Layers, pipes and patterns: detailing the concept of data stations as a foundational building block for federated data systems in healthcare

Daniel Kapitan
Health-RI, Dutch Hospital Data, Eindhoven University of Technology
daniel@kapitan.net

Jack Broeren
Health-RI

Jeroen Beliën
Health-RI, Amsterdam UMC

Niels Bolding
Health-RI

Stefan van der Loop
Cumuluz

Yannick Vinkesteijn
Dutch Hospital Data
y.vinkesteijn@dhd.nl

Joep de Ligt
Hartwig Medical Foundation
j.deligt@hartwigmedicalfoundation.nl

Abstract We describe …

1 A shift towards federated data systems as a design paradigm

The ambition for a seamlessly connected digital healthcare ecosystem, capable of leveraging vast quantities of patient data remains illusive. Designing and implementing health data platforms is notoriously difficult, given the heterogeneity and complexity of such systems. To address these issues, federated data systems have emerged as a design paradigm. This approach enables data to remain securely at its source and perform computations in a decentralized, distributed fashion.

Recent technological inventions offer important new enablers to implement federated data systems, most notably:

- **Significant increase in single-node computing capabilities**, whereby it is now possible to process up to 1 TB of tabular data on a single machine node thereby enabling increasingly large volumes of data to be processed in a decentralized, federated system [1, 2];
- **Maturity of federated analytics** and specifically federated learning as a means of performing analysis whilst 'hidding' the data from third parties, including training of deep learning models through aggregation of weights [3–5];
- **Privacy-enhancing technologies (PETs) such as secure multi-party computation (MPC)** that are now sufficiently mature to be used on an industrial scale, enabling com-

putations to be done under encryption (in-the-blind) thereby significantly improving security across a network of participants [6, 7];
- **The composable data stack** as a solution design that allows for unbundling of the venerable relational database into loosely coupled components, thereby making it easier and more practical to implement federated data systems using cloud-based components with microservices and thus opening up a transition path towards more modular and robust architectures [8, 9].

The architectural shift from centralized to federated data systems is not merely a technical evolution. Modern approaches to data governance are undergoing a similar paradigm shift towards federated solutions. Federated data systems are inherently more aligned with contemporary data governance frameworks, including the Data Governance Act (DGA), the European Health Data Space (EHDS) and the concept of data solidarity [10]. Within the context of large corporations, the concept of a data mesh is increasingly being adopted as well, which in essence is a federation of data producers and consumers within a commercial setting. From the perspective of sovereignty and solidarity, we believe that a commons-based, federated approach has distinct benefits in moving towards a more equitable, open digital infrastructure [11].

However, this ongoing paradigm shift towards is not without challenges. The notion of what constitutes a federated data system needs to be defined in more detail if we are to see the forest for the trees between different instantiations of the same concept. For example, 'federation' can mean any of the concepts:

- **Data federation** in de context of distributed databases addresses the problem of uniformly accessing multiple, possibly heterogeneous data sources, by mapping them into a unified schema, such as an RDF(S)/OWL ontology or a relational schema, and by supporting the execution of queries, like SPARQL or SQL queries, over that unified schema [12];
- **Federation within the context of a Personal Health Train (PHT)** refers to the concept where data processing is brought to the (personal health) data rather than the other way around, allowing (private) data accessed to be controlled, and to observe ethical and legal concerns [13–16], and is just one of many solutions designs that are collectively grouped as federated analytics [3];
- **Federation in Trusted Research Environment (TRE)** pertains to a mechanism for data sharing in a temporary staging environment within a network of research organizations through federations services such as localization and access [17];
- **Federation in the context of data spaces**, as described in the DSSC Blueprint 2.0, pertains to the support the interplay of participants in a data space, operating in accordance to the policies and rules specified in the Rulebook by the data space authority.

What then, is a viable development path out of this creative chaos?

