SPOTLIGHT

Sponsored by: Red Hat and Accenture

Organizations that make smart decisions around supplying enterprise data to generative AI applications will be in the best position to benefit from their investments.

Data: The GenAI Fuel That Generates Business Value

January 2025

Written by: Nancy Gohring, Senior Research Director, AI

Introduction

Since generative AI (GenAI) took the market by storm in late 2022, enterprises have focused on maximizing business value from the technology while keeping costs reasonable. Many have discovered that providing enterprise data to GenAI applications is key to driving value. However, supplying the data to applications is not straightforward, and data-related decisions have important cost implications that impact return on investment (ROI).

Factors such as where data is stored, data security and privacy requirements, and data quality, relevance, and quantity should influence critical decisions related to the following:

AT A GLANCE

WHAT'S IMPORTANT

The following are the most worrisome GenAI tech debt considerations:

- Compliance with data privacy and copyright laws (24%)
- Overinvestment in technologies that won't be needed as models become more efficient (22%)
- Limited data portability and integration across incompatible GenAI platforms (21%)

- **The best approach to adding data to GenAI:** There are several available techniques for adding data to GenAI, including fine-tuning, retrieval-augmented generation (RAG), and prompt engineering. Each approach can accomplish different goals, and organizations may use all of them to deliver high-performing GenAI applications. The data itself and the desired outcome should guide the decision about which method to use. For instance, for data sets that change frequently, RAG may be ideal for injecting enterprise data, as it retrieves information at the time of query. Fine-tuning may tap different data sets to enhance a language model's ability to perform specific tasks over a longer period of time.

- **Where to run different workloads throughout the AI life cycle:** For workloads including fine-tuning, RAG, and inferencing, organizations can choose from on premises, public cloud, edge, and AI PCs. The decision will depend on factors such as data gravity and privacy and security requirements. For instance, running a RAG implementation on premises may be best if the organization keeps the data locally and it is subject to regulatory requirements, as this approach can improve performance and reduce costs compared with moving the data or introducing latency via cloud deployment.

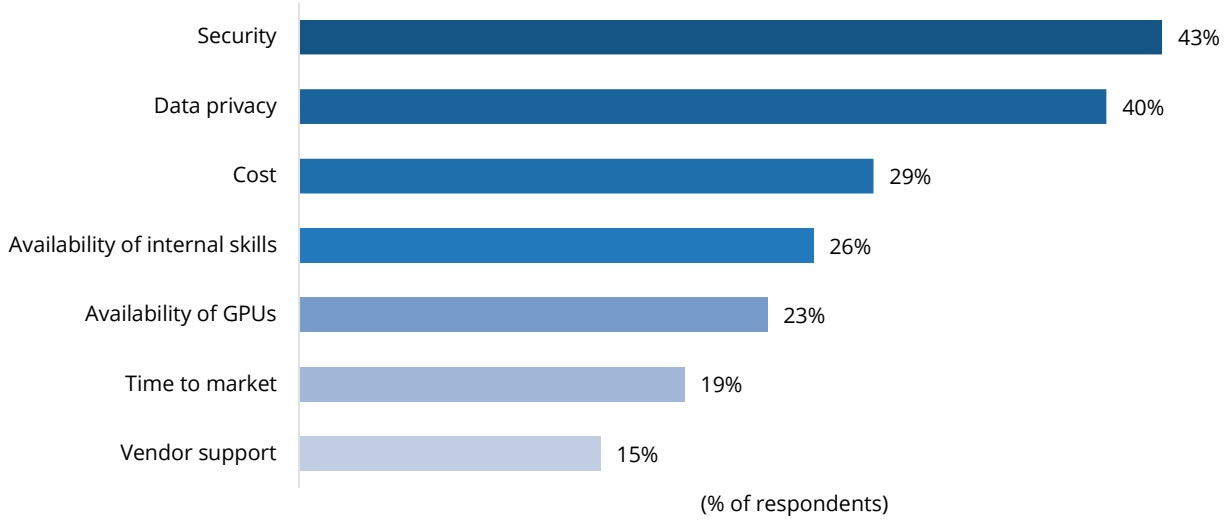
  The ideal inferencing location may depend on the model itself — some don't allow an organization to run them on premises. Latency should also be a consideration for inferencing; the geography of end users in relation to where inferencing occurs affects the time it takes to deliver an output to an end user.

