International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 1, 209-216, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I1P135

Original Article

The Convergence of Data Virtualization and Federated Learning in Pharmaceutical Real-World Evidence (RWE) Generation: A Survey and Gap Research in Architectures, Tools, and Governance Challenges

Pinaki Bose
Independent Researcher, USA.

Received On: 18/01/2026 ***Revised On: 19/02/2026*** ***Accepted On: 21/02/2026*** ***Published on: 23/02/2026***

Abstract - *The generation of pharmaceutical Real-World Evidence (RWE) is a critical imperative for modern healthcare, yet it is hindered by two fundamental bottlenecks: (1) stringent privacy regulations, such as HIPAA, which preclude data centralization, and (2) pervasive, systemic data fragmentation within healthcare institutions.2 Federated Learning (FL) has emerged as the consensus solution for the privacy challenge, enabling model training on distributed data. Concurrently, Data Virtualization (DV) is the industry-standard solution for data fragmentation, providing a unified logical view of data silos. The current scientific literature, however, investigates these two solutions in parallel, failing to address the critical gap at their intersection. FL research implicitly operates on a "Unified Node Assumption," presuming each participating hospital has its data in a single, queryable repository, which is practically false. This paper bridges this gap by conducting a systematic survey of both domains and proposing the first architectural taxonomy of "Federated Learning on a Virtualized Data Layer" (FL-on-VD). Three novel architectural models are proposed and illustrated: (1) FL with Virtualized Query Pushing (FL-VQP), (2) FL on a Logical Data Mesh (FL-LDM), and (3) FL on a Centralized Virtual-View (FL-CVV). This paper analyzes the unique, second-order challenges this convergence creates in federated query optimization, semantic interoperability, and "double-blind" governance. It concludes that this unified FL-on-VD architecture represents the only viable and scalable path toward national-level RWE generation.*

Keywords - *Federated Learning, Data Virtualization, Real-World Evidence (Rwe), Health Informatics, Data Architecture, Data Governance, Omop, Data Mesh.*

1. Introduction: The RWE Imperative and Its Dual Bottlenecks

The paradigm of pharmaceutical and healthcare research is undergoing a fundamental shift, driven by the imperative to generate Real-World Evidence (RWE). RWE, derived from the analysis of Real-World Data (RWD), provides critical insights into treatment effectiveness, patient outcomes, and safety profiles outside the confines of traditional randomized controlled trials (RCTs). Regulatory bodies, including the U.S. Food and Drug Administration (FDA), and key policy organizations like the Duke-Margolis RWE Collaborative, are actively working to integrate RWE into regulatory decision-making, aiming to accelerate drug development and ensure public safety. This shift is embraced by a wide range of stakeholders, from biopharmaceutical companies to payers, all seeking to understand the practical, real-world value of therapies [5].

Despite this enthusiasm, the path to scalable, national-level RWE generation is obstructed by two fundamental, co-existing barriers: regulatory constraints on data movement and systemic data fragmentation [1].

1.1. Bottleneck 1: Privacy and Regulatory Constraints.

The primary barrier is legal and ethical. U.S. federal laws, most notably the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, along with global regulations like GDPR, impose stringent restrictions on the use and disclosure of protected health information (PHI). These regulations make the traditional analytics model centralizing sensitive, patient-level data from multiple institutions into a single repository—legally untenable, logistically complex, and fraught with risk. This represents the critical *inter-institutional* data-sharing problem.

1.2. Bottleneck 2: Data Fragmentation and Silos.

The second, often-overlooked barrier is the "hidden" challenge of *intra-institutional* data fragmentation. Even within a single, sophisticated U.S. health system, patient data is not located in a single repository. It is fragmented across dozens of disconnected, heterogeneous systems: Electronic Health Records (EHRs), Laboratory Information Management Systems (LIMS), radiology archives storing DICOM images, pharmacy records, and financial billing systems [2]. These data silos, often a result of legacy systems

and departmental boundaries, prevent the creation of a comprehensive patient record even before the challenge of sharing it is considered [3].

