Futurum

IBM

From Data Chaos to AI Clarity: Activating AI Through High-Quality Enterprise Data

JULY 2025

Introduction: AI's Untapped Potential

In today's rapidly evolving and often unpredictable market, generative and agentic AI presents a transformative opportunity. Enterprises hoping to optimize operations and address complex challenges are increasingly investing in AI to handle everything from automating repetitive tasks to enhancing supply chain efficiency. Despite the potential, many organizations struggle to draw the anticipated value from their AI initiatives. Often, the primary impediment for businesses is not the technical complexity of training AI models or the cost of inference, but rather the fundamental difficulty in effectively leveraging their most valuable asset—enterprise data. Because AI systems are inherently data-driven, flawed or inaccessible data inevitably leads to flawed and unpredictable AI outcomes. To make AI more reliable, companies must modernize their data estates, making data easily accessible and contextually relevant at any scale.

The Data Bottleneck: Common Hurdles to AI Success

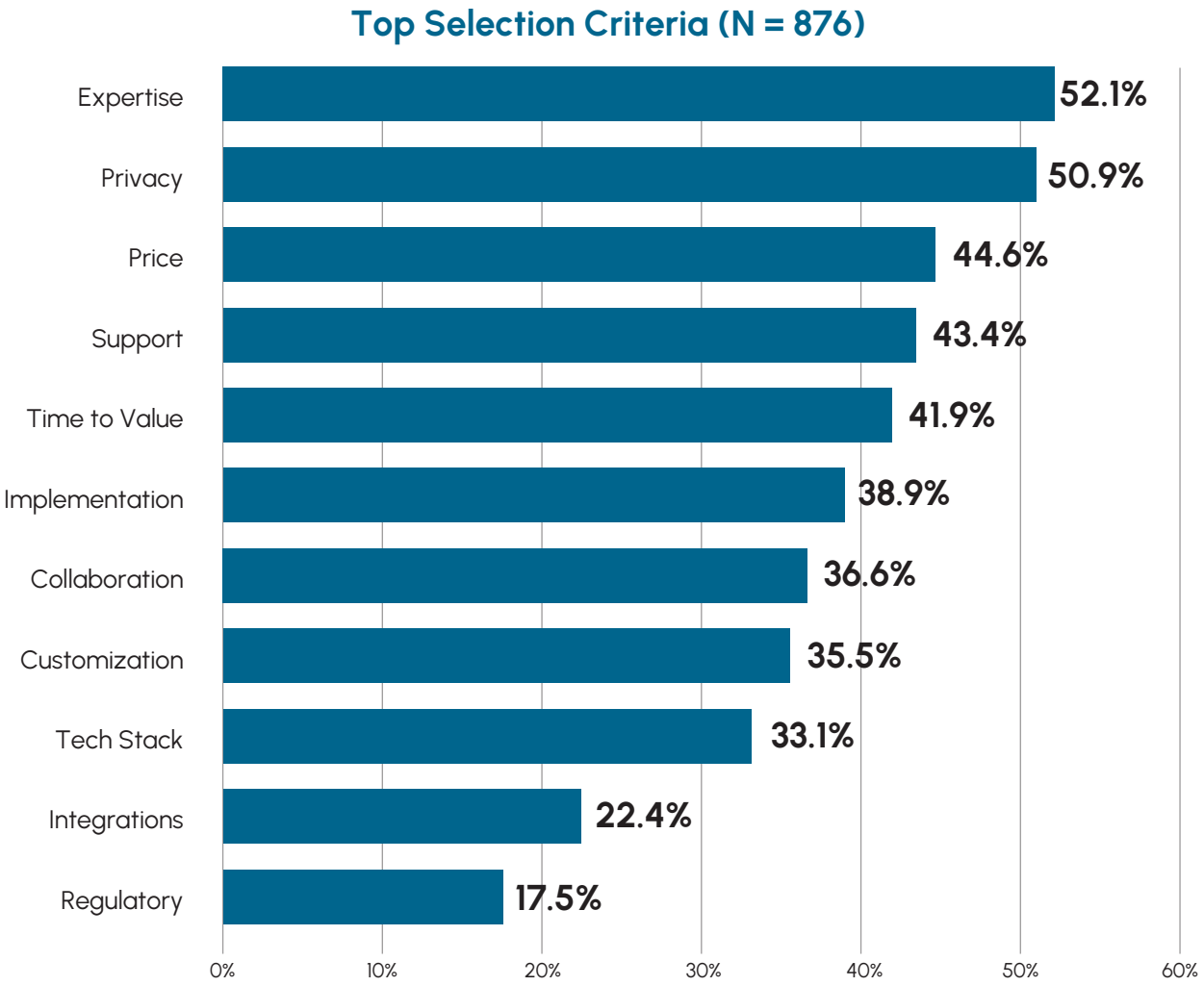
Realizing the full potential of AI requires overcoming several critical, data-centric obstacles. A significant challenge lies in poor data quality and inadequate governance. Incomplete, inconsistent, or biased data directly contributes to flawed AI models, unreliable predictions, poor decision-making, compliance gaps, and exposure to both seen and unseen risks.

