NVIDIA

The Rise of Enterprise AI: A Guide for the IT Leader

Table of Contents

Unlocking the Potential of Enterprise AI

Artificial Intelligence has emerged as one of the most impactful strategies that enterprises across industries can use to stay relevant and competitive—from smart factories in manufacturing, to recommendation engines in retail, to fraud detection in financial services. As AI adoption becomes more mainstream, AI in enterprises is shifting from experimental to actual business use cases with improved outcomes, unfolding the true value of enterprise AI.

While many organizations realize the importance of AI and the heightened pace of data-driven innovations, developing and deploying AI applications can be challenging. Unlike traditional enterprise applications, AI applications are anchored in rapidly evolving and often open-source code and lack proven approaches that meet the rigors of scaled production settings. To be successful, enterprises need a flexible AI platform that can be supported in their current data center environments and can evolve to support best-in-class AI tools and frameworks.

Enterprise AI Is Transitioning from Pilot to Production

Adoption is Increasing

96% of organizations have AI initiatives in pilot[1]

92% of organizations are increasing their AI investments[1]

Driving Real Business Value

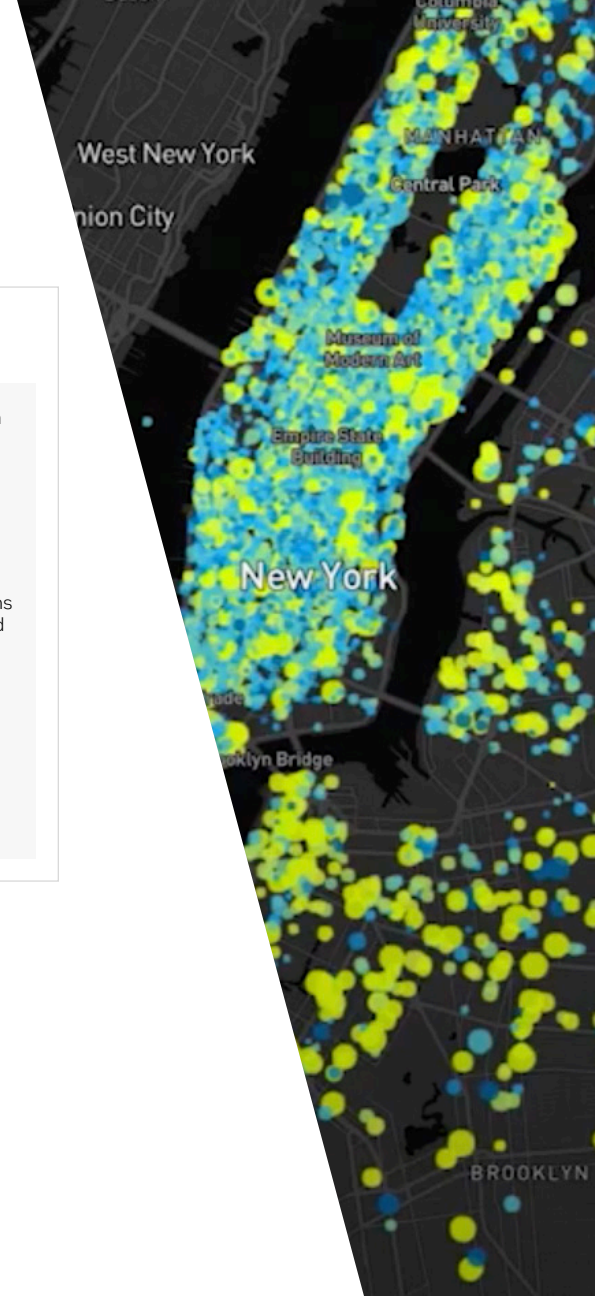
67% see increase in revenue[2]

79% see significant cost savings[2]

Starting Transition to Production AI

26% of organizations use AI in widespread production[1]

