White Paper

End-to-End AI Is Within Reach

Sponsored by: Dell Technologies

Peter Rutten Eric Burgener
September 2021

IDC OPINION

Over 90% of enterprises have already embarked on digital transformation — the transition to more data-centric business models. While IT organizations will maintain a number of legacy workloads when they make this transition, they will also be adding many more next-generation applications that are being developed and deployed specifically to meet the requirements of the new digital era. Big data analytics applications leveraging artificial intelligence (AI) will drive better business insights, fueled by the massive amounts of data that enterprises will be collecting from their products and services, employees, internal operations, and partners going forward. The massive amounts of data required, delivered with an increasingly real-time orientation, demand performance, availability, and scalability that legacy information technology (IT) infrastructure will be hard pressed to meet.

AI is made up of a number of different workloads, each of which generates a different I/O profile and has different storage requirements. To make the most effective use of AI-driven big data analytics, enterprises will need to create an "end to end" AI strategy that is well integrated across three different deployment models — from edge to core datacenter to cloud. Because of the many new requirements of this hybrid, multicloud strategy, almost 70% of IT organizations will be modernizing their IT infrastructure over the next two years. IDC has released the "artificial intelligence plane" model to help customers better understand how to create the right ecosystem to maximize the contribution AI-driven workloads deliver. The underlying storage infrastructure is a key component in that model, and it is already clear from end-user experiences over the past several years that legacy architectures will generally not provide the right foundation for long-term AI success.

To help its customers succeed with AI, Dell Technologies has put together Dell Technologies Validated Designs for AI. These engineering-validated stacks make it easy for enterprises to buy, deploy, and manage successful AI projects, offering not only the underlying IT infrastructure but also the expertise to create optimized solutions that drive real business value. With its broad IT infrastructure portfolio, including compute, storage and networking resources, and AI ecosystem partnerships, the vendor can bring the right resources together with an end-to-end AI focus that drives competitive differentiation for its customers.

SITUATION OVERVIEW

As enterprises begin to appreciate the value of AI to improve their products, customer experiences, business operations, and IT infrastructure (among many other possible AI use cases), the question they ask themselves is: *How and where do we develop and deploy our AI workloads?* "How" refers to

the infrastructure stack necessary to meet business requirements, while "where" concerns the best deployment scenario — edge, core, or cloud. "Develop" means training the AI models, while "deploy" refers to taking the AI models into production by inferencing on them as part of an application.

IDC defines an AI workload as a set of applications along with their primary and secondary data sets that can be categorized as follows:

- AI software and platforms — where the AI model training takes place
- AI applications — applications that perform AI inferencing as their primary function
- AI-enabled applications — applications that perform AI inferencing as a secondary function

Each of these AI workloads poses a different set of requirements and challenges to the enterprise.

Compute Requirements

Different AI use cases require different compute approaches. For example, both image recognition and natural language processing (NLP) are very compute intensive — both when training the AI model and when putting it into production. A compute approach that is gaining a lot of traction is the convergence of three workloads — data analytics, AI, and modeling and simulation — onto one infrastructure design, which IDC calls performance-intensive computing (PIC). PIC borrows many elements from what is generally known as high-performance computing (HPC), but it is more focused on the fact that in the case of AI, it is necessary to partition the workloads into many smaller chunks and distribute them in a parallelized fashion across compute resources within the server, between servers in the form of clusters, and between clusters across datacenters or clouds.

There are roughly five categories of compute for running AI: workstations, servers (scale-up systems, converged systems, and hyperconverged systems), clusters, supercomputers, and quantum computing. Except for quantum computing (which we will exclude from here due to its nascency), these systems can be accelerated with graphic processing units (GPUs), intelligence processing units (IPUs), field-programmable gate arrays (FPGAs), or application-specific integrated circuits (ASICs) that deliver in-processor parallelization with thousands of cores, or they can be solely dependent on the host CPUs. While the vast majority of data processing today takes place via server CPUs, there are opportunities to offload some tasks to data processing units (DPUs) or SmartNICs as well.

AI workload performance depends on how the application is written (e.g., whether it's written to take advantage of parallel processing with CPUs and accelerators, written to load images for recognition into the server memory and/or into the accelerator memory, and/or written to offload some tasks from CPUs). AI workload performance can also be significantly impacted by server and accelerator memory. For example, many applications attempt to load the entire image into GPU memory for processing and recognition, severely slowing performance. For those looking to optimize price/performance, MLCommons (a consortium of 50+ founding members and affiliates, including start-ups, leading companies, academics, and nonprofits from around the globe) publishes AI benchmarks for many AI training and inferencing workloads, sharing performance for a wide variety of technology combinations.

