Designing a Private Logging Pipeline

PEPR 2023

Mekhola Mukherjee, Thomas Vannet

Google

Best Practices that we think should be ‘*common practice*’.

~~Privacy Guidance~~

~~Legal Guidance~~

Not just Server Logs: Logging from Mobile Devices

Privacy

Personal data, sensors

No single party: OEMs, carriers, OS vendors, app developers

Critical

Outage disrupts lives

Billions of end user devices

Increasingly on-device logic

Technical

Low power

Low Bandwidth

High Latency


Launch “best” products with high reliability.


Have no data to test if the products are actually the “best”.

What are logs used for?

System Health

Memory Usage, Monitoring

Sessions Data

User onboarding journey

A/B Testing

Validate Hypotheses

Debugging

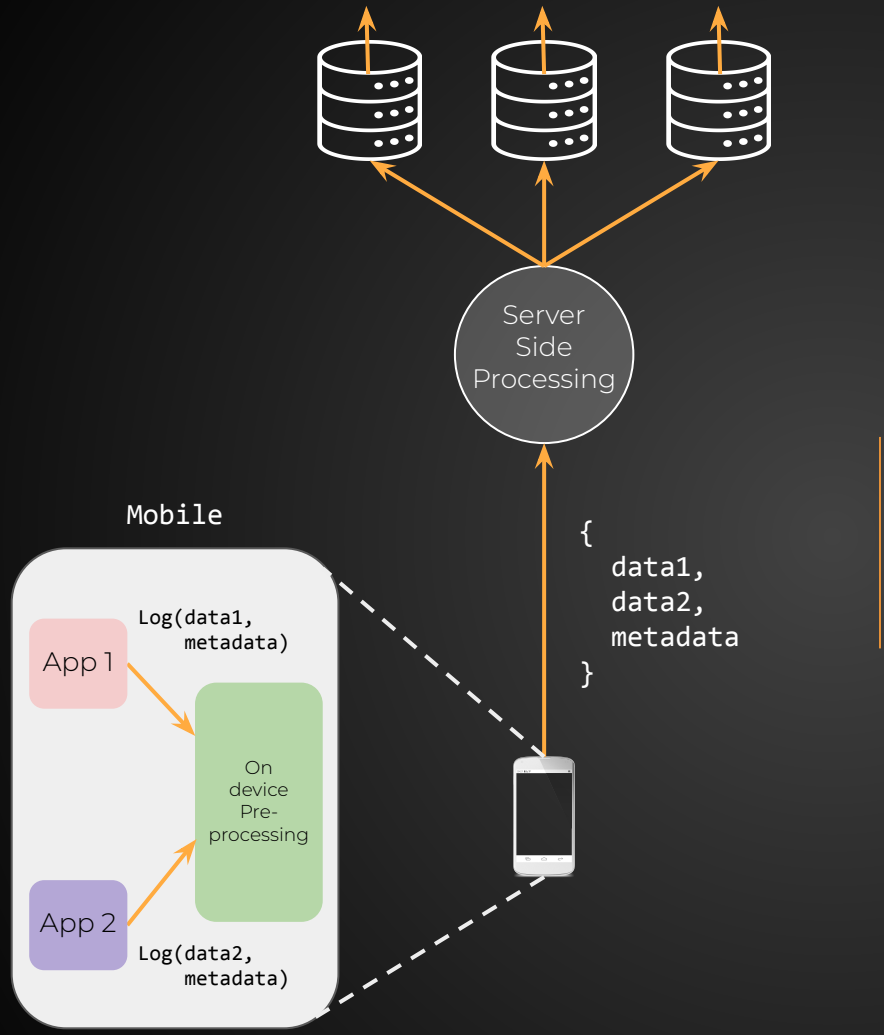
Fixing a voice call failure

Engagement Metrics

DAU, MAU

Abuse Prevention

Spam prevention, Detecting malicious apps



Journey of a Mobile Log


On-device policies and enforcements
Data minimization during collection

Understand your data

What

Type of data

How

Controls

Why

Collection purposes

Essential First Step: Annotation based policy enforcement

```
optional int32 caller_uid = 1 [(is_uid) = true];
optional string serial = 6 [ (android.privacy).dest = DEST_LOCAL ];
```

Define policies and enforce policy based collection

On-device Privacy Add ons

On-device data transformations make tangible privacy improvements

- Differential Privacy: Very niche use cases
- **String handling: Use enums instead of strings**
- Local Salting: Irreversible hashing server side
- Logging expiration: Automated metrics expiration

On-device Privacy Add ons

On-device data transformations make tangible privacy improvements

- Local processing: Relative and coarse timestamps, aggregations
- **Identifier choice: Remove or use least identifying**
- A/B testing: Prevent unique experiment assignments

Server side policies and enforcements

Analysts should only get access to what they need.

Data Policy Enforcements

Usage

Purpose restriction

Retention

Retention plans

Access

Access limitation

Essential First Step: Server side policies enforcement

Policy Auditing E.g., Access audits

Server side Privacy Add ons

- Shuffling: Removing identifiers and shuffling such that adjacent logs may not belong to the same user
- Random sampling: Collecting information from a random subset of devices
- Cross-device aggregations: Federated Analytics with Secure Aggregation

Server side Privacy Add ons

- K-anonymity: Enforce automatic aggregation thresholds for dashboards
- Allow/deny-lists: Prevent persistence of unneeded metrics
- Central DP integrated analytics tool: SQL DP, ease of querying

Smart Pipeline (futuristic outlook)

- ML based auditing of processing use case
- Multi party computation telemetry

Incentives to product teams for investing in private pipelines

- Integrated '*privacy by default*' features in systems aid accurate implementation
- Data minimization and transformations can reduce computation costs
- Transforming data helps with compliance

Questions?

Reach out if you are implementing these or if you have figured out new ways to achieve similar protections.