[1] NewVantage Partners, "Data and AI Leadership Executive Survey", 2022;
[2] McKinsey, "State of AI in 2021", 2021



Barriers to AI Adoption

An Explosion of Data

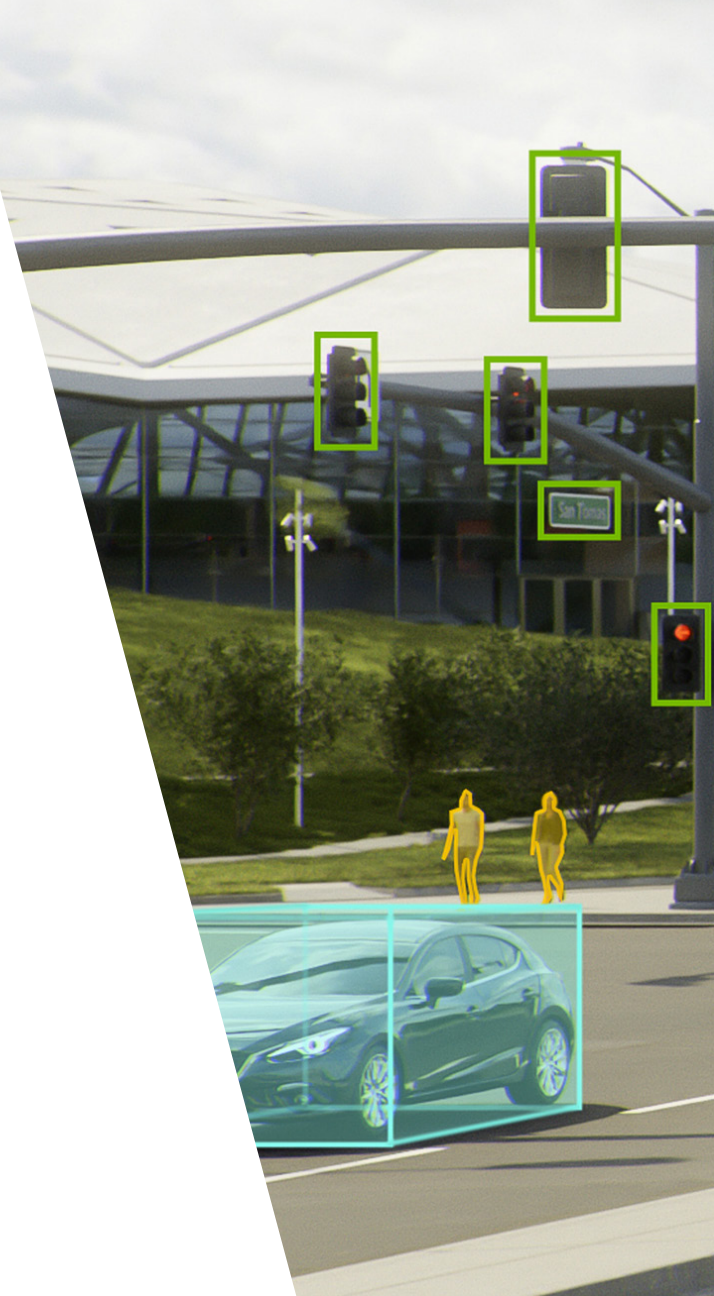
Companies are collecting more and more data that has the potential to transform their business. But as organizations accumulate increasingly large datasets, a problem arises: How do they store, evaluate, understand, communicate, and merge data into the insights that can help them make decisions? Modernizing infrastructure to support data management, preparation, integration, and training of AI models is critical to navigating the large volumes of data required for AI.

