WHITEPAPER

OPTIMIZED ARCHITECTURES

Designing Storage Architectures for Big Data and AI/ML Workloads

CONTENTS

Introduction

A thriving enterprise data economy is built on three imperatives: capture everything, manage it efficiently, and leverage its potential.

Artificial intelligence (AI), machine learning (ML), and other big data applications and systems that generate and use massive amounts of data are rising rapidly. According to recent surveys, well over half of enterprises say they are using ML today, and nearly all will within a few years. Big data and AI/ML workloads necessitate the ability to process and analyze massive volumes of data, both structured and unstructured; meanwhile, with hybrid cloud and multicloud strategies resulting in multiple locations for said data, including on premises, off premises, and edge, enterprises must reconsider numerous issues around data management, current and future storage capacity, and efficiency to define an architectural *sweet spot* that can manage these data sources and applications.

Rethinking your data strategy requires the right data architecture at its core. An ideal data architecture is optimized for infinite scale and efficiency, so you can capture everything: the data you need and the data you didn't know you needed. It's an architecture optimized for frictionless movement of your data, so you can get out of its way and move it to the right place at the right time—ready for your organization to leverage its full potential wherever it generates the most value.

But getting there is difficult, due to resource scarcity and limitations of current technologies and solutions. This paper considers hybrid cloud and storage architectures for enterprises seeking scalable and cost-effective IT infrastructure appropriate for big data and AI/ML workloads.

The Weight of AI and ML on Data Architecture

AI and ML are reshaping virtually every facet of modern enterprise operations. Natural language processing (NLP) and image recognition affect how consumers and companies alike interact with machines. New smart systems are reimagining manufacturing with the help of AI-aided solutions. Autonomous vehicles promise fundamental shifts in transportation and shipping. What's more, many organizations are experimenting with not just one or two, but many of these new and disruptive systems simultaneously.

According to Deloitte's State of AI in the Enterprise Report[1], 67% of respondents are using ML today; a full 97% are using or plan to use it within the next year. IDC projects[2] that by 2025, 41 billion IoT devices will generate more than 79 zettabytes (ZB) of data annually, and in the recent Seagate Rethink Data report, IDC notes 44% of data created in core data centers and at the edge will be driven by analytics, artificial intelligence, and deep learning[3]. The problem of unchecked data growth is not new, but it is expected to increase about 25% in 2020 alone, according to enterprise IT innovation firm 451 Research[4].

AI/ML is paving the way for data scientists and analysts to unlock previously hidden or unextractable insights within massive datasets. But it has also created new challenges for IT architects, who must consider how to maintain data availability, integrity, confidentiality, and durability at massive scale—while simultaneously optimizing for cost, governance, and compliance. Data sprawl, replication, and divergence across multiple locations are just a few of the challenges associated with this evolving landscape.

In order to mitigate these common problems and more efficiently harness the power of an ever-growing tidal wave of data, IT architects must consider several factors, including tiered storage mechanisms, evolving and emerging fabrics, and open-source orchestration tools, to strike the architectural sweet spot.

Solving for Increasingly Complex Data Management Needs

The cloud delivery model, which emerged more than a decade ago, opened the floodgates for a huge range of public and private platforms, as well as for manifold complexities regarding which cloud or data center resources fit specific workloads. No solution is uniform, and many even vary from workload to workload within a single enterprise.

Enterprises must determine what on-premises and private cloud systems and capabilities can complement public cloud storage services, as well as adapt to non-cloud business applications on an as-needed basis. The rise of AI/ML for both business operations and customer-facing services will necessitate that enterprises take a future-forward approach to data center and storage decisions. Data centers and storage architecture have evolved significantly over the past decade, reflecting increased reliance on hybrid and tiered storage implementations and the ascendant importance of disaggregation and composability.

Today most enterprises opt for a mix of public and private platforms as they seek to optimize scalability, data availability, performance, security, and cost. According to 451 Research's 2020 Cloud Confidence Report[5], more than half of enterprises (57%) are moving toward a hybrid IT environment that integrates both on-premises systems and off-premises cloud/hosted resources in an integrated fashion.

The 451 Research report provides a breakdown of common infrastructures being employed by enterprises today and in the future:

- **49%** of respondents utilize on-premises, non-cloud infrastructure today; **25%** plan to do so in the future.
- **36%** are implementing/plan to implement on-premises private cloud both today and in the future.
- **32%** are turning to Software as a Service (SaaS) and hosted applications today; **37%** plan to do so in the future.
- **31%** and **32%** respectively are using/will use hosted private cloud today and in the future.
- **28%** are relying upon Infrastructure as a Service (IaaS) public cloud today, and **33%** plan to use it in the future.
- **22%** are relying upon Platform as a Service (PaaS) today; **29%** plan to integrate it in the future.
- **15%** are using/plan to use a hosted, non-cloud environment both today and in the future.

