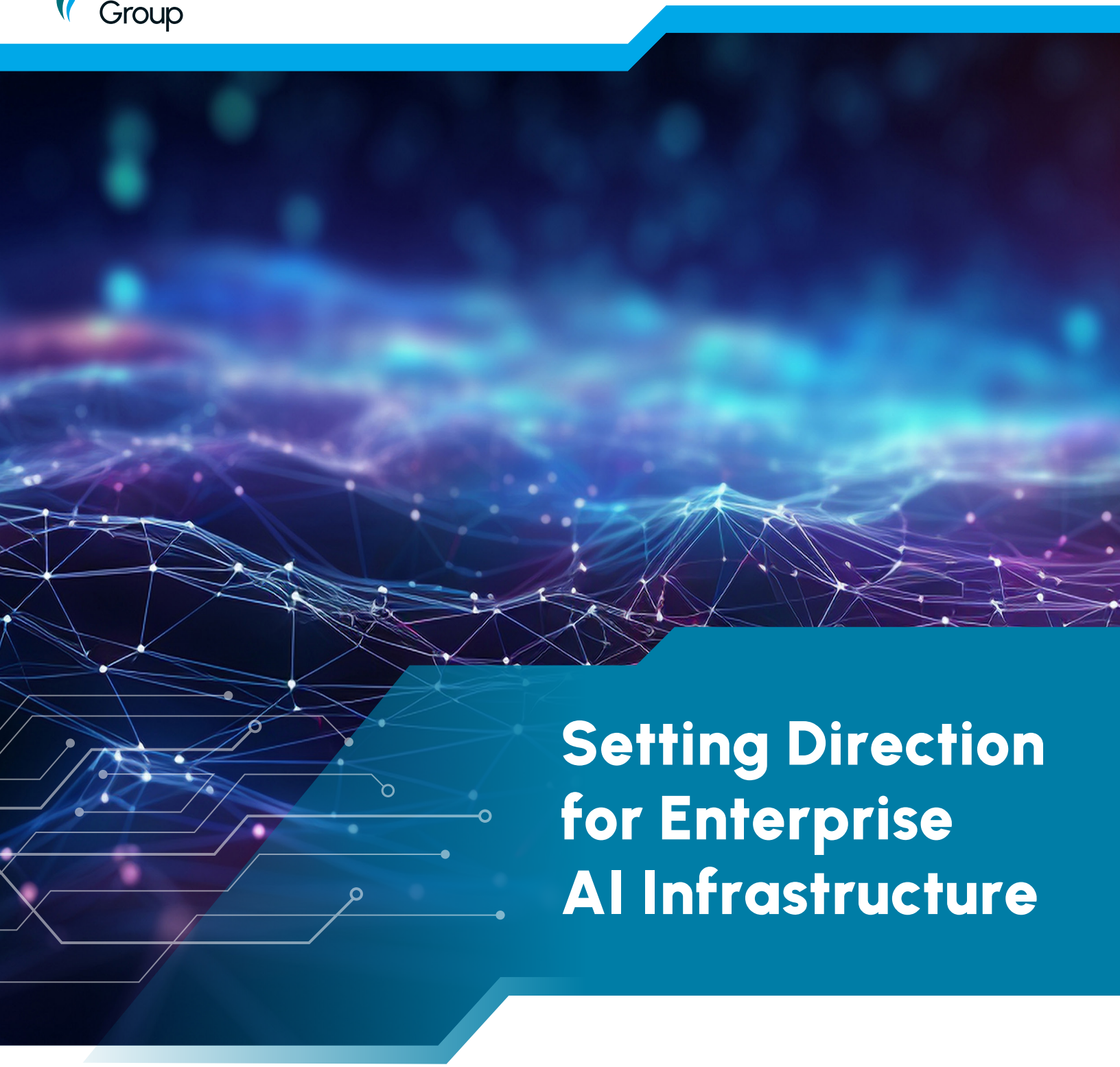
The Futurum Group



Setting Direction for Enterprise AI Infrastructure

AUTHOR

Randy Kerns
Senior Strategist and Analyst | The Futurum Group

IN PARTNERSHIP WITH

PURESTORAGE®

MARCH 2024

Overview

Pervasive publicity and discussions about artificial intelligence (AI) have brought heightened awareness of its possibilities to organizations in many areas. Organizations recognize that AI will benefit them through improved efficiencies or new solution offerings, to the point of understanding that investing in AI is a business competitive issue.