2 Towards data availability for secondary use

Inspired by previous calls to action to move towards open architectures for health data systems [18, 19] and the notion of the hourglass model [18, 20, 21], we hypothesize that

the concept of a 'data station' can be used as a foundational building block for federated data systems. A data station should provide a set of minimal standards (at the waist of the hourglass), thereby maximizing the freedom to operate between data providers and data consumers within the context of a health data space. Note that this approach has many similarities with the FAIR Hourglass [21, 22]. Our approach of data stations presented here focuses on enabling on secondary data use of routine collected clinical data using the architecture of the Personal Health Train as a starting point [13–16, 23]. The objective of this paper is to extend this architecture in order to address four design questions that are relevant in ongoing efforts to implement nation-wide federated systems.

2.1 Online transactional vs. analytical processing

First, it is well known data systems have different design an performance characteristics depending whether they are built for online transactional processing (OLTP) or online analytical processing (OLAP), as summarized in the Table 1 below (taken from Kleppmann M, Riccomini C [24]).

Table 1: Distinguishing characteristics between online transactional and analytical systems

Property	Transaction processing systems (OLTP)	Analytic systems (OLAP)
Main read pattern	Small number of records per query, fetched by key	Aggregate over large number of records
Main write pattern	Random-access, low-latency writes from user input	Bulk import (ETL) or event stream
Data modeling	Predefined	Defined post-hoc, either schema-on-read or schema-on-write
Primarily used by	End user/customer, via web application	Internal analyst, for decision support
What data represents	Latest state of data (current point in time)	History of events that happened over time
Dataset size	Gigabytes to terabytes	Terabytes to petabytes

Current efforts to design and implement the EHDS in fact aims to support primary (i.e. OLTP) and secondary use (OLAP) in one go [25, 26]. This Herculean endeavour has spawned many initiatives to develop a coherent architecture and support implementation across Europe that ultimately should lead to interoperability in the broadest sense of the word, most notably:

- The Data Space Blueprint v2.0 (DSB2) by the Data Spaces Support Centre ([27]) that serves as a vital guide for organizations building and participating in data spaces.

- The Simpl Programme ([28]) that aims to develop an open source, secure middleware that supports data access and interoperability in European data initiatives. It provides multiple compatible components, free to use, that adhere to a common standard of data quality and data sharing.
- TEHDAS2 ([29]), a joint action that prepares the ground for the harmonised implementation of the secondary use of health data in the EHDS.

We believe, however, that in order to successfully design and implement health data spaces, more detailed analysis and solution patterns are required that distinguish between primary (OLTP) and secondary (OLAP) data use. Although functional components can be shared between these two, it is a matter of the devil being in the details. Hence one of the objectives of this paper is to detail an open, technology agnostic architecture for secondary use, to complement existing efforts and guide the development in the field.

2.2 Centralized vs decentralized processing

A second design question pertains to the choice of single-node (centralized) or distributed (decentralized) platforms, which are not only be driven by technical considerations (scalability, elasticity, fault tolerance, latency) but are also strongly dependent on organizational, legal or regulatory requirements such as data residency. The general approach of EHDS and other data spaces is federative by nature, that is, decentralized. For example, DSB2 stresses the need for interoperability and federative protocols within and across data spaces.

Upon closer inspection, however, specific functional components that are foreseen within the EHDS are best characterized as centralized (sub-)systems. As an example, consider the secure processing environments (SPE) as defined in article 73 of the EHDS. Known examples of such SPEs include data platforms provided by national statistics offices (CBS Microdata environment), healthcare-specific national platforms (Finland's Kapseli platform) and Trusted Research Environments (TREs) within the domain of research (see EOSC-ENTRUST for examples across Europe). Given that healthcare data is often vertically partitioned (data elements of the same subject are scattered across various data holders), SPEs provide the most effective means to (temporarily) share, integrate and analyse such data. Hence many SPEs are best described as centralized systems, and thus we need to take into account that data spaces constitute a hybrid architecture that includes both centralized and decentralized components. Thus a more detailed analysis is required to arrive at scalable solution patterns that combine centralized vs. decentralized processing in a larger federated health data system.

