SPOTLIGHT

Sponsored by: IBM

There is no AI without data. Organizations are realigning their data strategy in the wake of GenAI's emergence, increasing focus on supporting AI use cases, improving the quality of data and analytics products, and enhancing data security and privacy.

Increasing AI Adoption with AI-Ready Data

October 2024

Written by: Stewart Bond, vice president, Data Intelligence and Integration Software

Introduction

For the past 18 months, hyper-experimentation with generative AI (GenAI) has dominated the conversations of business and technology leaders in organizations of all sizes, industries, and countries.

Leveraging data collected from CIOs and business leaders using IDC's monthly survey series, it has been found that a typical enterprise has identified hundreds of GenAI use cases and launched dozens of proofs of concept (POCs) but has only put a handful (fewer than six) into production. This approach is not sustainable for enterprises that care about converting POCs into a sustainable, long-term business. Factors such as a lack of skills, excessive costs, inadequate infrastructure, and poor coordination between IT and line-of-business (LOB) users are the major limitations to AI success. Conversely, successful organizations benefit from strong relationships with strategic AI partners, effective IT-LOB coordination, appropriate skills, and access to high-quality data sets.

AT A GLANCE

KEY STATS

- 83% of organizations have changed their data management strategy since the emergence of GenAI.
- 26% of organizations say data management is now their top focus.

WHAT'S IMPORTANT

Data is highly distributed, diverse, and dynamic.

KEY TAKEAWAY

Improving data intelligence will have a positive impact on business outcomes.

There is no AI without data, and organizations are therefore realigning their data strategies to embrace data management in the wake of GenAI's emergence. According to IDC's August 2024 *Office of the CDO Survey,* 83% of organizations have increased their focus on data management since the emergence of GenAI. Over one-quarter of these respondents have cited data management as their top area of focus. Their main strategic objectives in data management are supporting AI (generative, predictive, and interpretive) use cases, improving the quality of data and analytics products, and enhancing data security and privacy. These objectives are critical to making data AI ready in the modern data environment.

However, significant work remains even with this strategic realignment. Modern data environments are highly distributed, diverse, and dynamic. Data is distributed across operational and analytical repositories, both hybrid and multicloud, and throughout multiple lines of business. According to IDC's *Office of the CDO Survey*, operational data is sourced from an average of 35 different systems and integrated into 18 different analytical data repositories. One year ago, 48% of analytical data repositories were in the cloud; that number has risen to 53% and is expected to grow to 58%

in the next 12–18 months. Organizations utilize highly diverse data such as operational data, event data, master data, business documents, transactional data, communications data, and unstructured content for AI. In addition, data is highly dynamic, with organizations reporting that it changes faster than they can keep up with.

Taking control of modern data environments requires harvesting intelligence about data; stewardship and governance; data integration, federation, and unification; cleansing and mastering; continuous observability; and accessibility. These functions together can help organizations create AI-ready data to help move past AI experimentation and into a new phase of AI adoption, offering enterprises not just improved AI outcomes but also a source of competitive momentum.

Investing in Data Management for AI-Ready Data

Creating AI-ready data requires strategic investments in both AI technology for data and preparing data for AI. Organizations are investing in technologies with AI capabilities to improve the productivity of data workers, such as data stewards, data engineers, and data administrators, because managing highly distributed, diverse, and dynamic data requires automation. Investments in AI enable the identification and correction of data quality issues, thus enhancing the data quality used in AI applications. The issue of data quality has been long-standing, but its importance is amplified by GenAI, as it can minimize human intervention in scrutinizing data before it is used in analytics or decision-making processes. Humans are now tasked with assessing AI results and sometimes dealing with the consequences of incorrect, incomplete, and inconsistent data.

Investment in data governance is a key component of making data AI ready. Challenges such as overseeing personal or sensitive corporate information, managing the potential negative impact of AI outcomes, and addressing plagiarism and data bias are impeding the broader adoption of AI technologies. Data intelligence leverages metadata to describe data and information; tag, label, and classify data; and put data into context with lineage and business semantics. Similarly, model intelligence provides context about AI models, and combining data intelligence with model intelligence is a key capability in enabling AI governance.