Moving beyond the recognition of the potential of AI, organizations are evaluating what is necessary to achieve the advances to realize the expected value. Part of this evaluation is an understanding of the resources required, including people, systems, and processes. Technologies in key areas are also part of the evaluation, and a critical consideration is the selection regarding the handling of the information assets necessary to power the AI function.

Enterprise organizations are in the early stages of AI deployment, developing strategies on how to effectively bring their valuable information into AI models. This strategizing leads to important decisions regarding that information: providing the data efficiently to the AI processes, protecting it according to the governance requirements for information assets, and meeting the needs for an AI environment. For IT groups in these organizations, there is much to learn regarding AI, and it is a rapidly evolving area. Understanding the environment for AI and the specialists involved is necessary for successful creation of an AI production environment.

AI teams have gained experience developing models on public clouds using tool sets and cloud resources. Enterprises have presented them with new challenges in using proprietary information assets for models, now to be developed on-premises to maintain control of the assets. AI data scientists have not needed to deal with critical infrastructure for accessing and storing data in their cloud-based model development. Now, enterprise IT is being tasked with providing the infrastructure required to enable the AI team to continue development without having to deal with the unfamiliar infrastructure areas. Clearly, there is a great deal for IT organizations to learn and limited staff available. IT will lean on vendors for solutions and expertise.

Choosing technology to meet the evolving AI demands involves a number of key decisions for IT, affecting the success and expediency of delivering on the AI aspirations of the organization. While the required compute, graphics processing unit (GPU), and AI tools are well understood by data science teams, efficient, scalable storage that is optimized for AI can make a substantial difference in accelerating results, keeping costs contained, and improving data scientist productivity. IT is responsible for providing an infrastructure for AI that can meet the changing requirements including expected growth in the amount of information that will be used for training models. Understanding the requirements and determining the best solutions is expected from the IT organization by the AI team.

New Demands on Infrastructure for AI

An AI infrastructure is new for the IT team. There may not be extensive experience with GPUs or similar processors and the system architectures within IT. Certainly, there is experience with storing and managing information, but the usage for AI and the applications may be unfamiliar. AI environments operate on information pulled from many different sources with data engineers "conditioning" the data for use in creating the model. Data to be used in the process may come from different sources: extracted information from databases, file data from different applications, and externally acquired information. The data did not originate with the idea that it would be used in creating a model. Consequently, the different locations, storage formats, and access methods must be dealt with individually. The sheer potential size adds to the complexity of the task. The outcome from data conditioning, creating training datasets used in the AI framework, has specific requirements. The performance for accessing the data is critical in model delivery. Optimized performance may require access from GPUs or CPUs with accelerators, considerations that must be made by the IT infrastructure team.

Responsibility of the integrity of the data and availability are assumed by the data science team, but it is the responsibility of IT to deliver on that. For the data science team, data is expected to be available immediately, so ensuring data is provided as fast as is practical is another focus area for IT around the selection of storage systems and infrastructure. For AI developers, data services such as replicated copies, data protection, and database access become valuable capabilities delivered by the technology selections of the IT infrastructure team.

