VAST

White Paper

VAST on Cloud for Azure

VAST Data Fueling Cloud-Native AI Workflows in Azure

Table of Contents

Executive Summary 3

Introduction 3

VAST AI OS on Azure: A Consistent Data Foundation 4

- VAST on Cloud Deployment Model 5
- Advanced Data Reduction in Azure 6
- More Than Storage, An AI Data Platform 6

VAST DataSpace: Bringing Enterprise Context to Azure AI 7

Polaris: Global Fleet Management 8

Drive Azure AI Results with VAST Data 9

Executive Summary

AI is moving fast, and enterprises are rethinking where their cloud strategy should take them. Through services like Foundry Models, Foundry Agent Services, and Azure AI Search, Microsoft Azure provides businesses with the tools to evaluate leading AI models, orchestrate multi-agent systems, and run scalable inference and Retrieval-Augmented Generation (RAG).

However, AI and application services are only part of the picture. The real value of AI and analytics lives in data, and that data is usually scattered, tucked away in isolated storage systems, application silos, multiple databases, and specialized cloud storage services. That scattering keeps AI developers, data scientists, and applications from the data they need, and it multiplies copies, cost, and complexity.

VAST on Cloud (VoC) makes the VAST data platform accessible on Azure. The VAST AI Operating System (VAST AI OS) runs on Azure virtual machines (VMs) creating a unified data foundation for Azure AI services. Instead of building another isolated data silo, organizations can extend one data architecture across the enterprise and into Azure, allowing Azure applications and compute to work with a single source of truth.

Introduction

Hybrid cloud adoption has often led organizations to use a different data solution for each workload, one for file applications, another for object storage, and a separate layer for databases. AI pipelines compound that fragmentation with staging areas, vector stores, checkpoints, and duplicate copies of the same data.

For hybrid AI deployments in Microsoft Azure, this creates several challenges:

- Data movement: Large, fragmented datasets stored on-premises need to be migrated, staged, or copied before Azure AI services can use them.
- GPU inefficiency: GPU instances are only cost-effective when pipelines keep the hardware fully saturated with business-relevant content.
- Data duplication: Copies created for AI inference, RAG pipelines, and regional replication consume capacity and elevate the risk of data drift.
- Governance complexity: As data spreads across data centers and public clouds, security, access, and implementing consistent controls become increasingly difficult.

Each copy is three problems at once. It is time wasted waiting for data to be available. It is a staleness risk, because copies drift out of sync. And it is a potential governance risk, because you lose track of who can touch it.

