ARCHITECTING INTELLIGENT DECENTRALIZED DATA SYSTEMS TO ENABLE ANALYTICS WITH ENTROPY-AWARE GOVERNANCE, QUANTUM READINESS AND LLM-DRIVEN FEDERATION

Meethun Panda[1] and Soumyodeep Mukherjee[2]

[1]Associate Partner, Bain & Company, Dubai, UAE
[2]Associate Director, Genmab, New Jersey, USA

ABSTRACT

Enterprises pursuing AI-driven transformation face a critical tradeoff: centralized consistency vs. decentralized scalability. The "Data Platform Unification Paradox" captures this dilemma. Building on our prior NLPI 2025 paper, this extended version integrates technical depth, mathematical models, and concrete architectures, especially for integrating Data Mesh with Quantum Databases and LLM Agents. A federated architecture is proposed using graph-theoretic models and entropy-based data valuation. We introduce a formal structure to evaluate platform complexity and propose intelligent agent-based governance models to operationalize data sharing across domains. This work aims to move beyond conceptual frameworks by proposing actionable blueprints for next-generation, intelligent data ecosystems.

KEYWORDS

Data Mesh, Entropy, Federated Graph, Zero-Trust, Quantum DB, LLM Agents, Domain Ownership, Data Governance, Distributed Data Platforms, Decentralized Architecture, Centralized Architecture

1. INTRODUCTION

The rapid increase in data volume and heterogeneity challenges the scalability of centralized data architectures. While monolithic platforms offer control and standardization, they often create bottlenecks and delay innovation. Data Mesh has emerged as a viable alternative, decentralizing data ownership and enabling domain teams to manage their data as products. This paper builds on our original NLPI 2025 publication and focuses on formalizing the architecture, quantifying information value, and embedding intelligence using LLM agents within decentralized platforms. We explore how such architectures can scale analytics, improve AI model outcomes, and maintain governance in increasingly complex data ecosystems.

2. THE PLATFORM PARADOX: FORMALIZATION

Let the enterprise data platform be modeled as a bipartite graph $G = (D, C, E)$, where $D = \{d_1, d_2, ..., d_n\}$ are data domains and $C = \{c_1, c_2, ..., c_m\}$ are consumers (e.g., ML pipelines, BI teams). Edges $e_{ij} \in E$ represent data flow from d_i to c_j.

DOI: 10.5121/ijdms.2025.17202

Platform complexity is defined as:

$$\mathcal{C}(G) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} \cdot \log(1 + f_{ij})$$

Where w_{ij} is the perceived importance or size of data transfer, and f_{ij} is the frequency of access. Centralized systems try to minimize f_{ij}, but this can lead to overload on central nodes and inefficient scaling. In contrast, Data Mesh distributes ownership and flattens w_{ij} variation across nodes, reducing systemic fragility.

We define a **balance coefficient**:

$$\beta = \sigma(w_{ij}) / \mu(w_{ij})$$

Where σ is standard deviation and μ is mean. Lower β indicates a well-distributed platform.

3. ARCHITECTURAL FRAMEWORK

Modern decentralized data platforms necessitate a layered architectural design that integrates data ingestion, productization, governance, and embedded intelligence. The proposed architecture is structured across four layers:

Layer 1: Acquisition. This layer handles ingestion of structured, semi-structured, and unstructured data from various operational systems, APIs, real-time sensors, and third-party sources. Ingestion pipelines must support change data capture (CDC), batch ingestion, and streaming. Provenance metadata is captured at the point of entry to ensure traceability, supporting future audit and explainability.

Layer 2: Productization. Data within each domain is curated into products with clearly defined ownership, service-level agreements (SLAs), and metadata. The process includes data transformation, schema enforcement, enrichment, quality validation, and documentation. Each data product is described using metadata tuple M(p_k) = (schema, freshness, owner, SLA, access policy).

Layer 3: Federated Governance. This layer is the backbone of Data Mesh. It facilitates inter-domain coordination via a federated graph G_f = (P, E_f), where P is the set of all data products and E_f encodes lineage and governance relationships. Federated governance is realized using a combination of centrally defined standards and domain-level flexibility, enabling localized domain driven innovation without compromising compliance.