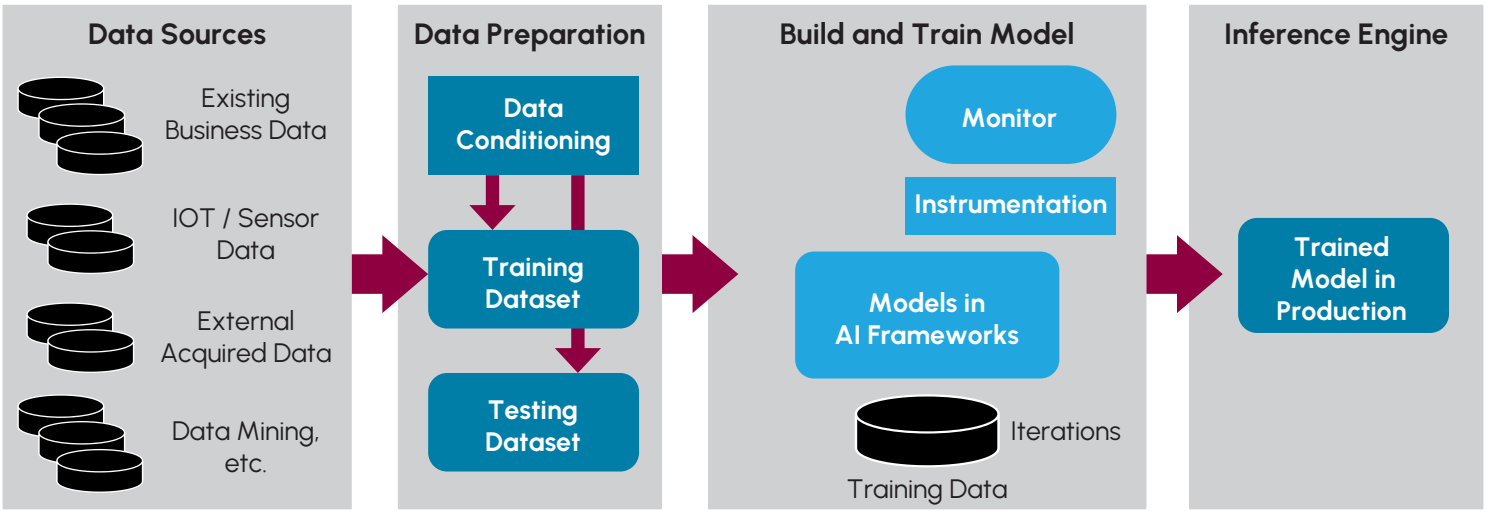


Figure 1. *Typical Data Flow for AI*

AI environments have primarily been developed on public cloud platforms because of the cost of building extensive model environments. Now, enterprises are bringing foundational models into their more modest environment and applying private data for further training and tuning. Stated differently, enterprises are using existing models from different providers as foundational models and then applying their private data on-premises to create a customized model. Generative AI Retrieval Augmented Generation (RAG) is an example gaining much traction with enterprises. RAG provides this customization with use of newer, custom or proprietary data, which enhances Large Language Models with greater accuracy, currency and relevance. The amount of training or tuning required varies depending on the foundational model selected.

Because of the prevalence of AI in public clouds, data science teams have the majority of their experience from that environment. The importance of selection of storage systems for performance, availability, and protection has not been a focus area in the cloud environment. Thus, IT must understand the needs, evaluate solutions, and communicate effectively the value from the choices made for on-premises deployments. Private data use requires the knowledge of the characteristics of where the data is stored and how it is accessed as well as the governance requirements including those for data protection and security.

Most organizations start AI projects in the cloud because of the cost of building extensive model environments and the ability to quickly get started. The cloud consumption model is also attractive as compared to as-needed capital investments. As an organization's AI initiatives grow, the data and cost of cloud infrastructure at significant production levels can become prohibitively expensive. Consistent availability of resources, security and governance concerns can also grow with ongoing use of AI in the cloud. These factors also motivate enterprises to bring AI into on-premises environments, or run a combination of cloud and on-premises resources. There are also innovative infrastructure and storage-as-a-service offerings that enable organizations to consume on-premises storage with an as-a-service cloud-operating model, providing a financial alternative to traditional capital investments.

