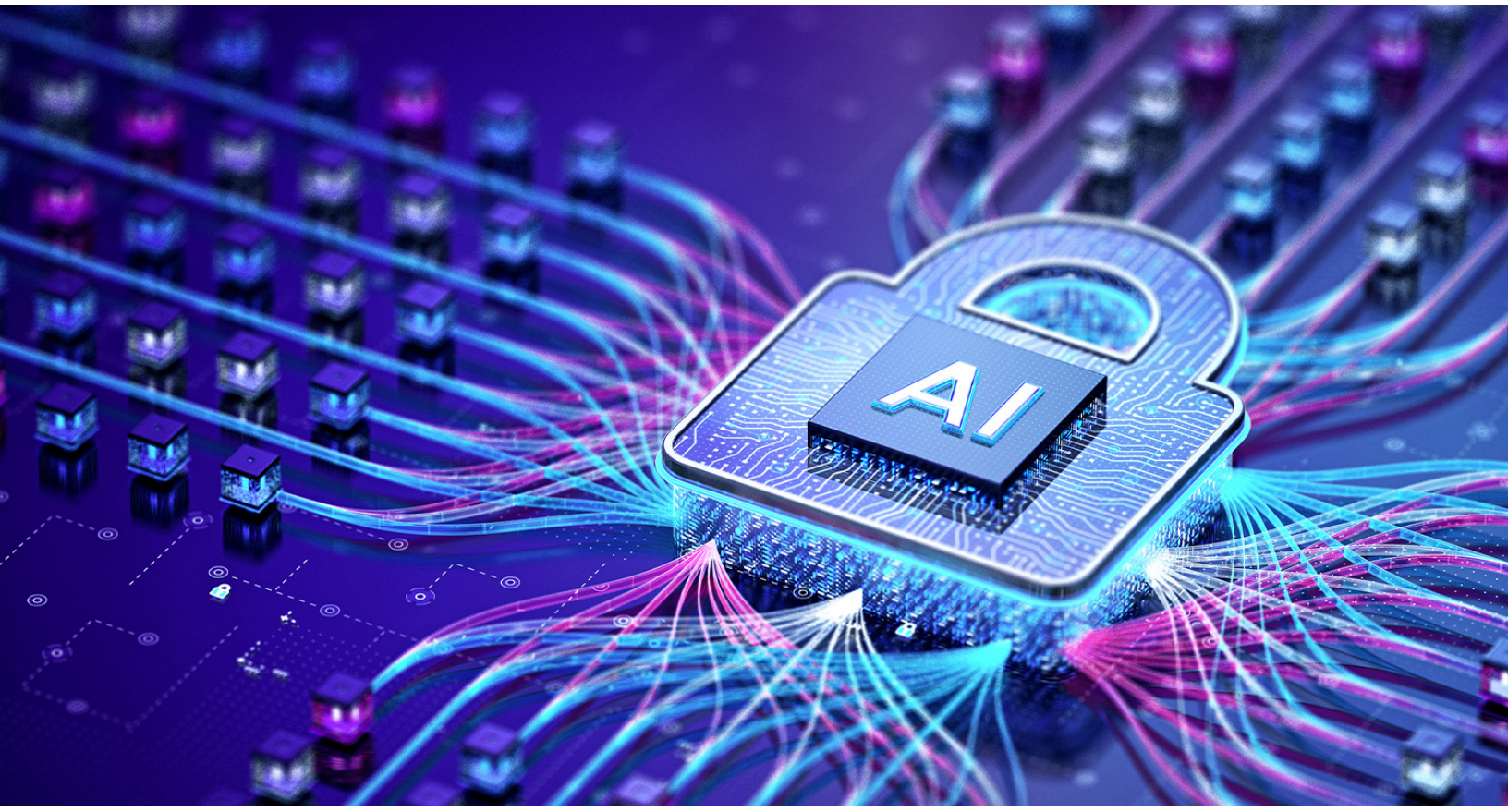
McKinsey & Company

McKinsey Technology Practice

AI data readiness: The key to scaling impact

As companies push their AI pilots to scale, data is emerging as a constraint. Leaders are prioritizing data readiness, connecting structured and unstructured data into a governed, reusable foundation.

