White Paper

# ONTAP – pioneering data management in the era of Deep Learning

Arindam Banerjee, NetApp
September 2024 | WP-4303

## Abstract

This white paper details the architectural vision for NetApp ONTAP to address the growing need for high throughput and performance for unstructured data to support Generative AI and deep learning workloads.

TABLE OF CONTENTS



# Technology trends – the age of data

## Deep Learning is driving infrastructure innovations

More than 400 million terabytes of data are generated every day. A staggering 80% of that data is unstructured. The explosion in unstructured data has triggered advancements even in one of the most modern fields in computer science and math, Artificial Intelligence. As more and more data generated is in the form of unstructured text, speech, images and videos, comprehending the data set needed advances in the field of AI. Machine Learning was superseded by Deep Learning driven architectures such as natural language processing (NLP), speech recognition, and computer vision.

Methods in Deep Learning resemble neural networks inspired by the human brain. While Deep Learning can efficiently process unstructured data and identify hidden relationships and patterns, it requires iterative and unsupervised learning techniques for developing its cognitive capabilities, much like the human brain. Just like the human brain, it requires large volumes of high quality data and enormous processing power to be able to consume and comprehend the data. As an example, a machine learning model may require tens to hundreds of data points per feature while a deep learning model may easily require thousands of data points per feature.

Deep Learning techniques will continue to advance, thereby driving the next wave of innovations to compute and storage architectures as it requires advanced recognition of data patterns or relationships for *learning*.

## Evolution of data driven architectures

The accuracy of deep learning model prediction improves through iterative training and feedback loops with large volumes of high quality data. Data, however, has gravity. The gravitational properties of data make it difficult to continuously move large volumes of data for AI. This can be attributed to multiple factors:

- The infrastructure cost to move large volumes of data at high speed. Data movement not only consumes network bandwidth but also requires significant compute power to copy the data across.
- Complexity of managing multiple platforms, each maintaining data in its own format. This includes duplication of data and increase in storage costs.
- Data privacy and governance regulations necessitates that data permissions and access controls be maintained across multiple platforms.
- The recurrent nature of the above steps since data is continually updated and AI platforms require the latest data sets to provide the most relevant predictions.

As deep learning models continue to evolve and drive business and research outcomes with a high degree of precision, they will continue to drive data processing at high velocity and data management at scale. Architectures that can manage such data operations in a simple, scalable, reliable and secure way will become the mainstay of data centers.

Data driven architectures, therefore, must demonstrate the following capabilities:

- Ingest data efficiently and reliably from multiple hybrid data sources (since data will be generated everywhere).
- Provide high speed and secure multi-protocol data access at the scale of modern parallel and grid computing needs.
- Elastic scaling of the underlying infrastructure as the volume of data increases.
- Policy based data lifecycle management that seamlessly integrates with the application lifecycle adhering to data security and privacy regulations.
- Provide a structured metadata namespace to derive insights from unstructured data assets.

## Architecting platforms for data driven architectures

Distributed storage and filesystem architectures have existed for a couple of decades. However, those architectures, especially those built on shared nothing topologies do not scale in a cost effective fashion. Existing architectures do not effectively decouple performance from capacity and hence scaling of one attribute necessitates scaling the other, thereby incurring additional cost.

Architecting storage and data platforms for the future requires a closer look at the secular trends around the core tenets of data center infrastructure, i.e. compute, network and flash. While CPU clock frequencies have started to plateau, network bandwidth continues to increase (we are talking 800 GbE today). Flash continues to become denser (addition of more dies) and faster. NVMe drives continue to gain bandwidth from advancements in PCIe technology. Given that different components of infrastructure are following different trajectories, large scale infrastructure for the future will need to be built on composable or disaggregated architectures.

A true composable or disaggregated architecture typically segregates compute and storage as independently scalable units connected over a high speed and low latency network. This enables independent scaling of performance and capacity, giving users the flexibility to scale on only one vector at a time. In data centers, the compute refresh lifecycle is typically shorter than storage refresh lifecycles. Disaggregation of compute and storage therefore simplifies tech refreshes. Disaggregated architectures also impart a simplified consumption model to the end user who can choose only one vector to optimize for and scale accordingly.

