White Paper

Accelerating AI Modernization with Data Infrastructure

Sponsored by: IBM Corporation

Ashish Nadkarni Sriram Subramanian
February 2021

EXECUTIVE SUMMARY

Artificial intelligence (AI), Machine Learning (ML) and often Deep Learning (DL) capabilities are now essential components of digital transformation initiatives. The business opportunities that can be achieved with AI are exceptionally promising. Enterprise organizations are increasingly aware that not acting on AI could potentially be a business disaster as competitors gain a wealth of previously unavailable insights and capabilities to grow and delight their customer base. Few if any businesses today believe that "AI is not for us" or that "AI is mostly hype". Rather, serious AI initiatives are being undertaken worldwide, across industries, and across company sizes.

Businesses are seeking an AI-led transformation and modernization initiative, which involves moving from experimenting to generating business value from their AI investments. The success business investments in AI-related digital transformation are directly tied to the breadth of expertise required to develop, implement, and maintain AI solutions at scale. Many organizations' lines of business (LOB), IT staff, data scientists, and developers have been working to learn about AI, understand the use cases, define an AI strategy for their business, launch initial AI initiatives, and develop and test the resulting AI applications that deliver new insights and capabilities using machine learning algorithms, especially deep learning.

As organizations scale these initiatives, new questions emerge. They know – indeed, they may have experienced first-hand – that they cannot use standard, general-purpose computing, and existing or legacy storage infrastructure. Also, they realize that AI training (the training of the AI model) and AI inferencing (the use of the trained model to understand or predict an event) require different types of scalable compute with an equally scalable storage infrastructure. While businesses have a better handle on compute, they often underestimate the value of storage in AI. Further, AI applications and especially deep learning systems, which parse exponentially greater amounts of data, are extremely demanding, require powerful parallel processing capabilities based on large numbers of computational cores, and standard storage systems cannot sufficiently enable the execution of these AI tasks. Finally, how such initiatives factor into modernization efforts that include Kubernetes and/or containers, and integration with one or more clouds by way of a hybrid cloud architecture.

IDC research shows that, in terms of storage infrastructure, improper or inadequate attention to detail can quickly derail AI transformation initiatives. To overcome this gap, businesses - that have experimented with existing infrastructure, and are now ready to scale this into production - must overhaul their infrastructure to obtain the required parallel processing performance and do so by investing in more modern storage solutions that scale-out for massive scale and integrated into the cloud, containers, and performance-intensive compute for both global deployment and data access.

This is where solutions like IBM Spectrum Scale and IBM Elastic Storage System (ESS) provide the necessary components for an AI information architecture. It is suited for AI workloads, containerized deployment and a hybrid cloud deployment specifically focused on AI workloads.

SITUATION OVERVIEW

AI/ML is Here and Now

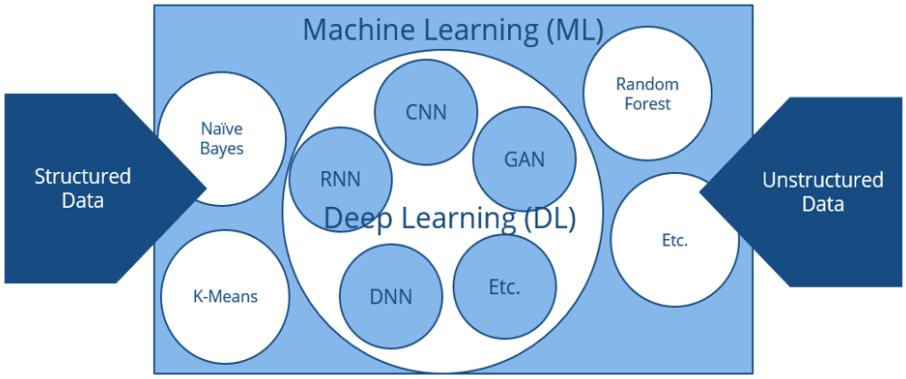
Businesses worldwide are responding vigorously to the new opportunities offered by investments in AI to catalyze their digital transformation initiatives. Artificial Intelligence are a set of technologies that use Natural Language Processing (NLP), image/video analytics, machine learning, knowledge graphs and other technologies to answer questions, discover insights, and provide recommendations. These systems hypothesize and formulate possible answers based on available evidence, can be trained through the ingestion of vast amounts of content, and adapt and learn from their mistakes and failures through re-training or human supervision.