The necessity and availability of GPUs may also drive workload location decisions. Not all AI workloads require GPUs, but for those that do, the volume and consistency of usage should influence the decision about whether to build out on premises or use a cloud service.

In IDC's recent *Global GenAI Technology Trends Survey,* respondents ranked security and privacy as the top drivers for deciding where to run GenAI workloads (see Figure 1). Cost was the third most important driver, notably trailing behind security and data privacy. Other drivers include availability of skills, availability of GPUs, and time to market, indicating that the decision about where to run GenAI workloads is complex.

FIGURE 1: ***Top Drivers Behind the Decision of Where to Run GenAI Workloads***

Q ***What is primarily driving your organization's decisions about where to run GenAI-related workloads, including RAG and inferencing?***



n = 624

Source: IDC's Global GenAI Technology Trends Survey, 2024

» **Model choice:** Data also influences which model is most appropriate for a particular application. An enterprise with a differentiated data set can fine-tune a large or small language model to support tasks that drive revenue or cut costs. Organizations often choose open models when they have data for fine-tuning, and they can run these models on premises or in the cloud. For open models, organizations should carefully consider the cost of inferencing and the privacy and security implications of cloud versus on-premises deployments. IDC's *Global GenAI Technology Trends Survey* found that data privacy and security requirements were the top drivers of model choice, with cost ranking second.

The starting point for data-oriented decision-making is data readiness. Many organizations require time and effort to prepare enterprise data for use by GenAI applications. An ideal data environment consists of a data platform that supports the secure, governed consumption of enterprise data by AI applications in use across the enterprise. Without this first step, enterprises will struggle to scale the use of AI in the organization or deliver business value through AI.

Smart decision-making related to data impacts the potential to achieve business outcomes for several reasons, including:

- **Cost:** Model selection, workload placement, and decisions about how to add data have notable cost implications. For instance, running a highly optimized workload on premises to leverage proximity to enterprise data may be more cost efficient than paying for public cloud usage. Overpaying for infrastructure decreases ROI and negatively impacts an enterprise's ability and willingness to invest in and benefit from GenAI.
- **Operational performance:** End users expect real-time response from applications, and GenAI applications are no different. To reduce latency, organizations must consider how model selection and workload location impact operational performance. Poor latency often leads to low usage, negatively impacting the potential to achieve targeted outcomes.
- **Quality performance:** Accuracy is crucial in GenAI applications, and model selection, along with the use of fine-tuning, RAG, or other techniques that incorporate enterprise data, contributes to quality performance. Applications that hallucinate excessively or deliver poor outputs will have limited adoption and may negatively impact the potential of other AI initiatives. Careful selection of technology to fuel GenAI applications with enterprise data should enable accurate outputs that deliver business value.
- **Risk:** Poor outcomes and leaked proprietary data are among the possible problems that put brand reputation or the ability to meet regulatory or legal compliance at risk. Model providers may offer different types and levels of indemnification to protect users against certain risks, including copyright infringement. Poorly managed risk can dramatically impact a GenAI application's ability to deliver desired outcomes.