The server accelerators may require special interconnects between them and the host CPU to provide sufficient bandwidth to the system memory (e.g., PCIe, NVLink, Infinity Fabric, or CXL). There are also ways to bypass server CPUs for faster communication directly to server components such as accelerators. When there is more than one server, high-speed networking is the unsung hero as Ethernet or InfiniBand switches make it possible to transfer data between servers and external

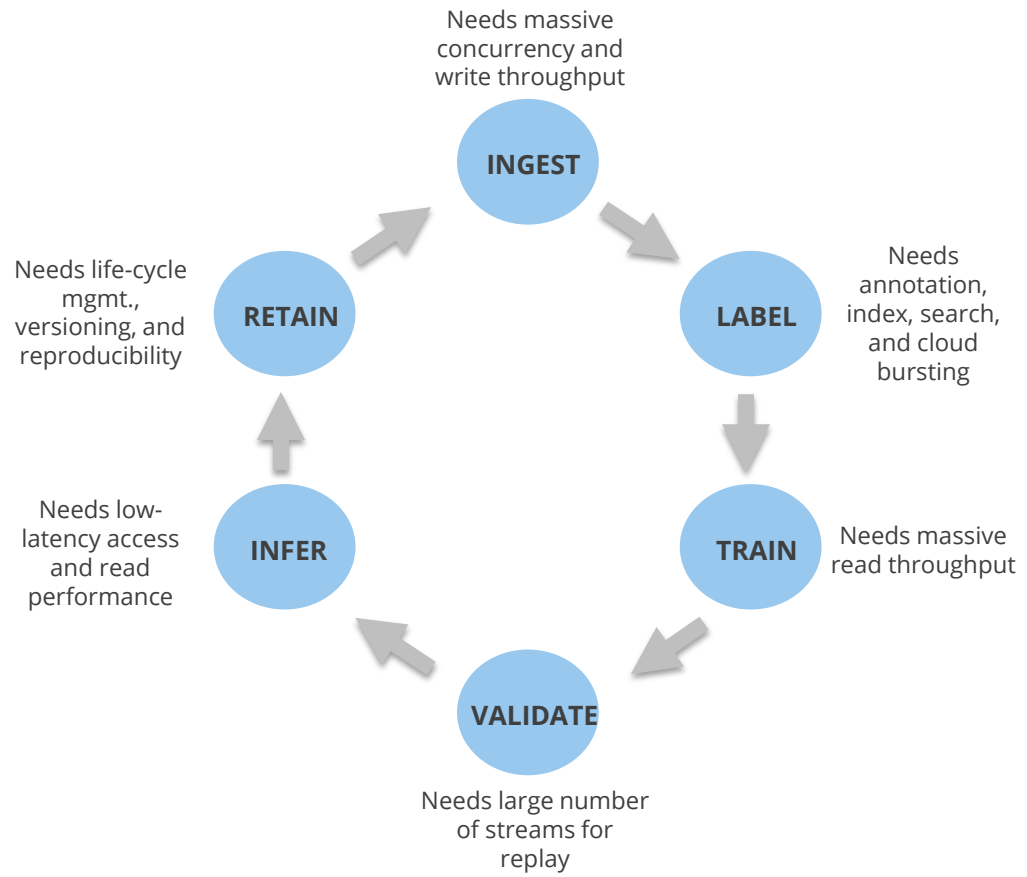
storage. It's also worth evaluating the distance between the data and the processing, along with the number of hops or transfers required to get the data to the processing. When AI speed is essential, some set up a data caching system, also called the data accelerator, with servers and direct attached storage to deliver data faster from storage to compute for processing. Today, there are fast storage servers with two high-bin CPUs and up to 100 hard drives, and there is fast data storage with higher-than-just-controller processors, blurring the lines between compute and storage.

Data Persistence Requirements

Storage requirements vary depending upon the type of AI workload. Figure 1 provides a quick summary of the key storage requirements based on the stage in the AI data pipeline. Each stage in the pipeline has a different I/O profile, and this wide range of storage requirements may drive the need for a different storage system or "silo" for each stage. A more flexible storage platform that can accommodate a wider range of I/O profiles may allow enterprises to consolidate storage workloads in the AI data pipeline on fewer systems — a capability that minimizes data movement requirements and can save considerable costs in both infrastructure and administration. The reader should note that the majority of data that will be driving AI workloads over the next five years is unstructured, so storage platforms that support file- and/or object-based storage are most relevant.