# AI Data Platform – powered by ONTAP

## Built on core principles of disaggregated infrastructure

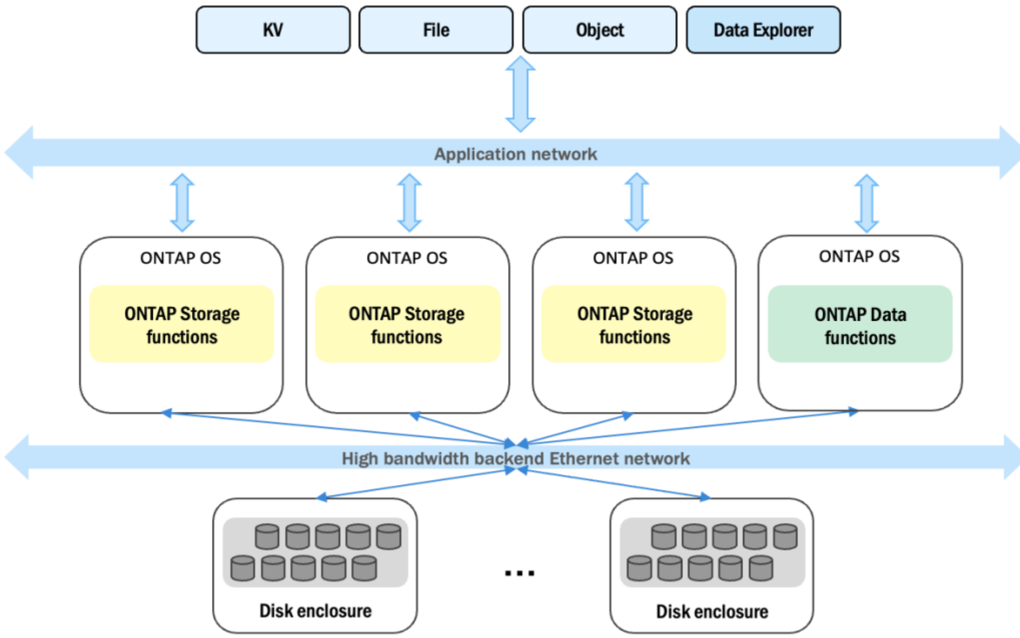
Today we are unveiling a new design center in NetApp® ONTAP® built on the tenets of disaggregation and composable architecture. This design center is perfectly suited for workloads that require high bandwidth and scale and will be a new entrant to the family of ONTAP-based products, which include NetApp AFF and ASA.

Built for the era of deep learning as well as high resolution images and video streaming, this new architecture will provide cost-effective scaling of the data infrastructure. It is composed of independent compute and storage units interconnected through high bandwidth and low latency Ethernet network leveraging RDMA capabilities. The modularity of the architecture ensures that it can operate at unprecedented levels of scale governed by the laws of physics.

The compute units will host the core software algorithms required to drive storage and data management functions. Each unit will be running an instance of ONTAP OS which will be embellished with additional metadata and data services. As more compute is added, the storage and data management functions scale performance linearly. The software logic automatically (and seamlessly) assimilates the new compute units and (re)balances existing workloads across the newly added compute units.

The storage units are high density disk enclosures that provide high levels of reliability (6x9s). They are connected using a high bandwidth and redundant Ethernet network required to address the high speed NVMe drives. If the system is running low on capacity, additional storage units can be added. The software algorithms running on the compute units can seamlessly detect and add new storage units to the global capacity pool and teach the allocation logic to allocate space from the new capacity pool.

