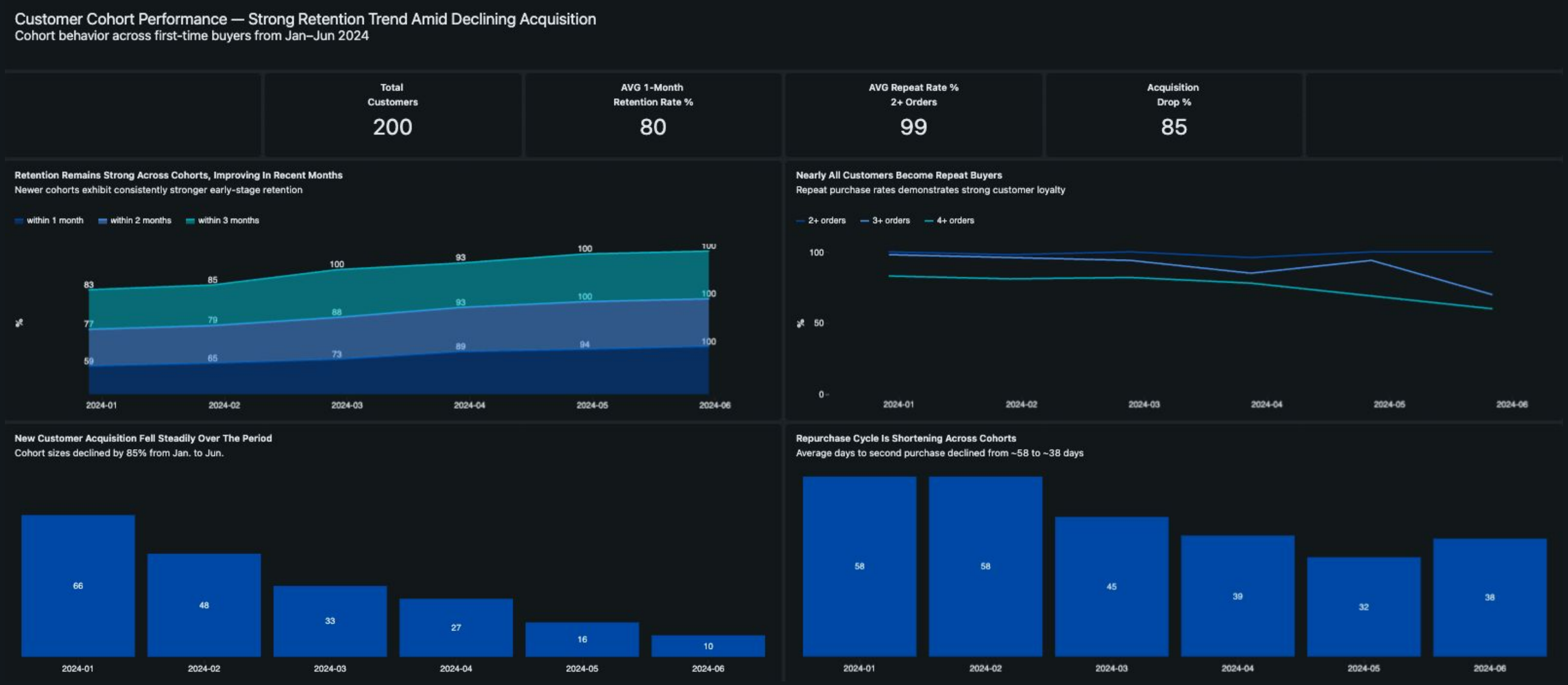
```sql
create or replace table workspace.bigquery_db_cohort_db.gold_retention_rate_by_cohort as

-- 1) Compute the month difference and cohort month
WITH customer_differences AS (
    SELECT
        customer_id,
        DATE_TRUNC('month', first_purchase_date) AS cohort_month,
        first_purchase_date,
        second_purchase_date,
        FLOOR(
            MONTHS_BETWEEN(second_purchase_date, first_purchase_date)
        ) AS month_diff
    FROM workspace.bigquery_db_cohort_db.silver_cohort_analysis
)

-- 2) Aggregate cumulative retention rates
SELECT
    DATE_FORMAT(cohort_month, 'yyyy-MM') AS cohort_month,
    COUNT(*) AS num_customers,
    ROUND(100.0 * SUM(CASE WHEN month_diff <= 1 THEN 1 END) / COUNT(*)) AS retention_rate_1m,
    ROUND(100.0 * SUM(CASE WHEN month_diff <= 2 THEN 1 END) / COUNT(*)) AS retention_rate_2m,
    ROUND(100.0 * SUM(CASE WHEN month_diff <= 3 THEN 1 END) / COUNT(*)) AS retention_rate_3m
FROM customer_differences
GROUP BY cohort_month
ORDER BY cohort_month;
```

