Presto-native noisy aggregations for privacy-preserving workflows

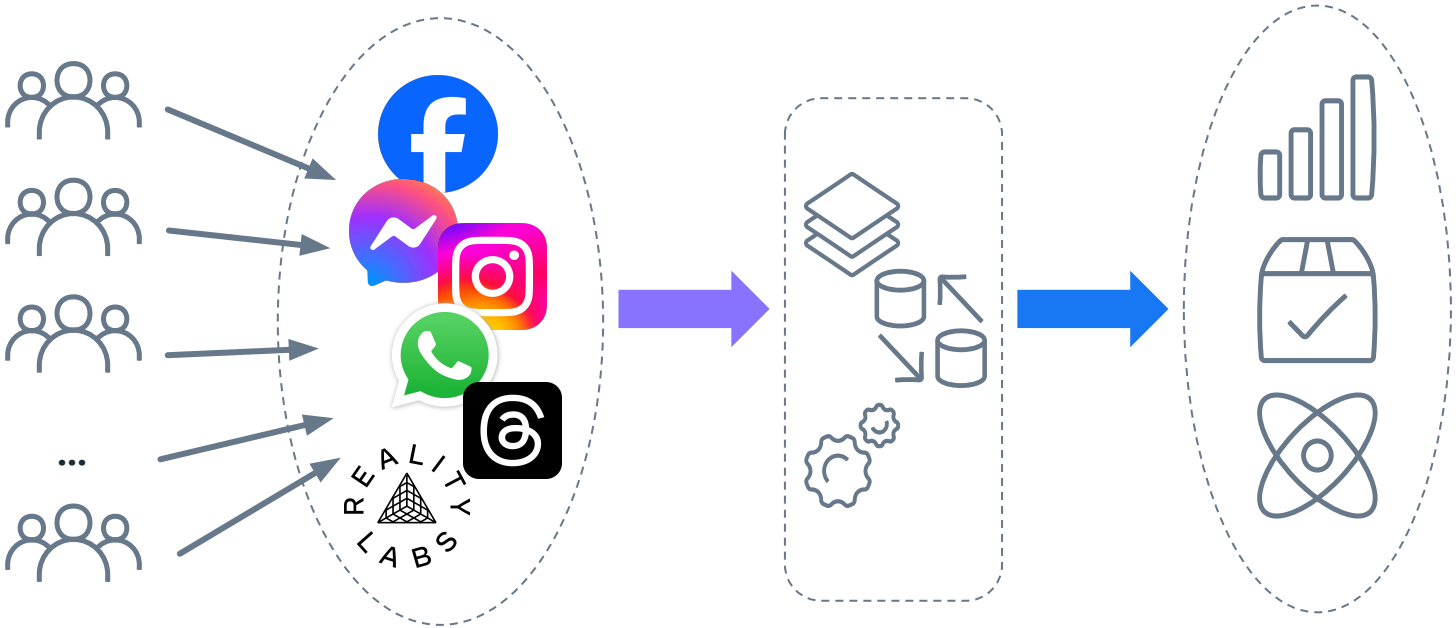
Privacy-preserving analytics at Meta scale