Figure 1: ONTAP on disaggregated infrastructure



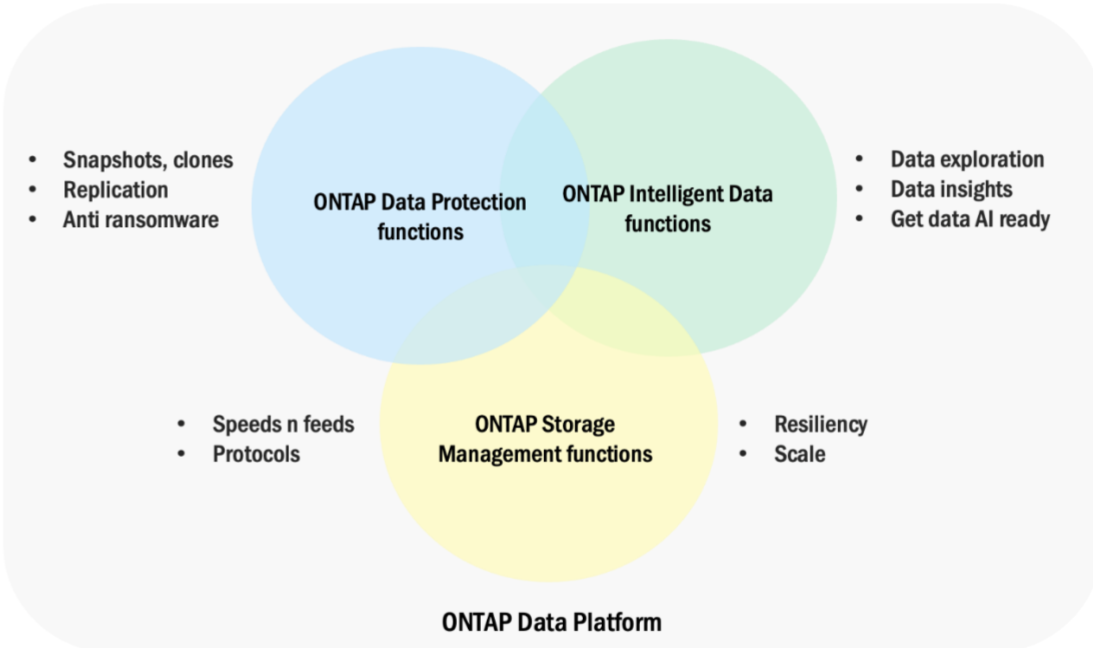

## Implements Intelligent Data functions in ONTAP

Deep learning operates on unstructured data. Unstructured data as it is defined has no structure that can be leveraged to classify, label, annotate, or even extract features directly from it. Therefore, unstructured data in its raw form is not directly useful for AI. It requires a contextual structure around it for it to be relevant for AI training or inferencing. Unstructured data therefore requires pre-processing or *data wrangling* to derive value from it. These operations are cumbersome, time consuming and often require data to be temporarily moved to different platforms, thereby increasing infrastructure cost, posing security risks and creating data siloes.

The new composable infrastructure framework enables compute units to be added dynamically to the system. ONTAP can leverage these additional compute units to perform *Intelligent Data functions* to help derive a structured view of unstructured data. This structured view of unstructured data enables seamless integration with AI software ecosystems. When combined with existing data management functions of ONTAP, it helps establish traceability of the AI model to the data the model was trained on.

The ONTAP Data Platform for AI is a comprehensive data platform that possesses capabilities to not only store and protect the data but also to derive insights from it. The three main architectural pillars include core storage management functions for high bandwidth storage and retrieval of data over file, object and block protocols, data protection functions to drive industry-leading data lifecycle management and the new Intelligent Data functions to provide a structured view of unstructured data as well as perform policy-based data classification. The combination of these three functions elevates ONTAP to a comprehensive data platform that helps bring AI to the data, in-place!

**Figure 2: Intelligent Data functions hosted on ONTAP**

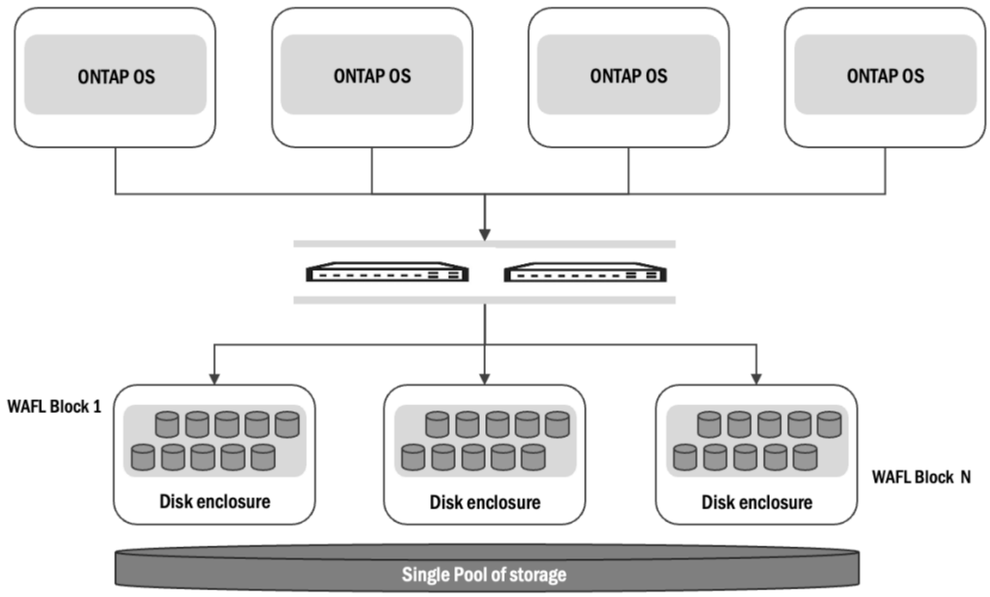



## Core architectural tenets

### Single pool of storage

One of the key tenets of this new architecture is the ability to construct a single pool of storage from disks in multiple storage units (enclosures). As new disk enclosures are added, additional capacity is automatically added to the system and the total capacity expanded. As such, the NetApp WAFL® physical block space is now distributed across multiple disk enclosures, thus creating a single extensible namespace. This architectural construct provides simplicity, flexibility and elasticity to the system. It simplifies storage provisioning, management and tech refreshes. New modular blocks can be added to the system and old blocks retired seamlessly.