Defining Storage Architectures for Shifting Priorities

With the cloud and its many manifestations evolving constantly, storage considerations are also in flux. Optimized storage tiering ensures that data is accessible in the right place and at the right time for a given application, analysis, or workload.

Historically, the best media for capacity has always been the least efficient for performance, and vice versa. From an economic standpoint, architects designing for AI/ML must ensure that the small amount of very frequently accessed data is as high up in the tiering as warranted, and that the very large amount of infrequently accessed data is as low down the pyramid as possible.

Tiered, hybrid storage that incorporates both HDD and SSD technology has become the de facto norm. Not only does this approach allow for overall better system performance of both HDD and SSD, but IT architects can cherry-pick components of each to reach the ideal combination of price, performance, and capacity.

Typically, mass-capacity storage devices will be a central component of data centers fielding AI/ML workloads, primarily for capacity and bandwidth purposes. But AI/ML applications may also require a small amount of non-volatile media devices for IOPS-intensive workloads. Ultimately, a blend of solutions may help handle unexpected variable demands and bursts of compute, as well as improve cost efficiencies.

Another important factor in the evolution of the data center is the continuing trend of disaggregation—not only of storage, but also emerging within the context of memory, CPU, and GPU. At the most basic level, composable architecture encompasses the ability to allocate a specific type, configuration, or number of physical components to create virtual machines that are custom-tailored for specific IT requirements.

Disaggregation also means that IT architects are no longer beholden to the closed architectures defined by CPU and motherboard makers. With a large, open-source community developing solutions for both software and hardware, easier data management may finally be within reach.

Ultimately, composable disaggregated architecture underpinned by an appropriate fabric, such as NVMe-oF, and supplemented with open-source orchestration tools is likeliest to provide the necessary flexibility for enterprises operating across a variety of cloud environments and types of data centers, as well as for those implementing a wide range of AI/ML applications.

Considerations and Solutions for Optimizing AI/ML Workloads

Proponents of the public cloud tout its many benefits—it's dynamic, user friendly, and capable of housing large amounts of data. Theoretically, enterprises only pay for what they need. But in the reality of AI/ML use cases, the public cloud can quickly become exorbitantly expensive. It's the *Hotel California* problem: "You can check out, but you can never leave." In other words, data extraction and retention at the scale required of a successful AI/ML application is often cost prohibitive.

In the private cloud, hybrid cloud, or multicloud environments, the bottom line once again comes down to the ability to store, harness, and retrieve massive amounts of data. The foundation for a robust and cost-effective architecture capable of enabling smooth AI/ML integrations begins at the storage device level.

Devices must have much more capacity than ever before. Software compatibilities are thus also imperative: Software must be updated to work in tandem with very large individual devices, which can sometimes fundamentally break assumptions—both implicit assumptions as well as explicit coding patterns. Software must also be consistently updated to search across data centers and allow users to consolidate more of their data in a single namespace.

Enterprise data centers will need large, mass-capacity devices and very scalable systems. An HDD-centric approach is apt for many AI/ML workloads; it functions well both for the human/data scientists' user experience, as well as for the labeling process and a small amount of random reads.

For AI/ML workloads involving very large amounts of random I/O, the focus must also expand beyond device capacity. While an HDD-dominant storage architecture will be suitable for many requirements—including searching metadata, bandwidth, and, of course, capacity—for IOPS-intensive workloads, it will be necessary to combine traditional storage with nonvolatile storage devices that are word-addressable—as opposed to simply 4K block-addressable.

This can, however, resurface the issue of data sprawl, data replication, and data siloing. To mitigate this, the data center must support an ability to leverage the appropriate media without difficulties in terms of locating data.