This article is a collaborative effort by Asin Tavakoli, Brian Goodman, Kayvaun Rowshankish, and Stephen Reddin, with Camila Cartafina and Maheshwar Muralidharan, representing views from McKinsey Technology.



June 2026

AI can make enterprise data look deceptively simple. A contract becomes a summary; a customer transcript becomes a recommended action; a policy becomes an answer. But none of this happens by magic. Underneath, AI systems pull data apart and put it back together again and again—across documents, systems, prompts, and workflows. Scaling AI means ensuring that all this data is treated as truth, so agents can act responsibly and users can trust outputs. Getting to this truth is anything but easy.

This bottleneck is a key reason only 7 percent of companies have fully scaled AI across their organizations. And solving for unstructured data alone won't be enough: AI data readiness requires companies to connect structured and unstructured data into a governed, traceable, and reusable foundation (see sidebar, "A case study in data readiness").

A case study in data readiness

A financial-services company recently rebuilt its unstructured data pipelines with the same rigor traditionally applied to structured enterprise data. The effort focused on documents, images, audio files, and other inputs that needed to be parsed, extracted, quality checked, enriched, and prepared for AI consumption.

The company recognized that a source artifact was no longer a single static object. A PDF, for example, could produce extracted text, tables, images, image summaries, metadata, sensitivity tags, quality scores, and other intermediate artifacts. These outputs needed to stay connected to the original source and to one another, preserving meaning, lineage, and control as they moved through the pipeline.

This decomposition mattered because the goal was not simply to load documents into a model. It was to make the right content discoverable, retrievable, and usable by AI applications. The company therefore created curated unstructured data products that could be accessed via text, metadata, and vector searches, as well as APIs. This enabled applications to retrieve the right documents, passages, tables, images, or entities before sending context to a model.

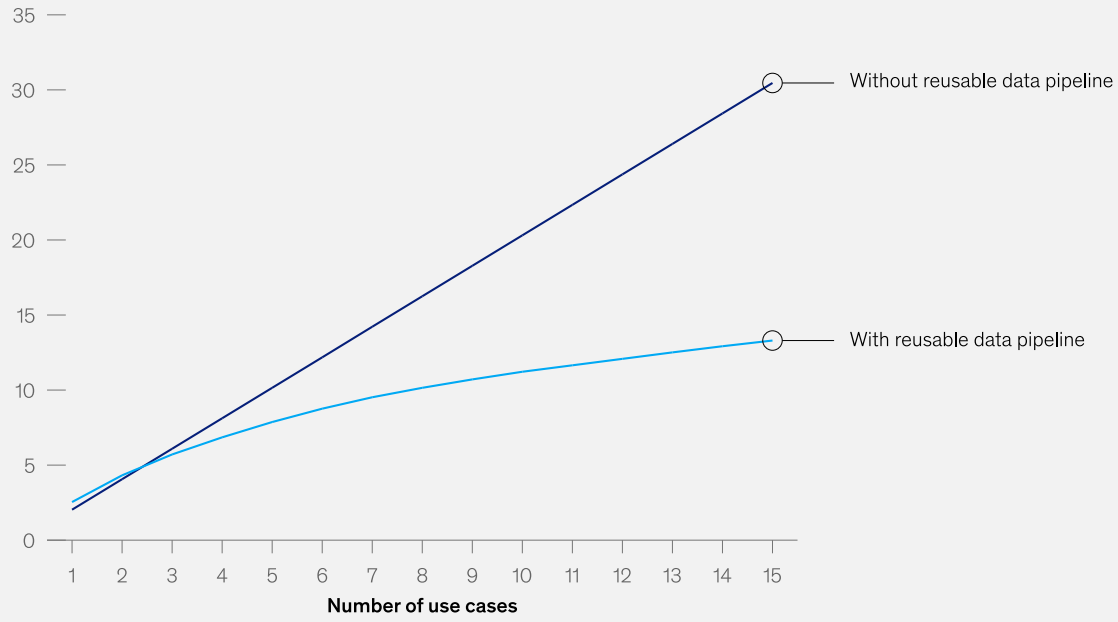
To support the transition, the company developed a common, extensible pipeline pattern. Teams could reuse the foundational mechanics for ingestion, extraction, quality checks, metadata, lineage, indexing, and exposure, while configuring only the steps needed for each curated data set. A video-focused use case, for example, needed different processing steps than an image and table-heavy document use case. Teams could also add rules at defined extension points rather than assembling a new pipeline for every use case.