2.3 Solution patterns to populate data stations

A third design issue pertains to the mechanisms through which the data stations are to be populated with data by the data holder. In essence the industry is converging towards solution patterns from data engineering and data warehousing. OpenHIE specification, for example, makes a distinction of a Shared Health Record as a OLTP system for primary health data use vis-a-vis FHIR data pipelines that populate data stations for analytical, secondary use. [TO DO: improve wording, add references]

2.4 Federation as system of systems

The fourth design challenge pertains to the need for having a 'system of systems' [30] in the federated health data system at large. In real-world setting, secondary health data sharing will need to take into account the limited resources and expertise of smaller health care providers. In fact, the EHDS explicitly addresses this issue in article article 50 that exempts micro enterprises from the obligation to directly participate in the EHDS, but that each member state may opt to form so-called data intermediaries to act as a go-between (recital 59). Along a different dimension, we foresee a system of systems of various domain-specific federated networks that are loosely coupled, which poses design challenges in terms of autonomy (to what extend can each sub-network make independent choices), connectivity (how to connect sub-systems) and diversity.

2.5 Design questions framed within realization of Secure Processing Environments for secondary use

By addressing these four design questions, our aim is to specify the solution designs for a ***federated health data system with national coverage that can support secondary data use within the context of the EHDS***. To clarify this objective, consider Figure 1 depicting the main concepts relevant to our paper in terms of an Archimate layered view. Starting from the top, the Secure Processing Environments (SPE) is taken as the key capability that needs to be realized as part of the EHDS. SPE can be realized using different types of platforms, namely centralized, federated or hybrid. At the time of writing, we observe that there are already quite a few products available in the market, be it commercial or as an established open source project. The secure processing platforms (note we use platform to designate the business product vs. SPE to denote the capability) are composed of various value creation services as defined in the DSB2 (Figure 2). These services in turn are realized by underlying data analysis processes, which are envisaged to be supported by (Data) Service Providers.

Figure 1: Layered view of Secure Processing Environments (SPEs) as capability in the European Health Data Space.