FIGURE 1

Stages in the AI Data Pipeline



Source: IDC, 2021

IDC has published research that provides considerable technical detail on how to select the right storage infrastructure for AI workloads. This research reveals that enterprises will need to consider five areas to make the best decision for their needs: performance, availability and resiliency, flexibility, ease of use and management, and infrastructure efficiency and cost. Flexibility may be the cornerstone around which each of these areas are evaluated. For example, under performance, IT decision makers need multiple media options to be able to cost effectively configure platforms for low latency, high bandwidth, high-capacity density, and/or an ability to handle more metadata-intensive workloads. Vendors that offer NVMe, SAS, solid state, and spinning disk media options give enterprises the flexibility to tailor the configuration to their particular I/O requirements. To meet high-bandwidth requirements, scale-out storage clusters running a truly distributed software platform where bandwidth within a single unified namespace can be scaled as nodes are added may provide a more cost-effective solution than scale-up storage appliances.

For extremely performance-sensitive workloads, enterprises will want to evaluate technologies that are particularly focused on delivering low latency. These include support for NVIDIA's GPUDirect Storage

(a published API that enables persistent storage to talk directly to accelerated compute that uses the vendor's GPUs), NVMe over Fabrics (which can deliver latencies similar to local disk for networked storage), and storage-class memory (which can drop storage device latencies to under 40 microsecond).

For availability and resiliency, the ability to support a variety of enterprise-class data services that enable administrators to balance latency, resiliency, and capacity overhead for data protection and recovery can result in real cost savings. Features to look for include host multipathing (with transparent failover), multiple RAID and/or erasure coding options, snapshots, replication, and underlying technologies (like NVMe) that can speed recovery and rebuild times. Scale-out storage clusters can offer nondisruptive expansion as well as multigenerational technology upgrades — a capability that allows systems to hit high overall availability requirements and can extend the storage life cycle well beyond the typical five years to drive additional cost savings.

In the area of flexibility, support for multiple access methods to the same underlying data store can have a strong simplifying effect on data management in AI workloads. To enable easy use by file system–based workloads, access methods like Network File System (NFS), SMB, FTP, and FUSE may be most important, enabling different workloads to effectively access the same data sets without having to perform any data migration. Less latency-sensitive, batch-oriented analytics can benefit from HDFS access. To enable easy use by object-based workloads and enable easy data movement to geodistributed and/or public cloud–based targets, access methods and APIs are important. Enterprises may want to understand how APIs are supported (look for features like multipart upload, versioning, cross-region replication, bucket life-cycle management, and object locking [to provide immutability]).

The range of scalability is an important consideration when buying storage for AI workloads. If an AI application is successful, it is likely to grow at a rapid rate. Administrators will need to carefully evaluate their growth potential to ensure that they will not outgrow systems for these workloads too quickly.

Enterprises will also need to consider their deployment strategies. Will workloads be deployed using bare metal, virtual machines (VMs), or containers, or some mix of the three? If containers will be used, enterprises should ensure they have storage systems that can deliver persistent storage in these environments that meet defined service-level agreements for performance and/or availability. This generally will require a third-party tool (e.g., Rancher and Robin) that supports Kubernetes, the de facto container orchestration platform standard. Enterprises need to ensure the storage they purchase can support their planned deployment strategies.

As enterprises craft their AI strategy, they will be considering deployment options at the edge, in the core (datacenter), and in the public cloud. Storage platforms that offer edge, core, and cloud-based deployment options can enable enterprises to build a more cohesive, well-integrated infrastructure that supports a common management paradigm across locations.

For a more in-depth technical discussion of what to look for in a storage platform for AI workloads, see *How to Evaluate Different Storage Options for Artificial Intelligence Workloads* (IDC #US47644021, May 2021).

Development and Deployment Locations

As to location, the obvious options are edge, core (datacenter), cloud services provider (SP), and managed services provider. However, given the increasingly distributed nature of AI, it is safe to assume that it is best to move compute closer to where the data is being sourced or generated versus the other way around. In recent times, the term *end-to-end AI* has become a common way of describing AI that is developed and deployed across edge, core, and cloud, whereby each location has a different workload profile and requires different infrastructure.