To solve these distinct challenges, the industry has adopted two powerful, yet separate, technological solutions. For Bottleneck 1 (privacy), Federated Learning (FL) has emerged as the widely accepted privacy-preserving paradigm, allowing for collaborative analysis without centralizing data [4]. For Bottleneck 2 (fragmentation), Data Virtualization (DV) has become the industry-standard architecture for integrating distributed data silos without physical replication [6].

This paper's novel thesis is that the true, un-surveyed challenge is not implementing FL *or* DV, but implementing Federated Learning *on top of* a Virtualized Data Layer (FL-on-VD). The current literature exists in two parallel streams, failing to address the practical reality that any FL implementation at a real-world hospital must first confront that hospital's internal data fragmentation. This paper provides the first systematic survey and architectural taxonomy to address this critical gap.

2. A Systematic Review of Federated Learning in Healthcare Informatics

Federated Learning (FL) has been rapidly adopted in healthcare informatics as the technical solution to the privacy bottleneck (Bottleneck 1). In this paradigm, a global machine learning model is trained collaboratively across multiple decentralized clients, such as hospitals or research institutions [8]. Critically, the raw patient data never leaves the client's firewall. Instead, each client trains a local version of the model on its private data, and only the resulting model updates (e.g., weights or gradients) are sent to a central aggregator. The aggregator updates the global model, which is then sent back to the clients for the next iteration [10].

This approach has seen prolific application in biomedical research and RWE generation. Studies have demonstrated its use for predictive modeling from EHR data , analyzing medical images, and supporting multi-center clinical studies. A key emerging application is federated causal inference, which allows researchers to estimate treatment effects and conduct post-marketing surveillance across diverse populations, a cornerstone of RWE, without violating patient confidentiality.

2.1. The Critical Gap: The "Unified Node Assumption"

Despite its focus on inter-institutional privacy, the existing FL literature harbors a critical, flawed premise. The academic and research communities have focused intensely on the problem of "statistical heterogeneity". This is the challenge that the data at each node is not independent and identically distributed (non-IID); for example, patient demographics, diagnostic practices, and outcome prevalence may differ significantly between hospitals [4]. Algorithms like Federated Averaging (FedAvg) and its derivatives are designed to achieve model convergence despite this statistical variance.

However, this focus on statistical heterogeneity has created a massive blind spot: it completely ignores the reality of "infrastructural heterogeneity**.**" The FL literature implicitly operates on what this paper formally defines as the "**Unified Node** Assumption": the assumption that each participating node (e.g., a hospital) has its local data available in a single, clean, integrated, and queryable repository.

This assumption is demonstrably false. As discussed in Section 1, a hospital's data is a fragmented mess of silos. Therefore, the current generation of FL frameworks is being designed for a data environment that does not exist in the real world. Before a hospital can even participate in a federated network, it must first solve its own internal data fragmentation problem. The solution to this problem, Data Virtualization, is explored in the next section.

3. A Systematic Review of Data Virtualization in Big Data Architectures

Data Virtualization (DV) is the established, industrial-strength solution to the "Unified Node Assumption" identified in Section 2. DV is a modern data integration architecture that creates a logical, abstract data layer, or "abstraction layer" [12]. This layer provides a unified, semantic view of data from multiple, distributed, and heterogeneous physical sources (e.g., databases, data lakes, APIs, and legacy systems) *without* requiring the costly, complex, and time-consuming process of physical data movement or replication (ETL) [6].

In modern big data architectures, DV functions as the enabling technology for "logical data warehouses" and "data fabric," accelerating self-service analytics [14]. It decouples the analytics and consumption layers (e.g., BI tools, ML models) from the complex, fragmented back-end systems, allowing analysts to access and query data as if it were in a single location [15].

3.1. Case Study 1: The Data Virtualization Platform like Denodo in Healthcare

The Data Virtualization platform like Denodo serves as a prime example of DV applied to high-stakes, regulated industries like healthcare [7]. Healthcare organizations use such DV technologies to create 360-degree views of patients by logically combining data from fragmented systems. This enables the creation of a "logical data warehouse" that can serve diverse stakeholders from clinicians needing a full patient history to administrators analyzing payment structures.

