White Paper

NetApp Solutions for AI Drive Business Outcomes Across Core and Cloud

Sponsored by: NetApp

Amita Potnis
November 2019

EXECUTIVE SUMMARY

Artificial intelligence (AI) is everywhere. Today, many organizations are looking to incorporate AI into their day-to-day business operations. IDC's *Enterprise Infrastructure Market Pulse: 3Q19 Market View — IT Infrastructure for Artificial Intelligence* (IDC #US45487919, September 2019) study indicates that 49% of respondents do not have an AI implementation today and are currently exploring AI initiatives or are actively planning AI initiatives in the next 12 months.

But what makes an AI initiative successful? A well-thought-out strategy is the mantra for a successful AI initiative. Organizations must note that any AI initiative built with specific business outcomes in mind, with data sets gathered from various sources and ingested and processed on reliable, flexible, and adaptable infrastructure that includes compute and storage, is bound to be successful. The AI infrastructure ecosystem includes a gamut of platforms and technology (servers, storage, and networking), component technologies (accelerators, processors, flash, etc.), and software as well as cloud-based services and converged infrastructure. The AI infrastructure and framework requirements will differ from one organization to another, and therefore, it is important to make intelligent knowledge-based choices when procuring infrastructure to drive the defined business outcomes. IDC recommends the following AI infrastructure considerations based on interviews with customers that have implemented infrastructure for AI initiatives:

- **Ease of use.** The deployment and management of AI infrastructure should be easy and user-friendly. Many organizations prefer to buy prepackaged infrastructure offerings that include compute, storage, and networking from trusted partners.
- **Support.** Organizations chose to procure prepackaged infrastructure to have a single point of contact for support for compute, storage, and networking.
- **Flexibility.** Because data driving AI initiatives is increasingly collected at the edge and processed at the core and in the cloud, it is important to consider an infrastructure supplier that offers a flexible edge-to-core-to-cloud capability.
- **Experience.** The experience and reputation of the AI infrastructure supplier (whether in compute, storage, or networking) in other markets or workloads is important because it lays the foundation for a solid offering.
- **Portfolio benefits.** An AI infrastructure supplier that offers supporting offerings for data tiering, caching, and so forth enables easy deployments with the peace of mind that the products are well integrated.

- **Partnerships.** Because every organization is different in its requirements based on the vertical the organization is in, any AI infrastructure supplier must have partnerships specific to a given workload for an enriching user experience and to drive business outcomes.

There are a few companies in the market today, such as NetApp, that offer these benefits. NetApp is developing an edge-to-core-to-cloud AI strategy that includes collecting and processing data at the edge, storing data in the core, and exploiting the AI tools of public cloud with the goal of simplifying and accelerating its customers' journey to AI.

IN THIS WHITE PAPER

In this white paper, IDC discusses the need for efficient and effective infrastructure supporting artificial intelligence workloads and the experiences of customers that deployed NetApp ONTAP AI.

SITUATION OVERVIEW

Artificial intelligence, machine learning (ML), and continual deep learning (DL) technologies are expected to drastically alter how enterprises and consumers operate on a day-to-day basis and gain insights. Data is now the basis of the new digital economy and is an invaluable asset for AI/ML/DL use cases. Successful organizations leverage AI/ML/DL to deliver meaningful insights and predictions to improve processes across industries and use cases.

IDC's *Artificial Intelligence Global Adoption Trends and Strategies Survey* indicates that nearly 50% of 2,473 respondents view AI as a priority and are executing on projects across business units within their organization. The proliferation of AI within organizations is expected to target IT automation, customer service and support, and fraud and risk management in the near future. AI is expected to create opportunities in several industries such as transportation through autonomous vehicles, life sciences through research projects, healthcare through AI-based clinical diagnosis, and government through AI-based surveillance. As AI makes its way into mainstream digital economy, many organizations find themselves in the initial proof-of-concept (POC) stage, with only a few in full production stage. IDC's *Artificial Intelligence Global Adoption Trends and Strategies Survey* indicates that 18% of organizations had AI models in production, 16% of organizations were in the proof-of-concept stage, and 15% of organizations were experimenting with AI.

