Enterprise Strategy Group by TechTarget

SHOWCASE

OCTOBER 2024

Simplifying and Optimizing AI Deployments With Integrated, AI-optimized Infrastructure

Mike Leone, Practice Director and Principal Analyst

Abstract: Organizations looking to adopt artificial intelligence for strategic use cases often see their best intentions undermined by a daunting combination of IT complexity, high implementation costs, and a lack of AI-related skills. Decision-makers can make their AI journey smoother by partnering with market-proven technology companies that integrate infrastructure stacks, orchestration software services, and AI software.

AI Deployments Accelerate as New Use Cases Emerge

AI adoption is rapidly picking up steam across all forms of the technology—generative AI, predictive AI, causal AI, machine learning, deep learning, and more. A key indicator of that progress is that organizations now see AI as part of their operations. According to research from TechTarget's Enterprise Strategy Group, 24% of organizations say AI is fully embedded in their culture and operations, and another 40% are expanding AI more broadly across the business to scale its use and support more use cases.[1]

Market Insight

72% of organizations said they saw value from their AI initiatives within three months of deployment.

Additional signs point toward further adoption of AI workloads and use cases, such as the growing support and sponsorship for AI projects from line-of-business leaders. Organizations also are heartened by early indications of positive return on investment: 72% of organizations said they saw value from their AI initiatives within three months of deployment.[2] As a result, organizations are allocating additional budget resources for AI investments necessary to move from pilot to production in a number of use cases. Take analytics, for instance: Research from Enterprise Strategy Group uncovered that 92% of organizations are allocating more budget to tools that enable them to better integrate, access, and analyze data, and 90% agreed that AI has enabled more users to do more with data.[3]

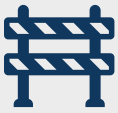
As organizations increasingly develop and deploy AI use cases for everything from creating better data insights and improving cybersecurity threat detection to generating cleaner software code, AI is hitting the mark. But could organizations deploy AI better, faster, more consistently, and more efficiently?

The Impact of Technical Complexity on Time to Value and Operational Efficiency

Although many organizations have experienced tangible benefits from their AI initiatives, several challenges prevent them from seeing value faster and deeming deployments as successful. And many of those challenges align

[1] Source: Enterprise Strategy Group Complete Survey Results, *Evaluating the Pillars of Responsible AI*, August 2024.
[2] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.
[3] Source: Enterprise Strategy Group Research Report, *Unleashing the Power of AI in Analytics and Business Intelligence*, May 2024.

Market Insight



Nearly one in three organizations cited high costs associated with implementation, difficulty scaling, and difficulty integrating with existing systems as top AI implementation challenges.

to the underlying infrastructure. While research highlights challenges associated with data quality and a lack of skills, infrastructure-specific challenges remain ever-present. In fact, nearly one in three organizations cited high costs associated with implementation, difficulty scaling, and difficulty integrating with existing systems as top AI implementation challenges.[4]

When it comes to optimizing their investments in AI, organizations must address four main areas:

- Identify the right use cases that can be delivered on time, on budget, and with maximum business value.
- Discover and audit all organizational data to best inform AI model development for training and inference.
- Reduce infrastructure complexity and enhance application performance for greater cost efficiency, easier deployments, improved competitive position, enhanced customer experience, and provable short- and long-term benefits to the organization.
- Address the full spectrum of challenges involved in creating end-to-end solutions—technical, operational, governance, risk, security, skills, and financial.

What makes developing and deploying AI workloads more complex than has historically been the case for decades with other IT initiatives? Several issues stand out, such as the massive data sets needed for AI training, the differing infrastructure requirements based on the AI workload (i.e. training vs. inferencing), the "model drift" phenomenon, the greater need for multi-departmental collaboration, and substantial—and rapidly evolving—requirements for governance, risk, compliance, and the responsible use of AI. The increased level of sophistication, precision, and flexibility requirements is unprecedented when compared with traditional data-center-based workloads that have fueled organizations for decades.

Those and other dramatic changes have had substantial impact on infrastructure design and architectural complexity, putting intense pressure on organizations to modernize their technical platforms and to overhaul how they develop and deploy their most strategic applications.