Enterprises are swimming in potentially transformative data, but they must grapple with the sheer volume and complexity of that data due to its unstructured nature. This data, often locked within diverse formats such as documents, emails, presentations, and videos, is incredibly difficult for AI to harness effectively. Conventional approaches like model fine-tuning and in-prompt contextual learning through basic retrieval augmented generation (RAG) can help, but they are often ineffective at extracting AI's full value. Worse, these methods frequently sit alongside but do not integrate well with more orthodox structured data access methods. The result? Very little enterprise data finds its way into large language models (LLMs) in a timely, consistent, and impactful manner.

Adding to these complexities is the typical enterprise reality that companies eventually end up using a fragmented collection of disconnected tools for data management. Disparate teams and departments often maintain their own data warehouses, data lakes, and governance and integration solutions. This creates a complex and cumbersome data stack that hinders agility

and efficiency. Over time, these technical debt and knowledge gaps expand, an issue that Futurum has noted in its recent study of nearly 900 enterprise AI practitioners who prioritize expertise over perennial solution metrics, privacy, and cost (see Figure 1).

Figure 1: Selection Criteria for AI Investments



Source: Futurum Research: AI Intelligence Decision Maker: Selection Criteria

Note: When asked to characterize purchasing criteria for AI platforms and tooling, over half of all survey respondents prioritized expertise as their top concern.

Furthermore, many organizations miss the forest for the trees, mistakenly focusing on the generative AI application layer (e.g., the chat interface and process flow), neglecting the crucial underlying data foundation. This prevents initiatives from reaching their full potential by falsely masking unseen, data-related issues such as accuracy and relevance. Until companies address this foundational data layer, generative AI initiatives such as AI agents will struggle to deliver on their promise.

A New Data Paradigm: Quality and Context for Powerful AI

The good news is that unlocking the vast potential within unstructured enterprise data does not necessarily require a complete overhaul of existing data management assets. Instead, organizations can shift their focus from feeding models with more data to ensuring they get the highest-quality, most-relevant, and contextually significant information.

This means adopting an overarching operational (Ops) mentality to build up data fabric architectures and practices, which emphasizes the unification and simplification of data access across diverse, disparate data formats and locations. This kind of approach enables organizations to seamlessly ingest, govern, and retrieve various data types at different velocities—whether that data lives on premises or in the cloud, and whether it comes from internal or external sources.

To illustrate, data fabric architectures and practices bring together several key elements.

Consistent APIs and access methods such as Anthropic's Model Context Protocol (MCP) and Google's Agent 2 Agent framework

The use of SQL as a standard analytical interface for both humans and agentic AI users

A rich semantic layer of metadata capable of unlocking both structured and unstructured data sources

Open data access standards and protocols, such as Apache Iceberg and Databricks Delta Sharing.

