Building an End-to-End De-Identification Pipeline for Advertising Activity Data at LinkedIn

Saikrishna Badrinarayanan

Chris Harris

Agenda

1. Introduction
2. Design and architecture
3. Implementation challenges
4. Privatization mechanisms
5. Results and future work

1. Introduction

Motivation

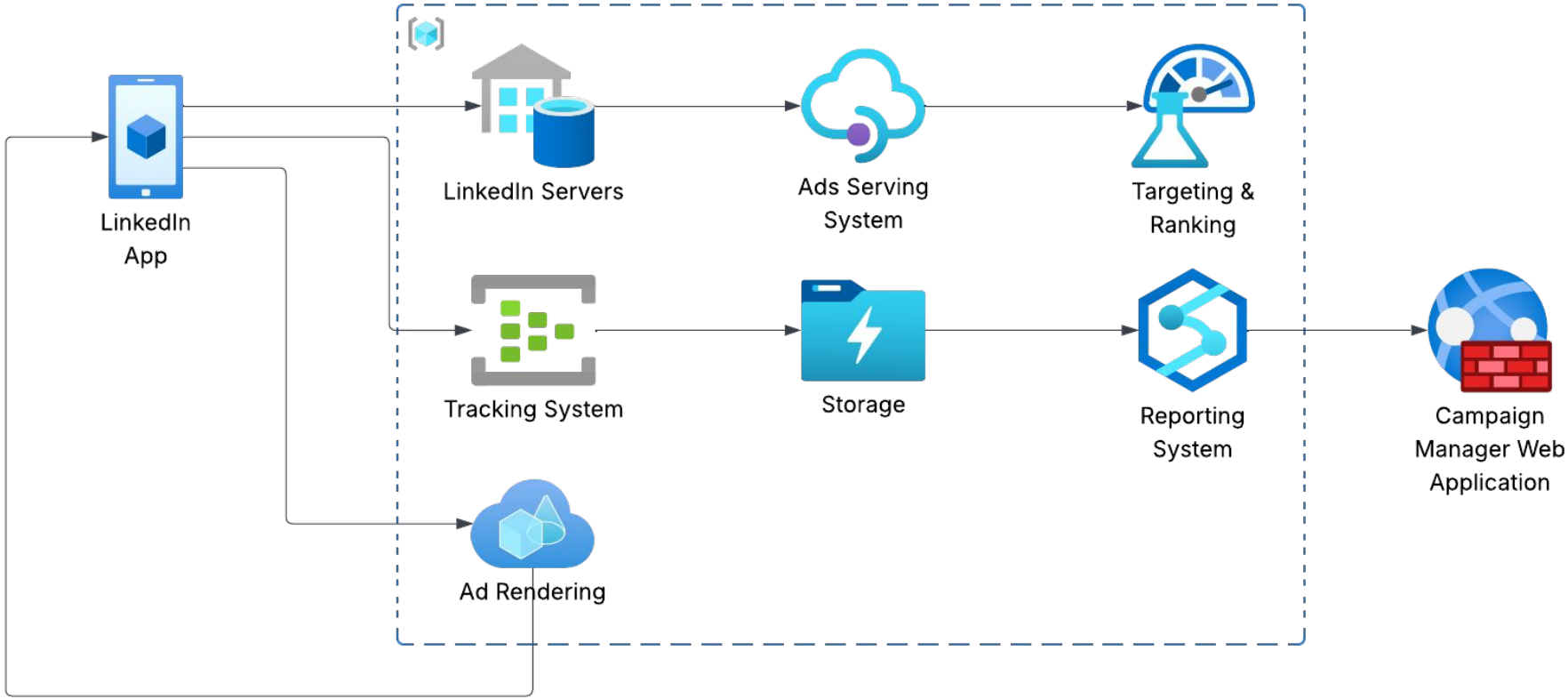
- Ads stack relies on members' activity data.