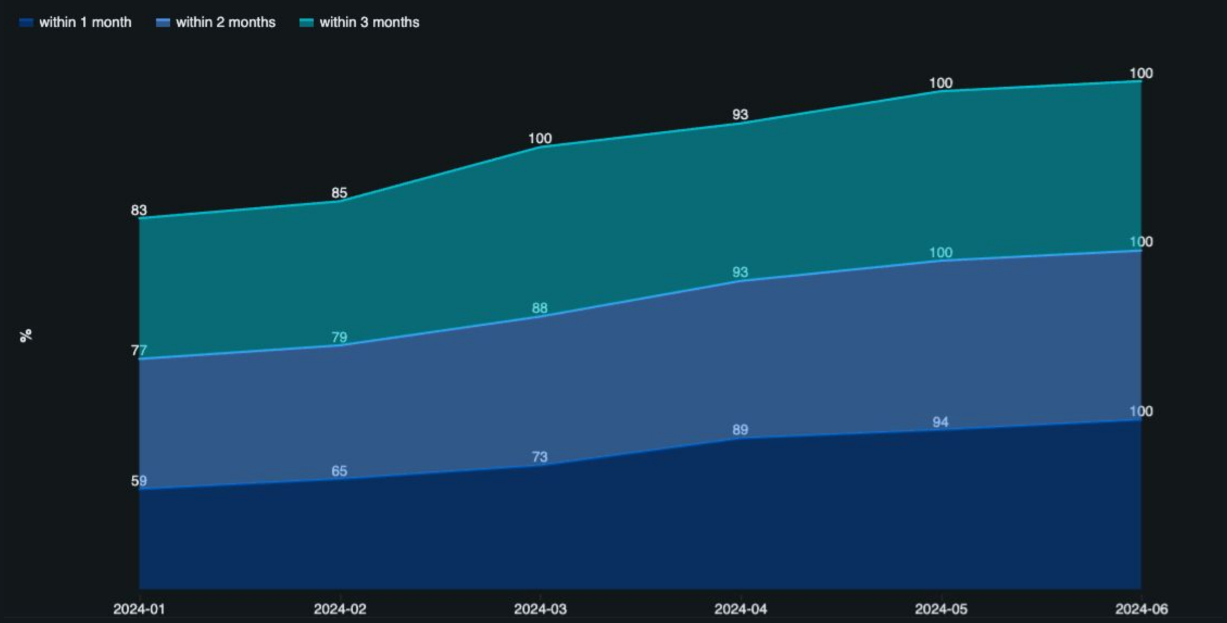
- Calculate month difference: 1st and 2nd purchase
- Compute cohort_month
- Aggregate cumulative retention rates

