Hewlett Packard Enterprise | NVIDIA

Technical white paper

Check if the document is available in the language of your choice.

ACCELERATING AI ADOPTION FOR THE MODERN DATA-DRIVEN ENTERPRISE

Leading-edge solutions and expertise to simplify the AI journey

CONTENTS

Intelligence is undeniably the most valuable currency in the global economy. As analytics become the driving force for business innovation, the adoption of artificial intelligence (AI) is escalating rapidly to pinpoint key patterns and insights and make predictions on endless streams of data. However, enterprises are struggling to harness the groundbreaking capabilities of AI, which leaves them searching for answers — how to get started, how to optimise their environments to handle demanding AI workloads, how to speed up production and how to derive valuable insights and outcomes.

Today's enterprises face numerous challenges in executing a successful AI strategy. Many enterprises lack the agility, flexibility and resilience that is mandatory for AI, as well as the budget to pursue rapid innovation. Infrastructure scalability and performance are critical to create the ideal AI environment. Without the proper infrastructure, enterprises are unable to satisfy the demands of AI and analytics or to implement larger-scale deployments, which leads to unreliable data access, increased time-to-production for AI models and slower, limited outcomes. Furthermore, enterprises often lack the necessary skills, talents and expert knowledge to integrate AI into core business practices in order to create a comprehensive, high-performing environment capable of cutting-edge AI analytics.

In this evolving technology landscape, enterprises are working to gain insights faster and more accurately than ever. In order to succeed, they must address the following components of AI adoption:

- Establishing a qualified AI architecture.
- Ensuring seamless scalability to handle diverse workloads.
- Ramping up the development cycle for faster time-to-insight.
- Accelerating AI applications to outpace the competition.
- Doing it all with better price performance.

The bottom line — enterprises of all sizes must innovate now to meet these requirements. Leading-edge solutions and expert guidance are essential to jump-start the AI journey and exploit the capabilities of AI analytics. Those who do will achieve significant competitive advantage in the race to intelligence.

MAPPING THE AI JOURNEY

According to the 2019 CIO Survey by Gartner, the number of enterprises investing in AI has grown 270% over the past four years. The rise of AI is fueling incredible growth across global markets. Yet it also poses major obstacles to enterprises that are struggling to transform their technology environments.

Strategic planning and deployment are key to operationalise the full power of AI. Understanding how to map your own AI journey is essential to selecting the right tools and technologies that are tailored to your unique requirements. First, enterprises must identify their biggest roadblocks to AI, along with the skills and resources that are required to overcome them. These insights provide crucial direction to design a framework comprised of the appropriate hardware and software solutions to meet their specific objectives.

The purpose of this journey is to optimise technology environments for faster AI deployment, build data foundations that can support multiple compute-intensive AI use cases and ensure greater data governance, quality, security and management. To achieve this, enterprises must invest in infrastructure that is extremely agile and flexible, while integrating and upgrading their current systems to ensure compatibility. Enterprises can accomplish these objectives by planning each stage of the AI journey, which allows teams to shift their focus onto areas where AI can extract maximum value.