Diverse Stakeholders

AI success requires more than a specialized team of AI practitioners and data scientists. Operationalizing AI requires blended teams to ensure that business priorities are aligned. The main stakeholders of enterprise AI projects include three teams that often have different needs and objectives: AI practitioners (developers, data scientists, AI engineers), enterprise IT (IT administrators, MLOps, DevOps), and lines of business (marketing and sales, customer support, operations). For example, AI practitioners desire an agile, cloud-native platform that can quickly evolve to support the latest in AI, while the IT team is concerned about the reliability and security of the company's infrastructure and data. Lines of business, on the other hand, want to see rapid time to value of their AI projects. As the AI pipeline grows, keeping these three teams aligned can be challenging without a flexible, scalable, and end-to-end AI platform.

Integrations and Security

AI-accelerated applications can have intensive resource demands. Data analytics often involves the transfer of data across multiple systems. Many AI models, while training with large datasets, may require a vast number of compute cycles, and AI inference demands real-time response. Companies often end up deploying these AI applications into one-off, single-purpose clusters or in the cloud, resulting in operational overhead and AI silos that don't adhere to IT standards for visibility, security, or governance. This lack of integration can create real barriers to managing data effectively and securely.



Building IT Strategies for AI

When used thoughtfully and efficiently, AI solutions deliver real business value by harnessing the power of data to develop intelligent solutions. While the benefits are clear, a variety of factors—including infrastructure, deployment, integration, and scalability—make building IT strategies challenging.