# Dashboard Overview



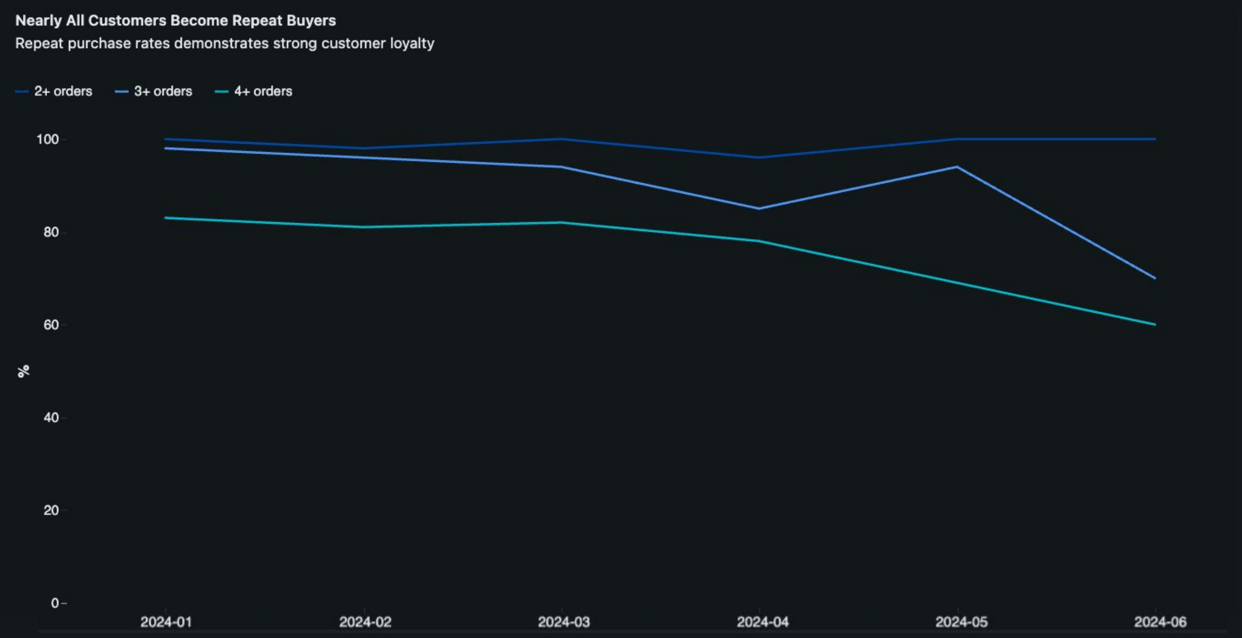
# Retention Remains Strong Across Cohorts, Improving In Recent Months

## Newer cohorts exhibit consistently stronger early-stage retention




- **1-month retention:** 59% → 100%
- **2-month retention:** rises steadily across cohorts
- **3-month retention:** stabilizes around ~90% in later months

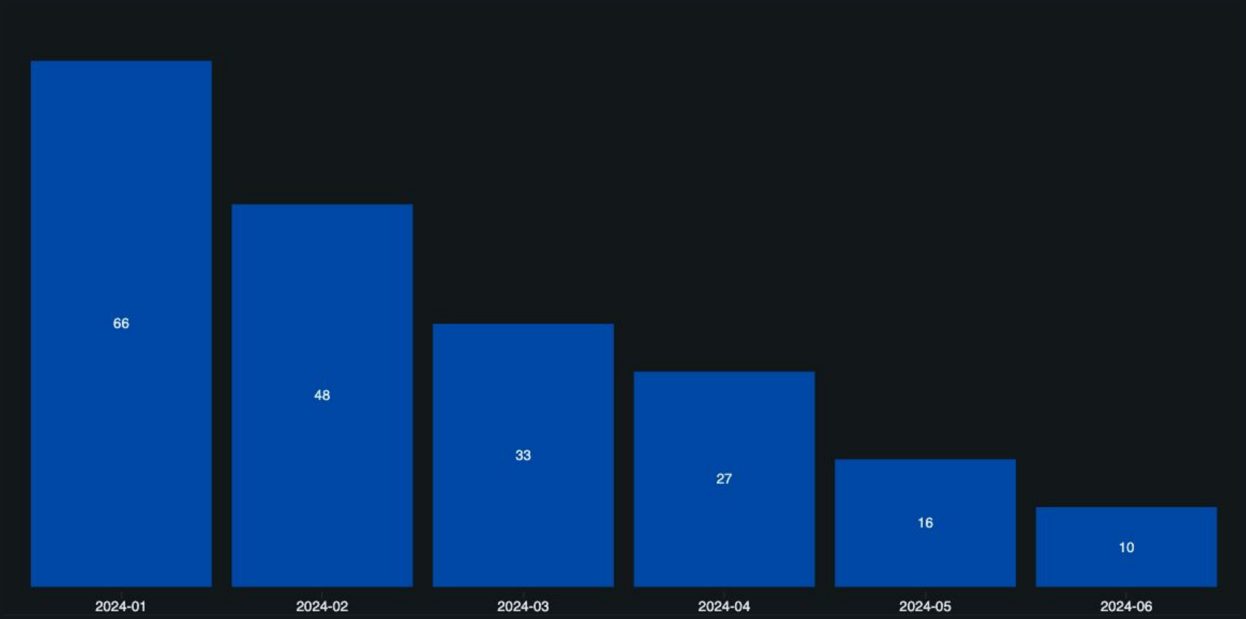
# Nearly All Customers Become Repeat Buyers
## Repeat purchase rates demonstrates strong customer loyalty



