NetApp's Vision for Enterprise AI

STEVE MCDOWELL, CHIEF ANALYST
9/30/24

CONTEXT

One of the biggest hurdles facing enterprises adopting AI is integrating disparate data sources with their AI initiatives. Data is often spread across various systems—on-premises, in the cloud, or a combination—leading to inefficiencies in accessing, processing, and analyzing this information for AI model training.

At its recent Insight customer event, NetApp shared its vision for how its solutions will evolve to address these challenges. NetApp's approach combines its existing ONTAP-based products with a new disaggregated architecture and new data manipulation capabilities that promise to deliver the efficiencies demanded by enterprise AI.

While NetApp shared its vision, it cautioned that it's not ready to ship products based on its new architecture. The company didn't say when it would ship products, but we expect the first wave of solutions to appear sometime in mid-2025.

AI & THE DATA INFRASTRUCTURE

One of the primary challenges organizations face in their AI journey is integrating disparate data sources into a cohesive AI workflow. Enterprises often store data across multiple systems, whether on-premises or in the cloud, leading to inefficiencies in managing and moving data for AI training. These inefficiencies become more pronounced as AI models become more advanced and require increasingly larger datasets.

NetApp clearly recognizes that for AI to succeed, the infrastructure must support seamless data access, discovery, and traceability. Data scientists need a unified view of all their data assets, regardless of location, to ensure that AI models can be trained effectively and with full context. NetApp will address these challenges with a sophisticated data management architecture that

streamlines data access across storage systems and integrates with enterprise AI workflows.

NETAPP'S VISION

AI workloads are heavily dependent on the efficient movement of data. Traditional legacy approaches can often deliver the raw performance required for AI but often struggle with scalability—this is where a disaggregated architecture can help.

Beyond raw performance, the AI lifecycle requires moving data between various external tools, such as a vector database. Moving data from storage to an external tool and back again is intrinsically inefficient. The process introduces latency and complexity that can make the AI workflow slower than it could be.

NetApp new vision for managing data for AI solves both of these challenges. NetApp said that it plans to deliver an intelligent data infrastructure built around a unified AI data management engine that not only stores data for AI but takes things a step further: Its new AI engine will contain data manipulation capabilities required for efficiently exploiting AI in the enterprise.

While data will always need to move between storage and external tools, NetApp believes it can dramatically improve the efficiency of AI workflows by moving some of the most common AI data manipulations directly into the platform. This changes the nature of what its data platform delivers and increases the efficiency of enterprise AI workflows. NetApp's new AI Engine accomplishes this.

Key elements of NetApp's approach include:

- **Disaggregated Storage Architecture**: To optimize storage resources, NetApp is developing a storage architecture that separates compute and storage, allowing enterprises to scale these components independently. NetApp chose this architecture to improve aggregate throughput, reduce costs, and provide flexibility for AI workloads.
- **Seamless Data Integration**: NetApp's AI data management engine promises to provide a comprehensive view of all data assets, enabling seamless data integration across on-premises and cloud environments. This unified approach should simplify AI workflows and reduce the complexity of managing data across hybrid environments.

- **Vector Embedding and Databases**: The NetApp AI data management engine will generate vector embeddings and store them in an integrated vector database, facilitating efficient data searches and enabling RAG workloads.
- **AI Ecosystem Integration**: NetApp promises to integrate its upcoming AI data services with the broader AI tool ecosystem to enable a streamlined AI workflow from data labeling and model training to deployment and monitoring.
- **Responsible AI**: NetApp emphasizes the ethical use of AI by incorporating model data traceability and governance features, ensuring that organizations can implement AI solutions that are both effective and transparent.

NEW DISAGGREGATED ARCHITECTURE

NetApp is building its AI vision on a new disaggregated storage architecture designed to address the specific needs of large-scale, compute-intensive AI workloads, such as training and refining LLMs.

Separating the storage backend from the compute infrastructure provides several key benefits:

1. **Maximized Utilization of Network and Flash Speeds**: Decoupling storage from compute allows storage resources to be used more efficiently. This maximizes network bandwidth and flash storage speeds, which is critical for handling the high data throughput required in AI and machine learning workloads.
2. **Cost Efficiency**: A disaggregated architecture allows for better resource allocation, reducing the need to over-provision expensive compute resources to accommodate storage demands. This leads to more economical use of rack space and power, essential for high-scale AI deployments.
3. **Improved Performance for AI Workloads**: The architecture significantly enhances performance by ensuring that the storage system can keep up with the high performance demands of AI model training. The storage system remains fast and efficient, even under the heavy loads typically associated with training LLMs and other AI applications.
4. **Optimized for Large-Scale AI**: A disaggregated architecture is particularly suited for enterprises running very large AI workloads that

require massive data storage and processing power. It supports scaling without sacrificing performance, ensuring that the storage system can handle the vast amounts of data AI workloads generate.

5. **Built-In Resilience, Security, and Governance**: Even though the architecture is disaggregated, it still retains the core benefits of NetApp's ONTAP operating system, including proven resiliency,

NETAPP AI ENGINE

NetApp's AI engine is deeply integrated with NetApp's storage and data management solutions to streamline and automate data handling for AI processes.