Regardless of the stage where organizations are in adopting AI, building, testing, optimizing, training, inferencing, and maintaining accuracy of models is always top of mind. Any ML and DL algorithm needs huge quantities of training data, and AI effectiveness depends heavily on high-quality, diverse, and dynamic data inputs. Data management of these ever-growing data sets is complex and challenging, given that data is generated and ingested from various sensors and devices. In the age of digital economy, a key consideration is to manage the data from edge to core to cloud, analyze data in near real time, learn from the data, and then act on the data to affect outcomes. IoT, mobile devices, big data, machine learning, and cognitive/AI all combine to continually sense and collectively learn from an environment.

While AI bears the promise of social impact through examples such as autonomous cars or diagnosis or treatment of rare diseases based on genomic research, it faces several infrastructure challenges and questions around the role of compute and storage resources in an AI infrastructure solution. The reality is that the ideal AI infrastructure is an optimal combination of compute and storage. At a stage where AI is a priority for several organizations, one of the key requirements for AI infrastructure is the ability to deliver a complete pipeline for big data, machine learning, and deep learning and incorporates the widest variety of data sources on the premises, at the edge, and in the cloud. Such an infrastructure can efficiently and effectively support AI workloads.

The Importance of Efficient and Effective Infrastructure for AI

Generally, when thinking of AI, most organizations primarily focus on GPU-based parallel processing compute resources for training and inferencing. All-flash array (AFA) offerings are largely thought of as optimal storage tiers to support the massive parallelism of the GPUs. According to IDC's *Cognitive, ML, and AI Workloads Infrastructure Market Survey* conducted in January 2018 (n = 405, 1,000+ U.S. employees and 500+ Canadian employees), today, traditional SAN/NAS is largely used for on-premises run of AI/ML/DL workloads because of their existing deployment footprint and earlier stages of AI adoption. However, with the need to scale dynamically, store large volumes of data at relatively low cost, and support high performance, it is likely that a combination of public cloud, software-defined storage, hyperconverged infrastructure, and all-flash arrays with newer memory technologies will gain traction for data pipeline stages of AI.

Recent years have seen several solutions geared toward AI workloads that boast a well-balanced solution based on innovative storage architectures that provide performance in terms of availability, capacity, throughput, latency, and IOPS. Some of the primary requirements of any AI storage infrastructure are:

- **Scalability.** IDC's 2018 survey of 405 IT respondents and decision makers who had completed an AI project in North America indicates that massive data volumes and associated quality and management issues are key AI deployment challenges.
- **Cost efficiency.** IDC's November 2018 *Enterprise Infrastructure Market Pulse* indicates total cost of ownership and capital costs as top criteria for selecting storage systems. The scale of data generated by AI/ML workloads is making customers and vendors consider software-defined object-based storage as a viable storage alternative.
- **Parallel architecture.** IDC's November 2018 *Enterprise Infrastructure Market Pulse* indicates that performance is the top characteristic when considering storage systems offerings. Specific to AI infrastructure, GPU compute layer provides massive parallelism, essential for AI, ML, and DL workflows. However, unless the storage layer is able to match with similar parallelism, expensive GPU cycles get poorly utilized, delaying the experiment time to value. At petabyte scale, the storage layer is expected to deliver hundreds of gigabytes of throughput.
- **Data durability and locality.** The storage layer needs to complement the compute layer, by enabling data locality, as well as provide extreme durability at petabyte scale, not just within the datacenter but across geographies. Traditional data protection and durability schemes such as RAID and backups fall short at petabyte scale.

- **Reliability and availability.** IDC's 2018 *Server and Storage Infrastructure Availability Survey* indicates that 47% of 358 respondents agree that there is significant difference between the availability features on different storage platforms, and they use it for making enterprise storage choices. IDC's November 2018 *Enterprise Infrastructure Market Pulse* also indicates that reliability is a top characteristic when considering storage systems offerings.
- **Hybrid cloud/multicloud data management.** Data is increasingly distributed across on-premises, colocation, and public cloud environments. For this data visibility, access, control, and single-pane-of-glass management, tools across all deployment locations is a must.

Today, the competitive landscape has a handful of infrastructure solutions that offer an end-to-end data pipeline for AI workloads. NetApp's ONTAP AI and its surrounding portfolio offerings are worth due consideration.

