DAP Use Cases

PPM - IETF 114 - Philadelphia
DAP overview

- Specifies execution of a VDAF over HTTP
  - Each VDAF defines the distributed computation of some aggregation function
    - Prio, Poplar, ...
  - Work-in-progress in CFRG: draft-irtf-cfrg-vdaf-02

DAP = "Distributed Aggregation Protocol"
VDAF = "Verifiable Distributed Aggregation Function"
DAP is three protocols in one

**Upload sub-protocol:** Client generates report (i.e., encrypted input shares) for its measurement and sends report to Leader.
DAP is three protocols in one

**Aggregate sub-protocol:** Leader picks a set of reports and interacts with the Helper to verify them and compute **aggregate shares**.
DAP is three protocols in one

Collect sub-protocol: Collector issues collect request to Leader. Leader and Helper send encrypted aggregate shares for the corresponding batch of reports to the Collector.
DAP is three protocols in one

Collect sub-protocol: Collector issues collect request to Leader. Leader and Helper send encrypted aggregate shares for the corresponding batch of reports to the Collector.

How to select the batch?
Batch selection in DAP-01: Time-series

Reports assigned to **time windows** according to timestamp

- Collector specifies a **batch interval** that defines the span of time windows of reports in the batch
  - Batch interval boundaries **MUST** align with time windows
  - Batch intervals **MUST NOT** overlap.

**Problem:** There are other use cases for which this scheme is not well-suited.
**Issue #183**: Select batch by Client property

Reports **grouped** by Client property (User-Agent, location, etc.)

- Collector specifies a **group predicate** that defines the properties of reports in the batch.
  - "Chrome or Safari", "!Firefox", "Chrome && US || Firefox && !CA", …

**Only partially supported**: Could "spin up" a new task per group, but this might not scale.
**Issue #273**: Partition reports into fixed-sized chunks

Reports sorted into arbitrary **chunks** such that:

- Chunks are pairwise disjoint
- Chunks are the same size (or at least within some small $\epsilon$)
  - Control sample size for statistical analysis and/or differential privacy
  - Latency-sensitive applications

**Not supported**: Collector has no control over the size of the batch.
DAP needs more flexibility. But how much?

- Due to privacy considerations, Collector is more constrained than in traditional database or telemetry systems. (See next talk.)
- Open questions:
  - What are the different "query types"?
    - Does your use case call for a specific way of batching reports?
  - Do we need to be able to compose query types?
    - "group == Firefox && timestamp in range(batch_interval)"
  - Would every DAP deployment need to support all possible queries?
Proposal for DAP-02

Incremental step that addresses the use cases we have on-hand. Requirements:

- Enumerate all possible **query types**
- Collect request includes a **query** and the Leader chooses a batch of reports that satisfies query:
  - Helper confirms that all reports in the batch satisfy the query
- Implementations don't need to support all query types:
  - Supported query types configured out-of-band (part of task config)
  - One query type per task

```c
+ enum {
+   reserved(0),
+   time(1), // Use case #1
+   group(2), // Use case #2
+   chunk(3), // Use case #3
+ } QueryType;
+
+ struct {
+   TaskID task_id;
+   Interval batch_interval;
+ } CollectReq;
```
Abstraction

collector picks "batch predicate"

batch is a subset of uploaded reports satisfying batch predicate

\( f(\text{batch\_predicate}, \text{reports}) \rightarrow \text{batch} \leftarrow \text{known to all parties, efficiently computable function} \)

- Aggregator wants to know: Given a sequence of batch predicates, might there be two \( N \) batches with the same report? (whether there *is* depends on the set of reports).
  - Does there exist a value of \( N > 2 \) such that allowing these requests would NOT be a privacy violation (with some application of DP or whatever)?
  - What is a reasonable value of \( N \)? Does it depend on DP being used? Does it depend on size of batch intersection count?

- \( M_1 \text{\ intersect} M_2 \text{\ intersect} \ldots M_n = \{\text{report1}\} \)
- \( M_1 \text{\ intersect} M_2 = M_1 \text{\ setdiff} \{\text{report1}\} \)
Considerations for queries

Privacy: [See next presentation]

Minimize storage cost: Not feasible to allow Collector to slice data arbitrarily. E.g.:

- **#1** Batch interval's start and end time MUST align with batch windows (see [draft-ietf-ppm-dap-01, Section 4.4.5](#))
- **#2** Upper-bound the total number of distinct groups
- **#3** Assign each chunk a unique identifier and distribute amongst the Aggregators

![Diagram showing Time-series, Groups, and Chunks](#)