A rising tide lifts all boats. With the shift of data management strategy and investment over the past 12–18 months to now focusing on AI for data and data for AI, organizations have reported significant improvements in five key areas:

- **Data security and privacy:** Strengthened measures protect sensitive information and comply with data protection regulations.
- **Quality of data and analytics products:** AI-powered data cleaning and automated error handling enhance the quality of data and analytics products, resulting in more reliable and trustworthy AI results.
- **Availability of data for generative, predictive, and interpretive AI use cases:** Improved knowledge of and access to data sources, quality scores, business context, and visibility into data flows facilitate advanced AI applications.
- **Regulatory data compliance:** Intelligence about data security and privacy classifications, combined with intelligence about model sensitivities, is enabling adherence to compliance requirements with less effort and greater efficiency.
- **Data worker productivity:** AI assistance in data intelligence and management software is helping with data discovery, documentation, quality improvements, and engineering while proactively identifying issues using continuous observability, elevating operational efficiency to previously unattainable levels.

Organizations that have invested in and become proficient in data management, from ingestion and transformation to consumption, are also seeing improvements in business metrics. According to IDC's *Office of the CDO Survey,* entities scoring highly in areas of data engineering, quality, cataloging, master data management, governance, and controlled access reported on average a 5.9% improvement in operational KPIs and a 4.4% improvement in financial KPIs over the past two years. Indicators related to time to market have surpassed benchmarks, with a 9.3% improvement in time to market, an 8.9% improvement in innovation, and an 8% improvement in customer acquisition and retention. Financial metrics above the average include a profit margin of 8.6% and revenue improvements of 6%.

AI-Ready Data Trends

Organizations should be aware of the following trends shaping the landscape of technology and data management:

- **Point solutions versus platforms:** Some organizations are taking a point solution approach to the procurement and implementation of technology investments enabling AI for data and data for AI, while others prefer one comprehensive data platform with all the capabilities. The point solution approach leverages different technologies from multiple or single software vendors that best align with the organization's requirements in each data intelligence and management function. The downside is that the organization often needs to perform the integration across all the different technologies. A single-platform approach removes the need for integration, but not all the functions may be the best fit for organization requirements. Each of these alternatives exists in the market at both extremes, and some fall somewhere in between, such as single software vendors with integrated multiple point solutions or multiple partners offering prebuilt integrations across portfolios of products, to address the full end-to-end scope of data intelligence and management.

- **Increasing focus on unstructured data:** Unstructured data has increased in importance because GenAI uses it to train, tune, and ground AI models intended for specific business domains and operations. This is increasing the need to understand more not only about where unstructured data exists in the organization but also what is in unstructured data and whether its use with AI is safe. Security, privacy, quality, context, and applicability are all elements of intelligence about structured data. This intelligence about unstructured data also needs to be understood. Organizations need to stop treating content differently than data and should instead unify intelligence about all data. Not only did GenAI drive the need for this, but it is also part of the solution, given how well large language models understand unstructured data.

- **Data lakehouse:** Data lakehouses emerge, in part, as alternatives to the point solution versus best of breed and the structured and unstructured data problem. The term *data lakehouse* refers to the ingestion of all types and formats of data into a data lake. Through data engineering, natural language processing, and intelligence functions, it is turned into structured data in warehouse-like constructs in support of analytics and AI use cases. Data lakehouses may also have data cataloging, quality, mastering, and observability capabilities at various levels of abstraction and maturity. Data lakehouses primarily focus on analytical workloads; therefore, not all of the organization's data will be in a lakehouse — and thus lakehouses alone may not be comprehensive enough to meet all of an organization's needs.

- **Treating data as a product:** Viewing data as a product involves abstracting and collecting data assets to enhance their business value and utility, thereby increasing usage by business-level users. As users continue to treat data more as a product, its increased utility could enable a broader range of users to derive greater business value from it. Respondents to IDC's *Office of the CDO Survey* who scored high in managing data as a product also demonstrated improvements in time to value and innovation and represented higher populations of organizations that also scored high in data governance, digital transformation, data utility, and successful use of AI.

