Data Management for GenAI: Building a scalable foundation

As highlighted in our 2025 Data Management Trends report, the rapid advancements of LLM capabilities are propelling GenAI models and agentic systems from experimental pilots into business-critical applications. Yet for these sophisticated models to produce real value, they must have consistent, high-quality access to your data—which often remains locked in silos or buried under legacy systems and updated at irregular intervals.

This situation introduces both immediate and long-term challenges for data management. If you don't deliberately embed robust data practices into your GenAI efforts, you may find yourself stuck after just one or two pilots, unable to scale or accommodate new GenAI-driven value cases —because the underlying data bottleneck remains.

The good news is that these smaller pilots present the perfect opportunity to face your data challenges head-on and establish sustainable best practices from the start. How, then, can you use data management principles to build GenAI solutions that truly scale?

Below, we'll first look at the data-focused challenges you'll likely confront in your first GenAI project as a result of its requirements, along with the key questions each one raises. We'll then explore how the right data management practices can help you solve these issues—ensuring your AI efforts remain both sustainable and future-proof.

Start smart: Data representability is key

Selecting your first GenAI value case is itself a critical skill—one that we've honed extensively through numerous successful projects. This process involves thoroughly validating the anticipated impact, carefully assessing feasibility, and ensuring sufficient data availability. Beyond immediate practicality, a critical criterion is representability with regards to data. In other words, how effectively does your initial use case surface data challenges and complexities that are representative of subsequent GenAI initiatives within your organization?

Your first pilot serves as a strategic test for your data landscape, revealing critical insights into underlying data bottlenecks and opportunities for improvement. Ideally, choose a value case that meaningfully mirrors the data-related issues you will face in broader future efforts. Avoid selecting an overly comprehensive scenario—doing so risks overwhelming your initial efforts. Instead, strategically select a targeted yet representative scenario, striking a balance between tangible early wins and the capability to sustainably scale your GenAI initiatives.

Don't get stuck: Answer these critical data questions

The targeted approach described above delivers an early tangible win to secure stakeholder buy-in—and, more critically, it exposes underlying data issues you might otherwise overlook. Here are typical requirements related to data for your GenAI-driven value case, linked with the key challenge it creates and the core questions you will then need to answer:

Requirement	Challenge	Key Questions
Integrated information	Information is often scattered across various systems in multiple formats.	*Where is all this data physically stored?* *Who maintains ownership of each source?* *Do you have the necessary permissions to access them?*
Searchable knowledge system	Documents need to be split into chunks, embedded and stored in a vector database	Can you find a reliable embedding model for capturing the nuances in your business? *How frequently must you re-embed to handle updates* *Are the associated costs manageable at scale?*
Context on how data is interconnected	Building a knowledge graph	*Is your data described well enough to construct such a graph?* *Who will ensure these relationships and definitions stay accurate over time?*
Retrieval from relational databases	Gathering extensive metadata so the model can generate valid SQL queries	*Are vital metadata fields consistent and detailed?* *Do you have the schemas and documentation the model needs to navigate these databases effectively?*

Defining prompts, output structures and tool interface	Building re-usable components	*How do you ensure these prompts and interfaces are reusable?* *Can you track changes for auditability?* *How do you log their usage to troubleshoot issues?*
Evaluate model outputs over time	Collecting feedback from users	*How do you store feedback?* *How do you detect dips in accuracy or relevance?* *Can you trace poor results back to specific data gaps?*
Access to multiple data sources	Safeguarding against threats such as prompt injection	*Which guardrails or role-based controls prevent unauthorized data exposure?* *How do you ensure malicious prompts don't compromise the agent?*

Avoid the pitfall: Apply data product thinking from day one

A common pitfall is solving these data hurdles only within the narrow scope of the first pilot—creating quick fixes that work in isolation but fail to generalize. That short-term success can rapidly degrade as you attempt more GenAI projects, each with slightly different data structures, metadata definitions, or security constraints. Instead, you want a strong, reusable strategy for the data from day one—so that each new GenAI solution builds on a consistent, stable foundation. Following data product thinking and aligning with your organization's data management strategy is the way to go.

Rather than chasing a quick fix for each new GenAI use case, treat your data assets—vector databases, knowledge graphs, relational metadata, prompt libraries—as products with clear lifecycle management. They turn into one or more data product, which means they should be accompanied with assigned owner, explicit documentation, stated quality criteria, and well-defined consumers.

Core data product ingredients for GenAI

Below, we outline the key principles for making that happen, along with examples, roles, and how to handle incremental growth.

