ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

Enterprise Data Harmonization Engine

Sophie Dubois
Asst Professor

Abstract

Federated insight engineering offers an industry-agnostic approach to federated analytics by exploiting a multi-tiered architecture for enterprise data within the larger FORGE methodology. The concept is rooted in the view of enterprise data landscapes as a collection of data sources susceptible to local analysis and insight extraction. Data sharing between these sources occurs where partners have a common business interest, and these ephemeral partnerships are governed by collaboration agreements.

Few industries have yet mastered the art of federated insight engineering. However, regulatory constraints and the privacy paradox mean that lessons learned in heavily regulated environments are applicable—often with minor adjustments—to other domains confronting different but equally pressing motivations. Tourism & hospitality, online marketplace, and finance & insurance exemplify the early movers in federated insight engineering within the analytics supply chain. Industry exemplars highlight not only common patterns of success and failure but also a reference architecture for federated analytics at the paradigm level.

Keywords : Discoverability: insight engineering; business intelligence; multi-tier architecture; data harmonization; enterprise data ecosystem; federated analytics; federated learning; data governance; metaphysics; ethics.

1. Introduction

Federated Insight Engineering enables organizations to leverage data from diverse sources across multiple administrative domains to discover insights, often by combining machine learning (ML) and analytic models. Industry examples underscore the practical value and feasibility of such approaches, offering lessons learned, key success factors, and actionable recommendations. Federated Insight Engineering is distinct from commonly used decentralization, distribution, federation, and cloud computing terms yet retains their underlying meanings. Federation provides an insightful perspective for the effective engineering of data that fuels analytics and ML at all scales. Data management, as a discipline, encompasses centralized, distributed, and federated strategies; the transition to and support for federation within Data Management functions at local, domain, enterprise, business ecosystem, national, and global scales warrant deeper investigation.

Industry implementations of Federated Insight Engineering span multiple economic sectors, prompting consideration of common architectural elements, operational patterns, governance frameworks, and ethical dimensions. Quantitative performance assessments and qualitative experience reports inform the identification of cross-industry lessons learned and best practices, encompassing both success factors and challenges. The observed industry breadth also facilitates comparisons with the insights drawn from other recent Federated Insight Engineering initiatives in academic and research contexts. Such analyses yield concepts, techniques, patterns, and recommendations that may benefit its application in any domain, industry, or business ecosystem, while also supporting similar endeavors in other contexts.

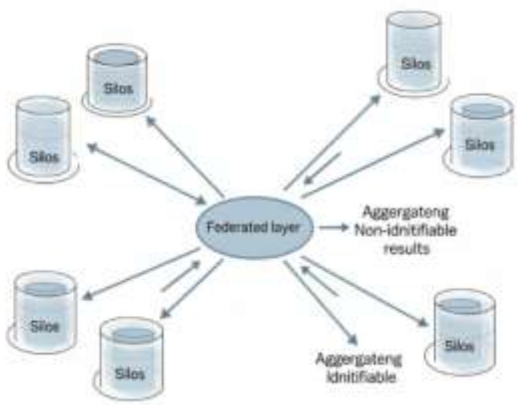
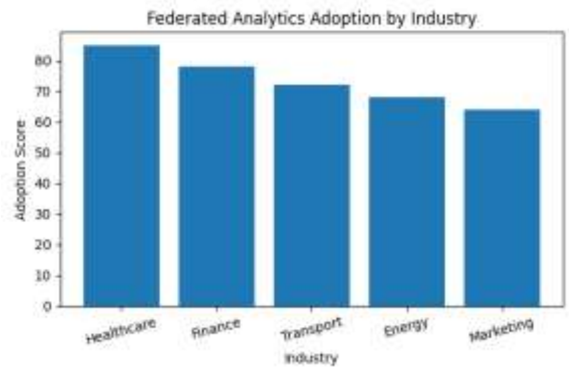


Fig 1: Federated Data and Interoperability

1.1. Industry exemplars

Knowledge gained from the examination of industry exemplars contributes to a deeper understanding of the federated data management paradigm and accelerates its deployment into new sectors by supporting successful proof-of-concept efforts. Examination of representative implementations demonstrates that clear sensitivity to application context, capability needs, and stakeholder roles has positively impacted delivery success. Such explorations yield both quantitative metrics and qualitative feedback controllable through success-factor analysis. Together, these streams of insight empower actionable recommendations with specified business justifications.

As suggested by the expression “Federated Insight Engineering”, federated data architectures provide a means for deriving insight and intelligence from regional and central domains alike. A broadly interpretable construction supersedes conventional definitions that restrict federation to a particular type of data source or analytics task. Industry interviews and case reports highlight federation’s fundamental principles—data locality, interoperability, governance, trust, and scalability. Emerging developments—including self-sovereign identity management and data wallets—help establish higher levels of decentralized trust, fostering federation between organizations across disparate operational domains. Services such as Databricks Unity Catalog support practical expression of the model even in Data Lakehouse environments.



1.2. Lessons learned and best practices

Many organizations are deriving production-grade business benefits from example implementations of federated insight engineering across a variety of industries and use cases. Quantitative metrics demonstrate substantial savings in time-to-insight, resource investments, cost exposure, risk management, and ROI. Moreover, experience gained from both implementing and consuming federated insights provides further invaluable lessons, possibly the strongest value proposition for federated capability.

A representative sample of industry use cases was collected, covering five meta-use case domains: Energy, Health, Law Enforcement, Marketing, and Transport. These were analyzed in terms of major architectural patterns and deployment models. Analysis was subsequently distilled into a set of quantitative factors contributing to success. Background patterns, together with the identified factors, suggest a first-order “Best Practice” configuration for federated models, providing a guide for federated realization across other industries and use cases. The analysis concludes by formalizing the understanding of federated initiative investment so as to create a Risk/Reward matrix that is

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

independent of a specific Industry or Meta-Use-Case. Nevertheless, the federated paradigm may impose a heavier burden of design. Integration makes widespread use of a central copy of the data for many purposes; hence, availability is by nature a simple service-level target across many consumers and uses.