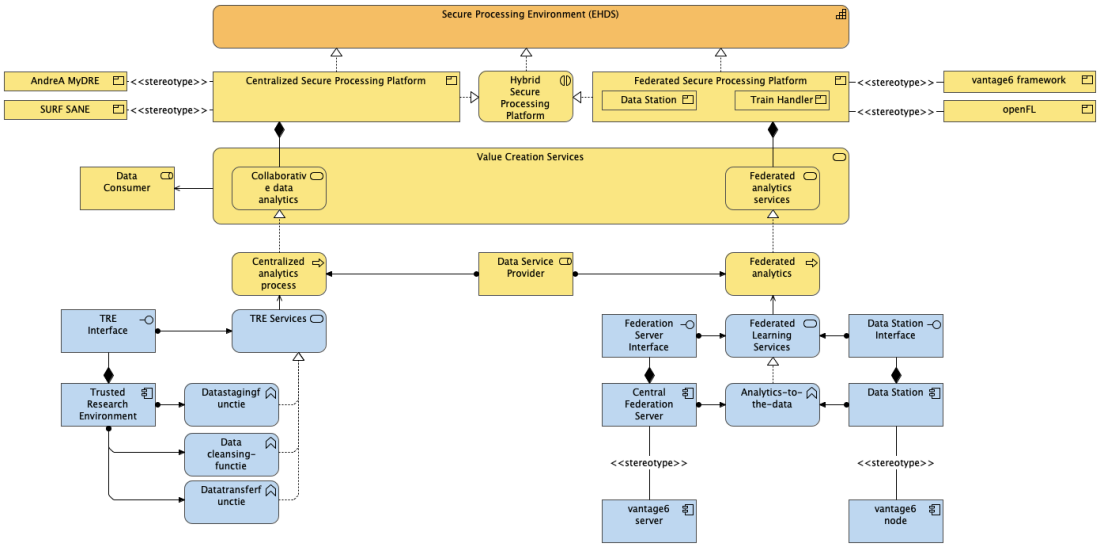
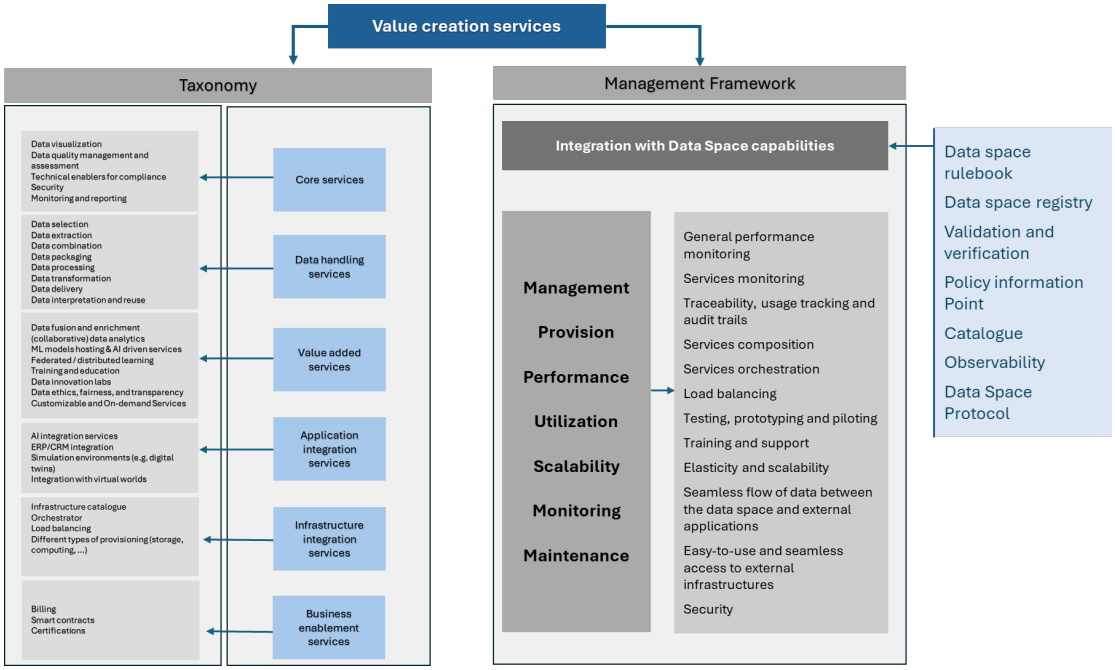


Figure 2: Taxonomy of value creation services as defined in the DSSC Blueprint 2.0. Source: dssc.eu.



2.6 Outline

With this overarching layered view in mind, we loosely follow an Action Design Research approach [31, 32] to address these design questions. Our main contributions are:

- A harmonized ontology of a data station and data hub, that integrates the PHT architecture [15, 23] and the DSSC Blueprint 2.0
- Comparative analysis of existing implementations
- Synthesis of the above into functional and technical description of a data station in Archimate, thereby focusing on two primary patterns [33]:
  - the layers pattern for addressing various aspects of interoperability across the stack, extending earlier work by Welten S, Arruda Botelho Herr M de, Hempel L, et al [34] who have introduced five layers of interoperability;
  - the pipes and filters pattern for addressing various solution designs for the extract-transform-load (ETL) mechanisms through which data stations are populated, following the current common practice of datalakes and lakehouse solution designs [35–39];
- Proposals for new designs for specific domains, such as genomics or imaging.
- A reference implementation of data stations using the trains based on containers as a generic infrastructure for federated learning and federated analysis.