Characteristics for AI Storage On-Premises

Data used for creating AI models comes from unstructured data sources as well as structured data in relational databases. The unstructured data is usually in file format but may also include data stored as objects. Data for conditioning by the data engineers may be in a data platform used for AI/machine learning (ML) and business analytics or may be in a custom-built arrangement using open-source tools. The storage used in a data platform is usually in a data lake or lakehouse. The storage system must support large scaling of capacity with performance to meet the usage demands

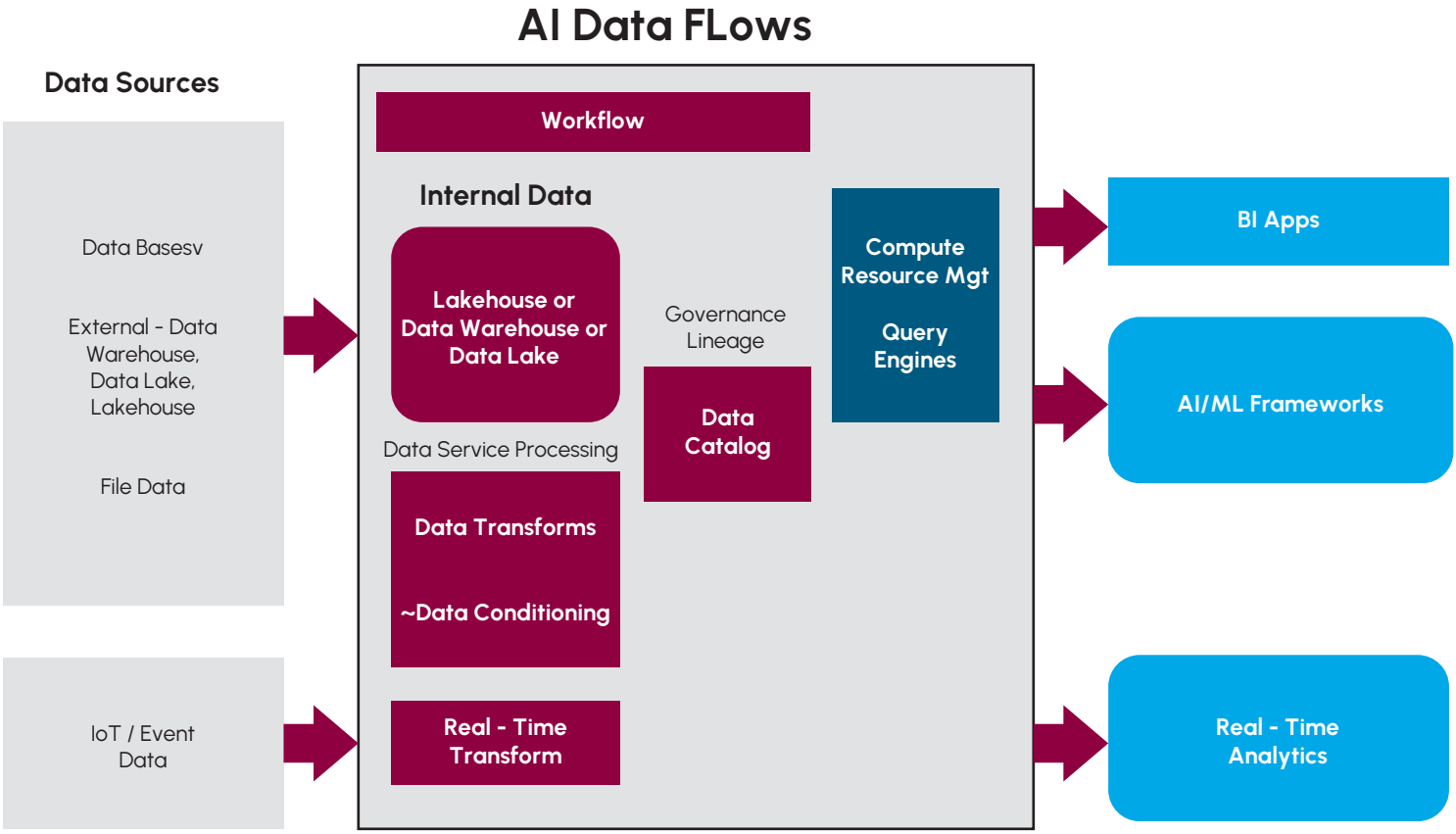


Figure2. *Data Platform for AI/ML antd Business Intelligence*

Data engineers create training and validation datasets as input to the AI frameworks for generating trained models. The performance focus of the data science team is on the aggregate memory and number of GPUs along with the bandwidth and latency in providing data from training datasets. As technology developments progress with AI framework software and GPUs/CPUs, the performance for storage must continue to excel and not be the element perceived to be delaying model delivery.