VAST AI OS on Azure: A Consistent Data Foundation

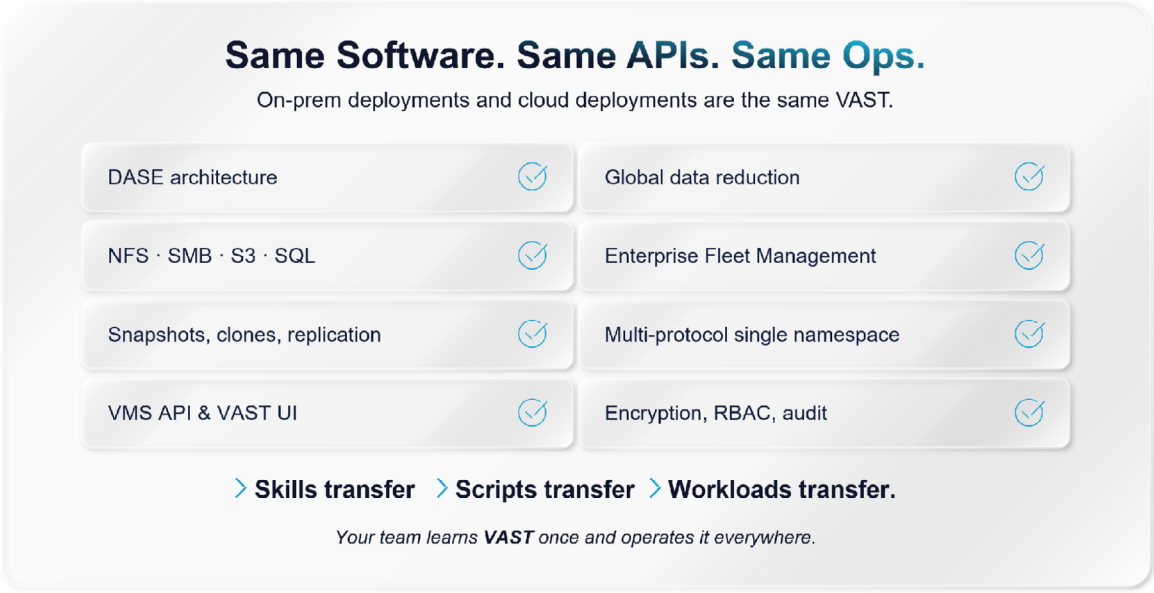


Figure 1. Same Software. Same APIs. Same Ops.

VAST AI OS brings VAST's data architecture into Microsoft Azure, extending a single, consistent data foundation from the data center to the cloud.

At its core is VAST's Disaggregated Shared Everything (DASE) architecture. With the virtual EBox (Everything Box), VAST combines the CNode data services and the DNode storage functions into a single Azure VM instance.

Each virtual EBox runs three VAST containers (Figure 2). The CNode container manages cluster metadata and serves incoming data requests across the environment. Working alongside it, two DNode containers link the VM's local SSDs to the NVMe fabric, routing storage commands directly between the drives and the cluster's CNodes. All EBoxes have simultaneous access to the exact same shared backend data. Should one DNode go offline, such as during an update, the other DNode will connect all the EBox's SSDs to the fabric.

To scale performance or capacity, you simply add more EBox VMs to the VAST cluster in Azure.

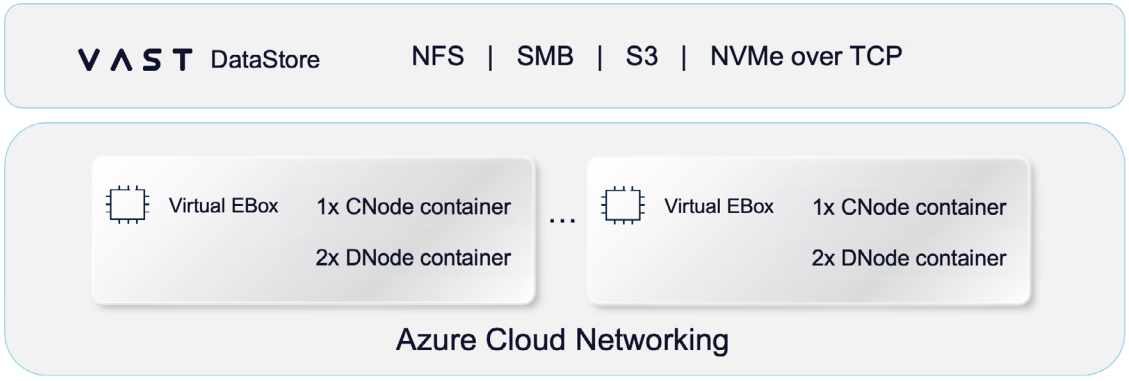


Figure 2. VAST on Cloud Architecture

The VAST DataStore stores unstructured and structured data. Within VAST DataStore, the Element Store maintains protocol-independent data elements, while the protocol layer provides multiprotocol access to those elements. Data written through S3, for example, can also be accessed through NFS when the same data is exposed through both protocols.

This unified approach extends to access control. VAST stores file and directory permissions internally in a protocol-independent format and translates them according to the security model configured for each view, helping maintain consistent authorization across NFS, SMB, and S3.

For vector retrieval, VAST InsightEngine enforces the access controls associated with source data during query execution, filtering vectors and chunks according to the requesting identity, so AI engineers and data scientists only receive information they are authorized to access.

