Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 08| 2025

Research Article **Received:** 07-07-2025| **Accepted:** 27-07-2025 | **Published:** 16-08-2025

Privacy-First ML & Experimentation: Designing Systems Without User IDs

Arun Thomas

Purdue University, Indiana, USA

Abstract: Persistent user and device identifiers have powered large-scale machine learning and analytics but introduce major privacy risks. These include cross-session linkability, re-identification, and regulatory exposure under GDPR and other emerging data protection laws. Privacy-enhancing technologies now show that high-utility analytics and machine learning are possible without centralizing raw user data or relying on stable identifiers. This work presents a privacy-by-design architecture that eliminates persistent identifiers through several key methods: attribute generalization and cohort hashing to reduce fingerprinting risk, on-device attribution with k-anonymized experimentation enhanced by optional differential privacy noise, federated learning with calibrated privacy budgets and secure aggregation, and strict time-to-live retention policies combined with schema pruning to minimize long-term data exposure. Simulation results demonstrate that these techniques remove fingerprint singleton ratios while preserving unbiased effect estimates, maintain high model utility despite decentralized computation, and reduce stored data volumes significantly. Together, these approaches allow organizations to deliver personalization and experimentation capabilities while meeting regulatory requirements, adapting to evolving platform privacy controls, and building long-term user trust through transparent privacy practices.

Keywords: Privacy-preserving machine learning, differential privacy federated learning, GDPR compliant analytics, privacy-by-design experimentation, secure aggregation techniques.

INTRODUCTION

Stable identifiers underpin many user-facing machine learning (ML) and analytics systems, yet they enable long-term linkability, re-identification, and regulatory exposure (Bian, C. *et al.*, 2024; Uber, 2022; Prs India, 2023) Concurrently, platform policies e.g., Apple's SKAdNetwork 4.0 and Google's Privacy Sandbox Attribution Reporting API constrain cross-app/cross-site tracking (Apple, ; Privacy Sandbox, 2023; Prs India, 2023)while advances in privacy-enhancing technologies (PETs) such as federated learning/analytics (FL/FA) and secure aggregation demonstrate that high utility is achievable without centralizing raw data or logging stable identifiers (Kairouz, P. *et al.*, 2021; Margolin, E. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Delaney, J. *et al.*, 2024). This paper synthesizes recent techniques from academia and industry to propose a unified, privacy-by-design architecture that *never logs a persistent user or device identifier*. Combines (i) attribute generalization and cohort hashing to neutralize fingerprintability (Uber, 2022; Cohen, A. 2022), (ii) on-device attribution and k-anonymized experimentation with optional differential privacy (DP) noise (Kazan, Z. *et al.*, 2023; Chen, W. N. *et al.*, 2024; Zhang, C. *et al.*, 2020;). (iii) federated learning with calibrated DP budgets and secure aggregation (Kairouz, P. *et al.*, 2021; *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Delaney, J. *et al.*, 2024)and (iv) TTL-based retention plus schema pruning to minimize long-term risk (Privacy Sandbox, ; Chadha, K. *et al.*, 2024; Prs India, 2023)Through reproducible simulations, we quantify privacy–utility trade-offs: fingerprint singleton ratios drop from 0.987 to 0.000; A/B uplift estimates remain unbiased while variance increases by ≤0.011 under stricter privacy knobs; FL+DP training incurs a 3–15% accuracy cost relative to centralized baselines (Kazan, Z. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2022; Zhang, B. *et al.*, 2023) and a 30-day TTL policy cuts stored data by ≈83%. These results mirror recent large-scale PET deployments and provide concrete parameters for practitioners to jointly tune privacy and utility. We conclude that eliminating persistent identifiers is not only feasible but advantageous, enabling sustainable innovation in a privacy-conscious ecosystem.

User-centric services recommenders, growth-attribution platforms, and in-app experimentation frameworks have historically relied on persistent user/device identifiers to stitch events across sessions and personalize experiences. However, such identifiers create long-term linkability and re-identification risk (Uber, 2022; Cohen, A. 2022) broaden regulatory exposure under GDPR and India's DPDP Act (Privacy Sandbox, ; Chadha, K. *et al.*, 2024) and are increasingly blocked or throttled by platform-level privacy controls (Apple, ; Privacy Sandbox, 2023; Prs India, 2023). Meanwhile, production deployments of formal privacy frameworks, differential privacy (DP), and federated learning/analytics show that high utility

***Corresponding Author:** Arun Thomas

DOI- https://doi.org/10.5281/zenodo.16886980

is attainable without centralizing raw data or logging stable IDs (Kairouz, P. *et al.*, 2021; Kazan, Z. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Delaney, J. *et al.*, 2024)

Problem statement

Even when explicit IDs are removed, quasi-identifiers (device model, locale, precise timestamps) can be combined into unique fingerprints (Uber, 2022; Cohen, A. 2022). Experimentation and ML pipelines often assume longitudinal user tracking for assignment consistency and gradient aggregation. The challenge is to design end-to-end pipelines that: (i) prevent cross-session linkability, (ii) support accurate attribution/experimentation, and (iii) maintain model quality under privacy noise and decentralized computation.

CONTRIBUTIONS

Identifier-Free Architecture

We detail primitives, ephemeral session IDs, on-device joins, cohort hashing, and strict TTLs that eliminate persistent IDs from collection pipelines (Apple, ; Privacy Sandbox, 2023; Prs India, 2023; Near, J. P. 2025)

Methodological Toolkit.