Bibliography

1. Raasveldt M, Mühleisen H (2019) DuckDB: An Embeddable Analytical Database. In: Proceedings of the 2019 International Conference on Management of Data. ACM, Amsterdam Netherlands, pp 1981–1984

2. Nahrstedt F, Karmouche M, Bargieł K, Banijamali P, Nalini Pradeep Kumar A, Malavolta I (2024) An Empirical Study on the Energy Usage and Performance of Pandas and Polars Data Analysis Python Libraries. In: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering. ACM, Salerno Italy, pp 58–68

3. Wang Z, Ji H, Zhu Y, Wang D, Han Z (2025) A Survey on Federated Analytics: Taxonomy, Enabling Techniques, Applications and Open Issues. IEEE Communications Surveys & Tutorials 1. https://doi.org/10.1109/COMST.2025.3558755

4. Rieke N, Hancox J, Li W, et al (2020) The Future of Digital Health with Federated Learning. npj Digital Medicine 3(1):1–7. https://doi.org/10.1038/s41746-020-00323-1

5. Teo ZL, Jin L, Liu N, et al (2024) Federated Machine Learning in Healthcare: A Systematic Review on Clinical Applications and Technical Architecture. Cell Reports Medicine 5(2):101419. https://doi.org/10.1016/j.xcrm.2024.101419

6. (2023) The PET Guide

7. (2023) From Privacy to Partnership

8. Pedreira P, Erling O, Karanasos K, et al (2023) The Composable Data Management System Manifesto. Proceedings of the VLDB Endowment 16(10):2679–2685. https://doi.org/10.14778/3603581.3603604

9. The Composable Codex. https://voltrondata.com/codex.html. Accessed 16 Oct 2024

10. Prainsack B, El-Sayed S (2023) Beyond Individual Rights: How Data Solidarity Gives People Meaningful Control over Data. The American Journal of Bioethics 23(11):36–39. https://doi.org/10.1080/15265161.2023.2256267

11. Krewer J, Warso Z (2024) Digital Commons as Providers of Public Digital Infrastructures

12. Gu Z, Corcoglioniti F, Lanti D, et al (2022) A Systematic Overview of Data Federation Systems. Semantic Web 15(1):107–165. https://doi.org/10.3233/SW-223201

13. Beyan O, Choudhury A, Soest J van, et al (2020) Distributed Analytics on Sensitive Medical Data: The Personal Health Train. Data Intelligence 2(1–2):96–107. https://doi.org/10.1162/dint_a_00032

14. Choudhury A, Soest J van, Nayak S, Dekker A (2020) Personal Health Train on FHIR: A Privacy Preserving Federated Approach for Analyzing FAIR Data in Healthcare. In: Bhattacharjee A, Borgohain SK, Soni B, Verma G, Gao X-Z (eds) Machine Learning, Image Processing, Network Security and Data Sciences. Springer, Singapore, pp 85–95

15. Silva Santos LO Bonino da, Ferreira Pires L, Martinez V, Moreira J, Guizzardi R (2022) Personal Health Train Architecture with Dynamic Cloud Staging. SN Computer Science 4. https://doi.org/10.1007/s42979-022-01422-4

16. Zhang C, Choudhury A, Volmer L, et al (2023) Secure and Private Healthcare Analytics: A Feasibility Study of Federated Deep Learning with Personal Health Train. https://www.researchsquare.com/article/rs-3158418/v1. Accessed 24 Oct 2024

17. (2024) European Network of Trusted Research Environments (EOSC-ENTRUST) Project. https://eosc-entrust.eu/. Accessed 26 Jun 2025

18. Estrin D, Sim I (2010) Health Care Delivery. Open mHealth Architecture: An Engine for Health Care Innovation. Science (New York, NY) 330(6005):759–760. https://doi.org/10.1126/science.1196187

19. Mehl GL, Seneviratne MG, Berg ML, et al (2023) A Full-STAC Remedy for Global Digital Health Transformation: Open Standards, Technologies, Architectures and Content. Oxford Open Digital Health 1:oqad18. https://doi.org/10.1093/oodh/oqad018