- Privacy regulations and platform restrictions

- Rebuild with privacy by design

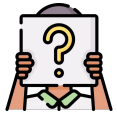
- Balance privacy with utility

Ads data flow

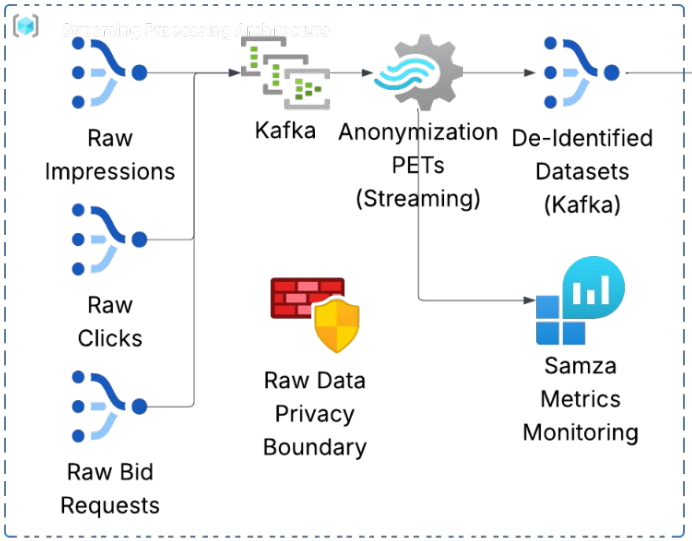


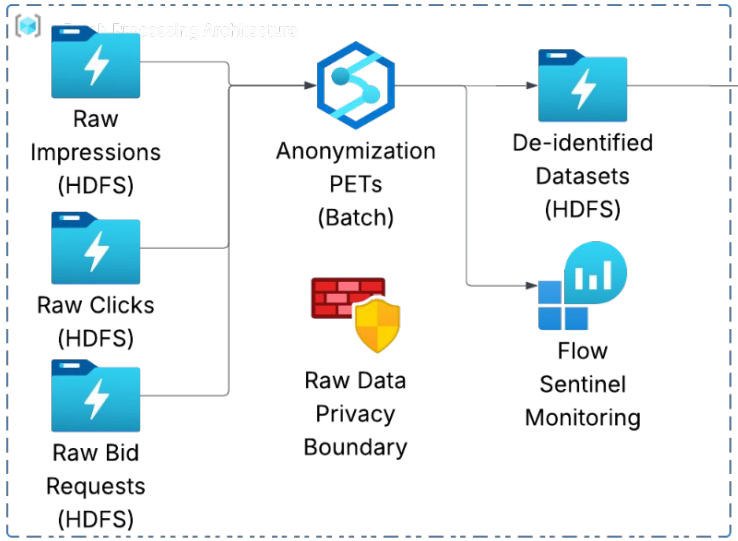
2. Design and architecture

Requirements

Functional	Non Functional	Use Cases
No re-identification 	Consumers SLOs maintained 	Billing 
Aggregated campaign performance metrics 	Privatization Streaming Data – 50ms p95	Reporting 
Consumer true norths	Privatization offline – finished by 3 AM PST daily	Forecasting

System design





Pros - Responsiveness
- Low Latency
- Frequent Checkpointing
- Smooth resource usage

