Evolving Your Infrastructure for AI: Considerations for IT Leaders

Constructing a scalable infrastructure to support artificial intelligence includes a range of important considerations, from strategic investment decisions to assembling the right team. By addressing these fundamental factors, IT organizations can lay a foundation that maximizes the potential of their AI initiatives spanning use cases such as agentic AI, generative AI, vision AI, AI reasoning, and beyond.

1. Consider the entire stack and strategize holistically.

Fast computation undoubtedly plays a significant role in AI infrastructure. However, any weak link within the overall solution can hinder productivity. IT leaders are increasingly recognizing the importance of a robust end-to-end stack, encompassing storage, networking, and software. This ensures optimal resource allocation and prevents projects and plans that could impact scalability and prolong time to solution.

AI applications have key differences from traditional enterprise applications, and this is especially true when it comes to security and compliance. Workflows for agentic and generative AI introduce new types of risks due to the way these applications are influenced by training data and retrieved information as well as user input. It's critical for IT leaders to consider these factors when planning AI solutions.

> Enterprises can centralize AI infrastructure and provide immediate, on-demand access for teams by adopting an AI center of excellence. This enables a holistic approach to AI development and deployment with the ability to reduce risk, scale faster, increase performance, and deliver better return on investment (ROI).

As you think ahead, it's also important to stay up to date on the latest advancements in AI and its limitless potential. Take advantage of opportunities for your teams to explore and engage with AI models and accelerated inference microservices, and identify solutions that best suit your organization's needs.

2. Decide on a cloud-first, on-premises, and/or hybrid approach.

A cloud-first approach offers quick access to powerful compute resources required for AI. This equips organizations with both the scalability and flexibility they need to train and deploy AI models quickly, regardless of project size and complexity.

Additionally, the pay-as-you-go model of cloud services eliminates the need for upfront investments for on-premises infrastructure. Enterprises can save money by paying only for the resources they use, avoiding overprovisioning and reducing the financial risks associated with underutilized infrastructure.

Moreover, the cloud's extensive ecosystem of AI tools and services empowers organizations to accelerate innovation and bring AI-driven solutions to market faster, without starting from scratch.

In situations where organizations deal with sensitive data or operate in highly regulated industries, data security and compliance become paramount. In such cases, an on-premises infrastructure is crucial, as it allows data to remain in house and tightly secured.

Although on-premises infrastructure may involve higher initial expenses, it provides long-term advantages in terms of reduced operational costs. This cost-effectiveness over time makes it an appealing choice for organizations seeking to retain control over their data while maintaining predictable costs.

Adopting a hybrid approach for AI infrastructure enables enterprises to leverage the scalability and flexibility of the cloud. This allows for better resource allocation and cost management in the short term for supporting pilots, while reducing long-term costs on premises for established models that are ready to scale. This approach also allows non-sensitive workloads to be offloaded to the cloud, while ensuring the protection of sensitive data to meet compliance requirements within a self-hosted deployment model.

3. Invest in accelerated AI infrastructure.

AI requires a departure from traditional corporate IT infrastructure, as it calls for specialized hardware, software, and AI algorithms that heavily rely on parallel processing and the power of accelerated computing. Conventional, non-accelerated data centers cannot effectively handle the increasing demands of AI workloads, which often involve processing and analyzing vast amounts of data that can be accessed quickly.

Modern AI infrastructure requires high-capacity, high-performance storage solutions capable of efficiently storing and retrieving large volumes of data. Consequently, it becomes imperative to build a dedicated infrastructure specifically tailored for AI, rather than trying to repurpose existing infrastructure. AI software purpose-built for accelerated infrastructure is necessary for saving costs while delivering the highest throughput across the AI pipeline.

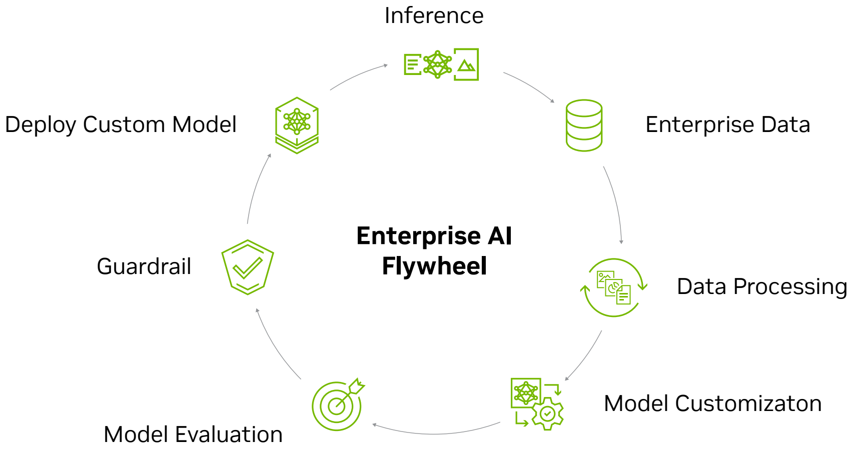
Unlike conventional data centers, AI factories go beyond just storing and processing data; they produce intelligence on a large scale, converting raw data into AI tokens that enable actionable insights in real time. For businesses and nations worldwide, this translates to a much quicker ROI, transforming AI from a long-term commitment into an immediate source of competitive edge. Organizations that invest in specialized AI factories today will be at the forefront of innovation, efficiency, and market differentiation in the future.