**Figure 3: Single pool of storage**




The single pool of storage provides a full view of all data and metadata blocks in the system. The capacity of the entire system can now be managed as a single global pool without any siloes. This is possible since each compute unit or node running the ONTAP OS has full view of and can directly communicate with the storage units providing the capacity. The WAFL block allocation algorithms will always select the best disk segments to direct the writes to at any moment

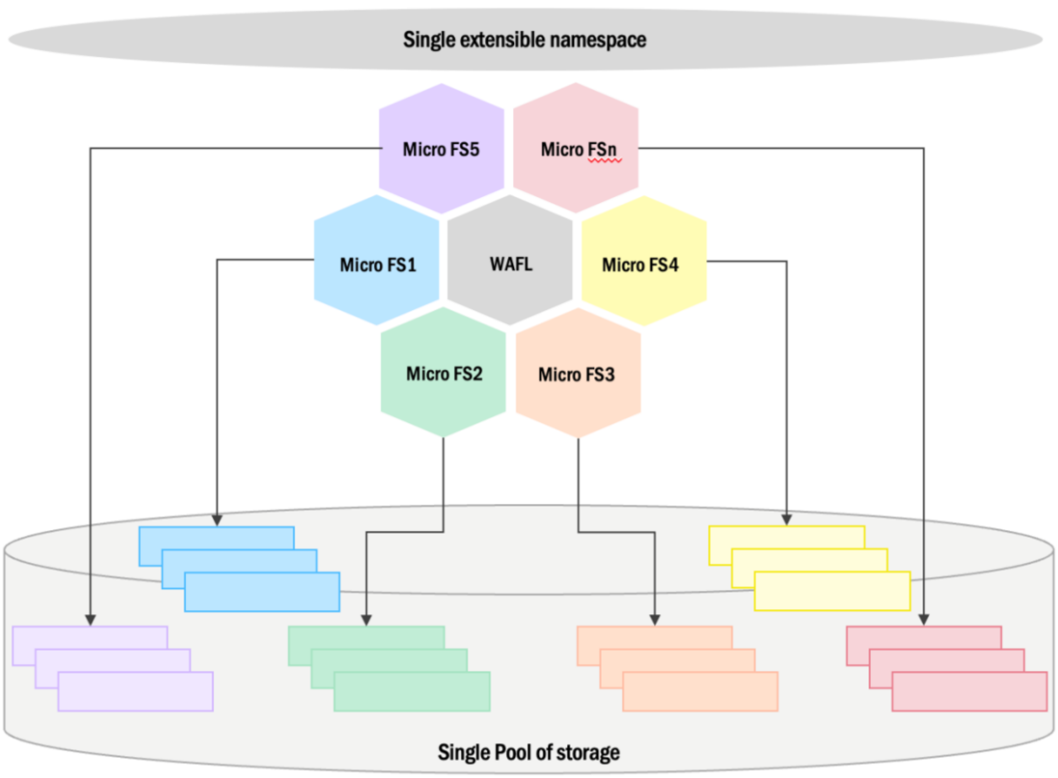
## Independently consistent and scalable micro filesystem instances

Traditional distributed storage systems leverage global transaction or lock management to maintain consistency of data (especially to handle read after write scenarios). This kind of design choice(s) often limits scalability of the system. Distributed transaction or lock management can be cumbersome. Establishing a write guard or a fence on a protected resource across nodes in a large scale distributed system requires idempotent communication protocols that can add significant overhead and impede both scalability and performance.

The sustainable way to build highly scalable distributed systems is by decentralizing the core operations and establishing loose coupling between the functions that maintain consistency at the highest level. This is critical when operating at the scale required by parallel computing, deep learning, and multi-tenanted AI services.

At NetApp, we are working along these core principles by introducing the concept of *independently consistent micro filesystem instances*. Each micro filesystem instance operates as a fully functional filesystem and provides consistency across data and metadata operations. Each instance operates with its own set of resources which includes a set of parity protected segments on disk known as allocation areas. Resource allocation across micro filesystem instances can be dynamically adjusted once certain thresholds are reached. These operations are performed in the background and have no implication to application performance. The system as a whole is constructed by cohesively aligning all of these micro filesystem instances across the single pool of storage. Resources shared between micro filesystem instances are reference counted and reconciled through an extremely low overhead idempotent technique to ensure that consistency is maintained across all micro filesystem instances. Since each micro filesystem instance has exclusive ownership of its resources at a given point in time, they can operate safely on filesystem internal data structures in parallel to other instances. This approach allows linear scaling of performance as new compute units (nodes) are added. As compute units are added, micro additional instances can be instantiated and re-distributed across the new compute units to scale performance.

**Figure 4: Independent micro filesystem instances**




These micro filesystem instances are decoupled from the front end or application facing constructs. As an example, a filesystem client mounting file shares and performing data and metadata operations has no visibility to which micro filesystem instance is processing the request. The client will communicate with the file server as per semantics prescribed during mount.

