NVIDIA

FOUR THINGS YOU NEED TO KNOW BEFORE UNDERTAKING AN AI PROJECT

The adoption of artificial intelligence in enterprises is growing worldwide, but its impact to the bottom line varies significantly. In a recent survey by McKinsey, only a small contingent of respondents across industries attribute 20 percent or more of their earnings before interest and taxes (EBIT) to AI.[1]

Contents

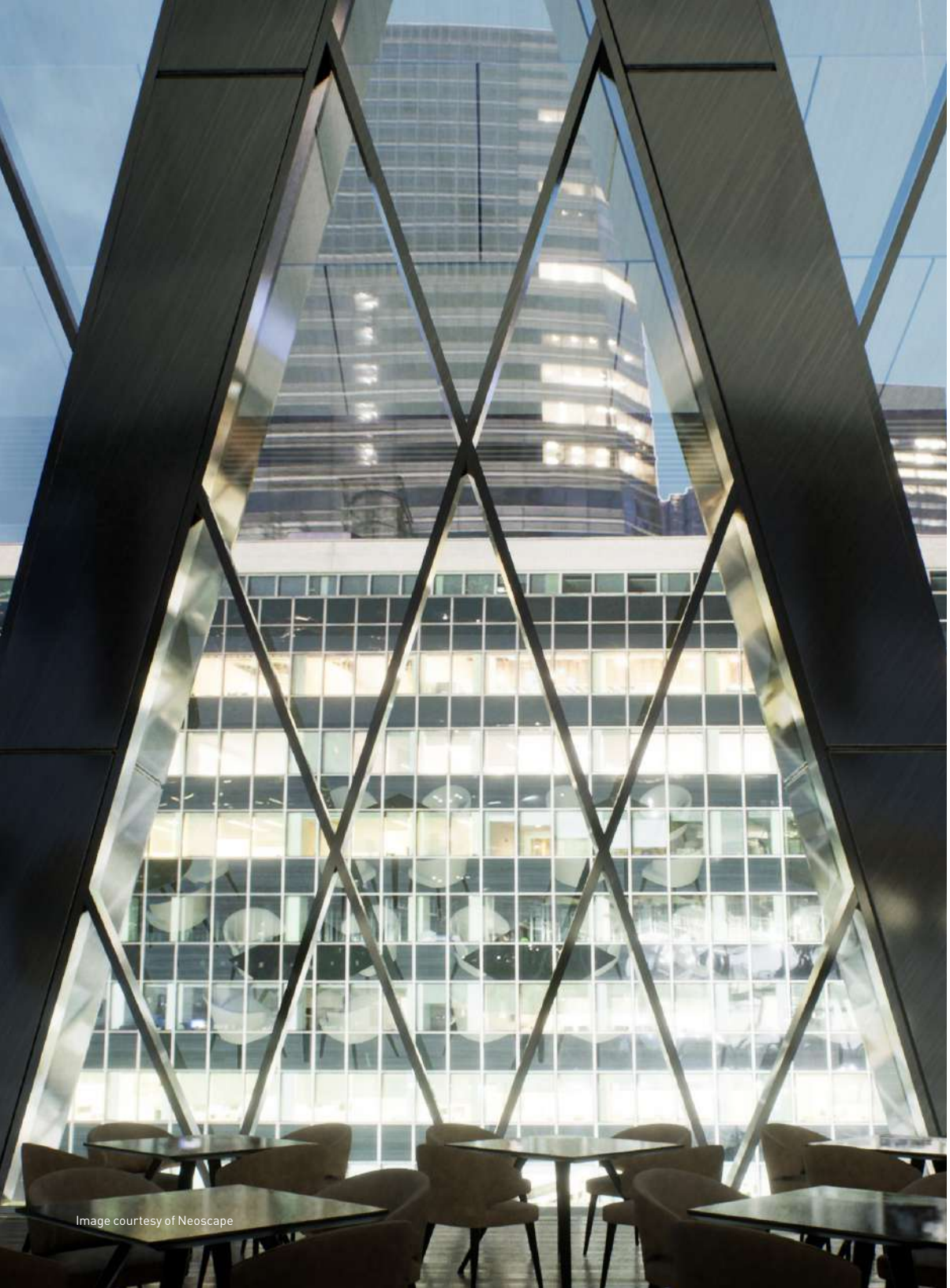


Image courtesy of Neoscape

AI high-performers attribute 20% or more of EBIT to AI[1]