AI is quickly becoming pervasive business-wide in the automation of processes and workflows. In 2019, IDC examined 176 digital transformation use cases across eight line-of-business functional areas, including customer experience, legal and corporate strategy, facilities, and procurement, and estimated that roughly 26% of these use cases were both dependent on AI and currently deployed across most organizations. By 2022, IDC expects that at least 60% of these AI-centric use cases will be deployed in at least 65% of Global 2000 organizations, representing a 34% growth from 2019. This means that soon, most leading organizations will be leveraging AI technologies, such as natural language processing, machine learning, and deep learning, and speech to text across the organization to scale operations, make sense of unstructured data, and deliver intelligent business insights. Meanwhile, organizations that have still not figured out how to move AI-based use cases from proof of concept (POC) to production will fall further behind, widening the digital divide.

Machine Learning (ML) is a subset of AI techniques that enables computer systems to learn and improve their behavior for a given task without having to be programmed by a human. Machine learning models are algorithms that can improve over time by testing themselves repeatedly using large amounts of structured and/or unstructured data until they are deemed to have "learned" a task, e.g., recognizing a human face. Figure 1 illustrates how Deep Learning (DL) is a subset of ML. Typical DL architectures are deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GAN), and many more.

FIGURE 1

Machine Learning and Deep Learning Applications



Source: IDC 2019

AI software platforms include: Conversational AI software, for example: digital assistants, predictive analytics to discover hidden relationships in data and make predictions, text analytics and natural language for recognizing, understanding and extracting value from text, voice/speech analytics for recognizing, identifying and extracting information from audio, voice and speech and image and Video Analytics for recognizing, identifying, and extracting information from images and video, including pattern recognition, objects, colors, and other attributes such as people, faces, emotion, cars, and scenery to name a few.

Many businesses are well on their way with AI initiatives and have reached a stage where they are ready to start deploying AI at production scale. Others are still experimenting with AI, while a third group is currently at the stage of evaluating what AI applications can mean for their organization.

Regarding the first group (ready to deploy), IDC is seeing a range of AI use cases that businesses, governments, and other organizations have begun to implement. The five most common use cases today are, ranked by the amount that businesses spend on them in terms of hardware, software, and services:

1. **Automated Customer Service Agents** - In the banking industry, for example, these AI applications provide customer service via a learning program that understands customer needs and problems and helps a bank reduce the time and resources needed for resolving customer issues. These agents are becoming widely used across industries.
2. **Sales Process Recommendation and Automation** – Used in in various industries, these are AI applications that work with Customer Relationship Management (CRM) systems to understand customer context in real time and recommend relevant actions to sales agents.
3. **Automated Threat Intelligence and Prevention Systems** – Becoming a critical part of threat prevention across governments and industries, these AI applications process intelligence reports, extract information from them, establish relationships between diverse pieces of information, and then identify threats to databases, systems, websites, etc.

4. **Fraud Analysis and Investigation** – In the insurance industry, but used widely elsewhere as well, these AI applications use rule-based learning to identify fraudulent transactions, and they automatically learn to identify various insurance-related fraud schemes.
5. **Automated Preventative Maintenance** – In the manufacturing industries, these AI applications are based on machine learning algorithms that build an accurate predictive model of potential plant and machinery failures, reducing downtime and maintenance cost.

Additional AI use cases that have gained traction in enterprises are (ranked in order of spending on hardware, software, and services) include:

- Program Advisors and Recommendation Systems
- Diagnosis and Treatment Systems
- Intelligent Processing Automation
- Quality Management Investigation and Recommendation Systems
- IT Automation, Digital Assistants for Enterprise Knowledge Workers
- Expert Shopping Advisors and Product Recommendations
- Supply and Logistics, Regulatory Intelligence
- Asset/Fleet Management, Automated Claims Processing
- Digital Twin/Advanced Digital Simulation
- Public Safety and Emergency Response
- Adaptive Learning
- Smart Networking
- Freight, asset, and fleet Management
- Pharmaceutical Research and Discovery

AI Transformation Requires a Solid Data Foundation

As illustrated in Figure 1, ML and DL initiatives rely heavily on a combination of various structured and unstructured data inputs. AI projects involve multiple steps including data collection, configuration, sanitization, verification, model building, training, testing, inference, and retiring data. With each step having varying requirements – from faster access, low latency storage to low cost, archival storage, enterprises need to select appropriate and cost-effective storage infrastructure for these steps. They also need to use disparate set of tools to manage AI dataset across their lifecycle.