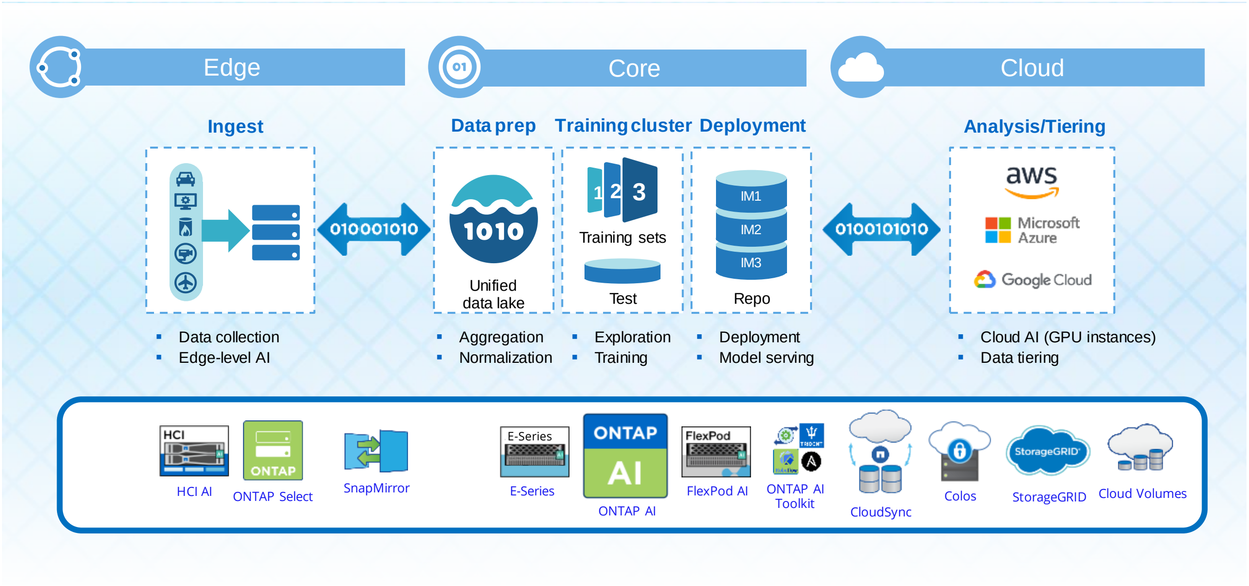
NETAPP SOLUTIONS FOR AI

NetApp's strategy is to make AI/DL data flow and management smooth, performant, and integrated from edge to core to cloud. By combining ONTAP data management/Data Fabric, AFF all-flash arrays, and Cloud Volumes, organizations can quickly deploy an optimized platform for AI/ML/DL workloads with high performance across the edge, core, and cloud.

The NetApp AI solutions are depicted in Figure 1.

FIGURE 1

NetApp AI Solutions



Source: NetApp, 2019

In detail:

- **NetApp ONTAP AI** is a fully tested and verified solution that is a combination of NVIDIA DGX-1 or DGX-2 supercomputers, NetApp AFF storage, and Cisco networking. NetApp's all-flash AFF (A200, A300, and A800) storage is available in various configurations to support smaller capacity needs to high-performance requirements.

  NetApp claims that this offering is designed to eliminate design complexities, enable independent scaling of compute and storage (can start small and scale seamlessly), and provide a range of storage options for various performance and cost point — all while managing data from edge to core to cloud. According to NetApp, the AFF A800 system has been verified with seven DGX-1 servers to demonstrate performance headroom to support two or more DGX-1 servers without affecting storage throughput or latency. For higher throughput and accelerated training and inferencing, the solution can be scaled to multiple racks by adding additional network switches and storage controller pairs to the ONTAP cluster. The offering supports the ability to scale compute and storage independently according to the size of the data lake, the deep learning models used, and the required performance metrics. The high-availability design is active-active, so maximum throughput can be sustained across all network connections in the absence of a failure.

