KB

BEYOND CLOUD

DEPLOYING GENERATIVE AI IN THE ENTERPRISE

March 2024

Jim Rapoza
VP & Principal Analyst, IT

ABERDEEN
STRATEGY & RESEARCH
ILLUMINATE. INSPIRE. IGNITE.

Overview

In this Aberdeen knowledge brief, we'll look at the challenges of implementing generative AI, the key strategies and practices that enterprises need to follow to optimize their infrastructure for generative AI, and the benefits that come with selecting the right full stack solutions and partners to build a strong base for creating generative AI now and in the future.

The Impact of Generative AI on Business Practices

No current trend has had a bigger impact on both current practices and future plans than generative AI. Nearly all businesses are reworking their organizations to best leverage generative AI to bring benefits and stay competitive.

However, as generative AI has begun its ascent, businesses have mostly been stuck with public, cloud-based implementations of the technology. This is due to the newness of generative AI and the perceived high requirements in compute, resources, and unique expertise. Enterprises looking to leverage generative AI internally, however, need capabilities that protect sensitive data and give them full control over the AI models, inferencing, and output.

What makes these problems even more frustrating is that they are totally unnecessary. Businesses are letting misconceptions about how generative AI "must" be done bring additional barriers to their ability to build their own.

To overcome these limitations, we are increasingly seeing new technology solutions that provide an entire infrastructure and technology stack to enable enterprises to build their own generative AI platform. By combining powerful server and GPU compute capabilities with scalable containers and cloud-native technologies, these platforms provide a strong base for generative AI while freeing enterprises from many of the enterprise complexities that can derail a generative AI effort.

Aberdeen research has shown that, when organizations work with partners who can provide the entire stack of technologies necessary for generative AI, along with much needed experience and expertise, they can have a platform in place that lets them bring generative AI that will increase innovation, improve competitiveness, and take their business into the future.

By combining powerful server and GPU compute capabilities with scalable containers and cloud-native technologies, these platforms provide a strong base for generative AI.

Overcoming Challenges and Myths about Leveraging Generative AI in the Enterprise

It's not an understatement to say that generative AI has massively transformed technology, business, and the way we all live more than any other recent technology innovation. Which is why nearly every business is

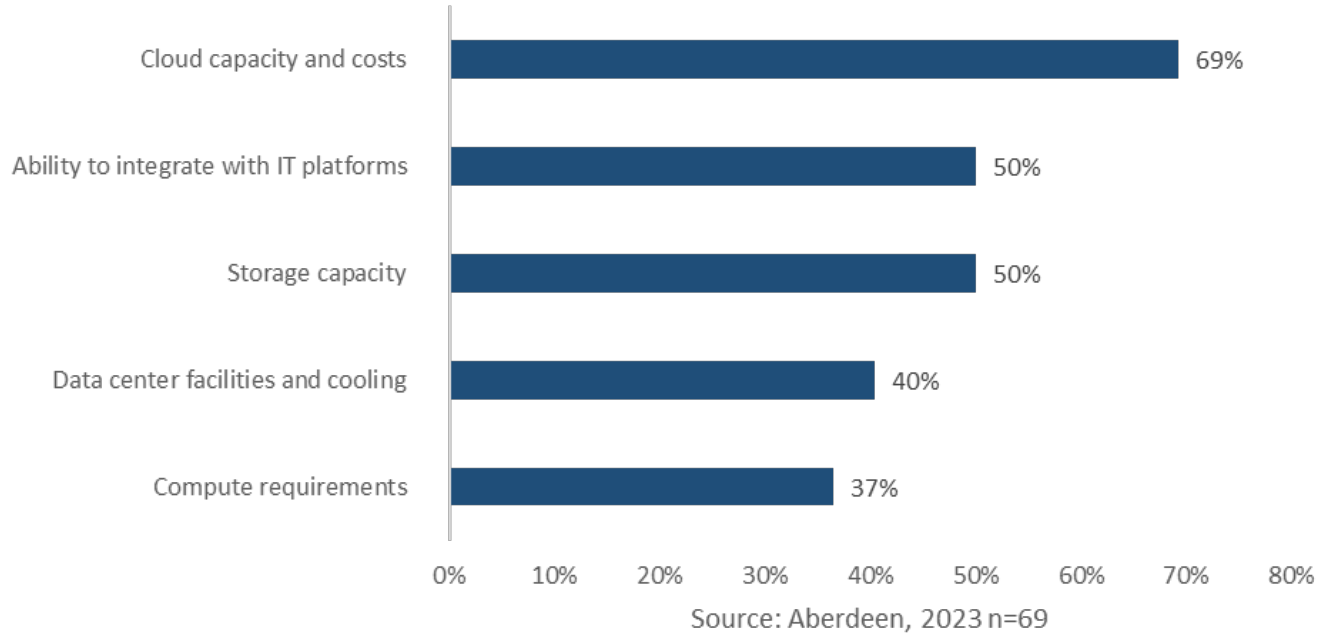
looking for ways to take advantage of generative AI, from creating content to enabling better communication to reducing workloads to predictive analytics and automation.