Most AI projects stall or don't achieve the highest return on investment (ROI). This is due to a number of reasons: Enterprises encounter roadblocks that prevent them from getting started sooner, don't have the right AI infrastructure and tools, are unable to enhance data scientist productivity, or fail to control escalating costs. Companies seeing the most value from AI have realized the importance of proven platforms and expertise that can speed the ROI of AI investments. Read on to discover the four things that companies are doing to achieve the highest bottom-line impact from AI.

[1] McKinsey Global Institute. The State of AI in 2020. November 17, 2020.

1. DE-RISK YOUR AI PROJECTS WITH THE RIGHT SOFTWARE AND TOOLING

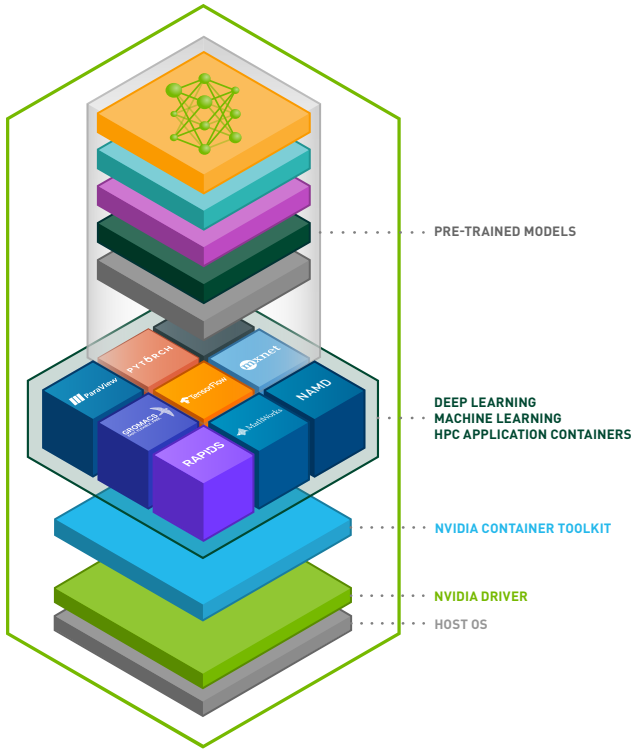
Ensure that your most valued resources don't waste time on systems integration, software engineering, or troubleshooting. Enable your data science talent to be productive from day one.

Organizations have been developing many machine learning models, but a recent study has shown only 47 percent of those models are going into production.[2] Having the right software, tooling, and practices in place is important as you get started and as you scale. Fully tested AI appliances and ready-made AI software—including pre-trained models and scripts—eliminate software engineering effort for fastest time to solution.

NVIDIA's decade-plus of AI leadership provides a known base to kick off your AI initiatives, so you can avoid roadblocks that have already been figured out. With a wealth of tools already available, you can jumpstart your AI development for a fraction of the cost and time it would take to develop them in house. For example, NVIDIA's state-of-the-art deep learning models are trained for more than 100,000 hours on NVIDIA DGX™ systems for speech, language understanding, and vision tasks. These pretrained models and scripts are freely available in the NVIDIA NGC™ catalog.

Often, data scientists, developers, and researchers not only have to do a lot of the heavy lifting—compiling and deploying AI models, optimizing AI software, and engineering code—but also must keep up with the latest updates. With the NGC catalog, customers get a 3-5X performance improvement on popular deep learning containers as new versions come out.[3] Monthly deep learning framework updates and stack optimizations deliver better performance on the same hardware, so you don't have to do this work.

