International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I3P107
Eureka Vision Publication | Volume 6, Issue 3, 43-47, 2025

Original Article

A Survey of Petabyte-Scale Data Architectures for Self-Serve Generative AI: From Foundational Systems to Intelligent Abstraction and In-Database Optimization

Pinaki Bose
Independent Researcher, USA.

***Received On:* 13/06/2025** ***Revised On:* 17/06/2025** ***Accepted On:* 21/07/2025** ***Published On:* 09/08/2025**

Abstract *- The integration of generative AI (GenAI) into self-service analytics platforms, such as Microsoft Power BI Copilot and Amazon Q, has necessitated a paradigm shift in data architecture. While these tools aim to democratize data access through natural language interfaces, their efficacy is contingent on an underlying data foundation that can manage the challenges of petabyte-scale data. This paper presents a survey and comparative analysis of three advanced architectural strategies designed to address the issues of query latency, cost, and semantic ambiguity inherent in such large datasets. The strategies examined are: (1) the Optimized Data Lakehouse, which focuses on foundational performance and cost-efficiency through open-source formats and high-performance query engines; (2) the Enterprise Semantic Layer and Knowledge Graph, an abstraction-first approach that ensures data consistency and mitigates AI hallucination by providing a structured context; and (3) the Instance-Optimized LLM (IOLM-DB), a specialized, in-database method that overcomes the high cost of per-row AI inference. A detailed analysis reveals that no single solution is universally optimal. Instead, the most effective approach for many organizations is a phased, hybrid architecture that combines the strengths of the first two strategies, reserving the third for specific, high-value applications. This framework provides a robust roadmap for building an AI-native data platform that is both performant at scale and semantically intelligent.*

Keywords *- Generative AI, Data Architecture, Data Lakehouse, Semantic Layer, Knowledge Graph, Petabyte-Scale, Self-Service Analytics, Large Language Models (LLMs).*

1. Introduction

The modern paradigm of business intelligence (BI) is undergoing a significant transformation, driven by the convergence of generative AI and self-service analytics tools. This evolution, exemplified by platforms such as Microsoft's Copilot for Power BI and Amazon Q, is shifting the responsibility of data querying from specialized data analysts to a broader audience of business users who can interact with data via natural language interfaces [1]. This democratization of data promises to deliver instant insights and reduce the time required for data-driven decision-making [2, 3]. However, the success of this shift is fundamentally dependent on the underlying data infrastructure. As documented, the effectiveness of these AI assistants is directly linked to the quality of the data foundation, as a poorly structured or ambiguous semantic model can lead to errors and diminished value [1].

The core challenge lies in navigating the "computational constraints" of petabyte-scale data processing [4]. At this immense scale, a single query is a complex operation that involves thousands of parallel tasks across numerous compute nodes. The primary computational constraints are:

- Data Locality and Movement, as separating computing and storage can lead to high latency and significant egress fees;
- Resource Management, where a single inefficient query can lead to cascading failures.
- Cost, where the financial implications of large-scale storage and computation can be prohibitive. These challenges have led many organizations to seek alternatives to traditional managed data warehouses, which may not offer the necessary flexibility or cost predictability.

2. This Paper Presents Three Distinct Architectural Strategies to Address these Challenges, Moving from Foundational Approaches to Highly Specialized Solutions:

- The Optimized Data Lakehouse: A foundational approach prioritizing raw query performance and cost-effectiveness.
- The Enterprise Semantic Layer and Knowledge Graph: A strategy focused on data consistency and context through abstraction.

- The Instance-Optimized LLM (IOLM-DB): A cutting-edge approach designed to solve the high cost of per-row AI inference.

A detailed comparative analysis evaluates these strategies based on performance, cost, governance, and technical complexity, concluding that a hybrid, phased implementation is the most robust solution for a majority of enterprises.

2.1. Strategy 1: The Optimized Data Lakehouse with High-Performance Query Engines

This strategy advocates for a foundational architecture that combines the scalability and cost-efficiency of a data lake with the transactional integrity and performance of a data warehouse, a pattern known as the data Lakehouse. A key principle of this approach is the use of open storage formats, such as Apache Iceberg, which allows data to be stored in a single, open-source copy while remaining accessible to various query engines. This separation of compute and storage provides significant advantages, including independent resource scaling and the avoidance of vendor lock-in [5, 6].

