White Paper

Comparing Storage Platforms for Big Data Applications

A massive shift is underway in data analytics. Workloads and applications are moving away from the Hadoop File System (HDFS) and towards more scalable and performant data storage platforms. In this paper, we will briefly touch on why this shift has taken place and compare several storage options for modern data analytics.

The End of the Hadoop File System

Hadoop along with the Hadoop file system was the premier data analytics platform of the previous decade. Hadoop was architected around the technical limitations of the time – where networking was the primary bottleneck and compute with local storage was relatively fast.

HDFS implemented the distributed shared-nothing architecture creating copies of data on multiple nodes with direct attached storage. This model allowed for local processing of data without the need to traverse networks and worked well for batch-oriented workloads along with the comparatively small data sets of the day.

Several factors have led to Hadoop's decline:

- Analytics are moving away from batch processing to real-time queries
- Data sets are growing beyond the scale that is practical to manage with HDFS
- Modern data analytics tools such as Spark, Presto, Elastic, and Kafka now support S3 and NFS

VAST Data Ranked #1 for Analytics Use Case in the 2022 Gartner® Critical Capabilities for Distributed File Systems and Object Storage Report

Requirements for Big Data Storage

The widespread adoption of analytics as well as artificial intelligence is putting new demands on storage infrastructure. Applying some of the lessons learned from HDFS can help you make an educated decision that will support evolving workloads well into the future

Predictable responsiveness regardless of file size and access patterns

HDFS struggled with small file IO limiting its flexibility, ensure the platform you choose won't limit future workloads

Scalable performance designed for massive concurrency

Your data repository must be able to handle access by multiple applications simultaneously as well as multi-threaded workloads for high performance

Real-time responsiveness as data sets grow larger

The method of scaling can have performance impacts: Solutions that implement tiering may suffer from inconsistent performance as data is retrieved from lower-performing tiers. Additionally, shared-nothing architectures or those that rely on federation ("cluster of clusters") may plateau or degrade performance beyond a certain point.

Support for all types of analytics including batch, Adhoc, ML, and others

An ever-expanding ecosystem of analytics applications requires more flexibility than those designed specifically for Hadoop

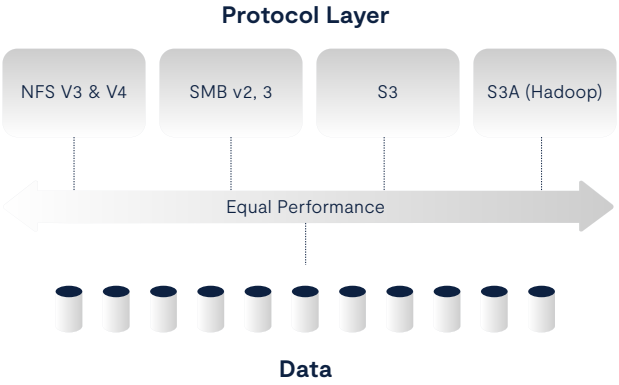
Low total cost of flash ownership

Choose a platform with all flash performance that will enable you to grow your data set without crushing your budget

Modern Data Workflows are multiprotocol

There is no denying that the momentum behind object storage and specifically the S3 API effectively make it the de facto choice for analytics workloads. That said, many analytics applications leverage the NFS protocol. The flexibility to support both S3 and NFS provides significant benefits:

- Support for applications that require or prefer POSIX
- NFSoRDMA + GPUDirect provides extreme performance
- Future-proof data access
- Avoid data sprawl and consistency problems



Modern Challenges Require a Modern Architecture

To evaluate storage options for analytics it's useful to review the storage architectures available today:

Shared Nothing

Shared nothing, scale-out architectures were first designed over 20 years ago – optimized for slow networks, mechanical hard drives, and data sets in the low 10s of petabytes. The shared-nothing model unifies storage media and computing into a single basic building block called a node. These nodes present NAS or Object API services to applications as well as perform data protection task such as replication and erasure coding. Shared-nothing storage systems today support flash media, but the architecture, designed around the constraints of mechanical hard drives can dramatically reduce the endurance under active workloads.

Each node of a shared-nothing cluster consists of compute with direct attached storage (DAS). As mentioned previously this allows for local processing of data, however IO also requires communication between the members of the cluster. This inter-cluster communication grows exponentially as clusters expand and is the limiting factor for scaling performance and capacity. The combination of storage and computing as a tightly coupled unit means that capacity and performance cannot be independently scaled. This lack of flexibility typically leads customers to over-provision one factor to meet the requirements of the other.