With the rollout of 5G and the proliferation of IoT-based sensors, more data is generated and consumed at the edge. Cameras and intelligent assistants are deployed at the edge, on the device to provide faster response and greater user experience. The need for local AI, processed at the edge of the network or on endpoints, is growing. Latency-sensitive AI applications running on edge devices with limited connectivity will require high scalability. AI-enabled edge computing use cases will be adopted.

IDC research shows that businesses often cite the following challenges with their AI initiatives from a data management perspective.

- Data ingest and preparation cycles are too time consuming.
- Silos of infrastructure for various analytics use cases.
- Multiple copies of same data without a single source of truth.

- A need to securely manage and protect data provenance for repeatability.
- A need for global accessibility (hybrid cloud) and collaboration.
- Data integrity once the data has been captured and curated.

An AI Data Infrastructure Accelerates AI Transformation

As organizations invest in AI, multiple infrastructure design and deployment factors are becoming critically important. Taking a page from the hyperscale and cloud service provider book, organizations are approaching the infrastructure requirements via the creation of a unified Data Infrastructure. A data infrastructure is a common foundation for AI initiatives including highly efficient and scalable compute and storage layer. Instead of approaching AI workloads as homogeneous, a data infrastructure treats them in a composite manner and connects a portion of that workload to the right compute layer powered by an appropriate storage layer depending on mix of structured and unstructured data sets. This composite approach accelerates AI Transformation along three dimensions:

- **Scale.** The scale dimension describes the scale at which the workload operates. Foundational subdimensions — compute, networking, and data persistence (storage) — are all hardware related. Crucially, software-related subdimensions such as orchestration are gaining equitability for maintaining balance with an increase in the size and complexity of the stack.
- **Portability.** This is the ability of the workload to be moved across core, edge, and endpoint deployments. Today, many of these workloads are static in nature (i.e., designed to run in a single deployment). Increasingly, companies are looking at developing workloads in one deployment (e.g., public cloud) and installing them (in production) at another one (e.g., edge). This is analogous to the current model of developing and deploying mobile apps.
- **Time.** This relates to the time continuity of the workload itself. Many AI workloads borrow their design from high-performance computing or big data and analytics deployment — they are designed to be batch in nature. Increasingly — and thanks to the proliferation of high-performance accelerators — AI workloads can analyze streaming data in a real-time or near-real-time manner.

The Myths and Must Haves of AI Data Infrastructure

An essential albeit often-overlooked foundation is storage infrastructure. Deploying AI at scale often places higher demands on the storage infrastructure in terms of capacity (growth) and performance (IOPS and bandwidth). Organizations often assume that internal server-based storage or enterprise storage that is used for other workloads is sufficient for running AI applications. And once the infrastructure is built out, they realize that storage is the weakest link in the chain. Each of these AI applications pose a different set of requirements and, hence, challenges to the IT organization. IT buyers and vendors should therefore avoid the proverbial "I have a hammer, therefore everything is a nail" situation.

A more prudent approach is to take a holistic approach to data infrastructure. While data persistence scaling and access mechanisms are table stakes, organizations need to expand the purview to include networking and integration between the computing, storage software and systems tiers. Businesses need to make a shift to a consistent end-to-end Data Infrastructure mindset and not just one that involves "plugging in another storage systems. IDC believes that Data Infrastructure requirements specifically around storage can be distilled down to the following key areas.

Compute Integration

It is often assumed – and this is a flawed assumption – that all AI workloads are containerized. Far from it, many AI workloads are run on bare metal or even virtualized. AI-enabled applications, especially, often running on bare metal or virtualized compute. Many AI workloads are being optimized to utilize accelerators. That does not mean all AI workloads are best running on accelerated compute – accelerated compute brings with it a different set of challenges for workloads in general.

Data Persistence and Access

If the compute requirements for AI workloads vary, so do their data persistence requirements. An underrepresented and misunderstood aspect of the AI workloads stack is the data persistence layer. It is often assumed – again a flawed assumption – that all AI workloads require a large amount of high-performance storage. The fact of the matter is that not all AI workloads are "big data sets" – they may be sampling lots of small data sets concurrently for a short period of time. Similarly, bare metal workloads running on open systems computing platforms often use scale-out block or file access. It is not that uncommon for virtualized workloads to run on hyperconverged infrastructure (HCI).

Multi-protocol access is a given with a mix of structured and unstructured data ingest into storage infrastructure. Many IOT and edge devices communicate via SMB or NFS, and a few use S3. Streaming data access is necessary in some cases. And in some cases, native parallel file system client could be used as well.

