

Successful generative AI initiatives align models, infrastructure, and data management investments with use case–specific requirements for data security and compliance across hybrid cloud architectures, offering a wide range of processor, cost, and performance choices.

Optimizing Data Management Solutions for Small Language Generative AI Models

January 2025

Written by: Mary Johnston Turner, Research Vice President, Digital Infrastructure Strategies, Worldwide Infrastructure Research

Introduction

IDC's research shows that on average, organizations are only able to unlock 56% of the potential value of business data due to operational complexity, data silos, and a lack of consistent data management, security, and integration capabilities. AI represents an opportunity for organizations to extract significantly more value from these data assets. Making this data more widely available also helps support improved processes, customer experiences, and decision-making activities.

Although AI capabilities have existed for some time, earlier generations of these technologies primarily focused on predictive analytics that use statistical analysis and machine learning (ML) to recognize patterns and anticipate outcomes using historical data. The recent emergence of generative AI (GenAI) has been a catalyst for a new wave of innovation around AI models that are able to ingest many types of content, including text, image, and software code, and create new content, code, and images using that information. GenAI-powered chatbots, natural language interfaces, search, and summarization capabilities are fueling innovation across different business processes and use cases.

Organizations have invested aggressively in testing and validating a wide spectrum of GenAI use cases over the past 18 months. For many, the cost and complexity of the computing and data management infrastructures necessary to train and support ongoing updates to GenAI large language models (LLMs) have presented roadblocks in moving from proof of concept (POC) to production. LLMs, such as the well-known ChatGPT, rely on trillions of parameters, ingest petabytes of data, require high-performance computing memory and processing resources, and consume vast amounts of power. In comparison, small language models (SLMs) built using billions of parameters and terabytes of data can often deliver more targeted results utilizing platforms that require significantly less memory and processing power than the average LLM.

AT A GLANCE

KEY STATS

- » IDC expects spending on all types of AI in 2025 to top \$330 billion and represent approximately 9.5% of IT spending worldwide.
- » GenAI SLM usage is increasing rapidly. IDC's recent survey found that 25% of enterprise respondents had already deployed SLMs, with a further 17% evaluating them and 13% testing small models.
- » Among organizations that have successfully moved multiple GenAI POCs into production, an estimated 45% of use cases rely on SLMs.

SLMs can be deployed on a wider range of computational platforms, including (in some cases) traditional CPUs that work on their own or in conjunction with accelerated computing resources, such as graphic processing units (GPUs). This allows organizations the flexibility to deploy different models in different locations, from public clouds to dedicated datacenters to edge locations, depending on data privacy, security, workload performance, and overall cost of ownership considerations.

Challenges to Unlocking the Value of Business Data with GenAI

IDC expects spending across all types of AI in 2025 to total \$337 billion and represent approximately 9.5% of IT spending worldwide (see *Worldwide Artificial Intelligence IT Spending Forecast, 2024–2028*, IDC #US52635424, October 2024).

Examples of GenAI use cases include the following:

- » **Internal productivity solutions**, such as search and summarization activities, can assist sales organizations in creating contracts and bids based on existing documents or enable marketing organizations to automate the creation of web and social media content. Software developers are able to decrease time to market and reduce defect rates by using GenAI to speed up code development, updates, and reuse.
- » **Customer engagement and loyalty use cases** often take advantage of AI chatbots that allow customers to use natural language interfaces to quickly gain personalized recommendations, local language support, or self-service access to training and documentation resources. Service desk support staff can leverage GenAI to provide customers with more targeted best practice information.
- » **Digital business innovation use cases** make use of GenAI content creation capabilities to rapidly develop and launch new products and services and offer customers more automated, interactive online activities. The ability to create highly personalized graphics, audio, and video content is driving new revenue opportunities across many organizations.

Traditional GenAI LLMs are built using public data sources and are optimized for general-purpose use cases. For many businesses, however, experiencing the full value of GenAI requires that models are able to access high-value internal data, such as customer information. In many industries, this data may be under the protection of regulations, such as HIPAA rules that protect patients' health information in the United States or European GDPR rules that provide individual privacy. Once a GenAI model ingests data, it can be difficult to determine the original source of that data or protect the data's confidentiality. Public models rarely offer copyright or intellectual property (IP) indemnification and can pose challenges in terms of validating the source of information used to build the model.