## Parallel access to filesystem data

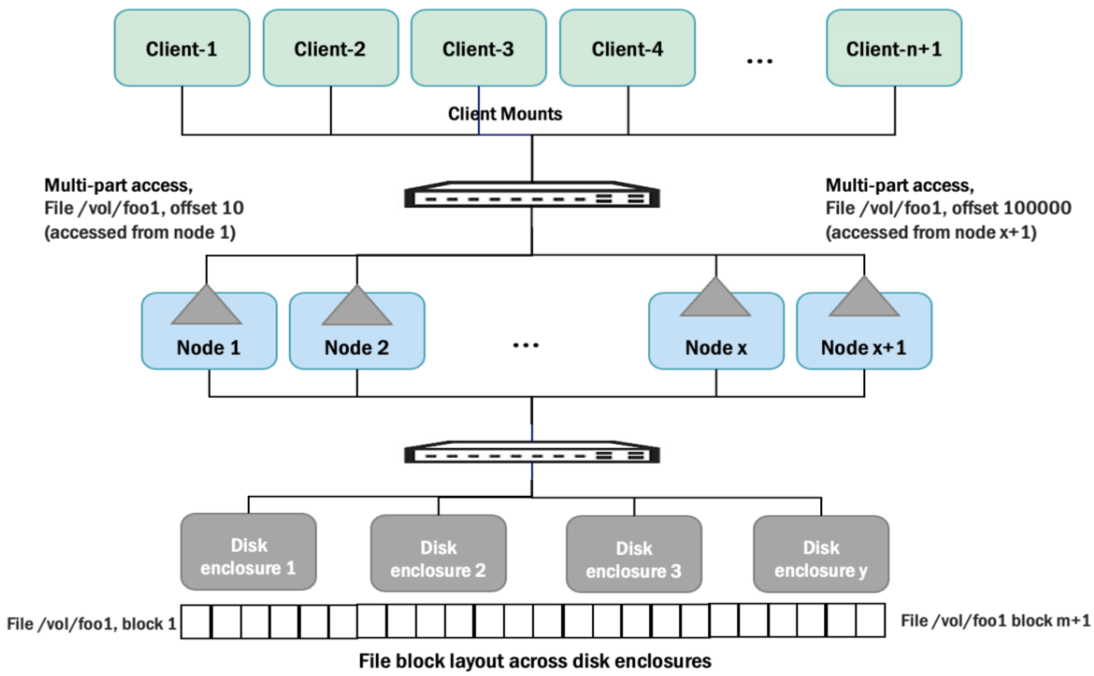
Enabling parallel or concurrent access to the filesystem data is one of the core design goals for any distributed filesystem. This applies not only for application driven data and metadata operations but also for background operations, especially ones that modify core data structures.

Driving parallelism in any large scale filesystem requires operations to be distributed in parallel across 3 subsystem layers:

- Client and server side protocol stack
- Filesystem namespace and object mgmt. subsystem
- File system block layer managing on-disk layout

With the new design paradigm, ONTAP can achieve this parallelism across all 3 subsystems. It can do so by leveraging standard access protocols without needing any special client side software.

**Figure 5: Parallel access filesystem over standard protocols**




End to end parallel access across the system will be enabled through the following methods and algorithms in each subsystem.

**Protocol layer**: Protocols like pNFS (parallel NFS) allow parallel access to the same exported file path through multiple data servers. Additionally, a file server can publish file layouts that enable even a single file to be accessed in parallel from multiple nodes through the protocol layer. The server can now publish different IP access points for different extents of a given file. The application IOs for that given extent will be automatically redirected to that node to get the fastest access path.

**Object level concurrency**: With ONTAP's multi-part access topology and algorithms, each individual file is a collection of multiple parts or extents of the file. Each of these file parts has its own unique identifier that enables a given file to be accessed in parallel across multiple compute units or nodes using these unique identifiers. This enables linear scaling of even a single filesystem object (aka file) as more compute units are scaled within the system.

**Access to data blocks**: WAFL on-disk layout will ensure that each individual file or a collection of files within a file share will have their data blocks distributed across multiple disk enclosures to drive massive parallelism and concurrency of access. Each instance of the ONTAP OS will have high bandwidth connectivity across the backend disk enclosures and can leverage RDMA constructs to maximize performance as well as ensure quality of service end to end.

With all 3 subsystems designed for highly concurrent access, ONTAP can maximize system utilization and enable extensive parallel processing across the system to meet the demands of AI training and inferencing workloads.

## Systemically reducing write amplification

Managing write amplification is an important function in system design especially when designing high throughput write intensive data storage systems. A high Write Amplification Factor (WAF) results in increased system cost, reduction of overall system throughput as well as reduction in the lifespan of flash media (SSDs) by adversely impacting their endurance.

ONTAP extends its existing write allocation and WAF management algorithms (aka WAFL tetris) and combines it with new distributed capabilities like global free space management to further optimize WAF within the system.

