Unlocking Innovation

How AI-Ready Businesses See Data Architecture

IDEA | INSTITUTE FOR DATA & ENTERPRISE AI

I. Introduction: Why Architecture Matters

Generative AI could unlock **US$2.6–4.4 trillion annually** for the global economy [McKinsey]. For a share of that staggering prize, most enterprises are racing to mature their data capabilities, but many are building on unstable and uncertain foundations.

Our first research brief explored the trust deficit where two-thirds of business leaders don't fully trust their data and why trust must be the foundation of all successful data initiatives.

This brief explores an adjacent gap: **how to build scalable data architecture systems that generate trust and innovation at the speed AI demands**. It also features insights and conversation highlights from a recent roundtable discussion convened by the IDEA Institute in London. This event brought together senior data leaders in an informal setting to explore practical approaches and different perspectives on data architecture.

Part of the architecture answer lies in the evolution from Big Data to better data—from Volume, Velocity, and Variety to the latest additions to the 5Vs of Validity and Value.

Before putting AI's full capabilities to work, we must fundamentally rethink how to architect, govern, and integrate data. And how to embed trust into every data interaction. As without a rock-solid architectural foundation, AI initiatives become exercises in amplifying data risk at scale.

Data fabric is the architecture that makes rapid progress possible with **active metadata as a key component**. Through active metadata and treating data as a product rather than an operational byproduct, we can create the architecture of trust that the future demands. Data innovation and data fabric working together, democratizing access while maintaining governance that builds trust and confidence.

But first, some large numbers about big and better data.

II. The Cost of Data Fragmentation

By 2025, organizations will be managing **175 zettabytes of data** [IDC] with **80% of it unstructured** [Forrester]. That's equivalent to streaming 960 billion hours of Netflix movies every day for a year.

Despite decades of investment in data infrastructure, **66% of business leaders don't fully trust their data** and **47% struggle with data fragmentation** [Qlik].

Poor data quality costs organizations an average of **US$12.9 million annually** [Gartner, Dataversity]. But the true impact—including missed opportunities and strategic missteps—is growing. As a roundtable guest posited, “How much benefit are you going to get from your AI if the underlying data needs constant validation?”

AI has simultaneously created demand for quality data while degrading overall quality. By 2025, some experts predict that up to 90% of online content could be generated by AI.

78%

As of mid-2025, 78% of organizations worldwide are using artificial intelligence in at least one business function, with adoption continuing to expand across multiple domains and functions.

Source: McKinsey State of AI

II. The Cost of Data Fragmentation

Where human analysts might catch data quality issues, AI systems can amplify them at machine speed. And legacy architectures that could barely handle traditional analytics now face the real-time, high-volume demands of AI workloads.

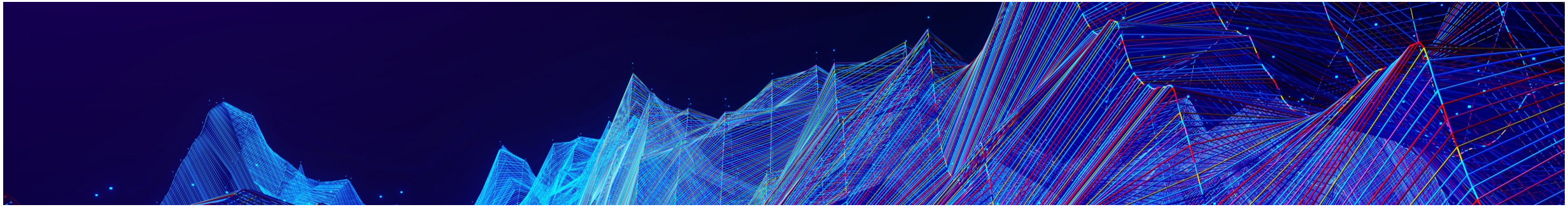
Point solutions layered upon point solutions have created friction instead of flow—systems that work in isolation just can't deliver what AI applications demand. A point that also arose early in the roundtable discussions, “The infrastructure around AI and the skills to drive that technology—that's where you have to have everything working together. The quality of the foundation, the supporting systems.”

Too much data? Or maybe AI is exposing data architecture gaps we already know exist?

