Federated Analytics and Agents Architecture

Dmitry Namiot

Lomonosov Moscow State University
dnamiot@gmail.com

GRID 2025

Outline

- The term “Federation” comes from the Latin word *foederis*, which can be translated as a treaty, agreement, or contract. And the term “federated” itself usually refers to the linking of autonomously operating entities.

- Data federation solves the problem of data access without compromising data security. How to aggregate information from independent data sources without revealing the exact details of individual participants?

Content

- Introduction
- Why is this necessary?
- Definitions
- Examples and protocols
- MCP servers
- Agents as proxy nodes in federated analytics

Introduction

- Data federation can be defined as a software process that allows multiple data sources (databases) to work together as a single entity.
- Unlike distributed databases, all data sources in such a system are completely independent.
- Federated analytics (the term itself was proposed in 2020) involves building analytical reports (queries) while hiding both the patterns of data in sources and the exact answers (for example, using differential privacy).

Introduction

- This is an alternative to a model in which data is moved or duplicated and then centrally hosted - when moving, data becomes vulnerable to interception, and moving large data sets is often very expensive for researchers.
- The use of this technology is very relevant for accessing sensitive biomedical data, since the data remains within the relevant jurisdictions (a medical institution does not share its data with others).
- 2020 - Covid has increased interest in this kind of technology

Introduction

- A similar problem occurs with the data of telecom operators, who cannot share their subscribers' data, but are interested, for example, in building a common subscriber churn model with other operators (who, of course, are in the same position regarding the sharing of their data).
- The same applies to financial institutions.
- In fact, when data becomes one of the significant sources of income for companies, the desire to share it will become less and less, even if we forget about legal restrictions.

Why?

- The World Economic Forum notes that, in fact, 97% of collected healthcare data is unused – it cannot be directly downloaded (uploaded), and there are no tools for federated analysis.
- The vast amount of data produced by healthcare institutions around the world remains underutilized.
- Current statistics show that 77% of healthcare systems do not have a coherent, integrated analytics strategy.
- The state of affairs with research data is likely even worse.

How does this work?

- Suppose that we have two data sets D_1 and D_2, and the querying party wants to get the number of records S that satisfy some given criterion.
- Let the true number of such records in the data sets be S_1 and S_2, respectively.
- In the federated analytics system, some random value M is generated, so that the first client returns the value S_1+M, and the second: S_1-M. By adding both answers, the querying party will get the true value S_1+S_2, while it will not know the exact answers of the surveyed clients.
- The scheme described in this example is, in fact, a protocol for exchanging data when solving a specific problem.

On definitions

- There is no single approach to federation, and different configurations entail different legal, regulatory, and technical requirements.
- Full federation refers to when both data access and compute access are federated across distributed computing and databases to enable querying and collaborative analysis of data.
- However, there is also potential for partial federation, where either compute or data access is federated, and compute or databases are distributed.
- The traditional model (no federation) is to centralize (download) all available data (datasets) and run queries on this common dataset.

On definitions

- In federated analytics, there is typically a querying party (question-asker) that wants to know some property or answer a question based on data distributed among different other parties (clients).
- Each of these clients owns a subset of the data that represents their local dataset. The querying party wants to answer the data analytics query through collaboration among multiple data owners.
- No raw data is exchanged or transmitted, but instead, intermediate responses to queries that are intended to be aggregated at the querying end are transmitted to answer the intended query.