Layer 4: Intelligence. The intelligence layer embeds LLM-based agents for a range of autonomous tasks such as generating documentation, answering user driven queries leveraging metadata and anomaly detection. Vector databases are used to store semantic embeddings of metadata and the actual data, enabling similarity-based search and retrieval. Agents are deployed as microservices and interact with a centralized orchestration engine.

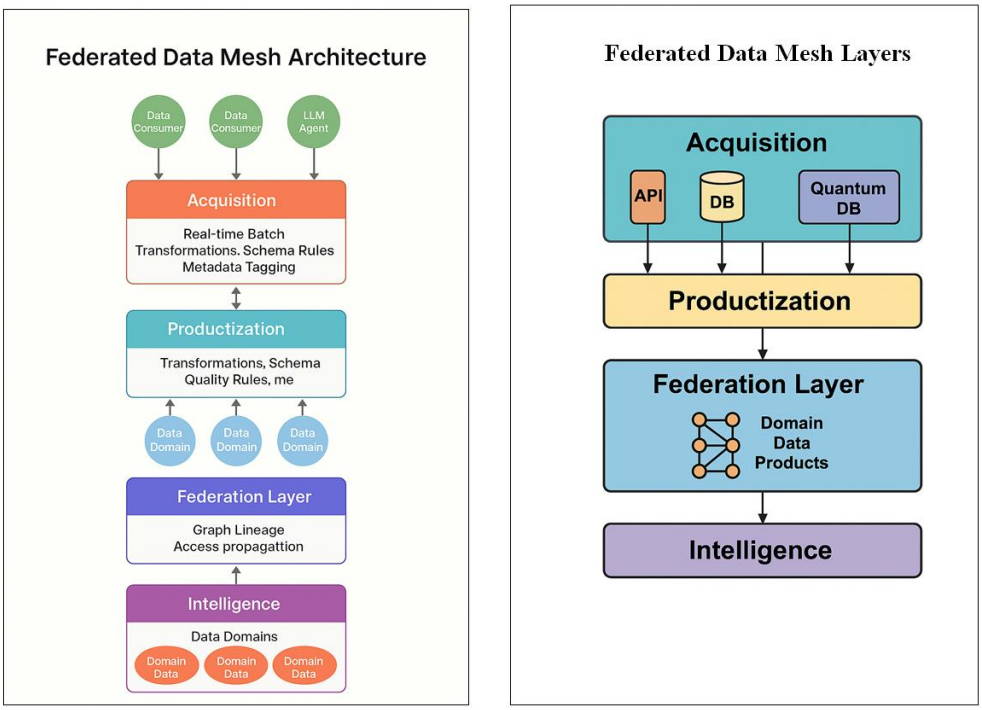

Figure 1. Federated Data Mesh architecture

4. DATA PRODUCT ENTROPY AND PRIORITIZATION

A key innovation in this architecture is the use of information theory to quantify the utility of data products. Each product p_k contains a set of variables X = {x_1, x_2, ..., x_n}. Shannon entropy is used to evaluate the information richness of a product:

$$H(p\_k) = -\Sigma\, P(x\_i) \cdot \log P(x\_i)$$

Products with higher entropy are more informative and are prioritized for high-value analytics tasks, such as model training. However, entropy alone is insufficient. A quality vector Q(p_k) = [accuracy, completeness, timeliness] is computed based on domain-specific criteria. The composite quality score is:

$$q(p\_k) = \alpha\_1 \cdot \text{accuracy} + \alpha\_2 \cdot \text{completeness} + \alpha\_3 \cdot \text{timeliness}$$

where α_i are weights summing to 1, determined via empirical studies or business priorities.

The final utility function U(p_k) = H(p_k) × q(p_k) helps rank data products. Products falling below an entropy threshold θ_H or utility threshold θ_U are flagged for archival or reengineering. This scoring mechanism ensures that storage, processing, and governance resources are allocated efficiently.