Kien Nguyen
Research Scientist

Chen-Kuei Lee
Software Engineer

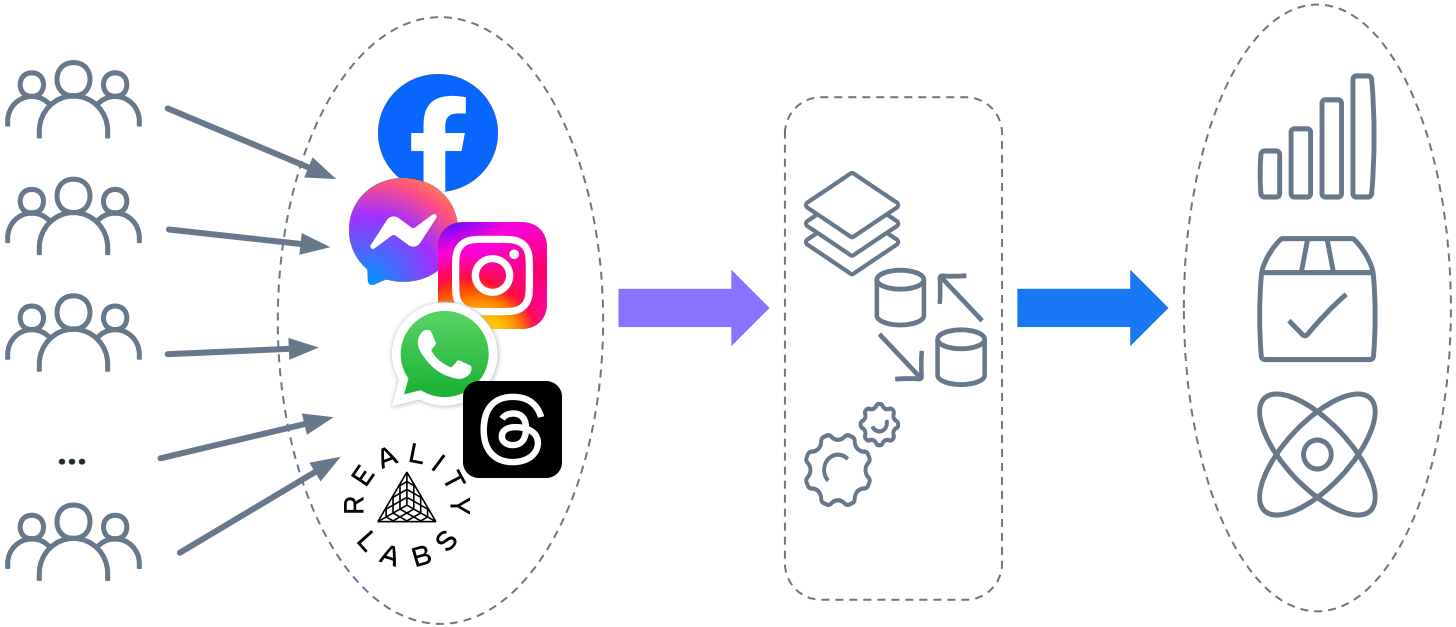
Meta

Differential Privacy at Meta scale



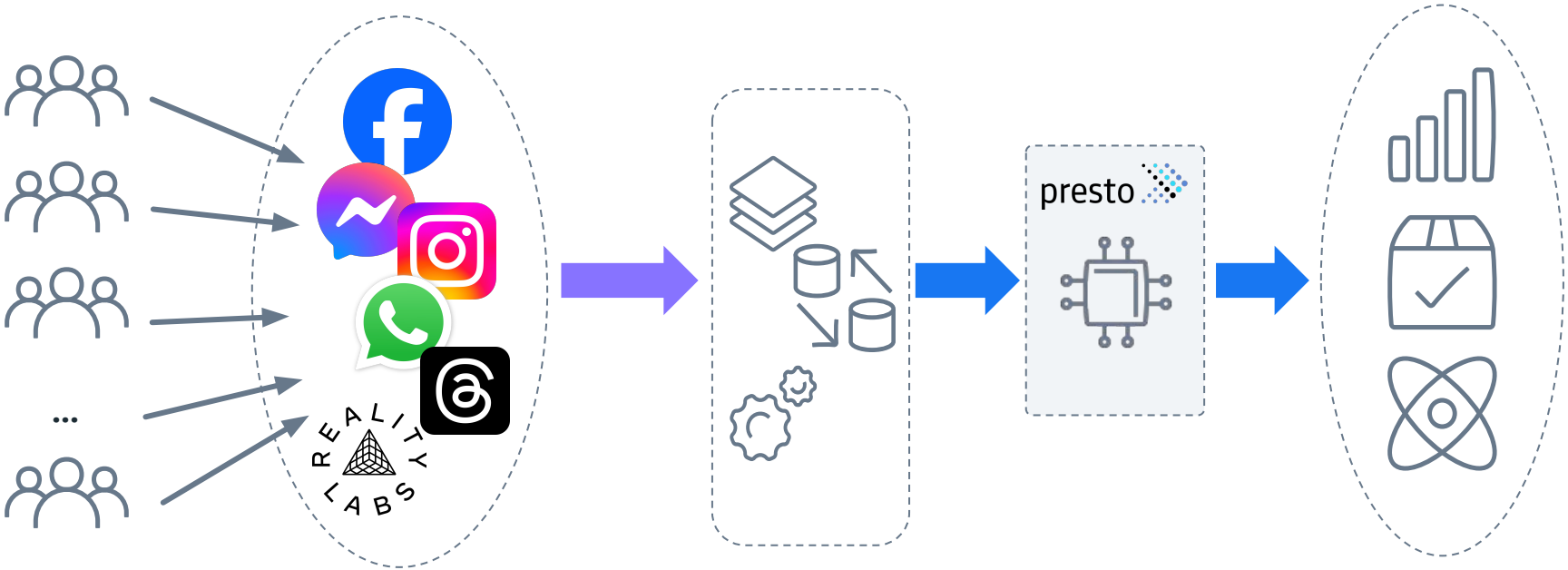
Differential privacy, due to its strong protection, is one of the privacy-enhancing technologies deployed by Meta to protect users' privacy while doing massive data analysis necessary for the business.