NVIDIA DGX Software Stack



[2] Matthew Budman, Blythe Hurley, Abrar Khan, Rupesh Bhat, and Nairita Gangopadhyay. Deloitte Insights. Tech Trends 2021.

[3] Based on BERT-Large and ResNet-50 v1.5 training performance with TensorFlow on a single node 8x NVIDIA V100 Tensor Core GPU (32GB) and NVIDIA A100 Tensor Core GPU (40GB). Mixed precision. Batch size for BERT: 10 (V100), 24 (A100), ResNet: 512 (V100, v20.05), 256 (v20.07). DLRM training performance with PyTorch on 1x V100 & 1x A100. Mixed precision. Batch size 32,768. DRLM trained with v20.03 and v20.07.

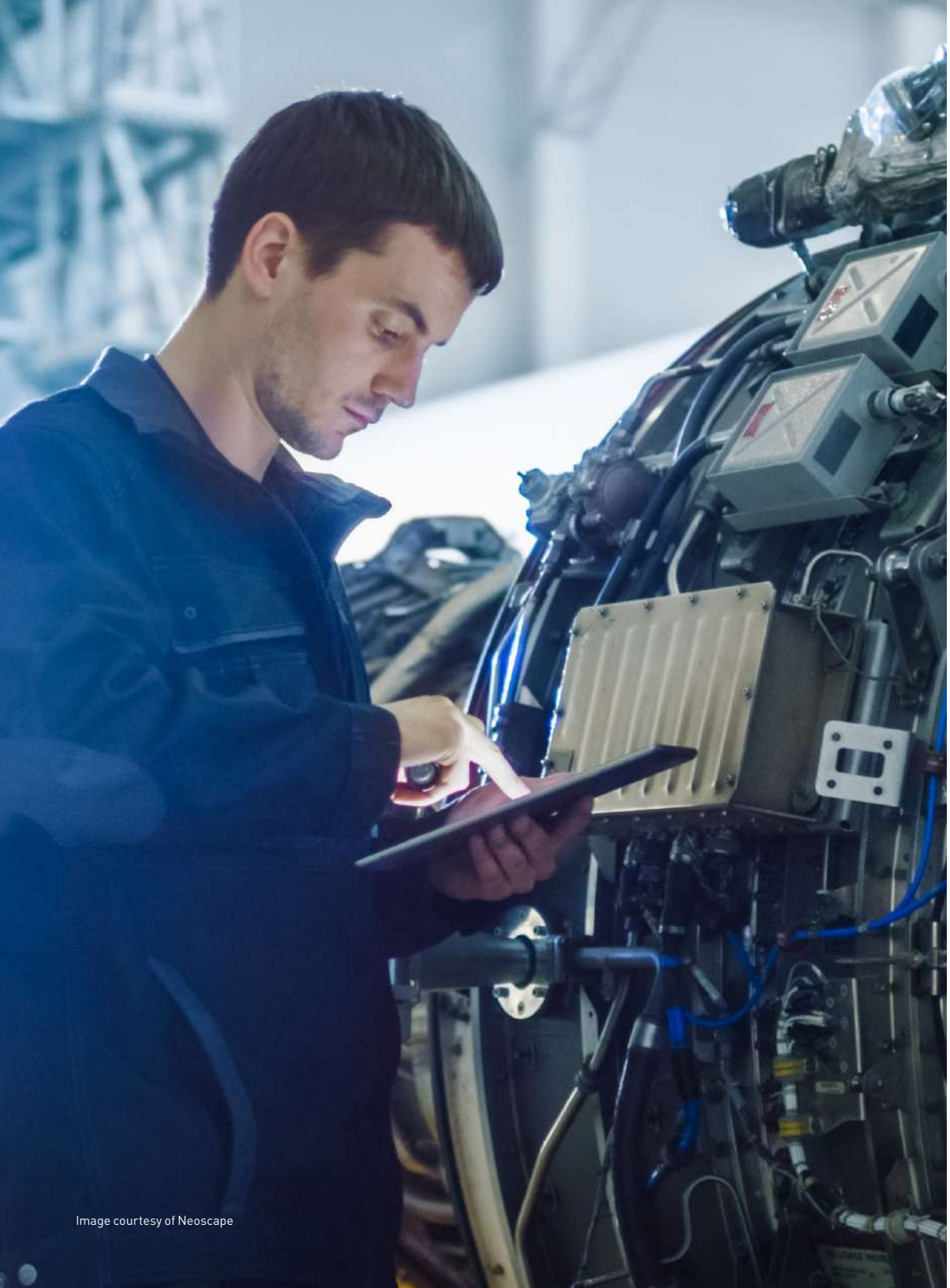


Image courtesy of Neoscape

LOCKHEED MARTIN

Lockheed Martin is using AI-based predictive maintenance to more accurately predict when to take a part out of service for maintenance, improving the availability of fleets. Using NVIDIA DGX, they experienced a **2X speedup** in training time compared to CPU-based servers with no change to architecture or code. "We achieved a **10 percent** boost in accuracy overnight because of the greater ability to train and tune parameters on the DGX," says Sam Friedman, senior data scientist in Lockheed Martin's Data Analytics Innovations Group.

Read the full case study >

2. START WITH AN AI PLATFORM THAT'S ALREADY POWERING ENTERPRISES AROUND THE WORLD

With purpose-built AI systems, your IT team doesn't need to learn a new set of disciplines to manage AI.

A recent study of AI adopters revealed that a lack of AI expertise is pervasive across the enterprise landscape; 25 percent of companies indicate they don't have enough data scientists and 23 percent say the same about machine learning experts.[4] Instead of using these valuable resources to build platforms and infrastructure, leverage the work that's already been done by leading experts in the field of AI. NVIDIA's expertise in building AI infrastructure since 2016 is incorporated into the NVIDIA DGX POD™ reference architectures (RAs), which provide prescriptive, validated approaches for building and scaling AI infrastructure in an enterprise setting. Each RA is tested at full scale and backed by industry leaders in storage and networking.