To meet the low latency demands of self-service GenAI, this architecture relies on high-performance, distributed query engines like StarRocks and ClickHouse. These engines are specifically designed for petabyte-scale Online Analytical Processing (OLAP) workloads and achieve sub-five-second query latency through advanced techniques such as vectorized query processing and advanced caching [5, 7]. Vectorized processing leverages modern CPU capabilities to process data in batches, dramatically accelerating query execution. Advanced caching uses local SSDs to store frequently accessed data, reducing reliance on main cloud storage. The ability of a single platform to serve as a unified data store for both classical analytics and AI workloads simplifies the data stack and reduces operational overhead, representing a new paradigm for data infrastructure [7].

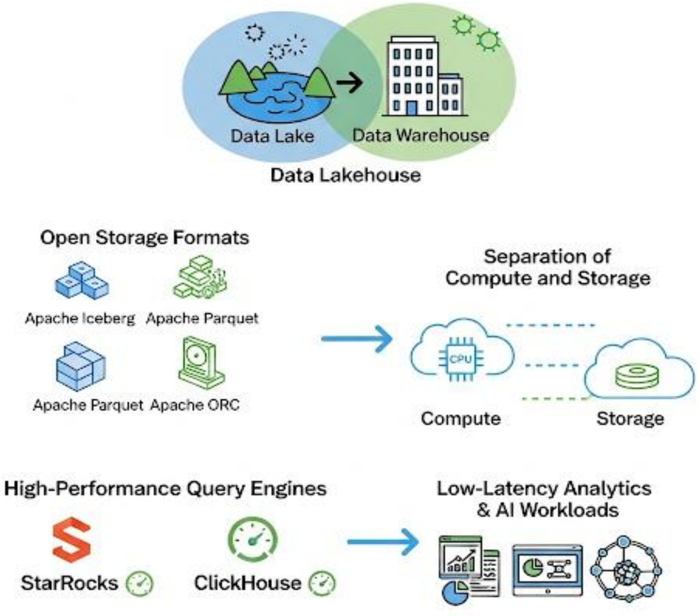


Figure 1. The Optimized Data Lakehouse with High-Performance Query Engines

2.2. Strategy 2: The Enterprise Semantic Layer and Knowledge Graph

In contrast to Strategy 1, this approach prioritizes intelligent abstraction over raw performance. The core component is the semantic layer, which functions as a centralized, machine-readable model of an organization's business metrics and rules [9]. This layer serves as a single source of truth, providing consistency in data definitions and mitigating the risk of AI hallucination by giving LLMs a clear, contextual framework.

The semantic layer is a dynamic system that learns from user interactions, operating in a closed-loop feedback cycle where prompts and clarifications are used to refine and update the model [10]. This transforms it from a passive metadata catalog into an active, self-improving reasoning framework.

A critical component of a robust semantic layer is the knowledge graph (KG). The KG acts as an orchestration and discovery layer, unifying structured and unstructured data and making enterprise knowledge understandable to both humans and machines [11]. Research indicates that knowledge graphs are a "missing piece" for modern data architectures like data fabric and data mesh, as they provide the dynamic context required for GenAI tools to reason and infer new insights [12]. The fusion of a data fabric with an enterprise semantic layer and knowledge graph is seen as the future of data strategy, particularly in complex industries where harmonizing disparate data sources is essential [13].