While traditional data centers are designed to handle a variety of tasks and support general computing needs, AI factories are specifically tailored to extract value from AI. They manage the entire AI life cycle, from data collection to training, fine-tuning, and, most importantly, low-latency AI inference.

> AI gives every company an opportunity to turn its processes into an enterprise AI data flywheel. This means identifying crucial processes, transforming them into AI-driven systems, enriching them with enterprise data, and establishing a continuous learning flywheel—enabling ongoing improvement and adaptation through AI and human feedback.

Enterprises Need to Customize and Continuously Update Models

Use enterprise data to drive productivity and business value.



4. Foster the growth of your team and cultivate AI proficiency.

With global technical staff and IT skill shortages, building a dedicated team with expertise in AI infrastructure can be a challenge. According to research, more than half (54%) of digital leaders say skills shortages prevent them from keeping up with the pace of change*.

AI agents can help bridge the talent gap by increasing efficiency, enhancing employee productivity, and automating tedious manual tasks, allowing IT staff to focus on high-priority initiatives. In security operations centers (SOCs), agentic AI can accelerate threat detection, improve operational efficiency, and streamline alert triage. By filtering and prioritizing alerts, AI agents ensure that IT and security teams concentrate only on the most urgent threats, optimizing their response time and effectiveness.

There are many AI training and certification programs available to help your team develop key skills and gain hands-on experience. Some solutions offer deployment and management services, enabling organizations to focus on business objectives, rather than building and managing infrastructure. If given the opportunity to hire more staff, IT leaders should consider those with expertise in managing infrastructure and cloud platforms, particularly those with a strong understanding of cloud technologies. Those skilled in DevOps practices and automation tools should also be considered. These individuals can streamline the deployment, monitoring, and maintenance of AI projects, providing smooth operations and minimizing downtime. IT security staff also need to learn how to manage the new risks associated with agentic and generative AI.

5. Weigh budget considerations with long-term AI goals.

Investing in infrastructure that'll work with unknown, future workloads is a crucial part of a long-term AI strategy. And with accelerated computing—which uses parallel processing on GPUs—demanding applications are sped up while increasing energy efficiency and cost savings in the long run. Fast, low-cost inference offers the key to profitable AI.

Cloud-based solutions offer a cost-effective way to start AI initiatives by reducing acquisition costs and shifting capital expenditures (capex) to operational expenditures (opex). Yet, while cloud solutions may have lower initial costs, long-term expenses can add up. IT leaders should evaluate the total cost of ownership (TCO) over time and consider factors such as data storage, compute resources, and ongoing maintenance.

In general, it's important to consider ROI as a key metric rather than the initial TCO. Building AI infrastructure requires dedicated resources, careful planning, and consideration of cloud and on-premises solutions. By leveraging the right blend of full-stack, optimized technology and strategy, organizations can navigate the challenges associated with building AI infrastructure and drive successful outcomes.

*Source: Nash Squared. *Digital Leadership Report 2023.*

Learn how NVIDIA can help future-proof your infrastructure and ensure that you're AI-ready.

Get Started With NVIDIA Leading Enterprise Solutions

Generative AI

Transformative innovation for organizations worldwide.

AI Inference

Faster, more accurate AI model deployment—from anywhere.

NVIDIA DGX Cloud

A unified AI platform on leading clouds that optimizes performance with software, services, and expertise for evolving workloads.

AI-Powered Cybersecurity

Zero-trust, real-time threat detection at scale.

NVIDIA Training

Online courses and instructor-led workshops for your teams.

Data Analytics

Accelerated analytics solutions from desktop to data center.

Request a Consultation

To learn more about how NVIDIA can help you address your business challenges, visit: nvidia.com/executives

Talk to an expert about your business goals.