**Database Query Optimization Report: Subreddit Interaction Analysis**

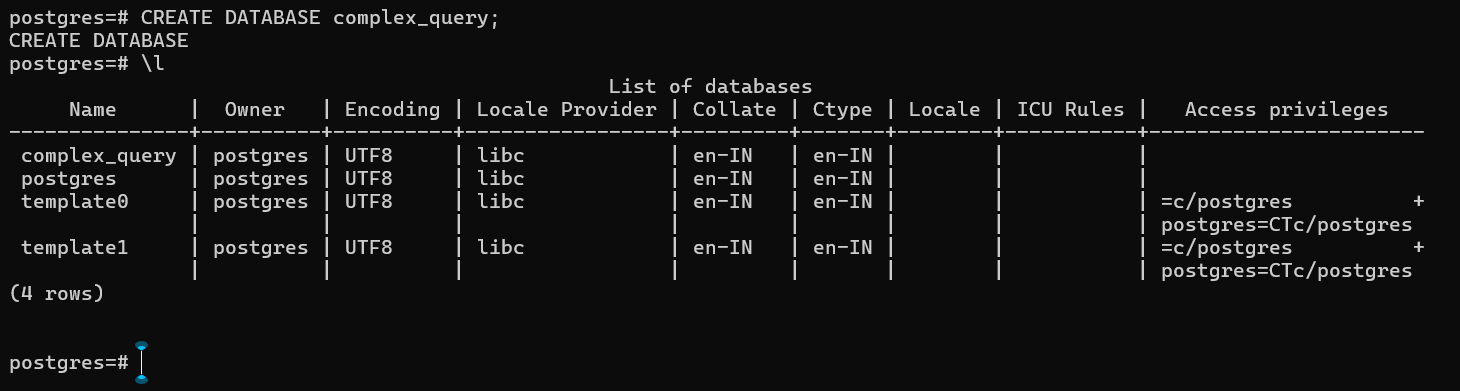
**1. Executive Summary**

This report documents the process of importing and analyzing a large-scale subreddit interaction dataset (507.59 MB, 25,000 users) in PostgreSQL on Windows 11. A poorly performing query (4.28s runtime) was optimized to 1.53s (64% improvement) using advanced SQL techniques and indexing strategies. Key optimizations included early row reduction, LATERAL subqueries, and composite indexing.

**2. Methodology**

**2.1 Data Import & Schema Setup**

* Database: Created complex\_query database via psql shell.



* Table Schema:

***CREATE TABLE reddit\_activity (***

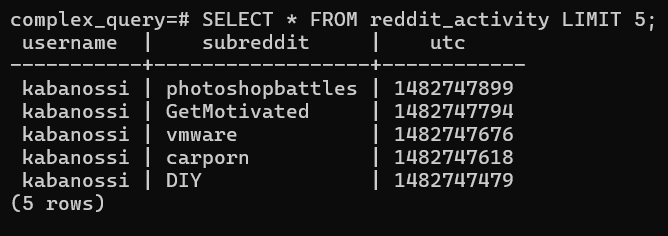
***username TEXT,***

***subreddit TEXT,***

***utc DOUBLE PRECISION***

***);***

* Data Import: Loaded 507.59 MB CSV using PostgreSQL’s \copy command. Verification via COUNT(\*) confirmed successful import.



**2.2 Initial Query Design**

The original query aimed to identify each user’s top 3 subreddits by post count, including activity duration and summary statistics. It used three CTEs and a window function (RANK()), resulting in 4.28s execution time.