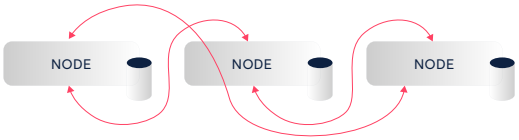
Disaggregated Shared Everything (DASE)

DASE is a redesign of scale-out storage architecture based on high-speed, low-latency networking and flash storage built for exabyte data sets. DASE addresses the key limitations of Shared Nothing by disaggregating the compute and storage across a high-speed low latency NVMe fabric.

In contrast to Shared Nothing nodes, every VAST CNode (Compute Node) has equal access to shared, persistent NVMe devices in highly available storage enclosures. This high-performance infrastructure enables VAST CNodes to be stateless machines that do not have to coordinate IO requests with each other thereby eliminating the scaling limitations and rigidity of Shared Nothing.

The NVMe fabric connects the CNodes to the storage enclosures in with just a few microseconds latency providing the benefits of Direct Attached Storage with none of the restrictions. This allows users to scale the performance of a cluster by adding and removing front-end CNodes from pools independently from the cluster's capacity, managed by adding storage enclosures.

Legacy, Shared-Nothing Cluster Architecture are Limited

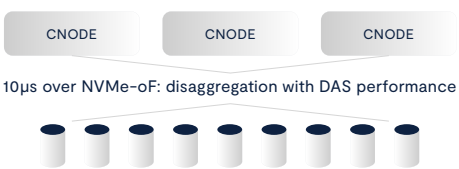


crosstalk, rebuilds and interdependencies increase geometrically with cluster size

SHARED-NOTHING

crosstalk, rebuilds and interdependencies increase geometrically with cluster size

Disaggregated, Shared-Everything Cluster Architecture



VAST DASE

zero cross talk, independent scaling: CPUs & drives, No rebuilds on server failure

The VAST Advantage

Comparing VAST Disaggregated Shared Everything against HDFS with Shared Nothing illustrates the benefits of moving to a modern platform. The efficiencies made possible with an architecture based on NMVe and compounded with data reduction plus low overhead data protection are dramatic when compared to HDFS storage with triple-copy data protection. Customers gain additional benefit from reduced rack space and power energy savings.

VAST Data vs HDFS

	HDFS	VAST
Total Rack Units	336	80
Raw Capacity	30 PB	6 PB
Available Capacity	10 TB	10.7 PB*
Storage Throughput	180 GB/s	360 GB/s
Racks	9	1
Cloudera Server Nodes	150	36

** 2:1 data reduction with data protection applied*

80%
Less Capacity Needed

2x
Faster Performance

88%
Less Rack Space

80%
Less SW License Cost

	VAST	Pure Flashblade	Minio	Cloudera Ozone	Dell PowerScale
Architecture	DASE	Shared-nothing	Shared-nothing	Shared-nothing	Shared-nothing
HDD/Hybrid	✗	✗	✓	✓	✓
Built for All Flash	✓	✓	✗	✗	✗
Low cost hyperscale flash (QLC)	✓	✓*	✗	✗	✓*
Data Protection	N+ 4	N+2+	N+4-N+8	N+3 or 3 x replica	N+1-N+4
Data Protection Overhead	3-11%	13.3-33%	25-50%	33-200%	20-66%
Asymmetric Scaling	✓	✗	✗	✗	✗
Protocol Support					
NFS	✓	✓	✗	✗	✓
HDFS	✗	✗	✗	✓	✓
S3	✓	✓	✓	✓	✓
Multi-Protocol File / Object Interoperability	✓	✗	✗	✓	✗
Data Reduction					
Duplicate Block Elimination	Inline, always on	✗	✗	✗	Post-process
Single Block Compression	Inline, always on	✓	✓	✓	Select models
Global, Cross-Block Compression / Similarity	Inline, always on	✗	✗	✗	✗

*Dell and Pure added QLC support in 2022 – QLC endurance information not available

Included Solutions

FlashBlade is Pure Storage's scale out Object and NAS platform. Flashblade repackages the shared-nothing, scale-out model into a chassis that holds 15 all-flash nodes in the form of plugin blades. Flashblade supports the S3 object protocol in addition to NFS and SMB but files and objects occupy independent namespaces. Flashblade is a follow-on to their FlashArray all-flash block offering which has recently introduced file services as well.

Minio is an opensource object storage platform based on classic shared nothing architecture. Available as software only Minio runs on industry-standard x86 servers with direct attached storage. Minio is strictly an object storage platform serving the S3 API with no support for NAS protocols.

Cloudera Ozone based on open source Apache Ozone is Cloudera's replacement for HDFS. Ozone seeks to address many of the most painful shortcomings of HDFS (small object, high file count scaling limitations) by shifting to an object storage architecture. The move to object architecture does not equal native support for the S3 API – rather Ozone supports S3 along with its own HDFS compatible OzoneFS via gateway nodes.