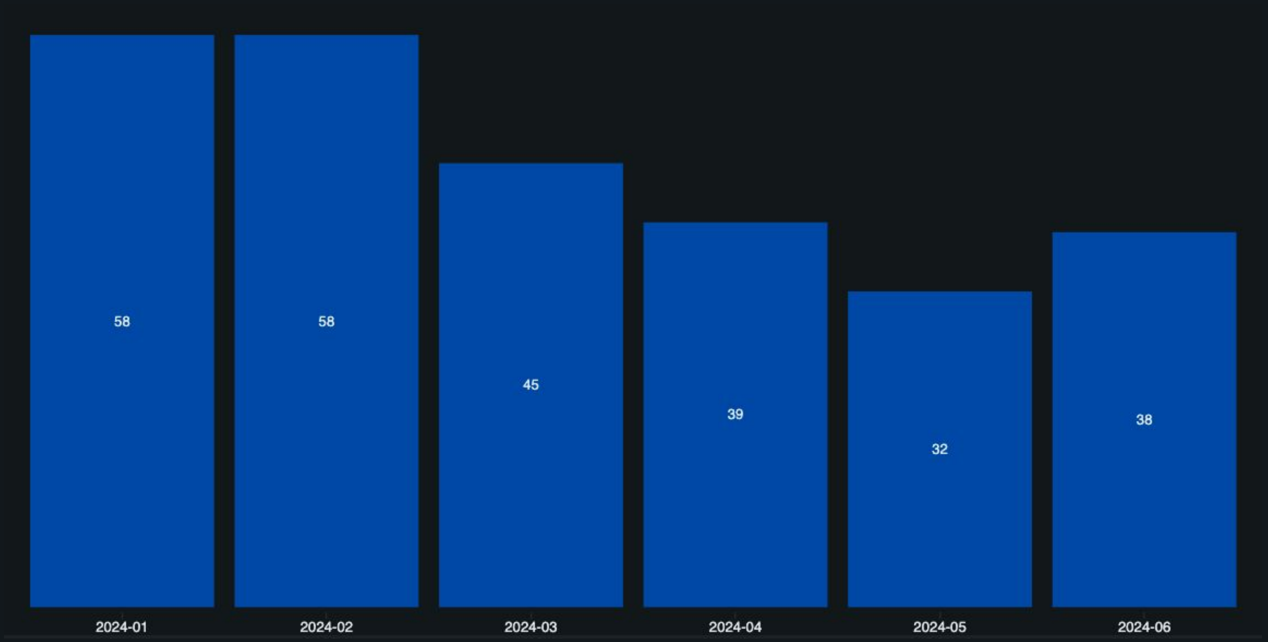

- **2+ orders:**
  ~100% across all cohorts
- **3+ orders:**
  85–96% for early cohorts; slight dip for June
- **4+ orders:**
  strong, but decreasing in latest cohort

# New Customer Acquisition Fell Steadily Over The Period

## Cohort sizes declined by 85% from Jan. to Jun.



# Repurchase Cycle Is Shortening Across Cohorts
## Average days to second purchase declined from ~58 to ~38 days



# Conclusions & Recommendations

### 1 — Retention appears strong

- Later cohorts show higher early-stage retention
- This may indicate improved onboarding, acquisition quality, or product-market fit

**Next step:**
Review whether internal process changes occurred during the period

### 2 — Drastically declining acquisition volume

- This might reflect seasonality, marketing shifts, or one-time events
- With only six data points, no trend can be confirmed

**Next step:**
Investigate channel performance, seasonality, and marketing spend

### 3 — Repurchase cycles seem to be shortening

- Could reflect better customer engagement or improved product experience
- But later cohorts have less elapsed time, so estimates may be biased downward

**Next step:**
Validate with larger time range and longer follow-up period

### 4 — Opportunity to strengthen analytics foundations

- Consider scheduling ETL workflows in Databricks once data volume grows
- Potential long-term enhancements:
    - LTV models
    - Churn prediction
    - Segmentation (behavioral cohorts)

# Thank you

Full project documentation on