Ebook

Four Things You Need to Know Before Undertaking an AI Project



AI adoption in enterprises is growing worldwide, but its impact on the bottom line varies significantly.

In a recent survey, only 8 percent of organizations attributed 20 percent or more of their operating profit to AI, while an overwhelming 92 percent of enterprises struggle to achieve this level of AI impact.[1]

Contents

4 Do Your Homework

6 De-Risk Your AI Projects With the Right Software and Tooling

9 Upskill Your Talent and Turn AI Into a Collaboration

11 Realize Your AI Vision With a Scalable, Purpose-Built AI Platform

AI High-Performers Attribute 20 Percent or More of Operating Earnings to AI[1]

Most AI projects stall or don't achieve the highest return on investment. 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 boost 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. Do Your Homework

Map out your AI journey, from problem definition to infrastructure planning.

Before starting an AI project, define your business problems, identify AI-solvable ones, and set success metrics. Understanding the problems will let you prioritize and create a transformational roadmap. Consider feasibility, quick wins, and data readiness when prioritizing to avoid frustrating data scientists. It's also important to establish early success metrics such as impact on revenue and efficiency to gain buy-in.

Having an AI strategy and roadmap will better enable you to determine your needs. For example, are you building mainstream applications like computer vision, or do you need to build a state-of-the-art generative AI model that's deep and broad with high accuracy to serve endless use cases such as a large language model for summarizing financial documents, performing multilingual translation, or creating brand specific content? If you're embarking on the latter, consider purpose-built AI infrastructure that's optimized for multi-node training and can better accelerate the training of large AI models.

IDC research surveys reveal that inadequate infrastructure performance and availability is one of the top factors impeding graduated success in generative AI initiatives.[2]

Scaling AI requires turnkey AI infrastructure, along with AI expertise and operational discipline. Instead of using your valuable resources to design and deploy infrastructure, leverage the platform already used by leading experts in the field of AI. Built from the ground up for enterprise AI, the **NVIDIA DGX™ platform** incorporates the best of NVIDIA software, infrastructure, and expertise in a modern, unified AI development and training solution. The DGX platform's inclusive access to NVIDIA AI expertise and support enables you to get answers in real time, instead of having to search through forums for information about infrastructure or tooling or tips for how to optimize your code.

