Hands-On Federated Analytics

A practical guide to secure collaborative data science

Daniel Kapitan

2025-07-06

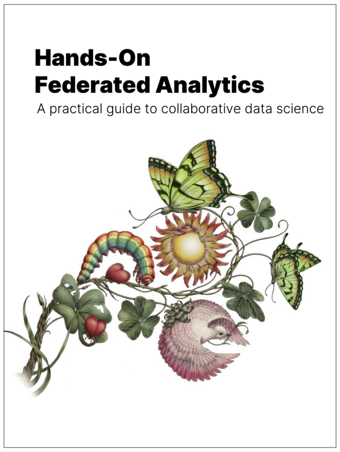
Table of contents

Welcome

It is a truth universally acknowledged, that a data scientist in possession of good skills, must be in want of more data.

Why I wrote this book:

- Federated analytics large potential but notoriously hard and often misunderstood. ‘Federation’ can mean many things.
- Motivated to support data science for the common good, I believe that federated analytics has a role to play in implementing data cooperatives, enable more wide-scale data sharing to support transformative change in healthcare, education, agriculture etc.
- Interesting from teaching modern data science: it touches on many separate topics that are interesting but hard to see in context. Cryptography, cloud engineering, standardization, predicate pushdown…
- Aim to enlarge the frame of thinking of those interested in the field. Side note denkraam with comic Maarten Toonder.



Scope: federated analytics at large. Having better understating of how federated queries work is also a huge step forward towards data commons.

Structure:

- stand-alone chapters that can be read on their own.
- chapter 1 to 4 explain basic principles and concepts of federated learning
- chapter 5 to 8 are specific usecases/applications
- chapter 9 and 10 are more technical and aimed at engineers and system administrators how to implement and maintain FA
- chapter 11 and 12 address ongoing and future developments
- accompanying code respository to run various examples and use-case
    - using vantage6 as the main platform

Image credit

In the spirit of the animal menagerie of O’Reilly books, I have chosen the *Fortuna Fragilis* as the preliminary cover. This species can be found on Terra Ultima by Raoul Deleo. I like to think that collaborative data science is fragile and fortuitous at the same time, and that it is worth kindling if we are to put data to use for the common good.

Part I

Principles

Where we lay out the principles and main concepts of federated analytics

- We introduce the taxonomy.
- We give toy examples to illustrate

Chapter 1

Why federated analytics?

We state the problem, the need and purpose of federated analytics

Chapter 2

Federated analytics

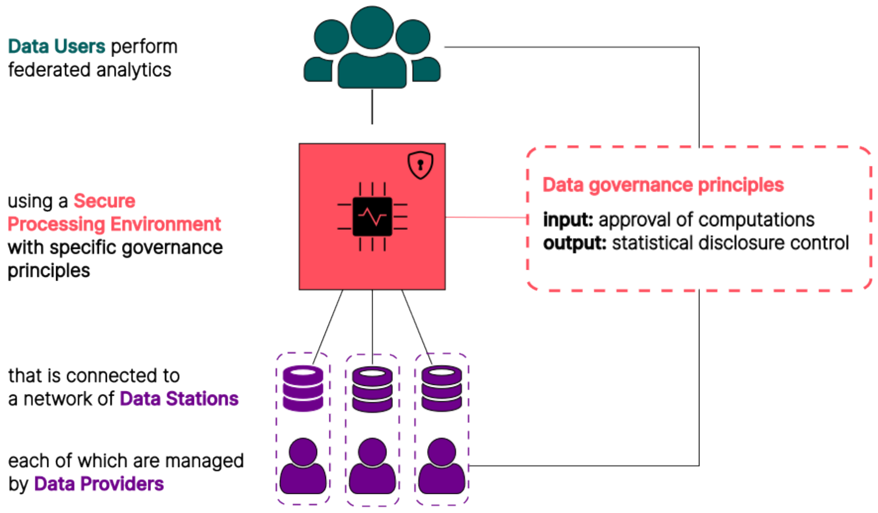
Where we explain the basic concepts and principles of federated analytics

2.1 Introduction

A survey on federated analytics (FA) defines it as “a paradigm for collaboratively extracting insights from distributed data that is owned by multiple parties (e.g., individual mobile devices or institutional organizations) under the coordination of a central entity (e.g., a service provider) without any of the raw data leaving their local parties or revealing information beyond the targeted insights. The core principles of this paradigm allow breaking the limitations for deriving analytics from limited centralized data, in terms of privacy concerns and operational costs.”[1]