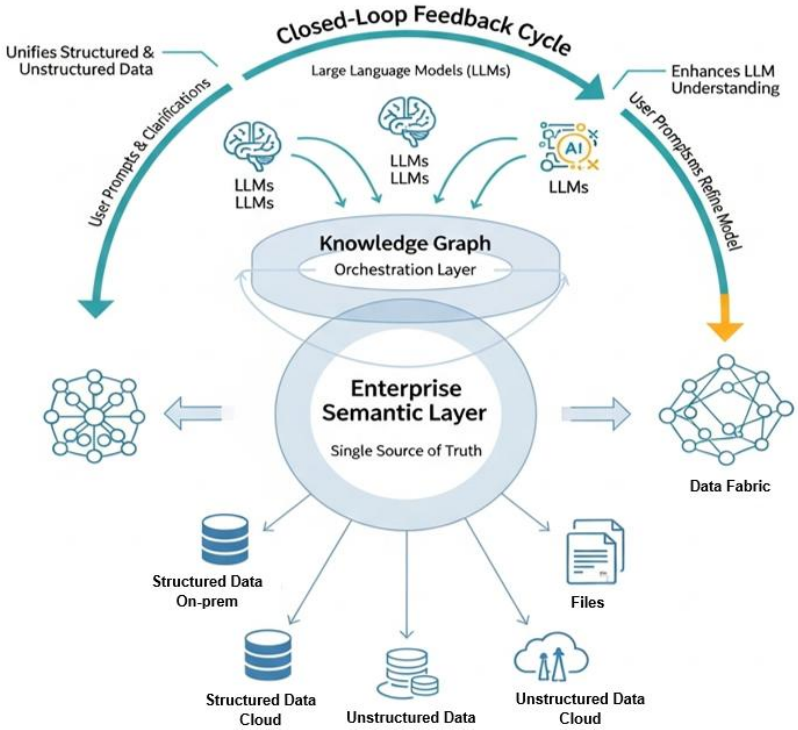

Closed-Loop Feedback Cycle
Unifies Structured & Unstructured Data
Large Language Models (LLMs)
Enhances LLM Understanding
User Prompts & Clarifications
User Promptsms Refine Model
LLMs LLMs
LLMs LLMs
AI
LLMs
Knowledge Graph
Orchestration Layer
Enterprise Semantic Layer
Single Source of Truth
Data Fabric
Structured Data On-prem
Files
Structured Data Cloud
Unstructured Data
Unstructured Data Cloud

Figure 2. The Enterprise Semantic Layer and Knowledge Graph

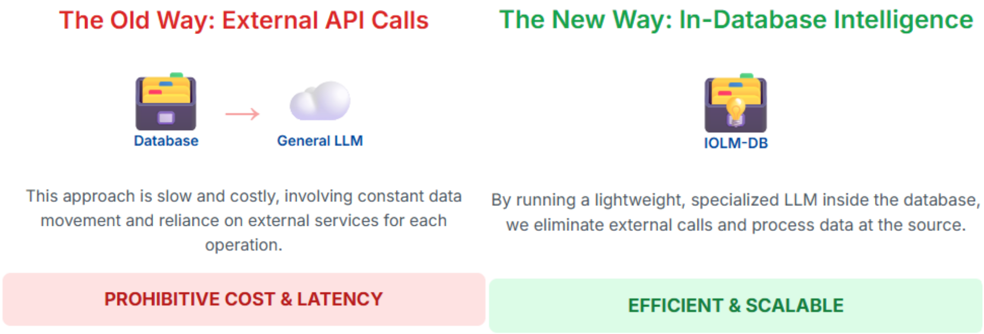
2.3. Strategy 3: The Instance-Optimized LLM (IOLM-DB) and In-Database Intelligence

This highly advanced strategy addresses the fundamental limitation of current GenAI applications: the prohibitive cost and computational overhead of using general-purpose LLMs for per-row data processing [15]. Instead of relying on external API calls, this approach proposes the creation of lightweight, specialized models tailored to specific query needs. These Instance-Optimized LLMs (IOLM-DB) operate as in-database functions, making LLM-enhanced queries practical for massive-scale analytics.

The technical foundation of this strategy relies on three key techniques:

- Quantization: Reduces the model's numerical precision to decrease memory footprint and computational overhead.
- Sparsification: Imposes sparsity patterns on model weights to reduce the number of active parameters.
- Structural Pruning: Removes entire components of the model that contribute minimally to a specific task.

These techniques are combined to aggressively reduce model size and cost while maintaining accuracy. Research has demonstrated significant throughput improvements with this approach, with one prototype showing a substantial increase in processing speed for a summarization task [16]. This strategy is best suited for a small subset of repeatable, high-value operations—such as data cleansing or fuzzy joins—and is not a universal solution for general self-service queries due to its complexity and specialized nature [17].



The Old Way: External API Calls

This approach is slow and costly, involving constant data movement and reliance on external services for each operation.

PROHIBITIVE COST & LATENCY

The New Way: In-Database Intelligence

By running a lightweight, specialized LLM inside the database, we eliminate external calls and process data at the source.

