Hadoop Modernization with the VAST Data Platform

Paving the Way for AI-Readiness and Real-Time Analytics on a Single Platform

Table of Contents

Executive Summary 3

Current State: What is Hadoop and Why It's Time to Move On 4

- What's Broken in a Hadoop Architecture 4
- Why These Challenges Matter Today 6

The VAST Data Platform Overview 7

- Past to VAST: A Single Data Platform for All Your Data and Workloads 8
- Hadoop Modernization with VAST Data by the numbers 9
- Benefits of Hadoop Modernization with VAST Data 11

Upgrading Beyond Hadoop is Easy 13

- 1. Strategic Planning and Architectural Design 13
- 2. Deploying VAST DataBase as the Core Data Platform 13
- 3. Migrating Legacy Data to VAST Storage for Long-Term Storage 13

Modernize for the AI-Era with VAST Today 14

Executive Summary

Modern enterprises are at a critical point where data growth and infrastructure constraints demand a strategic shift in data platform technologies. Legacy Hadoop infrastructures, particularly Hadoop Distributed File System (HDFS), are increasingly inadequate for handling the scale, complexity, and performance demands of modern data analytics and AI workloads.

To be prepared to take full advantage of AI for business, organizations must modernize their data platforms to ensure efficient data management, reduce operational costs, and unlock the full potential of AI and big data analytics. Moving beyond Hadoop with the right data platform enables analytics performance, scale, and security.

Since its launch in 2019, VAST Data introduced its revolutionary system architecture and has continued to build capabilities that align with modern data processing and management needs.

This whitepaper explores the benefits of Hadoop modernization and migration using the VAST Data Platform, which offers a modern, flash-based architecture with enhanced data reduction, improved performance, and streamlined data management.

	From Past	To VAST
Nodes	500 Hadoop nodes	**125 CPU & GPU nodes** **75% reduction**
Format	HDD-Based	**All Flash**
Overhead	200% Overhead (Triplication)	**3% Overhead** **(Locally Decodable Erasure Codes)** **99% Reduction**
Power Consumption	450 kW	**160 kW** **64% reduction in power consumption**
TCO	$21MM	**$12M** **60% reduction in TCO**

Current State: What is Hadoop and Why It's Time to Move On

Hadoop emerged in the mid-2000s as a revolutionary platform for big data, enabling enterprises to store and process massive datasets using distributed, commodity hardware. At its core, the Hadoop Distributed File System (HDFS) and MapReduce computational model provided a significant step forward for organizations managing large-scale unstructured data. However, while Hadoop transformed data management for its time, its design has become a hindrance in modern data centers.

Hadoop is now widely recognized as one of the most resource-intensive and inefficient systems in the enterprise landscape, with fundamental architectural decisions creating significant operational and performance challenges. As data volumes have grown and real-time analytics and AI workloads have become the norm, the inefficiencies of HDFS and Hadoop's broader architecture are impossible to ignore.

What's Broken in a Hadoop Architecture

Hadoop's Data Center Inefficiencies

Hadoop deployments are notorious for their heavy reliance on distributed servers that consume vast amounts of power and require significant physical floor space. The architecture leads to sprawling data center footprints, driving up operational costs for both energy and real estate. As workloads grow, so does the infrastructure, compounding inefficiencies and straining resources.

Hyperconvergence Creates Inflexibility

Hadoop's tightly coupled architecture forces organizations to scale compute and storage simultaneously, regardless of actual workload demands. This hyperconvergence means that when you add compute nodes to meet processing requirements, you also add unnecessary storage capacity—and vice versa. The result is significant resource overprovisioning, where one component inevitably becomes overkill, leading to wasted infrastructure investment and inefficiency.

HDD-Optimized Data Structures Waste Compute Resources

Hadoop's data structures were designed to accommodate the limitations of traditional hard disk drives (HDDs). Large row groups and partitions were necessary to achieve acceptable performance on spinning disks. However, this design creates inefficiencies when modern compute nodes must sift through massive chunks of data during MapReduce and shuffle operations. The CPU resources that could be used for analytics or AI workloads are instead wasted on processing oversized data chunks optimized for legacy storage.

Shared-Nothing Architecture Limits Modern Workloads

Hadoop's shared-nothing architecture, where each node operates independently and manages its own data, fundamentally limits the platform's ability to support modern transactional and real-time workloads. This model:

- Prevents efficient stream ingestion and appends, as cross-table updates and change data capture require a level of transactional consistency that Hadoop cannot provide.
- Hinders advanced workloads, such as AI-driven analytics, which demand seamless cross-node coordination and near-real-time data processing.
- Forces organizations to adopt complex and costly workarounds, such as re-partitioning and rewriting large data sets, to perform basic operations like updates and appends.