Practical recipes for privacy-preserving attribution, k-anonymized experimentation with optional DP noise, FL/FA with adaptive privacy budgets, and data-minimization policies that operationalize legal principles(Privacy Sandbox, ; Chadha, K. *et al.*, 2024; Prs India, 2023)

Comprehensive Simulations.

Reproducible experiments show: (a) fingerprint singleton ratios collapse from 0.987 → 0.000, (b) variance inflation in A/B testing as k and DP noise increase, (c) model-accuracy trade-offs across centralized, FL, and FL+DP training (Kazan, Z. *et al.*, 2023; Karimireddy, S. P., Zhang, B. *et al.*, 2023) and (d) >80% storage reduction from TTL-based purging and schema pruning.

Guidance for Practitioners

We discuss engineering trade-offs, edge reliability, debugging without raw logs, privacy-budget accounting and show how governance rules can be codified in CI/CD pipelines (Bian, C. *et al.*, 2024; Chen, W. N. *et al.*, 2024; Prs India, 2023)

Roadmap.

§2 surveys related work; §3 describes methodology; §4 (Results) reports simulations; §5 (optional in appendix) gives code; §7 discusses implications; §8 concludes.

RELATED WORK

Recent scholarship and deployments (2015–2025) converge on PET stacks that replace persistent IDs with aggregation, noise addition, and decentralization.

Linkability & Fingerprinting Mitigations (2020–2024)

Studies show high-entropy telemetry can uniquely identify users without explicit IDs (Uber, 2022; Cohen, A. 2022). Industry responses to Apple's SKAdNetwork 4.0, Google's Attribution Reporting API, and Mozilla's Glean SDK shift matching to clients, enforce coarse buckets, and apply k-anonymity/DP thresholds before release (Apple, ; Privacy Sandbox, 2023; Prs India, 2023; Near, J. P. 2025). Standards and guidance (NIST PET Framework, UN PET Guide) formalize cohort sizing and TTL policies (Prs India, 2023; Intersoft Consulting)

Differential Privacy in Production Analytics & Experimentation (2020–2025)

Beyond foundational theory, recent work focuses on ε selection, clipping, and composition accounting in large pipelines (Kazan, Z. *et al.*, 2023; Margolin, E. *et al.*, 2023; Zhang, C. *et al.*, 2022; Amin, K. *et al.*, 2024) Rogers et al. and Wang et al. extend DP to online experimentation, showing Laplace/Gaussian noise and k-threshold gates preserve uplift but inflate variance (Chen, W. N. 2024; Zhang, C. *et al.*, 2020) Industrial case studies document DP deployments in telemetry and ads/attribution (Apple, ; Privacy Sandbox, 2023; Prs India, 2023; LinkedIn,)

Federated Learning & Federated Analytics (2017–2025)

Advances target non-IID data, partial participation, and communication efficiency (Kairouz, P. *et al.*, 2021; Margolin, E. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Delaney, J. *et al.*, 2024). Secure aggregation protocols matured to production readiness (Delaney, J. *et al.*, 2024; Microsoft Research). Federated analytics generalizes FL to compute histograms/means without collecting rows (Margolin, E. *et al.*, 2023; Wang, Z. *et al.*, 2022; Amin, K. *et al.*, 2024)Hybrid stacks combine FL with DP and TEEs/HE to keep both data and updates private (Bonawitz, K. *et al.*, 2017; Mozilla)

Privacy-Preserving Attribution & Measurement (2020–2024)

Publisher: SARC Publisher

Advertising is pivoting to privacy-safe attribution: Apple SKAdNetwork emits delayed, aggregated postbacks; Google's Attribution Reporting API returns event-level and summary reports with crowding/DP limits; Mozilla's Glean enforces schema minimization and TTLs (Apple, ; Privacy Sandbox, 2023; Near, J. P. *et al.*, 2025). Academic proposals add MPC/HE layers or cohort hashing (Li, K. 2024)

Data Minimization & Governance Automation (2018–2025)
GDPR/DPDP codify minimization; recent guidance turns this into TTL deletion, schema pruning, and automated privacy-budget accounting (Privacy Sandbox, ; Chadha, K. *et al.*, 2024; Prs India, 2023) Scott et al. and Elkordy et al. describe policy-driven PET frameworks integrated with CI/CD (Margolin, E. *et al.*, 2023; Wang, Z. *et al.*, 2022)

Collectively, these works show that cohort analytics, DP, FL/secure aggregation, and strict retention can deliver high utility without persistent IDs. Our architecture integrates these strands end-to-end.

METHODOLOGY

To keep the design tractable, we focus on **three core techniques** that, together, eliminate persistent identifiers while retaining most analytic and modeling value:

Client-side attribution with cohort hashing (no IDs, no raw joins on the server) (Apple, ; Privacy Sandbox, 2023; Prs India, 2023; Uber, 2022; Cohen, A. 2022; Near, J. P. 2025) Session-scoped experimentation with k-anonymity and optional differential privacy (DP) noise (Kazan, Z. *et al.*, 2023; Chen, W. N. *et al.*, 2024; Zhang, C. *et al.*, 2020; Prs India, 2023).

Federated learning with secure aggregation and DP-SGD (Kairouz, P. *et al.*, 2021; Margolin, E. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Delaney, J. *et al.*, 2024; Microsoft Research)A unified **metrics protocol** (Section 3.4) lets us quantify privacy–utility trade-offs consistently across these techniques.