Global personalization @ Netflix

When you open Netflix in any of 190+ countries, the recommendations feel tailored just for you. Behind this experience sits a sophisticated data fabric that seamlessly blends viewing behavior, content metadata, A/B testing results, and operational metrics. The magic isn't in any single algorithm—the architecture treats each data domain as a product with clear APIs and service levels. Netflix stores viewing data in Parquet format, so any analytics tool to access the same datasets without conversion or copying for huge scale and agility.

Source: Netflix



III. From Big Data to Better Data

A brief history of best intentions

How did we get to a place where AI lacks the data it needs—high quality, governed, and trusted? It turns out that the evolution of data architecture is a series of solutions that created new problems.

First (in the mid 2000s), enterprise data warehouses promised centralized storage that would democratize access. Instead, it was too easy to dump data without context, making value extraction near impossible. But in practice, they proved costly to build, difficult to scale, and rigid to maintain, often creating bottlenecks instead of agility.

In response, data lakes emerged to offer more flexibility and lower storage costs. But without strong governance or metadata, many lakes turned into data swamps where data went to retire, lacking the governance and context needed for data discovery and trust.

The solution to data swamps isn't another architecture—it's 'open table' interchangeable formats like Parquet, Iceberg, Delta Lake, and Hudi. These transform raw storage by acting as universal translators that let any tool read any data, breaking vendor lock-in while maintaining performance.

> *"Technology is rarely the bottleneck—the problem is almost always organisational readiness and alignment."*
>
> — **MARCUS BEARDEN**
> ***Technology Partner, Gartner***

III. From Big Data to Better Data

It’s a less disruptive and more strategic approach. So, instead of dismantling data silos, zero-copy access and intelligent query layers work together to pull the right data, at the right time, from wherever it lives. This creates faster time to value, fewer data movement costs, and more flexibility for innovation.

McKinsey compares these historical architecture issues to railroad freight. It would be wildly inefficient to run a network with a different engine pulling each individual car. Instead, a standardized system with one powerful engine pulling any number of cars and different cargo to multiple destinations.

Architecture choice: Getting on the right train

Choosing the right architecture benefits everyone by building trust at scale and creating the conditions for innovation to thrive.

In the same way that passengers trust a well-managed railway system to deliver them safely and on time, business users and AI applications must trust that data fabric will deliver accurate, governed data consistently. Across a train network (a data product) that supports many journeys (use cases). Powered by an intelligent engine (active metadata) that adapts to different loads and conditions. And with an experienced conductor (governance) and robust coupling (integration) so the whole system moves as one.

While traditional single source of truth models made sense when data moved slowly and use cases were predictable, today's distributed, real-time demands need the best of all worlds. That’s where the concept of ‘data lakehouse’ combines the strengths of warehouses and lakes—all within a data fabric architecture.

III. From Big Data to Better Data

Data fabric and lakehouse synergy

Aspect	Data Lakehouse	Data Fabric	Integration Benefit
Primary focus	Unified storage/processing platform	Cross-platform data access & governance	Fabric enables lakehouse interoperability with other systems
Data movement	Centralizes data in one repository	Minimizes data copying via virtualization	Fabric queries lakehouse data in place
Governance	Built-in for its data	Enforces policies across all sources	Fabric extends lakehouse governance enterprise-wide
Use case scope	Optimized for analytics/ML workloads	Supports operational/real-time needs	Fabric routes workload-appropriate data to lakehouse

IV. Data Fabric Architecture for AI-Ready Businesses

Defining data fabric

Think of it as the connective tissue between disparate data sources, tools, and processes, binding these into a coherent, intelligent system. Core capabilities—integration, cataloging, curation, lineage, security, and abstraction—work together to solve the fragmentation challenge. Data fabric makes data both accessible and trustworthy at scale.

This collaborative architectural thinking extends beyond company boundaries through data spaces (ODI). This is the infrastructure that enables trusted data transactions between ecosystem parties based on shared frameworks. In data spaces, data remains at source while access is granted in controlled, secure ways.

Both data fabric and data spaces prioritize data sovereignty and federated governance, demonstrating how data architecture is moving toward distributed yet coordinated systems. And how important these types of data sharing (between companies and more widely with the public) will be in the future.

90%

By 2025, experts predict that up to 90% of online content could be generated by AI

Source: Nina Schick

IV. Data Fabric Architecture for AI-Ready Businesses

From Big Data to better data

Big Data has three dimensions: Volume (how much), Velocity (how fast), and Variety (how diverse).