VAST on Cloud Deployment Model

VAST AI OS on Azure uses L-family storage optimized virtual machines. They are engineered for workloads requiring high NVMe throughput, low-latency I/O, and high network bandwidth. This makes them well suited for AI data processing, model fine-tuning, production inference and RAG, and large-scale analytics.

VAST on Cloud for Azure provides reliability, availability, and durability through multiple mechanisms using Cloud-HA. The Element Store leverages erasure coding across the NVMe disks within the virtual machines providing continuous data availability in the event of a disk or node failure. Metadata and write buffers are synchronously persisted to Azure managed disks, while newly written data is continuously streamed to Azure Blob. In the unlikely event of an Availability Zone failure, upon service restoration, managed disks will be reattached to instances, making metadata available nearly instantaneously, and data will be lazy-loaded from Blob.

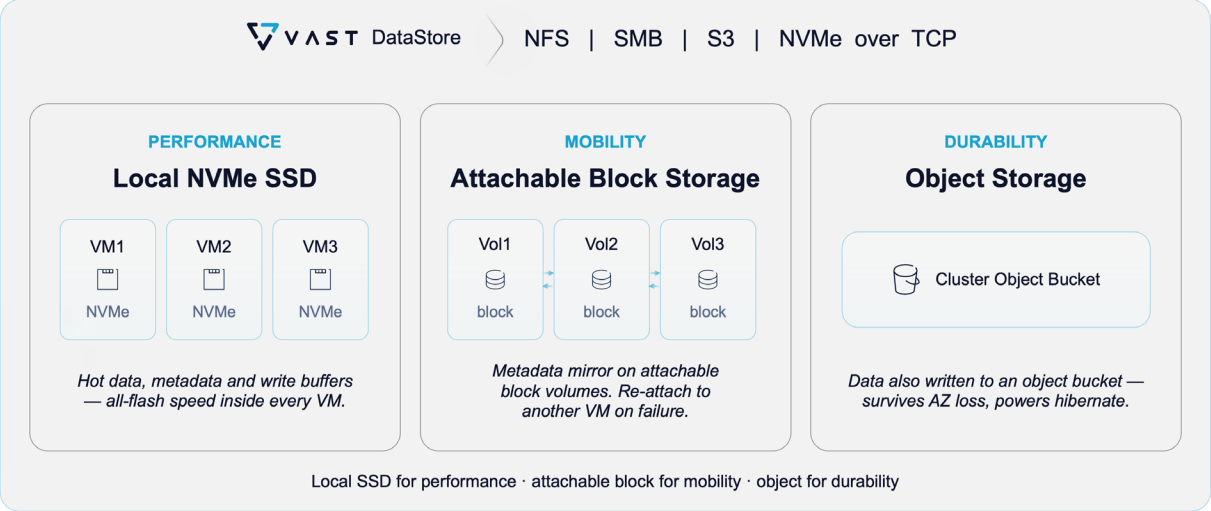


Figure 3. VAST Cloud-HA

A VoC deployment in Azure requires a minimum of three nodes. A single-node deployment is available for interoperability testing, but it is not intended for high performance and offers no resiliency.

Advanced Data Reduction in Azure

Beyond standard compression and deduplication, VAST DataStore adds similarity-based reduction. It shrinks the storage needed for data that is similar to what is already stored, even when it is not identical.

VAST combines several reduction techniques. It breaks data into variable-sized chunks, averaging 16 to 64 KB, and uses strong hashes to find exact duplicates. Similarity hashes then identify chunks that share much of their data even when they are not identical. For a similar chunk, VAST reuses the compression dictionary from a reference chunk and stores only the difference, not another full copy. This finds reductions that ordinary deduplication misses, where a small change makes an otherwise similar block look unique.

More Than Storage, An AI Data Platform

VAST AI OS on Microsoft Azure is more than file and object storage. VAST DataBase adds an analytical database for structured and vector data, while VAST DataEngine and VAST Event Broker support event-driven processing.