Three Dimensions of End-to-End AI Infrastructure

Infrastructure requirements for AI can be considered from three angles:

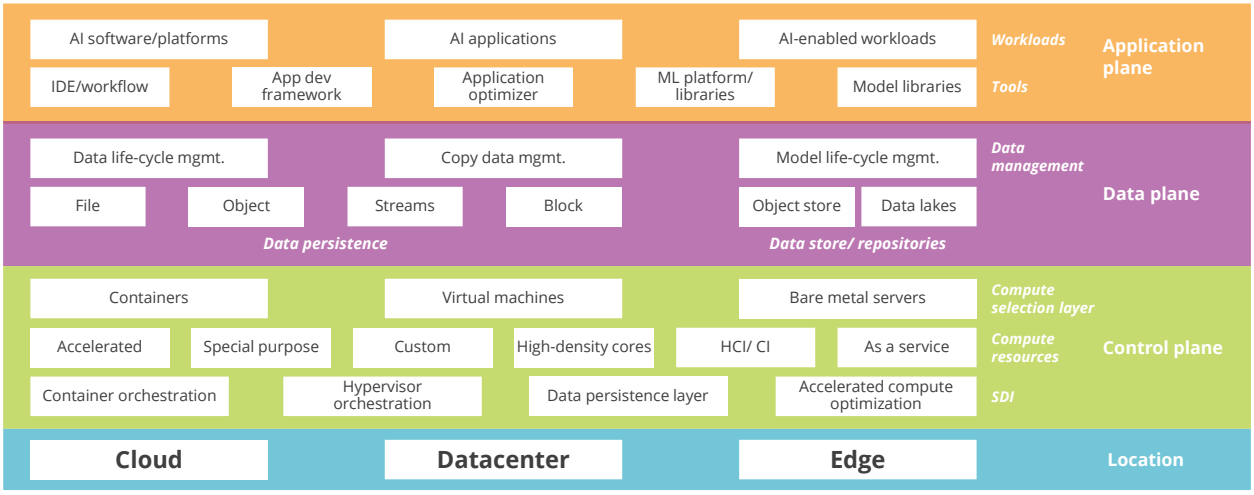
- **Scale.** The scale dimension describes the scale at which the workload operates. Hardware — compute, networking, and data persistence (storage) — plays a crucial role here, but software, such as virtualization and orchestration, is becoming just as important for managing ever larger and more complex AI models.
- **Portability.** Portability refers to the ability of the workload to be moved across edge, core, and cloud locations. Today, many AI workloads are still fairly static in nature (i.e., designed to run in a single deployment), but increasingly, businesses are looking to develop workloads using one deployment model (e.g., public cloud) and install them in production in another one (e.g., edge or core). As AI workloads grow, the best deployment model may change, and enterprises may choose to move them to a different one.
- **Time.** This relates to the time continuity of the workload itself. AI workloads are increasingly designed to analyze streaming data in a real-time or near-real-time manner (rather than operating in batch mode). Often, an AI application that may interact with an external client or user actually manages a number of separate but related internal workflows, which may execute independent AI algorithms as part of that interaction, all of which must occur in real time. An example of this might be an AI-driven online mortgage application that sequentially uses different AI algorithms to verify an applicant's identity, compile a relevant credit history, request additional information, interact through a natural language processing interface, and ultimately make a loan approval decision. This approach is referred to by IDC as the "AI hop effect" since the controlling AI application is essentially hopping between potentially different systems with the lowest possible latency.

The AI Plane

To address these AI requirements and dimensions, server OEMs, storage OEMs, and cloud service providers have architected ever more sophisticated AI infrastructure stacks. IDC has studied these various efforts and synthesized a formalized AI infrastructure plane from them, also referred to as the "AI plane" ("AIP"). The AIP can serve as a framework for organizations to assess the completeness of their AI environment based on commercial and/or open source components, particularly if they aim to deploy an end-to-end environment for AI. Figure 2 illustrates a generic AIP consisting of three layers — the application plane, the data plane, and the control plane — that are further discussed in the sections that follow.

FIGURE 2

The AI Plane



Source: IDC, 2021

The Application Plane

We have already described the top layer – the AI workloads – of the application plane previously in this white paper; below this AI workloads layer are the tools, which include:

- IDE/workflow tools for easier code and model development (a popular workflow tool is Jupyter)
- Application deployment frameworks for building AI applications, including libraries, SDKs, or reference code
- Application optimizer tools such as KubeFlow that help deploy, scale, and manage AI models
- Model libraries that serve as a marketplace for prebuilt ML models (e.g., NVIDIA GPU Cloud [NGC] or Google ML Marketplace)

The Data Plane

This is the infrastructure layer that provides storage resources to enable the AI pipeline. It consists of data management, persistent data access, and data stores/repositories.