Equation 1: Federated Averaging (FedAvg)

Used for aggregating local model updates from distributed domains.

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_k^{(t)}$$

Where:

- w_{t+1}= updated global model
- $w_k^{(t)}$= local model from client k
- n_k= local dataset size
- $n = \sum_{k=1}^{K} n_k$
- **Table 1. Core Concepts of Federated Insight Engineering**

Concept	Description	Key Objective
Federated Insight Engineering (FIE)	Cooperative analytics across distributed and autonomous data sources	Enable insights without centralizing data
Data Harmonization	Alignment of heterogeneous data for consistent interpretation	Improve interoperability and decision-making
Data Integration	Combining datasets into a unified structure	Support unified querying and analytics
Federated Analytics	Analytics performed locally with aggregated global results	Preserve locality and privacy
Federated Learning	Collaborative ML training without sharing raw data	Privacy-preserving AI development
Multi-Tier Architecture	Layered enterprise data ecosystem	Optimize latency, governance, and scalability

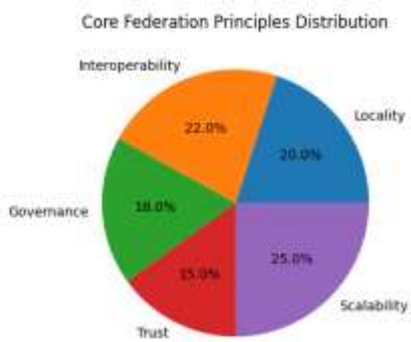
2. Foundations of Federated Insight Engineering

Federated Insight Engineering encompasses the federated identification, harmonization, analytics, and extraction of business insights through the cooperative use of interoperable data assets. These assets may originate from autonomous sources both within and beyond the organization. Use cases typically involve cross-domain, cross-organizational, or cross-time analyses that would be difficult or impossible with traditional solutions. A federated approach permits analytics-as-usual on local data, thus avoiding complex persistent data sharing arrangements, yet accommodates the exploration of bigger-picture questions that benefit from a more global data landscape.

Any desirable form of data federation may be deployed: analytical sources may remain entirely local, be managed in a decentralized configuration, or be provisioned in an enterprise data warehouse. Challenges such as latent data quality issues in remote sources are anticipated and resolved through the specification of an analytics contract. While contrastive situations tend to arise between a data quality disparity and a data availability guarantee, both can be maintained with sufficient visibility of the operational use of federated

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

analytics insight contracts, via stewardship mechanisms or business relationship management.



2.1. Definitions and scope

Federated insight engineering enables collaborative computing across diverse data sources with minimal prerequisites. Federation implies no single master or central point of control; each domain can independently manage its data as desired. Stakeholders can contribute, use, and gain insights from the shared environment without having full access to outside data, resulting in spontaneous and evolving knowledge networks. Online analytical processing (OLAP) and data mining capabilities can be distributed over the various data sources and federated learning conducted across them. Decision trees and other models can be trained and executed locally, while federated analytics extracts meta-results—a harvest of locally executed processing.

Federated insight engineering is particularly suited for sensitive information that cannot be copied to a central location, such as personal medical records or transaction history in financial systems. Distribution also supports minimizing latency. Where information stored locally by many is of limited age, such as call detail records or recommendation examples, insight can be extracted by executing requests close to their origin without the delay of transferring to a common storage area.

Equation 2: Differential Privacy Mechanism

Relevant to the section on privacy-preserving analytics.

$$M(D) = f(D) + \mathcal{N}(0, \sigma^2)$$

Where:

- $M(D)$= privatized query result
- $f(D)$= original analytical function
- $\mathcal{N}(0, \sigma^2)$= Gaussian noise added for privacy

2.2. Core principles of federation

The problem space of federated insight engineering encompasses the multi-tier harmonization of enterprise data: the design and operation of enterprise data landscapes that are justified, effective, and. ultimately, useful. To this end, a set of five core principles of federation is proposed, which concern the distribution of data within a federated data environment and the extent to which the data harmonization processes address the imperatives of locality, interoperability, governance, trust, and scalability. Illustrating the principles in practice, examples have been derived from the design and operation of data landscapes supporting the calculation of an enterprise-wide carbon footprint through the deployment of a consortium of organisations from the electricity sector, the defence sector, and cutting-edge research and innovation on Electric Mobility systems.

Over the past several years, the concept of federation has gathered momentum in a broad variety of application contexts. Here, it refers to the loose coupling of multiple organisations (or any non-homogeneous data producers) for the purpose of using (in the broadest sense) the data a single organisation does not directly own. The implementation of such federation models at the data level must overcome a number of challenges. These challenges, often classified as achieving appropriate local data governance, and appropriate data harmonisation at scale, are important for every federation, and are particularly acute when one of the main objectives of the federation is to operate at full data locality,

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

in a very responsive way either in terms of data availability or data latency.

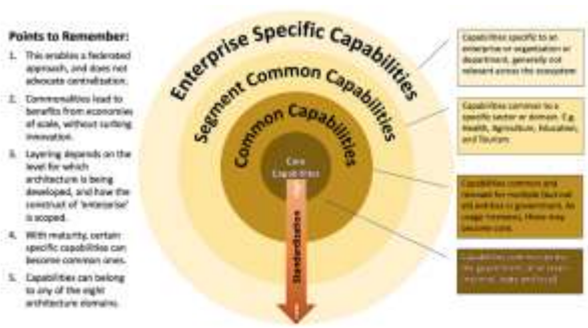


Fig 2: Federation in Architecture

Such challenges have become especially relevant in the specific context of industry use cases driven by sustainability requirements and obligations. Concrete industry deployments must adhere to sustainability requirements, which imply obligations to calculate carbon and GHG footprints. These are generally calculated at the enterprise level, requiring a full coverage of the footprint at the enterprise level either by collecting data in a centralised cloud-based system or by computing it in a data-harmonised way for data sources outside the enterprise system. The latter is the preferred method, to allow calculations at appropriate data locality and latency.