Critically, the DV platform is not just an integration tool but also a governance-enforcement layer. It is specifically designed to enhance HIPAA compliance by providing a single point of control for managing privacy and security. It offers granular, role-based access controls for clinical data and provides data lineage information, which is essential for audits and protecting against breaches. This proves that DV technology is already mature and trusted for managing sensitive, fragmented PII.

3.2. Case Study 2: Oracle Big Data SQL

Oracle Big Data SQL provides a deep-dive technical example of DV's advanced capabilities. Oracle explicitly markets this technology as a "data virtualization innovation".[34] Its architecture is designed to allow a single, standard SQL query to seamlessly and performantly join data across vastly different systems, including the Oracle Database, Apache Hadoop, Apache Kafka, and NoSQL databases [16].

The key technical mechanism that enables this performance is "Smart Scan" and "predicate push-down". When a user issues a query, the DV platform does not pull all the raw data from the sources to be joined centrally. Instead, it "pushes down" the computational logic of the query (e.g., the WHERE clause predicates) directly to the source systems. Each source system processes its fragment of the query locally, using its own compute resources, and returns *only* the filtered, relevant data to the Oracle engine for final aggregation.

This "predicate push-down" mechanism is a profound architectural concept. It is a form of distributed computation that minimizes data movement. This mechanism is conceptually identical to the core requirement of a Federated Learning client: performing local computation and returning only the result. This capability forms the technical foundation for the architectural models proposed in the next section.

4. The Architectural Nexus: A Proposed Taxonomy of "FL-on-VD" Models

The findings from Sections 2 and 3 establish a clear, unaddressed gap: FL requires a unified node, and DV creates a unified node. Therefore, a real-world architecture must integrate them. Industry-watch reports and technical documentation allude to this convergence, often listing FL frameworks (like IBM Federated Learning or NVIDIA FLARE) and DV platforms (like Denodo [18] or Trino/Starburst [20]) as components of the same "modern data stack." However, no formal survey or technical paper has defined the specific architectural patterns for *how* these components must integrate.

This section fills that gap by proposing the first taxonomy of "Federated Learning on a Virtualized Data Layer" (FL-on-VD) models.

4.1. Model 1: FL with Virtualized Query Pushing (FL-VQP)

This is the most computationally sophisticated model, directly leveraging the "predicate push-down" capabilities of advanced DV platforms and the concepts of "federated query processing" [21].

4.1.1. Architectural Block Diagram (Conceptual)

- A Global FL Aggregator sends a model-training request to a Local FL Client (e.g., a PySyft worker [22]).
- The Local FL Client does not query a simple table. Instead, it issues a complex computational request (e.g., "calculate gradients for these features").
- This request is intercepted by an advanced federated query optimizer or DV platform[17].
- The DV platform's optimizer translates this computational request into an optimized, federated query plan.
- The DV engine "pushes" query fragments down to the physical data silos (e.g. EHR, LIMS).
- The silos perform local computation, and the DV engine aggregates the results.
- Only the final aggregated result (e.g., the gradients) is returned to the Local FL Client.

Explanation: In this model, the DV layer acts as an active, distributed query and compute engine. The raw data *never* moves, not even to be joined in a single virtual table. The computation is pushed all the way down to the physical sources. This model maximizes privacy and minimizes data movement but places a heavy burden on the DV platform's federated query optimizer.