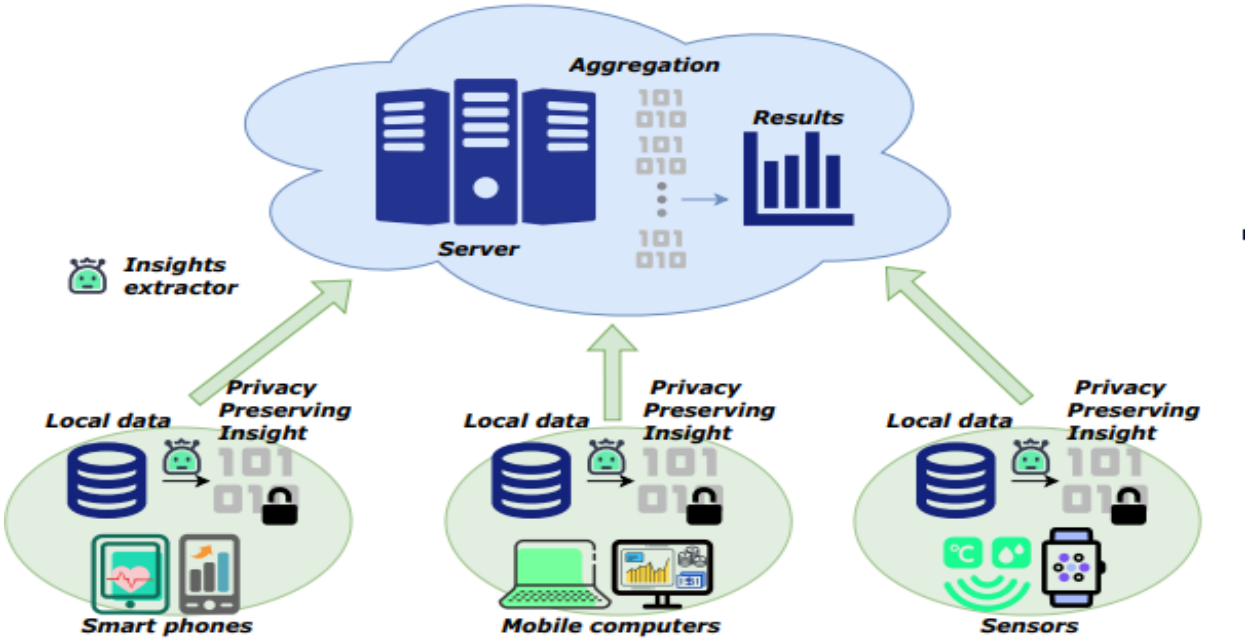
On definitions

- From this generalized perspective, the goal of federated analytics is for the querying end to obtain an answer to the following query Q

$$Q(D) = F\omega(D_1, D_2, \ldots, D_N)$$

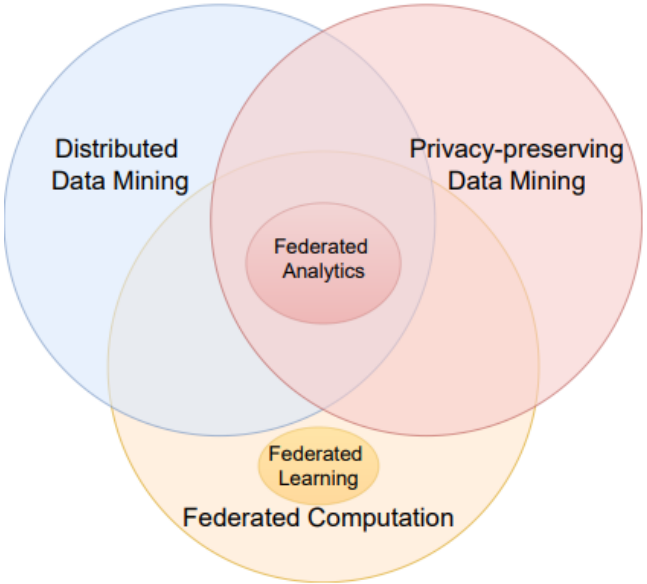
- Here $D = \{D_i\}^m$ are the private data sets of N data owners, and $F\omega$ is a parameterized function of the data that describes the target query.
- Accordingly, a taxonomy for federated analytics should consider (cover) the tasks to be solved, coordination models, and approaches to preserving privacy.