While these foundational dimensions still need to be tackled and brought under control, the AI era introduces two additional Vs, elevating data from merely big to genuinely better:

- Validity
  Ensuring data is accurate, complete, and contextually appropriate for its intended use. This goes beyond traditional data quality metrics to include lineage tracking, bias detection, and fitness-for-purpose assessments that change based on business context.
- Value
  The ultimate test of measurable business outcomes. More than monetization—it's better decisions, experiences, and efficiencies. Data fabric provides the governance, integration, and active management that turns raw data into value.

IV. Data Fabric Architecture for AI-Ready Businesses

1. Active metadata: The Intelligence Engine

Traditional metadata was like a thermometer—it could tell you the temperature but couldn't do anything about it. Active metadata acts like a thermostat, so it can automatically adjust data quality, routing, and access on the fly.

Intelligent orchestration engine (a Gartner term) uses AI and ML to automatically discover relationships between data and self-optimize. It generates quality ratings based on data lineage and the context of business use, not generic standards and metrics.

This intelligence enables the data fabric to self-regulate and improve over time. When a data source's quality degrades, active metadata can automatically re-route requests to better sources. It identifies patterns and suggests new ways to integrate the data and flag potential governance issues before they become problems. Active metadata's self-organizing properties make data much more dynamic.

2. Data fabric, deeper integration

In data fabric architecture, a single query might pull historical data through batch processes, current state through APIs, and real-time updates through streaming—all orchestrated automatically by the fabric.

In this **federated governance** approach, different domains own and manage data products while maintaining enterprise-wide consistency and discoverability. Rather than forcing all data through a central bottleneck, data fabric distributes governance intelligence throughout the system while maintaining coordinated oversight.

As a roundtable attendee pointed out, "Organizations are centralizing the definition of data—the metadata—rather than the data itself. The knowledge graph becomes decentralised with business groups, while data remains in different sovereign locations."

V. How Enterprise Leaders Are Scaling

Start composable, not set in stone

The most successful data fabric implementations prioritize modularity and API-first architecture. Think LEGO blocks, not concrete—each component serves multiple purposes and connects easily with others.

No data accountability bottlenecks

Distributed ownership means each data domain team owns their data products with clear contracts, quality guarantees, and SLAs. The ideal future state is data producers are responsible for quality, completeness, and context, so that users can trust the data without extensive validation.

Data readiness is always on

Financial reporting demands decimal-point accuracy. Customer experience initiatives prioritize real-time insights over perfect precision. Data fabric architecture enables these context-specific approaches through intelligent routing and active metadata's dynamic properties.

Real-time trading confidence @ Goldman Sachs

Goldman Sachs reduced trading risk by 30% by embedding data quality scores into trading systems. When market data arrives, active metadata automatically assess its lineage and reliability, providing traders with confidence indicators alongside price information.

Source: Goldman Sachs

70%

By 2026, generative AI will significantly alter 70% of the design and development efforts for new web applications and mobile apps

Source: Gartner

V. How Enterprise Leaders Are Scaling

Active metadata flags quality issues

Self-describing systems automatically catalog new data sources, suggest governance policies, and optimize performance based on usage patterns. This doesn't replace human judgment—it augments it with machine-scale monitoring and adjustment capabilities.

Knowledge graphs: Intelligent discovery

Knowledge graphs capture the data relationships and context that create value—like how CRM records connect to transactions, and how sustainability data links with supply chains. This semantic blueprint is the basis of automated discovery, intelligent routing, and contextual recommendations—and it makes data fabric architecture truly scalable.

Bringing data together, securely

Integrating structured data and unstructured content requires data fabric's abstraction layer. This makes diverse data types equally accessible while maintaining security and governance controls.

Observability into everything

Unlike traditional monitoring focused on technical metrics, data fabric observability tracks business outcomes, user satisfaction, and value generation. This wide-angle visibility enables proactive optimization and demonstrates ROI of data investments.

“Organizations that invested early in foundational capabilities like metadata management and data governance are starting to see tangible returns, the others are now playing catch up and many don't fully understand what they need to really make the most of where AI and data are taking them.”