For enterprises that are struggling on where to start and how to select the software, tools, and platform they need to deliver insights quickly, the NVIDIA AI Starter Kit can help them get to business-impacting results sooner. For enterprises that need an AI center of excellence to support their entire enterprise, the NVIDIA DGX SuperPOD™ Solution for Enterprise delivers a proven platform that has enabled organizations around the globe to centralize people, process, and platform for business-wide AI development. No matter what your deployment size, your team gets the same turnkey experience without having to wrestle with platform design and an IT skills gap that can delay time to insight.

Many businesses trust their mission-critical AI endeavors to the white-glove service and turnkey infrastructure experience provided by NVIDIA. NAVER, the leading search engine in Korea, and LINE, Japan's top messaging service, created the AI technology brand NAVER CLOVA. NAVER CLOVA needed powerful AI infrastructure to deploy very large language models for new conversational AI services and enhance their chatbot and contact center solution. They were able to stand up NVIDIA DGX SuperPOD built with 140 NVIDIA DGX A100 systems and start running their models in three months with support in three key areas:

- Deployment: NVIDIA helped NAVER with installation of the physical hardware, operating systems (OS), software stack, and monitoring and management tools. NVIDIA provided onsite and remote, around-the-clock support for NAVER's DGX SuperPOD hardware.
- Validation and testing: NVIDIA helped NAVER understand baseline performance, test individual nodes, and test at scale. These metrics allow NAVER to understand if their systems are performing well relative to each other.
- Knowledge transfer: After power-on, NVIDIA ensured that NAVER can operate and manage their DGX SuperPOD effectively, providing onsite and remote assistance to the customer.

[4] 451 Research, part of S&P Global Market Intelligence.
Voice of the Enterprise: AI & Machine Learning, Infrastructure—Advisory Report, August 2020.

Clova

With NVIDIA's team providing onsite and remote support, from the physical cabling of the 140 DGX A100 systems to installing deployment and cluster management software, it took NAVER CLOVA **only three months** from initial engagement to power on their NVIDIA DGX SuperPOD. It took **only one month** to go from an empty colocation data center to bringing the customer online. The large natural language model built using NVIDIA DGX SuperPOD will serve as a core platform for all NAVER services and will be provided through Naver Cloud Platform, a public cloud service.