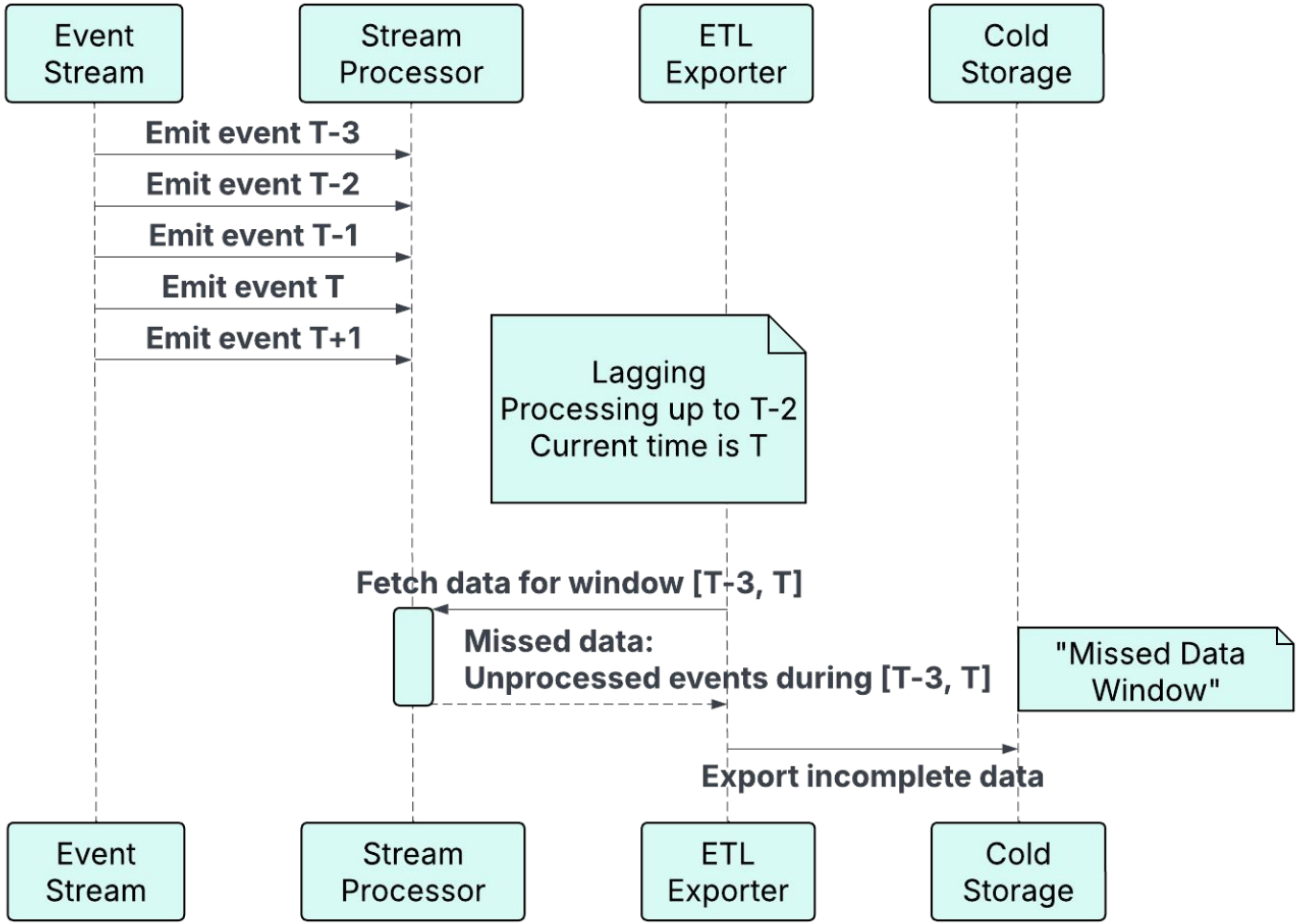
Cons - Complexity
- Operational Burden
- Development Complexity
- Backfills