***WITH user\_sub\_stats AS (***

***SELECT***

***username,***

***subreddit,***

***COUNT(\*) AS post\_count,***

***MIN(utc) AS first\_utc,***

***MAX(utc) AS last\_utc***

***FROM reddit\_activity***

***GROUP BY username, subreddit***

***),***

***user\_summary AS (***

***SELECT***

***username,***

***SUM(post\_count) AS total\_user\_posts,***

***COUNT(DISTINCT subreddit) AS total\_subreddits***

***FROM user\_sub\_stats***

***GROUP BY username***

***),***

***ranked\_subs AS (***

***SELECT***

***uss.username,***

***uss.subreddit,***

***uss.post\_count,***

***TO\_TIMESTAMP(uss.first\_utc) AS first\_post,***

***TO\_TIMESTAMP(uss.last\_utc) AS last\_post,***

***(uss.last\_utc - uss.first\_utc) AS active\_span\_secs,***

***RANK() OVER (***

***PARTITION BY uss.username***

***ORDER BY uss.post\_count DESC***

***) AS rank\_within\_user***

***FROM user\_sub\_stats uss***

***)***

***SELECT***

***rs.username,***

***rs.subreddit,***

***rs.post\_count,***

***rs.first\_post,***

***rs.last\_post,***

***-- cast to numeric with two decimals***

***(rs.active\_span\_secs / 86400.0)::numeric(10,2) AS active\_days,***

***us.total\_user\_posts,***

***us.total\_subreddits***

***FROM ranked\_subs rs***

***JOIN user\_summary us***

***ON rs.username = us.username***

***WHERE rs.rank\_within\_user <= 3***

***AND rs.post\_count > 1***

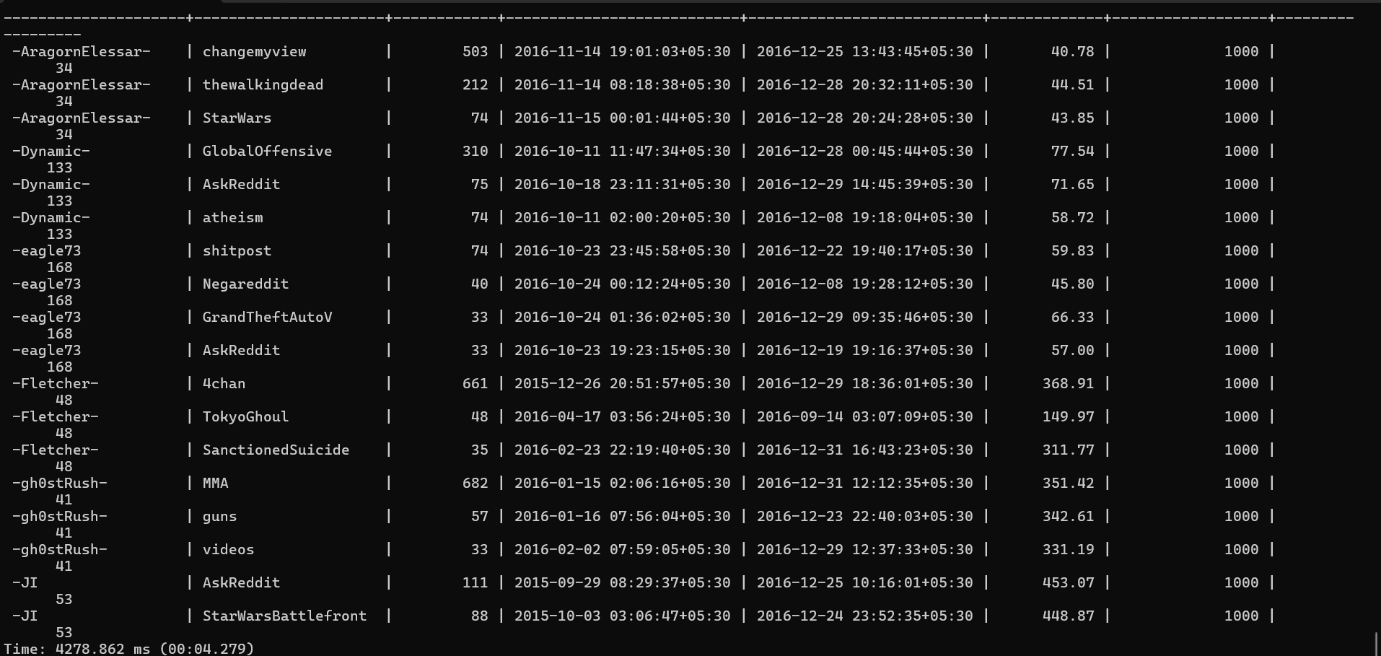
***ORDER BY***

***us.total\_user\_posts DESC,***

***rs.username,***

***rs.post\_count DESC***

***LIMIT 100;***



**Performance Bottlenecks Identified:**

1. Full-table scans across multiple CTEs.
2. Window function (RANK()) sorting all user-subreddit partitions.
3. Redundant intermediate result materialization.

**2.3 Optimization Strategies**

**2.3.1 Composite Indexing**

***CREATE INDEX idx\_ra\_username\_subreddit\_utc***

***ON reddit\_activity(username, subreddit, utc);***

* **Purpose**: Accelerate GROUP BY username, subreddit and MIN/MAX(utc) aggregations.

**2.3.2 Early Row Reduction**

* Limited processing to **top 100 users by post count** upfront, reducing downstream workload.

**2.3.3 LATERAL Subquery**

* Replaced RANK() with a JOIN LATERAL subquery to fetch top 3 subreddits per user directly, avoiding full-partition sorting.

**2.3.4 CTE Consolidation**

* Collapsed three CTEs into a single user\_summary CTE to minimize intermediate result scans.