**WAFL tetris**: WAFL write allocation continues to allocate the best file system extents to minimize read-modify-write operations on the underlying media. WAFL continues to employ tetris writes to ensure that writes to each SSD is always a multiple of flash blocks thereby limiting the number of erase cycles on flash.

**Global management of free space**: One of the benefits of a single pool of storage is that each instance of the ONTAP OS has full view of the free space available across the system. At the time of writes, the WAFL algorithms that allocate filesystem extents allocate the extents with the maximum number of contiguously free blocks. With a single pool of storage, this logic can be extended across the entire capacity pool. This improves the WAF management capability of the tetris algorithm.

**Efficient "*storage gardening*"**: WAFL's intelligent write allocation and free space management algorithms minimize the need for background operations like filesystem defragmentation and garbage collection that further amplify WAF. WAFL leverages remote reference counting mechanisms to augment the write allocation and free space management algorithms. The consistency of these reference counts is maintained by implementing a novel idempotent transactional mechanism leveraging disks attached to a NVMe over fabric network.

By combining the additive effects of WAFL tetris, global free space management and storage gardening, ONTAP minimizes write amplification across the system to ensure smooth and seamless operation with lower endurance flash media like QLC and even PLC in future.

## Dynamic load balancing

The ability to construct a single pool of storage over a many-to-many connection topology from the compute units to the storage units is the focal point of this architecture. Each compute unit or node that runs an instance of the ONTAP OS has full visibility across the entire storage pool and can access any data (or metadata) block across the system over an equal path length. This allows for even distribution of client workloads across all compute units within the system. The system can actively monitor overload conditions for a given compute node and redistribute the excess load amongst other compute nodes.

In the case of a node failure, the IO load serviced by the failing node can be evenly distributed across all surviving nodes in the system. If the client is accessing the data through a parallel access protocol like pNFS, the file server can also publish a new access layout for the client in such a scenario. The new layout will redirect the client to access the same file or file segment over a different path as advertised by the file server. The file server will intelligently map new layouts for each file system object or entity for one or more clients such that the load is evenly distributed amongst surviving nodes.

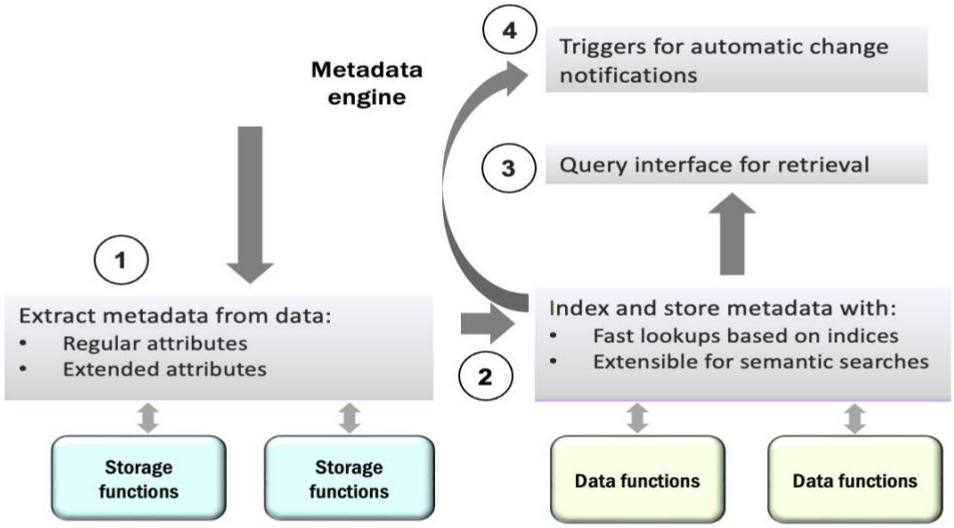
This ability to load balance automatically both under shifts in system load as well as under failure ensures efficient resource utilization in the system. Efficient resource utilization maximizes system utilization and improves the total cost of ownership of the system. It also maintains desired levels of headroom in the system to manage Quality of Service effectively.

## Structured metadata engine

The structured metadata engine is the pivotal element of the Intelligent Data functions introduced in ONTAP. This metadata engine manages attributes of the data, whether it is regular attributes, extended attributes or object tags (in case of object storage). Additional data classifiers may also be instantiated to process and classify the data based on user defined policies.

The metadata engine extracts the data attributes (or metadata) inline. Once the attributes are extracted, the metadata engine indexes and stores this metadata to enable fast lookups. A query interface allows applications to query for this metadata. The query interface is extensible to enable semantic searches on the data if exact key words are not known. Additionally, applications subscribing to be notified of any changes to the metadata can be alerted through a triggering mechanism implemented by the metadata engine.