Learn more about NVIDIA DGX SuperPOD solution for enterprise >

The DGX SuperPOD is helping NAVER CLOVA to build state-of-the-art language models for Korean and Japanese markets and evolve into a strong AI platform player in the global market. Built on a well-defined, long-standing methodology, from pre-staging to pre-deployment simulations to quality assurance (QA) tracking, NVIDIA can ensure customer success, backed by a full team—including a project manager, a data center site manager dispatched to the customer, and an escalation team. And with a global integration partner network, tens of resources per project are executed simultaneously around the world. NVIDIA makes scaled AI infrastructure turnkey with professional services that support the full lifecycle, from design to deployment to operations to optimization.

3. UPSKILL YOUR TEAM AND TURN AI INTO A TEAM SPORT

Deploy a system that includes direct access to experts who understand your full stack and have seen your impending issues before.

In a Deloitte study, 68 percent of surveyed executives described their organization's skills gap as "moderate to extreme," with 27 percent rating it as "major" or "extreme."[5] Those who have seen success in AI have addressed this gap; they have carefully chosen partners who have an extensive experience in AI infrastructure at scale, have thousands of systems in operation, and who understand the full stack. They have likely already seen your application, framework, model, GPU, storage, or network problem before and can easily troubleshoot, so you can achieve faster ROI.

NVIDIA can provide all the knowledge and partnerships you need to make your AI projects successful sooner. With every DGX comes a global team of AI-fluent practitioners who offer prescriptive guidance and design expertise to help fast-track AI transformation. This ensures mission-critical applications get up and running quickly and stay running smoothly, dramatically improving time to insights. NVIDIA DGXperts work directly with a customer's AI point person to make that person instantly productive.

Thousands of Leading Companies Deploy DGX Systems Today

- **9** OF THE TOP 10 **GLOBAL UNIVERSITIES**
- **6** OF THE TOP 10 **US BANKS**
- **8** OF THE TOP 10 **GLOBAL TELCOS**
- **7** OF THE TOP 10 **CONSUMER INTERNET COMPANIES**
- **7** OF THE TOP 10 **US HOSPITALS**
- **7** OF THE TOP 10 **GLOBAL CAR MANUFACTURERS**
- **10** OF THE TOP 10 **US GOVERNMENT INSTITUTIONS**
- **10** OF THE TOP 10 **GLOBAL DEFENSE COMPANIES**

[5] Deloitte Insights. Talent and Workforce Effects in the Age of AI: Insights from Deloitte's State of AI in the Enterprise, 2nd Edition survey. March 2020.

Scotiabank®

Scotiabank is using AI to develop more accurate scorecards that can determine whether they grant a loan to applicants. The customer worked directly with an NVIDIA DGXpert who helped them develop features to generate more complex scorecards while maintaining the model's explainability. The bank can now generate scorecards **6X faster** using a single GPU in a DGX system compared to what used to require 24 CPUs. "In a way, the best thing we got from buying that system was all the support we got afterwards," said Paul Edwards, director of data science and model innovation at Scotiabank.