Data Management

- Data life-cycle management tools and services for moving data sets in and out of the compute layer and for protecting and/or archiving data either on premises or off premises
- Copy data management (CDM) tools and services that reduce storage by eliminating duplication of production data, creating usable replicas without increasing the storage footprint
- Model data life-cycle management tools and services that manage the life cycle of the data used by AI models through the stages of data ingestion, sanitization, training, testing, and inferencing

Persistent Data Access

Another portion of the AI data plane is persistent data access, which can be achieved with:

- File access via file protocols such as NFS or SMB
- Object access via RESTful APIs such as S3 or Swift so that data and metadata can be accessed via HTTP/HTTPS
- Streaming data that is continuously generated by data sources such as sensors and that various streaming platforms can publish, subscribe to, store, or process in near real time

Data Stores/Repositories

And the AI data plane includes data stores and repositories:

- File- and/or object-based data stores, which are scalable repositories of unstructured data such as images, videos, and audio clips
- Data lakes, which are scalable repositories of both structured and unstructured data

The Control Plane

The control plane refers to the compute resources that enable the AI pipeline; it includes computing platform selection, compute resources, and software-defined infrastructure.

Computing Platform Selection

The virtualization choices for AI compute platforms are:

- Containers, which because of their scale, size, and portability, are increasingly being used for AI models and which – via Kubernetes – can be used across a heterogeneous AI pipeline with consistency and reusability
- Virtual machines, which have become more applicable for AI as GPUs can be virtualized and managed by the same hypervisor
- Bare metal, which is often used for workloads that need optimized performance without performance overhead from the virtualization layer

Compute Resources

While each phase of the AI pipeline requires some type of performance-intensive compute, AI model training is especially demanding due to the large amount of parallelism involved. There are various types of compute resources that are suitable for the different AI pipeline stages:

- General-purpose compute CPUs, which theoretically can be used to enable any phase of the AI pipeline but are not efficient for AI training and are generally only used for light AI inferencing
- Accelerated compute, which includes graphic processing units, field-programmable gate arrays, intelligence processing units, and application-specific integrated circuits (GPUs are particularly suitable for AI model training, thanks to their highly parallel structure, while lighter GPUs can be used for AI inferencing – they can be deployed in workstations and servers. FPGAs are suitable for inferencing at scale, and several vendors are developing AI-specific ASICs to support either AI training or AI inferencing.)
- Infrastructure as a service from cloud service providers enabling as-a-service availability of VMs, containers, GPUs, IPUs, FPGAs, and Google's ASIC, the TPU, allowing for pay-as-you-go subscription models

- Special-purpose compute with high-density cores and massively parallel computation resources, which can be an economical alternative to accelerated hardware

Software-Defined Infrastructure

Compute resources are becoming increasingly software defined, including for AI; for example:

- Converged infrastructure (CI) and hyperconverged infrastructure (HCI) platforms that provide a simplified and cost-effective infrastructure stack
- Disaggregated compute, storage, and networking resources connected via fast interconnects
- Reference architectures and/or converged infrastructure offerings that make it easier to buy and deploy AI infrastructure that is pre-validated by vendors
- Container orchestration platforms that provide a scalable, consistent, and interoperable environment to run, deploy, and manage containers with AI applications
- Hypervisors that provide the ability to abstract physical servers (including GPU-accelerated servers) into VMs through CPU and GPU virtualization
- A data persistence layer that enables access to persistent storage through the construct of logical volumes
- An accelerated compute optimization layer that provides platforms, tools, and abstractions to leverage accelerated hardware (e.g., NVIDIA CUDA)

The Location

In general, with developing, training, testing, and then deploying AI applications, data adjacency is critical as it reduces latency. Therefore, the public cloud is useful for AI or AI-enabled applications that operate on data residing in the public cloud or in a managed service provider, and an on-premises deployment, whether on servers or workstations, is useful for AI or AI-enabled applications that process data that reside on premises.

With regard to the public cloud, multicloud is a common strategy, wherein data is directly connected to multiple clouds to take advantage of cloud bursting for compute when necessary (model training, taking advantage of compute across clouds in an on-demand basis) and best-of-breed AI technologies across multiple cloud platforms at different steps of the AI workflow (e.g., taking advantage of TensorFlow from Google Cloud and SageMaker from AWS).

The need for data adjacency also extends to edge deployments of AI models, which should be close to where data is actually captured. IDC defines "edge" as the space between data collecting endpoints and the core and cloud locations. Edge is highly suitable for data ingestion and AI model inferencing, although some limited unsupervised AI training is also possible at the edge. Because of the various constraints that the edge poses for infrastructure, the opportunity to run a complete AIP stack at edge locations will be somewhat limited. Many vendors do, however, create scaled-down versions of their core infrastructure designed to fit the performance, cost, and ease-of-management requirements of edge deployments.