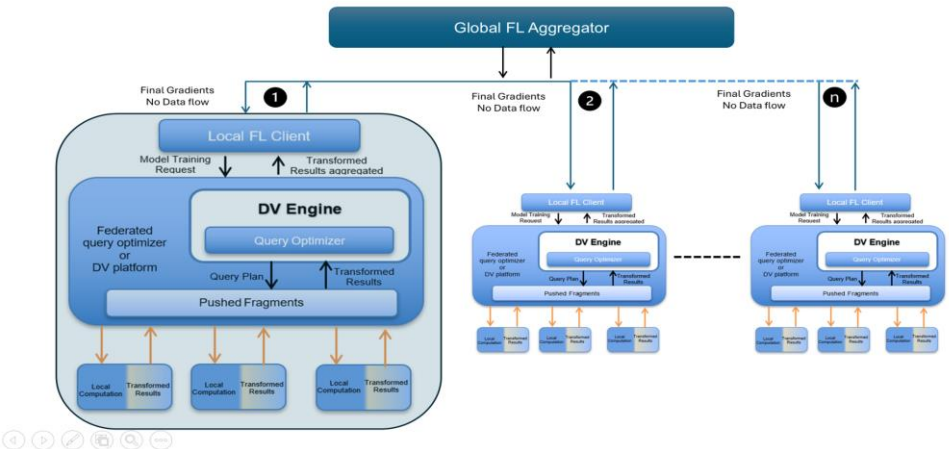

Fig 1: Model 1: FL with Virtualized Query Pushing (FL-VQP)

4.2. Model 2: FL on a Logical Data Mesh (FL-LDM)

This model adopts a decentralized, domain-driven architecture based on Data Mesh principles, as conceptualized in frameworks like "HealthMesh".

4.2.1. Architectural Block Diagram (Conceptual)

- A Global FL Aggregator communicates with a central Federated Computational Governance Layer.
- This layer does not talk to one "node" but to multiple, autonomous “Data Products”.
- One Data Product contains its own Virtualized View of a particular domain area (e.g. Clinical Data view) and its *own* Local FL Compute Engine.
- Another Data Product contains its own Virtualized View of another particular domain area (e.g. Billing Data view) and its *own* Local FL Compute Engine.
- Explanation: This architecture is decentralized both in its data and its governance. Each "Data Product" is a self-contained, sovereign unit responsible for its own data, virtualization, compute, and governance. The "Federated Computational Governance Layer" acts as a policy-enforcement and routing hub. When the aggregator requests a training step, the governance layer routes this request to the relevant, independent Data Products, which then execute the training on their virtualized data locally and return their results. This model, explicitly designed to enable and enhance secure analytical tasks... including Federated Learning, scales well organizationally but requires significant governance maturity.

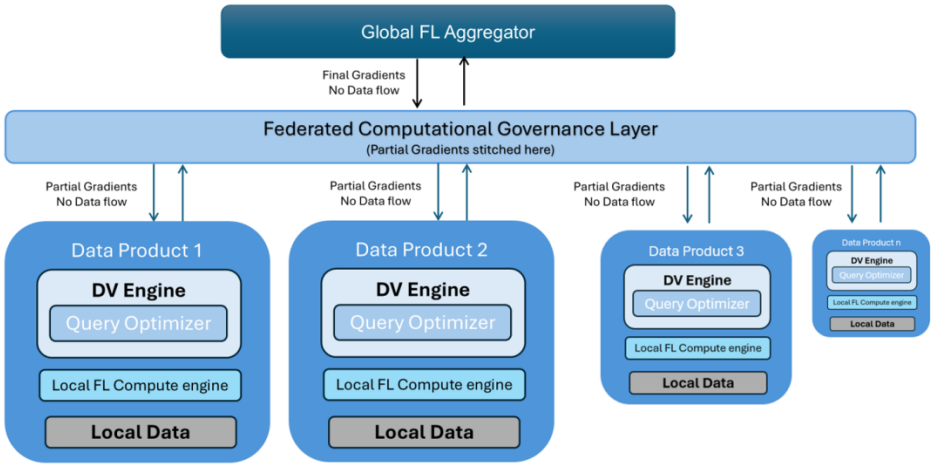

Fig 2: Model 2: FL on A Logical Data Mesh (FL-LDM)

4.3. Model 3: FL on a Centralized Virtual-View (FL-CVV)

This is the most straightforward and likely most common architecture. It treats the DV platform as a "black box" that creates a single, unified data source for the FL framework.