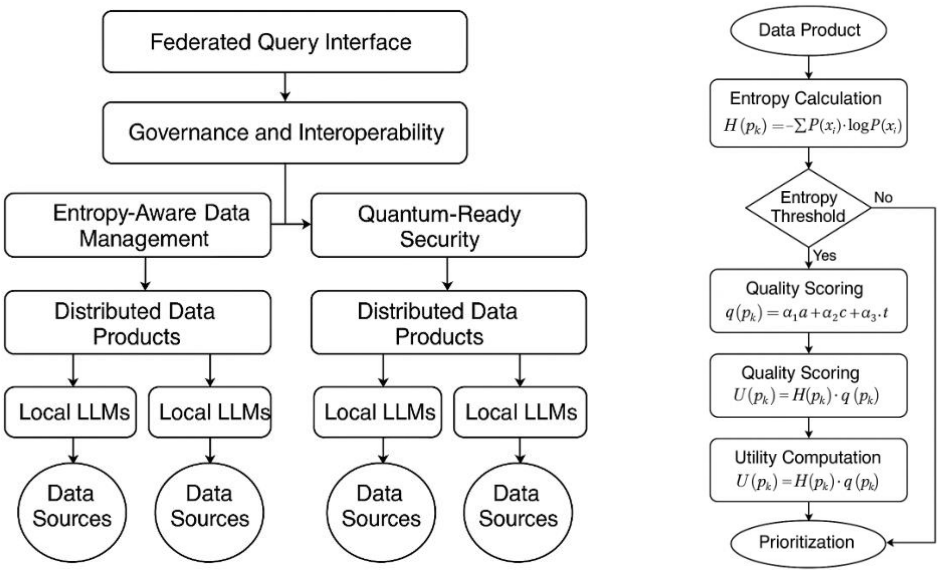

Figure 2. Entropy-Quality Trade-off for Data Product Prioritization

5. INDUSTRY APPLICATIONS

The proposed architecture finds utility across a wide range of domains:

5.1. Healthcare

Federated learning is employed to build predictive models using genomic and electronic health record (EHR) data without centralized data aggregation. Each hospital domain trains a local model M_i using private data D_i. A central coordinator aggregates models using Federated Averaging:

$$M\_global = (1/N) \Sigma M\_i$$

To preserve data privacy, cryptographic methods and differential privacy mechanisms are implemented. Ontologies are encoded as vectors for semantic interoperability.

5.2. Finance

Fraud detection systems are deployed using a multi-agent architecture. Each domain has agents that analyze transaction vectors T = [t_1, ..., t_m] using rule-based scoring functions Φ_i. The final fraud score is computed as:

$$S(T) = \Sigma w\_i \cdot \Phi\_i(T)$$

The mesh architecture allows for real-time correlation of suspicious activities across geographies and business lines.

5.3. Retail

In retail, decentralized forecasting models are built within each region. Time series data including promotional calendars, weather, and events are used to train hybrid ARIMA-LSTM models:

$$y_t = \alpha\, y_{t-1} + \beta\, f(x_t) + \varepsilon_t$$

Where f(x_t) includes contextual features. Forecast outputs are published to a shared data marketplace enabling collaborative planning across brands.

6. LLM AGENTS & ACCESS GOVERNANCE LLM

Agents are integrated within the data platform to manage and enforce access controls, improve user experience, and assist in data discovery. Access to a product p_k by an agent A_LLM is granted only if:

$$\varphi(A\_LLM, p\_k) = True \Leftrightarrow R\_A \cap R\_p\_k \neq \emptyset$$

Where R_A is the set of roles assigned to the agent, and R_p_k defines the permissible roles for accessing the data. To further enhance security, a contextual trust score $\zeta(A, p\_k) \in [0, 1]$ is calculated based on time, IP, user history, and sensitivity of the query.

If $\zeta(A, p\_k) < \tau$ (a threshold), access is denied or partially masked. This mechanism is enforced using policy-as-code frameworks like Open Policy Agent (OPA), and agents are continuously monitored for behavior drift using anomaly detection models.