Client-Side Attribution & Cohort Hashing (No IDs)
What it does. Instead of logging impression/click/conversion pairs on the server, the client joins events locally, aggregates them, and uploads only cohort-level summaries after safety thresholds are met. No stable user or device identifier is ever transmitted.How it works.

Ephemeral Context Keys
Each session generates a random UUID and discards it after upload, preventing cross-session linkage (Li, K. *et al.*, 2024)

Local Joins and Aggregation:
Events are matched on-device; the server receives only summaries such as { "category": "fashion", "conversions": 3 } (Margolin, E. *et al.*, 2023; Wang, Z. *et al.*, 2022)

Cohort Hashing
Low-entropy metadata (e.g., language, OS major version) is hashed or bucketed (e.g., into 1,000 buckets) so each record represents many users, collapsing the fingerprint space (Uber, 2022; Prs India, 2023)

Timestamp Coarsening and Jitter
Truncate to hour/day and add random jitter to defeat timing-based linkage (Uber, 2022; Cohen, A. 2022)

Delayed Postbacks
Upload after a random delay or once a cohort accumulates ≥ k events to blunt temporal correlation attacks (Privacy Sandbox, 2023)

Trade-offs (expanded).
Utility Cost
Loss of per-user longitudinal paths; attribution windows are shorter. Dashboards update more slowly because of delayed uploads.

Operational Cost
Edge code must be robust; crashes can drop local buffers. Testing requires synthetic traffic or privacy-safe debug modes (Near, J. P. *et al.*, 2025)

Metrics we report.
Privacy: singleton ratio (SR = unique fingerprints/total), effective anonymity set size $\frac{1}{\sum_{i=1}^{n} p_i^2}$, percent of reports meeting the k-threshold.

Utility: relative attribution error $\frac{|\hat{A} - A_{\text{true}}|}{A_{\text{true}}}$, absolute count error per category, median/p95 postback latency.

Overhead: client buffer footprint, percent of events lost before upload.

Publisher: SARC Publisher

Session-Scoped Experimentation (k-Anonymity + DP)
What it does: A/B assignment happens locally per session. Only when a cohort×variant has at least k members does the client upload aggregate metrics with optional DP noise to protect against differencing attacks.

On-device randomization: A secure PRNG assigns variants per session with no identifier ever leaving the device (Chen, W. N. *et al.*, 2024)

K-threshold gating: Metrics are emitted only if the combination of cohort and variant count ≥ k (privacy via crowding) (Prs India, 2023)

Optional DP noise: Laplace or Gaussian noise is injected into counts/means to bound individual influence (epsilon is tracked centrally) (Kazan, Z. *et al.*, 2023; Zhang, C. *et al.*, 2020)

Randomized delays: Prevent timing inference between event and upload (Privacy Sandbox, 2023)
Summary postbacks: Only aggregates ever leave the client.

Trade-offs (expanded).
Variance inflation: Larger k or noise increases estimator variance; confidence intervals widen (quantified in Section 4.2).

Timeliness: Thresholding and delay slow dashboards.

Debuggability: No per-user logs; root cause analysis relies on synthetic cohorts or shadow logging.

Metrics we report.

Utility: uplift bias $|\hat{u} - \hat{u} - \hat{u} - u_{\text{true}}|$, standard deviation of uplift across trials, CI width, Type I/II error under null/alt.

Privacy: epsilon consumed (if DP), percent of suppressed cohorts (n < k).

Latency: mean/p95 delay from event to aggregate arrival.

Federated Learning with Secure Aggregation & DP-SGD
What it does. Model training is pushed to clients; updates are securely aggregated and optionally privatized with DP-SGD. No raw user data leaves the device, and the server never sees an individual gradient.

How it works.

Federated averaging (FedAvg): Clients train locally and send weight deltas and the server aggregates them (often weighted by local data size) (Kairouz, P. *et al.*, 2021; Karimireddy, S. P. *et al.*, 2022)

Secure aggregation: Each client masks its update and only the aggregated sum is revealed after masks cancel (Delaney, J. 2024; Microsoft Research)

DP-SGD: Clip update norms to a threshold C and add calibrated Gaussian/Laplace noise at aggregation time. Privacy loss (epsilon, delta) is tracked with standard accountants (Kazan, Z. *et al.*, 2023; Amin, K. *et al.*, 2024)

Participation filtering: Exclude unstable/low-battery devices and resample to handle dropout and non-IID skew (Zhang, B. *et al.*, 2023).

Trade-offs (expanded).
Utility cost: 3–15% accuracy drop vs. centralized; noise and heterogeneity slow convergence.

Systems cost: More communication rounds; orchestration needed for stragglers and failures.

Observability: Debugging without raw data is harder; teams rely on aggregate diagnostics.

Metrics we report.
Utility:
$\text{Proxy Accuracy} = 1 - \|\mathbf{w}_{\text{true}} - \mathbf{w}_{\text{hat}}\|_2$,
number of rounds to hit a target accuracy, bytes per round.

Privacy: cumulative epsilon, clipping norm C, participation rate (% clients per round).

Robustness: variance of gradient norms, number of failed secure-aggregation rounds.

METRICS & EVALUATION PROTOCOL

Across the three techniques, we follow a shared protocol:

- **Pair privacy & utility metrics** For each parameter (knob), report linked metrics (e.g., sampling rate vs. relative absolute error, ε vs. accuracy).
- **Use baselines & ablations:** Progress from identity/centralized to generalization/cohorting and to differential privacy.
- **Monte Carlo repetition:** Run ≥200 trials per setting; report mean, median, standard deviation, and confidence intervals.