  NetApp has partnered with ScaleMatrix, a provider of high-density colocation and high-performance cloud services with four datacenters in the United States. The offering makes NetApp ONTAP AI with DGX-2 servers in a modular, DDC (Dynamic Density Control) liquid air–cooled cabinets from ScaleMatrix. This partnership gives customers the option of datacenter services through a network of colocation partners specific for AI workloads. NetApp is the first vendor in the market to publish a reference architecture with NVIDIA DGX-2 and Cisco. NetApp's ONTAP AI has reference architecture for autonomous driving and healthcare workloads. NetApp has also published additional reference architectures with ecosystem partners — allegro.ai (deep learning), H20.ai (machine learning), Parabricks (genomics), OmniSci (GPU accelerated analytics), and SQream (GPU data warehouse).
- **FlexPod AI** is a NetApp and Cisco solution that combines NetApp's AFF all-flash arrays and Cisco UCS C480 ML M5 with NVIDIA GPUs or UCS C-series servers with Tesla V100 GPUs and is managed with ONTAP OnCommand and UCS Manager. The ONTAP software provides fast access to data sets and can scale the infrastructure based on needs with 20PB+ in the same namespace.
- **NetApp Cloud Volumes ONTAP and Cloud Volumes Services** are fully managed cloud services that enable you to move your mission-critical workloads and applications to the cloud and manage them with ease. NetApp Cloud Volumes Services is offered as a native cloud marketplace service in AWS and Google Cloud Platform, and as a first-party service called Azure NetApp Files, it is delivered and supported by Microsoft, built on NetApp technology in Microsoft Azure. NetApp Cloud Volumes ONTAP is geared to enterprise workloads, with its data protection and storage efficiency features, and is available in AWS, Azure, Google Cloud, and hybrid environments.
- **NetApp ONTAP Select** is a software-defined storage offering that can be installed on commodity hardware with integrated data protection, storage efficiency, unified access (SAN/NAS), and scalability features. NetApp ONTAP Select is targeted toward edge environments to enable data aggregation and advanced data management.
- **NetApp Private Storage (NPS) for Cloud** is a family of enterprise storage solutions (offered by ScaleMatrix and Equinox) that customers use for multiple industry-leading clouds and maintain complete control over data on dedicated storage systems from NetApp.

- **NetApp Fabric Pools** enables automated tiering of data to low-cost object storage tiers either on-premises or off-premises (including major object-based public cloud storage services and NetApp's StorageGRID).
- **NetApp StorageGRID** is an object-based storage offering and is built to target hybrid and multicloud environments by leveraging S3.
- **NetApp FlexCache** is a remote caching capability that is focused on simplified file distribution, reducing WAN latency and lowering WAN bandwidth costs. It enables distributed product development across multiple sites, as well as branch office access to corporate data sets.
- **NetApp FlexClone** enables customers to make fast, space-efficient copies of flexible volumes (FlexVol volumes) and LUNs while reducing costs and accelerates virtual machine and virtual desktop provisioning among many benefits.
- **NetApp Data Availability Services** is a backup and recovery offering that allows customers to replicate data securely to any ONTAP storage or in the cloud in a simple and easy-to-use manner.

NetApp's vast product portfolio enables customers to tailor their choices based on AI infrastructure requirements.

NETAPP ONTAP AI CUSTOMER INTERVIEWS

IDC interviewed two customers using NetApp ONTAP AI. In the sections that follow, excerpts from in-depth interviews provide a real-world view into AI infrastructure deployments, the reasoning behind infrastructure choices, and NetApp ONTAP AI's benefits and challenges.

Customer 1: Consultel Cloud

Consultel Cloud is a managed service provider serving the Australian market. The company believes its offerings allow customers an easy path to embrace cloud-based services. For its Australian and APAC customers, Consultel Cloud offers a variety of services that include virtual machines, migration of legacy workloads (email and exchange, backup and disaster recovery, etc.), and supporting infrastructure as a service (IaaS) for AI offering. Consultel Cloud is predominantly a 100% NetApp shop offering NetApp HCI as a service to its customers and thereby offering a path to cloud with flexible consumption models and at the same time tapping into emerging technologies.

Today, the company boasts hundreds of customers and a handful of them are already using its IaaS for AI. Consultel Cloud's overall client base includes 50% in the commercial space, 40% in government, and 10% in other sectors such as retail or research marketplace. Today, Consultel Cloud supports multi-petabytes AI-related data on its IaaS for AI. Consultel Cloud believes that the company's value for IaaS for AI comes from the ability to provide more compute power than hyperscale environments, no egress charges as well as management capabilities for hybrid cloud across on-premises and public cloud.

The company states that 20% of its overall infrastructure investments are specific to infrastructure deployed in support of AI and that this spend is expected to increase over the next 18 months.
The sections that follow discuss the scenarios leading up to the deployment of NetApp ONTAP AI solution and the achieved business outcomes.