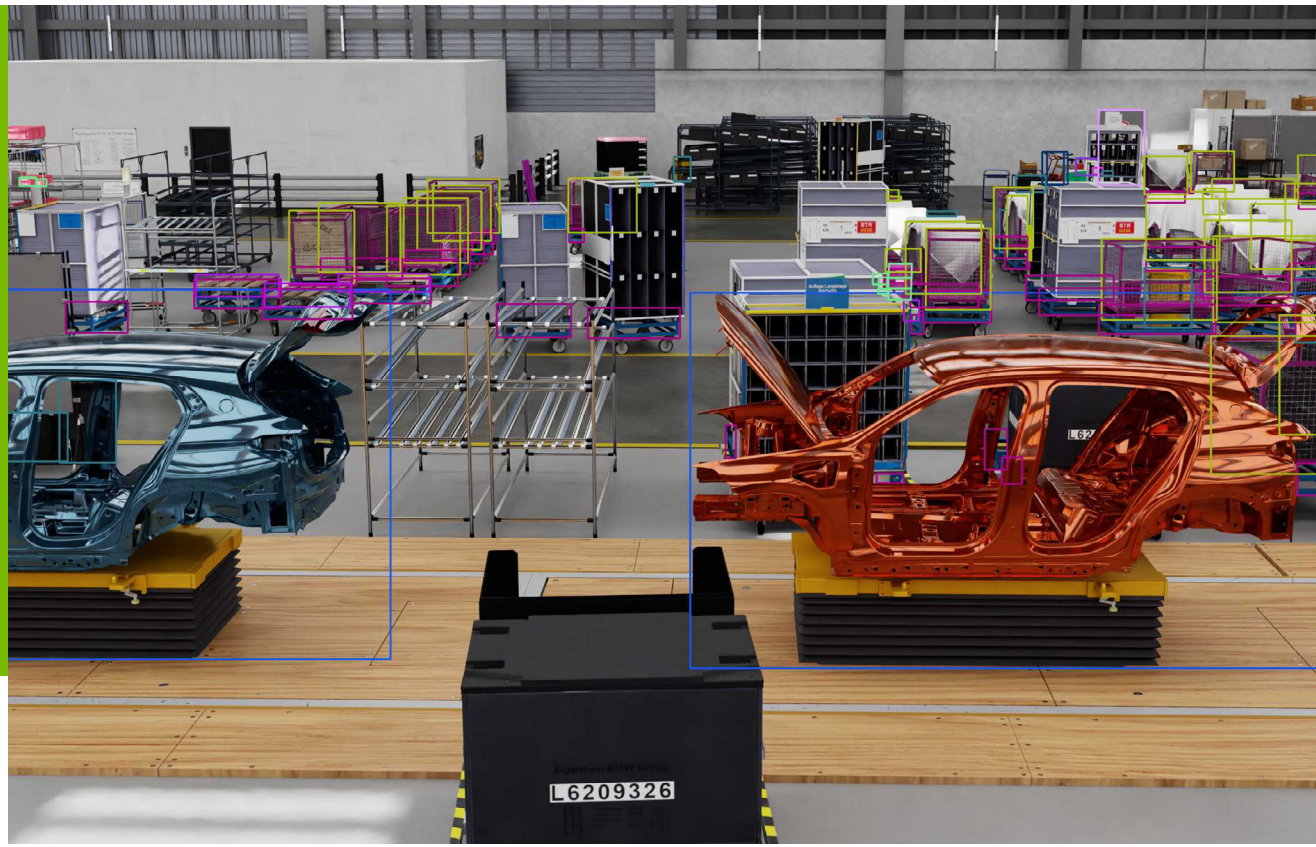


BMW GROUP

BMW is using DGX systems with the NVIDIA Hopper™ architecture as part of their large-scale sustainability strategy, called Green Physics AI. Each asset in their factory is supplemented with metadata from SORD.ai, the largest synthetic dataset in manufacturing. SORD.ai can segment and label images and continuously learns from the vast collection of images it has synthetically generated. The metadata used in BMW's factory includes an object's CO2 footprint, age, energy consumption, and distance traveled over a span of 15 years. The SORD.ai dataset lets manufacturers develop powerful AI models and create digital twins that optimize the efficiency of factories and warehouses.

The powerful computing capabilities of DGX systems enables BMW to handle complex data analysis and process vast quantities of metadata seamlessly. As a result, BMW can maximize energy and CO2 savings for the factory's products and the components that go into them. It's a remarkable leap toward greener and more efficient manufacturing.

Read the Blog



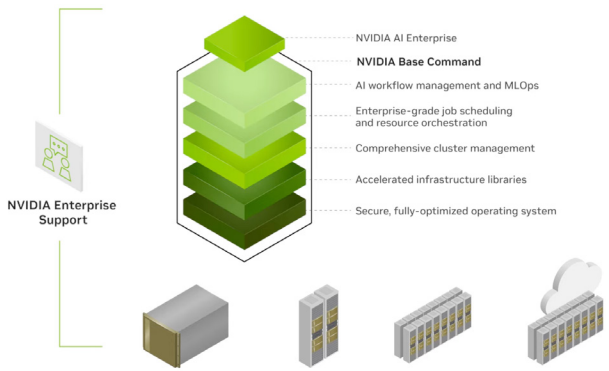
2. De-Risk Your AI Projects With the Right Software and Tooling

Ensure that your most valued resource—your data science talent—doesn't waste time on systems integration, software engineering, or troubleshooting.