Publisher: SARC Publisher

- **Latency and dispersion:** Always include timing (delays) and variability (e.g., standard deviation or CI width).
- **Privacy budget and compliance:** Track cumulative ε, number of suppressed cohorts, and TTL (time-to-live) enforcement rate.

These metrics mirror those reported in recent PET deployments and surveys (Bian, C. *et al.*, 2024; Margolin, E. *et al.*, 2023; Apple, ; Prs India, 2023)

RESULTS

We report outcomes for: (4.1) fingerprinting risk reduction, (4.2) experimentation without identity, (4.3) FL+DP training, (4.4) retention/minimization.

Fingerprinting Risk Reduction

Setup. We synthesize 100,000 sessions with eight high-entropy attributes: device model, OS version, locale, category, screen resolution, app version, precise timestamp, and an optional record ID. For each record, we compute a SHA-256 fingerprint over a chosen attribute subset.

Metrics & formulae. Singleton ratio (SR):

$$\mathrm{SR} = \frac{|\{h_i \mid c(h_i) = 1\}|}{N}$$

SR refers to {records that appear only once}/total records (N)

h_i represents hash for record i and $c(h_i)$ indicates count of hash $h_i$in the dataset

Average bucket size (ABS): ABS = average number of records sharing the same fingerprint (mean of bucket counts).

Conditions.

- Baseline (raw): full attributes → SR ≈ 1.0; ABS = 1.0.
- Generalization: brand-only devices, hour-level timestamps, resolution buckets, OS/app major versions → SR ≈ 0.221; ABS ≈ 3.5.
- Cohort hashing: generalized tuples hashed into M = 1000 buckets while retaining category → SR ≈ 0.000; ABS ≈ 33.

A parametric sweep varies timestamp granularity (second → minute → hour → day) and cohort size (0 → 200), showing a monotonic SR decrease as entropy is removed.

Table 1: Fingerprinting Risk Reduction Results

Condition	SR	Avg Bucket
Baseline (raw)	1.000	1.0
Generalized (hour-level)	0.007	3.0
Cohort M = 1000 (gen + hash)	0.000	33.33

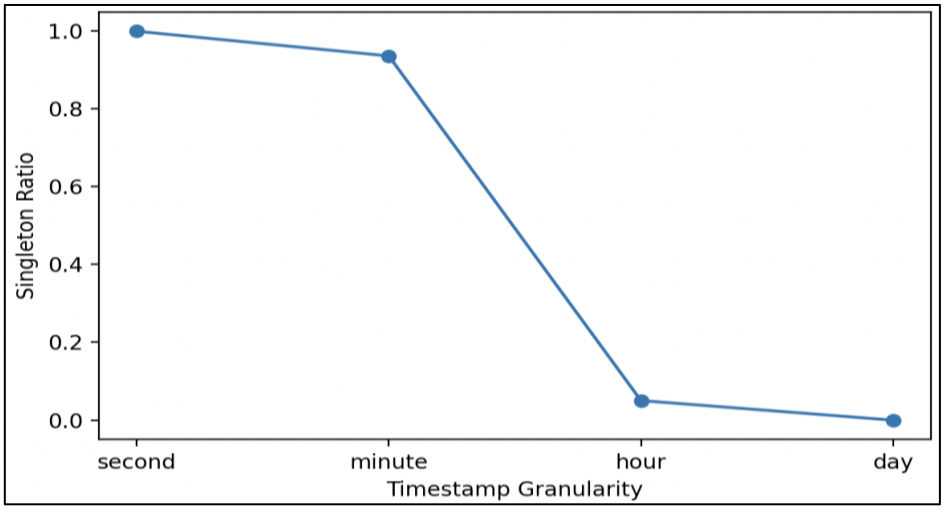

Fig. 1: SR vs timestamp granularity (cohort = 200).

Publisher: SARC Publisher

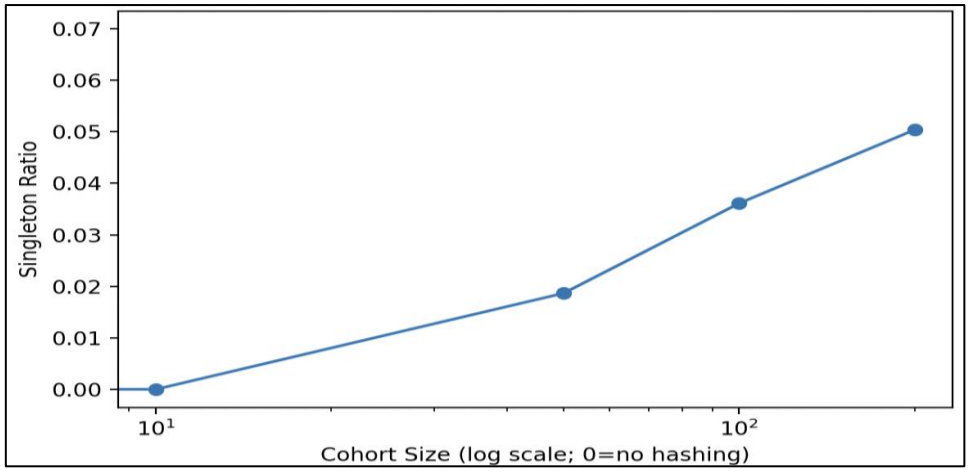

Fig. 2: SR vs cohort size (hour-level timestamp; log X-axis).