Differential Privacy at Meta scale is challenging



Scale Diversity Expertise Policy

Presto: Distributed SQL query engine for big data



Scale Diversity Expertise Policy

Rewrite SQL query to be DP using SQL

```
SELECT
  PERCENTILE(amount, 0.5)
    AS median_amount
FROM p2p_txns
```

```
WITH _bins0 AS ( SELECT * FROM (VALUES  (0),...,(99)) t (_bin)
), _input AS ( SELECT amount, 0.5 _f1 FROM p2p_txns
), _pdf0 AS (
  SELECT _bins0._bin, CASE WHEN t0._cnt IS NOT NULL THEN t0._cnt ELSE 0 END + 10.0 *
(LN(1 - 2*ABS(RAND()-0.5))) _cnt
  FROM (
  SELECT FLOOR((amount - 0.0) / 100000.0) _bin, COUNT(*) _cnt FROM _input
    GROUP BY FLOOR((amount - 0.0) / 100000.0)) t0
  RIGHT JOIN _bins0 ON t0._bin = _bins0._bin
), _cdf0 AS (
  SELECT _bin,
    SUM(_cnt) OVER (ORDER BY _bin NULLS LAST RANGE BETWEEN
      UNBOUNDED PRECEDING AND CURRENT ROW) _cdist,
    SUM(_cnt) OVER (ORDER BY _bin NULLS LAST RANGE BETWEEN
      UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) _tot
  FROM _pdf0
) SELECT CAST(0.0 + MIN(_bin) * 100000.0 AS DOUBLE) median_amount
FROM _cdf0
WHERE _cdist > _tot * 0.5
```

Rewrite SQL query to be DP using Presto-native noisy aggregations

```
SELECT
  PERCENTILE(amount, 0.5)
    AS median_amount
FROM p2p_txns
```

```
SELECT
    noisy_approx_percentile_qtree(amount, 0.5, 1, 1e-8, 1, 42.5)
      AS median_amount
FROM p2p_txns
```