The type of storage required may change as the AI environment evolves or matures. Greater proficiency and continued improvement will increase both capacity and performance demand for providing and storing data. In general, as is the trend across IT infrastructure, all-flash storage should be the foundation on which AI storage is built. Consistency of storage performance as provided with all-flash storage, is expected. The performance, longevity, reliability, and efficiency needed for AI mirror those from the most demanding IT workloads where all-flash has been the primary solution. If the AI environment is segmented based on the stage of maturity, the following characteristics apply:

- Initial/maturing AI environment: High-performance file storage along with object storage
- Production AI environment: Large-scale capacity file and object storage along with continued high-performance file storage

IT infrastructure teams are concerned with the characteristics of storage systems while AI platform architects have expectations given previous experience with public clouds. Important storage system characteristics of note are performance, reliability, data protection, and scalability to which simplicity, scalability and cost should be added to give a broader picture for use in AI. Table 1 summarizes the characteristics in these areas.

Table 1. *Important Storage System Characteristics for Use in AI*

Characteristic	Considerations
Performance	Predictable and consistent performance for AI workloads is valued. Variations where different, competing workloads are accessing storage are not expected. AI inference and training requires storage that is capable of providing low latency and high-bandwidth performance, best delivered with all-flash storage for file and object data.
Reliability and Data Protection	Built-in capabilities to protect data from individual element failure and location/site failure are expected such that there is no loss of data or access.
Security	Employ best-practice security for information in the storage system
Native Support/integration for K8s architectures	As Kubernetes is usually the default architecture/platform for modern AI/ML workloads, data management and storage systems should offer native support for this architecture
Accelerate ML Ops [w/Self-Serve Access]	Data Scientists and ML Engineers can access storage, vector databases, and ML services in a self-service, declarative fashion, accelerating ML model training and deployment.
Scalability	Systems with the ability to non-disruptively increase capacity and performance as the amount of data required increases alleviate issues as models continue to evolve and more training data is added.
Simplicity	Simplicity in configuration and operation reduces the effort involved, improving the time to delivery for AI projects. Simplicity in operations includes no requirement for complex tuning to achieve optimum performance across all platforms.
Cost	As AI expands and models scale, data storage becomes the major cost element. Cost for storage should scale with capacity demand without sacrificing other capabilities as the demand increases.
Power	Power is limited within racks and data centers, and GPUs create another level of demand. By utilizing power-efficient, yet high performance storage, more power is available for more GPUs in a limited amount of rack and data center space.

Pure Storage Solutions – Value to Customers

The different stages of AI maturity require a variety of storage capabilities.. With the different characteristics identified, Pure Storage offers a consistent data storage platform that fits the usage required for AI and supports the growth of that environment as it evolves. Information about Pure's data storage platform for AI solutions is available at this link. Additionally, as a major provider of storage for enterprise organizations, IT operations will have confidence in deploying Pure Storage systems in an AI environment.

- For the initial AI environment, the file and object storage system FlashBlade//S and //E would be the primary offerings to consider. Information about FlashBlade capabilities is available at this link. In addition, some data such as databases may be stored on unified block and file storage systems and FlashArray would be optimal for consistent performance and availability. Information about FlashArray is available at this link.
- When AI environments would be considered in production with processes and procedures for continued model development, training, and tuning, the FlashBlade//S scales and provides the performance while FlashBlade//E addresses capacity needs as a cost-effective answer for retaining ever larger amounts of data. The FlashBlade//E economics are described at this link.
- Software development by data scientists for AI is container-based, using Kubernetes to manage the workload execution. For this development, container native storage offered by Portworx from Pure Storage provides storage and data services specific to the Kubernetes development environment. Information on Portworx by Pure Storage capabilities for use with Kubernetes is available at this link.

