HAMMERSPACE

ERA

WHITEPAPER

Revolutionise your AI Workloads with ERA and Hammerspace

Leveraging Non-Disruptive Data Orchestration Across Multi-Vendor Storage Environments to Lower Costs, Ensure Data Governance, and Accelerate AI/DL Pipelines

Executive Summary

Seeking order out of data chaos

Imagine being the maestro of a world-class orchestra trying to create a beautiful symphony with musicians who are seated in separate auditoriums. Each musician holds one part of the masterpiece, but the walls that separate them make it almost impossible for the conductor to produce the combined harmonies needed for their performance to succeed.

For modern IT organisations in public and private enterprises of all sizes, digital assets play the role of these separated musicians. This is particularly so with unstructured data such as images, audio, text, or other files that don't fit neatly into traditional, structured databases.

The problem is that for many years unstructured data has grown much faster than structured data has, and makes up over 80% of the digital assets in all verticals. What's worse, unstructured data is typically dispersed across the edge, in multiple on-premises and cloud-based storage silos from different vendors, and often across multiple geographic locations. The issue is that unstructured data is growing faster than can be meaningfully analysed or utilised. So instead of adding value to organisations, unstructured data often becomes a burden and a steadily growing cost centre they must manage and store with little to no return on investment.

Organisations simply can't afford to throw out existing infrastructure and migrate their unstructured data to a new platform in order to implement an AI strategy.

80%

The Problem

Unstructured data has grown much faster than structured data and makes up 80% of the digital assets in all verticals becoming a burden and a steadily growing cost centre.

HAMMERSPACE

The use of data analytics, BI applications, and data warehouses for structured data is a mature industry, and the strategies to extract value from structured data are well known. But the emerging explosion of generative AI and related deep learning (DL) technologies now hold the promise of extracting hidden value from unstructured data as well.

Such AI-inferencing workloads will not only help data owners figure out what they have and what they should keep or can discard, but AI/DL use cases also hold the promise of helping enterprises create new outcomes by gaining insights previously hidden in large volumes of unstructured file data. In this way organisations will finally be able to utilise both their structured and unstructured digital assets to create new business value as well as to drive efficiencies.

Silos, data governance and other problems with AI workloads

The problem is that the barriers separating unstructured data silos now become a serious limitation to how quickly enterprises can implement AI pipelines without costs and complexity spiralling out of control. They need the flexibility to use any or ALL of their data to feed AI/DL workflows, which traditionally has meant consolidating files from different resources into a unified repository.

In addition, AI application models are rapidly evolving and may require different subsets of data over time. This not only creates an operational problem of needing to copy data from silo to silo, but it also creates serious data governance problems, with added risks for compliance, proper access controls, auditability, and ensuring data integrity as copies proliferate.

For AI strategies to succeed within reasonable costs and timeframes, organisations need the flexibility to break down data silos securely and with proper controls to get direct global access to the data where it is today.

One look at the technology industry trade press and you'll see that there is a feeding frenzy among storage vendors touting one-size-fits all solutions to address this problem. But organisations simply can't afford to throw out existing infrastructure and migrate their unstructured data to a new platform in order to implement AI strategies. Additionally, very few data environments are consolidated into a single storage silo to satisfy all phases of the data lifecycle. Existing siloed storage architectures were not designed for the cross-platform requirements of AI data pipelines.

HAMMERSPACE

Creating actionable structure out of unstructured data

In this white paper we show how ERA and Hammerspace solves these problems with a software-defined solution that automates high-performance unstructured data orchestration across existing decentralised storage silos from any vendor, and even across multiple locations and clouds. With its high-performance Parallel Global File System that can bridge data across silos of existing storage, Hammerspace is ideally suited to feed the many differing performance requirements needed at different phases of AI/DL workflows.

ERA & Hammerspace's high performance Parallel Global File System can bridge data silos on any existing storage, with the performance to solve AI/DL workflows at any scale.

We create structure out of otherwise chaotic and difficult-to-categorise data with advanced capabilities to automatically index file metadata across any storage silo. This metadata catalog can be enriched with user-generated or automated custom metadata, which can simplify cross platform data orchestration and data governance. Global metadata is the key to knowing which data you have, where it should be located, and which projects or programs it is connected with. Such metadata is made actionable in Hammerspace, and is an essential requirement to achieve adequate data governance and data quality in AI/DL pipelines.

In this way, ERA and Hammerspace provides an actionable structure for disparate silos of unstructured data wherever they are today, without the need for wholesale data migrations or to replace existing storage infrastructure.

Creating structure out of chaos

Create structure out of chaotic and difficult to categorise data

HAMMERSPACE

Create actionable structure out of unstructured data

By providing analytics, artificial intelligence, deep learning and machine learning workflows with unified access and automated control to all data on any storage type anywhere, ERA and Hammerspace help data owners not only leverage their existing storage resources, but also to dramatically accelerate AI workloads for distributed unstructured datasets. By eliminating the need to copy data to new repositories or to consolidate to a single location, we can reduce the time to inference from weeks to hours for large data environments, and in the process also significantly reduce costs associated with setting up an AI pipeline.