2.3. Data harmonization versus integration

Fulfilling data management goals requires both data harmonization and data integration—two related yet fundamentally different concepts. Data harmonization seeks context-specific alignment of datasets for effective decision-making within a limited local or domain-specific data landscape; data integration prepares datasets from different data sources for joint analysis or querying, regardless of the originating context. Because harmonization and integration require different approaches, achieve different properties, and rely on different outcomes, the natural tension between them rises to the forefront of design decisions.

Data integration represents a classical problem in data management: combining records from diverse sources into a coherent whole, often with the goal of answering a complex query. Broader than federated analytics, the goal is to normalize disparate elements into a coherent one by joining and merging or by ontology-based approaches—often utilizing record linkage techniques to resolve duplicates, and the end result is a single dataset or five datasets or specific merging between datasets. Harmonization that enables integration at any level needs to be achieved at the metadata level before data integration.

Harmonization enables not just integration but any handling of disparate or heterogeneous data sources. The goal is to enable decision-making within a localized or authorized domain that actually requires data index rather than standardized search. Data harmonization can happen at various levels: semantic or by-dimensional modeling and canonical representations.

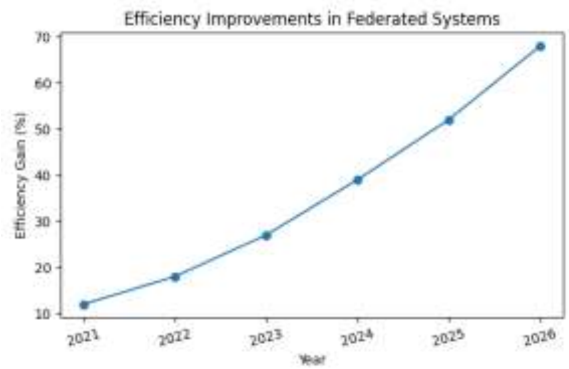
Table 2. Multi-Tier Enterprise Data Architecture

Tier	Ownership	Characteristics	Latency Requirement	Typical Functions
Local Tier	Operational systems	Real-time localized data	Very Low	Transaction processing, local analytics
Domain Tier	Business/Data Science teams	Analytical and cross-functional data	Low	Data science, operational dashboards
Enterprise Tier	Enterprise analytics teams	Harmonized canonical datasets	Medium	Enterprise reporting,

Tier	Ownership	Characteristics	Latency Requirement	Typical Functions
				strategic analytics
External Tier	External organizations/ APIs	Third-party or partner data	Flexible	External enrichment, ecosystem collaboration

3. Multi-Tier Architectures for Enterprise Data

A multi-tiered landscape and system architecture compose enterprise data functions. The landscape consists of four strata: local, domain, enterprise, and external. The local tier is owned by operational systems and handles real-time processing with localized data requirements. These data remain within source systems for insight engineering. The domain tier, owned by data science or business teams, supports data science and analytics with data sourced from operational systems, enterprise repositories, and other external sources. Latency must be low and changes reflected quickly. Interaction with other tiers is largely invisible to consumers. The enterprise tier addresses longer-latency reporting and analytical requirements. Data are harmonized into canonical forms, typically on dedicated enterprise analytical platforms. The external layer comprises external sources and services. Latency is unimportant, but compatibility with consumption requirements and standards becomes critical.



3.1. Tiered data landscapes

Four strata (local, domain, enterprise, and external) describe federated insight engineering (FIE) data landscapes, providing a basis for separating ownership and operational responsibilities, optimizing latency, and determining access patterns. Local data includes primary data sources owned and created by federated members, located within their information technology (IT) perimeter and governed by the institution's standard operating procedures. Domain data represents a time-optimised flow of consumed and produced information within an enterprise governance structure.Execute a federated cycle driven by local data generation or external data consumption, taking place within an enterprise and governed by its internal policies. External data arise from coordinated flows crossing enterprise boundaries, especially when time is secondary to data quality considerations. Domain landscapes contain a subset of the local data and are subject to a stricter latency strategy. These landscapes may also include external data that are cached for short- or medium-term subsumption.The four strata of federated insight engineering data landscapes reflect the different owners and layers of latency strategies.

Data latency considerations primarily influence other reasons for limiting access to domain data: (1) ideally, local data should be accessed as close as possible to their generation source; (2) time is the distinguishing factor for data when information quality is paramount; (3) external information

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

sources answer questions that cannot be addressed within local landscapes.Ideally, federated analytical computations should ingest data from their local landscapes or those of a neighbouring federation partner. This rule is difficult to enforce in practice because of the inherent uncertainty associated with data availability and quality.

Equation 3: Data Harmonization Similarity Score

Useful for semantic alignment and ontology matching across domains.

$$Sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

(Jaccard Similarity)

Where:

A, B= semantic attribute sets or ontology terms

3.2. Layered governance and stewardship

Four roles are defined for data governance and stewardship, along with the respective policies, structures, and compliance controls governing the data at each tier. For local data sources, data owners are responsible for defining and enforcing internal data governance policies and controls, thus ensuring the quality and security of local data. These controls often include data access privileges for local users. At the domain data tier, domain data stewards underwrite the relevance, quality, and accessibility of the domain data landscape. The roles and responsibilities of the different data stewards across the enterprise are documented in a governance framework outlining policies to be enforced at the enterprise level.