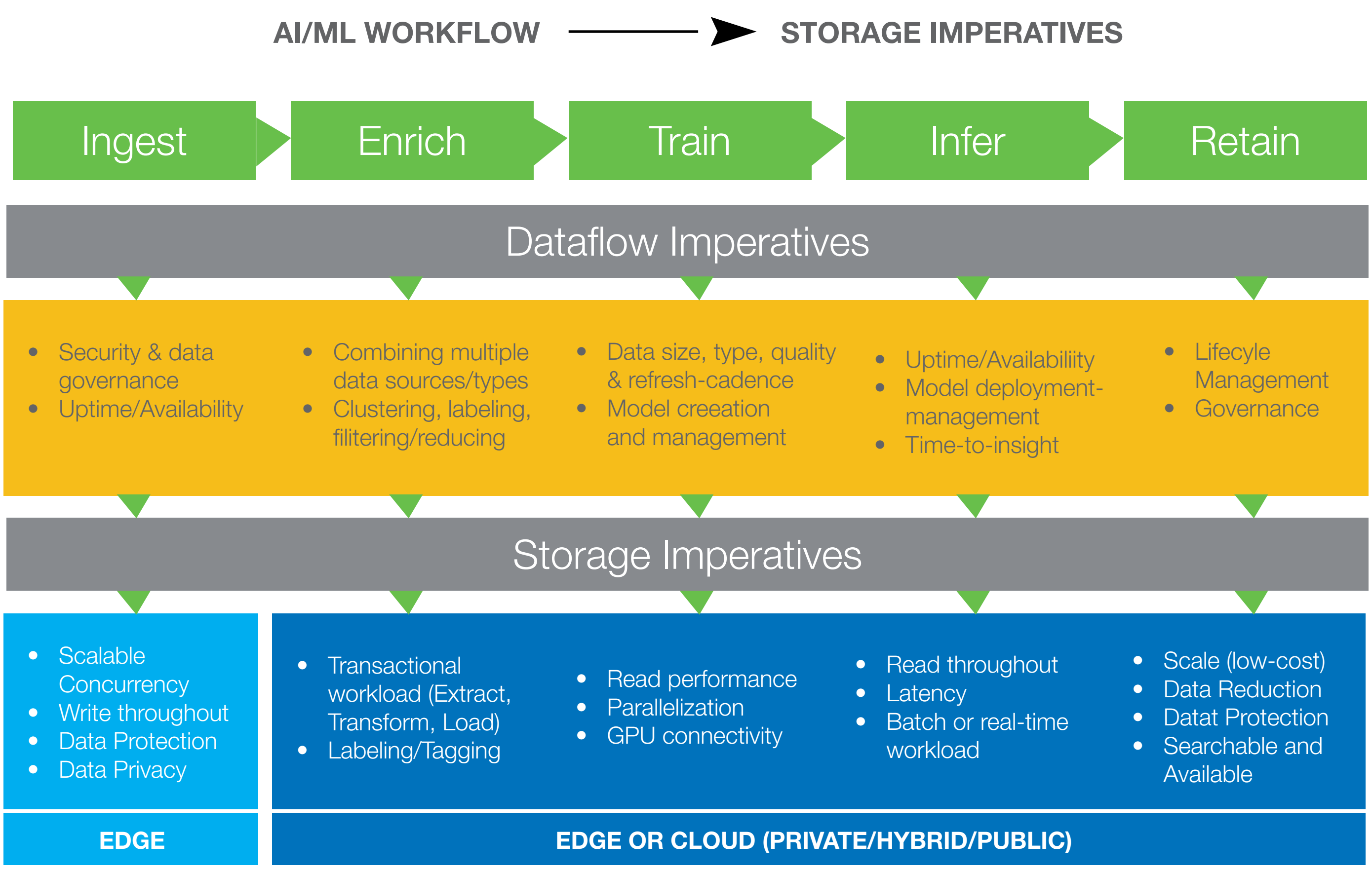
Different Dataflow Imperatives for Different Phases of the AI/ML Workflow

AI/ML workloads do not progress linearly; they might be more appropriately termed *workflows* than workloads. It is not a simple matter of connecting the dots between the broad stages during which data is in motion—from data collection to decision to short- or long-term storage.

Architectures and data management needs will vary dramatically based on the functionality and directives of AI/ML systems. A basic image recognition system will, for instance, have much different needs than that of a fleet of autonomous vehicles.

While there is no one-size-fits-all recommendation for AI/ML workflows, they can be broadly broken into phases, which can help IT architects make key decisions regarding the appropriate fabric, orchestration software, and storage solutions.

AI/ML workflows typically include data ingest, enrichment, training, inference, and retention stages. IT architects should consider the dataflow and storage imperatives of each one of these phases.



At the ingest stage, the primary concerns from a dataflow perspective are security and governance, availability, and uptime. In this phase, the source of the data becomes paramount in decisions about infrastructure and storage. An IoT application, for instance, will need to aggregate very small files coming from very disparate sources. The data is likely to have different formats—image data will have one format; a sensor's time series data will have another. The infrastructure must support taking a massive flow of information from different sensors and making sense of it.