VAST DataBase keeps structured information close to the files and objects associated with it. Applications can work with business records, metadata, and unstructured content without first moving that information into a separate database environment.

Vector search is built natively into VAST DataBase. Embeddings created from business-relevant content can be stored with the source data they represent, ready to be queried for semantic search and RAG. Vector search scales from millions to trillions of vectors with sub-millisecond retrieval latency. This removes the need to operate and synchronize a separate vector database for Azure-based AI applications.

VAST DataEngine and VAST Event Broker support workflows that respond as data changes. New or updated information can trigger downstream processing, including analytics and AI pipelines.

This gives organizations a way to run storage, database, vector search, and event-driven services within the same VAST environment rather than stitching together separate data systems.

VAST DataSpace: Bringing Enterprise Context to Azure AI

For many organizations, adopting Microsoft Azure does not mean moving every dataset to the cloud. Large repositories often stay in the corporate data center, but AI developers and data scientists want to take advantage of Azure GPUs and AI services.

VAST DataSpace bridges the divide. It extends VAST data elements across clusters in the data center, edge locations, and in Azure regions, making business-related data available to compute wherever it is needed.

At the heart of DataSpace lies VAST Global Access. Data stored on a VAST cluster in the data center can be marked as a global folder and made accessible to a satellite cluster in Azure.

The VoC satellite cluster:

- Serves the data to local applications.
- Caches frequently accessed data in its own VAST DataStore.
- Fetches or prefetches data based on access patterns instead of copying a whole file or dataset up front.

To keep data strictly consistent, Global Access uses read and write leases. A write lease guarantees that only one VAST cluster can change a data element at any given time, and a read lease makes sure cached data is current before an AI developer or application uses it.

When any cluster, including the origin cluster, receives a request to write to a global folder, it requests a write lease from the origin cluster. If no other cluster holds a write lease to the element, the write lease will be issued to the requesting cluster.

As a result, one location never overwrites another's changes, and companies are spared the manual reconciliation of conflicting versions, and the lost edits that last-writer-wins systems allow.

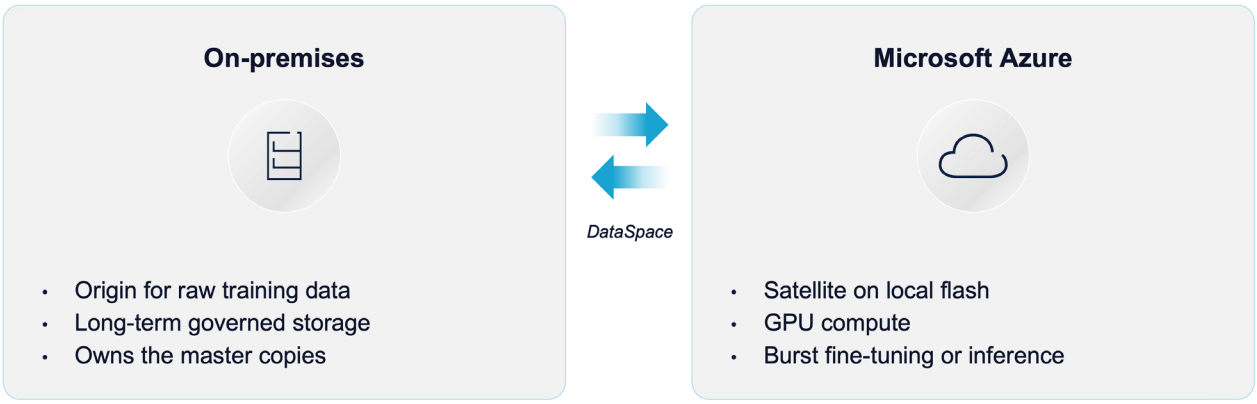


Figure 4. Hybrid AI with VAST DataSpace

Cached Global Clones are another option for Azure workloads. A VAST Global Clone creates a clone on one cluster from a snapshot on another. With a cached clone, only the data read is transferred, and new writes stay local.