4.3.1. Architectural Block Diagram (Conceptual)

- A Global FL Aggregator sends a request to a Local FL Client.
- The Local FL Client is configured to see *one* data source: a single unified data source.
- The client issues a simple query (e.g., SELECT * FROM VIRTUAL_OMOP_TABLE).
- The DV layer intercepts this query. Its internal optimizer handles all the complex, real-time joining and harmonization
- A single, unified, virtual table is streamed to the Local FL Client for local training.

Explanation: This model leverages a DV platform like Denodo to create a "Logical Data Warehouse". The FL framework (eg., NVIDIA FLARE or IBM FL) is completely unaware of the underlying fragmentation. It simply queries a clean, virtual database. This model maximizes simplicity for the data science and FL teams, as the DV platform's administrators handle all the complex data integration work. This architecture is directly implied by platforms that list connectors for both Denodo and FL frameworks [19].

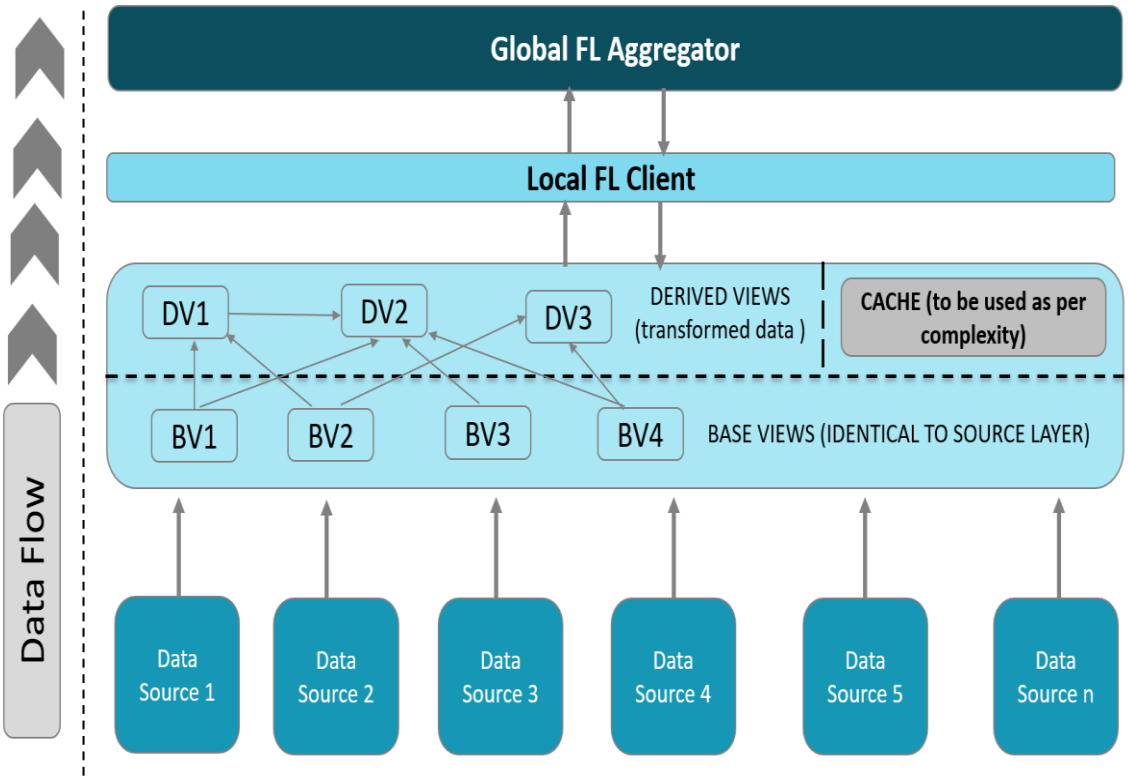

Fig 3: FL on A Centralized Virtual-View (FL-CVV)

Table 1: A Taxonomy of Proposed FL-On-VD Architectural Models

Model	Core Principle	Query/Compute Flow	Governance Model	Key Challenge	Example Technologies
FL-VQP	DV as Active Query Engine	FL client issues a computation; DV "pushes" query logic down to physical silos.	Centralized DV query governance; Federated FL governance.	Complex federated query optimization.	Trino/Starburst, Oracle Big Data SQL , PySyft
FL-LDM	Decentralized Data Products	FL aggregator requests; autonomous Data Products execute locally.	Federated computational governance layer.	High organizational complexity; ensuring semantic consistency.	"HealthMesh"
FL-CVV	DV as Black-Box View	FL client queries a single, unified virtual database. DV platform handles all complexity.	Centralized DV access control ; FL governance.	Potential performance bottlenecks; semantic mapping effort.	Denodo, IBM FL , NVIDIA FLARE.