Discerning characteristics:

- Apply some form of statistical disclore control (SDC) and/or privacy-enhancing technologies (PETS). Note we SDC can include differential privacy (DP), but we take a more generic approach that we want to protect the raw data as such, for example to provide guarantees to data holders with respect to commercial interest, trade secrets and the like. data privacy/security through PETs. Practically, all approaches to FA should have output control
- governance on the computation and not on data access as such


Data Users perform federated analytics
using a Secure Processing Environment with specific governance principles
Data governance principles
input: approval of computations
output: statistical disclosure control
that is connected to a network of Data Stations
each of which are managed by Data Providers

2.2 Key resources

Monographs, very detailed:

- Federated Learning Systems (2021) by Rehman and Gaber[2]
- Federated Learning (2022) by Ludwig and Baracaldo[3]

Review papers

- Elkordy, A.R. et al. (2023)[1]
- Wang, Z. et al. (2025)[4]

Applications in healthcare

- Rieke, N. et al. (2020)[5]
- Joshi, M. et al. (2022)[6], Fig 2 is nice diagram explaining horizontal, vertical partitioning and transfer learning

Hands-on - vantage6 workshop (link)

Chapter 3

Federated learning

From peer-to-peer to client-server federated learning approaches

Chapter 4

Secure multiparty computation

Where we explain ...

There are different solutions to prevent the reconstruction of raw data. One solution is to make sure that no party other than the data owner is actually able to see the intermediate data. One branch of techniques that can be used for this is Secure Multiparty Computation (MPC). With MPC, computations are performed collaboratively by multiple parties. Data is encrypted in such a way that other parties cannot see the original values, but values of multiple parties can still be combined (e.g. added or multiplied). A classic technique from the field of MPC is secret sharing. With this technique data is encrypted, after which pieces of the encryption are sent to the other parties. No single party will be able to reconstruct the original value. Only when a certain minimum of parties work together (n-1 in many cases) the original value can be retrieved.

When combining multiple values using secret sharing, this will result in the parties owning new puzzle pieces that when put together will reveal the result of the computation.

Part II

Applications

Applications of federated analytics

Chapter 5

Image classification

Training image classifiers with federated learning

Work and show-case Dekker et al.

Chapter 6

Text classiciation

Federated supervised learning with text

Work DHD AOIC

- F-NLP covers supervised learning with text
  - classification using text, often in combination with tabular data. NLP/Language models tools are used as pre-processing step for embedding
  - not in scope: training/tuning of LLMs (separate chapter?)
- PLUGIN-ML library

Chapter 7

Graph-based analytics

Federated graph-queries

7.1 Omics domain

- FAIR Data Cube
- Hartwig

7.2 Business analytics

- KIK-V

Chapter 8

Private set intersection

Using MPC for private set intersection

Part III

Engineering

Chapter 9

Composable data stack

The composable data stack as the foundation for federated analytics systems

Chapter 10

Semantic interoperability

Beyond mappings and syntacting interoperability

- Work ‘ontologies of ontologies’
-

Part IV

Future developments

Perspectives on future developments

Chapter 11

The paradox of open

Towards open, federated data platform

11.1 The engines of data science and the paradox of open

The Computer History Museum recounts that the history of the Information Age is driven by three 'engines', namely the silicon engine, the storage engine and the internet as the networking engine. Each of these engines have been instrumental in the development in of data science as we know today. Without the silicon engine, we would not be able to train foundational large language models with over 10^{23} FLOPs.[^1] Without the storage engine and the internet, we would not be able to store and collect petabyte-sized datasets such as Common Crawl as input for these models.

Besides the technological developments in and of themselves, the internet also gave birth to the open movement.[^2] Numerous organisations and initiatives have been launched with a belief in openness and free knowledge. Their proponents placed their bets on the combined power of networked information services and new governance models for the production and sharing of content and data. Members of this broad movement believed it possible to leverage this combination of power and opportunity to build a more democratic society, unleashing the power of the internet to create universal access to knowledge and culture. Such openness meant not only freedom, but also presented a path to justice and equality.