EFFICIENT & SCALABLE

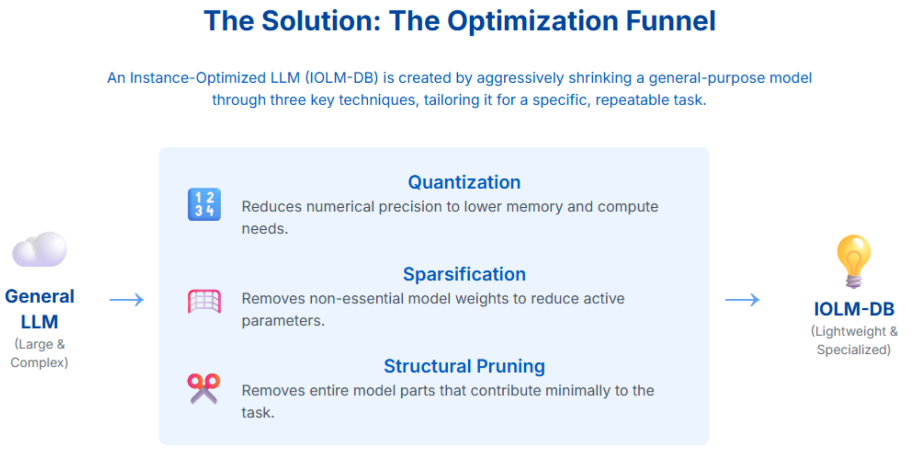

The Solution: The Optimization Funnel
An Instance-Optimized LLM (IOLM-DB) is created by aggressively shrinking a general-purpose model through three key techniques, tailoring it for a specific, repeatable task.
General LLM
(Large & Complex)
Quantization
Reduces numerical precision to lower memory and compute needs.
Sparsification
Removes non-essential model weights to reduce active parameters.
Structural Pruning
Removes entire model parts that contribute minimally to the task.
IOLM-DB
(Lightweight & Specialized)

Figure 3. The Instance-Optimized LLM (IOLM-DB) and In-Database Intelligence

3. Comparative Feasibility Analysis & The Strategic Trade-offs

A multi-dimensional framework is necessary to evaluate these three strategies, considering not just technical pros and cons but also performance, cost, governance, talent acquisition, and time-to-value.

Table 1. Comparative Feasibility Analysis & The Strategic Trade-offs

Feature	Strategy 1: Optimized Data Lakehouse	Strategy 2: Semantic Layer & KG	Strategy 3: Instance-Optimized LLM
Primary Value Proposition	Raw performance and cost control.	Data consistency and business context.	Per-query efficiency and scalability.
Core Technologies	Lakehouse (Iceberg), Query Engines (StarRocks, ClickHouse).	Semantic Layer (YAML), Knowledge Graph.	Instance-Optimized LLMs (IOLM-DB), Quantization.
Key Strength	Handles ultra-low latency, high-concurrency BI and ML workloads; avoids vendor lock-in.	Eliminates data ambiguity, prevents hallucinations, and enables democratized access.	Dramatically reduces cost and latency for repetitive, high-volume LLM tasks.
Key Weakness	Requires deep data engineering expertise; does not solve the semantic ambiguity problem alone.	Does not solve raw query performance issues on petabyte scale; requires strong governance.	Highly complex, research-level; not a universal solution for all query types.
Ideal Scenario	An organization with a strong data engineering team, strict performance needs, or a multi-cloud/on-premise strategy.	An organization seeking to democratize data for a broad, non-technical audience.	An organization with specific, high-volume, and repeatable LLM-enhanced data tasks (e.g., data cleansing).
Operational Overhead	High. Requires continuous optimization.	Medium. Requires governance and modeling.	Very High. Requires specialized R&D talent.

The strategies represent different architectural philosophies. Strategy 1 is a "build-it-yourself" approach focused on foundational performance and control, while Strategy 2 is an "abstraction-first" approach that prioritizes business enablement and consistency. The IOLM-DB approach (Strategy 3) is a highly specialized optimization rather than a general-purpose solution.

The optimal choice depends on an organization's specific goals and technical maturity. For a "democratization-first" enterprise, prioritizing Strategy 2 offers immediate value. For a "performance-driven" organization, Strategy 1 is a prerequisite for any advanced capabilities. Strategy 3 is reserved for the "AI-native" enterprise with specific, high-cost use cases and specialized technical talent.

4. Conclusion

The future of a petabyte-scale data architecture for self-serve GenAI is not defined by a single technology but by a unified, multi-layered platform. This platform must address the core challenges of performance, semantic ambiguity, and computational cost. The most robust and defensible long-term strategy for any enterprise is a phased, hybrid implementation. This roadmap begins by building a modern data Lakehouse (Strategy 1) to establish a foundation for petabyte-scale data ingestion and low-latency querying. Once this foundation is stable, an enterprise semantic layer and knowledge graph (Strategy 2) should be layered on top to provide the business context and governance required for GenAI tools to function effectively. Finally, for specific, high-value operations, targeted optimizations like the IOLM-DB approach (Strategy 3) can be implemented to achieve new levels of efficiency. This phased approach allows organizations to build a data platform that is flexible, governed, and intelligently optimized, empowering a new era of data-driven decision-making.

