Data Engineering Technical Assessment
Part I: Data Exploration and SQL Optimization

Mohamed Amin Soumih

May 27, 2025

Objective

Analyze user behavior to:

- Track Weekly Active Users (WAU)
- Measure Revenue per Product Category
- Optimize SQL queries and benchmark their performance

Data Preparation

The PostgreSQL database was populated using synthetic e-commerce data generated in Python. Three tables were used:

- `users(user_id, signup_date, country)`
- `products(product_id, category, price)`
- `events(user_id, event_type, product_id, timestamp)`

Query 1: Weekly Active Users

Goal: Count distinct users who interacted each week.
Optimization: An index was added on `events.timestamp` to improve grouping performance using `DATE_TRUNC`.
Benchmark: Executed via Python using `pandas` and `sqlalchemy`. Query returned in under 0.2 seconds.

Query 2: Revenue per Category

Goal: Calculate total revenue grouped by product category.
Optimization: An index was added on `events.event_type` to speed up filtering for purchases.
Benchmark: Query performance was measured and returned results in under 0.2 seconds.

Conclusion

Indexes significantly improved performance of both queries. Clean separation between data generation, querying, optimization, and benchmarking was maintained using layered Python scripts.