Terre des Hommes

Data Engineer - Assignment
Minh Truong
2025.08.25

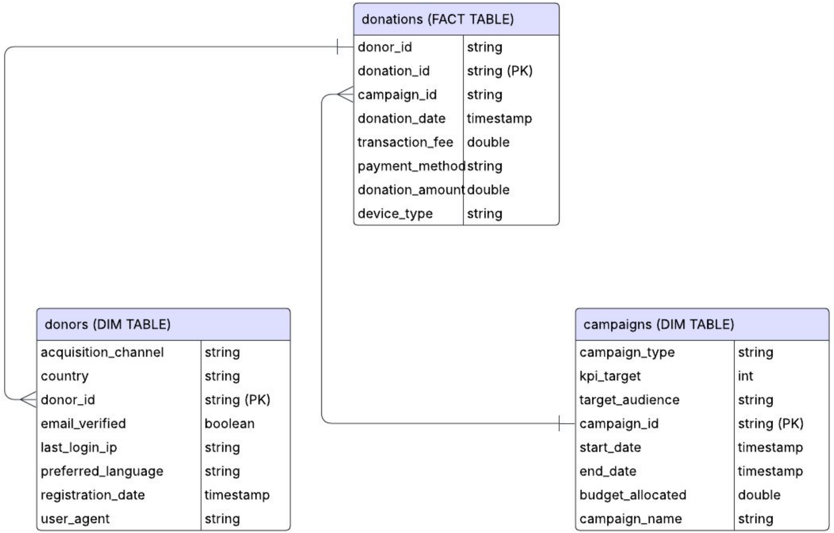
1. Building the Data Model

donations (Fact Table): one row per donation; measures = donation_amount, transaction_fee; links to donors and campaigns.

donations (Dim Table): donor attributes (channel, country, language, registration, tech info).

campaigns (Dim Table): campaign attributes (type, budget, KPI target, start/end dates).

Enables analysis of donations **by donor profile** and **by campaign**.



donations (FACT TABLE)	
donor_id	string
donation_id	string (PK)
campaign_id	string
donation_date	timestamp
transaction_fee	double
payment_method	string
donation_amount	double
device_type	string

donors (DIM TABLE)	
acquisition_channel	string
country	string
donor_id	string (PK)
email_verified	boolean
last_login_ip	string
preferred_language	string
registration_date	timestamp
user_agent	string

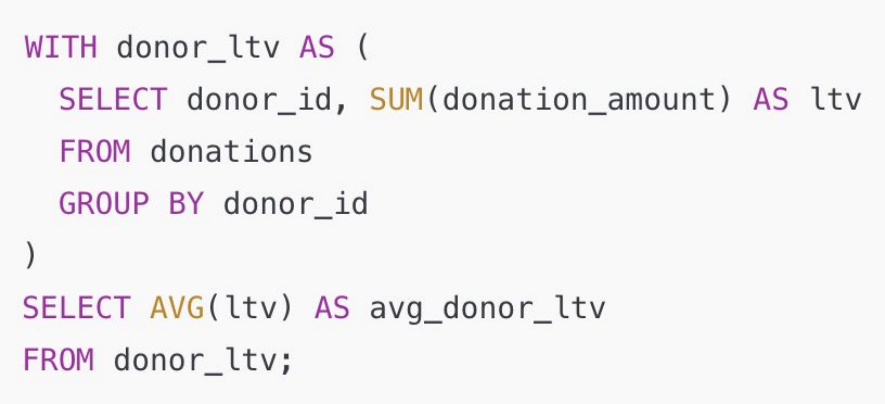
campaigns (DIM TABLE)	
campaign_type	string
kpi_target	int
target_audience	string
campaign_id	string (PK)
start_date	timestamp
end_date	timestamp
budget_allocated	double
campaign_name	string

Star schema

2. Lifetime Value (LTV)

The following query calculates each donor’s total donated amount, then averages across all donors.