**Figure 6: Metadata engine workflow**




One of the characteristics of AI workloads is the increase in system object density per system (aka files and directories). This is a function of consolidating all the required artifacts for model training or inferencing. A side effect of such consolidation is that traditional filesystem crawls take an inordinately long time to complete. The metadata engine solves that problem by providing a fast index and search capability through the metadata set. AI software ecosystems deployed for data labeling, classification, feature extraction or even a RAG framework deployed for generative AI inferencing use cases can significantly speed up time-to-value of their data by leveraging the structured view of unstructured data presented by the metadata engine.

## NetApp brings AI to the data through disaggregation

Fast forward to the future. Deep Learning models will continue to evolve and push for exponentially increasing data points to strive for more precision and accuracy. New model architectures may also evolve to learn from more complex data sets. The infrastructure challenges and platform complexity with moving extremely large data sets is going to get arduous especially with security and privacy regulations getting more mature.

At the same time, AI will be everywhere and will have a transformative impact on human life. Most devices that either store, receive, transmit data will need to be equipped with AI. The democratization of AI is causing a lot of edge devices to be equipped with AI accelerated compute especially for inferencing. Model and vector quantization techniques are evolving such that complex models can operate with fewer resources. AI is finding its way to the edge, core and more and more data centers. It will be ubiquitous.

AI inferencing capabilities can be and will need to be delivered in-place, right where your data is. With disaggregation, NetApp is bringing AI accelerated compute close to where the data is and making customers derive full value of their data assets, in-place!

## Challenges in establishing a Generative AI framework for inferencing

Training Large Language Models (LLMs) requires significant capabilities from the underlying compute and data infrastructure. The cost of re-training a foundational model is very high. Fine tuning LLMs with enterprise or organization specific data is also cumbersome. Since data is constantly changing, the fine-tuned model may not be up to date with the most recent context or information. Retrieval Augmented Generation (RAG) is a very common technique employed for inferencing when LLMs are deployed. It is the process of contextualizing the output of a LLM with a data source that is outside the scope and more relevant than the data source the LLM was trained on.

Unstructured data cannot be used for RAG as is. It requires to be made ready for RAG. The challenges with establishing a RAG framework includes:

- Maintaining the recency of the data used for RAG
- Classifying the data as per security and privacy policies
- Chunking and embedding the incremental data efficiently for storage and retrieval
- Managing the vector database

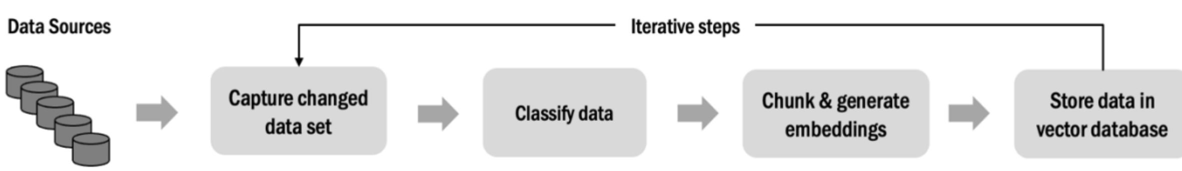
**Maintaining the recency of data**: One of the big challenges with establishing a RAG framework is maintaining the recency of the data used for “Retrieval”. Doing it repeatedly and at scale is demanding. The challenge lies in identifying the incremental changes to the data set. If the changed data set cannot be accurately identified, the LLM may not be able to provide the most updated responses to the user prompts. This can be construed as hallucinations.

**Classifying the data**: Security and privacy policies mandate that data used for AI be classified in case it has sensitive or PII information. Some users may have policies to anonymize the data sets in such scenarios. Policies may also mandate that the data is anonymized prior to embedding creation and storage in a vector database.

**Chunking and embedding efficiencies**: This is a very important step to get the data ready for “Retrieval”. Identifying the right chunk size and the dimensions of the vector embeddings is key. Tradeoffs may need to be made between response accuracy, good query performance and the overhead of managing the continuously growing vector database.

**Managing the vector database**: As AI practitioners are quickly identifying, managing the content of the vector database can be quite challenging. Depending on the chunking and embedding methods, the vector databases can quickly grow to sizes which are at least an order of magnitude higher than the actual data set. This has significant impact on the query performance since the database indexes also grow along with the data and hence index lookups cannot be served from memory. To overcome this, AI practitioners overprovision for memory in their servers which leads to significant increase in infrastructure cost.

**Figure 7: Getting data ready for RAG**




To exemplify, Figure 7 identifies the iterative steps required to make unstructured data ready to be consumed for RAG driven inferencing. This is the backend of the RAG framework which makes sure that the response to the generative AI application query is relevant. The data processing functions deployed in this phase are compute intensive and may require AI accelerated compute. The vector database requires storage that needs to be provisioned, protected and managed while enabling for rapid growth as new embeddings are created. Performing these operations repeatedly and at scale is a challenging task. More importantly, the data sets need to be versioned as well since a given response to a generative AI application query is a function of the *retrieved* data set at that point in time. This would require versioning the database for auditing and compliance reasons. These tasks take a significant amount of time and expertise to automate. Currently deployed solutions often require data movement to new platforms to perform different operations.