Situation

The decision to target the AI market was based on internal assessment that indicated a strong need of such services in the Australian markets. As a part of its due diligence process, Consultel Cloud engaged with existing and potential customers and strategized a new IaaS offering for AI. The main objectives of bringing IaaS for AI to market was to position Consultel Cloud as a pioneer in this space in Australia and APAC and thereby increase revenue in a high-growth market. Consultel Cloud indicated that its AI customers use the infrastructure to drive customer/product insights, improved operational efficiency, employee productivity, and risk management. The specific AI infrastructure requirements were as follows:

- Consultel Cloud was keen on working with a trusted infrastructure provider. Given that the company had an existing relationship with NetApp having deployed its HCI offering, Consultel Cloud was comfortable with expanding this relationship for new services, including IaaS for AI by considering NetApp ONTAP AI.
- Consultel Cloud wanted to adopt the latest and most successful compute technology, and therefore any AI infrastructure solution that could be considered needed to include NVIDIA compute.
- A prepackaged solution geared toward AI was a must, given that Consultel Cloud needed to focus on expanding its IaaS for AI offering and having one point of contact for all things related to infrastructure was a necessity.

NetApp's ONTAP AI is a packaged offering that included NVIDIA DGX-2 compute servers, Cisco networking, and NetApp AFF all-flash storage that addressed Consultel Cloud's requirements. And thus, having made the decision to procure the NetApp ONTAP AI solution, the company began the deployment process.

Solution

In 2018, Consultel Cloud brought to market its IaaS for AI offering. The company's IaaS for AI is fully supported by NetApp's ONTAP AI solutions with NetApp's AFF800 all-flash arrays and NVIDIA's DGX-2 servers, thus offering its customers cloud-based compute and storage infrastructure services for AI. An existing relationship, a solution-based approach (prepackaged design), and partnership with NVIDIA were three things that were key for Consultel Cloud to adopt NetApp ONTAP AI. The company claims that edge computing in Australia is in its formative stages, and therefore, it has not yet adopted NetApp Select for edge deployments. However, it expects edge investments to increase in the next 12-18 months as potential customer requirements mention capturing and processing data at source instead of the cloud.

Business Outcomes

Consultel Cloud's deployment of NetApp ONTAP AI drives several important outcomes for the company and its customers.

Business outcomes for Consultel Cloud:

- First, with this offering – IaaS for AI – Consultel Cloud becomes the first provider of such service in APAC and therefore a pioneer in this space.
- Second, the company plans to grow its offering to support more customers and drive revenue.

For its customers, this service will drive three key business outcomes:

- It enables customers to automate business processes without the hassle of owning and managing infrastructure resources.
- Customers can gain cognitive insight through data analysis by taking advantage of highly performant infrastructure geared for AI workloads and reduce time to market.
- Users of this service can increase their engagement with customers and employees through AI efficiently and in a time-sensitive manner.

Customer 2: Cambridge Consultants

Cambridge Consultants is a leader in technology-based consulting, helping its clients transform their businesses.

Breakneck innovation makes it harder than ever for organizations to gain and maintain a competitive edge, which is why companies are turning to Cambridge Consultants.

To get ahead of the trends in AI, Cambridge Consultants built its own research lab and development facilities, known as the Digital Greenhouse. With this investment, the company is able to discover, develop, and test machine learning approaches in a fast, secure environment. Researchers are continually assessing algorithmic models developed in the Digital Greenhouse to understand how to solve their clients' challenges across diverse fields such as medical diagnostics, security, and industrial automation.

Recently Cambridge Consultants adopted NetApp AFF A300 all-flash storage system for its Digital Greenhouse ONTAP AI deployment, built on NVIDIA DGX supercomputers. This enhancement is helping Cambridge Consultants (and its customers) converge big data, machine learning, and compute capacity to advance the utility of AI.

Situation

The need for compute resources accelerated significantly in the past two years, which was the reason Cambridge Consultants chose the all-flash array (AFF300 100TB usable) from NetApp. NVIDIA DGX-1 and NVLink for data movement between GPUs have been very efficient compared with GPU-based white-box workstations. The specific AI infrastructure requirements were as follows:

- Performance was an important factor for both compute and storage resources as Cambridge Consultants was focused on quickly getting through jobs and delivering insights to customers.
- Cambridge Consultants also preferred a prepackaged solution-based approach that included compute, storage, and networking resources from trusted partners. This approach allowed Cambridge Consultants to focus on the services it was bringing to market with the comfort of one point of contact for all AI infrastructure resources.
- A flexible solution was of utmost importance. Because Cambridge Consultants needed to handle multiple file formats, the supporting AI infrastructure needed to be flexible.