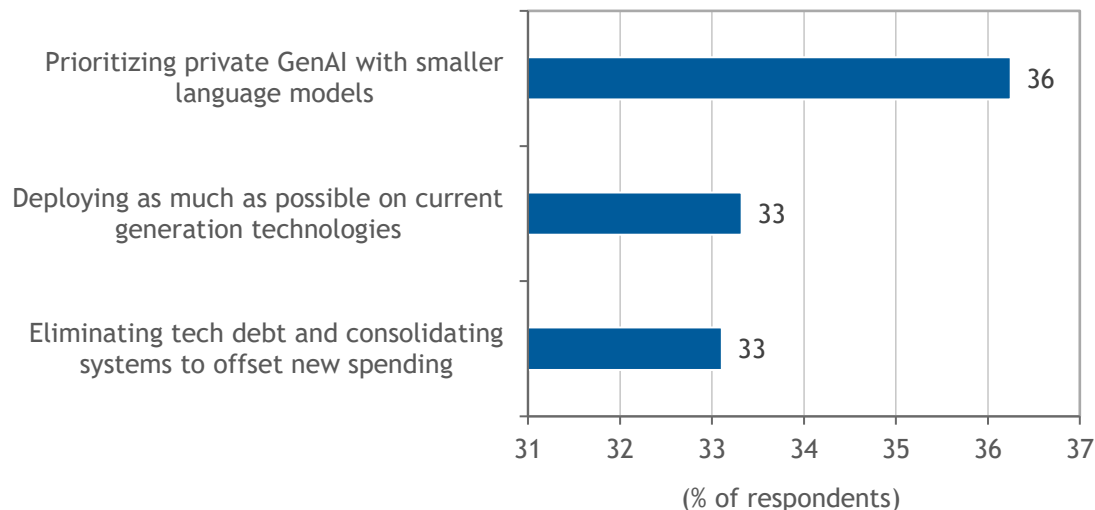
SLMs, often working in conjunction with more traditional AI/ML models and analytics, offer AI users an opportunity to train or fine-tune models using internal data and then make deployment decisions based on the specific use case requirements for performance, security, compliance, and cost. As a result, SLMs are becoming important building blocks for organizations that want to quickly harness the value of AI in targeted ways that protect a company's intellectual property.

IDC's July 2024 *GenAI Technology Trends Survey* found that 25% of surveyed enterprises have already deployed SLMs, with a further 17% evaluating them and 13% testing small models. Among organizations that have successfully moved multiple GenAI POCs into production, an estimated 45% of use cases rely on SLMs (source: IDC's *AI-Ready Infrastructure*

Tech Buyer Adoption Trends Multiclient Study, 2024). As Figure 1 shows, 36% of organizations are prioritizing SLMs, at least in part, to help better control the GenAI infrastructure costs.

FIGURE 1: ***Small Language Models Expected to Reduce the Cost of GenAI Infrastructure***

Q What are the top steps that your organization is taking to control the cost of infrastructure needed for high-performance, data-intensive workloads, such as generative AI, over the next 18 months?



n = 887

Note: Data is weighted by country IT spend.

Source: IDC's Future Enterprise Resiliency and Spending Survey, Wave 3, March 2024

The business value of SLMs is expected to increase as model orchestration and agentic AI capabilities are deployed to allow models to interact with minimal levels of human intervention. As more models are deployed into production, an organization's ability to effectively locate, access, transform, distribute, secure, and automate data management across hybrid cloud infrastructure environments will be critical to unlocking the full value of AI across the business.

IDC's recent surveys of organizations that have successfully moved AI POCs into production show that the most successful are planning for flexible, hybrid infrastructure strategies that allow them to develop or tune models in one location and deploy them in others, protect confidential data, and optimize cost and performance. Among organizations that have successfully moved multiple GenAI POCs into production, an estimated 45% of use cases rely on SLMs.