5. Analysis of FL-On-VD Convergence Challenges

The synthesis of Federated Learning and Data Virtualization into a unified FL-on-VD architecture solves the dual bottlenecks of privacy and fragmentation. However, this convergence introduces new, second-order challenges that must be addressed.

5.1. The Federated Query Optimization Problem

In a traditional FL setting, the only optimizer is the FL algorithm itself, which is concerned with the speed and stability of global model convergence [11]. In a DV setting, the optimizer is concerned with query performance, network latency, and join strategies.

In an FL-on-VD architecture, these two optimizers are "stacked." This creates a new, complex problem: How does the DV platform's query plan (e.g., which joins to run first, which data to cache) affect the statistical properties of the data batch returned to the FL client? And how does that, in turn, affect the FL model's convergence rate and final accuracy? This is a significant open research question. The development of "adaptive query optimization techniques for virtual data layers" and ML-based optimizers that learn the

performance of underlying systems as "black boxes" will be critical for the viability of high-performance models like FL-VQP.

5.2. Semantic Interoperability: The Critical Role of CDMs

This is perhaps the most critical and non-obvious challenge and benefit of the FL-on-VD architecture. Federated Learning requires semantic interoperability. For a global model to be meaningful, all participating nodes must agree on the definition, format, and vocabulary of the data. The "blood_pressure" feature from Hospital A must mean the same thing as the "blood_pressure" feature from Hospital B [9].

The physical data silos that the DV layer connects to are semantically heterogeneous. The DV layer's true, higher-order purpose in this architecture is not just data access, but semantic harmonization. The DV platform functions as the "semantic layer" [13]. It is the ideal place to perform the on-the-fly mapping, transformation, and harmonization from disparate, proprietary source-system schemas into a Common Data Model (CDM).

For RWE, the de facto standard is the Observational Medical Outcomes Partnership (OMOP) CDM, which is maintained by the OHDSI collaborative [23]. Therefore, the most robust FL-on-VD architecture (particularly Model 3, FL-CVV) solves the semantic problem: the DV platform (e.g., Denodo) implements the complex mappings to create a "VIRTUAL_OMOP_TABLE". The FL client only queries this virtual, pre-harmonized OMOP layer, ensuring semantic consistency across the entire federation.

5.3. The "Double-Blind" Governance Problem

Finally, the FL-on-VD architecture creates a multi-layered governance complexity. This is not a single governance model, but two distinct layers that must co-exist:

- DV Governance (Local Layer): This is the intra-institutional governance, managed by the hospital's data officer. It defines data ownership, enforces data quality standards, and, most importantly, manages the fine-grained access rules that "blind" the FL client to any PII. This layer satisfies the local HIPAA compliance officer.
- Federated Governance (Global Layer): This is the inter-institutional governance, managed by the research consortium. It defines what models can be trained, what aggregated results can be shared, and ensures the "institutional data sovereignty" of each member. This layer satisfies the consortium's legal and ethical framework.
- This paper terms this the "double-blind" governance model. This two-layer structure should not be seen as a burden, but as a critical feature. It provides a robust, defense-in-depth framework for privacy and security. It creates a practical and auditable system that addresses the concerns of both the local, risk-averse data owner and the global, collaborative research network.