End-to-End AI

Because of the increasing maturity of the AI plane for all locations (edge, core, and cloud), there is an emerging opportunity to not only develop and deploy AI near the data but also connect these locations with low-latency, high-bandwidth networks and only move data that is used for reads/writes by an AI model across these networks. This way, the bulk of the data can always remain in place, regardless of

where it is being processed. The benefits of leaving data in place are substantial. Data movement is minimized, reducing infrastructure, administrative, network bandwidth, and potential egress costs (when public clouds are in use). The time to move through the stages of the AI data pipeline to get to better business decisions is shorter. And the data security and/or governance implications of moving the data are minimized or entirely avoided.

DELL TECHNOLOGIES' AI SOLUTIONS

Dell Technologies is positioning itself as not just a server, storage, and workstation company but as a vendor that can support businesses' end-to-end AI needs from proof of concept (POC) all the way to large-scale production. Dell takes the point of view that within this trajectory from POC to production, many hurdles await businesses, specifically with determining what the right infrastructure is for each workload and, to be more precise, for each stage of a given workload. For example, the company does not necessarily subscribe to the notion that every single AI task requires high-end processors, one or more coprocessors, and a large amount of memory. Whether these additional components are really needed depends on the workload.

There's a distinct difference between AI training and AI inferencing, of course, but there are also AI workloads that inference quite well on just general-purpose CPUs, whereas others execute better with a combination of CPUs and certain types of GPUs or on CPUs and specifically programmed and optimized FPGAs. Ultimately, what businesses need is an ability to find the right infrastructure match for their AI workload, which requires access to an extensive set of test data. Dell has conducted AI workload benchmarking on more than 100 different hardware configurations that businesses can leverage to determine whether, for example, a workload would inference better on a light GPU or on an FPGA.

Dell Technologies markets a range of systems for every AI scenario, allowing businesses to grow their capabilities at their own pace as their needs shift and as their data sets grow. Dell's Precision workstations for AI, for example, are designed with powerful host processors. Dell EMC PowerEdge servers can be equipped with as many as 16 GPUs. Dell also has a wide range of PowerEdge servers for AI, with a variety of configurations that include multiple coprocessors and can be clustered for any production size. The company also offers "Dell Technologies Validated Designs"-certified reference systems for AI. And for businesses that need AI supercomputing capacity, there is the Dell Zenith supercomputer with more than 27,000 cores and support for up to 10PB of storage. Deployment scenarios with Dell Technologies solutions include datacenter, edge, cloud, and multicloud, with the compute brought to the data rather than the other way around.

From POC to Production

Every AI initiative begins with a business problem (or opportunity). Data scientists convert a use case into a data science problem and then develop a data science solution, at which point IT takes over to make the solution production ready. In all likelihood, the data scientists are working with very large data sets needed for training the AI model. This poses an immediate challenge as large volumes of data need to be moved in and out of various systems while complying with the organization's data governance. Businesses need to determine where the data resides, how they get an application from one place to another, and how to ensure compliance.

Dell Technologies offers an AI starter bundle that consists of a Dell Precision 7920 Data Science Workstation (tower or rackmount) and the Dell EMC PowerScale scale-out NAS storage platform,

allowing data scientists to start small and build out a data lake as large as needed as their data volumes grow while keeping that data in place. Data scientists have shared access to these large data sets without the need for time-consuming and often expensive data migration. They can use various protocols simultaneously, including Hadoop Distributed File System (HDFS) and Network File System, and run analytics and AI applications.

Dell EMC PowerScale is an unstructured data storage platform that can start with a cost-effective 3-node cluster for edge and distributed locations and scale all the way up to a 252-node cluster that provides up to 15.75M IOPS, 945GBps of aggregate throughput, and over 30PB of raw storage capacity. The system supports significant software flexibility, multiple file- and object-based storage access methods, and NVMe-based all-flash as well as SAS-based hybrid and archive nodes that allow customers to effectively configure performance, capacity, and low cost when and where they need it for their AI workloads. Throughput is important because the Dell Precision 7920 Data Science Workstation can be outfitted with as many as 4 GPUs, which means that the I/O processes on the PowerScale become critical for overall performance. For AI, storage can sometimes become a bottleneck when reads/writes to the storage device cannot keep up with what the GPUs are executing. Data governance is included in this package.

