Executive Summary

Unlocking AI and GenAI Use Cases with AI-Ready Hybrid Cloud Infrastructure

Sponsored by: Nutanix

Dave Pearson
July 2024

Emerging AI and generative AI (GenAI) technologies are revolutionizing modern digital businesses by enhancing efficiency, optimizing operations, and deriving insights from extensive data sets. These technologies are pivotal in automating tasks, forecasting trends, and personalizing customer experiences to transform workflows across various industries. According to IDC's 2024 *Future Enterprise Resiliency and Spending (FERS)* survey, a significant majority of respondents anticipate that GenAI foundation models and technologies will disrupt their competitive positioning and operating models within 18 months. However, the journey toward AI transformation is fraught with challenges, including high costs and new risks, underscoring the importance of careful planning and execution.

Investments in AI and GenAI have become critical organizational priorities, along with security, risk, compliance, infrastructure, and IT optimization. Despite the rapid integration of AI and GenAI into digital business transformation plans, managing costs during AI model development has emerged as the top tech strategy risk factor in IDC's 2024 *FERS* survey. The shift toward generative AI and early adopters' exploration of key GenAI use cases indicate a growing maturity and confidence in AI strategies.

Addressing the lack of available skills at all stages of AI transformation is crucial, as end users seek partnerships with IT consultancy, systems integration, cloud, and digital infrastructure providers. These internal and external skills are critical to success, as organizations face challenges in determining the types of AI, use cases, sizing of models (which can inform infrastructure performance and capacity needs), and deployment methodologies. This makes AI transformation a strategic priority yet also a significant risk factor due to AI model costs.

Identifying suitable use cases for AI and GenAI and establishing an efficient, productive, and cost-effective infrastructure foundation are essential for businesses of all sizes and industries. IDC categorizes GenAI use cases into three broad types: productivity, business function, and industry-specific. Productivity use cases focus on tasks such as document search and analysis, while business function use cases integrate models with corporate data for departmental use. Industry-specific use cases often require custom work and may involve building proprietary GenAI models.

Transforming organizations have concerns despite the potential for AI initiatives based on internal data to deliver value. Privacy, security, regulatory compliance, and the need to protect internal competitive data may limit the use of public cloud resources for certain AI activities, highlighting the need for dedicated infrastructure.

For efficient and effective AI transformation, organizations must weigh the "build vs. buy" question, considering the pros and cons of each approach to deploying AI-ready infrastructure for each workload

and stage of the AI pipeline. Collaborative governance across IT, cloud infrastructure teams, and business decision-makers is vital for coordinating infrastructure decisions and ensuring optimal investment in AI-ready infrastructure. Not every use case needs overwhelming performance and scale, as smaller models and reduced data sets can satisfy many initiatives. For some workloads, CPUs with integrated AI accelerators may be more cost-effective than GPU-intensive systems. Similarly, AI-accelerated CPUs in more affordable infrastructure deployments may support particular tasks within AI and GenAI initiatives, such as fine-tuning, RAG, or inferencing at the edge.

The partnership between Nutanix, Cisco, and Intel offers a full-stack AI-ready infrastructure solution that simplifies AI deployments and manages AI workloads with pre-validated designs and Intel AI accelerators that don't require additional licensing.

Organizations navigating the complex landscape of AI transformation must prioritize GenAI use cases based on business value, cost, and potential risk. Deploying AI and GenAI workloads efficiently is crucial, especially for organizations that cannot manage more silos of complex infrastructure but need full control over their data or are looking to avoid the costs or power and cooling requirements of GPU-accelerated compute resources. By laying the appropriate AI and GenAI technology foundations and adopting a flexible approach to deployment choices, businesses can maximize their return on AI investments while still adapting to changing business and technical requirements as needed.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com