Benefits of Adopting a Platform Strategy for AI Data Operations

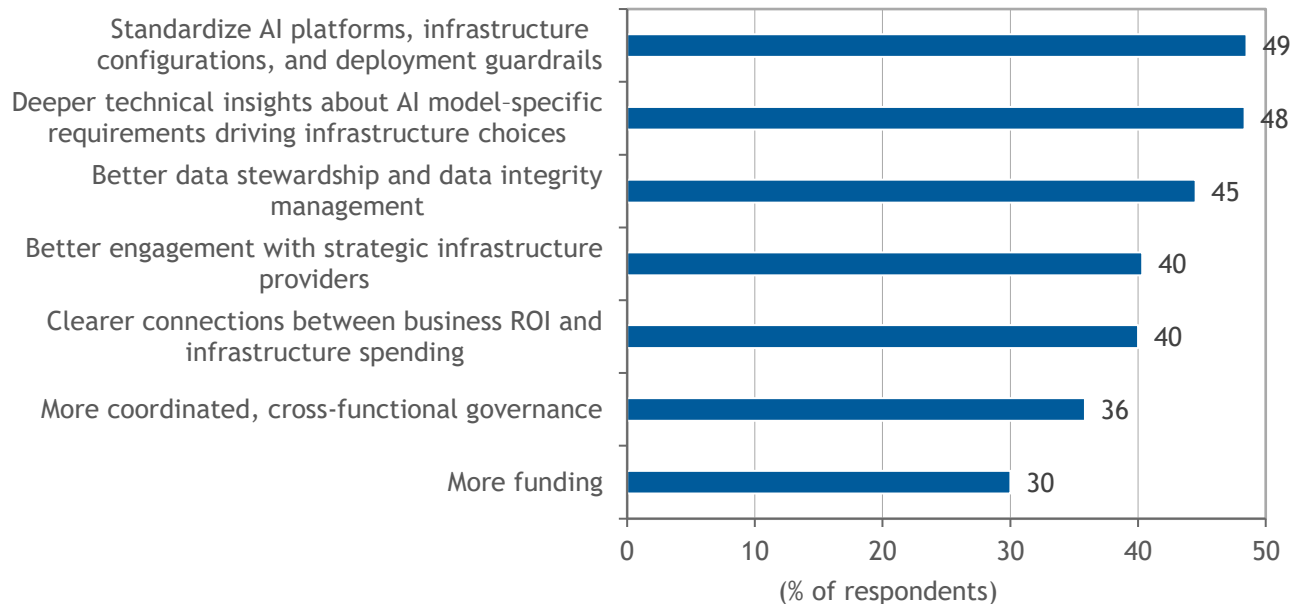
To ensure that all models can access necessary data in a manner fully compliant with company and legal requirements, organizations will need to invest in appropriate management, security, and automated controls across the board. Implementing each of these capabilities using purpose-built tools and workflows often creates data silos and makes it difficult to share data quickly and efficiently across models and systems. Platform-based approaches overcome these

challenges by centralizing data access, streamlining the process of preparing and analyzing data for machine learning models, and enabling faster development and deployment of AI applications.

According to IDC's research, AI decision-makers from organizations that have successfully moved multiple AI POCs into production identified "standardized AI platforms, infrastructure, and guardrails" as the number 1 priority in delivering the full value of AI to their organizations by 2026 (see Figure 2). They also cited improving the ability to match AI models to specific infrastructure choices and improving data stewardship as critical enablers.

FIGURE 2: ***Most Important Investment to Ensure AI Infrastructure Is Ready for Business Over the Next Two Years***

Q What are the most important things your organization needs to do to ensure its AI infrastructures is ready for business in 2026 and beyond?



n = 600

Source: IDC's AI Infrastructure Survey, September 2024

AI data platforms can provide an effective way to address these requirements by providing organizations with a consistent, scalable method for managing data across the full range of processing options and physical deployment choices. Important AI data platform capabilities to consider include:

- » **Data lakes** to provide multiple AI engines with on-demand access to data stored in various formats, including structured, semi-structured, and unstructured data
- » **Data warehouses** that store already processed and structured data optimized for specific types of business intelligence queries and reporting purposes

- » **Data transformation capabilities** that prepare raw data for use by structured machine learning models, including functions such as data normalization, log transformation, categorical encoding, data cleaning, aggregation, time series manipulation, text processing, and handling missing values
- » **Data distribution and stream processing** functions to enable customers to analyze streaming data for complex patterns and gain actionable insights, with use cases that include real-time fraud detection, network threat analysis, and manufacturing optimization
- » **Data engineering capabilities** to build and manage data pipelines, ensuring data is properly prepared and transformed for analysis by checking that clean, structured data is fed into AI models
- » **Data security analytics** built on ML and AI and incorporated into the platform to support advanced anomaly detection, behavior analytics, and enhanced threat analysis and prevention across all enterprise data resources and use cases
- » **Data governance and compliance** to allow organizations to manage data quality, security, and lineage throughout the entire AI life cycle to ensure and document the responsible and compliant use of data for AI models

Over the next several years, many organizations are expected to develop sophisticated agentic AI business processes that orchestrate insights and actions across multiple AI models and analytics. The level of orchestration and increasingly sophisticated use of AI will continue to drive demand for comprehensive and consistent data management across different processing platforms and hybrid deployment locations.

Considering the Cludera and AMD Partnership for Hybrid GenAI

Cludera AI (formerly Cludera Machine Learning powered by AMD) aims to help enterprise data science teams collaborate across the full data life cycle with immediate access to enterprise data pipelines, scalable computing resources, and access to preferred tools. The solution's centerpiece is the Cludera Open Data Lakehouse that delivers scalable and secure data management capabilities across all types of data formats and is designed to perform consistently across dedicated cloud deployment locations. The Cludera Open Data Lakehouse uses Apache Iceberg as the underlying table format to enable efficient data management with features such as time travel and ACID compliance, which are important for AI development and model evaluation. It supports open integrations with popular AI tools and frameworks such as Hugging Face and can support a wide range of model training and inferencing requirements at production scale.