Pros - Simplicity
- Development complexity
- Efficiency via scale
- Backfills

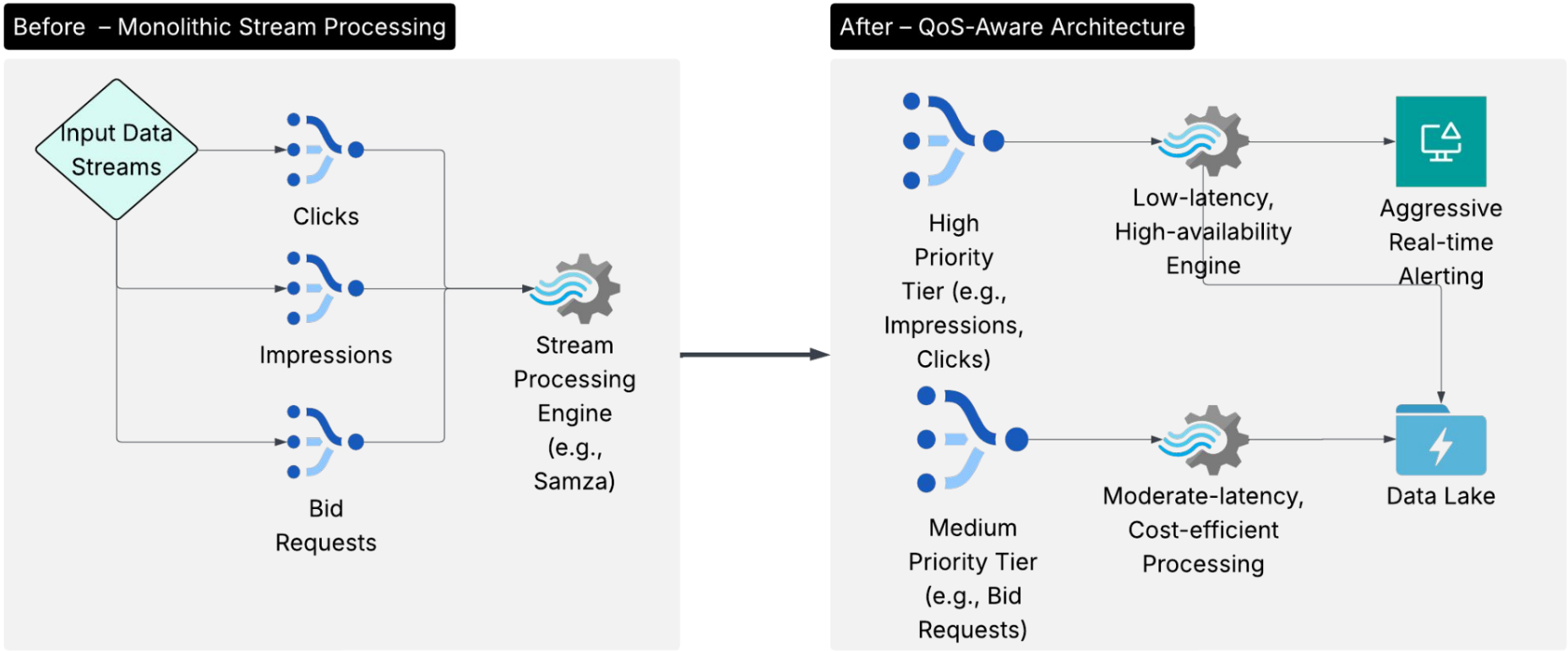
Cons - Freshness
- Latency
- High failure cost
- Spiky resource usage