The result was a more repeatable way to deliver the last mile of data access for AI. Business and application teams could consume governed content through standard interfaces, rather than receiving sanitized data and having to determine how to retrieve, rank, and use it themselves (exhibit).

Exhibit

A financial-services company built reusable data foundations that could deliver $10 million to $20 million in cost avoidance as use cases multiply.

Cost of deploying data-driven use cases, illustrative,[1] $ million



[1]Costs reflect best-estimate assumptions; break-even point and total cost avoidance are indicative and may change depending on inputs. Accounts for build costs across all use cases; run costs are excluded given their dependence on team structure and duration. Up-front cost includes building 2 new use cases and retrofitting 2 existing ones.

McKinsey & Company

AI systems, including gen AI and agentic AI, rely on large volumes of unstructured content files such as documents, emails, call transcripts, and video to create outputs. Each file can expand into multiple representations—including text, tables, and images—increasing both the volume of data and the complexity of managing it. AI systems also reuse outputs across applications, causing small data issues to spread and quickly become large-scale problems. And as data is transformed and recombined across multiple steps of the data supply chain, it becomes harder

to trace, validate, and defend outputs.

Many companies try to solve the unstructured data problem by digitizing their content and making it searchable. But for AI systems, searchability does not equal usability. To function reliably, AI systems also need data with clear versioning, structure, and context, including links between unstructured content and the structured enterprise data that defines customers, products, contracts, assets, policies, and transactions.

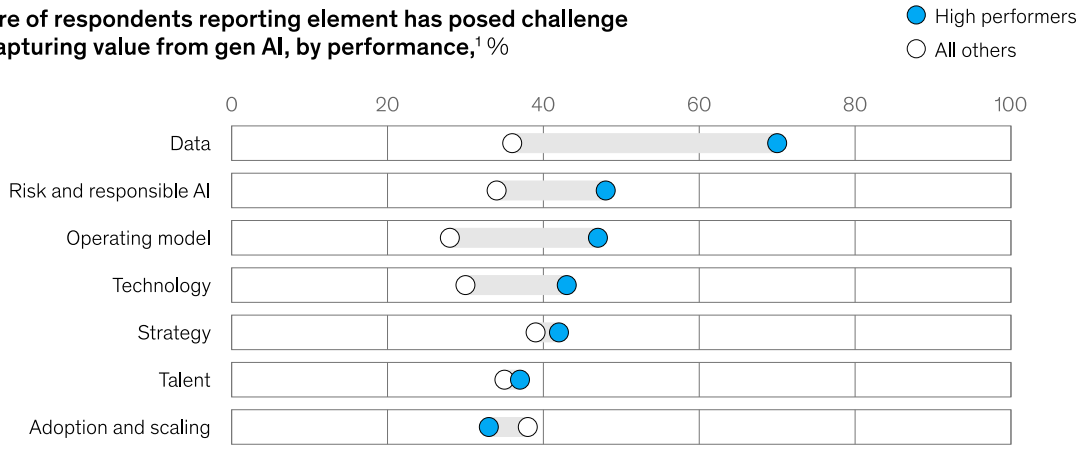
Other companies try to solve the problem with tools. They invest in vector databases, model gateways, and retrieval pipelines. Yet they still struggle to explain AI outputs, trace answers to source documents, or prevent sensitive information from being accessed. It soon becomes clear that tools alone, including those based on retrieval-augmented generation, are not a silver bullet.

As a result of these challenges, more than two-thirds of high-performing companies say data is the primary obstacle for enabling AI (Exhibit 1).

Exhibit 1

High-performing companies cite data as the main challenge to scaling gen AI.