## ONTAP improves time-to-value for inferencing by solving the data challenges

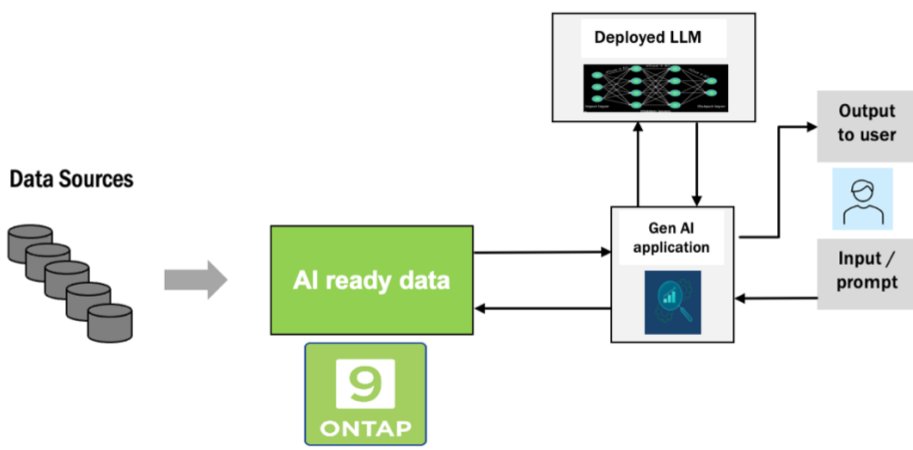
The flexibility of a composable infrastructure is the ability to add compute power to the system independently. ONTAP will leverage the additional compute to make the data "*AI ready*". Data Scientists can bring in their own Generative AI application and their LLM to establish a RAG framework. ONTAP will keep the data ready to serve the "**R**etrieval" phase of **R**AG in a simple, economical and scalable fashion.

With a systemic view of solving the data engineering challenges, ONTAP's Intelligent Data functions make it simple for Data Scientists by making the data AI ready, in-place, inside the ONTAP data platform! NetApp's powerful *SnapDiff®* API will track incremental changes to data in the most efficient manner. The metadata engine in ONTAP will record these changes and leverage its *trigger* functionality to initiate downstream operations for data classification, chunking and embedding creation. Specialized algorithms within ONTAP will generate highly compressible vector embeddings that significantly reduces both the on-disk and in-memory footprint of the vector database (significantly shrinking infrastructure cost). A novel in-memory re-ranking algorithm during retrieval ensures high precision semantic searches.

Since the generated embeddings are stored in an integrated vector database backed by ONTAP volumes, the vector database will inherit the ONTAP data management model. This enables NetApp Snapshot™ copies and versioning of data in conjunction with the deployed model. When integrated with well-known AI ecosystem software, it will allow tracing of RAG based predictions to the data set that the **R**etrieval was based off.

By solving the key data challenges to make unstructured data inference ready, ONTAP not only improves time-to-value for AI inferencing but also versions the data for traceability or *explainability*.

**Figure 8: ONTAP Intelligent Data functions makes your data AI ready**




### Built into ONTAP by ONTAP engineers!

Finally, all these innovations will be an integral part of ONTAP. They are built on ONTAP's proven resiliency features. ONTAP's data integrity checks, incremental checksums, lost write protection, automatic re-parity and the full suite of WAFL error detection and recovery scans are available with this new architecture.

All of ONTAP's multi-tenancy, security, anti-ransomware and the highly interoperable set of data protection features (leveraging NetApp SnapMirror® technology) will continue to be supported and more importantly through the same ONTAP API. This is going to change the game on data management!

## Let's recalibrate

ONTAP is the premier storage operating system in the world for unified and unstructured data workloads. It provides high throughput access to data over the most comprehensive list of protocols. It ensures the highest levels of reliability through both software and hardware innovations. It has industry leading data management and data protection features that includes the most efficient Snapshot and cloning technologies combined with the reliability, efficiency and interoperability of SnapMirror technology. It is considered the safest storage on the planet and can detect ransomware attacks inline and with high precision.

With a new design center based on modularity, ONTAP is reliving the vision of *data driven architectures*. It is taking a systemic view towards providing the best infrastructure for AI with industry leading performance while economizing on rack space (density) and power (sustainability).

By leveraging disaggregation of storage and compute, ONTAP is adding Intelligent Data functions that brings AI to the data and enables users to derive incisive insights about their data assets. Further it is transforming unstructured data to AI ready data, in-place without the complexities and cost of introducing another platform. It is time we recalibrate our thinking on data management!