Investing in these basic elements can unlock and maximize existing corporate data sources and engineering tool chains, enhancing data quality from the start. And over the long term, these will engender strategic vendor independence, enhance integration flexibility, optimize operational and governance costs, and reduce exposure to security, regulatory, and compliance pitfalls.

Solution Spotlight: IBM watsonx.data – Your Foundation for Trusted AI

IBM watsonx.data was built specifically to meet these critical requirements and is rapidly evolving to address emerging AI opportunities directly.

Fundamentally, IBM watsonx.data functions as an open data lakehouse platform imbued with data fabric capabilities; IBM's approach seeks to unify, govern, and activate data across database silos and data formats. A key differentiator is its hybrid architecture, supporting deployment across on-premises environments and multiple clouds. Standing as a unique capability opposite cloud-only platform options, watsonx.data enables interoperability with existing data investments and avoids vendor lock-in. The platform acts on data stored in open formats such as Iceberg, Parquet, and Avro, residing on object storage from different providers.

As previewed at IBM Think in May 2025 and currently available, watsonx.data has evolved to radically simplify the data-for-AI stack, emphasizing unstructured data. New capabilities enable organizations to ingest, govern, and retrieve unstructured data at scale, laying the foundation for more accurate and scalable generative AI solutions.

IBM watsonx.data facilitates workload optimization by offering fit-for-purpose query engines, including Apache Presto for business intelligence (BI) or Apache Spark for data engineering, that operate directly on the data in open formats. This multi-engine approach provides essential optionality and optimizes performance vs cost ratios for a wide range of data, analytics, and AI workloads. It can also help reduce expensive data warehouse costs, augmenting existing environments such as Snowflake, particularly with write-intensive workloads.

For generative AI, watsonx.data incorporates the open-source vector database, Milvus, built by Zilliz, while also accessing native vector support from IBM Db2. These capabilities are expanding with technologies from the recently finalized acquisition of DataStax already available to watsonx.data customers. With AstraDB, Hyperconverged Database (HCD), and DataStax Enterprise, for example, IBM watsonx.data users can blend generative AI semantic search with NoSQL capabilities, which are geared toward handling large amounts of data at scale with high availability and fault tolerance.

Further, a key contribution from DataStax is Langflow, an open-source framework tuned to RAG pipelines, MCP data and tool access, as well as multi-agent application workflows. With Langflow, enterprise practitioners can orchestrate complex agentic AI solutions, delivering those as an API endpoint or a highly observable project that can itself be delivered as an MCP server instance.

Working with IBM watsonx.data, these new assets will provide robust, production-ready database and application capabilities essential for trustworthy AI development projects.