Share of respondents reporting element has posed challenge in capturing value from gen AI, by performance,[1] %



[1]Respondents could select more than 1 response. Top performers defined as respondents who said that at least 11% of their organizations' EBIT in 2023 was attributable to their use of gen AI. For respondents at AI high performers, n = 46; for all other respondents, n = 830.
Source: McKinsey Global Survey on AI, Feb 22–Mar 5, 2024 (n = 1,363)

McKinsey & Company

So, how can AI leaders transform unstructured data from a roadblock into a core asset? It starts by ensuring that data is reliable, clearly understood, traceable, and reusable—so that outputs can be produced consistently and trusted across applications. Unstructured data must become part of a governed and reusable data foundation that also includes structured enterprise data, metadata, lineage, and the tools and skills AI systems use to act on data consistently. Without this foundation, each step toward scaling could increase risks and erode trust in outputs, limiting AI's impact.

There is an emerging myth that data quality matters less in AI environments. In fact, AI often amplifies both the risk of insufficient data quality and the complexity of addressing the problem. The objective is not to achieve perfect data before starting, but to define what "good enough" means for each use case based on business needs and the risk profile of the data and related business processes (Exhibit 2).

In this article, we outline the challenges caused by data and examine how technology leaders can overcome them to scale AI effectively.

Four enterprise technology shifts stalling AI scaling

Unstructured data is a major driver of four major shifts that require readiness across both structured and unstructured data, and across the links between them. These shifts are making data management more complex and stalling AI scaling.

Unstructured data is becoming a harder enterprise challenge

As AI systems rely more heavily on unstructured data, gaps in standards, tooling, and governance are becoming visible in many organizations. Customer conversations, emails, contracts, and internal documents now drive many of the decisions companies make. AI systems actively use this unstructured data to generate answers in real time, and this content doesn't stay whole.

Unstructured data gets transformed during the entire life cycle of extraction, chunking, and embedding. Each transformation alters the context of each data point, especially as it gets reused across systems. That means the results an AI system produces depend on which pieces of the puzzle it pulls and how they are put together.

That's where things can start to break. In the past, data quality was checked once, when data entered the system. Now, something can be correct as a full document and still lead to an incorrect answer if the wrong section is used or key context is missing. Unlike traditional software, AI doesn't follow a fixed path through clean data sets but rather pulls together bits of information on the fly. Small changes in what it retrieves can lead to very different results.

When AI accesses unstructured data, it becomes harder to explain where answers come from. A single response may combine pieces from many documents or fragments of documents, so

Exhibit 2

Companies can define what ‘good enough’ data means for each AI use case.

Data quality and governance requirements by AI use case type, illustrative

Use case example	Quality threshold	Governance rigor	Human oversight	Typical examples
Internal productivity copilots	Moderate	Basic controls	Human review	Drafting emails, summaries, content generation
Knowledge search/ summarization	Moderate-high	Citation, lineage, access controls	Human validation	Enterprise search, policy Q&A, synthesis
Customer-facing assistants	High	Runtime monitoring, guardrails, personally identifiable information controls	Escalation controls	Chatbots, service agents, onboarding
Regulated decisions (eg, financial, legal)	Very high	Full lineage, compliance, auditability	Human approval	Loan decisions, claims, field operations, customer billing, contracts, regulatory reporting

Operational definition of ‘good enough’

1. Enables reliable business outcomes for the intended use case
2. Data quality checks produce errors that are measurable, monitored, and bounded
3. Maintains appropriate controls for sensitive and regulated information
4. Provides sufficient traceability to explain how outputs were generated
5. Applies human oversight proportionate to business and regulatory risk
6. Data readiness improves continuously as models, workflows, and usage evolve

McKinsey & Company

traceability is not straightforward. This shift toward unstructured data entering and exiting AI systems with little traceability is a major challenge for companies. When companies cannot track how content is broken down, used, and recombined, outputs become indefensible. Teams cannot govern a consistent version of truth across transformations. They cannot reliably reproduce which source, version, or transformation logic produced a given answer. In regulatory audits or legal discovery, that gap becomes visible and material.

The risk surface area is expanding as AI retrieves, recombines, and generates data

In the past, technology leaders mitigated risk by monitoring who could access, control, and consume data. With AI, the bigger challenge is managing risk after AI accesses data. AI systems draw from many sources at once, including databases, documents, internal tools, and external

systems. They don't rely on fixed data sets but instead select and combine pieces of data, both structured and unstructured, to generate answers. Much of this happens in a black-box way, with limited visibility into the AI's deterministic "thought process."