Moving on to large-scale production, with the data remaining in place, businesses can grow into Dell Technologies' Validated Designs for AI, which include the Dell EMC PowerEdge platform for compute, the Dell EMC PowerScale platform for storage, and the Dell EMC PowerSwitch switches with InfiniBand for scaling.

Dell Technologies' End-to-End Solutions

AI data is typically generated across various locations. A business may have points of sale (POSs), for example, where data captured locally is used for inferencing. Businesses also want to take advantage of their hybrid cloud leveraging AI development and deployment services in public clouds (e.g., AWS SageMaker or Google Cloud AI platform with TensorFlow). The question is how to take advantage of such AI services without moving data in and out of the cloud service providers' datacenters.

Dell Technologies enables this by offering multicloud services with Dell PowerScale storage provided by a managed services provider that is connected to the public cloud SPs, including AWS, Google Cloud Platform (GCP), Oracle Cloud, and Microsoft Azure. Data remains in place in the Dell EMC PowerScale-based data lake on the managed SP, and the managed SP provides a high-speed direct connect to the cloud SPs (essentially a fat pipe), allowing AI initiatives to scale across various cloud-based analytics and AI compute services. Only read/write data moves in and out of the cloud SPs with a latency that is comparable with native compute storage networks within a cloud. With two of these cloud SPs, Oracle and Microsoft Azure, there are no egress fees for moving the data in and out, regardless of whether the data is in a customer's on-premises Dell EMC PowerScale deployment or on the managed SP.

These services from Dell Technologies are powered by multicloud services provider Faction with whom Dell has partnered for this specific purpose. Faction has datacenters in multiple locations close to the datacenters of the major cloud services providers, and businesses simply purchase the capacity that they need in the location of their preference. This partnership significantly expands Dell's storage offerings for AI. Businesses can have their data on premises (under an outright purchase or a subscription model), move their data into the managed SP from on premises using data replication, spin up a disaster recovery center in the cloud, or take advantage of any of a number of public cloud-based services.

The advantages for businesses are that they have easy access to high-performance, low-latency, scale-out storage where the data resides, combined with data management and data governance. Dell Technologies states that for AI training, where fast reads/writes are required, the network can meet these latency requirements. Businesses can start small and grow their environment to petabytes or even exabytes. Furthermore, the data is available for cloud use cases, allowing businesses to take advantage of the myriad of AI cloud services that the cloud SPs offer. And, just as critical, businesses are avoiding the cost of moving data in and out of the cloud SPs. Data generated at the edge, for example, can be extremely costly to move into a cloud SP.

Indeed, Dell Technologies is also an Internet of Things (IoT) provider enabling data to move from, for example, sensors to a Dell EMC edge server — such as the Dell EMC PowerEdge XE2420, XR11, or XR12 — for processing and AI inferencing. With Dell's Streaming Data Platform, the data can then be efficiently streamed between locations. The data can be streamed to a datacenter, for example, for use in an AI training model and then sent back to the edge for AI inferencing, or it can be streamed to a public cloud where it is made available to a network of multiple sites. Another example might be streaming data from sensors that monitor the performance of complex machinery to the cloud where various companies that design parts for the system use computer-aided engineering to improve those parts based on the data.

Ultimately, the data that moves from endpoints to edge servers to cloud or datacenters feeds two pipelines: real-time analytics with, for example, Kafka and storage where it is made available to multicloud services. Dell's solutions cover this entire trajectory in an end-to-end fashion with its servers, workstations, storage platforms, and software.

Examples

A few use cases for the environments that Dell Technologies helped build are:

- Dell Technologies' own environment consists of a data lake with the Greenplum database to handle structured data and Hadoop for unstructured data. Dell feeds data into this environment for its AI-driven Dell EMC CloudIQ solution, which monitors the health and configuration status of its installed base to improve system performance and availability, speed problem resolution, and improve systems management. Data analysts can run queries on that data and use the results to help improve the overall customer ownership experience for Dell Technologies' end-to-end AI solutions. The infrastructure consists of Dell EMC PowerEdge servers, Dell EMC PowerScale storage, Dell EMC CloudIQ software, a layer of microservices, and TensorFlow.
- A healthcare company needed an environment to analyze medical imaging, especially related to the COVID-19 pandemic. The company looks for patterns in images as the virus changes to identify weaknesses in its armor. The company uses the Washington University supercomputer, with PowerEdge servers, PowerEdge networking, PowerScale storage, and Dell workstations.
- Another healthcare company collaborated with the Dell Technologies HPC and AI innovation lab and used a public data set of chest x-rays to improve the speed and accuracy of diagnoses, an objective whose purpose is to drive better overall patient care.