Dell Powerscale, formerly Isilon is an almost 20-year-old shared-nothing, scale-out NAS that supports all-flash, HDD/flash hybrid and all HDD nodes in heterogeneous clusters. Isilon is, as one would expect from a 20-year-old product, is feature rich including automated tiering between Isilon pools and to the cloud.

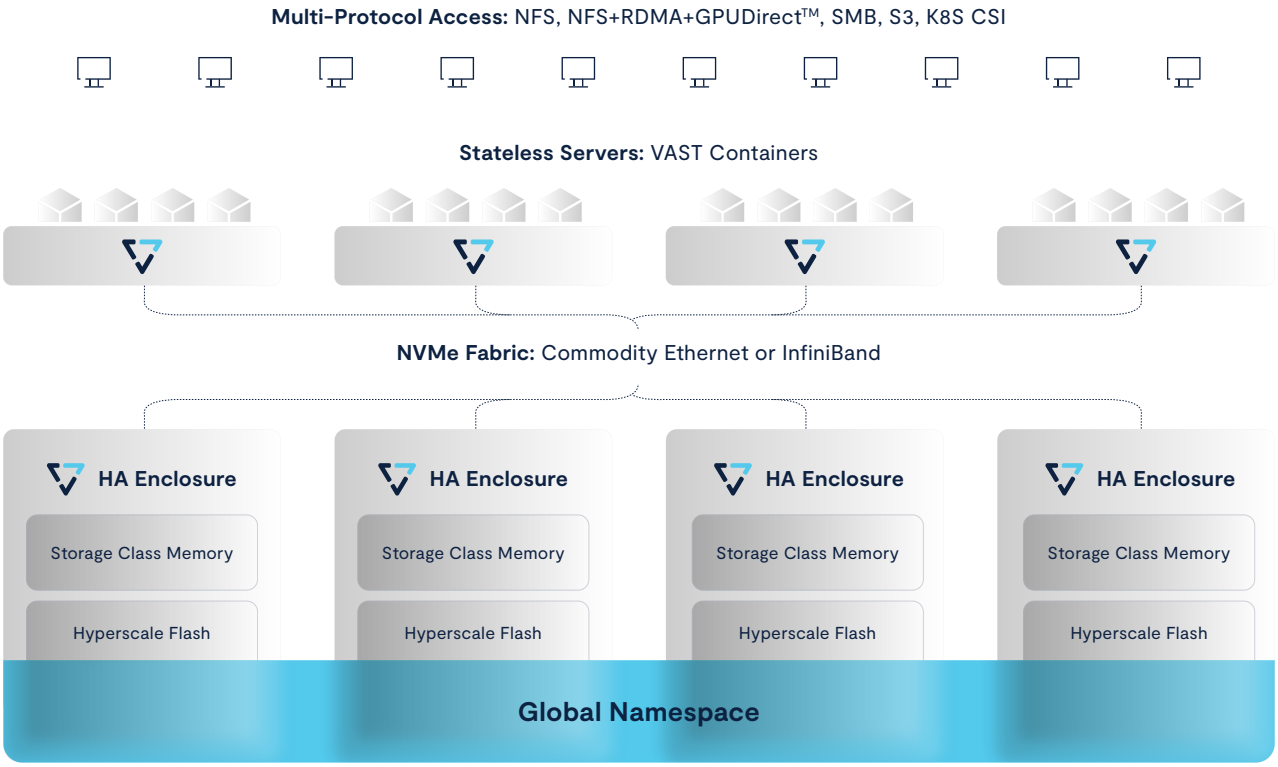
Archive Tier S3 Considerations

We have not included archive grade S3 solutions in this comparison guide. Analytics workloads using the S3 API have evolved to expect higher performance. Many of the on-premises S3 solutions were optimized for low cost over performance and are generally not suitable for Big Data.

A Brief Look at Universal Storage

VAST's Universal Storage reduces the cost and complexity of all-flash file, and object, to create a data platform for all workloads. Built to take full advantage of NVMeOF, Storage Class Memory and dense hyperscale flash (QLC) VAST drives efficiencies that aren't possible with legacy architecture:

- Locally decodable erasure codes provide N+4 data protection with as little as 3% overhead
- Similarity data reduction yields greater reduction than any other storage system
- Global flash translation using Storage Class Memory write buffer extends hyperscale flash endurance for up to a decade.



The result is a file and object system that delivers all-flash performance for the most demanding big data workloads, scales from petabytes to exabytes all with economics that make it affordable for any enterprise workload.

For more information on Universal Storage and how it can help you solve your application problems, reach out to us at hello@vastdata.com.

02/23