At the enrichment stage, before it can be used for AI training the data must be combined from multiple sources and from among multiple types, and will then be clustered, labeled and filtered, and reduced to create a subset of the most relevant information. It's imperative at this stage that storage systems can manage a transactional workload involving the extraction, transformation, and loading of the data while tagging it and labeling it for the next stage.

Next, training and inference represent distinct workloads. Training requires taking data—sometimes a great deal of it—and using that information to train a model. Once a model is established and deployed, perhaps on an edge device, the inference phase begins. This requires using real-time data to generate insights and make decisions. Here, time-to-insight, uptime, and availability are particularly critical requirements.

The retention element is another animal entirely—data may need to be retrieved for retraining, or it may need to be archived or stored for months, or even years, for regulatory and compliance purposes. This phase is perhaps the thorniest territory when it comes to cost. While using the cloud for data ingest and training is one thing, data extraction or repatriation into the private cloud can get expensive quickly—à la the *Hotel California* problem. Data retention requirements, in fact, will be central to the thought process around other architectural components—or else the solution may become expensive down the line.

The challenge with AI/ML is that each of these phases has relatively unique requirements. Below are two sample use cases that illuminate some of the benefits and trade-offs architects need to consider when designing for different workloads:

Use Case: Smart Manufacturing

Typically, the data requirement for training models for a smart manufacturing application isn't that high. The dominant aspect of the architecture in this instance comes down to the data scientists' time.

At the training stage, the primary concern is iterating data through the training process as quickly as possible. Using the cloud for training may be attractive; elasticity of the cloud allows scientists to spin up as much compute as they want, very quickly.

On the inference side, however, an on-premise choice becomes better for several reasons. Efficient factory operations necessitate real-time dataflow, and there are significant issues with relying exclusively on cloud architecture for that. One is uptime: If the cloud is fueling inferencing, or decision-making, and the link to the cloud is lost for any reason, the factory goes down. On-premise infrastructure enables better control over destiny (at least as it's relative to hardware). In addition, the compute requirements associated with inferencing are typically much lower—and therefore the cost of that hardware is lower.

Use Case: Autonomous Vehicles

This use case has a much higher data requirement.

Looking at the ingest stage, the economics of getting data—including real-time video—into the cloud is very challenging.

Vehicles will also need to constantly recirculate data to identify and learn from mistakes. Inference hardware in an AV system is typically located in the vehicle itself, at the edge. (There is still some debate as to whether training data should be in the cloud or on the edge, simply because of the sheer volume involved in this phase.)

In this use case, moving data along the wire can get very expensive and time-consuming, given the bandwidth between an edge location and a cloud or a colocation. Physical, mobile arrays of devices at the edge can help solve for this: Vehicles may contain a flash-based solution or an SSD on the edge for data aggregation; data may then be transferred via shuttle devices and moved to the data center for more efficient ingestion.

For any given use case, it's critical to understand the whole workflow relative to the amount of data that needs to be both ingested and retained, the time-to-insight on the inference side, and the cadence required for re-training models. It's a mistake to look at any one of these components in isolation.

Choose an Object-Based Software Architecture

One commonality shared by all AI/ML systems is the need for data to be modified in a meaningful way, whether by clustering, filtering, reducing, or labeling data. ML sets cannot be trained on raw images alone; there must still be some element of human oversight for initial labeling.

For these reasons, it's critical that software architecture has excellent support for labels, in addition to being economically efficient for mass capacity. Object-based or object storage, which spans private and hybrid and public cloud, is the best interface for economic storage of large amounts of unstructured data, making it an appropriate abstraction for AI/ML. Companies with data-intensive AI/ML systems are increasingly moving toward object storage because of its fundamentally more scalable nature, as well as apt underlying data structures, abstractions, and assumptions about workloads.

Object data is also a very powerful tool for gleaning information from metadata. Once the data is stored in an object storage system, it can be more easily found and filtered; there's a richness to the ability to work with data that is very compatible with HTML.

Future Fabrics Will Facilitate More Seamless AI/ML

AI/ML machines typically require large pools of DRAM—but today's maximum capacity for DRAM in a single server is about two terabytes. Revisiting the autonomous vehicle example: If a car returns from a day's worth of driving with six terabytes of data, it must be spanned across multiple servers, which ultimately results in a hyperconvergence situation.