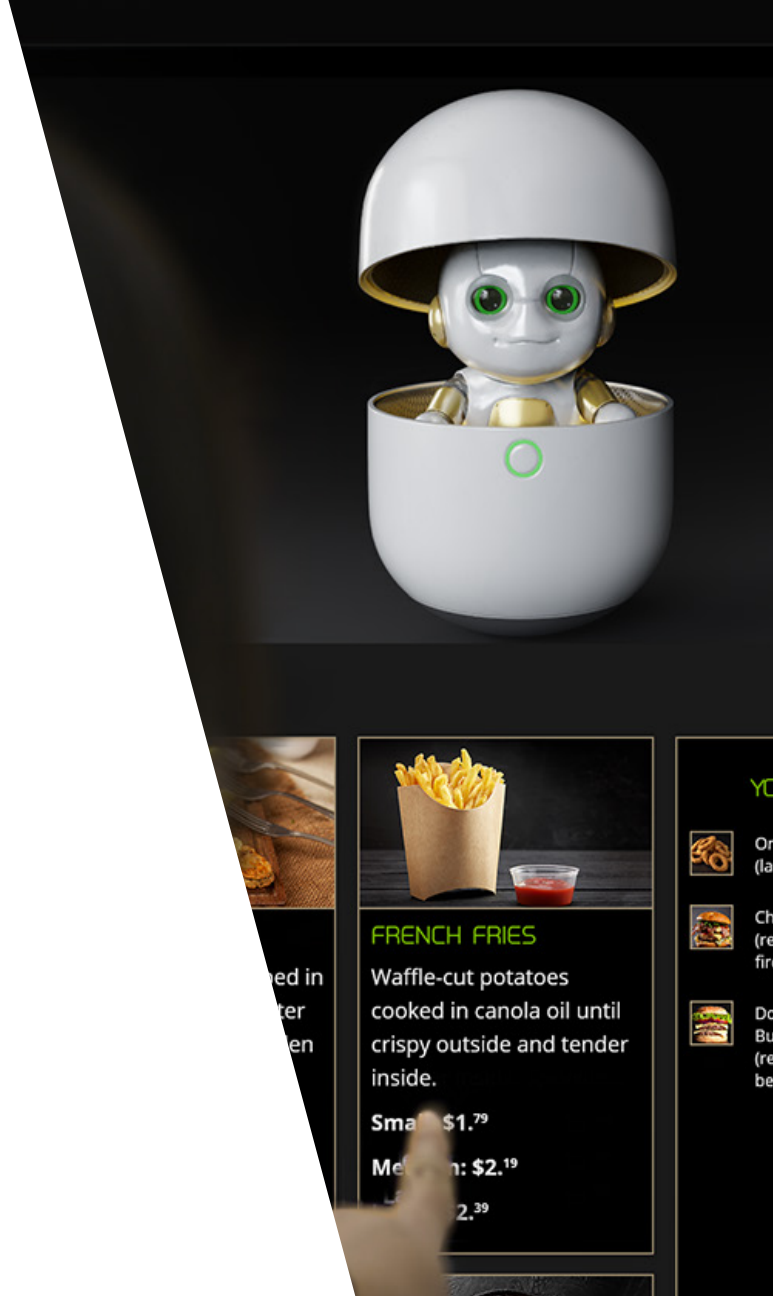
AI-Ready Infrastructure

Because AI compute requirements often lie outside standard IT practices, AI development and workloads are often deployed on siloed, bare-metal platforms or public cloud instances. Two frequently asked questions arise when an IT organization is building an AI-ready infrastructure: What's required to be AI-ready, and how can AI be integrated into the existing infrastructure?

AI is an end-to-end endeavor, from data preparation, AI training, and inference to deployment. To enable the entire pipeline, traditional enterprise data centers need to evolve to support large amounts of structured or unstructured data, the development of complex AI models for workloads like natural language processing, and the processing of incoming data to deliver AI decisions in near real time. An AI-ready infrastructure consists of an optimized operating system layer of AI frameworks and tools, compute-intensive resources, and a streamlined platform for fast time to deployment.

Enterprise-Wide AI from a Unified Data Center

Investing in accelerated data centers that streamline the development and deployment of AI workloads can fast-forward AI innovation. It can also eliminate the need for expensive, specialized hardware and the creation of infrastructure silos for running AI. Using virtual machines for an enterprise-wide scaling of AI makes the best use of distributed server, cloud, and/or edge availability. Resources can be pooled and aggregated based on AI processing demands. It also allows enterprises to bring AI to the data, rather than the other way around. With existing virtualized infrastructure, delivery models don't always have to be created from scratch and can be integrated based on the organization's needs.



A Flexible AI Platform for Hybrid and Multi-Cloud

For many organizations, the enterprise accelerated platform is distributed and runs on heterogeneous infrastructure, extending from on-premises and colocation facilities to the public cloud. While virtualized infrastructure is common, a container-based architecture is lightweight and provides additional agility. Most AI workloads perform best on accelerated infrastructure, but that isn't always readily available. Finally, orchestration and management tools are often Kubernetes-based but can vary among development teams even within the same enterprise. For these reasons, an AI platform needs to be flexible, optimized, and future-proofed to evolve with the enterprise infrastructure across virtual, container, and cloud environments.

Enterprise Support for Fast-Growing AI Pipelines

Mainstream AI frameworks and containers are mostly developed and available through open-source communities, which rely on individual developers to maintain software updates and fix security vulnerabilities, leading to support challenges for enterprises when integrating multiple open-source applications. As AI initiatives move into the preproduction stage, the need for a trusted, scalable support model for enterprise becomes vital to ensuring AI projects stay on track.

Implementing AI at Enterprise Scale

AI adoption is on the rise across every industry and has outpaced predictions for growth. To get ahead of this, enterprises need to anticipate their future AI needs and transition current workloads. Whether you're implementing your first AI project or integrating AI workloads into your infrastructure, here's a look at how to set your AI journey up for success.

Identify One AI Use Case to Start with

Many enterprises are seeing AI unlock business benefits; however, not every project can benefit from AI. Implementing AI inappropriately can not only consume time and resources but also cause strain in collaborating with cross-functional teams. Instead of initiating a large transition, start with one use case that can directly impact a business problem and garner executive support without budget challenges.

When identifying a use case, keep some factors in mind: Can this problem be easily tackled with available data that can be mined for insights using AI algorithms? Will this use case scale to a larger problem your organization is trying to solve? Is the use case simple enough to perform and explain to your business partners?