In addition, we'll show how the flexibility of ERA and Hammerspace architecture enables customers to break free of being vendor-locked into a single infrastructure solution that limits their ability to adapt to changing requirements. This is key, because not all steps in the AI journey have the same performance requirements.

ERA and Hammerspace helps data owners dramatically accelerate AI workloads for distributed unstructured datasets, even with their existing infrastructure.

HAMMERSPACE

With the scale-out/scale-up capabilities of Hammerspace's Parallel Global File System creating unified access to data across existing storage silos, this means Hammerspace enables data owners the freedom to pivot at any time to adapt to new AI use cases or technologies. No need to manage data copies to new platforms, or to disrupt data access for existing users or applications. All such changes are completely transparent to users as background operations.

For example, emerging techniques like LoRA (Low-Rank Adaptation) allow for fine tuning of existing models with much lower performance requirements than are needed by current technologies. So having the flexibility to pivot existing infrastructure without expensive over provisioning will provide significant improvements to ROI as AI technologies improve over time.

Navigating the AI Journey

There are multiple phases in AI workflows, and of course different AI/DL use cases can vary greatly depending on the industry or desired outcome. For unstructured data in particular, medical image analysis for disease detection will differ from activity recognition in video data, or sentiment analysis in text data to determine targeted ad placement. Inferencing workloads for analysing satellite imagery for crop yields or to drive decisions on irrigation or water management will differ from prediction models used on video and other sensor data for refining autonomous vehicle behaviour, or to streamline manufacturing automation.

But as diverse as the AI use cases are, the common denominator of them all is the need to collect data from many diverse sources and often different locations. In typical workflows this may mean largescale data movement with manual file migration tools such as rsync or other point solutions, especially to feed the high-performance requirements for HPC-style inferencing workloads that require serious computing horsepower.

File system fragmentation is the key problem

The fundamental problem is that access to data by both humans and AI/DL applications is always funneled through a file system at some point. That is, a file system organises the raw bits on the storage media into the files and folder structures that humans can understand and that applications need to access. This is done via the metadata in the file system, which is the interface between the raw data and the file structure that users/applications see.

The issue is that since the 1990s and the introduction of network-attached storage, file systems have been embedded within the storage infrastructure. Even though different vendors will present the file/folder structure via industry standard NFS or SMB file access protocols, the underlying file systems that contain this metadata are siloed into vendor-specific variations that are incompatible with each other.

The result of this storage-centric approach is that when data outgrows the storage platform it is on today, or if different performance requirements or cost profiles dictate the use of other storage types, users and applications must navigate across multiple access paths to incompatible systems to get to their data.

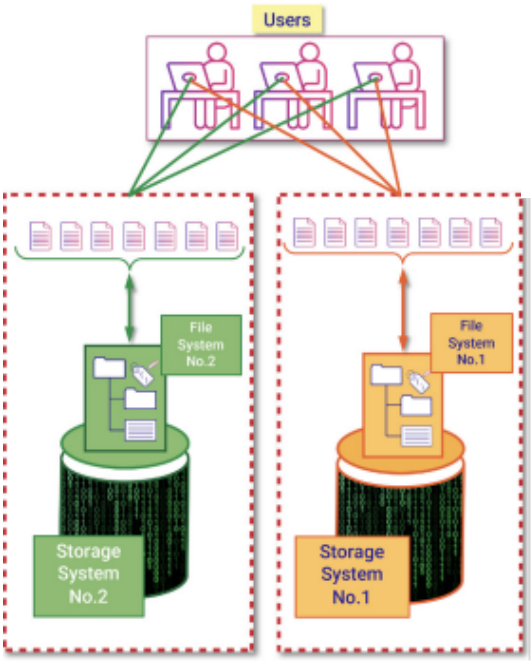


Fig 1

Fragmented data becomes out of control

Over time, and as volumes of unstructured data have exploded to span multiple silos, locations, and the cloud, this problem of fragmented data has gotten out of control. Indeed, bridging these silos, sites, and clouds has spawned an industry of point solutions solely dedicated to data migration, file copy management, the use of tiering solutions or cloud gateways, and other techniques to overcome the fragmentation of data access and control across multiple file systems within storage vendor silos.

