DP-accurate DAU/MAU Counter Under Deletions

Dev Sanghvi Lazeen Manasia

October 9, 2025
580-Probability Algorithms and Data Structures

Abstract

We present a proof-of-concept system that maintains differentially private estimates of Daily Active Users (DAU) and 30-day Monthly Active Users (MAU) for high-velocity turnstile event streams with retrospective deletion guarantees. The platform ingests JSON events in real time, hashes user identifiers via keyed HMAC with scheduled salt rotation, and materializes per-day sketches backed by an extensible interface. We ship both an exact set-based sketch for correctness testing and approximate Theta and HyperLogLog++ variants for scalability, allowing operators to swap implementations by setting {{`SKETCH_IMPL`}}. Erasure requests propagate through an SQLite-backed ledger, triggering dirty-day rebuilds so that deletions retrospectively excise a user from all future DAU/MAU releases. Differential privacy is enforced through Laplace or Gaussian mechanisms tuned by a documented flippancy bound {{`W_BOUND`}}, while a budget accountant applies naïve composition and rejects requests exceeding monthly caps.

The public interface is a FastAPI service with REST endpoints for event ingestion, DAU/MAU queries, Prometheus metrics, and health checks, complemented by a Typer-based CLI for local workflows such as dataset ingestion, budget resets, and salt rotation drills. Evaluation scripts generate synthetic and adversarial workloads, benchmark sketch accuracy versus privacy budgets, and export reproducible plots; an accompanying Jupyter notebook ties the results into narratives for stakeholders. Continuous integration hardens quality through placeholder-ledger enforcement, style and type checks, unit tests, and coverage reporting, while Docker artifacts and Make targets streamline deployment.

Beyond correctness, the documentation suite—including README, HANDOFF, and AGENTS guides—captures operational runbooks, placeholder governance, extension hooks for gRPC, and migration notes for Postgres and Kafka. Collectively, the architecture balances academic rigor in probabilistic data structures with practical engineering constraints, offering an end-to-end baseline for privacy-preserving analytics teams to extend toward production-grade continual release systems.