WHITE PAPER

Fueling GenAI Innovation with the Right Data

Making LLMs Deliver Real Business Value Through Solid Data Foundations

October 2025
By Michael Farrugia and Max Struever

GenAI has sparked a wave of bottom-up innovation: employees aren't waiting to be given tools or doing things manually—they're building their own AI-powered productivity apps. This excitement for scaling an employee's abilities with hyper-personalization holds huge potential. But for a business to enable workers to create fully-customized, high-performing individual apps, the effort must be built on a robust, AI-native foundational data platform. Only with a solid data foundation can these interfaces go beyond demos and deliver consistent, enterprise-grade results.

Fortunately, there's plenty of optimism and demand from companies to empower this platform-based experimentation and creation. Coding platforms like v0 and Replit are already in use in the real world, making creation fast and fun for anyone building tools. An explosion of similar platforms, plus the **rise of GenEng** and autonomous agents, allow users and teams to scale their ability to build tremendously. But get the data fundamentals wrong, and no amount of newly spun-up apps or clever design will save your GenAI initiatives.

A strong data platform also lets you go further with more model capability, allowing LLMs to derive second-order analytical insights, support Q&A about the organization, and summarize and co-relate information. Ultimately, your data architecture determines the speed and direction of your innovation.

Leaders consistently point to a lack of trustworthy data as a challenge to this kind of innovative, forward-thinking AI implementation—a clear reason why 75% of AI transformations and 95% of GenAI initiatives stall[1]. Creating and consistently maintaining a strong data foundation is like having to eat your vegetables and going to the gym: while not flashy work itself, it not only helps you draw more value from today's AI projects but also prepares the business for the future.

This level of innovation isn't a problem just for technical leaders. The business side also needs to think about a strong foundation for LLMs—especially as agents become critical assets in an organization's AI strategy. It boils down to an essential question: "Has your organization's approach to data and foundational requirements readied you to harness the capabilities of LLMs, or are foundational weaknesses holding you back?"

1. https://www.bcg.com/publications/2024/most-large-scale-tech-programs-fail-how-to-succeed, https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/

Key Components of a Foundational Data Platform

The technology world is investing heavily in data foundations. Examples include Salesforce's $8 billion acquisition of data management platform Informatica[2] and ServiceNow's strategic purchase of Data.World. These moves are more than technology investments. When industry titans put this level of capital into data services and governance companies, it's a message to the market: data foundations aren't optional—they're the engine of the future and AI advantage.

It also signals a reality for all businesses. Modern GenAI depends not just on powerful foundational models, but on proprietary, current enterprise data connected by smart infrastructure. Delivering this requires LLM-ready platforms—beyond traditional analytics—to rethink how data is organized, governed, and accessed. The next section outlines four core components of an AI data foundation.

Discoverable Data: The first step is to find the right data at the right quality and granularity. Making data discoverable and available in a scalable medium has long been a problem in supporting effective data analysis and business intelligence. In fact, Forrester reports that data professionals still struggle to find needed information. Often, the root cause is flawed thinking that buying a tool is enough to solve the data discoverability problem.

Providing the right data discovery capability begins with building tools to identify, categorize, and describe data—in other words, to query self-describing data. This effort itself can be augmented with GenAI: active metadata generation can use specialized LLMs that can infer schema, identify relationships, and extract semantic meaning from both structured and unstructured sources.

The benefits are twofold. Both LLMs and human users will find that well-documented data in natural language improves discoverability. Modern discovery platforms leverage embedding models as part of solutions like Retrieval Augmented Generation (RAG) to create unified semantic spaces. This makes database tables, API endpoints (as described through OpenAPI or GraphQL specifications), document repositories, and streaming data sources equally discoverable through natural language queries.

The future of data discovery lies in agentic systems that proactively maintain data readiness. These can automatically document new data sources, identify quality issues, suggest enrichment opportunities, and even negotiate access permissions on behalf of users.

2. https://techcrunch.com/2025/05/27/salesforce-acquires-informatica-for-8-billion/

Access Control, Authorization, and Authentication: The shift from human-centric to agent-centric access control represents one of the most critical and underestimated challenges in enterprise AI adoption. While many organizations have deployed LLMs across their operations, **IBM's 2025 research** reveals that 97% of breached AI systems lacked proper access controls, with shadow AI breaches costing organizations an additional $670,000 per incident.

Traditional identity management frameworks, designed for static machines and predictable human behaviors, fail when confronted with autonomous agents. These agents can dynamically assume both human and non-human identities, execute complex workflows across multiple systems, and make real-time decisions at machine speed. Such capabilities are difficult for traditional frameworks to manage.

Forward-thinking enterprises are responding by architecting zero-trust frameworks specifically designed for AI's unique authentication challenges. **The Model Context Protocol's 2025 authorization specification, which extends OAuth 2.1 with dynamic Protected Resource Metadata**, provides a starting foundation for securing agent-to-data interactions, but because it is just optional, you may end up with a patchwork of unreliable and inconsistent integrations. Implementation of authentication for your data, especially for MCP, requires fundamental rethinking of access governance.