Because AI assembles context in real time, its outputs depend on what information is selected and how it is combined. If the system does not fully understand the data it is using, even correct data can lead to flawed results.

This creates new risks. Rules applied at the document level may not hold if only parts of that content are used. Sensitive information buried in emails, transcripts, or images can surface in prompts or outputs if proper guardrails are not in place. Even when two AI systems use the same underlying data, they can produce different answers based on how they are set up. The risk expands to inconsistent reasoning, unintended exposure, and indefensible outputs.

At the same time, responsibility shifts to the application layer. Outputs must be continuously validated, monitored, and evaluated over time, with human oversight where appropriate, to ensure correctness and prevent regression as models and prompts evolve.

AI is generating and reusing data faster than governance models can keep up

AI systems don't just consume data; they generate it constantly. Prompts, responses, summaries, and decisions add to a company's already massive data pile. Much of this output ends up back in core systems; for example, as summaries in customer relationship management or decisions in enterprise resource planning. That generated content that could then feed future actions, creating feedback loops that build over time.

Each time AI interacts with data, it also sets off multiple downstream steps, such as retrieval, model runs, and searches. Data processing work is no longer handled in batches. Instead, it happens continuously, often embedded in daily workflows through copilots and agents. AI systems increase both data volume and speed, which changes how much oversight is possible. Errors can spread faster than they are caught. Controls designed for slower, batch-based systems can't keep up with always-on AI workflows.

Manual monitoring cannot keep pace with this velocity. Governance mechanisms designed for periodic review fail in environments where interactions occur continuously, and each data element expands into multiple representations across systems. Small inaccuracies scale into systemic distortion as both volume and complexity increase.

Fragmentation is making chief data officers central to AI enablement

As AI adoption spreads, applications are built in parallel across business units. Data is no longer tied to a single use case. It's shared across workflows, systems, and decisions, and depends on shared layers that pull, process, and control data from across the enterprise.

Without clear ownership, these capabilities fragment. The same source content is processed differently, tagged differently, and accessed through different retrieval methods. As a result, the

same input produces different outputs across applications. Governance becomes inconsistent, costs multiply, and trust in AI declines.

These challenges are expanding the mandate of the chief data officer (CDO). The role now emphasizes ensuring that data can be reused, traced, and governed consistently wherever AI systems operate. Increasingly, that mandate also extends to linking structured and unstructured data, and to the reusable tools and skills that allow AI systems to act on that data consistently, not just access it. CDOs will also need to collaborate more closely with product engineering teams, as data is crucial in the AI-driven development process. As expectations of the CDO role broaden and increase, talent requirements shift accordingly. Data, engineering, product, and governance roles are starting to blend. Organizations must build multifaceted skill sets rather than treating them as separate domains.

What must change structurally

Traditional data architectures were designed for stable data sets and predictable workflows. AI systems operate differently. They dynamically retrieve and generate information across many systems in real time. As a result, organizations must evolve core data disciplines—not replace them—to support AI at scale. That means CDOs must rewire these capabilities to operate in new ways. Achieving AI data readiness requires CDOs to apply six disciplines across structured data, unstructured content, derived artifacts, and the governed tools and skills that allow large language models (LLMs) and agents to use data consistently, especially where more deterministic behavior is required.

Observability

Observability makes data processing visible end to end, enabling teams to detect issues early and intervene before flawed data affects AI outputs. It monitors data ingestion and transformation to prevent incomplete content, failed processing, or corrupted data from reaching downstream systems.

With AI, that visibility must extend further. It must make the assembly of context and the production of outputs observable, not just the movement of data. This includes detecting stale or incomplete content influencing answers, monitoring retrieval behavior over time, identifying retrieval or orchestration failures that distort outputs, and tracking whether generated responses remain aligned with current source material.

Pipelines must still run reliably. In addition, retrieval logic, answer quality, citation integrity, and content freshness must be continuously evaluated as source documents and use cases evolve. This has to extend beyond the pipeline to how content is seeded into search indexes, vector stores, and APIs, and how it is delivered at runtime. Observability must also track newly generated artifacts as they are created and reused across systems. It builds on traditional monitoring to cover the full life cycle through which AI outputs are assembled, generated, and reused.