Scaling and Tiering

Supporting AI and ML applications means that storage systems must deliver performance at scale. For unstructured data repositories it is storage systems that make use of parallel file systems with network access. For structured data it is the use of flash-based storage systems. Scaling is essentially the dialing up of performance and capacity independent of other to meet the requirements of AI and ML applications.

Further, for future-proofed infrastructure, the system also needs to be able to simply and cost-effectively handle data that is older or colder onto a low-cost object storage with a known object storage interface like S3.

Software-defined Storage

AI and ML act as catalysts for software-defined storage. They enable infrastructure as code and automation via a heterogeneous software control layer above the hardware. This helps better integration with the AI/ML workflows, thus ensuring that the storage seamlessly scales along with the demands of the application.

Deployment Agility and Flexibility

The applications that address these use cases may be custom developed by an organization, may be based on commercial AI software, or they may be delivered as AI SaaS. Deployment considerations for the custom developed and commercial software are on-premises, in the cloud on IaaS, or as a hybrid cloud, wherein the on-premises environment interacts with a public cloud environment via a common automation and orchestration layer.

Given the 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 than the other way around. In recent times, the core-edge-

endpoint model has become a de facto way of describing AI (where core includes cloud and endpoints include embedded intelligence). It is important to note that the workload profiles for each location vary, and therefore the underlying infrastructure requirements.

For the various deployment scenarios, solutions must be considered for:

- Securely processing the volume of data that is required for training AI models with extremely high performance. The performance requirements for Deep Learning training involve the ability to execute massively parallel processing using GPUs combined with high-bandwidth data ingestion.
- Securely processing the volume of data that the AI model will perform inferencing on with extremely high performance. Performance related to inferencing means the ability to process incoming data through the trained AI model and deliver near real-time AI insights or decisions.

For data scientists and developers, it can sometimes be easier to start an AI initiative in the cloud, saving them from having to arrange on-premises compute that, for Deep Learning, typically needs to be accelerated. Accelerated AI cloud instances are available on most public clouds, usually with open-source AI stacks. Of course, with accelerated cloud instances for AI training, the CSP dictates what is available to the end user in terms of processors, co-processors, interconnects, memory sizes, I/O bandwidth, etc. Not all CSPs offer the most optimized combinations of these components, which ultimately determine the speed and quality with which data scientists can develop training models. As a result, many organizations opt for on-premises deployment.

During their AI experiments in the past few years, many organizations found themselves "hitting the wall" with their standard infrastructure or with the basic cloud instances. Training models took too long, and inferencing was too slow. IDC research shows that 77.1% of respondents say they ran into one or more limitations with their AI infrastructure on-premises and 90.3% ran into compute limitations in the cloud.

EXTENSIBLE GLOBAL AI INFORMATION ARCHITECTURE USING IBM STORAGE

IBM Storage solutions for Data and AI enable customers to seamlessly introduce AI initiatives at production scale in hybrid cloud environments. IBM continues to drive leadership for scalable high-performance workloads as well as efficient, secure, scalable, capacity storage for high performance AI and big data solutions. IBM's storage portfolio provides integrated storage and data management from the edge, the core data center and the public cloud accelerating AI modernization. It comes ready with broad support for and integration with Kubernetes containers and the Red Hat OpenShift platform and can be deployed and accessed in the public cloud or for data center workloads. IBM Storage for Data and AI seeks to reduce complexity and cost by providing deep increased integration with an AI information architecture that can be deployed at scale for the entire organization.

IBM Spectrum Scale

IBM Spectrum Scale is built on a distributed computing architecture that is designed for any performance-intensive workloads such as AI/ML, modeling and simulation and analytics. It is a clustered, parallel file system that reduces OPEX via simple management and scalability and CAPEX via policy-based data optimization and transparent data lifecycle management. IBM Spectrum Scale provides concurrent access to a single file system or set of file systems from multiple nodes. The nodes can be direct attached, network attached, a mixture of direct attached, and network attached, or

in a shared nothing cluster configuration. This enables high performance shared access to this common set of data to support a scale-out solution and to provide a high availability platform.

IBM Spectrum Scale supports data replication, policy-based storage management, and multi-site operations. IT operations teams can create a cluster of Kubernetes container nodes, IBM AIX nodes, IBM Z or LinuxONE nodes, Linux nodes, Microsoft Windows server nodes, or a mix of all five. IBM Spectrum Scale can run on virtualized or containerized instances providing common shared data access in environments, leverage logical partitioning, or other hypervisors. Multiple IBM Spectrum Scale clusters can share data within a location or across wide area network (WAN) connections for global data collaboration and data access.