Privacy and Classification: In an AI-native architecture, privacy and data classification evolve from compliance checkboxes into the intelligence infrastructure that determines what data your AI could or should not access.

The technical foundation begins with sophisticated classification engines that understand context beyond simple pattern matching. These systems need to recognize when a customer service note contains sensitive health information, when a contract clause implies confidentiality obligations, or when seemingly innocuous data combinations could reveal protected attributes.

These capabilities rely on hierarchical classification taxonomies; automated PII detection across structured and unstructured data; dynamic data lineage tracking that follows information as it flows through AI pipelines; and versioning systems that maintain classification integrity even as data evolves. Modern implementations leverage LLMs themselves as classification engines, creating a virtuous cycle where AI systems help govern their own data access.

Unstructured Data Management and Knowledge Systems: GenAI's strength lies in processing unstructured data: documents, emails, presentations, audio transcripts, images, and other multimodal content. Retrieval-Augmented Generation (RAG) is a common pattern to dynamically pull relevant information from a proprietary knowledge base at the right moment, though it has some limitations, such as scalability. Think of it as giving AI photographic memory of every document, decision, and insight an organization has ever produced.

From an architecture standpoint, this requires organizations to implement vector databases that can store and search semantic embeddings of their content. However, implementing RAG isn't just about connecting AI to documents. It requires maintaining real-time synchronization between source materials and their semantic representations (the meaning behind the words).

Most organizations stumble at this critical juncture: creating coherent meaning from chaos. This is where formal ontologies—an organization's semantic blueprint—become strategic assets rather than academic exercises. An ontology defines how your organization understands its world: what a "customer" means across different divisions, how "risk" relates to "opportunity," and why "innovation" connects to "market position." Without this semantic scaffolding, AI systems produce technically correct but strategically meaningless outputs. With it, they become extensions of your organization's expertise and judgment.

The implementation of formal ontologies and knowledge graphs represents a strategic inflection point in enterprise AI adoption. After two decades of academic promise, semantic web technologies are experiencing a renaissance through LLM integration. Organizations are deploying LLMs as 'ontology oracles' that automatically extract and validate domain knowledge, creating proprietary semantic scaffolding that encodes the unique characteristics of a business.

This knowledge architecture enables powerful capabilities: multi-hop reasoning that connects disparate information sources, temporal knowledge layers that track relationship evolution, and cross-domain federation that allows AI agents to seamlessly navigate between different knowledge domains.

Alongside the technical questions, there's a clear advantage in such foundations for businesses. These options allow a business to expect better results for their agent strategy. Over 75% of businesses predict agents to become more important over the next several years. These technical choices can guide a company toward better implementation of these make-or-break assets.

Empowering LLMs Through Internal APIs

Deploying LLMs and using them as effective autonomous agents hinges on accessing clear internal APIs as tools and function calls (potentially via MCP). To transfer data and power these resources, APIs must be organized and uniformly accessible, but this is often not anywhere near the reality for an organization's internally developed APIs.

API-wrangling has many best practices. Setting up an API management tool, either native to your cloud provider of choice or a third party, is often the first step. Documentation is just as critical here, and tools like OpenAPI docs or strongly typed and auto-documenting GraphQL can help.

Consider tooling that can combine and organize disparate APIs into a single retrievable portal, such as a Supergraph[3]. Eliminate the need to custom-build an API with tools like Hasura[4] or Supabase[5], which can make an API available against a data source instantly. MCP servers often already exist for a given data source type, like Google's MCP Toolbox for databases[6], which can potentially avoid the need for building an API at all, though, again, authentication and access limiting must be strongly considered at every step.

3. https://supergraph.io/
4. https://hasura.io/ddn
5. https://supabase.com/docs/guides/api, https://docs.postgrest.org/en/v13/
6. https://googleapis.github.io/genai-toolbox/getting-started/introduction/

The Steps to Get Foundational Data Right

BCG has found that people make technology adoption a success. **Sixty-four percent of executives expect their human workers to engage with AI in the future**—and several moves around data will be key to that collaboration.

Establish Data Literacy: Without a culture of data awareness, even the most advanced AI implementations fall flat. Investing in training, workshops, and ongoing education ensures that employees across functions can confidently engage with data and GenAI tools

Roll Out a Use-Case-Driven Strategy: BCG has long urged companies that are planning tech transformations to start **small and targeted**. Focus on high-impact GenAI use cases tied to real business problems. Roll out with limited pilots to validate feasibility and outcomes, then scale successful efforts. This approach keeps scope manageable, accelerates time to value of GenAI initiatives, and provides quick wins that build internal momentum

Consider a Minimum Viable Architecture Approach: Avoid overengineering at the beginning. Build a "just-enough" data architecture that supports immediate needs while remaining flexible for future growth. Gradually add complexity, while ensuring visibility and control with modern observability tools.