Read the blog >

4. CONTROL COSTS BY CONSIDERING HOW INFRASTRUCTURE ADDRESSES DATA GRAVITY

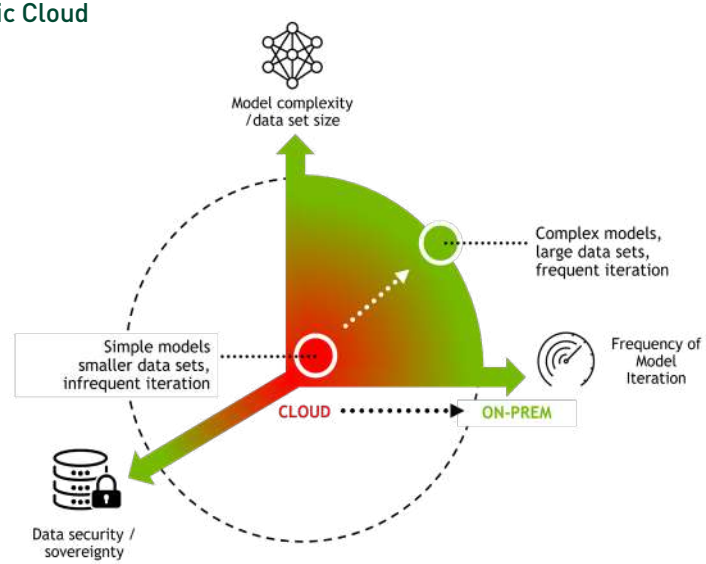
Train where your data lands in order to achieve the lowest "cost-per-training-run."

AI has transformed the traditional software development life cycle (SDLC), from enabling rapid prototyping to automating data analysis. Data plays a vital role in ensuring an AI model is accurate and maintains accuracy over time as new data is added. Because of this, data gravity—an analogy of data's ability to attract additional applications and services—comes into play. In a recent IDC survey, 84% of businesses are repatriating workloads from the public cloud as the costs of data gravity are driving workloads on-premises.[6] While cloud-first or cloud-only works for conventional application development, AI apps are uniquely disadvantaged by data gravity. If your compute resources and the data they need to act on are separated by distance and network latency, then this data gravity is working against your workflow, and more time and money is spent resisting it.

Some organizations turn to the cloud for the early phase of AI projects, as this is dominated by experimentation and sporadic GPU spikes. But as models become more complex, data sets start growing exponentially. And with more frequent model iteration, teams and data scientists hit an inflection point where data gravity starts to significantly drive up costs. Organizations are starting to realize that they need to train where their data lives, using a purpose-built co-resident AI infrastructure to achieve the lowest cost-per-training run.

Hybrid architectures that let organizations own the base and rent the spike offer the best of both: lowest infrastructure cost for ongoing demands paired with the cloud for temporal spikes. Many customers today are taking advantage of an easy-to-scale, fixed-cost infrastructure with NVIDIA DGX systems. For customers who don't have a data center, colocation facilities are available to house their infrastructure where their data lives. And with financing options like leasing or as-a-service (aaS) offerings that combine the simplicity of the cloud with the performance of a dedicated system, NVIDIA is making it easier than ever for customers to deploy and scale AI.

Top Drivers: Why Enterprises Are Repatriating AI Workloads from the Public Cloud



[6] IDC 2020 Cloud Pulse Survey and IDC 2020 Workload Repatriation; Placement Best Practices.

Milwaukee School of Engineering (MSOE) needed tremendous computational resources and an optimized software stack to meet growing AI workloads. As cloud instances were limiting experimentation, they turned to NVIDIA DGX systems and NVIDIA T4 Tensor Core GPU-based servers. Today, **80 percent** of their computer science students are actively using the cluster, and faculty GPU usage has increased by **10X**. "With NVIDIA DGX systems, our students had access to the best-in-class AI infrastructure and no longer had to worry about the cloud 'odometer' always running and limiting experimentation," said Dr. Derek Riley, associate professor and program director at MSOE.

Read the full case study >

THE RIGHT FORMULA FOR AI SUCCESS

Successful enterprises who have adopted AI are distinguished by their ability to de-risk their AI projects with the right tools, software, and AI infrastructure from the start. With the proper tools and infrastructure in place, these enterprises know how to make their data scientists productive immediately, enabling them to innovate without worrying about escalating costs.

By adopting these learnings, you can uncover insights faster and ensure higher ROI for your AI projects, sooner.

Fast-track your AI journey with NVIDIA AI solutions on NVIDIA DGX™ systems, powered by NVIDIA A100 Tensor Core GPUs and second-generation AMD EPYC™ CPUs, at: **nvidia.com/dgx**

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, DGX, DGX POD, DGX SuperPOD, NGC, and RAPIDS are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. All other trademarks are property of their respective owners. JUL21