VIEW POINT

RETHINKING PHARMA DATA COLLABORATION: A DATA MANAGEMENT PRIMER ON FEDERATED LEARNING

Infosys®
Navigate your next

In the age of data-driven pharmaceutical innovation, managing sensitive healthcare data while enabling meaningful collaboration remains a complex challenge. With rising demands for real-world evidence, faster drug development, and compliance with stringent privacy regulations like GDPR and HIPAA, the pharmaceutical industry must reimagine how it shares and utilizes data. The traditional approaches to data sharing simply aren't cutting it. It's clear: the industry must fundamentally reimagine its data strategy. This is precisely where **Federated Learning (FL)** emerges as a game-changer. It allows organizations to collaboratively train sophisticated AI models without ever centralizing sensitive datasets, effectively bridging the chasm between collaboration and privacy.

Understanding Federated Learning's Power

FL is a decentralized machine learning approach where data remains localized, and only model updates are transmitted to a central server. This method ensures that raw patient data never leaves its origin, mitigating privacy risks and regulatory concerns. Instead of creating massive, centralized datasets, FL enables local model training across institutions like hospitals, research centers, and pharma labs.

Core Principles of FL

Decentralized Training: FL trains models locally, keeping data on devices and sending only model updates.

Secure Aggregation: A central server aggregates model updates, not raw data, to build a global model.

Privacy Preservation: FL ensures sensitive patient data remains local, minimizing breach risks.

Iterative Refinement: Continuous model improvement through repeated local training and global aggregation.

Key technologies powering FL

Secure Aggregation: Encrypts updates during aggregation, revealing only the combined result, ensuring individual privacy.

Differential Privacy (DP): Adds calibrated noise to data or updates, preventing individual data inference while maintaining model accuracy.

Homomorphic Encryption (HE): Enables computation on encrypted updates, allowing secure aggregation without decryption until the result.

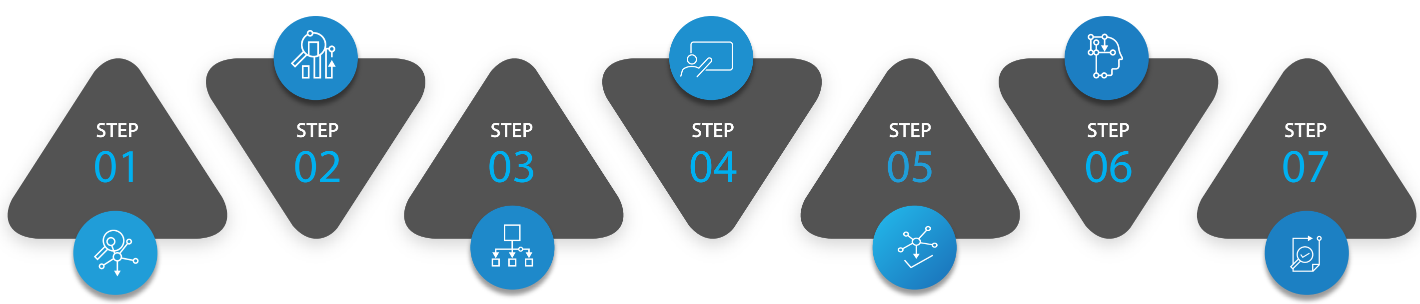
Key Types of FL

Horizontal FL: Same features, different data subjects (e.g., multiple hospitals with similar patient data).

Vertical FL: Same data subjects, different features.

Federated Transfer Learning: Limited overlap in both features and subjects; used for knowledge transfer.

FL operates through **federated round,** which refers to an iteration of the training process where a model is trained across distributed local nodes and then aggregated at a central server. Below is an illustration of how the federated learning process works:



Initialization

The central server selects and distributes a machine learning model (e.g., linear regression, neural network) to local devices, which wait for tasks.

Client Selection

A subset of local devices is chosen to participate in training, while others remain idle until the next round.

Configuration

The server instructs selected devices on training specifics (e.g., mini-batch size, optimization algorithm) using local data, ensuring privacy.

Local Training

Devices train the model on their local data, improving the model without sharing raw data.

Model Update

After training, devices summarize changes and send encrypted updates back to the central server.

Aggregation

The server aggregates updates from selected devices, refines the global model, and sends it back for the next round.

Termination

The process repeats until a stopping criterion is met, with the model continuously improving through iterative training and real-time updates.

Reshaping Data Management: The Federated Learning Impact

FL fundamentally shifts the landscape of data management, moving from centralized control to a distributed, collaborative approach. This has profound implications across various dimensions:

Data Quality and Interoperability

Governed Consistency: Quality checks, standardized schemas (e.g., FHIR, OMOP CDM), and shared governance ensure reliable model inputs.

Robust Preprocessing: Advanced techniques support fair and accurate model training across sources.