The value from a storage system for customers goes beyond the ability to store and retrieve data and includes other selection criteria in an evaluation. For use in AI, many of the same value propositions for storage in IT apply but there are some areas where there is additional focus:

- The ability to scale performance and capacity independently as the AI environment grows is valuable for IT when they do not have to change or replace systems and infrastructure. AI engineers and data scientists benefit from being able to accelerate model training and inference without interruption, shortening the turn-around times required for AI workflows. Reconfigurations and moving data are very disruptive and consume staff time. FlashArray and FlashBlade have the ability to deliver the non-disruptive scaling necessary for growing AI environments, meeting a highly valued requirement.
- The portfolio of all-flash scale-up and scale-out storage arrays from Pure Storage provide multi-dimensional performance, efficiency, and future-proof storage that enables IT to maximize the operational efficiency of AI data pipelines and avoid technical debt. Pure Storage has been the pioneer and continues to be a leader with a technology update program termed Evergreen. The technology update offered by Pure Storage allows customers to continue to keep pace and avoid having museum quality systems in use. This functionality becomes especially important in rapidly developing/performance demanding areas such as AI.
- A combination of the scaling and performance capabilities of FlashSystem//S provides the opportunity for IT to consolidate multiple sources of data in the AI environment to reduce the number of connections for data sources. This consolidation is possible when those data sources are, or could be, on-premises – the consolidation simplifies operations – both from managing the different connections for data and for ensuring protection of the data sources with Pure Storage data services such as replication and snapshots.
- Pure Storage DirectFlash Modules and custom chassis, managed by intelligent built-for-flash software, delivers impressive density and energy savings, resulting in less power consumption than other solutions. The Pure Solution consumes less than 1W/TB power consumption and 95% less rack space which provides greater capacity, better energy efficiency and a smaller footprint that leaves more room and power for GPU's to drive larger AI workloads.

Summary

AI presents new challenges for IT. There are new demands on the infrastructure for storing and managing information. Many of these challenges are new or at least different for IT. Operational procedures and the criticality of storage in an IT environment may also be unfamiliar to AI platform architects whose experience for AI comes primarily from work in public cloud environments. With this understanding, organizations must make critical IT choices for deploying and evolving an AI environment.

The types of storage supporting specific characteristics are probably the most important selections. Pure offers a data storage platform that accelerates adoption of AI with the necessary capabilities for early-stage deployments through the evolution to a mature AI production environment. The requirements for performance, efficiency, reliability, data protection, scaling, and simplicity across a consistent product line to meet different usage and cost needs are available from Pure Storage. Importantly, Pure Storage systems provide value to customers in rapid deployment and use, simplicity of operations with no complex training and maximum efficiency, and advanced capabilities like multi-dimensional performance and multi-protocol access. AI infrastructures are mostly containerized, with Kubernetes the most popular orchestration framework. Portworx unifies the Pure platform and makes managing stateful applications in terms of storage classes and service levels simple.

The most compelling factor for use of Pure Storage in AI environments is the time to realization of an AI environment capable of developing or adding to a trained model with enterprise private data. The Pure Storage systems have the characteristics for high confidence in minimizing the time required for delivering the storage infrastructure.

Important Information About this Report

CONTRIBUTORS

Randy Kerns
Senior Strategist and Analyst | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

PURESTORAGE®

ABOUT PURE STORAGE

Pure Storage, Inc., based in Santa Clara, California, develops all-flash data storage solutions. Founded in 2009, it is transforming the storage experience and empowering innovators by simplifying how people consume and interact with data.

The Futurum Group

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

The Futurum Group

CONTACT INFORMATION

The Futurum Group LLC | futurumgroup.com | (833) 722-5337 |