Deeper Insights: Leveraging AI Within watsonx.data for Enhanced Value

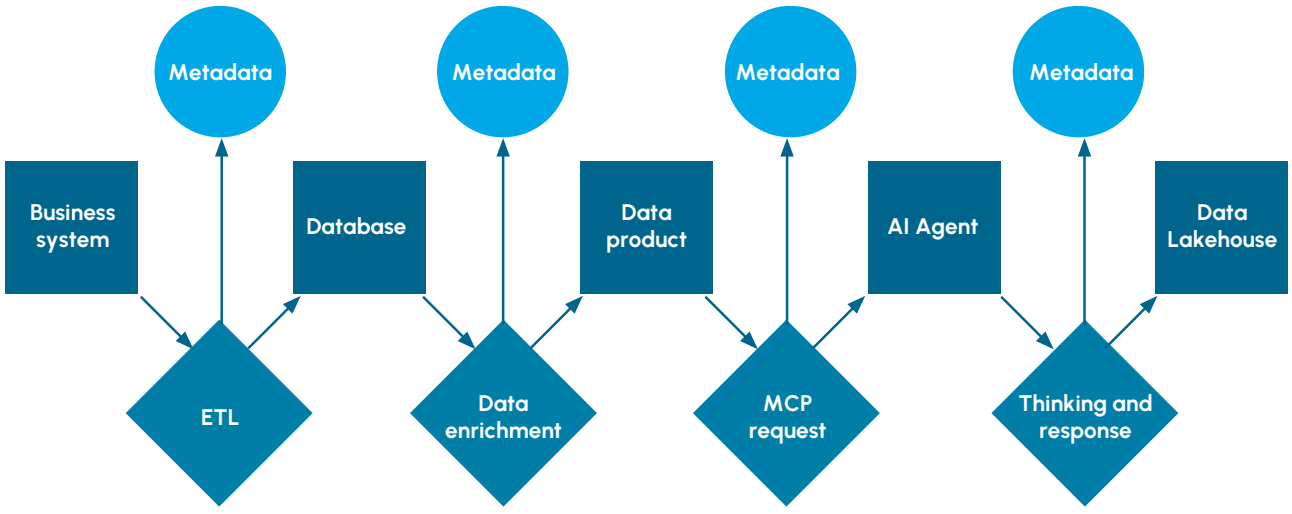
A key differentiation for IBM watsonx.data is around the integrated use of AI within the platform to enhance data management and governance workflows.

This is particularly evident in the new watsonx.data integration component, which introduces the idea of adding structure to unstructured data. Beyond simple vectorization, this new component enables the extraction of queryable insights. For example, it can automate entity extraction (e.g., identifying cost of sale or bill of sale information), mask personally identifiable information (PII), summarize documents, and classify data.

With this blended approach, data practitioners (and increasingly AI agents) can more effectively leverage structured data methods using SQL queries on information extracted from unstructured data to perform more deterministic calculations. Doing so can significantly improve the accuracy of AI model results when combined with traditional RAG techniques, which often struggle to utilize unstructured data for context.

Integrated data governance capabilities, including cataloging, data lineage, and data sharing, also play a vital role in watsonx.data and watsonx.data intelligence (see Figure 2). For example, generative and agentic AI solutions often fail to preserve and propagate proper access control throughout supportive data pipelines, particularly for unstructured data pulled from source repositories such as Box or SharePoint. IBM watsonx.data addresses this issue by ensuring that users querying data through watsonx.data only access authorized information based on the original source permissions. Standard RAG pipelines and vector databases alone lose this chain of custody, cutting vectorized data off from the controls associated with that data (file system, folder, document, etc.).

Figure 2: Data Lineage: Mapping the Flow of Data from Source to Destination



Source: Futurum Research

Note: As data moves from source to solution, critical metadata emerges at each stop along the way. When surfaced in a data catalog, this data can be used to reverse engineer exactly "how" an agentic AI solution arrived at a particular response.

Conclusion

Successful AI implementations are based on a solid, operational data foundation characterized by quality, accessibility, and optionality. The ongoing maturation of open data fabric architectures and operational practices unlocks measurable value from all data types, especially the vast stores of unstructured data within every enterprise that all too often remains inaccessible to AI. An emphasis on open standards and flexible software deployment patterns promises to unleash this inaccessible enterprise data for AI. Moreover, focusing on well-governed, secure, and collaborative access to data at scale empowers enterprises to maximize their AI investments. The result? Companies can more effectively navigate market uncertainties and uncover new strategic opportunities.

Important Information About This Report

AUTHORS

Brad Shimmin
Vice President & Practice Lead, Data Intelligence, Analytics, & Infrastructure | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT IBM

IBM has a long history of helping organizations manage their most mission-critical data and applications. We have extensive experience with innovative enterprise data and AI offerings, including market-leading database and enterprise-ready AI solutions. We help our clients unlock their enterprise data across cloud and on-premises environments and believe that our clients' data belongs to them.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337