Table 2: Analysis of Key Convergence Challenges In Fl-On-Vd

Challenge	Impacted Layer	Problem Description	Mitigation Strategy / Solution
Federated Query Optimization	DV Platform & FL Aggregator	"Stacked" optimizers: DV query plan impacts FL model convergence.	ML-based query optimizers that treat silos as black boxes; adaptive FL algorithms.
Semantic Interoperability	Virtualization Layer	FL requires harmonized data. Physical silos are heterogeneous.	Implement a Common Data Model (e.g., OMOP) *within* the DV semantic layer.
"Double-Blind" Governance	Both Layers	Must satisfy both local (HIPAA) and global (federation) rules.	A two-layer model: (1) Local DV governance for PII access ; (2) Global Federated Governance for model sharing.
Performance & Latency	Virtualization Layer	Real-time federated queries (FL-VQP) or virtual view generation (FL-CVV) can be slow.	Intelligent caching in the DV platform, query optimization , and selecting the right architecture (Model 1 vs 3) for the use case.

6. Conclusion: A Unified Architecture for National-Scale RWE

The generation of Real-World Evidence is at a crossroads. The healthcare industry's goals are ambitious, with thought leaders and policy groups envisioning a future of patient-centric drug development and large-scale, international research collaborations. Organizations like the Duke-Margolis RWE Collaborative are actively building the policy frameworks for a national learning health system. However, these goals are unattainable if we only solve part of the problem. This paper has systematically demonstrated that the two primary bottlenecks to RWE inter-institutional privacy concerns and intra-institutional data fragmentation are currently being addressed by two separate, un-integrated technologies.

Federated Learning (FL), as detailed in Section 2, is an incomplete solution. It ably addresses privacy but fails by operating on the "Unified Node Assumption"—a flawed premise that ignores the fragmented reality of hospital data. Data Virtualization (DV), as detailed in Section 3, is also an incomplete solution. It masterfully solves internal fragmentation but does not, by itself, provide the framework for secure, inter-institutional model training.

The "Federated Learning on a Virtualized Data Layer" (FL-on-VD) architectural taxonomy proposed in Section 4 provides the first comprehensive blueprint for solving the *complete, real-world problem*. By unifying these two technologies, the FL-on-VD architecture creates a system where:

- A virtualized layer solves the "Unified Node Assumption" by harmonizing internal silos into a single, logical, query able source.
- This same virtualized layer functions as the "semantic layer," mapping heterogeneous data to a Common Data Model like OMOP to ensure federated interoperability.
- The FL framework can then securely and privately train models on this logical, harmonized layer, satisfying the privacy requirements of all participants.
- A "double-blind" governance model ensures that both local HIPAA officers and global research consortiums have their security and sovereignty requirements met.

This unified architecture, in one of the forms presented in this paper's taxonomy, is the critical enabler for the future of RWE. It is the only practical, scalable, and secure pathway to bridge the gap from fragmented, siloed data to the collaborative, national-scale learning health system that patients and researchers are waiting for.

References

[1] DelveInsight, "Artificial intelligence in drug commercialization: Accelerating market success through data-driven precision," DelveInsight Blog, 2024. [Online]. Available: https://www.delveinsight.com/blog/artificial-intelligence-in-drug-commercialization

[2] N. D. Heiger, C. R. Thompson, and J. S. Brown, "The current landscape and emerging applications for real-world data in diagnostics and clinical decision support and its impact on regulatory decision making," Clin. Pharmacol. Ther., vol. 113, no. 1, pp. 31–36, Jan. 2023, doi: 10.1002/cpt.2783.

[3] M. R. Al-Zahrani and M. M. S. Al-Majeed, "Convergence of integrated sensing and communication (ISAC) and digital-twin technologies in healthcare systems: A comprehensive review," Healthcare, vol. 6, no. 4, Art. no. 51, 2024, doi: 10.3390/healthcare6040051.

[4] S. Zhang et al., "Federated causal inference in healthcare: Methods, challenges, and opportunities," arXiv preprint arXiv:2505.02238, 2025.

[5] Duke-Margolis Institute for Health Policy, "Real-world evidence," Duke-Margolis Healthcare Topics, 2024. [Online]. https://healthpolicy.duke.edu/topics/real-world-evidence

[6] P. K. Suri and P. Singh, "A theoretical exploration of data management and integration in organization sectors," Int. J. Data Mining Knowl. Manag. Process, vol. 11, no. 1, pp. 31–45, Jan. 2021, doi: 10.5121/ijdms.2019.11103.