IBM Spectrum Scale delivers a rock-solid foundation for AI infrastructure built on years of servicing the high-performance computing industry. As Figure 3 illustrates, IBM Spectrum Scale provides a global namespace, shared file system access among IBM Spectrum Scale clusters, simultaneous file access from multiple nodes, high recoverability and data availability through replication, the ability to make changes while a file system is mounted, and simplified administration even in large environments. Key differentiators of IBM Spectrum Scale are:

- Shared file system access among IBM Spectrum Scale clusters allows data sharing between separate clusters within a location or across a WAN.
- Improved system performance using IBM GPFS file systems can improve system performance.
- File consistency via concurrent and detailed access to clients across the cluster by utilizing token management.
- Increased data availability and reliability via features like file system logging and configurable features like intelligent mounts.
- Enhanced system flexibility enables the addition or deletion of disk resources while the file system is mounted.
- Simplified storage management helps achieve information lifecycle management (ILM) through powerful policy-driven, automated tiered storage management.
- Simplified administration via many standard file system interfaces that can be executed directly from most applications.
- Hybrid cloud deployment provides data availability, integrity, security and optimized container native storage and integration with Red Hat OpenShift.

IBM Elastic Storage System (ESS)

IBM Elastic Storage System (ESS) is a modern implementation of software-defined storage designed as easy to configure and manage building blocks, making it easier for IT organizations to deploy fast, highly scalable storage for performance intensive computing applications, include AI, big data, and analytics applications. IBM ESS:

- Is built with the NVMe flash storage to deliver exabyte-level scalability and consistent service quality throughout entire infrastructure.
- Can be integrated with file management and data services capabilities of IBM Spectrum Scale to deliver a federated global storage system.

- Enables businesses to consolidate storage requirements from the edge to the core data and integrate with the public cloud thus reducing inefficiency, lowering acquisition costs, simplifying storage management, and eliminating data silos.
- Delivers consistent high-performance access for multiple demanding applications deployed as bare-metal or in virtualized environments. It supports Kubernetes and Red Hat OpenShift integration making it easier to deploy cloud-native applications.

IBM ESS is available in two form factors – 3000 and 5000.

- IBM Elastic Storage System 3000 (ESS 3000) is designed to meet and beat the challenge of managing data for analytics. Packaged in a compact 2U enclosure, ESS 3000 speeds time to value for artificial intelligence / deep learning and performance-intensive computing applications thanks to its all-NVMe storage and simple, fast containerized software installation and upgrade. The hardware and software design of the ESS 3000 gives organization access to industry-leading performance required to keep data-hungry compute fully utilized.
- IBM ESS 5000 is an Elastic Storage System, offering scale-out PB capacity nodes with high throughput, that combines the software-defined IBM Spectrum Scale storage with IBM POWER9 processor-based I/O-intensive servers. By consolidating storage requirements across the organization onto IBM ESS 5000 and the NVMe-based ESS 3000, IT teams can reduce inefficiency, lower acquisition costs, and support the demanding AI, HPC, analytics, and/or high-capacity storage requirements that are typical in the fields of healthcare, media, government, or financial services. The ESS 5000 can start with TBs and grow to hundreds of PBs or even EB scale with a single unified namespace that eliminates costly data silos. IBM Spectrum Scale is the parallel file system at the heart of IBM ESS 5000 that expands throughput as it grows. It allows for integration with previous ESS models for investment protection and provides lower cost options such cloud storage and IBM Tape. With IBM Spectrum Scale, IT can eliminate data silos and bottlenecks, simplify storage management, and get faster access to data.

IBM Cloud Object Storage (COS)

IBM Cloud Object Storage is an industry-leading software-defined highly scalable and cost-effective storage solution for storing unstructured data on the edge, the core data center or in private or public clouds. IBM Cloud Object Storage is ideal for deploying or modernizing performance intensive infrastructure for AI, analytics, IoT, video and image repositories. It also provides unprecedented value allowing organizations to lower storage costs up to 12%, using new 18 TB SMR (shingled magnetic resonance) drives while increasing throughput up to 55 GB/s in a 12-node cluster. Businesses can protect their data with our local or geo-dispersed data protection that can be customized for demanding data lakes and large capacity requirements.

IBM Cloud Object Storage is a foundational system for AI infrastructure with key benefits such as:

- Scalability: Supports exponential data growth scaling performance and capacity from terabytes to exabytes.
- Security. Built in encryption and policy enabled lockable WORM (write once, ready many) storage.
- Simplicity. Access data concurrently from any location. Provides automatic failover, data rebuild, auto expansion and rebalancing.
- Savings efficiencies. Delivers geo-protected data with Information Dispersal Algorithm (IDA) efficiency. Is available as software only or fully supported appliance solutions.