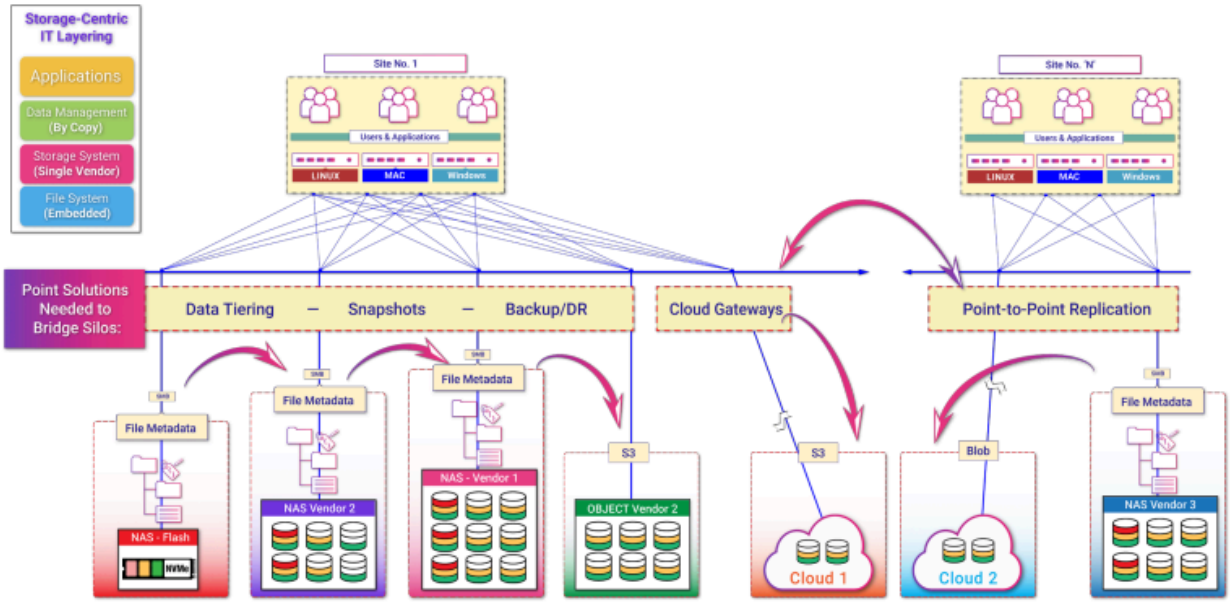


Fig 2 - Storage silos caused by multiple file systems fragment access for users and applications, and require multiple point solutions to copy data from silo to silo. This adds cost and complexity to AI workflows where global access to consolidated datasets is required.

Silos are even worse for AI/DL workloads

This problem is particularly acute for AI/DL workloads, where a critical first step is to consolidate data from multiple sources to enable a global view across them all. AI workloads must have access to the complete dataset to classify and/or label the files as the first step to figuring out which of them should be refined down to the next steps in the process.

With each phase in the AI/DL journey the data will be further refined. This might include cleansing, classifying & labeling, and eventually large language model (LLM) training and tuning. Each of these steps have different performance requirements for compute and storage, ranging from slower, less expensive mass storage systems and archives, all the way to high-performance GPU clusters with NVMe storage.

Silos are even worse for the AI/DL workloads

The problem for data owners is to figure out how to accommodate the multiple performance requirements with a single system. That is, how to manage both the data classification and/or labeling steps, which do not require high performance, and then also to feed GPUs for training/tuning and then inferencing, which typically are high-performance workloads that need NVMe storage.

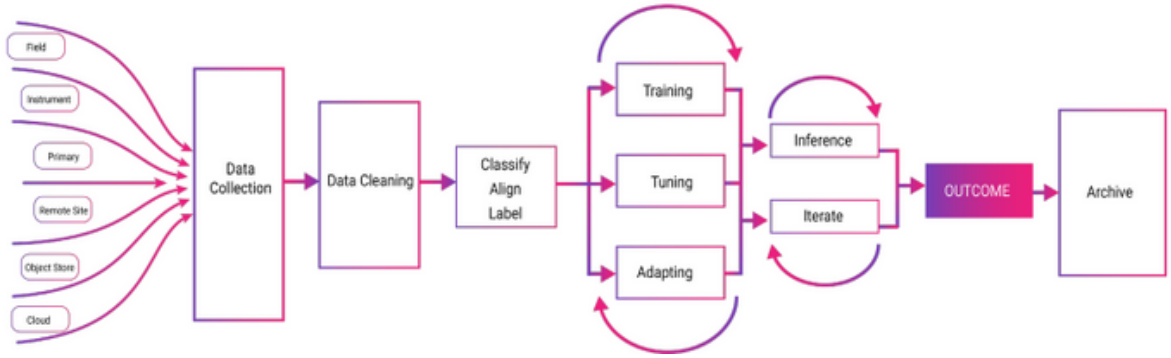


Fig 3 - AI Pipelines pull from multiple data sources, and then proceed through multiple steps, each of which have different compute and storage requirements.

Organisations are faced with a Hobson's choice of either over-provisioning their infrastructure with enough high-performance storage so all the data can be in one place for all phases of the AI journey, or to pay the 'data copy tax' of shuffling file copies between storage silos, and thus increasing the time to outcome. When data is separated across multiple sites or the cloud, this copy penalty becomes even worse, and may also result in expensive GPUs or other HPC systems sitting idle waiting for data to be copied into high-performance storage to begin the processing runs.

What makes this all the more painful is that organisations already have a significant investment in existing infrastructure, such that it is cost prohibitive to simply replace existing systems with a new vendor-locked dedicated storage platform that can handle all performance requirements. Moreover, AI technologies are advancing so rapidly that locking into one storage solution that works with today's AI pipelines may prevent organisations from taking advantage of new, emerging technologies that may be better suited to their use cases.