Fig 3: Data Stewardship in a Federated

3.3. Interoperability standards and schemas

Automatic compatibility checks can be performed when interoperability standards and schemas are established across the data landscape. Supported and supported schemas, ontologies, and APIs are specified at each layer of the local and domain data landscapes. Supporting schemas express the structures of the data produced at the layer during data processing. Small-scale mappings are applied to locally harmonize the data. Supported schemas residing at the domain layer can be formed by routine inspections of the available domain data or the associated data contracts. These mappings enable compatibility checks during data usage. Data residing at each domain layer can also be published in the supported schemas to facilitate data inquiring from external data consumers.

In addition to the supported schemas, mappings can be expressed among the domain-owned data that belong to different domains. These mappings provide indications for the consumable pattern of these data and thus ease domain-wide data harmonization. When the local and the domain data layers are established, ontologies are created to represent a domain and are especially useful when imposing a compatibility requirement on a domain.

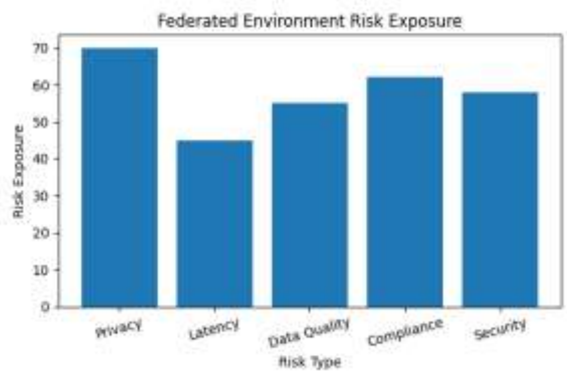
Table 3. Core Principles of Federation

Principle	Description	Importance
Locality	Data processed near its source	Reduces latency and transfer overhead
Interoperability	Systems communicate using common standards	Enables seamless federation
Governance	Policies and stewardship mechanisms	Ensures compliance and accountability
Trust	Secure and reliable collaboration	Supports cross-organizational cooperation
Scalability	Expansion across domains and organizations	Enables enterprise-wide deployment

4. Methods for Data Harmonization

Unlike integration, which produces a new amalgamated artifact, harmonization is about reconciling the descriptions of pre-existing data artifacts so that they can be understood and related to each other. This reconciliation is often achieved through semantic matching between datasets. Different types of semantic alignment carry different degrees of expressiveness and explanation, and so of the artefacts produced, it is the creation of ontologies, or domain models, that is most powerful. Once semantic alignment is in place, dimensional models can be defined for the distributed datasets, allowing a view of the data across the different domains, and a canonical representation can be developed. Instances of the canonical representation are then produced by ETL or ELT processes. When these constitute local fact populations, the processes may also provide the necessary metadata for schema mapping and synchronisation.

Semantic Alignment and Ontologies. An essential aspect of federated data harmonisation is semantic alignment, which encompasses the assignment of synonyms, synonyms, and mappings between heterogeneous source schemas. In a semantic landscape, a central resource, called a domain ontology, provides a vocabulary for conceptualising the domain and enables additional semantic services, such as explanation through inference.



4.1. Semantic alignment and ontologies

Design and maintain semantic alignment across data from different sources, including domain terminologies and synonyms, mappings across synonyms, and ontology-based reasoning methods. Domain ontologies provide a central point for semantic alignment of local and enterprise-level data landscapes. The semantic alignment is structured into four parts: domain ontology development, synonymous terms storage and translation, term mappings across sets of synonyms, and ontology-based reasoning methods. Domain ontologies align local terminology with enterprise understanding and provide reasoning capabilities to derive additional terms and relations.

The domain ontologies encapsulate keys terms, relations, and categorization structures for the domain and its analytic usage. An ontology development methodology ensures that terminology supports analytic requirements and decision-making users. For example, semantic translation engines such

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

as the e-business ontology specify B2B and B2C exchanges through various online services such as banks, auctions, and marketplaces.

Synonyms ensure that the meaning of a term is preserved, even when it is expressed differently by the source, enterprise, or consumers. For example, a warehouse may be referred to as a depot, storeroom, or other terms specific to the organization. Domain-dependent synonym sets are designed for steadily changing data sources such as news feeds or blogs, which are continuously mined for sentiment analysis and discussion topics.

Mappings among sets of synonyms determine compatibility of terms with similar meanings in order to perform union or join operations. For example, an enterprise may wish to aggregate news sentiment for its competitors, even though different web sources may use different terms for the same competitor (for example, “Ford” , “Ford Motor Company” , or “Ford Motor”). These mappings are enriched as new terms enter the data sources and become associated with company or competitor names.

Fig 4: Ontology Alignment

4.2. Dimensional modeling and canonical forms

Dimensional modeling can also contribute to harmonization across data tiers without direct integration. A reliable gateway approach identifies local facts (or observables) supported by discrete dimensions; designates a canonical form, schema, or cube for the local facts; and elaborates an ETL submission checklist. For each fact of interest, local dimensions are combined to form the canonical set, which is populated by effortless routines for those local facts offered for harmonized aggregation—"an always-ready, best-effort route to strong federation."

Dimensional modeling can also streamline data migration, as "moving an entire cube is a no-brainer" for an associate or apprentice with limited experience. In this case, a new country has built a full OLAP cube of inbound shipments and a single-dimensional COD (Cede d'Or) Cube of outbound deliveries to observe cargo surges and assess demand on receding spikes, for an average effort of 25% of production, but a significant economic shortfall relative to the country of hope that has no IPR."