- Search capabilities. Offers custom insight and search to save time. Organizations can create custom metadata to enhance value.
- Enhanced file access. A new file access software gateway can seamlessly attach to any Windows or Linux file system using SMB or NFS access, easily connecting file-based applications to object storage.
- High-speed transfer. The IBM Aspera high-speed data transfer option makes it easy to transfer data, and flexible storage class tiers help manage costs while meeting data access needs.

FUTURE OUTLOOK

In 2022, a total of 65% of global GDP will be digitalized, driving $6.8 trillion of IT spending worldwide from 2020 to 2023 (see IDC FutureScape: Worldwide Digital Transformation 2021 Predictions, IDC #US46880818, October 2020). Digital infrastructure is not limited to traditional central enterprise services nor to discrete cloud datacenters. It includes all assets and resources that enable the shift of applications and code for transformation. It will be the foundation for improved customer experience (CX). It also enables embedding intelligence/automation into business operations, and it supports ongoing innovation right to the digital edge of the enterprise and the industry. Successful digital strategy must transform the digital infrastructure to eliminate silos, break down technology-imposed barriers, and go beyond just supporting traditional tools and applications.

IDC believes that AI will be the foundation of digital infrastructure. IDC's Customer Insights and Analysis Group recently conducted a survey to understand current and future IT spending and adoption plans across organizations of all sizes from various industries. The study found that almost 76% of respondents – which consisted of 3600 key IT Decision Makers at IT organizations from various industries worldwide - indicated that Artificial Intelligence is a key part of their DX strategy or expected over the next one to two years. Only 22% of survey respondents indicated it will be in the next three to five years. Of those who expect AI to be a key component of their DX strategy in the next one to two years, Telecommunications, Utilities, Education and Professional Services are most likely to turn to AI in their digital transformation efforts.

AI is about "time to value" – the value that companies can obtain from data in the fastest time possible. IDC's FutureScape on Artificial Intelligence estimates that "Artificial intelligence is the most disruptive innovation of our lifetime. AI is not just 'nice to have' anymore. The global pandemic has accelerated AI adoption, and it is becoming ubiquitous across all business processes. AI solutions powered by machine learning, conversational AI, and computer vision are at the forefront of business resiliency, accelerated innovation, and transformative customer and employee experiences. About 51% of respondents in the above indicated that they are currently evaluating artificial intelligence or already in production; this is up from 34% of respondents in 2019. AI's greatest impact is in helping employees to get better at their jobs. AI enterprise uptake will continue to expand as benefits from full deployments become more tangible.

Artificial intelligence is quickly becoming pervasive business-wide in the automation of processes and workflows. In 2019, IDC examined 176 digital transformation use cases across eight line-of-business functional areas, including customer experience, legal and corporate strategy, facilities, and procurement, and estimated that roughly 26% of these use cases were both dependent on AI and currently deployed across many organizations. By 2022, IDC expects that at least 60% of these AI-centric use cases will be deployed in at least 65% of G2000 organizations, representing a 34% growth from 2019. This means that soon, most leading organizations will be leveraging AI technologies, such

as natural language processing, machine learning, and deep learning, and speech to text across the organization to scale operations, make sense of unstructured data, and deliver intelligent business insights. Meanwhile, organizations that have still not figured out how to move AI-based use cases from proof of concept (POC) to production will fall further behind, widening the digital divide.

AI is also expanding across a variety of functional areas. With an increase in remote work, staff are less likely to be collocated and less able to perform manual handovers for many tasks, thereby requiring the use of AI tools for document processing, process automation, and question answering. Increasing use of AI-supported data ingestion, translation, and retrieval is projected as structured and unstructured data volume continues to rise exponentially.

Intelligent search and other AI-based digital assistance tools will help workers derive insights from this deluge of data in their moment of need, augmenting and enhancing human capabilities. In addition to providing a foundation for these kinds of use cases as the future of work becomes increasingly digital, artificial intelligence will be a critical component of contactless experiences to support safer in-person and onsite interactions. For example, in facilities management, predictive AI can support dynamic hoteling to manage safe distribution of in-office employees, while speech recognition, digital assistance, and intelligent automation technologies combine to provide voice-based controls for in-room utilities and meeting systems. Voice-based interfaces were the top planned investment area for contactless experiences, according to IDC's *COVID-19 Impact on IT Spending Survey* of 572 global organizations (conducted August–September 2020).