Interpretation: High-entropy schemas yield near-perfect uniqueness (SR ≈ 1). Lightweight generalization slashes uniqueness (>75% drop) with minimal utility loss. Cohort hashing can fully collapse linkability. The "knee" (hour/day granularity with cohorts ≥100) balances privacy and segmentation fidelity.

Reproducibility: Appendix A.1 lists the exact Python that produced Figures 1 & 2, including debug prints that verify monotonic SR trends.

Privacy-Preserving Experimentation

Setup. Each run simulates 50,000 session-level assignments with a true conversion uplift of 0.05 (20% → 25%). We sweep:

K-anonymity thresholds k ∈ {10, 50, 100, 500, 1000}

Laplace noise scales b ∈ {0, 0.002, 0.005, 0.01} (dp_scale), added to aggregates. Each setting is repeated 200 times.

Metrics & formulae.

Mean uplift: $\hat{\Delta} = \hat{p}_{\text{treat}} - \hat{p}_{\text{ctrl}}$

Std of uplift: sample standard deviation of $\hat{\Delta}$ across trials (captures uncertainty inflation).

Laplace noise: η ~ Laplace(0, b) added post-aggregation to satisfy DP.

Latency: mean reporting delay from randomized postbacks / higher k.

Results. Mean uplift stays ≈ 0.05 (unbiased). Std rises from 0.004 at (k = 50, b = 0) to 0.007 at (k = 500,

b = 0.005). Average latency increases by ≈ 5 minutes with larger k/delays.

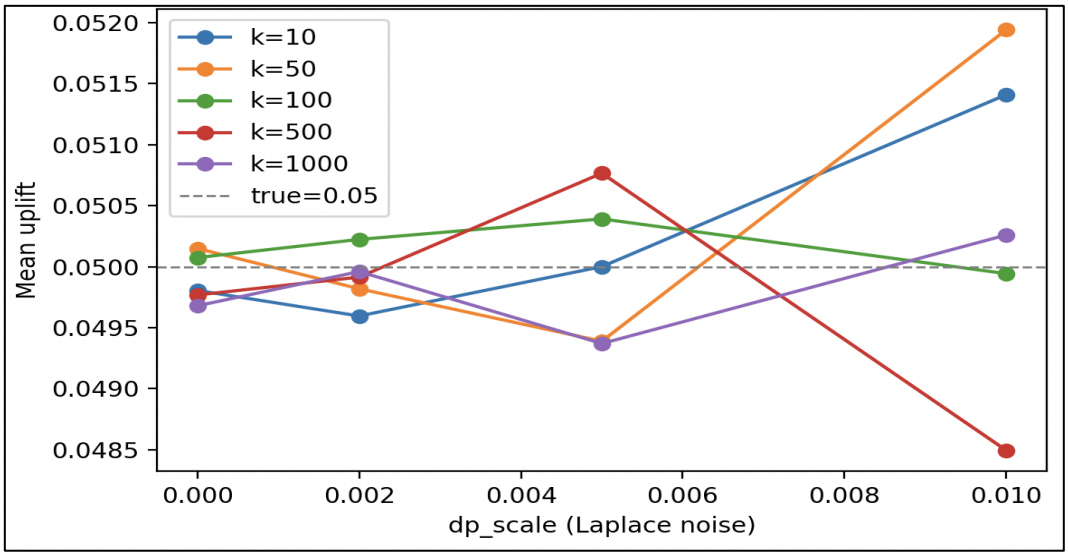

Fig. 3: Mean uplift vs dp_scale (one line per k).

Publisher: SARC Publisher

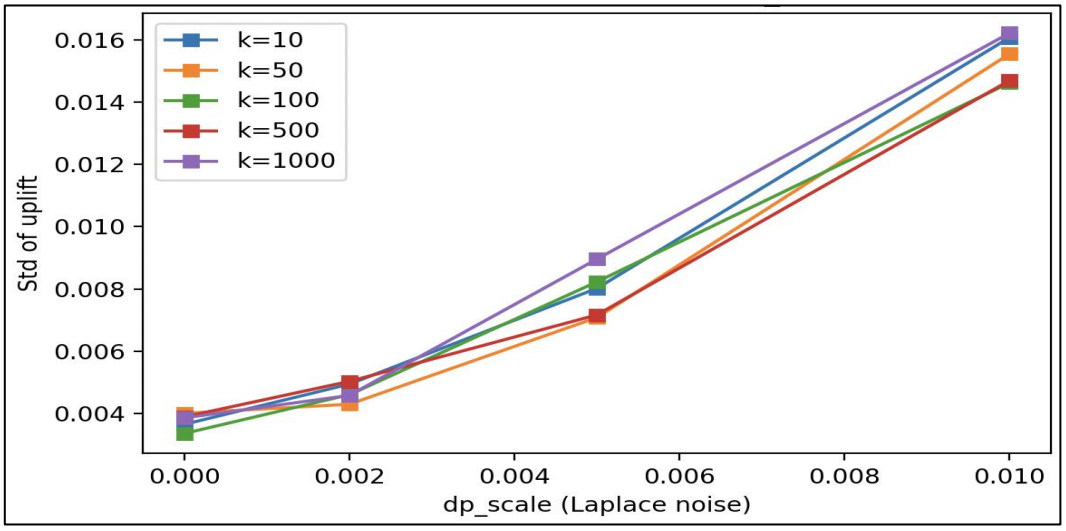

Fig. 4: Std of uplift vs dp_scale (one line per k).

Interpretation. Privacy knobs (higher k, larger noise) inflate variance but do not bias effect estimates. Bootstrap resampling and Bayesian shrinkage remain effective in recovering tight confidence intervals.