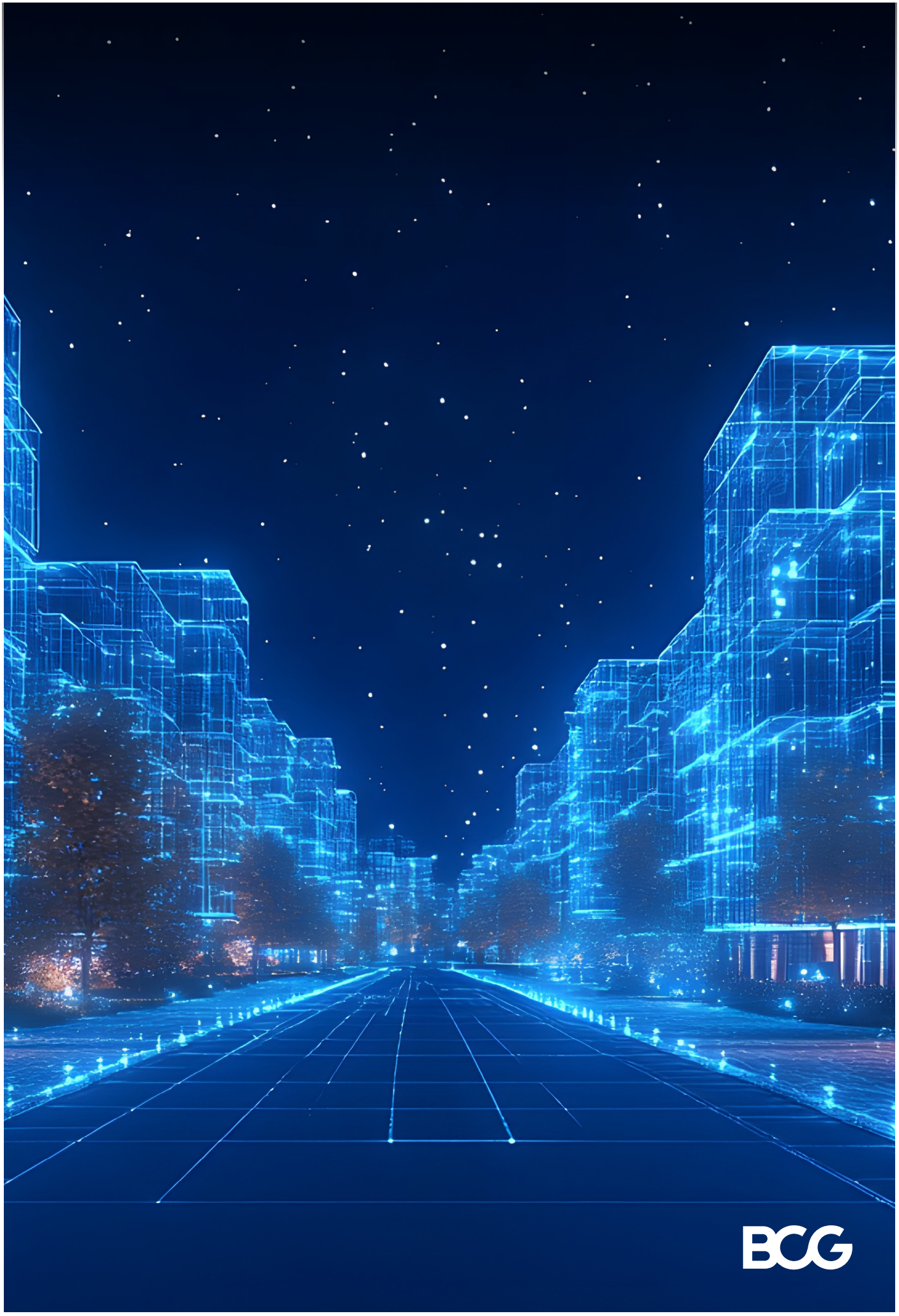
Leverage LLMs to Help With the Actual Build Effort: GenAI isn't just the goal—it's also a tool. Developer copilots like GitHub Copilot, Cursor, and Claude Code can accelerate foundational build-outs through **GenAI-assisted Coding** ("GenEng"). LLMs can also support processes like labeling datasets, writing metadata, categorizing content, and automating documentation.

Remember That Experimentation Is Invaluable: Enable and encourage your teams to test, iterate, and explore. Experimentation fosters learning, reveals edge cases, and ensures your GenAI strategy adapts to real-world needs. Experimentation isn't just a side activity—it's how strategy becomes reality, use case by use case.

Conclusion

GenAI has created a lot of great excitement and offers significant opportunities to scale a user's or business's abilities, but only if businesses commit to solid underlying systems and data strategies. Investing time and resources in foundational data management and clear API integration will ensure that LLM capabilities deliver true, lasting value. Starting with a practical, use-case-driven approach can let you reach certain ROI sooner.

Ultimately, the payoff from getting the basics right can extend well beyond immediate AI projects, supporting innovation long into the future. This can be both a revolution and an evolution - where decades of software engineering principles and the flexibility of LLMs can transform enterprise software today.


BCG