Data quality management

Data quality ensures that only complete, correct, and current data flows from the source to downstream systems. Traditional controls often focused on validating fields, enforcing schema rules, and preventing stale or corrupted data from entering reports or models.

With AI, those controls must extend across transformations. Extracted objects, chunks, and embeddings become the artifacts that influence outputs. Quality must therefore be maintained not only at ingestion but also across extraction, chunking, retrieval, and generation. A document may be accurate in full yet produce incorrect answers if outdated or incomplete fragments are retrieved.

Quality now includes semantic integrity. Superseded content must be prevented from influencing responses. Updates to source documents have to propagate predictably through embeddings and indexes, and validation must operate at the artifact level to avoid silent inconsistencies that surface only in outputs.

Metadata management

Metadata has traditionally made data discovery easier to govern. It identified ownership, sensitivity, and usage rules, enabling structured data to be found, interpreted, and used safely.

With AI, metadata must become the control layer for unstructured artifacts. Autonomous and agentic systems rely on contextual signals to determine what content can be used and how it should be interpreted. Ownership, sensitivity, intent, and allowed usage must be explicit for extracted objects, not only for source documents.

Organizations need to move beyond managing files and folders to managing reusable data objects with fine-grained schemas. For example, in an audio transcript, these objects might include speakers, clauses, time stamps, and attributes. Enterprise graphs must anchor unstructured artifacts to structured enterprise data and core entities such as customers, contracts, products, assets, policies, and transactions. Without those links, the same document, clause, transcript, or image can be interpreted differently across applications.

Data lineage

Data lineage documents how structured data moves from source systems into curated tables and then into reports or models. A transaction record is transformed, stored in a table, and used by a finance or risk model. Lineage ensures that teams can trace the output of that model back to the original record and transformation logic.

In an AI environment, the chain becomes more complex. A single PDF may contain text, tables, and images. Text is extracted, tables are parsed, and images may be converted into additional text. Content is segmented into chunks, embeddings are generated, and retrievals are fragments. Prompts assemble context.

Lineage, therefore, has to capture not only the original source document but also each derived

artifact that influences the final output. It must track which version of a document was indexed, how it was segmented, which chunks were retrieved, how prompts were constructed, and which tools were invoked. Lineage must also capture how unstructured artifacts are linked to structured records, master data, and business entities, because those relationships often determine how an AI system interprets and uses the content.

Without this artifact-level traceability, the organization cannot explain how an answer was produced, assess the impact of updating a document, or confidently manage change. Lineage shifts from tracking table transformations to tracing dynamic assembly across multiple generated layers.

Governance and controls

Governance has traditionally enforced access rights and policy at the storage layer. A document or table was classified as confidential. Access was restricted by role, and sensitive fields were masked in reports. Data elements were flagged for quality issues. Compliance was focused on who could see the data.

With AI, critical control decisions are no longer made at the storage layer. They also arise at the point where content is retrieved, assembled, and generated into outputs. Consider a sensitive contract stored in a document repository with restricted access. In a traditional system, restricting access to that document would be sufficient. In an AI system, portions of that contract may have been extracted, chunked, embedded, and indexed. If retrieval logic does not enforce policy at the embedding and prompt layer, fragments of sensitive clauses can surface in model outputs even when document-level access controls are in place.

Governance must therefore extend beyond storage to runtime. Controls must apply to embeddings, prompts, memory layers, and generated outputs. Sensitive information must be filtered during retrieval and generation, not only when the document is stored. Policies must govern how information is assembled, not just who can open a file.

This shifts governance from static access management to dynamic control over how information is interpreted and surfaced. Without that extension, compliant storage does not guarantee compliant outputs.

Platform and tooling architectures

Platform and tooling architectures have traditionally helped standardize data ingestion, storage, and analytics pipelines to support reporting and structured use cases.

With AI, the platform must standardize how unstructured content becomes AI-ready. A single document may require text extraction, table parsing, image conversion, data segmentation, and indexing for retrieval. If each team builds that pipeline independently, duplication multiplies.