Equation 4: Risk Exposure Model

Applicable to federated governance and risk assessment.

$$Risk = Probability \times Impact$$

Expanded form:

$$R = P(E_i) \times I(E_i)$$

Where:

- $P(E_i)$= likelihood of event i
- $I(E_i)$= business impact of event i

4.3. Schema mapping and metadata management

ETL/ELT processes transfer data across tiers in local or domain storage systems. These processes have metadata describing mappings between internal and external schemas. Schema mapping languages such as the Object-Relational Data Mapping (ORM) framework provide a formal rule-driven description language for mapping data stored in a database with different schemas. The statements contained in ORM mappings are compiled into a format comprised of a collection of SQL statements that extract data from a source database and load it into a destination database. These

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

mappings can be made by a data steward or other knowledgeable personnel or by a data discovery and/or profiling mechanism. Where data lineage tracking tools exist, they should be consulted and used in the mapping discussion. Although data discovery and profiling should allow an ETL/ELT process to work with unknown and auto-discovered source-data availability and structure, a schema mapping language should be used whenever a source schema is known beforehand.

A metadata repository stores mappings between schemas, tables and columns of data sources, ETL/ELT lineage (i.e., the mapping of source data tables to destination data tables), and dimensions. As data contracts define the KPIs, the technical, syntactic and semantic mapping of the Data Supply Chain allow for compliance evaluation of these KPIs, enabling the discovery of quality problems as well as the augmentation of monitoring capabilities.

Table 4. Data Harmonization vs Data Integration

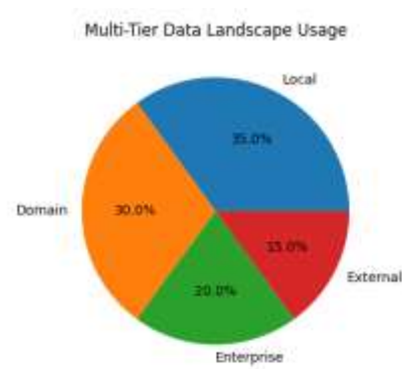
Aspect	Data Harmonization	Data Integration
Purpose	Align meanings and structures	Merge datasets
Focus	Semantic consistency	Unified data repository
Output	Compatible datasets	Combined dataset
Scope	Context-specific	Cross-system analytics
Techniques	Ontologies, schema mapping	ETL, record linkage
Federation Suitability	Highly suitable	Often centralized

5. Federated Analytics and Insight Extraction

Data harmonization face multiple challenges but it is necessary to leverage the entire system. Analytics could be distributed on a specific perimeter but could be coordinated across the domains to avoid redundant and incongruous results. Also, Machine Learning could be used to exploit all data in a privacy-preserving way.

Local processing solutions implemented in a federated analytics scenario will produce potentially disparate results. In such a case, an additional phase should be considered where the “flat results” delivered by the domains are aggregated, analyzed, and synthesized to achieve a global overview. The aggregation phase is in charge of the integrator data governance. Several techniques could be exploited to manage local results aggregation.

Data Analytics and Machine Learning require sensitive data and the stakeholders have different data privacy rules. In some cases, regulations could even explicitly forbid the sharing of structured data. Privacy-Preserving Machine Learning Techniques aim at enabling the distributed training of machine learning models with non-IID private data without sharing information between data holders and making it possible to evaluate the models’ performances without revealing test data.



ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

5.1. Local processing with global coordination

Local processing with global coordination is a prevalent use case for federated analytics. Individual sites perform the data analytics relevant to their domain on the data they steward, generating interim results that are shared with a coordinating hub for high-level aggregation or cross-domain reasoning.

The key concern when employing these methods is to manage the merging of interim results produced at different sites without introducing inconsistencies, contradictions, or unjustifiable conclusions. Horizontal aggregations, such as sums, can be safely performed; however, structural differences in the contributing responses can pose issues. Solutions include employing a common response schema augmented with a collection of "others" slots in the response to account for variations across domains or defining dedicated summary concepts, such as the dimensions of a "federated" facts table, to reconcile disparate structures.

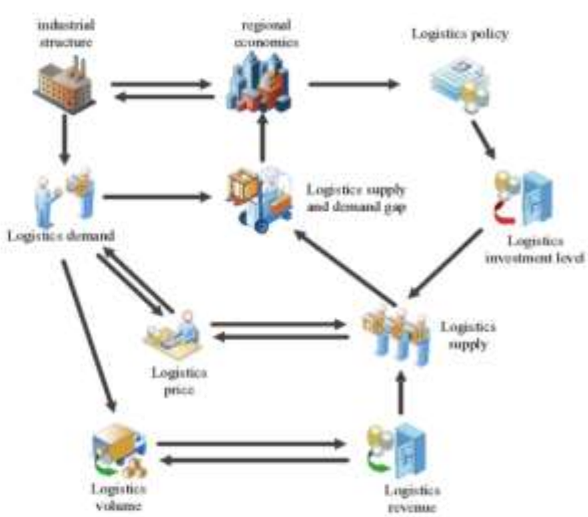


Fig 5: Local processing with global coordination

5.2. Federated learning and analytics

Federated model training across decentralized domains enables collaborative AI without centralized data. Participants own data and model parameters and disclose only orchestration-processed updates for aggregation. Contributions during distributed training are weighed using data size and/or performance accuracy. Although federated architectures were coined for private data and regulatory compliance, similar concepts also benefit AI development in domains with scarce data availability or high deployment costs, such as healthcare and environmental sciences.

Federated analytics describes the application of analytic methods on decentralized data without data transfer. Participants apply a chosen analytic method locally on their own data and communicate results within a safe framework for external evaluation or insight extraction. Such methods have gained traction in recent years both because of strong data privacy requirements (e.g., in the EU) and because of data governance regulations (e.g., GDPR) breaking classic data exchange practices. Banks and insurance companies are some of the first adopting federated analytics, where the results are compared to detect fraud or anti-money laundering activities.