Pilot an AI Project

Getting started with an AI project can involve many steps, from building a team to choosing the right tools, AI workloads, and infrastructure. Successful AI projects require forming a team of subject matter experts to support this new initiative, in addition to recruiting data scientists and AI engineers. Modern AI workloads like machine learning, deep learning training, and deep learning inference demand different integrations of software stacks and frameworks. Make sure to choose AI tools that are performance-optimized and verified for your workloads to avoid inaccurate outcomes. Lastly, provide a robust, accelerated computing platform with an end-to-end software solution stack. It can be as simple as taking an "off-the-shelf" approach and leveraging pretested and integrated GPU-accelerated mainstream hardware.

Simplify Proof of Concepts

When conducting a proof of concept that can be used to create a preproduction application, it's important to optimize resources and infrastructure. The most successfully scalable AI solutions will be those that can be folded into existing enterprise infrastructure. Organizations can get started with minimal initial investment and leverage state-of-the-art AI software suites to simplify the development workflow.

Scale Incrementally

After the production pilot proves the effectiveness of AI, implementation will move into production by adding more AI users and expanding the production framework. Scaling AI is the most elusive challenge for enterprises and will remain a work in progress as the complexity and scope of your projects grow. Incrementally scaling your AI deployments by integrating them into your enterprise infrastructure and avoiding AI shadows will yield better flexibility, IT budgeting, and manageability.

Continuously Improve AI Models and Processes

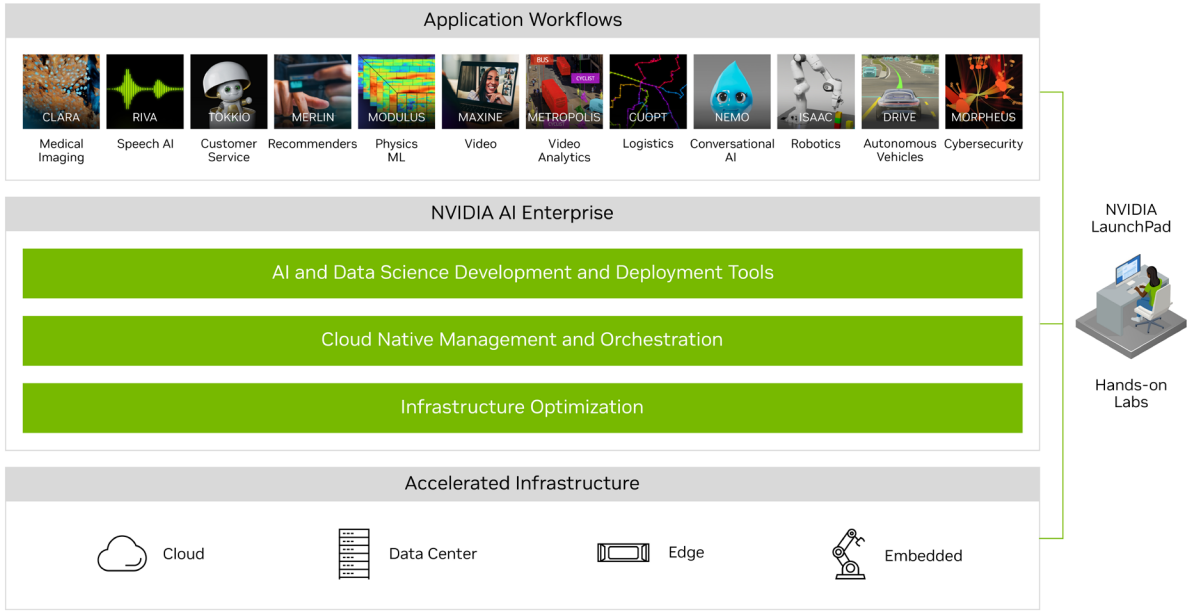
AI-driven automation and prediction are altering business processes and triggering restructure. In some areas, AI is creating new opportunities that never existed before. Transforming an organization's receptivity to AI can take time. By continuously monitoring deployments to see the gaps in workflows and optimizing your AI processes, you'll be poised to unlock AI's vast potential in your organization.