Consider two business units building customer support copilots. Both ingest call transcripts and knowledge base articles. Without a shared platform, each team defines its own extraction logic,

chunking strategy, embedding model, and retrieval configuration. This can lead to differing outputs and duplicated costs.

Platform maturity in an AI context means creating reusable extraction pipelines, shared embedding and indexing infrastructure, common retrieval layers, and standardized guardrails that can be used safely across applications. Tooling does not replace governance or quality disciplines. It enables them to scale consistently. For workflows that require deterministic behavior, governed tools and reusable skills become especially important because they give LLMs and agents controlled ways to retrieve, validate, calculate, apply policy, and update systems consistently across applications.

The call to action: What data leaders must do now

Scaling AI requires CDOs to shift their mindset from owning data pipelines, models, and warehouses to owning the standards, control plane, reusable data products, and governed tools and skills that make AI outcomes reliable and repeatable—regardless of where applications are built. Here are six concrete steps CDOs can take to lead that shift.

— *First, include structured and unstructured data in data products.* Unstructured artifacts must be treated as governed data products, not temporary pipeline outputs. This includes defining canonical schemas, entity alignment, quality thresholds across transformations, artifact-level lineage requirements, and explicit rules for handling sensitive content. For example, every contract PDF is converted into the approved schema, with named entities, sensitivity tags, and lineage captured before any model can use it. These standards must apply consistently across all AI applications.

— *Second, establish shared foundation services.* Core capabilities such as governed retrieval layers, reusable tool and skill services, runtime policy enforcement across prompts and embeddings, artifact-level lineage tracking, and observability must be built once and reused across applications. For example, a new copilot uses the enterprise retrieval, policy, and monitoring stack by default rather than building its own prompt filters and audit logs. These services form the control plane for enterprise AI. Without them, every new use case recreates governance and risk logic independently.

— *Third, enable federated delivery on top of common infrastructure.* Business units should be able to build AI applications within their functions, but only on top of standardized extraction pipelines, reusable embeddings, shared metadata models, and common guardrails. For example, the HR and legal functions each launch their own assistants, but both run on the same extraction pipeline, metadata model, embeddings, and guardrails. Innovation scales when the foundation is consistent.

— *Fourth, manage derived artifacts as enterprise assets.* Extracted objects, embeddings,

indexes, and other generated artifacts must be versioned, auditable, and explicitly retired with clear ownership and service levels. For example, each embedding index has an owner, version, refresh cycle, audit trail, and retirement date, just like any other production data asset. If these artifacts are treated as transient outputs, governance breaks at scale.

— *Fifth, govern semantic consistency across access paths and modalities.* Ensure that meaning is preserved whether data is accessed through SQL and warehouse analytics, keyword and metadata search, or vector and semantic retrieval by linking structured records, unstructured artifacts, shared entities, and policy rules across all access methods. For example, an "active customer" returns the same underlying set of data whether queried in SQL, filtered in search, or retrieved through a vector-based assistant.

— *Sixth, measure readiness and reduce risk.* A company can assess its data readiness by measuring four key metrics: reuse, reliability, governance, and scalability. Reuse shows whether the company is succeeding at building capabilities once and using them many times—rather than rebuilding pipelines for each use case. Reliability indicates whether outputs remain accurate and traceable as content evolves over time. Governance confirms whether controls operate where data-use decisions are made—including across retrieval and generation layers, not only at storage. Scalability reveals whether data expansion reduces marginal cost and time to deploy applications, while also reducing complexity and duplication. Without measuring these metrics, companies risk continued data fragmentation that could accumulate unnoticed as AI scales.

CDOs who want to scale AI throughout their organizations will need to treat data as a core enterprise asset. That means putting in place the standards and controls that make data reliable, traceable, and usable across AI applications. Those who succeed with this balancing act can scale AI with consistency, safety, and speed.

Asin Tavakoli is a partner in McKinsey's Düsseldorf office; **Brian Goodman** is a partner in the New York office, where **Kayvaun Rowshankish** is a senior partner and **Camila Cartafina** is a consultant; **Stephen Reddin** is a partner in the Toronto office; and **Maheshwar Muralidharan** is a consultant in the Boston office.

This article was edited by Kristi Essick, an executive editor in the Bay Area office.