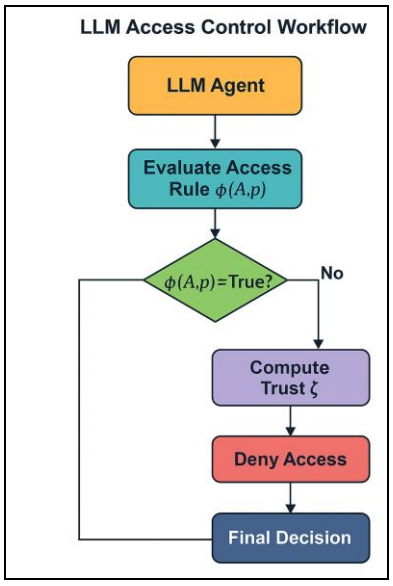

Figure 3. LLM Access Control Workflow

7. QUANTUM-DRIVEN DATA MESH

Quantum databases (QD) offer a new frontier in high-performance data systems. In a quantum-enabled Data Mesh, each domain hosts a quantum node storing entangled data states:

$$|\Psi\rangle = \Sigma\, c\_i\, |d\_i\rangle$$

Cross-domain queries are executed using quantum channels that preserve entanglement. The key challenge is maintaining coherence:

$$\Delta H = H\_pre - H\_post \leq \varepsilon$$

Where ε is the allowable decoherence. Quantum error correction codes such as Shor's or surface codes are applied to protect against noise. Applications include genome similarity search, financial Monte Carlo simulations, and supply chain optimization. Integration with classical mesh nodes is achieved via hybrid quantum-classical orchestration protocols.

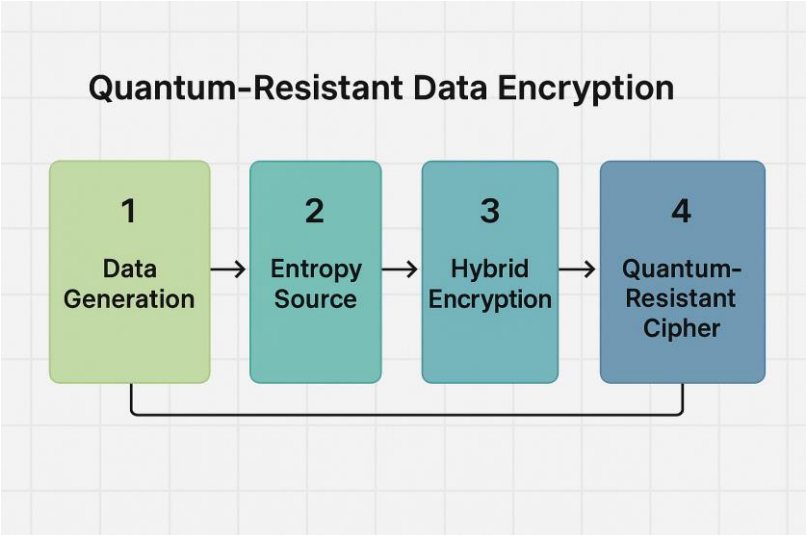

Figure 4. Quantum-Ready Data Mesh Node Network

8. CONCLUSIONS

This paper enhances the NLPI 2025 foundation with rigorous modeling of data platform complexity, entropy-based data valuation, and agent-based governance logic. We proposed a four-layer architecture that integrates LLM agents and anticipates quantum evolution.

Future Research Directions:

- Development of dynamic, learning-based governance agents that evolve access policies in response to organizational changes.
- Application of reinforcement learning to optimize LLM agent workflows in data discovery, access, and compliance.
- Research into quantum indexing structures to reduce query latency in entangled data systems.
- Use of blockchain for trust orchestration in cross-organizational data mesh collaborations.

Emerging Industry Applications:

- **Pharmaceutical R&D:** Secure sharing of clinical trial data to accelerate multi-site drug discovery.
- **Smart Cities:** Real-time coordination of urban services like energy, traffic, and pollution control.
- **Finance & Regulation:** Real-time compliance auditing by regulatory bots embedded in the data mesh.