Organizations have been developing many machine learning models, but a recent study shows that only 47 percent of those models are going into production.[3] Having the right software, tooling, and practices in place is important as you get started and as you scale. A fully tested AI platform and ready-made AI software—including pretrained models and scripts—eliminate software engineering effort for fastest time to solution.

With **NVIDIA Base Command™**, you get the same software that supports NVIDIA's thousands of in-house developers, researchers, and AI practitioners. Base Command includes valuable MLOps tools for your IT team such as cluster management, job scheduling, and orchestration, so you can ensure developers working on shared resources get the computing resources they need. And with programming libraries that accelerate compute, storage, and network infrastructure, as well as system software optimized for AI, you get the best performance for your AI workloads. No more cobbling together multiple vendor or open-source products or searching for answers on forums to support a "DIY" software stack.

When it comes to system optimization, developers often need to do a lot of the heavy lifting, such as adapting their code to take advantage of multiple GPUs or even multiple systems in a cluster. NVIDIA Base Command understands the underlying topology and how to leverage it to achieve the 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.

NVIDIA AI Enterprise is also included with the DGX platform, helping organizations jump-start their AI development and deployment for a fraction of the cost and time it would take to create pretrained models and frameworks themselves.

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.

NVIDIA AI Enterprise also undergoes regular deep learning framework updates and stack optimizations to deliver better performance on the same hardware. Customers have routinely experienced a 3–5X performance improvement on popular deep learning containers with new version releases.[4]

NVIDIA AI Enterprise and Base Command work together to simplify and streamline productivity, allowing developers to start with NVIDIA-optimized frameworks, libraries, and pretrained models, as well as manage end-to-end training workflow and job orchestration. With NVIDIA AI Enterprise, enterprises can train on their DGX systems and deploy at scale with **NVIDIA NIM** inference microservices, supporting an end-to-end AI workflow.



LOCKHEED MARTIN

Lockheed Martin is using AI-based predictive maintenance to more accurately determine 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 DGX,” says Sam Friedman, senior data scientist in Lockheed Martin’s Data Analytics Innovations Group.

Read More Customer Use Cases



3. Upskill Your Talent and Turn AI Into a Collaboration

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

A recent study of AI adopters revealed that a lack of AI expertise is pervasive across the enterprise landscape: 44 percent of respondents stated insufficiencies in technical skills in scaling AI initiatives. And even after implementation, 50 percent of respondents stated challenges in maintenance or ongoing support after initial launch.[5] Those who have seen success in AI have addressed this gap. They've carefully chosen partners who have extensive experience in AI infrastructure at scale, have thousands of systems in operation, and who understand the full stack. Odds are, they've already seen your application, framework, model, GPU, storage, or network problem before and can quickly troubleshoot, so you can achieve faster ROI.

With every DGX comes a global team of AI-fluent practitioners who offer prescriptive guidance and design expertise to help fast-track AI transformation.

NVIDIA can provide all the knowledge and partnerships you need to make your AI projects successful sooner. NVIDIA offers training to ensure your IT organization is fully prepared to leverage your DGX investment, covering deployment, optimization, management, and more. And every DGX system comes with access to **NVIDIA DGXperts** and Enterprise Support, ensuring that mission-critical applications get up and running quickly and stay running smoothly, dramatically improving time to insights.

"Speed of innovation is important, as time to market is critical for us. One of the things that was most striking to us is NVIDIA's ability to support us—from optimizing natural language processing models to testing new framework features for rollout to sizing of AI infrastructure."