Either way, whether investing in a new infrastructure, or adding the complexity and delays of the 'data copy tax', there is a significant added cost that makes the calculus of whether there is a true return on investment for the AI journey very difficult to answer.

HAMMERSPACE

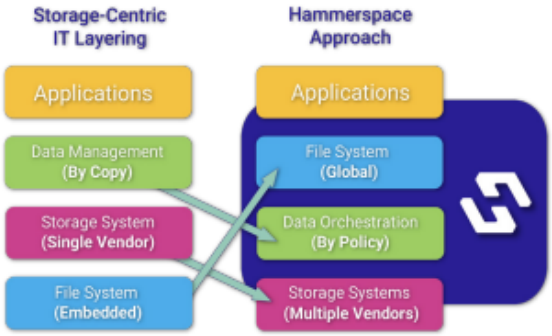
Solving the Silo Problem to Unlock the AI Puzzle

Decoupling the file system from the infrastructure layer

ERA and Hammerspace has solved this problem by investing years of development to reimagine from first principles a standards-based file system that is independent of proprietary storage infrastructure, but which is still compatible with existing storage systems from any vendor.

Unlike conventional storage platforms that embed the file system within the infrastructure layer, Hammerspace is a software-defined solution that is compatible with any onpremises or cloud-based storage platform from any vendor.

In effect, Hammerspace creates a high-performance file system that is elevated above the storage system infrastructure layer. In this way it creates a high performance, Parallel Global File System that spans otherwise incompatible storage silos from any vendor across one or more locations, including the cloud.



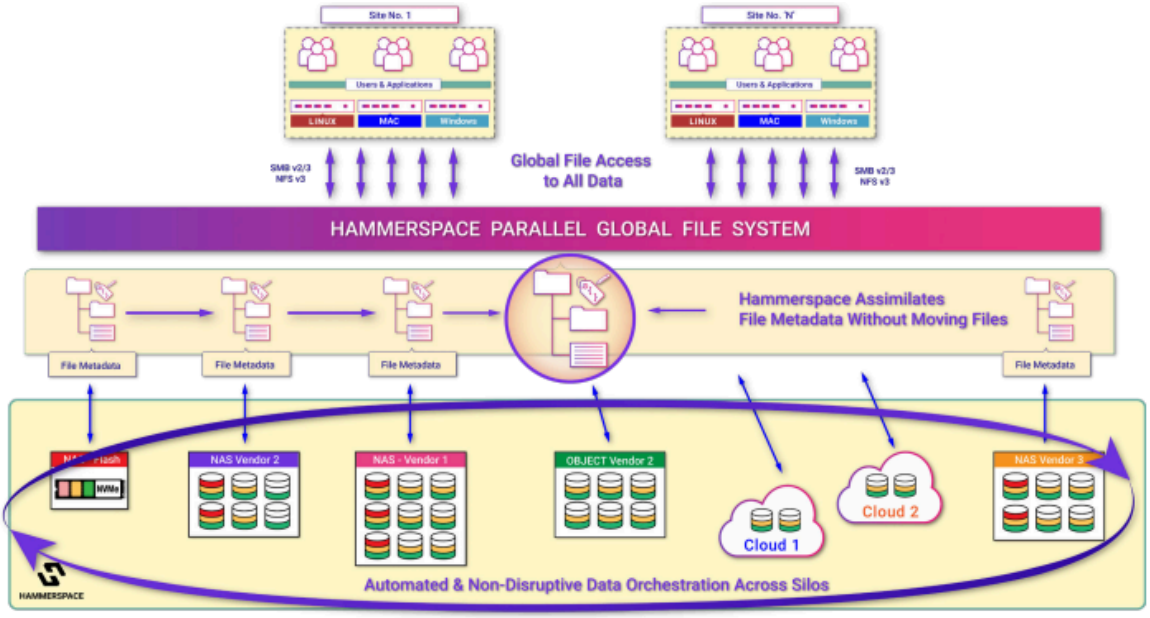


Fig 4 - Hammerspace assimilates file system metadata while leaving data in place on existing storage. In this way all users and applications access all data globally using standard file protocols via this high-performance file system. No agents or proprietary clients are required.

It does this by assimilating file system metadata from data in place on existing storage systems, and then presenting the global file system to users and applications via standards-based file protocols.

No agents are needed on the storage side. No client software needs to be installed on user systems to access the data. To users and applications, the Hammerspace Parallel Global File System presents industry-standard SMB or NFS mount points exactly like any enterprise NAS. But unlike any other solution, with Hammerspace users and applications now have global access to all of their data, and can span multiple data silos, locations, and cloud storage platforms in a single global namespace.

The ability to maintain a persistent audit trail across all file copies and locations is difficult if not impossible in traditional siloed systems.

Automated data orchestration

the key to data governance

In addition, with the file system decoupled from the underlying infrastructure, Hammerspace is able to automate data orchestration at any performance level as a background operation between storage types or to feed AI pipelines. This also means that workflow-driven data placement, data protection, or other data services can be automated without interruption to user or application access. Data orchestration can even be automated on live data that is actively being used by applications or users, without disrupting access or workflows.