The AI-Ready Platform From NVIDIA

NVIDIA AI Enterprise - An End-to-End, Cloud-Native AI Software Suite

Optimized for every organization to excel at AI, certified to deploy anywhere from the enterprise data center to the public cloud, and includes global enterprise support so AI projects stay on track, allowing organizations to focus on harnessing the business value of AI.

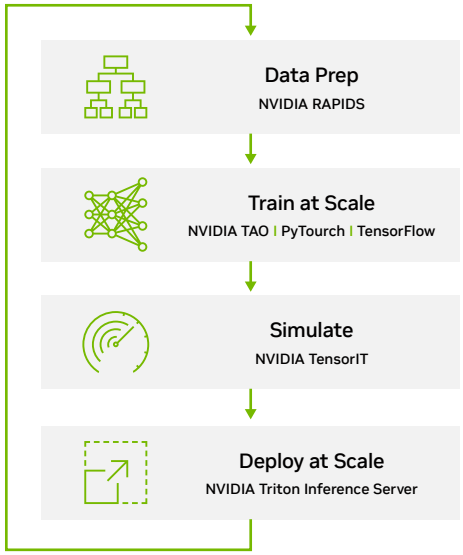
The operating system of the NVIDIA AI platform, NVIDIA AI Enterprise is essential for production and support of applications built with the extensive NVIDIA library of frameworks such as Riva for speech AI, Merlin for recommendation engines, Clara for medical imaging and more. Certified to deploy on broadly adopted enterprise platforms including multi cloud instances on AWS, Azure, and Google Cloud and NVIDIA Certified Systems from leading server vendors. Sold separately and included with NVIDIA AI flagship products, the NVIDIA H100 for mainstream servers and NVIDIA DGX Systems, to ensure organizations have access to state-of-the-art AI software.



THE KEY FEATURES OF NVIDIA AI ENTERPRISE

Essential AI Frameworks and Tools

- NVIDIA RAPIDS™ for data preparation
- PyTorch and TensorFlow for training at scale
- NVIDIA® TensorRT™ for optimized inference
- NVIDIA Triton™ Inference Server for deployment at scale
- NVIDIA TAO Toolkit for AI model development



Container Orchestration and Flexible Deployment

AI development has been accelerated in a container environment to gain portability and scalability. Certified to run on mainstream NVIDIA-Certified servers accelerated by GPUs, or CPU-only, NVIDIA DGX Systems, or in the public cloud, NVIDIA AI Enterprise can be deployed nearly anywhere and enables AI projects to be portable across today's increasingly hybrid data center. NVIDIA AI Enterprise is also certified to run on common virtualization and container orchestration platforms such as VMware vSphere with Tanzu and Red Hat OpenShift so enterprise IT can integrate AI into the data center while still relying on familiar tools and management solutions. Organizations with a hybrid cloud strategy also have the flexibility to run NVIDIA AI Enterprise on GPU-accelerated public cloud instances on AWS, Azure, and Google.