FUTURE OUTLOOK

IDC expects that within the next few years, AI will start to permeate business processes for most enterprises. In general, more data will drive better products and services, improved customer experience, and more relevant business insights. There will be a proliferation of data capture points as enterprises glean data from edge devices, their own products and services, employees, supply chain partners, and customers. All these data sets are generated by the thousands of operations that together constitute a business, but they themselves are not yet integrated or understood in relation to each other. What does a message from a sensor in a bottling machine in the United States mean for a soda drinks truck driver in Japan? What do social media messages about a U.S. airline's political stance mean for the fares on a domestic route? To deliver critical insights like these at scale requires end-to-end, real-time intelligence across an organization and its ecosystem. Data needs to stream freely, and where it naturally settles in a storage environment after having been leveraged for insights, it then needs to be joined by compute to perform more analysis, not the other way round. This is the kind of environment that businesses will start to build in the coming years.

CHALLENGES/OPPORTUNITIES

For Businesses

As business models become much more data driven, the key challenge for enterprises will be to identify and capture the data they need to improve their offerings and then use that data effectively to drive value for the business and its customers and partners. As they craft their strategies, enterprises will need to ensure that they respect the privacy of their constituents, effectively safeguard the data they do collect, and stay within various governments' regulatory requirements. A key consideration will be to build a cost-effective IT infrastructure that can continue to meet business performance, availability, and capacity requirements as their AI workloads and data continue to scale over time. Successful AI applications tend to grow very rapidly, and businesses will need to ensure they don't outgrow their supporting IT infrastructure. All enterprises will not meet these requirements, but there is an excellent opportunity to grow the business for those that do.

For Dell Technologies

As enterprises modernize their IT infrastructure to better meet the workload and business requirements of digital transformation, infrastructure requirements are evolving. Compute, networking, and storage infrastructure must provide higher performance and availability for workloads that will depend on data sets in the tens of petabytes range and beyond. In addition, the preferred deployment model for IT infrastructure will be hybrid multicloud going forward, and technology providers need to support this strategy by offering products and services that can be deployed effectively and simply, either on premises or in public cloud environments. They will also need to support the ecosystem that is developing around AI workloads with a focus on providing end-to-end AI solutions that are easy to buy and deploy and offer the features needed across the spectrum of AI environments. Given its broad portfolio and focus on delivering end-to-end AI, Dell Technologies is in a good position to drive value for customers deploying AI workloads.

CONCLUSION

AI workloads will become increasingly common in enterprises over the next several years, and they will demand many new capabilities from the underlying IT infrastructure. Enterprises should consider

the infrastructure requirements for AI from three angles — scale, portability, and time — as they modernize their IT infrastructure for the data-centric digital era. IDC's AI infrastructure plane provides a framework for enterprises to use when crafting their AI infrastructure. Enterprises will be building their infrastructure using both general-purpose and accelerated compute, distributed unstructured storage platforms, a mix of different storage technologies (including NVMe, SAS, and different media types), and AI-driven systems management as well as new AI framework tools like PyTorch and TensorFlow. Getting the underlying infrastructure right is a key determinant of success as enterprises look to AI to help drive better business decisions.

Enterprises successfully deploying AI will have their AI infrastructure distributed across edge, core, and cloud deployment locations, each of which will exhibit different workload profiles. Rather than thinking about AI infrastructure as a series of point deployments in different locations, enterprises should strive to craft a well-integrated end-to-end AI infrastructure strategy that leverages each of these deployment locations effectively (based on what each has to offer).

As a vendor of IT infrastructure, Dell Technologies has taken this point of view with the Dell Technologies Validated Designs for AI. Drawing on its broad infrastructure portfolio, which includes compute (both general purpose and accelerated), storage, and networking, Dell's solutions are based around proven AI expertise that reduces complexity and gets customers to more relevant business insights faster, validated by real-world customer success stories. It offers highly scalable unstructured storage platforms that can be deployed on premises or in the public cloud, along with a unified management dashboard that leverages AI operations that make administration simple and efficient. Dell Technologies has also created an AI ecosystem supporting the key partners and technologies necessary to deploy and manage effective AI-based solutions across edge, core, and public cloud-based infrastructure. Dell has the right focus and technology to help its customers transition to the digital era. With worldwide Customer Solution Centers, customers and partners are free to evaluate and test AI technologies and solutions to find the best fit. Dell Technologies can then provide a wide range of flexible consumption models.

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

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com