In traditional storage architectures where the file system is embedded in the storage platform, if files need to be moved to another storage type or location, a copy of both the file metadata and the file essence are created and sent.

That action creates a second, forked copy of the file that must be later reconciled, and consumes additional storage capacity.

In addition, the proliferation of data copies adds risk to data governance concerns on how data is being accessed, and by whom. The ability to maintain a persistent audit trail across all copies is difficult if not impossible in traditional siloed systems.

Because the Hammerspace Parallel Global File System is independent of the storage layer, the need to wrangle such forked file copies is no longer required. With Hammerspace, all users and applications in all locations have read/write access to all data everywhere. Not to file copies, but to the same files via this unified global file system, just as they would if they were accessing all data on a single network share on a local NAS.

HAMMERSPACE

Data Governance and Auditability in AI Pipelines

Another key component when bridging silos and locations is ensuring data is not only accessible to the AI workflows, but that data placement policies to feed AI engines don't break data governance or compliance rules. In the same way that Hammerspace automates data placement and other services in the background across silos, the global reach of the Hammerspace file system also provides global audit of all file system operations.

Data governance is maintained with global audit logs of all file system operations. This is persistent even if files are moved or copied to other sites or platforms.

For example, Hammerspace supports System ACLs across both SMB and NFS shares, creating a global audit log of file system operations such as file/folder deletes, renames and other actions. This is a critical security innovation for decentralised environments to enable persistent System ACLs to be applied across multi-siloed environments, regardless of which storage type or location the file instances reside.

Since the Hammerspace Parallel Global File System manages the data placement across all silos, sites and clouds, this capability also ensures that security enforcement is not broken by moving or copying data to other sites or platforms. Hammerspace Service Level Objectives then can be tuned to maintain alignment with data governance and compliance rules.

Custom metadata to streamline data classification

Of critical importance to AI/DL workflows, data classification can be significantly enhanced and automated within Hammerspace. The system includes powerful metadata management capabilities that enable files and directories to be manually or automatically tagged with user-defined custom metadata, creating a rich set of information that can be used to streamline the classification phase of AI/DL workflows, and simplify later iterations.

Custom metadata may include virtually any information that data owners need to classify the files and help identify the subset of data that is appropriate at each phase of the AI journey

HAMMERSPACE

Reducing the problem of human error
with metadata tagging

A common problem with custom metadata in other solutions is when they must rely on humans to remember to tag things. Even the best indexing system on earth will not help you if a user forgets to apply a custom metadata tag.

Hammerspace solves this problem with automated metadata inheritance, which can be easily customised by administrators or authorised users to assign any combination of metadata tags or labels to a folder hierarchy in the file system.

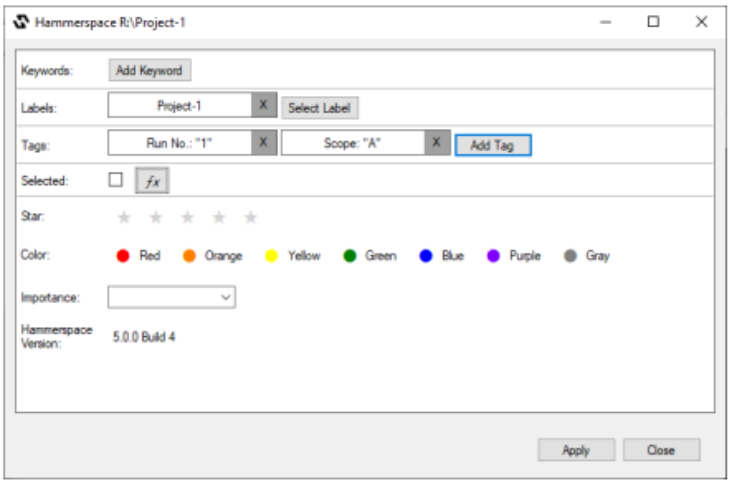


Fig 5 - Hammerspace enables custom metadata to be applied by users, or automatically. No client software is needed, since this capability is part of the standards-based Hammerspace file system.

Once these custom metadata tags or labels are applied at any level in the folder hierarchy, from the root level on down, any file or folder that lands in that hierarchy automatically inherits the custom metadata.

This means that data being generated by an instrument, or created by a user, can automatically inherit crucial identifying information as it is created, based upon controlled vocabularies of metadata variables specific to their workflows and business needs.

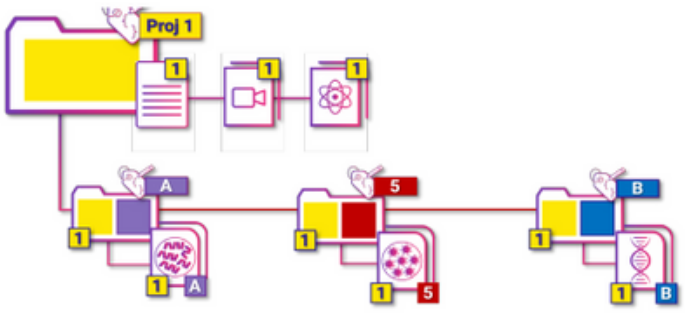


Fig 6- Custom metadata can be automatically inherited throughout the file system hierarchy to minimise human error. This streamlines data classification, and enhances automated data orchestration at a file-granular level.