Considering IBM for AI-Ready Data Solutions

IBM's key platform offerings for making data AI-ready include:

- IBM Datastage, a data integration solution for moving, integrating, transforming, and reshaping data from multiple disparate sources or contained within single repositories, such as data warehouses, lakes, and lakehouses
- IBM watsonx.data, a data lakehouse solution for managing structured, semistructured, and unstructured data for use in analytic and AI use cases
- IBM Knowledge Catalog, a data catalog and quality solution to improve the accuracy of and provide context for data used in analytic and AI use cases to improve the accuracy and relevancy of outcomes
- IBM Manta Data Lineage, a data lineage solution to provide intelligence about the origins and journey of data through an organization to the point of consumption
- IBM Databand, a data observability solution to monitor changes in data flow metrics, schema drift and shift, and data value anomalies occurring in data pipelines, identifying data quality issues in real time
- IBM Data Product Hub, an internal data marketplace/sharing solution to help streamline the discovery and delivery of trustworthy data products
- IBM StreamSets, a real-time data integration solution that enables users to ingest, enrich, and harness the potential of streaming data, regardless of its structure or complexity

IBM offers a versatile suite of solutions designed to function either independently or as an integrated part of its Data Fabric platform, which is a comprehensive platform that integrates point solutions to satisfy a wide range of organizational preferences.

The IBM Data Intelligence portfolio offers tight integration and unified user experiences across IBM Knowledge Catalog, IBM Manta Data Lineage, and IBM Data Product Hub, all leveraging a common set of shared platform services and built on a common design system to provide consistent user experiences across all products.

Across this portfolio of data intelligence and management capabilities, IBM is integrating AI for automated integrations, data preparation, metadata enrichments, and governance-based automation for visibility into lineage and data policy enforcement for transparency and trust. IBM's tools for data incident management, pipeline monitoring, quality assurance, and related error monitoring equip users with a broad tool set. This ensures high confidence in data quality throughout the data life cycle and contributes to enabling data product disciplines.

Challenges

Preparing data for AI applications is a rapidly evolving field. Successfully managing structured, semistructured, and unstructured data is essential, though it remains a challenging and often misunderstood task because of the historical differences in how these data types have been managed and utilized. Customers often need assistance in identifying the products and capabilities they will need to meet their business goals. While IBM's portfolio of data products is extensive, it can be a complex portfolio to understand when trying to align with customer requirements. IBM will need to collaborate closely with customers to help set priorities based on the current data management capabilities and guide them through an agile journey to making their data AI ready.

Conclusion

Modern data environments are highly distributed, diverse, dynamic, and dark. Organizations will need to take control of data in data platforms that provide contextual, quality, protected, and secured data for use in AI to improve the accuracy and relevancy of AI outcomes in AI-fueled businesses. IDC believes that the market for data intelligence, integration, and management capabilities will continue to grow as demand for AI-ready data increases, and to the extent that IBM can address the challenges described in this document, the company has a significant opportunity for success.

Modern data environments are highly distributed, diverse, dynamic, and dark.

About the Analyst

***Stewart Bond,** Vice President, Data Intelligence and Integration Software*

Stewart Bond is vice president of IDC's Data Intelligence and Integration Software service. Bond's core research coverage includes watching emerging trends that are shaping and changing data movement, ingestion, transformation, mastering, cleansing, and consumption in the era of digital business. Having worked in the IT industry for over 30 years, from early experience in database and application development through solution design and deployment to strategic architectural consulting, Stewart has worked through some significant changes in the IT industry. His depth of field experience coupled with market insight gives him a unique perspective, valued by his customers and peers.

MESSAGE FROM THE SPONSOR

At IBM, we recognize the transformative power of data in driving business success, especially in the era of generative AI. Organizations today face challenges like fragmented data stacks and data preparedness. IBM Data Fabric is designed to overcome these hurdles by streamlining data integration, curation, governance, and delivery. This innovative solution empowers organizations to build a robust data architecture, enhancing productivity for data teams and enabling faster, data-driven decision-making.

Our hybrid, AI-ready platform works seamlessly across on-premises and cloud environments, supporting various integration styles. By leveraging IBM Data Fabric, businesses can unlock the full potential of their data, paving the way for groundbreaking insights and innovation. We are pleased to sponsor this paper, emphasizing the critical role of modern data management solutions in today's digital landscape. Together, we can harness the power of data to drive business excellence and achieve strategic goals.

Learn more at https://www.ibm.com/data-fabric.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
blogs.idc.com
www.idc.com

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.