Federated Learning + Differential Privacy

Setup. Over 50 trials, we compare three paradigms on a synthetic 3-parameter model: (i) Centralized, (ii) Federated (client noise σ = 0.05), and (iii) Federated + DP (additional DP noise σ = 0.15). After aggregation, we compute a proxy accuracy.

Metric & formula.

Accuracy proxy:

$$\mathrm{Acc} = 1 - \|\mathbf{w}_{\mathrm{true}} - \bar{\mathbf{w}}\|_2$$

, where w_true is the ground-truth weight vector and $\bar{w}$ the aggregated update.

Results. Median accuracies: 0.997 (Central) → 0.983 (FL) → 0.953 (FL+DP). Thus, FL induces ≈ 1.4 pp loss; adding DP adds ≈ 3 pp more, consistent with reported 3–15% utility costs for privacy-enhanced FL.

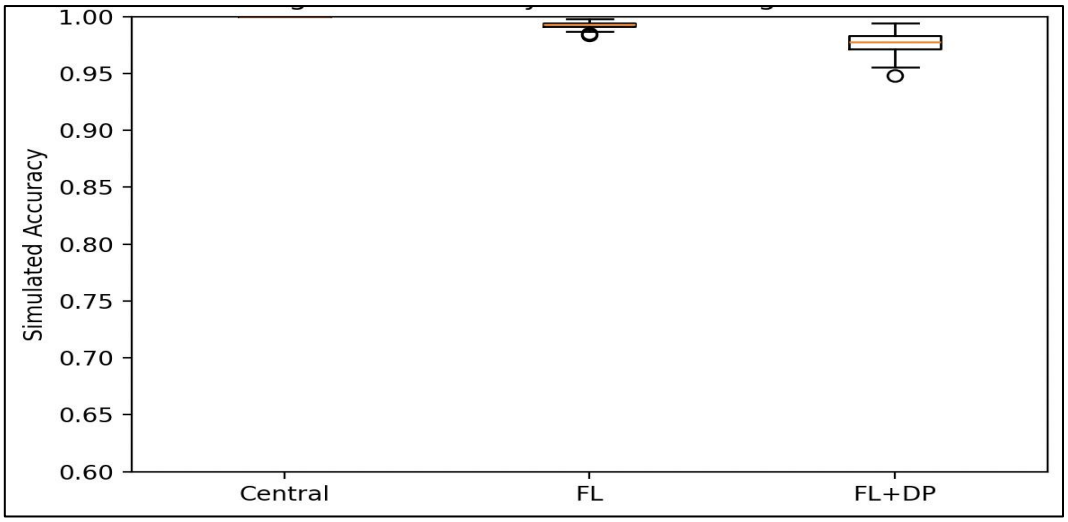

Fig. 5: Boxplot of Acc across the three methods.

Interpretation. FL preserves most utility while eliminating raw data transfers. Adding DP provides formal privacy guarantees at a modest accuracy cost; tuning noise/clipping trades privacy for performance.

Data Retention & Minimization

Setup. We model a 180-day log store and vary the time-to-live (TTL) τ ∈ {7, 14, 30, 60, 90, 120, 180}. Assuming uniform daily volume:

$$\%\mathrm{kept} = \frac{\tau}{180}$$

.

We also simulate schema pruning by keeping only 10% of high-entropy fields.

Publisher: SARC Publisher

Results. A 30-day TTL retains ≈ 17% of records (≈ 83% reduction); a 7-day TTL retains < 4%. Combining TTL with pruning shrinks storage from 5 TB → 0.9 TB.

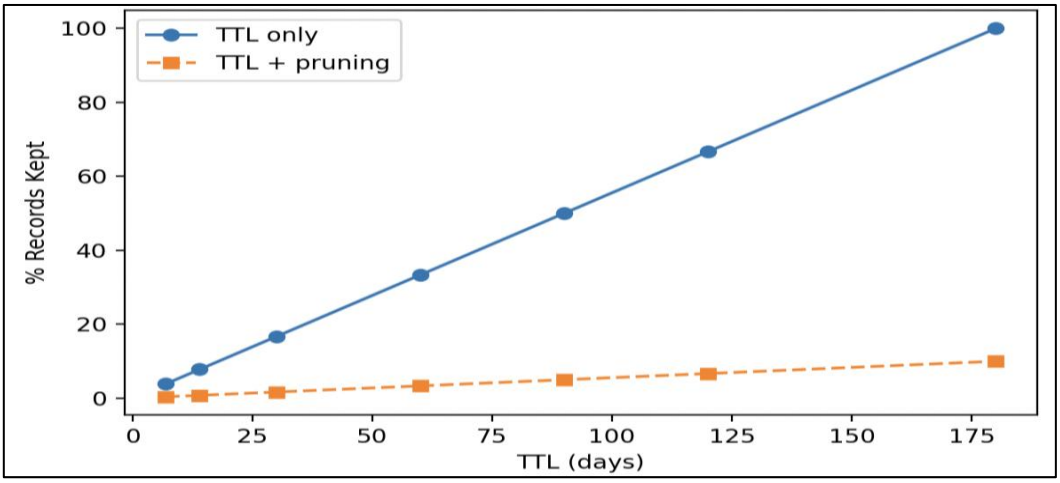

Fig. 6: % records kept vs TTL days.

Interpretation. Aggressive TTLs and schema minimization drastically cut exposure and regulatory scope with minimal impact on aggregate analytics. Teams should tier TTLs by sensitivity and build automated purge checks into CI/CD.