3. Implementation challenges

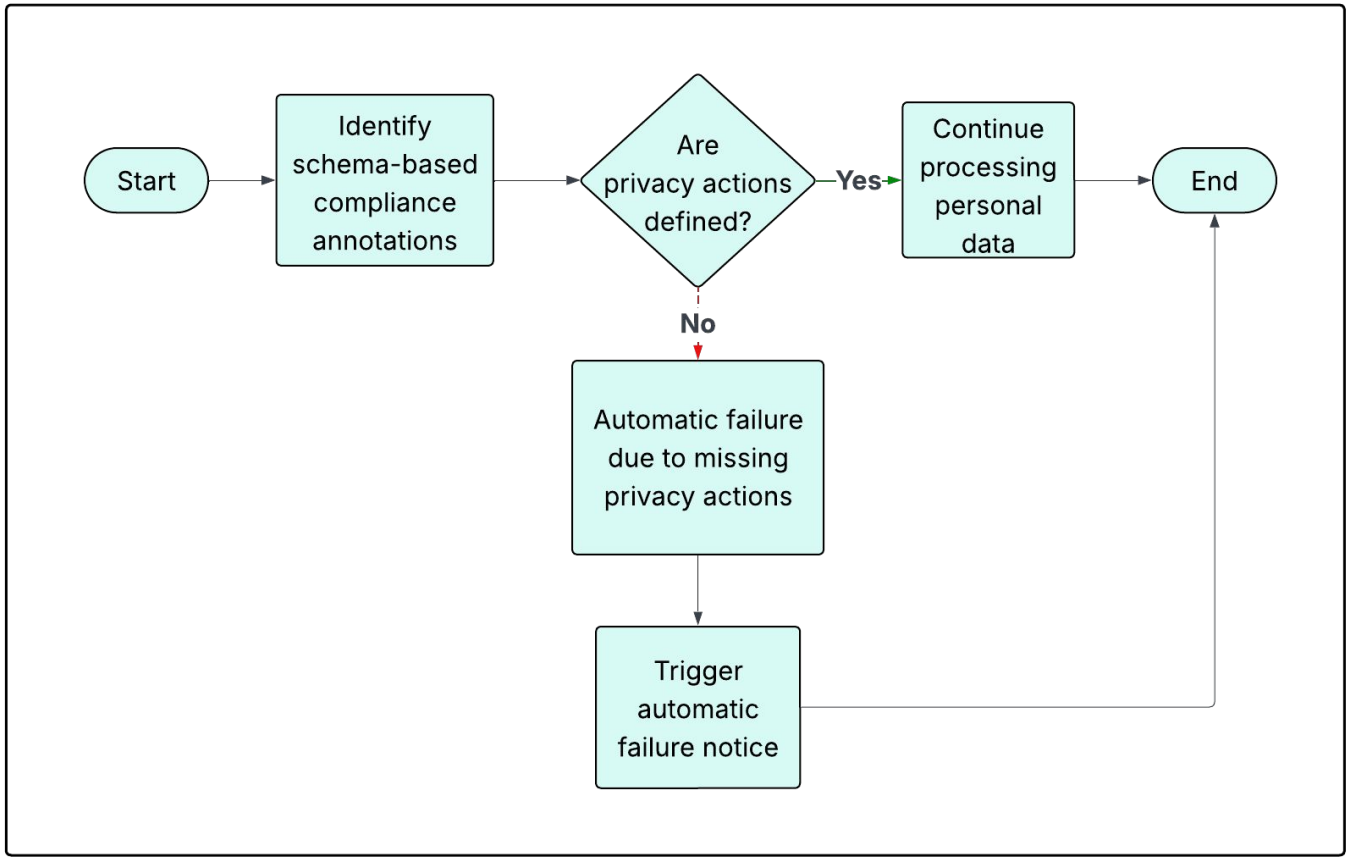
Lag


Event Stream
Stream Processor
ETL Exporter
Cold Storage
Emit event T-3
Emit event T-2
Emit event T-1
Emit event T
Emit event T+1
Lagging
Processing up to T-2
Current time is T
Fetch data for window [T-3, T]
Missed data:
Unprocessed events during [T-3, T]
"Missed Data Window"
Export incomplete data
Event Stream
Stream Processor
ETL Exporter
Cold Storage