Enterprise Support by NVIDIA

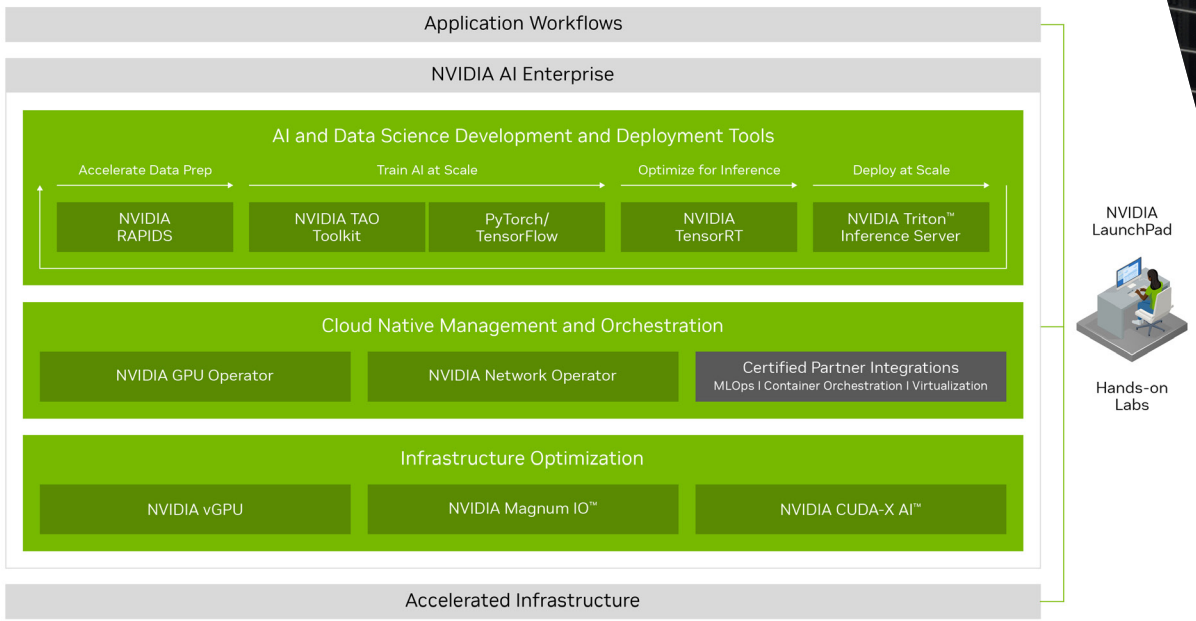
With NVIDIA AI Enterprise, organizations can get full enterprise support, including platform certification, access to NVIDIA experts around the world, ticket prioritization, and coordinated support across the full solution and partner products. They can also control upgrades and maintenance schedules with long-term support (LTS) options and access the latest customer training and knowledge base resources. NVIDIA AI Enterprise also supports NVIDIA domain-specific frameworks that developers can leverage to build innovative business solutions

NVIDIA-Certified Systems

NVIDIA-Certified Systems™ create the essential platform for the evolution of enterprise data centers, delivering infrastructure that can handle a diverse range of accelerated workloads. The certification process exercises the performance and functionality of a configured server by running a set of software that represents a wide range of real-world applications in AI and data science. Key technologies include NVIDIA Ampere architecture-based GPUs, NVIDIA ConnectX® smart network interface cards (SmartNICs), the NVIDIA BlueField® data processing unit (DPU) for accelerating networking and security, and NVIDIA converged accelerators for secure AI systems in data centers and at the edge.

A Full-Stack AI Platform for Every Enterprise

With the NVIDIA AI Enterprise software suite, AI is accessible to organizations of any size. By leveraging the existing infrastructure they've already invested in, enterprises can adopt AI into their workflows and scale the deployment with ease. The optimized, certified, supported performance and tools enable a streamlined AI platform that brings data scientists, AI engineers, IT organizations, and lines of business together to focus on creating AI business value.



Fast-Track Your AI Journey

AI implementation needs to address all the use cases that the organization may need. Whether your organization is adopting AI as a new workload or looking to build a future-proof AI environment, making confident design and purchase decisions can be daunting.

NVIDIA LaunchPad offers free, short-term access to the NVIDIA AI Enterprise software suite, letting IT professionals, AI practitioners, and data scientists experience an end-to-end AI solution workflow running on private accelerated computing infrastructure. NVIDIA LaunchPad also includes a set of hands-on labs for AI practitioners and IT staff. IT administrators can learn best practices for deploying NVIDIA AI Enterprise, and AI practitioners can learn how to optimize training and inferencing workloads using AI tools and frameworks.

Get started now: **www.nvidia.com/try-ai**

Partner Logo

NVIDIA