Rewrite SQL query to be DP using Presto-native noisy aggregations

```
SELECT
    count(*),
    count_if(age > 10),
    sum(age),
    avg(age),
    approx_distinct(age),
    min(age),
    max(age),
    approx_percentile(age, 0.5)
FROM
    input
```

```
SELECT
    noisy_count_gaussian(*, 2.4),
    noisy_count_if_gaussian(age > 10, 2.4),
    noisy_sum_gaussian(age, 42.5),
    noisy_avg_gaussian(age, 42.5),
    noisy_approx_distinct_sfm(age, 1),
    noisy_approx_percentile_qtree(
          age, 0.001 1, 1e-8, 1, 42.5),
    noisy_approx_percentile_qtree(
          age, 0.999 1, 1e-8, 1, 42.5),
    noisy_approx_percentile_qtree(
          age, 0.5, 1, 1e-8, 1, 42.5)
FROM
    input
```

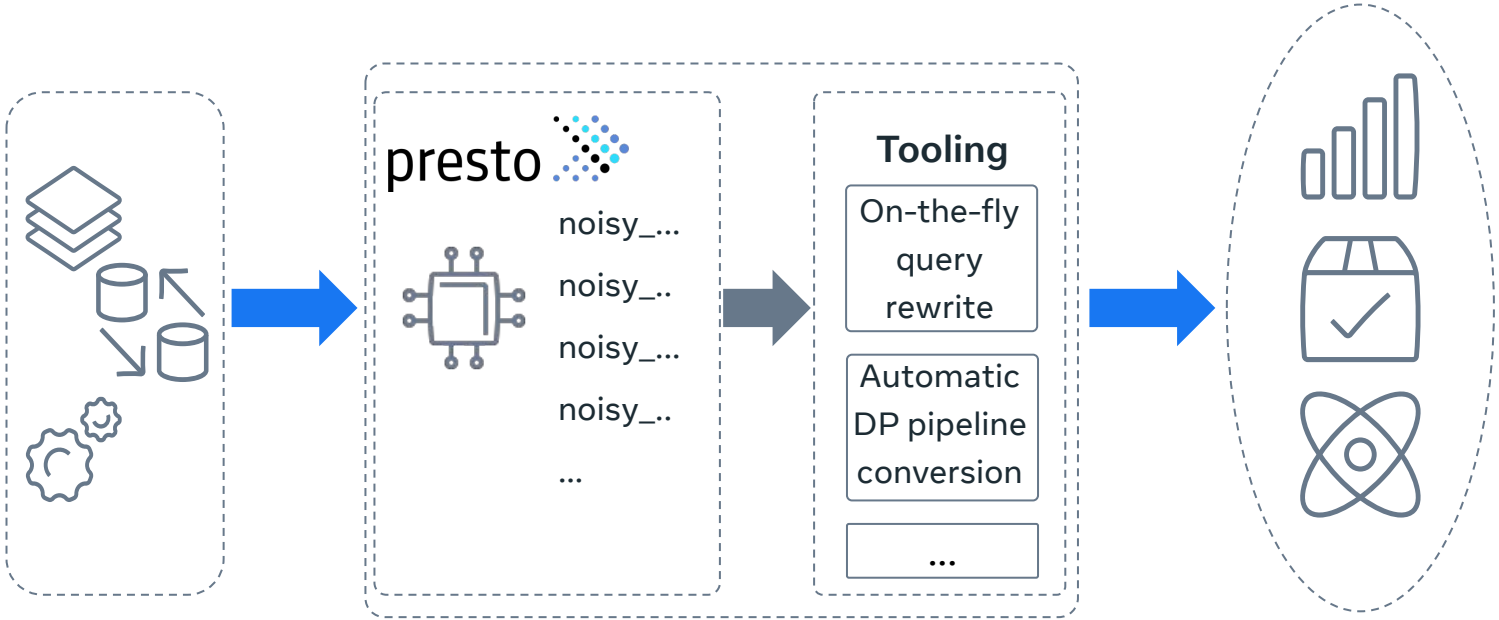
Presto-native noisy aggregations

Build lightweight “noisy aggregation” functions into Presto that can do the heavy lifting required by our varied private workflows