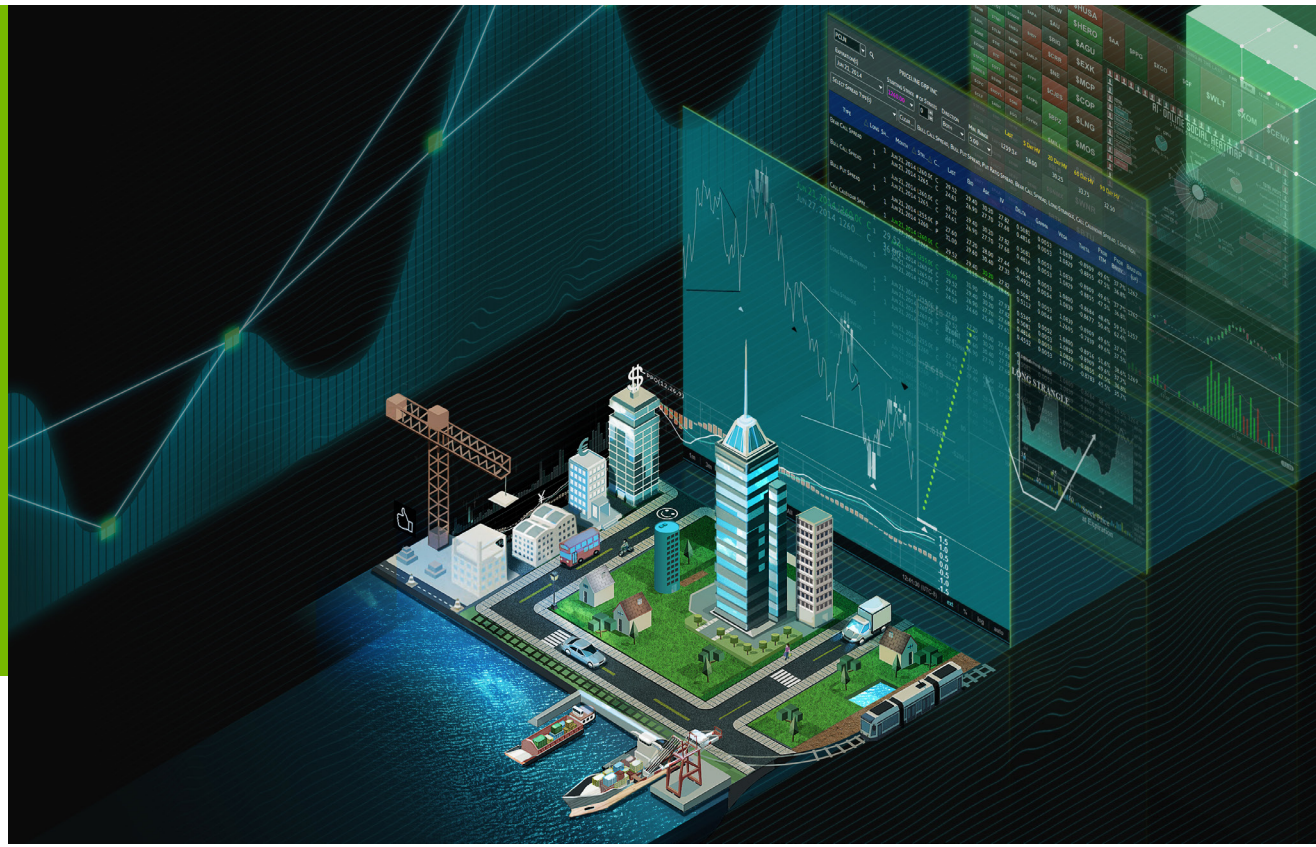
— Prashant Kukde, Associate Vice President, **RingCentral**



Scotiabank®

Scotiabank is using AI to develop more accurate scorecards that can determine whether they grant a loan to applicants. They 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. And by integrating **RAPIDS™**, part of the NVIDIA AI Enterprise, into their workflows, the bank is able to eliminate some manual work of tweaking and testing. For example, with GPU-accelerated hyperparameter tuning, a developer can let a computer test 100,000 model parameters while she is having her lunch. "In a way, the best thing we got from buying that system was all the support we got afterwards," said Scotiabank's director of data science and model innovation.

Read the Blog



4. Realize Your AI Vision With a Scalable, Purpose-Built AI Platform

Simplify and speed AI development with a comprehensive solution that eliminates the complexity of at-scale AI development.