System robustness

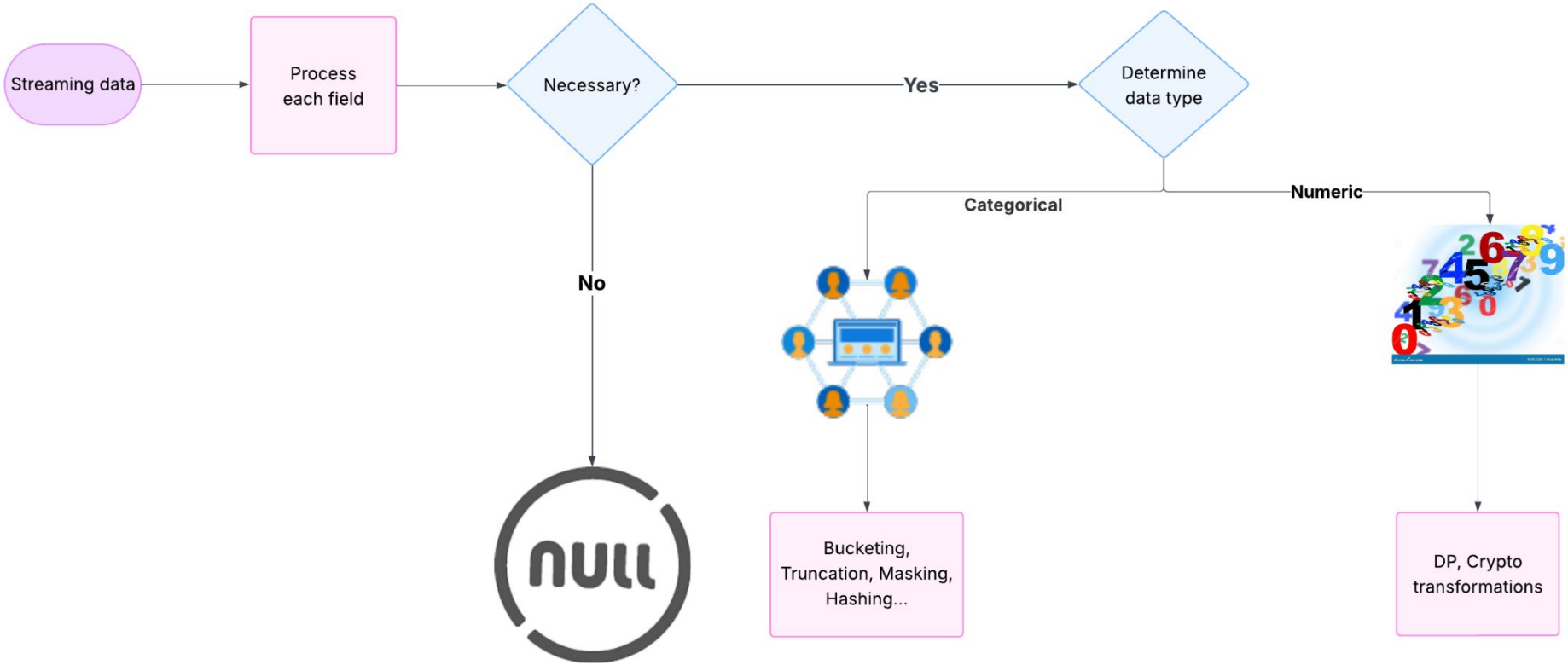


Other challenges



4. Privatization mechanisms

Privatization mechanisms



Eg 1: Timestamps

- Time of impression, how long ad was seen.
- 🤔 Innocuous?

- Rounding?

- **Local DP**
  - Clamp values.
  - Bound sensitivity.
  - Gaussian noise with eps = 4 and Delta = 0.0001
  - Product constraints

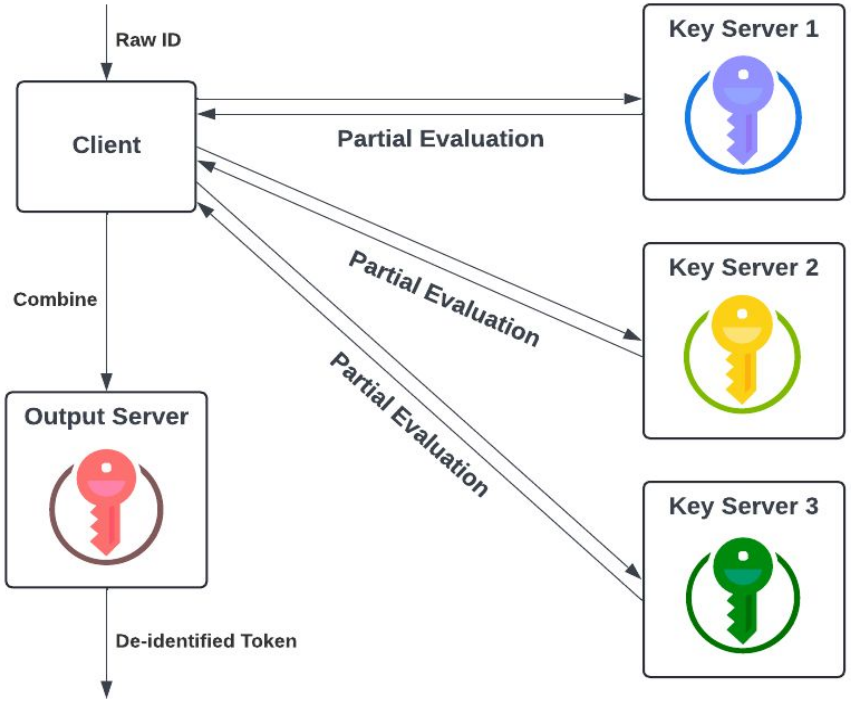
- Formal *metric DP* guarantee.