Presto-native noisy aggregations

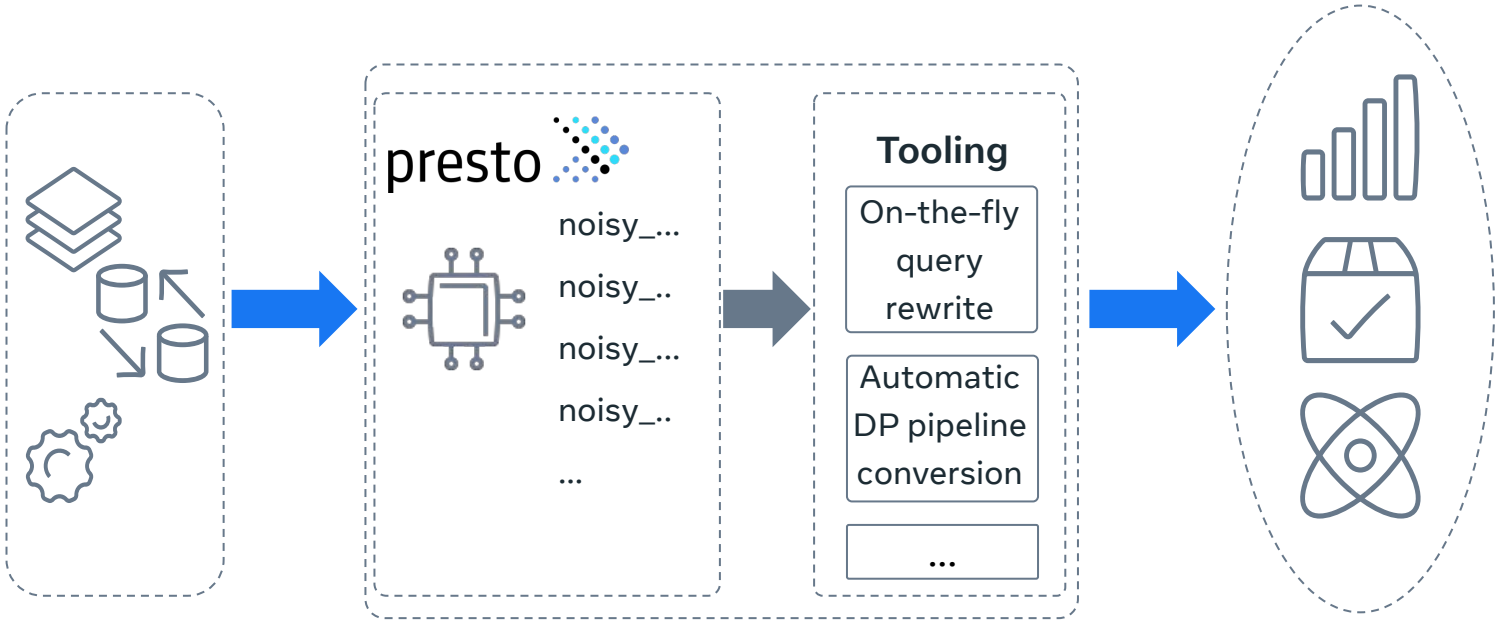
- Noisy `count`, `count_if`, `sum`, `avg`: Gaussian mechanism
- Noisy `approx_distinct`: sketch-flip-merge algorithm
- Noisy `min`, `max`, `approx_percentile`: Private quantile sketch

Presto-native noisy aggregations and tooling



Build tools to automatically convert queries/requests from analysts into a privacy-preserving form and to enforce privacy policies

Benefits



- Deploy DP at scale
- Flexibility for teams to use DP with minimal change to current workflows
- Minimizing inconsistencies and errors
- Enforcement and guarantee

Takeaways

Deploy DP at scale

Non-trivial problem

Build primitives and tooling

Build DP primitives directly into the compute engine (e.g., Presto) and build supporting frameworks and services to streamline the DP operations and to enforce privacy policies

Benefits

Enable DP at scale, increase flexibility, reduce errors, and enable privacy enforcement and guarantee


Meta