IDC believes that by 2024 deployment of ML-based automation for intelligent optimization will replace rigid processes, resulting in the ability for G2000 to rapidly react to external market factors.

IDC estimates that AI infrastructure investments will continue to be strong in the next few years. IDC projects the AI hardware (server and storage combined) revenues to reach $13.4 billion in 2020, representing 10.3% year-over-year growth. Within the hardware market, AI Storage is forecast to grow 11.4% this year. The overall hardware market is forecast to have a strong recovery in 2021 with 35.5% year-on-year growth next year, led by AI Storage, which is expected to grow 43.1% year over year.

Storage will continue to be the bedrock on which AI initiatives can scale now and in the future. Investments in AI, and data modernization initiatives due to AI will drive investments in scale-out file storage and unstructured data. IDC recently surveyed 624 IT practitioners and operations teams globally to determine IT infrastructure adoption trends. In this study, IDC found that over 65% of respondents prefer scale-out file systems accessed locally or via NFS for their high-performance workloads like AI. In fact, for AI workloads specifically, which includes training and inferencing workloads, performance was the top requirement for storage. This was followed by ease of deployment in a hybrid cloud and quality of service.

ESSENTIAL GUIDANCE FOR IT BUYERS

Technology buyers are rightly confused about the process of building their own AI infrastructure stack. They have defined use cases, launched AI initiatives, and hired or trained data scientists and application developers but are suddenly finding themselves constrained by the infrastructure to develop AI models on. Often, existing infrastructure is leveraged briefly, followed by investments in accelerated infrastructure. Data scientists are finding themselves putting stacks together on the accelerated infrastructure and trying to make it work, which is ultimately not part of their job

description. IT infrastructure teams are not familiar with the stacks that data scientists need and cannot put them together and optimize them. This has led to a severe skill gap that server vendors and cloud SPs have attempted to fill with their own stacks, each in their own way. Today, there are as many stacks as there are vendors, often overlapping when they are developed by multiple members in the same value chain.

Start with Business Outcomes

Organizations must begin by tying together service constraints and use cases to identify business outcomes that benefit from an investment in AI infrastructure. They must seek to quantify and measure the benefits of such investments. For example, when seeking competitive differentiation, the question would be by how much and by when? These criteria should then lead to the selection of an application architecture. Businesses must consider AI when looking at enhancing their brand through better understanding of and focused responsiveness to customer sentiment, wants, and needs, which in turn lead to increased revenue and profits.

Take a Holistic Approach

It is important to view the entire picture i.e., take a comprehensive view when implementing any AI initiative. Looking at only one problem by itself can result in the creation of (yet) another silo, or worse still increase complexity in the environment due to lack of integration and interoperability between multiple architectures and solutions. A data infrastructure must be looked at as a global solution from edge to core to cloud, and across use cases like AI and other mission critical applications.

Develop Right Application and Data Architecture

Developing an AI application and data architecture is a complex task. This involves conversion of business requirements and outcomes into a deterministic AI-enabled workflow. The workflow needs to describe the way AI capabilities enhance the behavior of that application, how data is ingested and analyzed, and how the application interacts with other business applications and with users. The focus here must be on how data is consumed, produced, and analyzed by applications and the implications on the hardware. When creating a greenfield plan, it should focus on the blend between custom (open source or proprietary) and off-the-shelf software components.

Choose the Right Reference Stack

Several vendors and service providers have put out reference stacks for implementing an AI infrastructure. Many of these are "open" in nature, allowing a modular "plug and play" experience, and can be consumed on a pay-as-you-go service for capex-friendly implementation. This is an important consideration as AI infrastructure investments can get expensive quickly. IDC plans to publish its perspective on popular vendor reference stacks in an upcoming document.

IT benefits to keep in mind when examining reference stacks are reduced costs, data and application availability, effective infrastructure utilization and consolidation and, where possible, a single interoperable application delivery platform.

Create an Information Architecture for AI

Enterprises need a data management strategy to provide flexible, organized access to all data, of every type, regardless of where it lives, and addresses the above concerns. A modernization effort would define and deploy an information architecture that provides an open, extensible foundation, with choice and flexibility, capable of communicating with other cloud platforms. IBM’s hybrid data

management strategy to accelerate the Journey to AI is a prescriptive approach defined by a 4-step Ai ladder: Collect, Organize, Analyze, and Infuse.