Equation 5: Latency-Aware Federated Processing Cost

Relevant to multi-tier enterprise data architecture.

$$C_{total} = \sum_{i=1}^{N} \left(C_{compute}^{(i)} + C_{transfer}^{(i)} + \lambda L_i \right)$$

Where:

- $C_{compute}^{(i)}$= local computation cost
- $C_{transfer}^{(i)}$= communication cost
- L_i= latency at node i

λ= latency weighting factor

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

5.3. Privacy-preserving techniques and compliance

Privacy risks are especially pressing when insights from sensitive data are shared across organizational boundaries. Compliance with privacy regulations is therefore a major concern of very large-scale federations in which various organizations contribute results from their own proprietary and sensitive data but have little visibility or control over the overall federated system. A variety of privacy-preserving techniques can be applied in conjunction with federation to address these concerns.

Secure multiparty computation allows several parties to jointly compute a function over their inputs while keeping those inputs private. In the case of training machine learning models, each party can locally compute a secret share of its contribution to the global model, and these contributions can be combined without revealing any of the contributions. Different techniques can be used to achieve the same goal, including homomorphic encryption and garbled circuits. Differential privacy allows a data owner to release insights from a data set while guaranteeing that the output does not significantly depend on any individual data contributor. It achieves this by adding noise based on a sensitivity determination to the output so that it meets the differential privacy condition with respect to a chosen privacy parameter ε. Numerous practical mechanisms exist to provide differential privacy guarantees for model training in a federated context, as in federated learning with differential privacy. The privacy and other risk aspects of federation-based insight engineering are neither an intrinsic nor a supporting pattern; rather, they need to be embedded back into the federation architecture as design-time and run-time considerations.

Table 5. Governance and Stewardship Roles

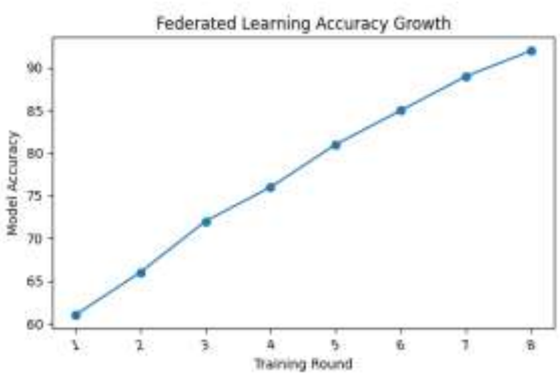
Role	Responsibilities	Governance Scope
Data Owners	Maintain local data quality and access controls	Local Tier
Domain Data Stewards	Validate relevance and accessibility	Domain Tier
Enterprise Compliance Officers	Ensure privacy and regulatory compliance	Enterprise Tier
Supply-Chain Data Stewards	Manage external partner data quality	External Tier

6. Governance, Ethics, and Risk Management

Federated Insight Engineering requires attention to governance, ethics, risk, and societal impact. While regulatory compliance is a fundamental dimension, effective governance also aligns stewardship across the federated landscape, ensures compliance with contracts, and preserves reputation, while risk assessment and mitigation help to identify and proactively address potential pitfalls. Furthermore, fairness and bias in algorithmic predictions have gained attention across society, and federated machine learning often raises additional ethical considerations.

Data Governance Frameworks Effective data governance requires a dedicated organisation structure, clearly defined roles and responsibilities, policies aligned to business objectives, and key performance indicators (KPIs) suitable for monitoring of governance effectiveness. Distinguishing between data governance and data stewardship emphasises that management of data quality, access, availability, and privacy is distributed across the federated data landscape and ultimately rests with the various owners of data at every layer.

Risk Assessment in Federated Environments Risk is defined as “the effect of uncertainty on objectives”. Therefore, risk associated with an insight engineering effort can be defined as the likelihood that insights will not meet their stated goals. Risk assessment provides an understanding of how various factors might contribute to this risk. Developing a risk assessment involves identifying risk categories—the elements that present a risk to the effort—deciding how to assess each risk category, and developing measures to mitigate against each category.



6.1. Data governance frameworks

Effective governance comprises the organizational structures, policies, roles, and responsibilities necessary to ensure that data are properly stewarded and that they meet quality, availability, and compliance requirements. The governing body is ultimately accountable for the fulfillment of all three aspects and ideally provides an explicit data strategy that articulates its vision for the use, management, and protection of data in the organization.

The strategy should assign responsibility for each data element referred to in legal and compliance documents, specify important external compliance requirements, and define the organization's risk appetite with regard to data availability and quality. A dashboard of key performance indicators (KPIs) can indicate success and help to identify potential areas for improvement. Data-centered initiatives, including enterprise data management, data governance, and data quality improvement, can then be driven by the strategy with impact assessments conducted to demonstrate value.

Table 6. Interoperability Components

Component	Purpose	Example
Schemas	Define data structures	JSON/XML schemas
Ontologies	Define semantic relationships	Domain ontology
APIs	Enable system interaction	REST APIs
Metadata Repositories	Store mappings and lineage	ETL metadata catalog
Semantic Mapping	Align equivalent concepts	“Depot” = “Warehouse”

6.2. Risk assessment in federated environments

Risk assessment encompasses the identification, evaluation, and mitigation of risks that could hinder federated insight-engineering objectives. Of particular interest are risks that impact data quality and availability and associated legal, regulatory, and contractual obligations. Undetected changes in data assets can yield unanticipated analysis outputs, degrading service-level performance. Evaluation of risks inherent in federated environments should draw on frameworks applied in both centralized data ecosystems and traditional enterprise-wide risk assessments.

Risk exposure can be plotted on a matrix based on qualitative and quantitative assessments of likelihood and impact. Some risk categories are unique to federated systems, while others can be classified according to established frameworks—such as information security, business continuity, change management, and project risk assessment—augmented as

needed with details pertinent to the specific case. Impact mitigation requires appropriate skills, experience, and oversight throughout the risk assessment process.