> We learned that open approaches flourish under two types of conditions:
>
> 1. **Projects where many people contribute to the creation of a common resource** – this is the story of Wikipedia, OpenStreetMap, Blender.org, and the countless free software projects that provide much of the internet's infrastructure.
> 2. **Circumstances where opening up is the result of external incentives or requirements, rather than voluntary actions** – this is the story of publicly-funded knowledge production like Open Access academic publications, cultural heritage collections in the Public Domain, Open Educational Resources (OER), and Open Government data.

However, the open revolution not happen. At least not on the scale that we and many other proponents of free culture expected. Although the copyright wars are almost over, conflicts about access to and control over informational resources have been superseded by conflicts about privacy, economic value extraction, the emergence of artificial intelligence, and the destabilising effects of dominant platforms on (democratic) societies. Instead of access to information, the control of personal data has emerged in the age of platforms as the critical contention.

Today, open is both a challenge to and an enabler of concentrations of power. Over the last decade, we have witnessed a wholesale transformation of the networked information ecosystem. The web moved away from the ideals and the open design of the early internet and turned into an environment that is dominated by a small number of platforms.

11.2 From closed digtal platforms towards open, federated data platforms

[7],[8]

Table 11.1: Types of data sharing and in relation to new standards and technology enablers to create openness. Taken from de Reuver et al. (2022).

	Type of data sharing	Technology enablers to create openness of health data platforms
1	Data at the most granular subject level, which is persisted and used to provide a longitudinal record.	Decentralized catalogs, semantic interoperability

	Type of data sharing	**Technology enablers to create openness of health data platforms**
2	Aggregated data, for example statistics for policy evaluation and benchmarking	**Federated learning**
3	Data analytics modules, that provide access to work and access the data.	**Federated learning (FL)**[5] and **privacy- enhancing technologies (PETs)**[9,10]: new paradigms that address the problem of data governance and privacy by training algorithms collaboratively without exchanging the data itself. Models can be trained on combined datasets and made available as open source artifacts for decision support. Data analysts can use FL and PETs to work with the data in a collaborative, decentralized fashion.
4	Trained models that have been derived from the data and can be used stand-alone for decision support.	ONNX, HugginFace …

11.3 Image credit

In the spirit of the animal menagerie of O'Reilly books, I have chosen the *Fortuna Fragilis* as the preliminary cover. This species can be found on Terra Ultima by Raoul Deleo. I like to think that federated data science is fragile and fortuitous at the same time, and that it is worth kindling if we are to put data to use for the common good.

Chapter 12

Data spaces and data solidarity

Can Europe realize its dream of free movement of data?

References

1. Elkordy, A.R. et al. (2023). Federated Analytics: A Survey. SIP *12*. 10.1561/116.00000063.

2. Rehman, M.H.U., and Gaber, M.M. eds. (2021). Federated Learning Systems: Towards Next-Generation AI (Springer International Publishing) 10.1007/978-3-030-70604-3.

3. Ludwig, H., and Baracaldo, N. eds. (2022). Federated Learning: A Comprehensive Overview of Methods and Applications (Springer International Publishing) 10.1007/978-3-030-96896-0.

4. Wang, Z. et al. (2025). A Survey on Federated Analytics: Taxonomy, Enabling Techniques, Applications and Open Issues. IEEE Commun. Surv. Tutorials, 1–1. 10.1109/COMST.2025.3558755.

5. Rieke, N. et al. (2020). The future of digital health with federated learning. npj Digit. Med. *3*, 1–7. 10.1038/s41746-020-00323-1.

6. Joshi, M. et al. (2022). Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. ACM Trans. Comput. Healthcare *3*, 40:1–40:36. 10.1145/3533708.

7. de Reuver, M. et al. (2018). The Digital Platform: A Research Agenda. Journal of Information Technology *33*, 124–135. 10.1057/s41265-016-0033-3.

8. de Reuver, M. et al. (2022). The openness of data platforms: A research agenda. In Proceedings of the 1st International Workshop on Data Economy DE '22. (Association for Computing Machinery), pp. 34–41. 10.1145/3565011.3569056.

9. Scheibner, J. et al. (2021). Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis. Journal of Medical Internet Research *23*, e25120. 10.2196/25120.

10. Jordan, S. et al. (2022). Selecting Privacy-Enhancing Technologies for Managing Health Data Use. Front Public Health *10*, 814163. 10.3389/fpubh.2022.814163.