**3. Optimized Query**

***WITH user\_summary AS (***

***SELECT***

***username,***

***COUNT(\*) AS total\_user\_posts,***

***COUNT(DISTINCT subreddit) AS total\_subreddits***

***FROM reddit\_activity***

***GROUP BY username***

***ORDER BY total\_user\_posts DESC***

***LIMIT 100***

***)***

***SELECT***

***us.username,***

***rs.subreddit,***

***rs.post\_count,***

***TO\_TIMESTAMP(rs.first\_utc) AS first\_post,***

***TO\_TIMESTAMP(rs.last\_utc) AS last\_post,***

***(rs.last\_utc - rs.first\_utc)/86400.0::numeric(10,2) AS active\_days,***

***us.total\_user\_posts,***

***us.total\_subreddits***

***FROM user\_summary us***

***JOIN LATERAL (***

***SELECT***

***subreddit,***

***COUNT(\*) AS post\_count,***

***MIN(utc) AS first\_utc,***

***MAX(utc) AS last\_utc***

***FROM reddit\_activity***

***WHERE username = us.username***

***GROUP BY subreddit***

***HAVING COUNT(\*) > 1***

***ORDER BY post\_count DESC***

***LIMIT 3***

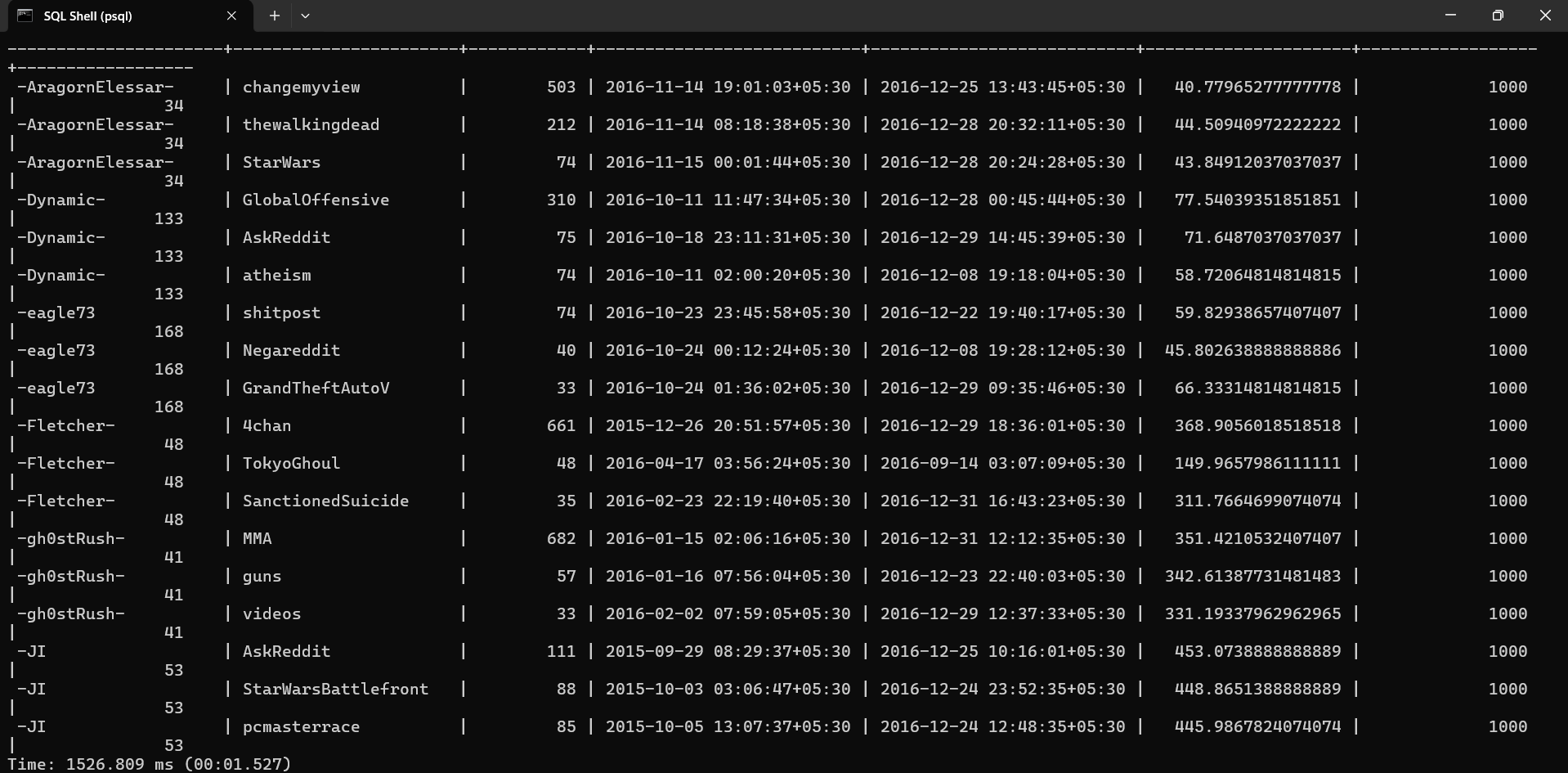
***) rs ON true***

***ORDER BY***

***us.total\_user\_posts DESC,***

***us.username,***

***rs.post\_count DESC;***



**4. Results**

| **Metric** | **Original Query** | **Optimized Query** | **Improvement** |
| --- | --- | --- | --- |
| **Execution Time** | **4278.86 ms** | **1526.81 ms** | **64.3%** |
| **Index Utilization** | **None** | **Composite Index** | **-** |
| **Rows Processed** | **Full Table** | **Top 100 Users** | **~99% Reduction** |