Solution

Over the course of several years, Cambridge Consultants has established relationships with systems integrators for its infrastructure needs. These trusted systems integrators that had previous experience with NVIDIA technology made recommendations, which resulted in Cambridge Consultants deploying NetApp ONTAP AI for AI services.

In its deployment of NetApp ONTAP AI, Cambridge Consultants has currently deployed AFF A300 for storage with roughly 100TB usable capacity, NVIDIA DGX-1 servers, and Cisco networking. In addition, Cambridge Consultants have also deployed NetApp FlexCache and NetApp FlexClone in support of its AI infrastructure.

Business Outcomes

Cambridge Consultants is experiencing growth across its global businesses. The company now completes over 400 projects a year. It has offices across North America, Europe, and Asia, and 90% of its more than 850 staff members are expert engineers, designers, and scientists.

NETAPP ONTAP AI BENEFITS AND CHALLENGES

NetApp's vast product portfolio supporting AI infrastructure has several benefits that have been confirmed through customer interviews. The challenges NetApp's ONTAP AI and the rest of the portfolio face are mainly driven by market dynamics.

Benefits

- **Ease of use.** Customers interviewed for this research have indicated that having a prepackaged offering that includes compute, storage, and networking makes NetApp ONTAP AI a relatively easy-to-use offering. NetApp claims that simplified NetApp ONTAP AI configuration with Ansible can deploy a full stack in less than 20 minutes.
- **Support.** Customers indicate that NetApp's ONTAP AI offers support for compute, storage, and networking resources with a single point of contact and that NetApp's support staff have been keenly supportive of their needs.
- **Flexibility.** NetApp's extended portfolio supports an end-to-end flexible solution that addresses the needs of edge (NetApp ONTAP Select), core (NetApp ONTAP AI), and cloud (NetApp Cloud Volumes Services and NetApp Cloud ONTAP). Customers can thus take advantage of existing investments and future proof their infrastructure road map.
- **Experience.** NetApp has long-standing experience in storage, and over recent years, it has significantly expanded to address hybrid cloud. NetApp's ONTAP is a robust offering that has served several industries and use cases. This experience makes NetApp a trusted provider.
- **Portfolio benefits.** NetApp's product offerings in support of AI infrastructure include NetApp Fabric Pools (storage tiering), NetApp FlexClone (efficient data copying), NetApp FlexCache (caching capabilities), StorageGRID (low-cost object storage), and NetApp Private Storage (NPS offering colocated infrastructure), which allow customers to tailor a solution that best suits the requirements of the AI project.
- **Partnerships.** NetApp boasts an array of partnerships with ecosystem providers that are specifically geared for AI workloads such as genome sequencing and data warehousing.

Challenges

For Customers

- **One size does not fit all.** One of the challenges for Consultel Cloud is to efficiently identify its customers' usage patterns on NetApp ONTAP AI. Given that Consultel Cloud is a provider of IaaS for AI, the projects it serves for its customers almost always vary in requirements. Therefore, it is a constant challenge to cater to these ever-changing requirements, but having a solution that in itself is flexible has helped the company.
- **Lack of standardization.** A challenge that both customers faced is the lack of standardization (various data formats, file systems, ingestion methods, etc.) across projects. This challenge, as previously highlighted in the section, is driven by the market and does not reflect on NetApp ONTAP AI.

For NetApp

- A challenge particular to NetApp is that since AI is in its nascent adoption stages, it is imperative that the company sharpens its messaging across its product portfolio by suggesting specific solutions for certain verticals and workloads. IDC believes that this initial hand-holding will go a long way for NetApp, allowing the company to penetrate the AI market and establish itself as a leader in this space.

CONCLUSION

As AI becomes mainstream and organizations move AI initiatives to production, IDC recommends that organizations consider the following:

- **Business outcomes.** Monitor and prioritize project timelines and assess financial impact of AI initiatives to effectively deliver on specific goals.
- **Automation.** Leverage automation software to allow data scientists and IT to focus on important tasks at hand.
- **Architecture.** Choose infrastructure that will enable an end-to-end AI solution that includes edge to core to cloud. This infrastructure should be easy to configure and maintain, self-healing, and integrated with predictive analytics tools for valuable insights.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com