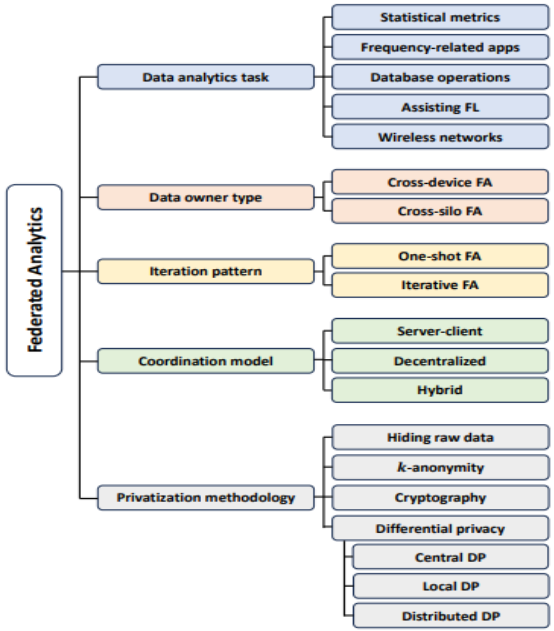
On architecture



Wang, Zibo, et al. "A survey on federated analytics: Taxonomy, enabling techniques, applications and open issues." IEEE Communications Surveys & Tutorials (2025).

On related areas





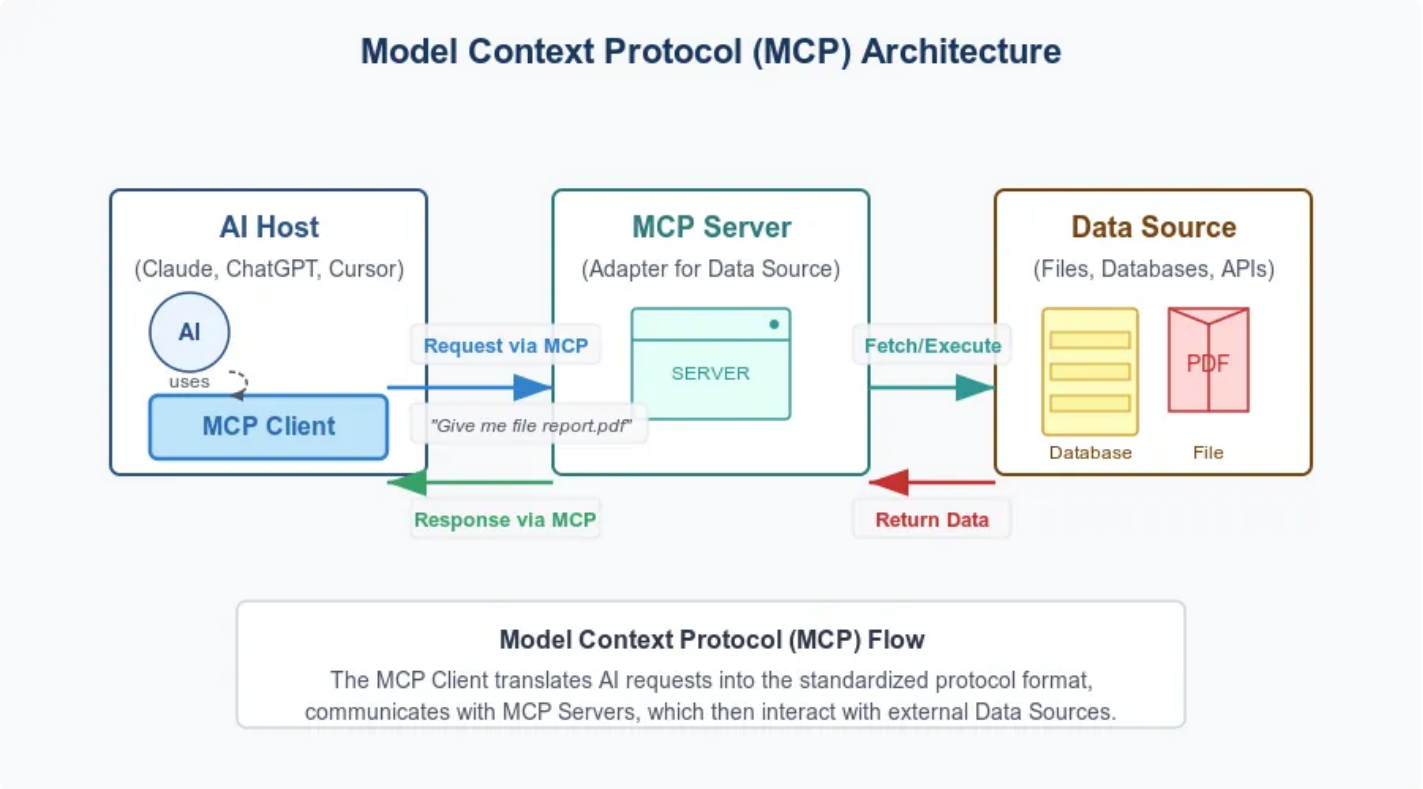
Wang, Zibo, et al. "A survey on federated analytics: Taxonomy, enabling techniques, applications and open issues." IEEE Communications Surveys & Tutorials (2025).