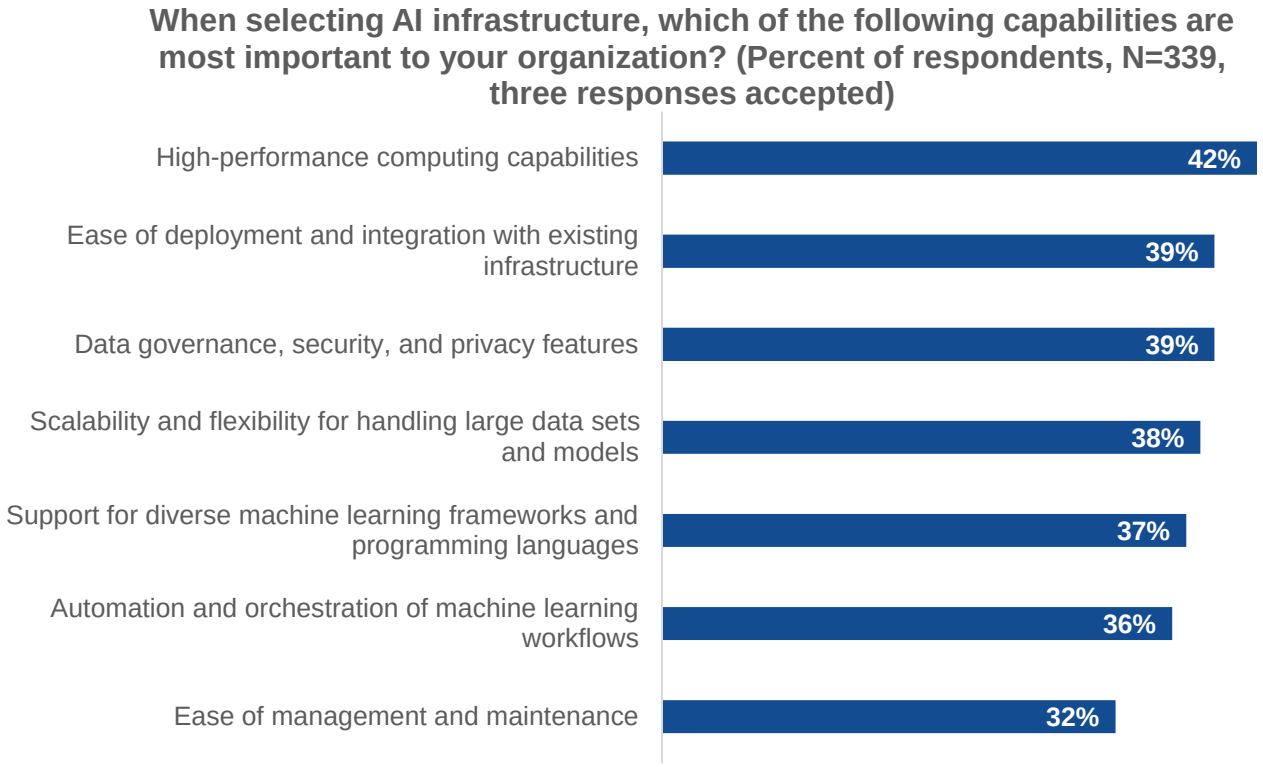
Another factor creating big challenges for organizations looking to successfully deploy AI is the widespread lack of experienced technical and business professionals comfortable with AI's requirements. This has put organizations in a position to aggressively seek out partnerships for a wide range of requirements, such as selecting the right use cases, determining how to connect AI project progress with key business goals, modernizing infrastructure to better support new AI workloads, rapidly and reliably deploying AI models into production, and enhancing the AI application development process.

Simplifying AI Infrastructure Procurement and Deployment

Unquestionably, AI infrastructure has taken on dramatically heightened importance, given the compute-intensive nature of AI use cases. This is a big reason why, of all capabilities considered when selecting AI infrastructure, high-performance computing capabilities was the top response, cited by 42% of organizations. The rest of the list clearly highlights a desire to simplify and optimize implementation, reduce risk, and deliver scale and flexibility. As shown in Figure 1, 39% of respondents reported that both ease of deployment and integration with existing infrastructure and data governance, security, and privacy features were important to their organization when selecting AI infrastructure, while 38% cited scalability and flexibility for handling large data sets and models.[5]

[4] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.
[5] Ibid.