Even subfolders inherit the custom metadata tags, and may have additional tags added to them. And when those files or folders move from the initial storage location to different storage types or to the cloud based upon workflow requirements, the custom metadata tags and labels are persistent and will remain associated with the files as they are moved.

The Hammerspace data orchestration system then can automate data actions based upon any combination of metadata, including standard file system variables and any custom tags.

Automating the AI journey across silos and locations

Automating the AI journey across silos and location

With this global view and control of data and metadata across otherwise incompatible data stores, locations and the cloud, Hammerspace can now provide the automation needed to feed AI pipelines from beginning to end, across all phases of the process and to all the necessary resources.

A key capability within the Hammerspace data orchestration system is the ability to define explicit, plain-language policies called Service-Level Objectives to control everything about how data should be accessed, placed, and protected, how storage resources from any vendor should be utilised, in addition to other critical data services.

In AI workloads, data placement to centres of excellence for cleansing, or to a remote data centre for training workloads, or to high-performance computing resources in the cloud or another site for inferencing workloads can all be automated in the background without disrupting user or application access, even on live data.

In AI workloads, data placement can be automated in the background without disrupting user or application access, even on live data.

This is particularly important because AI workloads typically require several iterations through multiple independent datasets. For inferencing workloads, this will often require extremely high-performance compute (HPC) infrastructures and GPU clusters that may be on-premises or as part of a temporary cloud-based resource cluster set up for the job. And with each step in the process, additional custom metadata tags can be automatically applied that identify the algorithm that was used, or other variables needed to track or recreate the workflow.

The problem is that at any given time only 15-20% of the data is active for any given step in the process. As noted above, this creates the problem of either over-provisioning the most expensive resources, or paying the “data copy tax” to wait for the inevitable churn as files are copied using manual tools such as rsync between storage silos. While the files are being copied, HPC and GPU clusters are sitting idle, and training or inferencing timelines drag out.

HAMMERSPACE

Fragmented data becomes out of control

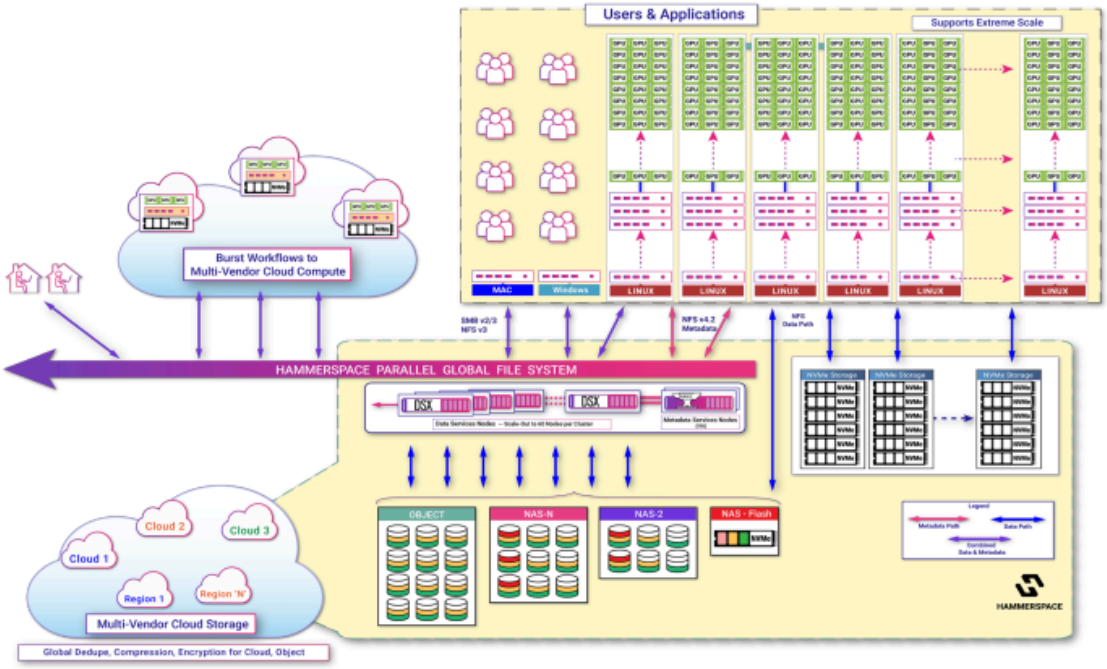


Fig 7 - Hammerspace is a software-defined solution that can start small, but scale out to accomodate even extreme performance requirements, enabling automated data orchestration to feed AI pipelines on-premises, across multiple site, and one or more clouds.