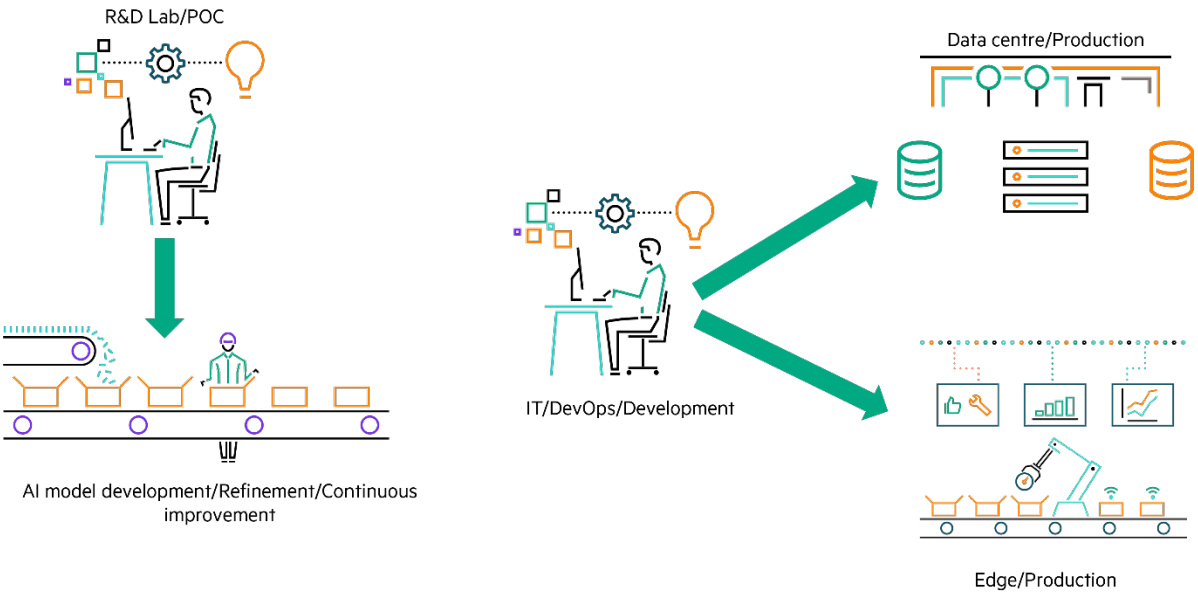

FIGURE 1. Critical phases of AI deployment

AI adoption is an enterprise-wide initiative. Mapping the AI journey is a strategic way to streamline and simplify AI usage from end to end — for data scientists monitoring sophisticated analytics programmes, to IT teams focussed on optimising vital resources for AI projects, to executives deriving business value from AI by leveraging data-driven outcomes.

HARNESSING THE FULL POWER OF AI CAPABILITIES

Training and inference are the cornerstones of machine learning and deep learning model development and use, both of which are compute-intensive and data-heavy workloads that require exceptional capacity, acceleration and durability. AI training is the crucial step to derive insights, where algorithms and models, using massive amounts of data, are created for a variety of applications. AI inference workloads use trained models to make predictions, forming the backbone of any AI-enabled application.

The new frontier of AI is driving a major paradigm shift in countless markets worldwide. Major industries such as Financial Services, Government, Energy, Retail, Healthcare, Media & Entertainment and Manufacturing are making AI a top priority. Enterprises across these industries are pursuing rampant automation and innovation, positioning themselves to lead the market with AI proficiency and revolutionary insights.

In the Financial Services space, AI applications are being deployed to proactively monitor insider trading, collusion and improper investment management practices. With the combination of vigorous compute platforms and analytics tools, financial institutions have the ability to train AI models with a constant stream of data inputs. Once validated, the models are used to perform inference in order to identify and stop suspicious activity in real-time. These pioneering capabilities are currently used to monitor hundreds of traders across email and IM, with voice monitoring applications coming in the near term. The next stage of innovation will enable monitoring thousands of traders around the globe in a dozen languages.

In Healthcare, enterprises are utilising machine learning and deep learning to usher in a new era of patient care. AI tools are processing troves of sensitive information, using the data to train sophisticated AI models to recognise patterns and abnormalities in treatment plans, medical imaging, patient care, medical histories, drug referrals and numerous other use cases. Once trained, these models are used to predict outcomes with extreme accuracy. This transformation is paving the way for personalised medicine in which healthcare institutions can treat patients remotely, perform autonomous robotic surgery and much more. Deep learning workloads are highly compute-intensive, so powerful technology tools and modern IT infrastructures are a pre-requisite for healthcare providers looking to augment their current capabilities with AI. These high-level techniques require not only extreme, massively parallel processing performance for training deep neural networks, but also cost-effective compute platforms for inference. This means that customers still operating with legacy IT infrastructures or limited technology capabilities are typically locked out from using AI and deep learning, and are left to continuously struggle with poor performance, low efficiency and slow time-to-value.

Retailers are also experiencing massive changes as AI capabilities enable the next wave of smart retail. With intelligent video analytics (IVA) and real-time customer and store analytics, companies are dramatically improving margins and enhancing the shopping experience. AI-powered solutions are empowering the retail industry with deeper insights into their operations, paving the way for game-changing innovations — from intelligent forecasting and inventory management, predicting customer preferences to deliver online and in-aisle promotions, to robotics and smart assistants freeing up human resources to focus on customer needs. The possibilities of AI are limitless.

LEADING-EDGE SOLUTIONS THAT SIMPLIFY AND ACCELERATE AI

HPE and NVIDIA® offer next-generation technologies and expert guidance to streamline the most complex and demanding phases of AI workflows.