For many organizations, taking advantage of generative AI has seemed pretty simple (at least from a technology infrastructure standpoint) as all the top players offer cloud-based platforms to develop, manage, and deploy generative AI applications and services. But many businesses, especially larger enterprises, have quickly seen the limitations of these cloud-based offerings, including costs, privacy, and loss of control.

It's no wonder that these innovative enterprises are now looking for ways to create and run generative AI models and services within their own data centers. By having these platforms in house, they can better protect data and intellectual property while avoiding potential compliance, legal issues, and AI hallucinations that could come from using a public cloud AI solution.

However, deploying generative AI in-house comes with its own set of issues and hurdles. In Figure 1, we see the top challenges that enterprises deploying generative AI have run into.

Figure 1: Top Hurdles to AI Deployment



Examining this data, it's no surprise that cloud capacity and costs comes in as a top pain point. Enterprises are quickly discovering that the levels of data and compute that come with generative AI can quickly lead to staggering scloud costs and demands.

The next two challenges in the top five clearly relate to those organizations looking to leverage AI in their own infrastructure, as they run into challenges integrating generative AI with their existing systems (which most likely aren't designed to integrate with modern AI) and they are having difficulties with storage capacity in their data center to support AI.

Rounding out the top five challenges, we see more expected issues in data center power and cooling management and compute requirements. Basically, all of these pain points relate to the problems with cloud and on-premises systems that were never designed to work with generative AI.

Building a Generative AI Infrastructure Inside Your Business

Successful enterprises are working to overcome these hurdles by internally developing generative AI platforms to drive innovative capabilities. Our research increasingly shows that leading businesses are turning to solutions and providers who can offer an end-to-end hardware and software stack designed from the ground up to support generative AI development and deployment.

These solutions integrate state-of-the-art inferencing servers, powerful GPUs, AI development and management software, and all of the core data center capabilities needed to support powerful generative AI models. Enterprises deploying internal generative AI are looking for partners who have deep expertise to help them overcome complexity and knowledge gaps.

With these generative AI ready platforms in place, enterprises can build customized foundation models from their own secure, private data and develop applications that run anywhere. By leveraging these capabilities, enterprises overcome challenges and reap significant benefits from their investment in an internal generative AI platform.

In fact, when Aberdeen filtered our recent AI infrastructure research to identify enterprises who had adopted platforms similar to these, we found that they saw significant gains over competitors and addressed some of the key challenges to deploying internal AI, as seen in Table 1 at the top of the following page.

From an agility standpoint, enterprises with an internal generative AI solution were more likely than competitors to report faster AI application deployment and they were also more likely to have higher availability for their AI applications and services.

Table 1: Benefits of a Generative AI-ready Infrastructure

Compared to all other businesses, organizations that deploy an AI-ready infrastructure are:

20%	more likely to have faster AI app and service deployment
40%	more likely to report high management satisfaction with AI projects
45%	more likely to reduce power consumption
24%	more likely to reduce IT expense for AI
26%	more likely to report less downtime

We also found that, despite expectations of the high costs of running generative AI, these enterprises actually saw benefits in these areas. With a platform tuned to effectively run generative AI, they were 45% more likely than peers to reduce power consumption and 24% more likely to lower AI costs, all of which leads to them reporting high management satisfaction with AI projects.

Key Takeaways

How does a successful enterprise replicate the generative AI platforms built by some of the largest and most innovative companies in the world? The answer is by selecting the right technology solutions and partners, who have strong generative AI infrastructure capabilities and the knowledge and guidance to help enterprises optimize their infrastructure for generative AI.

An effective internal deployment of generative AI gives enterprises the agility and the freedom to innovate and evolve while maintaining high levels of data security and internal controls. With an end-to-end infrastructure for an enterprise level generative AI platform, these organizations can build a strong foundation for generative AI that frees them from the costs and lack of control of the cloud.

About Aberdeen Strategy & Research

Aberdeen Strategy & Research, a division of Spiceworks Ziff Davis, with over three decades of experience in independent, credible market research, helps **illuminate** market realities and inform business strategies. Our fact-based, unbiased, and outcome-centric research approach provides insights on technology, customer management, and business operations, to **inspire** critical thinking and **ignite** data-driven business actions.

18680