With Hammerspace this interruption can be eliminated, since data placement can be automated as a background operation on a file-granular level so data is staged on the appropriate resources just in time for the inferencing run. These can be scheduled using standard third-party tools, or directly within the Hammerspace software. Because the Hammerspace Parallel Global File System spans all resources, this data movement is transparent as a background operation, and can bridge across multiple sites, or burst to cloud-based resources seamlessly.

Empowering data scientists with self-service workflow automation

In addition, since many industries such as pharma, financial services, or biotechnology require training data as well as the resulting models to be archived, the ability to automate placement of this data into low-cost resources is critical. With custom metadata tags tracking data provenance, iteration details, and other steps in the workflow, recalling old model data for reuse or to apply a new algorithm is a simple operation that can be automated in the background.

In this way, with Hammerspace data scientists may be given direct, self-service control over all stages in the AI pipeline, across multiple locations, storage silos, and the cloud without needing to request data retrieval from IT administrators or needing to get into IT infrastructure management. And because the data can be seamlessly accessed from existing storage resources, these workflows can leverage data in place without the need to replace legacy storage systems with new infrastructure.

Scalability and Performance

As noted above, not all phases of the AI journey require high-performance compute or storage. But when they do, extreme performance is essential. Hammerspace is designed as a software-defined solution that can scale out without compromise to saturate the performance of even the most demanding network and storage infrastructures.

As a software-defined platform, Hammerspace is hardware agnostic and may be deployed on bare-metal servers, VMs, and in cloud machine instances. It is loaded from a single installer that handles both the Anvil metadata services nodes, and the DSX data services nodes types.

There is no one-size-fits-all specification for the server requirements for Anvil or DSX nodes, which means the system can be tuned to the specific load requirements of the customer's use cases. This enables the system to be dialled in to minimise unnecessary infrastructure expenses, and also to dynamically scale out when needed to saturate to high-performing infrastructure without disrupting user or application access. This includes the ability to scale out to support extreme performance environments where GPU-direct access is needed across very large on-premises or cloud based clusters.

This ability to scale-out the performance when needed, plus the automation of workflows to eliminate the delays normally needed to manage data copies means ERA and Hammerspace can increase utilisation of GPUs and other resources. The direct impact of this is to reduce the aggregate numbers of GPUs needed to perform a given workload and/or increase the throughput of an existing cluster, both of which directly impacts overall system ROI.

The direct impact of this is to reduce the number of GPUs needed and/or increase the throughput of an existing cluster, both of which directly impacts overall system ROI.

This ability to scale-out the performance when needed, plus the automation of workflows to eliminate the delays normally needed to manage data copies means Hammerspace can increase utilisation of GPUs and other resources. The direct impact of this is to reduce the aggregate numbers of GPUs needed to perform a given workload and/or increase the throughput of an existing cluster, both of which directly impacts overall system ROI.

Some Hammerspace customers have begun with initial cloud-based implementation to rapidly provision an initial workload, for example, which they later convert to on-premises systems. User access is seamless, and never impacted by this change in infrastructure.

In addition, this decoupling of the file system layer from the storage layer enables independent scaling of I/O and IOPS at the data layer. Extremely high-performance NVMe storage can now co-exist with lower cost and lower performing tiers including cloud in a global data environment. Data orchestration between tiers and/or locations is controlled transparently as a background operation based upon workflows or objective-based policies.

In Summary

In Summary

As noted throughout this white paper, the ERA and Hammerspace solution has been designed from the ground up to solve the problems caused by fragmentation of data across silos in the data centre, and increasingly across distributed systems that may span multiple data centres and the cloud.

The rapid shift to accommodate AI/DL workloads has created challenges that exacerbate the silo problems that IT organisations have faced for years.

And the problems have been additive:

- **High Performance & Massive Scale:** AI pipelines need the ability to scale up and out to extreme performance requirements. The ability to do this without overprovisioning infrastructure, plus the ability to burst to the cloud is essential.
- **Multiple Data Sources:** To be competitive as well as manage through the new AI workloads, data access needs to be seamless across local silos, locations and clouds.
- **Data Governance:** And there is the need to be agile in a dynamic environment where fixed infrastructure may be difficult to expand due to cost or logistics. This means the ability for companies to automate data orchestration across different siloed resources while maintaining auditability and controls for compliance, security, and other data governance requirements.
- **Standards Based:** Enterprises need to bridge their existing infrastructure with these new distributed resources based upon industry standard protocols, without requiring proprietary vendor clients or agents to install. When combined with automated data orchestration, this ensures that the cost of implementing AI/DL workloads does not crush the expected return.
- **Seamless Burst-to-Cloud:** Whether due to the difficulty in procuring GPUs, or because some AI pipelines only need short-term compute/storage resources, the ability to rapidly burst to temporary cloud-based resources is a key requirement. To do so by extending the Hammerspace file system and minimising data movement is a key to the flexibility and adaptability needed.

ERA and Hammerspace solutions are ideally suited to provide customers a solution to these problems, leveraging their existing infrastructure.