The rise of generative AI has brought models that demand significantly more resources than traditional AI tasks. They require high computational power, extensive data, longer training times, substantial memory usage, and overall greater complexity. To unlock AI innovation and stay competitive, enterprises must future-proof their operations with a scalable, purpose-built platform that can grow alongside their business.

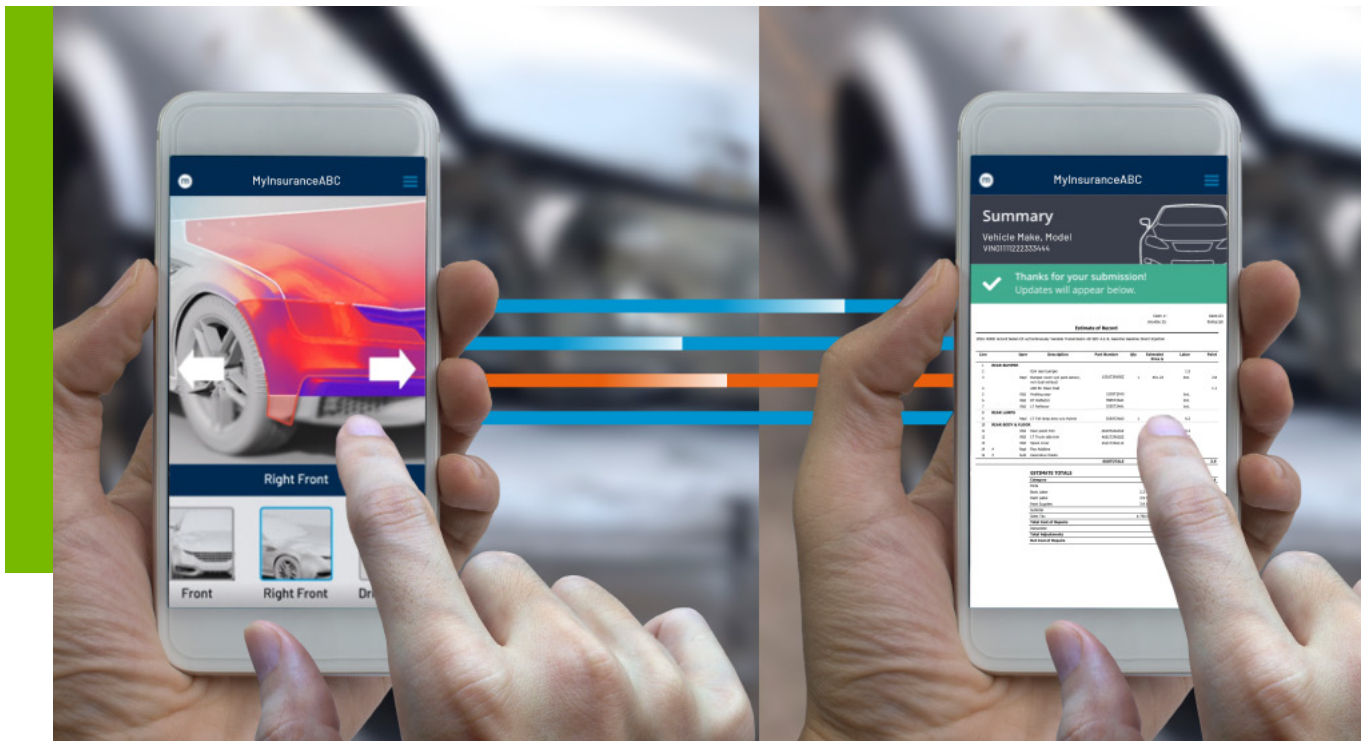
NVIDIA Base Command powers the DGX platform, enabling organizations to leverage the best of NVIDIA software innovation. Enterprises can unleash the full potential of their DGX infrastructure with a proven platform that includes enterprise-grade orchestration and cluster management, libraries that accelerate compute, storage, and network infrastructure, and an operating system optimized for AI workloads. This lets you easily scale as your requirements evolve.