Lifetime Value (LTV): 1482.77

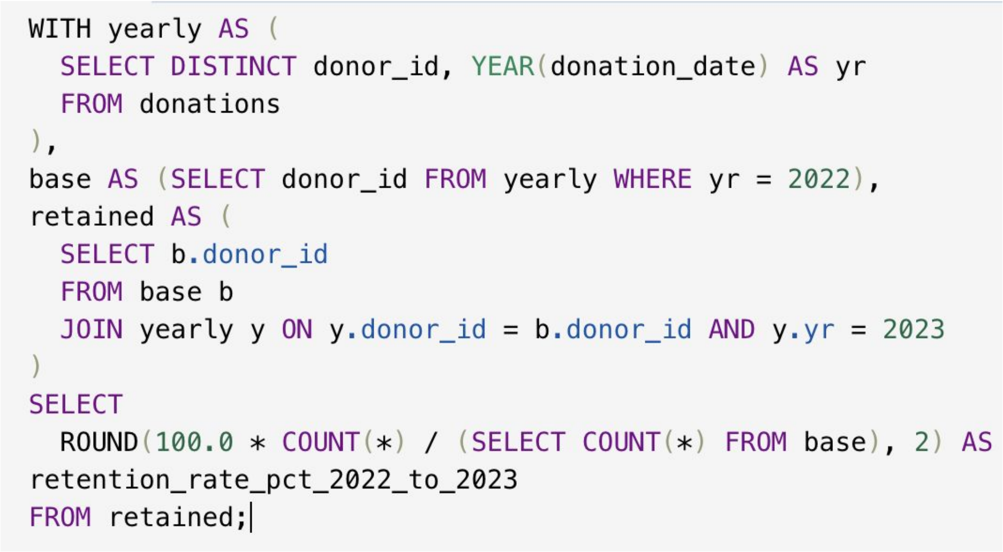


```sql
WITH donor_ltv AS (
  SELECT donor_id, SUM(donation_amount) AS ltv
  FROM donations
  GROUP BY donor_id
)
SELECT AVG(ltv) AS avg_donor_ltv
FROM donor_ltv;
```

2. Retention & Churn

yearly CTE: extracts each unique donor and the year(s) in which they made a donation.
base CTE: selects all donors who donated in **2022** (the cohort we're tracking).
retained CTE: keeps only those donors from the 2022 base who also appear in **2023**.
Final SELECT:

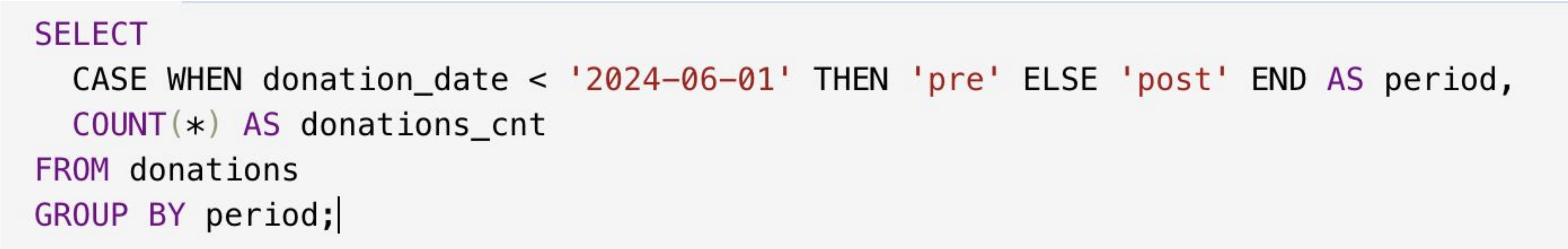
- Counts the number of retained donors (2022 → 2023).
- Divides it by the total number of 2022 donors.
- Multiplies by 100 and rounds to 2 decimals → giving the **retention rate %**.



```sql
WITH yearly AS (
  SELECT DISTINCT donor_id, YEAR(donation_date) AS yr
  FROM donations
),
base AS (SELECT donor_id FROM yearly WHERE yr = 2022),
retained AS (
  SELECT b.donor_id
  FROM base b
  JOIN yearly y ON y.donor_id = b.donor_id AND y.yr = 2023
)
SELECT
  ROUND(100.0 * COUNT(*) / (SELECT COUNT(*) FROM base), 2) AS
retention_rate_pct_2022_to_2023
FROM retained;
```

Retention rate: 94.08%

2. Causality Analysis



```
SELECT
  CASE WHEN donation_date < '2024-06-01' THEN 'pre' ELSE 'post' END AS period,
  COUNT(*) AS donations_cnt
FROM donations
GROUP BY period;
```

Not enough time sorry!

```
+------+-------------+
|period|donations_cnt|
+------+-------------+
|  post|         9091|
|   pre|         7327|
+------+-------------+
```

If I had more time

I would have used Databricks Free Edition.

I am not used to using Spark on a local computer.