On protocols

Reference	Data analytics task	Privacy	Note
Cormode *et al.* [31]	Mean computation	LDP	One-bit client upload
Ding *et al.* [32]	Mean computation	LDP	Handling continuous data collection
Harmony [33]	Mean computation	LDP	Laplace mechanism
PrivKV [34]	Mean computation	LDP	Means of key-value structure data
Honeycrisp [37]	Mean computation	Cryptography&CDP	Sparse vector theorem
FEVA [24]	Mean computation	Hiding raw data	Federated Computation Builders architecture
SecAgg [38]	Mean computation	Cryptography	Utilization of two kinds of masks
Bell *et al.* [39]	Mean computation	Cryptography	Interacting with only a small random part of clients for each client
FastSecAgg [35]	Mean computation	Cryptography	Multi-secret sharing scheme based on fast Fourier transform
Turbo [36]	Mean computation	Cryptography	Circular communication topology
Zhao *et al.* [52]	Mean computation	Cryptography	Trusted third party
LightSecAgg [53]	Mean computation	Cryptography	a Secure and dropout-resilience secret sharing scheme
Liu *et al.* [54]	Mean computation	Cryptography	Lightweight and dropout-resilience secure aggregation
Wei *et al.* [45]	Gradient aggregation	LDP	Lightweight and dropout-resilience secure aggregation
Kim *et al.* [46]	Gradient aggregation	LDP	Adjustable query sensitivity
HFL-DP [47]	Gradient aggregation	LDP	Server-edge-client architecture
Geyer *et al.* [42]	Gradient aggregation	CDP	Hiding the participation of each participating client
Mcmahan *et al.* [43]	Gradient aggregation	CDP	Recurrent language models
Hu *et al.* [44]	Gradient aggregation	CDP	CDP-enhanced aggregation in personalized FL
Aggarwal *et al.* [40]	Medians and percentiles	Cryptography	Transforming the problem into a combinatorial circuit
Tueno *et al.* [41]	Medians and percentiles	Hiding raw data	
Bohler *et al.* [51]	Medians and percentiles	Cryptohraphy& LDP	Exponential Mechanism
[56]	Medians and percentiles	Hiding raw data	Model-based optimization approach in a federated way
Dennis *et al.* [27]	Clustering	Hiding raw data	Utilizing a hierarchical structure
Zhou *et al.* [58]	Clustering	Hiding raw data	Applying kernel functions to transform data points into feature vectors
UIFCA [59]	Clustering	Hiding raw data	Replacing the models used to be generative models that capture the distribution of one cluster
Lubana *et al.* [57]	Clustering	Hiding raw data	Uploading both the cluster model parameters and the local centroids to the server
Servetnyk *et al.* [60]	Clustering	Hiding raw data	Uploading a vector representing the number of samples falling into each bin of the grid
FedWalk [48]	Graph metrics	LDP	FA version of a random walk
LF-GDPR [49]	Graph metrics	LDP	Performing local computation by deriving the adjacency vector and the degree scaler each client
Liu *et al.* [50]	Graph metrics	LDP	Protecting the information of vertex neighborhood within local graph data

Wang, Zibo, et al. "A survey on federated analytics: Taxonomy, enabling techniques, applications and open issues." IEEE Communications Surveys & Tutorials (2025).

On protocols



Conclusion

- Federated data analysis is a direct way to bring private distributed data back into research.
- This will help to fully democratize access to data in many areas.
- Data federation can also help support global collaboration in data research. Federated data analysis maximizes cost efficiency by avoiding expensive data transfers.
- Federated platforms, with the attendant benefits of lower cost, can help make big data analytics more accessible.