Eg 2: MemberID

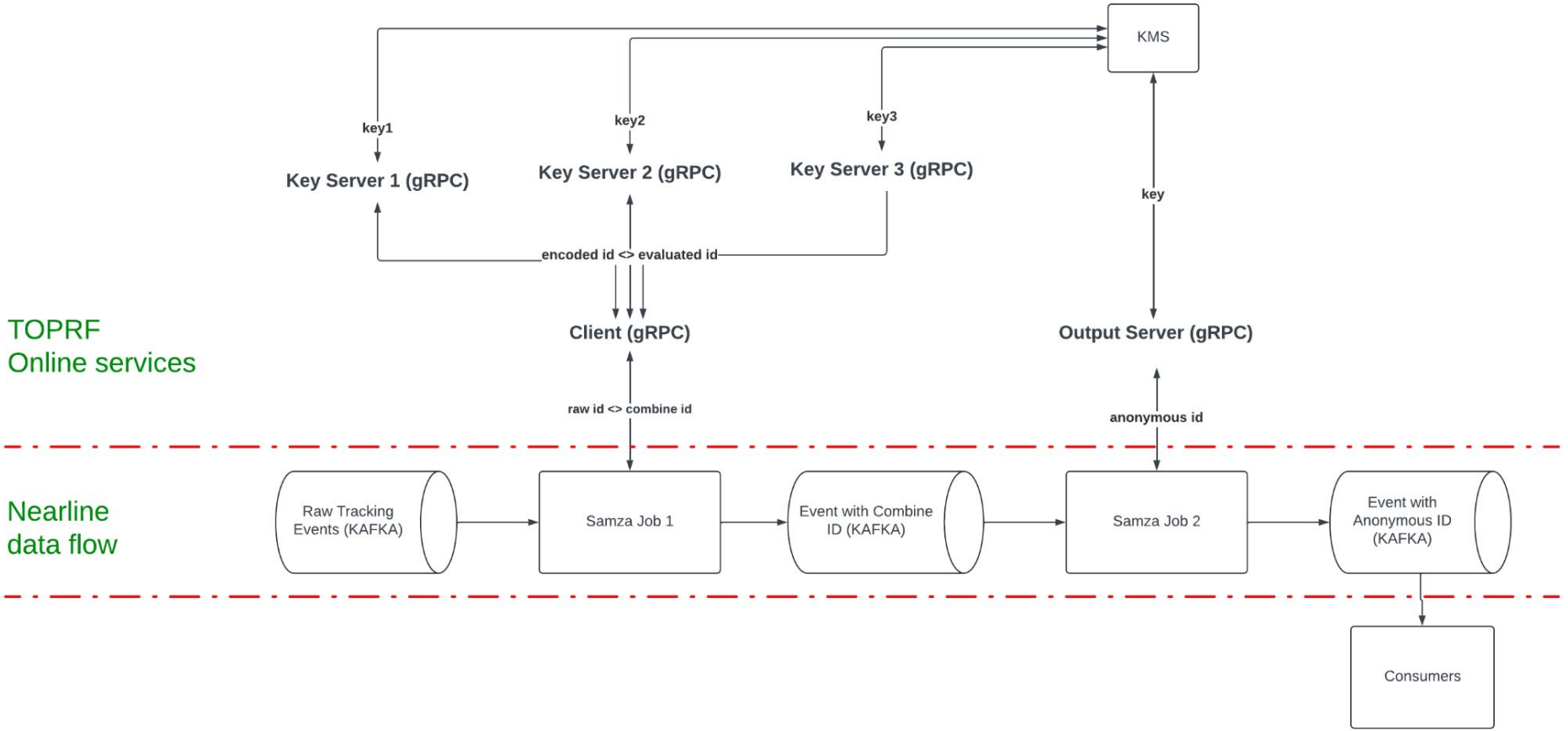
Principles

- Raw ID shouldn’t leave client.
- Raw ID and Token shouldn’t co-exist.
- Token can’t be constructed by one entity.