References

[1] Microsoft, "Prepare your data, your semantic model, and your users for Copilot for Power BI," learn.microsoft.com. Available: https://learn.microsoft.com/en-us/power-bi/create-reports/copilot-semantic-models.

[2] Seisma Group, "Smarter Data with Microsoft Fabric Copilot for Power BI," seismagroup.com. Available: https://www.seismagroup.com/news/smarter-data-with-microsoft-fabric-copilot-for-power-bi.

[3] Silicon Republic, "Artificial semantic layer is a missing piece of the GenAI puzzle," siliconrepublic.com. Available: https://www.siliconrepublic.com/enterprise/business-intelligence-artificial-semantic-layer-genai-data.

[4] S. Saifi, "Beyond Snowflake: What Actually Happens When You Query Petabytes of Data," Medium, 2024. Available: https://medium.com/@sohail_saifi/beyond-snowflake-what-actually-happens-when-you-query-petabytes-of-data-7bd57cbc52df.

[5] T. O'Sullivan, "From BigQuery to Lakehouse: How We Built a Petabyte-Scale Data Analytics Platform," trmlabs.com. Available: https://www.trmlabs.com/resources/blog/from-bigquery-to-lakehouse-how-we-built-a-petabyte-scale-data-analytics-platform-part-1.

[6] B-Eye, "Modern Data Platform Blueprint," b-eye.com. Available: https://b-eye.com/blog/modern-data-platform-blueprint/.

[7] ClickHouse, "Use cases: Machine Learning and Data Science," clickhouse.com. Available: https://clickhouse.com/use-cases/machine-learning-and-data-science.

[8] H. PMP, "Building the Complete Modern Enterprise Data Architecture: A Comprehensive Guide," Medium, 2023. Available: https://hamidpmp.medium.com/building-the-complete-modern-enterprise-data-architecture-a-comprehensive-guide-2c48f003942b.

[9] A. Johnson, "The Ultimate Guide to Semantic Layers for AI," promptql.io. Available: https://promptql.io/blog/the-ultimate-guide-to-semantic-layers-for-ai.

[10] Orange Business, "How AI is transforming self-service analytics and BI," perspective.orange-business.com. Available: https://perspective.orange-business.com/en/how-ai-is-transforming-self-service-analytics-and-bi-and-what-you-need-to-get-right-first/.

[11] Enterprise Knowledge, "Data Management Trends in 2022: Data Fabric v. Data Mesh v. DataOps," enterprise-knowledge.com. Available: https://enterprise-knowledge.com/data-management-trends-in-2022-data-fabric-v-data-mesh-v-dataops-what-is-right-for-your-organization/.

[12] Ontotext, "How Knowledge Graphs Power Data Mesh and Data Fabric," ontotext.com. Available: https://www.ontotext.com/blog/how-knowledge-graphs-power-data-mesh-and-data-fabric/.

[13] Ontoforce, "Gartner: Semantic Technologies Take Center Stage in 2025," ontoforce.com. Available: https://www.ontoforce.com/blog/gartner-semantic-technologies-take-center-stage-in-2025.

[14] Amazon Web Services, "Amazon Q: The generative AI assistant for business," aws.amazon.com. Available: https://aws.amazon.com/q/.

[15] A. Z. et al., "The Case for Instance-Optimized LLMs in OLAP Databases," arXiv preprint arXiv:2507.04967v1, 2025. Available: https://arxiv.org/html/2507.04967v1.

[16] A. G. et al., "The Case for Instance-Optimized LLMs in OLAP Databases," ResearchGate. Available: https://www.researchgate.net/publication/393477387_The_Case_for_Instance-Optimized_LLMs_in_OLAP_Databases.

[17] "Decoding LangChain's Structured LLM Calls for Model Fine-Tuning," Medium, 2024. Available: https://blog.gopenai.com/decoding-langchains-structured-llm-calls-for-model-fine-tuning-eaea34710783.

[18] Data Science Collective, "Comprehensive Guide to Fine-Tuning LLM," Medium, 2024. Available: https://medium.com/data-science-collective/comprehensive-guide-to-fine-tuning-llm-4a8fd4d0e0af