20. Beck M (2019) On the Hourglass Model. Communications of the ACM 62(7):48–57. https://doi.org/10.1145/3274770

21. Schultes E (2023) The FAIR Hourglass: A Framework for FAIR Implementation. FAIR Connect 1(1):13–17. https://doi.org/10.3233/FC-221514

22. Silva Santos LO Bonino da, Guizzardi G, Prince Sales T (2022) FAIR Digital Object Framework. https://fairdigitalobjectframework.org/. Accessed 19 Jun 2025

23. Choudhury A, Volmer L, Martin F, et al (2025) Advancing Privacy-Preserving Health Care Analytics and Implementation of the Personal Health Train: Federated Deep Learning Study. JMIR AI 4(1):e60847. https://doi.org/10.2196/60847

24. Kleppmann M, Riccomini C (2026) Designing Data-Intensive Applications, 2nd Edtion (Early Release). O'Reilly

25. (2025) European Health Data Space Regulation (EHDS). https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en. Accessed 9 Jun 2025

26. Cascini F, Pantovic A, Al-Ajlouni YA, Puleo V, De Maio L, Ricciardi W (2024) Health Data Sharing Attitudes towards Primary and Secondary Use of Data: A Systematic Review. eClinicalMedicine 71:102551. https://doi.org/10.1016/j.eclinm.2024.102551

27. Data Spaces Blueprint v2.0. https://dssc.eu/space/BVE2/1071251457/Data+Spaces+Blueprint+v2.0+-+Home. Accessed 9 Jun 2025

28. Simpl Programme. https://simpl-programme.ec.europa.eu/. Accessed 9 Jun 2025

29. TEHDAS2. https://tehdas.eu/. Accessed 9 Jun 2025

30. Gorod A, Sauser B, Boardman J (2008) Paradox: Holarchical View of System of Systems Engineering Management. In: 2008 IEEE International Conference on System of Systems Engineering. pp 1–6

31. Sein M, Henfridsson O, Purao S, Rossi M, Lindgren R (2011) Action Design Research. MIS Quarterly 35:37–56. https://doi.org/10.2307/23043488

32. Venable J, Pries-Heje J, Baskerville R (2016) FEDS: A Framework for Evaluation in Design Science Research. European Journal of Information Systems 25(1):77–89. https://doi.org/10.1057/ejis.2014.36

33. Buschmann F, Meunier R, Rohnert H, Sommerlad P, Stal M (1996) A System of Patterns. Wiley

34. Welten S, Arruda Botelho Herr M de, Hempel L, et al (2024) A Study on Interoperability between Two Personal Health Train Infrastructures in Leukodystrophy Data Analysis. Scientific Data 11(1):663. https://doi.org/10.1038/s41597-024-03450-6

35. Harby AA, Zulkernine F (2025) Data Lakehouse: A Survey and Experimental Study. Information Systems 127:102460. https://doi.org/10.1016/j.is.2024.102460

36. Schneider J, Gröger C, Lutsch A, Schwarz H, Mitschang B (2024) The Lakehouse: State of the Art on Concepts and Technologies. SN Computer Science 5(5):449. https://doi.org/10.1007/s42979-024-02737-0

37. Hai R, Koutras C, Quix C, Jarke M (2023) Data Lakes: A Survey of Functions and Systems. IEEE Transactions on Knowledge and Data Engineering 35(12):12571–12590. https://doi.org/10.1109/TKDE.2023.3270101

38. AbouZaid A, Barclay PJ, Chrysoulas C, Pitropakis N (2025) Building a Modern Data Platform Based on the Data Lakehouse Architecture and Cloud-Native Ecosystem. Discover Applied Sciences 7(3):166. https://doi.org/10.1007/s42452-025-06545-w

39. Mazumdar D, Hughes J, Onofre JB (2023) The Data Lakehouse: Data Warehousing and More. http://arxiv.org/abs/2310.08697. Accessed 11 Jan 2024