Threshold Oblivious Pseudorandom Functions



Eg 2: MemberID



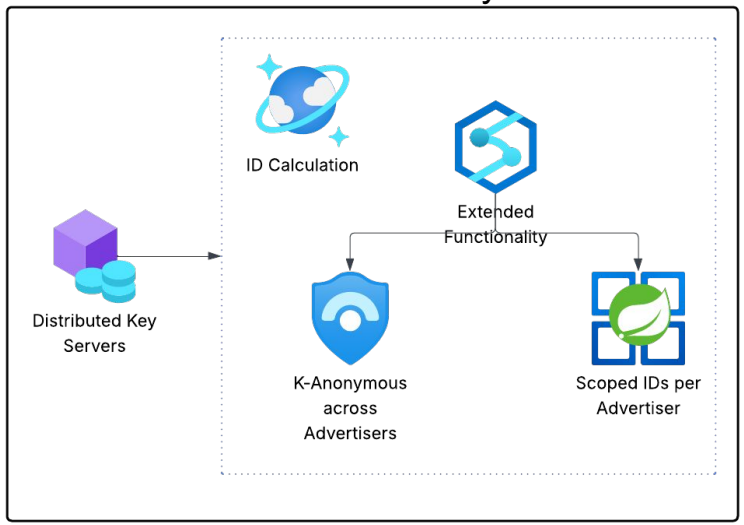
4. Results and future work

Results

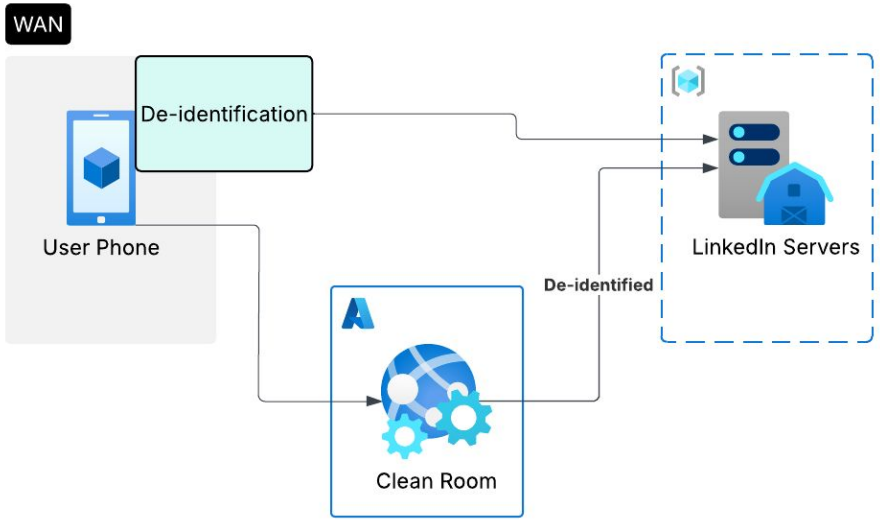
Privacy	Business
User data protected ⛑️	Reporting and billing onboarded 🚀
Smaller retention periods ⏳	No net negative impact 💰
Privacy monitoring 👀	SLA maintained ⏰

Future work

- Extended Functionality



- De-identify early



- 3P Signals
- Support modeling use-cases

Acknowledgements

- Ben Doan
- Bhavik Shah
- Brandon Poole
- Catalin Cosovanu
- Celeste Cassel de Camps
- Chao-Ping Lin
- Ian Koeppe
- Joe Xue
- John McCarthy
- Kamola Kobildjanova
- Mario Ortega
- Navi Mehrok
- Rahul Tandra
- Ryan Tecco
- Saranya Visvanathan
- Sidd Singh
- Siyao Sun
- Ting Dai
- Vinit Parakh
- Xiaolan Gu
- Yuliia Lut

Q&A

Media attributions

- Privacy icon made by freepik on flaticons.com
- Measurement icon made by Nuricon on flaticons.com
- Hourglass icon made by Ghozi Mutharom on flaticons.com
- Billing icon made by Kiranshastry on flaticons.com
- Magnifying icon made by surang on flaticons.com