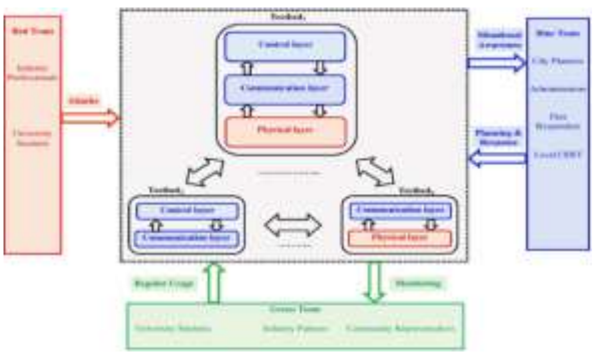


Fig 6: Federated Environment - an overview

6.3. Ethical considerations and bias mitigation

Federated Insight Engineering practices strive to deliver fair outcomes for all stakeholders. Achieving fairness is a much-debated topic, which requires careful handling of bias in the underlying data. Federated Analytics and federated learning offer various solutions, but fairness and correctness should be examined during the design stage. One such direction involves defining business rules that explicitly disallow certain kinds of bias, such as gender-related decisions. Data controllers or trustees should oversee data acquisition to guarantee that the data used for model training is clean and representative.

Few scholars publish on bias mitigation, and even fewer address privacy-sensitive domains. To minimize negligence in such domains, data contracts must define fairness expectations explicitly. In the case of supervised learning, training performance and generalization performance for sensitive groups should lie within specific bounds. In the presence of disparate impact, careful population selection is needed during data acquisition. Compliance checks should audit fairness, allow analysis of features that contribute to unfairness, and assess features that support fairness. Finally, for maintenance and monitoring, any detected disparity requires justification to evaluate whether it results from changes in fundamental attributes or has occurred due to unfair influence.

Like any engineering discipline, research and innovation in insight engineering draws heavily on established architectural patterns: proven solutions to recurring problems. Such patterns also facilitate the communication of complex ideas since, just as words and phrases are combined into sentences and paragraphs to express longer and more complex concepts, patterns are often combined into reference architectures that depict large parts of the discipline being explored. In the case of federated insight engineering, four architectural patterns have been identified: a framework for managing multi-tier enterprise data, support for federated machine learning with privacy preservation, a typical deployment topology, and a reference architecture for federated insight engineering.

Multi-Tier Federated Data Harmonization

7. Engineering Practices and Toolchains

Data analytics is evolving rapidly, driven by advances in computing, algorithms, infrastructure, and increasingly

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

sophisticated data consumers. Consequently, most organizations can no longer deliver timely, relevant analytics effectively. As a response, enterprises are looking beyond their boundaries into their business ecosystem for complementary insight. Federated Insight Engineering supports these external-facing requirements, enabling data-driven decision making across a wider data landscape. Architectural patterns and reference architectures guide Federation Engineering projects. Work contracts formalize expectations around data quality, accessibility, and availability, while automation orchestration streamlines workflow instrumentation.

Federated analytics, like any form of analytics, is concerned with generating insight from data. In this context, insight is drawn from data across the entire federation but residing within the federation boundary. That said, any one organization is unlikely to have the entire set of data sources required for a particular analytic use case. Therefore, to realize more demanding use cases that cross Federation boundaries, federated insight engineering provides the foundations for local analytics with aggregation and control within a global context. Data compute and analytics run in situ at each local domain, while a coordinating role facilitates management of the overall process. The bank example from the financial services domain illustrates the automation of data monitoring, validation controls, and data quality issue root cause analysis, and remediation, for a specific risk domain, based on infrastructure components (such as Micro Focus IDOL) designed for enterprise scalable deployments.

Federated Analytics and Governance Flow

7.1. Architectural patterns and reference architecture

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

Just as an information system gathers vast quantities of data to support broad-scale analytics for strategic insight at the enterprise level, so too does a multi-tier architecture for enterprise data harmonization. Such an architecture not only mirrors the operational model of an information system, with the corresponding gain in data locality and response speed, but also supports varied ownership and stewardship patterns across an organisation's digital ecosystem. At its simplest, a multi-tier architecture for enterprise data is stratified into local, domain, enterprise, and external landscapes, each defined by data ownership, access patterns, norms and latency tolerances, and prospects for machine-to-machine orchestration. Local landscapes are established and governed by the enterprise's business units and functions, domain landscapes by their groups and domains, the enterprise landscape by enterprise-wide data and analytics teams, and the external landscape by APIs that channel data into and out of the enterprise, including mobile apps and an information system's data lake.

8. Conclusions

Federated Insight Engineering enables data-driven decisions that abstract away the complexities of enterprise data management. Instead of requiring data scientists to expend effort on the intricacies of accessing, understanding, preparing, and integrating data custodians' resources, it delivers signatures of the data sources combined with an architecture that allows data harmonization at different layers. The approach offloads some of the demands placed on executive actions by data personnel relations and cooperation; mandates and procedures emerging from the data governance and from the compliance areas of the enterprise thereby facilitate and structure the environmental situation for maximizing value creation, which data science and analytics activities are proving to be so efficient at doing.

Centralizing enterprise data sources under a data lake or warehouse resolves access and understanding frictions and enables processing on large data volumes. But under these centralized models, speed of access and processing becomes an issue. Time-sensitive use cases need low-latency sources, such as operational data stores. Federation handles the resulting discrepancies and conflicts in light of the analytical task requirements, provides one actually specified logical model that can support the intended analytics across data custodians, and anticipates a distributed implementation scenario through data combinability sampling checks performed as part of the actual analytics preparation work.