Figure 1. Most Important Capabilities When Selecting AI Infrastructure



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

When it comes to AI initiatives that deliver fast time to value, support for a wide and growing number of use cases, and simplified deployment, organizations are looking for several important capabilities. These include:

- High-performance compute, storage, and network bandwidth infrastructure that can be quickly deployed and easily scaled. Particularly important is the need for improved performance and density from graphics processing units (GPUs) to support compute-intensive AI workloads.
- Automated deployment capabilities for fast time to value.
- Simplified management to reduce operational overhead, minimize risks and disruptions, and obviate the need for manual monitoring and management of AI infrastructure.
- Support for multiple storage options and protocols.
- Access to industry experts in the form of AI services companies that understand such issues as vertical industry domain knowledge, AI responsible use guidelines, compliance mandates, safe data-handling requirements, and AI security demands.
- Support for multiple types of deployment options, including on-premises, hybrid cloud, public cloud, private cloud, and multi-cloud environments.

These capabilities are essential in an increasingly AI-centric era of application development and deployment, especially for those focusing on containers as a facilitator of application modernization. The high performance, scalability, security, and manageability requirements of AI workloads demand a modern infrastructure platform that directly addresses the key challenges laid out earlier in this paper, including reducing infrastructure complexity, simplifying and speeding deployment, and ensuring that AI models and tools follow responsible use protocols.

Integrated, AI-optimized Infrastructure, Orchestration, and Deployment Services With Dell, NVIDIA, and Red Hat

To help organizations achieve their AI, modernization, and innovation goals, Dell has teamed with NVIDIA and Red Hat to develop and deliver Dell APEX Cloud Platform for Red Hat OpenShift, including several key NVIDIA-centric reference designs. The integrated platform makes it easier for AI application developers to deploy and run Red Hat OpenShift Kubernetes containers on bare-metal infrastructure, while supporting full-stack automation for standing up initial deployments.

It also optimizes data storage requirements—a major area of importance for handling massive data sets and various protocols (file, block, and object) typically associated with AI applications—by using Dell's software-defined storage tools and a universal storage layer. This makes it easier to move data across hybrid cloud environments that are increasingly prevalent in enterprise settings. The solution utilizes a bare-metal implementation of Red Hat OpenShift Container Platform with Red Hat Enterprise Linux Core OS to help automate infrastructure operations and simplify management and orchestration of container-based application development environments,.

The platform has been validated with several tools and use cases, including NVIDIA Blueprints, which speed up development and deployment of foundation models in any data center or cloud environment. This not only helps to increase organizations' ability to deliver demonstrable value on AI investments faster, but it also helps create a safer and more secure AI application development environment. Dell also offers a validated reference design for AI digital assistants, which is easily customizable for a wide range of use cases. These can be used in technical settings such as code generation or application testing, as well as in business environments where natural language queries using an organization's own data are a vital capability. Another key capability is the combination of Dell reference design and NVIDIA Riva speech services, which enables AI-driven audio transcription and translation that is widely used across geographies and industries. It uses Riva's automatic speech recognition and natural language processing to promote global collaboration, regardless of where stakeholders work..

Conclusion

As AI adoption accelerates, organizations seek innovative, flexible approaches to improving time to value as well as the demonstration of tangible results to justify further investments. But many organizations need tools, services, infrastructure, software, and know-how to ensure their initiatives are successful.

Increasingly, those organizations benefit from an integrated, AI-optimized infrastructure platform, augmented by proven software and AI services, industry expertise, and use case knowledge. Dell, in collaboration with fellow industry leaders and innovators NVIDIA and Red Hat, has put together a tightly integrated, end-to-end solution with comprehensive outcomes-led services offerings that simplify and accelerate organizations' AI journeys.

Dell APEX Cloud Platform for Red Hat OpenShift supports swift, reliable, and safe development and deployment of containerized AI applications. It combines on-premises infrastructure, Red Hat OpenShift Container Platform Plus, and Dell professional and technical services for AI to provide a full-solutions approach to AI based upon the overarching pillars of deployment location choice, consistent operational and developer experiences, and control over governance and risk challenges with a centralized management framework.

NVIDIA Blueprints and NVIDIA Riva speech services on OpenShift Container Platform, as well as the Dell Digital Assistant, which uses OpenShift AI on top of OpenShift Container Platform, is further enhancing the utility and value of Dell APEX Cloud Platform for Red Hat OpenShift. This deep integration and collaboration between leading AI vendors across the technology stack offers organizations the necessary architecture and simplified deployment to advance their AI initiatives faster than ever before.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com