ARCHITECTURE

Goal. Provide a privacy-by-design data path in which no persistent identifier is ever collected while analytics/experimentation and model training still function.

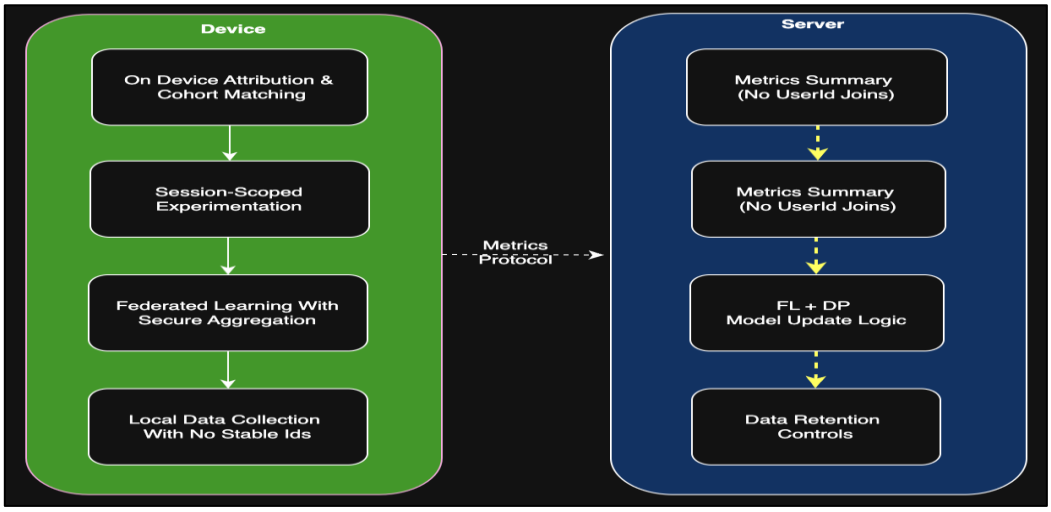

Fig. 7: High-level system diagram.

This architecture is composed of on-device processing, cohort hashing/generalization, secure aggregation, and strict retention so identifiers never exist on the server-side and linkability is neutralized while utility is preserved. (Kairouz, P. *et al.*, 2021; Margolin, E. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Bonawitz, K. *et al.*, 2017; Delaney, J. *et al.*, 2024; Apple, ; Privacy Sandbox, 2023; Prs India, 2023)

DISCUSSION

Our experiments and architecture surface four broader themes, echoing recent PET deployments and surveys (Kairouz, P. *et al.*, 2021; Wang, Z. *et al.*, 2022; Apple, ; Privacy Sandbox, ; Amin, K. *et al.*, 2024; Prs India, 2023).

Privacy Knobs are Interdependent

Generalization depth, cohort size, k-thresholds, ε/noise scale, FL participation rate, and TTL windows jointly define the privacy–utility frontier. Tuning one knob in isolation yields diminishing returns (e.g., shrinking ε without clipping). Automated privacy-budget accounting and utility monitors should surface these knobs as first-class

Publisher: SARC Publisher

configs in CI/CD pipelines. (Margolin, E. *et al.*, 2023; Amin, K. 2024; Prs India 2023)

Edge Complexity Rises

Moving attribution, aggregation, and assignment to clients reduces server risk but increases SDK complexity, battery/CPU cost, and observability challenges. Without raw logs, teams must lean on synthetic traffic, shadow pipelines, and privacy-safe debug channels. (Apple, ; Privacy Sandbox, 2023; Near, J. P. *et al.*, 2024).

Inference under Constraints

Thresholding, censoring, and DP noise violate assumptions of classical estimators. Bootstrap resampling, Bayesian shrinkage, and hierarchical models help recover power and calibrated intervals. Delayed postbacks slow iteration. (Kazan, Z. *et al.*, 2023; Chen, W. N. *et al.*, 2024; Zhang, C. *et al.*, 2020; Amin, K. *et al.*, 2024).

Compliance Becomes Code

TTL purging, schema pruning, and cohort gates can be enforced as policy-as-code and audited automatically, but those controls need their monitoring. (Privacy Sandbox, ; UNBigData, ; Prs India, 2023)

Limitations

We use synthetic data and proxy metrics (e.g., parameter distance for accuracy). Real traffic has heavy tails, device churn, and adversaries adapting to privacy knobs (Uber, 2022; Cohen, A. 2022) We did not model long-term temporal linkage or coordinated attacks. Future work should include adversarial simulations, sanitized production traces, and explicit compute/bandwidth cost models.

Opportunities

Integrate TEEs/MPC/HE for higher-assurance aggregation (Bonawitz, K. *et al.*, 2017; Zhang, C. *et al.*, 2020; Mozilla); explore adaptive privacy budgets; standardize PET dashboards that visualize both privacy posture and model/experiment health; investigate hybrid FL/MPC pipelines and privacy-preserving synthetic data generation. (Delaney, J. *et al.*, 2024; LinkedIn, 2022)

CONCLUSION

We showed that end-to-end ML, attribution, and experimentation pipelines can operate without persistent user/device IDs by composing a small set of proven techniques: (i) attribute generalization and cohort hashing to collapse fingerprintability (Uber, 2022; Cohen, A. 2022)(ii) on-device attribution and k-anonymized, optionally DP experiments (Kazan, Z. *et al.*, 2023; Chen, W. N. *et al.*, 2024; Zhang, C. *et al.*, 2020; Apple, ; Privacy Sandbox, 2023; Near, J. P. *et al.*, 2024) (iii) federated learning/analytics with secure aggregation and calibrated privacy budgets (Kairouz, P. *et al.*, 2021; Margolin, E. *et al.*, 2023; Karimireddy, S. P. *et al.*, 2020; Zhang, B. *et al.*, 2023; Delaney, J. *et al.*, 2024) and (iv) aggressive TTL-based retention plus schema pruning to minimize long-term exposure (Privacy Sandbox,; Chadha, K. *et al.*, 2024; Prs India, 2023)