Thousands of Leading Companies Deploy DGX Systems Today

- 10 of the Top 10 **Global Universities**
- 8 of the Top 10 **Global Telcos**
- 7 of the Top 10 **Global Pharmas**
- 10 of the Top 10 **U.S. Aerospace and Defense Companies**
- 10 of the Top 10 **U.S. Government Institutions**
- 8 of the Top 10 **Consumer Internet Companies**
- 9 of the Top 10 **Global Car Manufacturers**
- 6 of the Top 10 **U.S. Banks**

CCC Intelligent Solutions wanted to minimize low-value, high-volume, repetitive tasks in claims estimations. They established an end-to-end hybrid pipeline for training AI models using a DGX Cloud-based infrastructure and additional on-premises DGX systems. With NVIDIA Base Command integrated into their development pipeline for dataset management and orchestration, they experienced a **2X speedup** in running their data scientists' experiments. This AI pipeline has enabled CCC to unleash new innovations in the market, including their CCC Estimate-STP technology that provides line-level claim estimates in seconds based on insurer rules.

Read the Story



Unlock Turnkey Enterprise AI With Enterprise Services

NVIDIA's expertise in building AI infrastructure since 2016 is incorporated into the **NVIDIA DGX BasePOD™** reference architectures, which provide prescriptive, validated approaches for building and scaling AI infrastructure in an enterprise setting. Each reference architecture is tested at full scale and backed by industry leaders in storage and networking.

For enterprises that need an AI center of excellence on premises to complement their hybrid AI cloud, the **NVIDIA DGX SuperPOD™ Solution for Enterprise** is a proven platform that has enabled organizations around the globe to centralize people, processes, and infrastructure 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.

Built on a well-defined, long-standing methodology, from prestaging to predeployment simulations to quality assurance tracking, NVIDIA infrastructure specialists ensure customer success with full-team support—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 **NVIDIA experts** that support the full lifecycle, from design to deployment to operations to optimization.

Many businesses trust their mission-critical AI endeavors to the white-glove service and turnkey infrastructure experience provided by NVIDIA.



Clova

Naver, the leading search engine in Korea, and Line, Japan's top messaging service, created the AI technology brand Naver Clova. DGX SuperPOD is helping Naver Clova build and deploy state-of-the-art language models for new conversational AI services to enhance their chatbot and contact center solution. With NVIDIA's team providing onsite and remote support, from the physical cabling of 140 DGX systems to installing deployment and cluster management software, it took Naver Clova **only three months** from initial engagement to power on their DGX SuperPOD. It took **only one month** to go from an empty colocation data center to bringing the customer online. Naver Clova received support in three key areas:

- **Deployment:** NVIDIA helped Naver with installation of the physical hardware, operating systems, 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.

Learn More About the NVIDIA DGX SuperPOD Solution for Enterprise



The Right Formula for AI Success



Successful enterprises that 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. Adopt these learnings to uncover insights faster and ensure higher ROI for your AI projects sooner.

Ready to Get Started?

Learn more about NVIDIA DGX at **nvidia.com/DGX**

1. McKinsey & Company. The State of AI in 2022—and a Half Decade in Review. Operating earnings in the citation refer to earnings before income and taxes (EBIT). December 6, 2022.
2. IDC, Generative AI, Copilots, and AI Agents, doc # US52347824, June 2024.
3. Deloitte Insights. Tech Trends 2021.
4. 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 and 1x A100. Mixed precision. Batch size 32,768. DRLM trained with v20.03 and v20.07.
5. Deloitte's State of AI in the Enterprise, 5th Edition Report. Fueling the AI Transformation: Four Key Actions Powering Widespread Value From AI, Right Now. October 2022.

© 2024 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, Base Command, DGX, DGX BasePOD and DGX SuperPOD are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 3306943. JUL24

NVIDIA