- **Collect**: Make the data simple and accessible at the right location, from any database or storage facility.
- **Organize**: Ensure that data is trustworthy, complete, and consistent at all stages of the information lifecycle: profile, cleanse, and catalog the data, provide protection and compliance, enable policy-driven visibility, detection, and reporting.
- **Analyze**: Build, deploy, and manage AI models using integrated tools to explore and analyze both structured and unstructured data, and deploy them securely.
- **Infuse**: achieve trust and transparency in model-recommended decisions, explain decisions, detect bias, etc., using provided solutions and services.

The journey to AI is about moving data from ingest to insights with an information architecture that can easily be infused throughout the organization. It is important that each part of the AI ladder provides an integration to the entire journey. Storage typically has been implemented in tactical ways with specific storage solutions that create silos of data and solutions that are not integrated together or with a comprehensive set of infrastructure solutions. Customers may store data on a large file or object storage system but then not have details about that data or use that data for additional insights. Customers can still start or focus a project on one part of the journey, but each project should consider an overall AI information architecture to optimize resources and modernize your infrastructure for expanding AI workloads.

Organizations that treat AI as a high-performance composite application (made of several applications that are interconnected), borrowing crucial elements such as scale-out file systems, heterogeneous computing and high-speed interconnects for distributed compute and storage access are the ones that can ultimately scale their AI infrastructure.

Leverage Right Partnerships

For IT buyers, partnering with an end-to-end solutions provider is crucial to long-term success, but IDC does not currently believe that any vendor in the market today is providing that end-to-end environment yet, even as the vendors are working hard on getting there. Nevertheless, by engaging with a trusted partner, organizations can better use AI approaches to grow their business. They can become agile and adaptive and take advantage of internal synergies to drive profitability. Finally, they can get ahead of the disruptive forces in their industry by reinventing themselves. An ideal partner would provide:

- Proven solutions that scale from small lab deployments up to massive global deployments
- Vertical segment expertise focused on segments that align to their business
- Integrated access to a diverse ecosystem of ISVs and infrastructure providers
- A data-first viewpoint to ensure you secure and maximize the long-term value of their investments in new data sources
- Verified success simplifying the hardware, software, and security aspects of large projects

CHALLENGES/OPPORTUNITIES

For Organizations

This paper has discussed a range of challenges organizations face when they are ready to scale their AI applications for production. From data preparation to model development to runtime environments to training, deploying, and managing AI models, the requirements for the underlying infrastructure defy the old models of general-purpose hardware. Investments in an infrastructure that is designed for data-intensive workloads, with superior performance, scaling, data access and integration and can blend into a hybrid cloud environment provide long-term value and service quality. Organizations will need to make decisions about replacing or supplementing existing general-purpose storage platforms with storage systems that are geared towards AI specific processing tasks. This will serve as the foundation for developing and run cutting-edge AI applications.

For IBM

The challenge for IBM is always one of market recognition. IBM offers exceptional and comprehensive AI infrastructure solutions that are integrated with well infrastructure software stacks (like Red Hat OpenShift) and the public cloud, but that potential customers – incorrectly – perceive as complex or more costly. The subsequent knee-jerk reaction to go with one of the large commodity storage systems vendors for extremely data-intense workloads is depriving these organizations of AI infrastructure solutions that they could truly benefit from. The IBM Storage for Data and AI solutions with IBM Spectrum Scale and ESS, for example, is a supercomputing building block for many of the largest datacenters. Now that new AI workloads are starting to seriously mimic supercomputing deployments with HPC requirements and challenge the on-premises and cloud infrastructure they run on, this is the moment for IBM to take the stage and win new customers.

CONCLUSION

In the past several years, IDC has witnessed the AI Transformation in many organizations - how they started to develop a wide range of AI capabilities. Initially launched as experiments by relatively inexperienced staff and executed on whatever infrastructure was available, these initiatives have now started to gain critical mass. Many organizations have developed extensive AI expertise and they are experiencing first-hand the speed with which their AI capabilities are becoming a critical aspect of their business.

At the same time, IT too has undergone a transformation through a learning curve with respect to the infrastructure to run AI on. There is today much greater clarity as to the infrastructure requirements for Deep Learning training or inferencing and how to scale those environments for production. That Deep Learning training requires different infrastructure than other applications is pretty much a given. Deep Learning training wants clustered nodes with strong processors, powerful co-processors, scalable storage enabled fast interconnects, large I/O bandwidth, and plenty of memory.

Today, the biggest decision that IT must make is how to best design and deploy their Data infrastructure AI applications with best of breed systems, connect and optimize them together. IDC believes that IBM storage solutions for AI which include Spectrum Scale, Elastic Storage System and Cloud Object Storage Power offer unprecedent value and performance.

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