Now, these technology leaders are helping enterprises realise the value of AI faster, with proven strategies to create new AI applications and achieve breakthrough results. With an industry-leading portfolio of high-performance computing (HPC) solutions, HPE provides robust compute capabilities and durability that AI demands. Backed by the processing power of NVIDIA graphics processing units (GPUs), enterprises can execute increasingly dense workloads at breakneck speeds to drastically improve time-to-insight. To save even more time and resources, NVIDIA GPU Cloud (NGC) software provides easy-to-deploy AI frameworks and HPC application containers, so users can focus on building their ideal use cases. NVIDIA NGC offers containers for GPU-optimised deep learning frameworks, machine learning algorithms and HPC applications, so users can focus on building their solutions.

To help enterprises make the most of their AI deployments, HPE has compiled solutions that are transforming AI training and inference workloads. These solutions are purpose-built for extreme HPC compute, and are NGC-ready, so enterprises can manage AI applications remotely and execute intensive workloads with the scalability of the cloud. This approach allows enterprises to accelerate solutions to business, gain deep insights from machine learning or artificial intelligence, power virtual environments and create seamless, scalable visualisations. The AI training solution is powered by the HPE Apollo 6500 Gen10 server which features eight NVIDIA V100 Tensor Core GPUs per server to accomplish quicker and more economical deep learning model training. The NVIDIA V100 is named 'The Most Advanced Data Centre GPU Ever Built,' with 32X faster training throughput than a computer processing unit (CPU).[1] Additionally, this solution contains NVIDIA NVLink™ GPU interconnect, which connects GPUs at up to 300 GB per second, making it one of the top computing servers in the world. As a result, AI models that would consume weeks of computing resources on previous systems can now be trained within a few days.

In order to address the full spectrum of needs in the market, HPE has developed a solution specifically engineered to execute AI inference. Powered by the HPE Apollo 2000 Gen10 with HPE ProLiant XL190r servers, users can deploy with confidence and gain unprecedented performance in AI inference. Each server utilises four NVIDIA T4 GPUs to enable faster time-to-insight for competitive advantage while also providing the scaling needed to add more resources as requirements change. Coupling compute power within the HPE Apollo 2000 Gen10 server with industry-leading NVIDIA GPUs provides better-together performance.

These comprehensive AI solutions for training and inference ensure resilience, cost-effective performance and usability, regardless of your AI expertise. For those needing additional guidance, experts with HPE Pointnext Services offer remote support as well as in-person trainings to effectively implement these solutions. NGC Support Services provide direct access to NVIDIA's subject matter experts so enterprise IT teams can deploy NGC-Ready systems faster and maximise user productivity.

AI TRAINING STARTER KIT

Built for machine learning/deep learning workloads delivering extreme compute for AI model training.

HPE Apollo 6500 Gen10 Server

NVIDIA V100 Tensor Core GPU

- NGC-Ready
- Apollo 6500 Gen10 Server
- Intel® Xeon® Gold 6230 (2.1 GHz/20-core/120W)
- NVIDIA V100 32GB SXM2 GPUs (orderable with four or eight GPUs)
- NVIDIA NVLink 2.0 GPU Cross-connect
- 768 GB RAM
- 8.64 TB Storage
- Two 10/40 GB Ethernet Ports
- InfiniBand EDR/Ethernet 100 GB Ports
- HPE Deployment and integration services for NVIDIA GPU Cloud
- Three-year Foundation Care Next Business Day
- HPE training credits

AI ENTRY STARTER KIT

Built for AI inferencing and light training workloads and optimised for high performance results.

HPE Apollo Gen10 Server

NVIDIA T4 GPU

- NGC-Ready
- Apollo 2000 Gen10 Server (XL190r)
- Intel Xeon Gold 6230 (2.1 GHz/20-core/125W)
- NVIDIA T4 16GB GPUs (orderable with four or eight GPUs)
- 192 GB RAM/384 GB RAM
- 4.8 TB TB Storage/384 GB Storage
- One 10/25 GB Ethernet Ports
- HPE Deployment and integration services for NVIDIA GPU cloud
- Three-year Foundation Care Next Business Day
- HPE training credits

FIGURE 2. AI solutions for training and inference

[1] ResNet-50 training, data set: ImageNet2012, BS=256 | NVIDIA V100 comparison: NVIDIA DGX-2™ server, 1x V100 SXM3-32GB, MXNet 1.5.1, container=19.11-py3, mixed precision, throughput: 1,525 images/sec | Intel® comparison: Supermicro SYS-1029GQ-TRT, 1 socket Intel Gold 6240 @ 2 GHz/3.9 Hz Turbo, TensorFlow 0.18, FP32 (only precision available), throughput: 48 images/sec

INCREASING YOUR COMPETITIVE EDGE WITH DEDICATED SUPPORT

Deployment and Integration Services from HPE Pointnext Services for NGC offers HPE server infrastructures that are purpose-built for AI workloads, incorporating NVIDIA GPU accelerators and container repository to deliver a comprehensive deployment stack. NGC is a container registry that promotes cloud-like agility in AI workloads, scalability and simplicity in the utilisation of GPU hardware on-premises. This offering enables GPU-accelerated cloud capabilities within those containers, as well as the option to deploy performance-optimised containers directly from NVIDIA's repository.