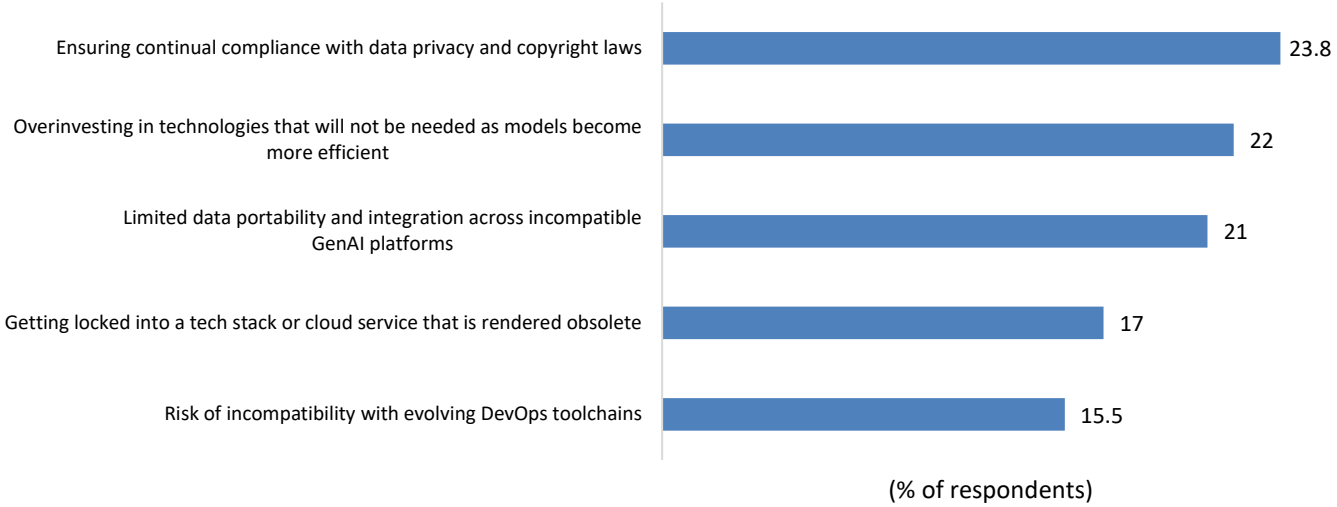
Benefits

The key to leveraging enterprise data as fuel for GenAI applications is building flexibility into the organization's technology stack so that it's possible to make changes regarding harnessing data. The benefits of this flexibility include:

- **Less accumulation of technical debt:** When enterprises create an application architecture that supports workload deployments across hybrid infrastructure and allows for easy model swapping, they can prevent the accumulation of technical debt that can result from workarounds associated with lock-in. A recent IDC study found that the top tech debt concerns for enterprises using GenAI were ensuring ongoing compliance with data privacy and copyright laws, along with overinvesting in technology that won't be needed as models become more efficient (see Figure 2). Limited data portability across GenAI platforms also ranked high as a technical debt issue. Adopting a flexible technology stack from the start can help eliminate some of these top concerns.

FIGURE 2: ***The Most Worrisome GenAI Infrastructure Tech Debt***

Q What is the GenAI infrastructure debt consideration that most worries you?



n = 571

Source: IDC's Future Enterprise Resiliency and Spending Survey, Wave 2, February 2024

- **Lower spending:** Having the ability to choose where a workload runs and selecting the model that fits the organization's criteria based on enterprise data availability and factors such as data gravity allow for cost optimization. For instance, enterprises should fine-tune a model on premises for data gravity and security reasons but run inferencing in the cloud or at the edge for potential cost efficiency.
- **Improved quality performance:** When enterprises make smart decisions about infrastructure, model selection, and data adaptation, output quality improves. Hallucinations, where language models return inaccurate information, become less frequent, and responses are more useful, mitigating risks related to data privacy, security, and compliance. The result is a more powerful application that is more likely to achieve goals such as improved productivity and efficiency.

All of these benefits reduce risk for organizations. Higher-quality applications are less likely to put brand reputation at risk because they are more secure and deliver better outcomes for users. Similarly, lower spending and the avoidance of technical debt ensure that applications can evolve and produce value to the business, eliminating the risk of unnecessary investment.