Today it's easiest to put that data into disaggregated memory to be shared across all the CPUs actively working on the system. The CPUs can share in access, paging back and forth from DRAM to storage-class memory. But there's a cost to moving back and forth. In the future, new, low-latency fabrics may free up memory and put it in a pool that's very close to those devices, and share it across devices. This evolution very well may be the next step in faster, more capable machine learning.

This iteration, however, is likely still several years away. New fabrics are very complex to create. The Gen Z evolution from CPU-centric computing to memory-centric computing has been ongoing since about 2012. Within the next ten years, it's likely that Gen Z fabrics will become a standard part of the ecosystem, as well as part of the hypervisor and some enterprise solutions. CXL 2.0 and 3.0 may start showing up in labs by mid-2025—and in the following decade, CXL 2.0 and 3.0 may start being deployed in data centers.

Open Environments Are Critical to Flexibility, Security, and Optimization

Open environments further facilitate flexibility and opportunity for innovation and growth. This has long been true in software, and it's becoming true for hardware too. It will be particularly relevant as enterprises continue to experiment with AI/ML technologies.

Open architectures include integrated collections of composable compute, networking, and storage resources. Scalable, open hardware infrastructures make it easier to collaborate within and innovate upon existing systems. Peer review enables an open source project to be examined from different lenses; it ultimately accelerates optimization, as many different authors with shared use cases and goals can come together to debate and identify solutions, security vulnerabilities, and more.

Today IT professionals are slowly moving toward hardware architectures that can be scaled at a high level, and that are less dependent upon vendor-unique components. Groups like the Open Compute Project (OCP) collaborate to debate and ultimately identify ideal solutions—for instance, the height or length of a chassis—thus democratizing the process by which hardware standards evolve.

Open source—from both a software and hardware perspective—can help enterprises manage the fluctuating data flow imminent with AI/ML and other big data applications. Working on a data set at scale can be debated in the open, and architects can share common problems plaguing optimization efforts.

Open standards hardware may also ultimately improve the accountability and security of hardware architectures. It allows systems to be designed from the ground up with security top of mind. RISC-V International, for instance, has created an open standard for next-generation processors with native security capabilities.6 Open-source implementations of those RISC-V processors can be evaluated for security and used as a baseline for more secure system designs. OpenTitan7 is another example: the open-source, silicon-based Root of Trust (RoT) project is the first of its kind to develop a transparent, high-quality reference design and integration guidelines for a RoT. Open-hardware standards and open-source hardware projects, such as RISC-V and OpenTitan, allow us to rethink hardware security for data storage.

On a more granular level, open sourcing object storage software that is appropriate for AI/ML workloads—and especially software that is optimized for the mass-capacity devices that are necessary for them—is important for two primary reasons: cost and flexibility to handle hardware idiosyncrasies without sacrificing efficiency.

Conclusion

Advancements in AL/ML are simultaneously unlocking opportunities for innovation and presenting challenges for data management and storage. The increasing number of cloud and multicloud environments available further exacerbates the complexities that must be considered when designing for AI/ML workloads.

While a tremendous amount of data is already being created and will continue to mount over the coming years, a relatively small amount of that data is being stored, typically due to cost constraints. While on the software side of the equation, advancements in autonomous management via systems like Kubernetes allow applications to scale out and back down on an as-needed basis, the hardware world has not traditionally been as flexible.

The emerging trends of composability, disaggregation, and open source/open standards may help shift the tide. In a composable future, pools of storage, memory, and compute should be able to scale as necessary to complement ongoing software innovations. In the context of AI/ML, architects will be more able to implement the exact ratio of servers and storage necessary for performance and capacity, versus considering servers as an availability path to devices.

Seagate understands how to architect for the whole spectrum of hybrid cloud options and blends. Learn more about solutions tailored to your enterprise's AI/ML needs.

References

1. **https://www2.deloitte.com/content/dam/insights/us/articles/6462_state-of-ai-in-the-enterprise/DI_State-of-AI.pdf**

2. **https://www.idc.com/getdoc.jsp?containerId=prUS45213219**

3. **https://www.seagate.com/our-story/rethink-data/**

4. 451 Research, “Balancing Hybrid Storage as Part of a Hybrid Cloud Data Management Strategy”

5. 451 Research, “Cloud Confidence Report,” March, 2020

6. **https://riscv.org/**

7. **https://opentitan.org/**

Resources

High Performance Computing Solutions: **https://www.seagate.com/solutions/high-performance-computing/**

Big Data Analytics Solutions: **https://www.seagate.com/solutions/data/big-data/**

Ready to Learn More?

Visit us at **seagate.com**

SEAGATE

TP716.1-2009US, September 2020