HPE Pointnext Services co-ordinates the installation, configuration and validation of your new GPU-enabled systems. This process begins with service planning, which is delivered remotely and provides you with a detailed plan and schedule for their NGC Deployment and Integration Service. During service planning, HPE Pointnext Services experts communicate the list of service activities and confirm that pre-deployment pre-requisites have been addressed. Experts are available throughout the service — remotely or on-site — to ensure a seamless deployment and integration process that helps you quickly leverage the benefits of your solution and accelerate time-to-value. This service is available for each HPE server configuration with three separation options to empower core, cluster and core-to-edge environments:

- Container platform installation and configuration.
- Container scheduler installation and configuration.
- Cross-environment integration.

Additionally, HPE gives you the option to buy enterprise-grade NGC software support, serviced by NVIDIA. This provides direct access to subject matter experts to resolve issues faster. And with the ability to easily submit service requests with a clear escalation path, you can deploy their systems with confidence.

HPE Pointnext Services for NGC drastically improves time to production, allowing you to quickly spin-up containers to accelerate the development cycle. These services complement your IT team with globally available advisory and professional service assistance from HPE Pointnext Services, and they can be easily attached to hardware offerings and delivered alongside your new infrastructure. Working with HPE data and AI experts, enterprises can take advantage of these innovations to gain a competitive edge in this dynamic open source AI ecosystem.

HPE AND NVIDIA ARE EXPANDING AI

HPE and NVIDIA are committed to addressing the challenges of AI adoption, equipping enterprises with innovations designed to meet their existing infrastructure standards and requirements in order to transform their technology environments easily and cost-effectively. With the development of two comprehensive and tightly integrated solutions, these top innovators are revolutionising AI capabilities for training and inference.

Designed for a range of AI workloads, HPE's best-in-class people, technology and partners can be leveraged at each stage of the AI journey, from building AI proof-of-concept projects to adding to existing AI compute resources. With our joint expertise and the introduction of these game-changing solutions, enterprises can achieve superior performance per dollar for GPU-intensive workloads, allowing them to swiftly extract and employ vital business insights without exhausting IT or budgetary resources. Together, HPE and NVIDIA are putting the infinite potential of AI at your fingertips.

Best people

Years of delivery expertise: global projects across multiple technologies and AI workloads

- Data scientists
- AI ambassadors/presales
- Solution architects
- Advisers/consultants
- AI centres of excellence/POC
- Data centre technologists
- AI benchmarking/engineers

Best technology

Integrated AI solutions leveraging HPE's comprehensive portfolio

- Servers, storage, software, networking, and services that work 'better together' for the full end-to-end workflow

Best partners

Proven global AI ecosystem, ideal for strategic planning/ performance optimisation

- ISVs
- System integrators
- Value-added resellers
- Service providers

FIGURE 3. HPE for your AI projects

CONCLUSION

Today's enterprises must find ways to leverage increasing speed, durability and intelligence to prepare for the next wave AI. To help them get started, HPE is delivering cutting-edge AI innovations that promise to simplify the AI journey. HPE's industry-leading AI solutions combine the performance of HPE compute with the acceleration of NVIDIA GPU technologies, backed by proactive support from experts around the globe to empower the modern data-driven enterprise. With personalised planning and turnkey deployment, enterprises are equipped to harness the full power of AI with ease, at optimal price performance. Together, HPE and NVIDIA are enabling enterprises to rapidly deploy and scale their AI frameworks to solve some of the world's greatest challenges.

Let us help you start your AI journey today.

LEARN MORE AT

hpe.com/uk/en/solutions/artificial-intelligence.html
hpe.com/uk/en/services/consulting/big-data.html
nvidia.com/en-us/deep-learning-ai/
nvidia.com/en-us/deep-learning-ai/solutions/

Follow us:
HPE HPC: @HPE_HPC
NVIDIA Data Centre: @NVIDIADC
NVIDIA AI: @NVIDIAAI

a50001405EEW, July 2020

Hewlett Packard Enterprise