HAMMERSPACE

In Summary

ERA and Hammerspace solutions are ideally suited to provide customers a solution to these problems and requirements, without the need to retool their data centres with new storage & compute infrastructure. At the same time, customers no longer need to manually shuffle file copies between vendor silos and pay the resulting 'data copy tax'.

Because the Hammerspace high performance file system is global, and data orchestration can be automated seamlessly across all silos and locations, AI workloads can now be optimised and rapidly adapt to new AI applications as they occur and to meet even extreme performance requirements.

To keep up with the many performance variables for AI pipelines, a new paradigm was necessary that could effectively bridge the gaps between one or more on-premises silos and clouds. Such a solution required new technology and a revolutionary approach to lift a high-performance file system out of the infrastructure layer to enable AI pipelines to utilise existing infrastructure from any vendor without compromising results. It is a revolution as important as when network-attached storage vendors lifted the file system out of the operating system in the 1990s.

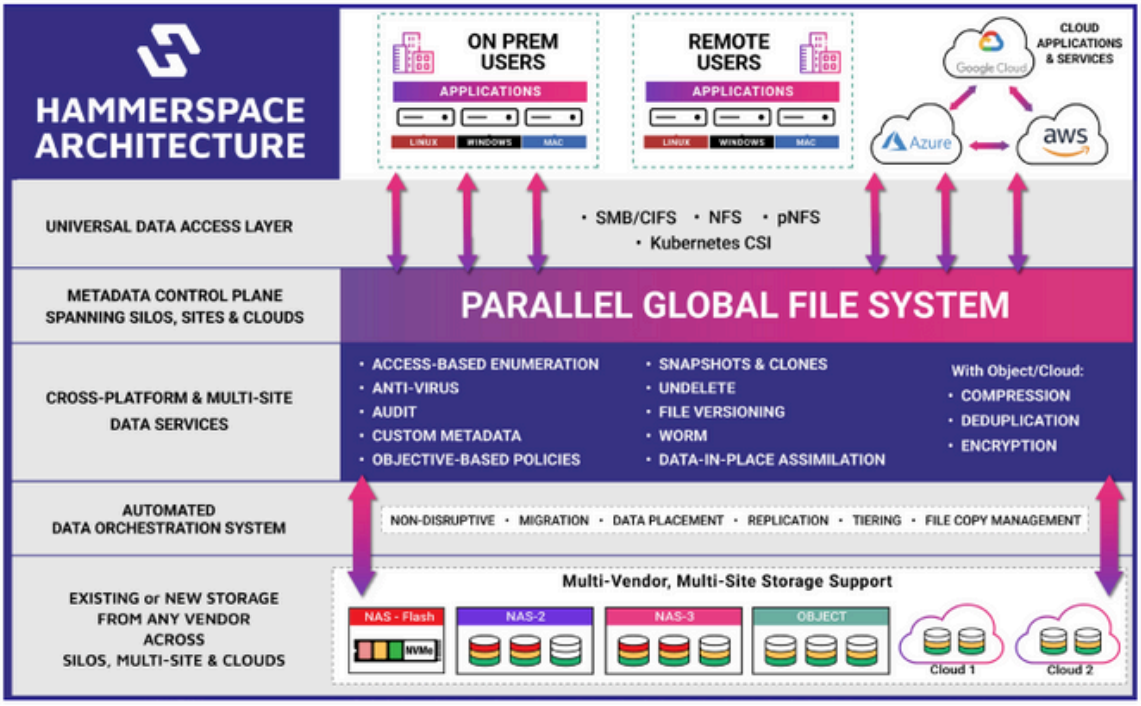


Fig 8 - A logical view of the capability stack that comes with Hammerspace software, anchored on its high-performance Parallel Global File System.

About Us

Based in Chesham, Buckinghamshire and delivering IT and workflow solutions since 1998, ERA is one of the UK's leading independent providers of IT workflow solutions for clients in the media, VFX, post-production and broadcast industries. Solutions cover all aspects of Infrastructure as a Service (IaaS), cloud archiving, remote workstations, storage, maintenance support and other managed services to meet the needs of complex and demanding media workflow projects.

By taking a vendor-neutral design approach and understanding no two customers are the same, ERA has developed a large and strong customer base. With customers including Jellyfish Pictures, The Look, Evolutions, The University of Salford and Vice Media; ERA have helped with challenges around storage, impractical infrastructures, accessibility issues, scalability, and integration.

Hammerspace and ERA

Together we are revolutionising the way our customers tackle storage challenges:

- **Consolidated storage under a single name space** – Say goodbye to scattered data.
- **Global Access, unified dataset** – Access your data anywhere in the world, simultaneously.
- **Rapid Access for rendering needs** – Our burst infrastructure has you covered with fast access for your projects.
- **Bridging on-premise with hosted services** – No high-speed connectivity? We make it easy for customers with on-premises infrastructure to tap into hosted services effortlessly.

Thank you to our partner **Hammerspace** for providing valuable insights and information in this whitepaper. The content presented is based on the findings and analysis outlined in their original publication.