Cludera AI provides data engineers and AI leaders with native and robust tools for deploying, serving, and monitoring AI models. Security enabled by shared data experience (SDX) is a core element of the overall Cludera AI architecture. Independent from compute and storage layers, SDX provides an integrated set of security and governance technologies built on metadata and delivers persistent context across all analytics as well as public and private clouds. Consistent data context simplifies the delivery of data and analytics with a multitenant data access model that is defined once and seamlessly applied everywhere. SDX reduces risk and operational costs by delivering consistent data context across deployments. IT can deploy fully secured and governed data lakes faster, giving more users access to more data while maintaining compliance and security policies.

Building on a long history of supporting sophisticated ML and advanced data visualization use cases, Cludera AI provides data engineers and data scientists with a unified interface that provides automated self-service access to a variety of data

sets and project environments, including elastic on-demand compute resources. The recently released Cloudera AI Inference service streamlines AI management and governance across public and private clouds.

Cloudera is working with AMD to develop and optimize solutions for SLMs deployed across hybrid environments. The two companies are collaborating to provide full AI data management support for models running on AMD EPYC multi-core x86-64 microprocessors for smaller models and AMD Instinct GPUs designed to support medium- to large-scale models and inferencing use cases. AMD's capability as both an x86 CPU vendor and a high-performance GPU vendor provides an opportunity for the two companies to co-innovate on mixed workload environments using a blend of general-purpose processing and accelerated computing. By leveraging the power efficiency features of AMD processors, along with collaboration from many AI ecosystem partners including infrastructure platform vendors such as Dell Technologies and cloud service providers such as AWS, the partnership is striving to provide best-in-class AI solutions to ensure that customers can implement consistent AI data management and engineering capabilities across a wide range of deployment models including on-premises datacenters, edge, and public cloud locations.

Challenges

Given the rapid pace of evolution and innovation impacting GenAI capabilities and enterprise investment priorities, the Cloudera and AMD partnership is expected to rapidly evolve over the next several years. For most enterprises, the cost and availability of AMD-powered Cloudera solutions will be a key consideration as they evaluate the different deployment models and options across on-premises, cloud, or hybrid infrastructure choices. Customers will need to engage closely with both companies to assess how their specific needs are being supported by current implementations and planned road maps for future investment.

Conclusion

By 2028, IDC predicts that 75% of enterprise AI workloads will be deployed on hybrid fit-for-purpose infrastructures to turbocharge time to value while optimizing performance, cost, and compliance (see *IDC FutureScape: Worldwide Digital Infrastructure 2025 Predictions*, IDC #US51665124, October 2024). Model training, tuning, and inferencing activities will be deployed across hybrid environments to match models with the most appropriate computing and storage options based on cost, performance, security, and compliance considerations.

Small language GenAI models will likely be important enablers of domain-specific AI use cases that work with confidential and regulated data to support high levels of customer engagement, digital business innovation, and internal productivity improvements. The ability to apply consistent data access, security, engineering, and governance across all data sources, formats, and models will be vital to enabling efficient, scalable, and cost-effective AI usage in large-scale production environments. To the extent that Cloudera and AMD are able to ensure that SLMs have access to cost-effective processors and effective data management capabilities, their partnership will provide customers with important options for realizing the full value of internal data resources across many diverse types of AI models and use cases.

By 2028, 75% of enterprise AI workloads will be deployed on hybrid fit-for-purpose infrastructure to turbocharge time to value while optimizing performance, cost, and compliance.

About the Analyst



Mary Johnston Turner, Research Vice President, Digital Infrastructure Strategies, Worldwide Infrastructure Research

Mary Johnston Turner is research vice president within IDC's worldwide infrastructure research organization and global research lead for the Digital Infrastructure Strategies practice. Mary's coverage tracks enterprise tech buyer sentiment related to compute, storage, edge, operations and cloud platforms and deployment models. Current research priorities emphasize the impact of rising requirements for data-driven AI-ready infrastructure, fit-for-purpose hybrid and multicloud architectures, autonomous operations, edge integration, and collaborative business and IT governance.

MESSAGE FROM THE SPONSOR

AMD is the high performance and adaptive computing leader, powering the products and services that help solve the world's most important challenges. Our technologies advance the future of the data center, embedded, gaming and PC markets. AMD has grown into a global company setting the standard for modern computing, with many important industry firsts and major technological achievements along the way. AMD products are built on scalable, power-efficient, and adaptable architectures designed for workloads ranging from large-scale AI model training to real-time inferencing. AMD is the only technology provider in the industry that delivers a comprehensive product portfolio to support AI deployment from the cloud to the edge to endpoints. For more info, visit www.amd.com.

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. Visit www.cloudera.com to learn more.



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

Copyright 2025 IDC. Reproduction without written permission is completely forbidden.

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