Here are the key components and features of NetApp's proposed AI engine:

1. **Integrated AI Data Pipeline**:
    - The NetApp AI engine automatically and iteratively prepares unstructured data for AI tasks. It captures incremental changes in the data, enabling continuous data readiness for AI applications without manual intervention. This simplifies data preparation for AI workloads by automating data classification, transformation, and processing.
2. **Policy-Driven Data Classification & Anonymization**:
    - The AI engine incorporates policies for classifying and anonymizing data. This ensures that sensitive information is properly managed and protected while ensuring the data is ready for AI applications. These policies ensure compliance with data governance and privacy regulations.
3. **Vector Embedding Generation & Storage**:
    - The engine can generate highly compressible vector embeddings, which are mathematical representations of data used for semantic search and RAG workflows. The embeddings are stored in a vector database that integrates directly with NetApp ONTAP, enabling fast, scalable AI processing.
4. **High-Scale, Low-Latency Semantic Search**:
    - By leveraging vector embeddings, the AI engine allows efficient high-scale semantic search across large datasets, which is

particularly useful for AI applications that require fast and accurate information retrieval, such as natural language processing (NLP) and machine learning inferencing.

5. **Support for RAG Inferencing**:
    - The AI engine enhances AI models with RAG capabilities. RAG combines generative AI with information retrieval techniques, allowing the AI to produce more accurate and contextually relevant responses by incorporating real-time data from vast datasets.
6. **Seamless Integration with NetApp ONTAP**:
    - The AI engine is tightly integrated with NetApp's ONTAP storage operating system. This ensures that AI workloads benefit from ONTAP's proven data management capabilities, including security, governance, and high availability, without compromising performance.
7. **Disaggregated Storage Architecture**:
    - The AI engine enhances performance by working in conjunction with NetApp's disaggregated storage architecture. By separating compute and storage resources, the engine optimizes data processing for AI, enabling efficient scaling and reducing infrastructure costs.
8. **Cloud Integration**:
    - The AI engine integrates with NetApp's cloud services, allowing AI workloads to seamlessly operate in multi-cloud and hybrid environments. Its cloud-native approach ensures that AI data can be ingested, processed, and stored across various platforms, facilitating AI workflows across on-premises and cloud infrastructures.

COMPETITIVE ENVIRONMENT

NetApp is following the lead of competitors like VAST Data and WEKA with its comprehensive approach to AI. It's also pushing well beyond traditional storage boundaries with features like its AI Engine.

VAST Data was the earliest mover in adding data manipulation capabilities directly into the data platform, recognizing that integrating data management capabilities directly into the platform could bring greater efficiencies to the AI pipeline. VAST's vision goes beyond what NetApp is promising, delivering an extensive set of data manipulation (and database) technologies applicable to a range of AI and data science problems.

WEKA is another early mover in building a disaggregated AI data platform. The WEKA Data Platform software is built around an innovative disaggregated architecture with a resilient parallel file system. This allows WEKA to deliver the performance and scale required for enterprise AI training and inference workloads across distributed edge, core, and cloud environments.

While WEKA has yet to unveil integrated data manipulation capabilities like VAST and NetApp, it wouldn't be surprising to find that they're aggressively working on it.

NetApp's traditional competitors, like Pure Storage, Lenovo, HPE, and Dell Technologies, aren't yet following the new path. Instead, they're so far choosing to focus on scalable performance while leaving data manipulation to third-party tools.

While this approach may be less performant and scalable, it allows customers to adopt tools independent of underlying storage. Many customers will want such an unopinionated storage stack.

ANALYSIS

NetApp is one of the industry's most conservative companies, rarely announcing products or services until those offerings are available. Its conservatism, coupled with its long-time presence in nearly every enterprise data center, gives NetApp a higher level of credibility than might be given to some of its less-enterprise-tested competitors.

Companies like VAST will likely find deals stall as customers wait to see how NetApp will ultimately deliver on the vision, while other competitors will be asked to defend their more traditional approaches.

There is a risk for NetApp in announcing new capabilities so far in advance. It pressures the company to quickly deliver features aligned to the vision. The longer NetApp waits, the more opportunities traditional storage competitors have to reveal their solutions. At the same time, forward-looking competitors like VAST and WEKA will benefit from NetApp's implicit endorsement of the

disaggregated approach each has long championed. Not all customers will be willing to wait.

The vision NetApp described isn't as radical as it may seem. The company has been quietly building toward this moment for years as it delivered increasing levels of integration between its products, including building a data fabric bridging on-prem and cloud storage. Leveraging these same core technologies into a new disaggregated architecture and AI tools to evolve NetApp's Intelligent Data Infrastructure is a compelling and natural next step for the company.

Will enterprises choose the flexibility of legacy storage, or will they prefer the performance and scalability of a highly integrated data platform? It's too soon to predict. The beauty of NetApp's approach is that it delivers a data infrastructure to match nearly any enterprise data needs, on-prem or in the cloud.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.