This feature is useful for development, testing, and AI projects in Azure. Teams can quickly make writable copies of production data near Azure compute without first copying and maintaining a full second production dataset.

Polaris: Global Fleet Management

VAST Polaris provides a centralized control plane for VAST on Cloud provisioning and management. Administrators can use it to deploy clusters, add capacity, manage upgrades, replace nodes, hibernate systems, and maintain a consistent desired state across deployments.

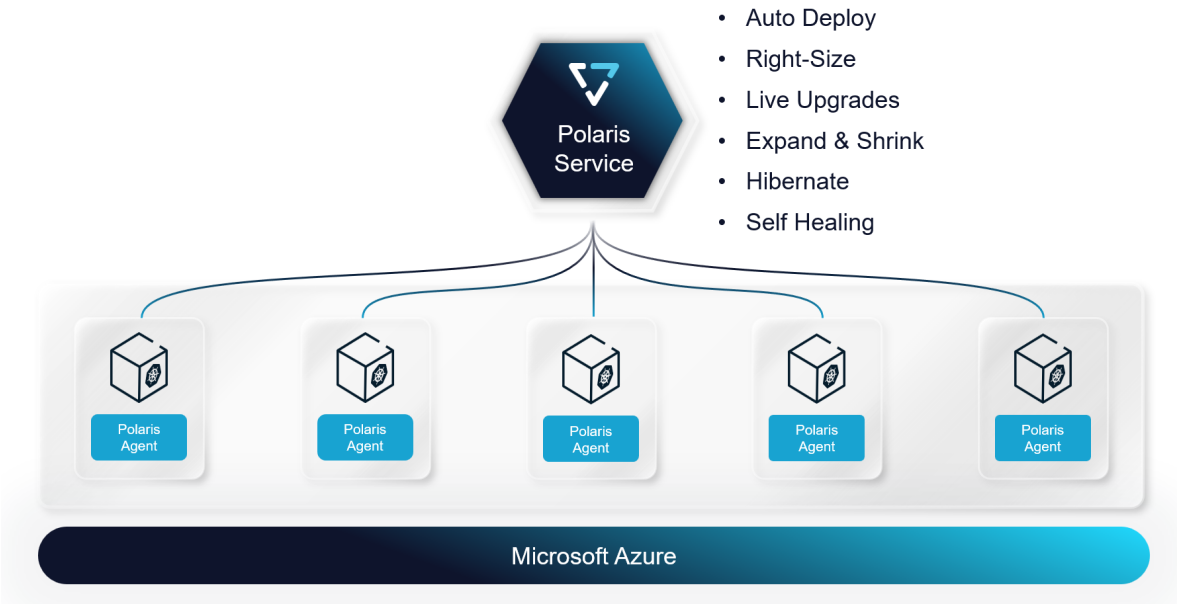


Figure 5. Polaris

As soon as a VAST on Cloud cluster is deployed in Microsoft Azure, day-to-day management is done through the Polaris Service, which talks to the cluster over a lightweight Polaris Agent.

Polaris uses an intent-driven architecture to monitor VAST AI OS and maintain a consistent configuration that is defined by the VAST administrator.

Drive Azure AI Results with VAST Data

AI and application services depend on current, reliable data. The challenge is that enterprise data is often spread across on-premises storage systems, storage tiers, data lakes and databases, and cloud services.

VAST on Cloud brings VAST AI OS into Microsoft Azure, providing a unified platform for file, object, table, and vector data. VAST DataSpace extends access from the data center to Azure, ensuring data remains where it makes sense while making it accessible to cloud-based workloads. And Polaris pulls it all together, providing centralized management of distributed VAST clusters.

The result is a single source of truth that gives AI developers, data scientists, and applications broader access to enterprise data without compromise.

One Data Fabric

AI workloads can run wherever compute capacity is available while maintaining control of data and security policies.

Maximum Utilization

GPUs can access data from VAST storage with performance consistent with local NVMe devices.

Consistent Policies

VAST DataSpace and Polaris enforce a single trust zone across environments.

©2026 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.