Across simulations, these primitives eliminated linkability (SR → 0.000) and reduced retained data by >80%, while maintaining acceptable utility (≈3–15% accuracy cost; uplift std ≤ 0.011), consistent with recent empirical reports on privacy–utility trade-offs (Kazan, Z. *et al.*, 2023; Chen, W. N. *et al.*, 2024; Zhang, B. *et al.*, 2023; Amin, K. *et al.*, 2024). Organizations adopting this architecture can meet evolving regulations (GDPR, DPDP), adapt to platform privacy shifts (ATT, cookie deprecation), and earn long-term user trust, without abandoning personalization or rigorous experimentation. Future work should deepen adversarial modeling, integrate TEEs/MPC for high-assurance aggregation, and standardize PET dashboards and privacy-budget accounting. (Bonawitz, K., *et al.*, 2017; Mozilla, ; Amin, K. *et al.*, 2024 ; Prs India, 2023)

REFERENCES

1. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. "Advances and open problems in federated learning." *Foundations and trends® in machine learning* 14.1–2 (2021): 1-210.
2. Bian, C., Cheu, A., Chiknavaryan, S., Gong, Z., Gruteser, M., Guinan, O., ... & Yi, R. "Mayfly: Private Aggregate Insights from Ephemeral Streams of On-Device User Data." *arXiv preprint arXiv:2412.07962* (2024).
3. Kazan, Z., Shi, K., Groce, A., & Bray, A. P. "The test of tests: A framework for differentially private hypothesis testing." *International Conference on Machine Learning*. PMLR, (2023).
4. Margolin, E., Newatia, K., Luo, T., Roth, E., & Haeberlen, A. "Arboretum: A planner for large-scale federated analytics with differential privacy." *Proceedings of the 29th Symposium on Operating Systems Principles*. (2023).
5. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. "Scaffold:

Publisher: SARC Publisher

Stochastic controlled averaging for federated learning." *International conference on machine learning*. PMLR, (2020).
6. Chen, W. N., Cormode, G., Bharadwaj, A., Romov, P., & Ozgur, A. "Federated experiment design under distributed differential privacy." *International Conference on Artificial Intelligence and Statistics*. PMLR, (2024).
7. Wang, Z., Carrion, C., Lin, X., Ji, F., Bao, Y., & Yan, W. "Adaptive experimentation with delayed binary feedback." *Proceedings of the ACM Web Conference 2022*. (2022).
8. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. "Practical secure aggregation for privacy-preserving machine learning." *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. (2017).
9. Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., & Liu, Y. "{BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning." *2020 USENIX annual technical conference (USENIX ATC 20)*. (2020).
10. Zhang, B., Doroshenko, V., Kairouz, P., Steinke, T., Thakurta, A., Ma, Z., ... & Spacek, J. "Differentially private stream processing at scale." *arXiv preprint arXiv:2303.18086* (2023).
11. Delaney, J., Ghazi, B., Harrison, C., Ilvento, C., Kumar, R., Manurangsi, P., ... & Raykova, M. "Differentially private ad conversion measurement." *arXiv preprint arXiv:2403.15224* (2024).
12. Apple, "SKAdNetwork."
13. Privacy Sandbox, "Feedback Report - 2023 *Q4*," (2023)
14. Privacy Sandbox, "Protected Audience API overview."
15. Mozilla, "Glean SDK: Telemetry Without Identifiers."
16. Chadha, K., Chen, J., Duchi, J., Feldman, V., Hashemi, H., Javidbakht, O., ... & Talwar, K. "Differentially private heavy hitter detection using federated analytics." *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, (2024).
17. Microsoft Research, "Experimentation Platform (ExP),"
18. Uber, "Supercharging A/B Testing at *Uber*," (2022).
19. LinkedIn, "Our Approach to Research and A/B Testing." (2022).
20. Li, K., Chen, X., Leng, L., Xu, J., Sun, J., & Rezaei, B. "Privacy Preserving Conversion Modeling in Data Clean Room." *Proceedings of the 18th ACM Conference on Recommender Systems*. (2024).
21. Amin, K., Kulesza, A., & Vassilvitskii, S. "Practical considerations for differential privacy." *arXiv preprint arXiv:2408.07614* (2024).
22. Cohen, A. "Attacks on deidentification's defenses." *31st USENIX security symposium (USENIX Security 22)*. (2022).
23. UNBigData, "UN Guide on Privacy-Enhancing Technologies for Official Statistics."
24. Near, J. P., Darais, D., Lefkovitz, N., & Howarth, G. S. "Guidelines for evaluating differential privacy guarantees." *US Department of Commerce, National Institute of Standards and Technology*, (2025).
25. Intersoft Consulting, "General Data Protection Regulation."
26. Prs India, "The Digital Personal Data Protection Bill" (2023).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Thomas, A. "Privacy-First ML & Experimentation: Designing Systems Without User IDs." *Sarcouncil Journal of Engineering and Computer Sciences* 4.8 (2025): pp 619-628.

Publisher: SARC Publisher