References

1. Kokash, N., Wang, L., Gillespie, T. H., Belloum, A., Grosso, P., Quinney, S., Li, L., & de Bono, B. (2025). Ontology- and LLM-based data harmonization for federated learning in healthcare. arXiv.
2. Davuluri, P. S. L. (2023). AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems (December 15, 2023).
3. Stripelis, D., & Ambite, J. L. (2023). Federated learning over harmonized data silos. arXiv.
4. Zhang, F., Kreuter, D., Chen, Y., Dittmer, S., Tull, S., Shadbahr, T., Preller, J., Aston, J. A. D., Schönlieb, C.-B., Gleadall, N., & Roberts, M. (2023). Recent methodological advances in federated learning for healthcare. arXiv.
5. Li, M., Wang, Y., Zhao, X., & Chen, H. (2025). Implementing federated learning in healthcare: Challenges and opportunities. Medical Image Analysis, 92, 103112.
6. Kolla, T. (2024). Graph Neural Networks for HCC Risk Adjustment and Interoperability. International Journal of Science, Research and Technology, 7(6), 13244-13255.
7. Luo, H., & Ji, C. (2025). Federated learning-based data collaboration method for enhancing edge cloud

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

AI system security using large language models. arXiv.

8. Tölle, M., Burger, L., Kelm, H., André, F., Bannas, P., & Engelhardt, S. (2024). Multi-modal dataset creation for federated learning with DICOM structured reports. arXiv.
9. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A., Bonawitz, K., Charles, Z., Cormode, G., & Cummings, R. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.
10. Inala, R. (2023). AI-powered investment decision support systems: Building smart data products with embedded governance controls. Journal for ReAttach Therapy and Developmental Diversities, 6(10), 2251-2266.
11. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.
12. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19.
13. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. Proceedings of the 37th International Conference on Machine Learning, 5132–5143.
14. Kolla, S. K., & Mangalampalli, B. M. (2024). Edge-Based Deep Learning Systems for Point-of-Care Diagnostic Intelligence. Journal of Neonatal Surgery, 13(1), 2387-2399.
15. Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., & McMahan, H. B. (2021). Adaptive federated optimization. International Conference on Learning Representations.
16. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2020). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 92–104.
17. Inala, R. Designing Scalable Technology Architectures for Customer Data in Group Insurance and Investment Platforms.
18. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.
19. Karargyris, A., Umeton, R., Sheller, M. J., Aristizabal, A., & George, J. (2023). Federated benchmarking of medical artificial intelligence with MedPerf. Nature Machine Intelligence, 5(8), 799–810.
20. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., & Van Overveldt, T. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374–388.
21. Kolla, S. K., & Reddy, V. A. R. (2024). Evaluating Cloud-Native vs. Hybrid Architectures for Health Benefit Administration Systems. International Journal of Medical Toxicology and Legal Medicine, 27(5), 1042-1053.
22. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.
23. Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, B., & Thorne, B. (2017). Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv.
24. Overman, T., Blum, G., & Klabjan, D. (2024). A primal-dual algorithm for hybrid federated

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 03 Issue: 01
RECEIVED: JANUARY 21 REVISED: FEBRUARY 03 ACCEPTED: FEBRUARY 23 PUBLISHED: MARCH 08

learning. Proceedings of the AAAI Conference on Artificial Intelligence, 38(12), 13544–13552.

25. Guo, J., Mu, H., Liu, X., Ren, H., & Han, C. (2025). Federated learning for biometric recognition: A survey. Artificial Intelligence Review, 58(2), 1–38.
26. Eden, R., Chukwudi, I., Bain, C., Barbieri, S., & Callaway, L. (2025). A scoping review of the governance of federated learning in healthcare. NPJ Digital Medicine, 8(1), 144.
27. Gottimukkala, V. R. R. (2024). Federated Learning Approaches for Fraud Detection in International Payment Systems. https://www. jisem-journal. com/download/118_JISEM. pdf.
28. Barbereau, T., Delgado Fernandez, J., & Potenciano Menci, S. (2025). The governance of federated learning: A decision framework for organisational archetypes. Data & Policy, 7, e25.
29. Liu, Y., Zhang, X., Kang, Y., Li, L., Chen, T., Hong, M., & Yang, Q. (2022). FedBCD: A communication-efficient collaborative learning framework for distributed features. IEEE Transactions on Signal Processing, 70, 4277–4290.
30. Zhang, X., Yin, W., Hong, M., & Chen, T. (2020). Hybrid federated learning: Algorithms and implementation. NeurIPS Workshop on Scalability, Privacy, and Security in Federated Learning.
31. Sheth, A., & Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys, 22(3), 183–236.
32. Hedberg, T. D., Bajaj, M., & Camelio, J. A. (2020). Using graphs to link data across the product lifecycle for enabling smart manufacturing digital threads. Journal of Computing and Information Science in Engineering, 20(2), 021008.
33. Inala, R., & Somu, B. (2024). Agentic ai in retail banking: Redefining customer service and financial decision-making. Journal of Artificial Intelligence and Big Data Disciplines, 1(1), 1-19.
34. Zhang, Q., Liu, J., & Chen, X. (2024). A literature review of the digital thread: Definition, key technologies, and applications. Systems, 12(2), 78.
35. Katta, T. B. (2025). Federated learning for enterprise data integration: Examining the application of federated learning to integrate AI models without centralizing enterprise data. Journal of Information Systems Engineering and Management, 10(63s).
36. Na, S., Rouček, T., Ulrich, J., Pikman, J., & Krajník, T. (2022). Federated reinforcement learning for collective navigation of robotic swarms. IEEE Transactions on Cognitive and Developmental Systems, 14(4), 1517–1528.
37. Yu, X., Queralta, J. P., & Westerlund, T. (2021). Towards lifelong federated learning in autonomous mobile robots with continuous sim-to-real transfer. Procedia Computer Science, 184, 406–412.