Considering Red Hat and Accenture

Red Hat offers a portfolio of AI products and services that aims to speed time to market and reduce the operational cost of delivering AI solutions across hybrid cloud environments. It is designed to enable efficient tuning of small, fit-for-purpose models with enterprise-relevant data and provide the flexibility to deploy wherever the data resides. Red Hat's AI portfolio includes:

- **Red Hat Enterprise Linux AI:** RHEL AI is a platform for developing, testing, and running the IBM Granite family of models. It includes an AI-optimized Linux instance, Granite models, open source AI tooling, and InstructLab, the method developed by IBM and Red Hat for essentially programming a language model. RHEL AI also includes indemnification with the use of Granite models, an assurance from Red Hat that copyrighted works weren't used to train the models. RHEL AI can run on a laptop, or a single server, and is designed to ease development and deployment of AI.
- **Red Hat OpenShift AI:** For larger-scale deployments, OpenShift AI provides an integrated MLOps platform for building, training, deploying, and monitoring AI-enabled applications, as well as predictive and foundation models at scale across hybrid cloud environments. The solution accelerates AI/ML innovation, drives operational consistency, and promotes transparency and flexibility when implementing secure, trusted AI solutions across the organization. With hybrid cloud capabilities, organizations can bring model training and serving capabilities to their data, no matter where the data resides.

Accenture, building on its experience with GenAI platforms, offers services for enterprises looking to scale AI development and deployment, including:

- **AI Refinery:** Accenture's AI Refinery comprises services and assets designed to enable a foundation within enterprises for leveraging AI. Core capabilities include:
  - **Autonomous agents:** Accenture helps enterprises use AI to develop agentic workflows that can autonomously execute workflows with reduced human intervention.
  - **Enterprise data:** Accenture supports organizations in building an enterprisewide repository of corporate data to power GenAI applications across the organization.
  - **Model selection:** Accenture's Switchboard offering is designed to make it easier for organizations to select the right models and swap models to deliver improvements in cost and accuracy.
  - **Model customization:** Accenture's AI Refinery enables customizing foundation models using corporate data.

Challenges

For some organizations, adopting AI is daunting because they have limited or no experience building or using AI applications. However, internal and external pressure to adopt AI is high, with competitors sharing success stories and executives concerned about falling behind. The fast pace of AI development, along with the proliferation of AI-powered tools, technologies, and services, presents significant challenges for enterprises navigating the AI space. Enterprise technology and service providers such as Red Hat and Accenture must be able to break through the noise and serve as partners to guide enterprises in applying AI to achieve a competitive advantage.

Conclusion

As organizations gain experience with AI, it's becoming clear that the ability to derive business value directly links to efficiently leveraging enterprise data. Corporate data fuels the use of AI to win rather than just survive. Smart decision-making on how to supply enterprise information to GenAI is key to benefiting from investments in GenAI.

About the Analyst

Nancy Gohring, *Senior Research Director, AI*

Nancy Gohring is a senior research director, clearing IDC's Generative AI Strategies program alongside Ritu Jyoti. Nancy covers big picture trends related to enterprise AI adoption (including GenAI). Key research themes include business, organizational, and technology architecture transformation, in the context of AI and GenAI. As part of the Worldwide AI, Automation, Data and Analytics Research practice, Nancy supports a range of clients across the technology stack.

MESSAGE FROM THE SPONSOR

Red Hat and Accenture have built a strategic alliance to help clients achieve better business results with innovative solutions for complex enterprise projects. As strategic partners, Red Hat and Accenture work hand-in-hand to bring world-class innovation to enterprise clients, while facilitating and accelerating their transformation journeys. As partners, Red Hat and Accenture bring their own unique set of capabilities — coming together in an extremely powerful value proposition — to significantly increase agility, accelerate innovation, reduce costs and improve quality and security.

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