- **Discoverability:** Implement a comprehensive data catalog that clearly documents what data exists across your organization, how it's structured, who owns it, and when it's updated. This catalog should work in conjunction with—but not necessarily be derived from—any knowledge graphs built for specific GenAI use cases. While knowledge graphs excel at modeling domain-specific relationships, your data catalog needs broader enterprise-wide coverage. Include the metadata and examples needed for generating SQL queries, as well as references for your vector database. Define a scheduled process to keep these inventories current, ensuring that new GenAI projects can quickly discover existing datasets or embeddings—avoiding needless re-parsing or re-embedding of data they already have.
- **Quality Standards:** Define explicit criteria for what "good data" entails in your GenAI environment—this might include chunk-level completeness for embeddings, consistent naming conventions, or standardized templates for prompt outputs. Involve domain experts to clarify which fields or checks are critical, and align those standards in your data catalog. By enforcing requirements at the source, you prevent silent inconsistencies (like missing metadata or mismatched chunk sizes) from undermining your GenAI solutions later on.
- **Accessibility:** Ensure your data and systems are accessible to GenAI models through standardized interfaces. Models need consistent, reliable ways to retrieve information across your enterprise landscape without requiring custom integration for each data source. The Model Context Protocol (MCP) offers a solution by providing a standard specification for models to access external data and tools. With major providers like OpenAI now adopting this protocol, implementing MCP-compliant data access layers allows your models to interact seamlessly with your organization's systems while maintaining flexibility to switch between different model providers as needed.
- **Auditability:** Log every step—from embedding generation and knowledge graph updates to prompts, reasoning, and final outputs—so you can pinpoint which data, transformations, or model versions led to a given response. This is especially vital in GenAI, where subtle drifts (like outdated embeddings or older prompt templates) often go unnoticed until user trust is already compromised. From a compliance perspective, this comprehensive logging becomes critical when GenAI outputs influence business decisions. Organizations may need to demonstrate exactly which responses were generated, what processing logic was applied, and which data was used at specific points in time—particularly in regulated industries or during audits. Ensure each logged event links back to a clear version ID for both models and data, and store these audit trails securely with appropriate retention policies to meet regulatory, privacy, and governance requirements.

- **Security & Access Control:** Ensure appropriate role- and attribute-based access is enforced on all data sources your models can access, rather than granting blanket privileges. Align access consistently with the user's identity and contextual attributes, preventing the model from inadvertently returning unauthorized information. Be particularly vigilant about prompt injection attacks—where carefully crafted inputs manipulate models into bypassing security controls or revealing sensitive information. This remains one of GenAI's most challenging security vulnerabilities, as traditional safeguards can be circumvented through indirect methods. Consider implementing real-time checks, especially in autonomous agent scenarios, alongside an "LLM-as-a-judge" layer that intercepts and sanitizes prompts exceeding a user's permissions or exhibiting malicious patterns. While no solution offers perfect protection against these threats, understanding the risk vectors and implementing defense-in-depth strategies significantly reduces your exposure. Resources like academic papers and technical analyses on prompt injection techniques can help your team stay informed about this evolving threat landscape.
- **LLMOps:** While this topic deserves a full blog post on its own, we summarize the key points here. Generic organization-wide LLMOps pipeline templates and toolbox ensures every GenAI project can inherit a stable, proven setup—from LLM vendor switching and re-embedding tasks to a shared library of connectors and tool interfaces. Consolidating these resources avoids ad-hoc solutions, fosters collaboration, and enforces lifecycle management for retiring outdated components. As a result, different teams launch new GenAI initiatives faster, standardize best practices, and retain the flexibility to experiment with multiple models or toolchains.

The takeaway: Combine your GenAI implementation with proper data management

Implementing GenAI models and agentic systems isn’t just about prompting an LLM. It’s a data challenge and an architectural exercise. By starting with a high-impact use case, you secure immediate value and insights, while robust data management practices ensure you stay in control—maintaining quality and trust as machine intelligence spreads across your organization.

By adopting these principles in your very first GenAI initiative, you avoid the trap of quick-win pilots that ultimately buckle under growing data complexity. Instead, each new use case taps into a cohesive, well-governed ecosystem—keeping your AI agents current, consistent, and secure as you scale.

Those who excel at both GenAI innovation and data management will be the ones to transform scattered, siloed information into a unified, GenAI-ready asset—yielding genuine enterprise value on a foundation built to last.

If you haven’t already established data management standards and principles, don’t wait any longer—take a look at our blog posts on data management fundamentals part 1 and 2.

If you’d like to learn more about building a robust GenAI foundation, feel free to reach out to us.