— STUART COLEMAN
Director, Open Data Institute

VI. Pragmatic Steps to Building Better Data

Data Fabric Readiness Checklist

Step 1: Foundation

- [x] **Map your data landscape**

  Audit critical data sources and identify 2-3 high-impact domains where fragmentation undermines business performance

- [ ] **Establish ownership**

  Create cross-functional teams (including data producers and users) with shared accountability and link data stewardship to business outcomes, as well as technical metrics

Step 2: Core Architecture

- [x] **Deploy intelligent cataloging**

  Implement augmented data catalogs with knowledge graphs for priority domains to capture relationships and enable semantic discovery

- [ ] **Automate governance**

  Build compliance and quality checks directly into data pipelines with automated policy enforcement

VI. Pragmatic Steps to Building Better Data

Data Fabric Readiness Checklist

Step 3: Scale and Culture

- [x] **Enable self-service**

Launch data product marketplace with certified datasets and built-in guardrails for business user access

- [ ] **Build innovation capacity**

Establish data champion programs and innovation sandboxes where teams can rapidly experiment with new data combinations

Standardization with flexibility @ Toyota Motor Europe

Toyota's "freedom in a box" approach unites integration methods, governance frameworks, and quality tools while allowing adaptation to local market requirements. This balance helped the company ride out supply chain disruptions when regional teams needed to adapt quickly while maintaining coordination with global operations.

Source: Atscale

60%

By 2026, 60% of leading enterprise intelligence companies will have identified data products, and 15% will have attributed business value to the products

Source: IDC

“Data fabric architecture does more than solve modern data management challenges—it builds the foundation of trust that makes innovation possible at enterprise scale.”

— DAN YU
CMO, SAP Data & Artificial Intelligence

VII. The Architecture Advantage

Data's definition is shifting from a single source of truth to contextual truth. So, organizations need architecture that consistently delivers the right data and the right quality at the right time—where trust becomes automatic rather than earned. Then teams stop questioning reliability and start innovating with confidence—while Big Data becomes genuinely better data.

This creates a powerful cycle: architecture generates trust and trust accelerates innovation. Or as a roundtable participant put it, "Data fabric doesn't just store and move data, it manufactures trust." The result is trust at scale that supports hundreds of use cases simultaneously rather than requiring continuous validation.

In this way, teams can rapidly experiment, AI models train on rock-solid foundations, and insights flow directly into business decisions without bottlenecks. Meanwhile, organizations with fragmented architectures find every innovation project stalled by trust deficits that their architecture perpetuates.

Architecture has evolved from IT plumbing to trust infrastructure—determining how confidently businesses can act on data-driven insights. The key takeaway from roundtable discussions was the urgency for organizations to build this foundation of trust now or be trapped by architectures where trust is too expensive to achieve or too fragile to scale.

In the AI era, data architecture doesn't just support strategy—it determines what's strategically possible.

> "Advancing data maturity requires as much investment in people and mindset as it does in platforms and tools."
>
> **— London roundtable participant**

Stay Connected:

Membership - IDEA INSTITUTE

IDEA | INSTITUTE FOR DATA & ENTERPRISE AI

Appendix

Glossary of data terms

Data fabric: An integrated, metadata-driven architecture providing seamless data access, integration, and governance across hybrid and multi-cloud environments.

Active metadata: Self-regulating, intelligent data management systems that use AI and automation to provide proactive recommendations and automatic optimization rather than reactive monitoring.

Data lakehouse: Modern architecture combining data lake flexibility with data warehouse structure and performance, supporting both analytics and AI workloads on unified platforms.

Data mesh: Decentralized, domain-oriented approach to data architecture emphasizing data as a product with federated governance and self-serve infrastructure.

Data product: Data treated as a product with clear ownership, defined service levels, usage metrics, quality guarantees, and continuous improvement based on consumer feedback.

Intelligent query layer: An abstraction that orchestrates, optimizes, and federates queries across diverse data sources and formats, enabling real-time access, streamlined security, and performance tuning without moving or replicating underlying data.

Zero-copy access: An approach that allows analytical and processing tools to directly query and use data where it resides—without creating duplicate copies—eliminating unnecessary data movement, reducing costs, and maintaining data consistency across platforms.

Data fabric elements and examples

Data fabric architectures use these elements to help organizations break down data silos, automate governance, and optimize data operations. These examples show how enterprises across industries are driving innovation and data-driven decision-making.

Data Fabric Element	Use case	Case study reference
Active metadata	Automated connector generation, data product creation	Nexla, data integrator with extensive use cases of active metadata - LINK
Auto-cataloging	Inventory and enrichment of new data sources	Modern data catalogs and how they help get your data in order LINK
Unified data source	Health data management for better patient outcomes eg 360-degree patient view	Healthcare use cases LINK
Real-world integration	Unified data for analytics and operations	Kroger, food manufacturer LINK