[7] Denodo Technologies, "Healthcare data management: Modernizing healthcare with data virtualization," 2024. [Online]. https://www.denodo.com/en/solutions/by-industry/healthcare.

[8] A. Sharma et al., "An advanced data fabric architecture leveraging generative AI and metadata-driven automation," arXiv preprint arXiv:2402.09795, 2024.

[9] L. Wang et al., "Ontology- and LLM-based data harmonization for federated learning in healthcare," arXiv preprint arXiv:2505.20020, 2025.

[10] TEHDAS, "Report on EHDS architecture and infrastructure implementers expectations and experiences," Joint Action Towards the European Health Data Space, Rep., 2022. [Online]. https://tehdas.eu/app/uploads/2022/06/tehdas-report-on-ehds-architecture-and-infrastructure-implementers-expectations-experiences.pdf

[11] M. A. Alzahrani, "Framework of big data analytics in real time for healthcare enterprise performance measurements," Ph.D. dissertation, Dept. Industrial Eng. & Management Syst., Univ. Central Florida, Orlando, FL, USA, 2021.

[12] R. S. S. S. Prasad, "The best practice of big data architecture in a health care organization," ResearchGate, Oct. 2014. [Online]. https://www.researchgate.net/figure/The-best-practice-of-big-data-architecture-in-a-health-care-organization_fig1_266613537

[13] Orion Innovation, "Data virtualization: Accelerating self-service analytics with a unified semantic layer," Orion Case Studies, 2023. [Online]. https://www.orioninc.com/case-studies/accelerating-self-service-analytics-with-a-unified-semantic-layer/

[14] enodo Technologies, "Data virtualization can deliver ROI of 408% according to new independent research study," Denodo Press Release, Nov. 30, 2021. [Online]. https://www.denodo.com/en/press-release/2021-11-30/data-virtualization-can-deliver-roi-408-according-new-independent-research

[15] Oracle Corporation, "Oracle big data SQL," Oracle Datasheet, 2020.

[16] Online].https://www.oracle.com/docs/tech/database/bigdatasql-datasheet.pdf

[17] Capgemini, "TechnoVision 2024: CTIO report," Capgemini Technical Report, 2024. [Online]. https://www.scribd.com/document/733241240/TechnoVision-2024-CTIO-Report-Web-Version

[18] Gathr.ai, "The silent revolution of invisible AI," Gathr Blog, 2024. [Online]. https://www.gathr.ai/blog/the-silent-revolution-of-invisible-ai/

[19] World Bank, "Harnessing data for better lives," World Development Report 2021, World Bank Group, Washington, DC, USA, 2021. [Online]. https://openknowledge.worldbank.org/handle/10986/35218

[20] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Federated query processing for big data in data science," in Proc. IEEE 6th Int. Conf. Big Data Comput. Serv. Appl. (BigDataService), 2020, pp. 115–118, doi: 10.1109/BigDataService49289.2020.00025.

[21] A. B. Researcher et al., "Rethinking pluggable federated query optimization: From laptops to data warehouses," in Proc. VLDB Workshops 2025 (CDMS), 2025. [Online]. https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/CDMS/CDMS25_07.pdf.

[22] Horizon-Trustee Project, "D2.1 Live doc conceptualisation, use cases and system architecture V1," EU Horizon Europe Deliverable, 2024. [Online]. https://horizon-trustee.eu/wp-content/uploads/2024/06/D2.1-Live-doc-conceptualisation-use-cases-and-system-architecture-V1.pdf.

[23] IBM Cloud, "Denodo connection," IBM Cloud Pak for Data Documentation, 2024. [Online]. https://eu-gb.dataplatform.cloud.ibm.com/docs/content/wsj/manage-data/conn-denodo.html.

[24] Journal Press, "Federated data governance for cross-institution anti-money laundering," London J. Eng. Res., vol. 25, 2024. [Online]. https://journalspress.com/LJER_Volume25/Federated-Data-Governance.pdf.