REFERENCES

[1] Zhamak Dehghani, (2020), "How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh," ThoughtWorks.

[2] Fowler, M., (2003), "Patterns of Enterprise Application Architecture," Addison-Wesley.

[3] Kiran, B., Vohra, D., & Sengupta, S., (2019), "Data Lake for Enterprises: Leveraging Data Lakes for Advanced Analytics," Springer.

[4] G. Piatetsky-Shapiro, (2019), "The Evolution of Data Warehousing and Big Data," KDnuggets.

[5] Z. Li & J. Zhang, (2020), "Federated Data Governance: Balancing Local Autonomy and Global Standards," IEEE Transactions on Data Engineering, Vol. 15, No. 3, pp. 234-245.

[6] Srivastava, J., (2021), "Quantum Databases: Advancing Beyond Classical Data Storage," Journal of Quantum Computing, Vol. 7, No. 2, pp. 95-110.

[7] T. Nguyen & L. Johnson, (2020), "AI-Driven Data Architectures for Business Intelligence," Data Science Quarterly, Vol. 6, No. 4, pp. 88-101.

[8] Zhamak Dehghani, (2022), "Data Mesh: Delivering Data-Driven Value at Scale," O'Reilly Media.

[9] Y. Chen, S. Wang, & R. Patel, (2018), "Decentralized Data Platforms and the Role of Blockchain," ACM Transactions on Information Systems, Vol. 36, No. 5, pp. 423-437.

[10] H. J. Watson, (2018), "Big Data Analytics: Concepts and Techniques," Communications of the ACM, Vol. 61, No. 2, pp. 22-25.

[11] D. Laney, (2012), "The Emerging Role of Data Governance in Modern Organizations," Gartner Research Report.

[12] B. Stonebraker, (2016), "The Case for Data Warehouses in an Era of Data Lakes," IEEE Data Engineering Bulletin, Vol. 39, No. 2, pp. 3-7.

[13] A. Gawande, T. Shroff, & L. Peters, (2019), "Domain-Centric AI Models: Enabling Innovation through Localized Data Architectures," Journal of AI Research, Vol. 12, No. 1, pp. 55-70.

[14] Soumyodeep Mukherjee and Meethun Panda, "General-Purpose Quantum Databases: Revolutionizing Data Storage and Processing", International Journal of Data Engineering (IJDE), Volume (9), Issue (1), 2024 (ISSN: 2180-1274)

[15] Soumyodeep Mukherjee, "The Rise of Multi-Agent LLMs: Insights from Agent Smith and the Challenges of Distributed Data Processing in AI Systems", International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (13), Issue (1), 2024 (ISSN: 2180-124X)

[16] Meethun Panda and Soumyodeep Mukherjee, “Empowering AI and Advanced Analytics through Domain-Centric Decentralized Data Architectures”, 6th International Conference on NLP & Information Retrieval (NLPI 2025), Vol. 15, No. 2, pp. 75-85. (DOI : 10.5121/csit.2025.150506)

AUTHORS

Meethun Panda, Associate Partner at Bain & Company is a thought leader having deep expertise in technology, cloud, Data, AI, LLM, and Quantum computing. He brings 15+ years of experience across technology realms leading and delivering large-scale data and analytics transformations. One of the leading Data/AI consultants in North America by CDO Magazine. Meethun’s key focus is to drive Tech/AI strategy and large-scale transformation cases for fortune 500 clients.

Soumyodeep Mukherjee, Associate Director of Commercial Data Engineering at Genmab (an international biotech company specializing in antibody research for cancer and other serious diseases) is a seasoned data professional with over 14 years of experience in data engineering, architecture, and strategy. Currently steering commercial data initiatives at Genmab, Soumyodeep’s key focus is on crafting innovative data and analytics strategies to drive commercialization efforts.

Previously, he served as a Project Leader at BCG.X and a Data Specialist at McKinsey & Company, where he led teams in implementing robust, end-to-end data solutions across healthcare, insurance, and retail sectors. His expertise includes deploying machine learning models and leveraging Generative AI to streamline data management and enhance organizational efficiency.