Sharing and Collaboration

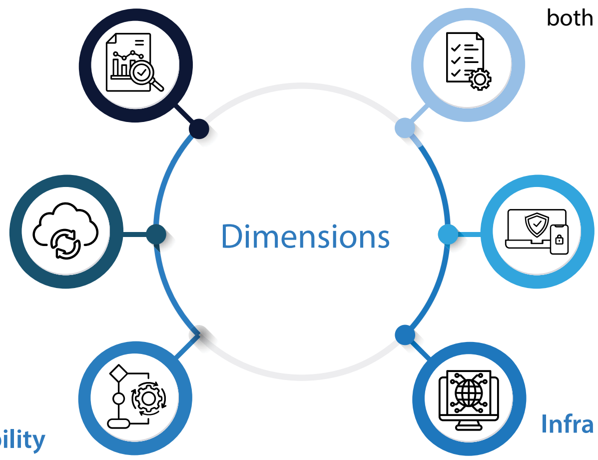
Breaking Down Data Silos: FL breaks data silos, enabling access to previously unavailable data for collaborative training.

Collaborative Model Development: FL enables multi-stakeholder research without data centralization.

Lifecycle Management and Traceability

Local Stewardship: Institutions manage their own data according to internal policies.

Model Lifecycle Management: Requires versioning, update security, and audit trails for full traceability.



Governance and Compliance

Decentralized Control: Data remains within source institutions, easing compliance with GDPR, HIPAA, and local laws

Clear Accountability: Requires formal agreements, audit trails, and clarity around both data and model ownership.

Security and Privacy

Privacy-by-Design: FL shares model updates—not raw data—while employing techniques like differential privacy, homomorphic encryption, and secure aggregation.

Integrity Assurance: Local and global validations preserve data quality across nodes.

Infra and Architecture

Distributed Processing: FL demands scalable, decentralized compute systems.

Edge Computing Integration: Compatible with edge computing, reducing reliance on central servers.

Data Standardization: Handles heterogeneous data via preprocessing, transformation, and feature harmonization.

The tabulation below highlights the distinct advantages of FL over conventional centralized or purely decentralized models, providing clear comparative analyses that underscore its unique benefits.

Aspect	Centralized Learning	Traditional Decentralized Learning	Federated Learning (FL)
Data Sharing	Requires full data centralization	No data sharing or coordination	Shares only model updates; no raw data transfer
Privacy & Compliance	High risk of data exposure; complex regulatory compliance	High privacy, but lacks coordination	Strong privacy by design; aligns with GDPR, HIPAA
Collaboration	Limited due to data silos and IP concerns	Rare, due to lack of integration	Enables secure, multi-party collaboration across institutions
Model Accuracy	May lack generalization if data is biased or siloed	Varies; may suffer from inconsistent training	High generalizability via diverse, distributed data
Infrastructure Needs	Requires large, centralized storage and compute	Low; independent systems only	Requires distributed compute with secure aggregation
Governance & Auditability	Central authority controls governance	Limited visibility or coordination	Shared governance with audit trails and data lineage
Adaptability to Pharma	Challenging due to IP sensitivity and patient confidentiality	Not well-suited for regulated, collaborative environments	Ideal for pharma due to privacy, regulatory alignment, and secure collaboration

The Road Ahead

Federated Learning isn't just redefining data management in pharma; it's unlocking the next frontier of innovation. For any pharmaceutical company serious about modernizing its data strategy, FL isn't merely a compelling path forward – it's an imperative. As the Pharma sector rapidly becomes more data-centric, embracing Federated Learning means more than just adopting new tech; it's about harnessing our collective intelligence to drive truly scalable, compliant, and profoundly impactful advancements in drug development and, most importantly, in patient care.

About the Authors

Saarthak Gupta

Consultant, ICLS, Infosys Consulting

Saarthak is a consultant in Infosys Consulting's Life Sciences practice within the LS Data Transformation Team with more than 6 years of professional experience. He has experience in Life Sciences, Healthcare, Information Technology and Ed-Tech industries. He has worked in multiple engagements on Data Analytics, Data Migration and Transformation, Data Quality Management, Reference Data Management and Net Revenue Management.

Pragnya Koya

Consultant, ICLS, Infosys Consulting

Pragnya is a Business Consulting professional with 4+ years of experience in the Healthcare and Life Sciences sectors. She brings a robust blend of expertise in product development, performance reporting and improvement, and customer experience enhancement, consistently leveraging strong data management and analytical skills to drive tangible results.

Nitisha Nitin Patil

Analyst, ICLS, Infosys Consulting

Nitisha is an experienced Analyst with a strong background in the Healthcare and Life Sciences industries. She brings expertise in stakeholder management, data analysis, and business process understanding. Her proficiency spans stakeholder engagement, requirement analysis, data interpretation, and cross-functional collaboration to deliver impactful business solutions.

Ramya Gunza

Principal Consultant, ICLS, Infosys Consulting

Ramya is a seasoned Principal Business Consultant with around 10 years of experience within the Healthcare and Life Sciences industries, possessing deep expertise in data management, data quality, business intelligence and AI. Her proficiency extends to Data Office initiatives, Commercial Excellence strategies, and delivery of impactful data-driven solutions.

For more information, contact askus@infosys.com

Infosys®
Navigate your next