Inefficient Data Protection Bloats Storage Costs

Hadoop relies on a triplication model for data protection, storing three copies of every file across the cluster. While this approach was appropriate for networks and drives in the early 2000s, it is now outdated. Modern techniques like erasure coding offer higher resiliency with far less storage overhead, but Hadoop has no support for such efficiencies. This results in excessive storage consumption and increased costs.

Disaster Recovery Strategies Double Infrastructure Costs

Hadoop's shared-nothing and triplication model also creates inefficiencies in disaster recovery. Operational clusters often require replicated clusters to achieve failover capabilities, effectively doubling infrastructure costs without providing comprehensive protection for all workloads.

Lack of Global Data Reduction

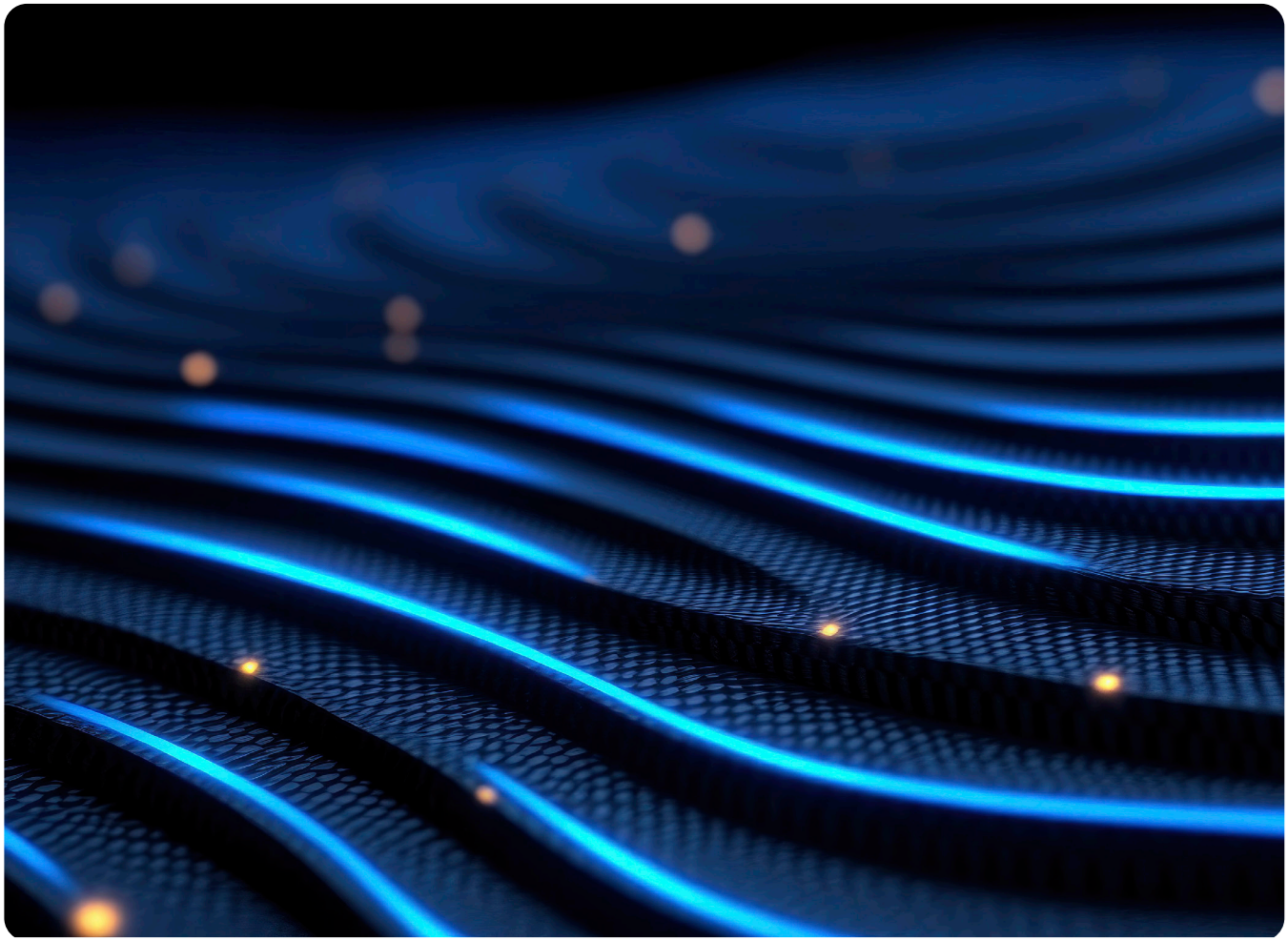
Hadoop lacks built-in mechanisms for advanced data reduction, such as fine-grained deduplication or compression. This leads to massive duplication across the environment, particularly when creating and aggregating multiple reports. The result is wasteful data sprawl, further straining storage and compute resources.

Why These Challenges Matter Today

As enterprises transition to AI-driven applications and real-time analytics, the limitations of Hadoop become increasingly problematic. Modern workloads demand a platform that can:

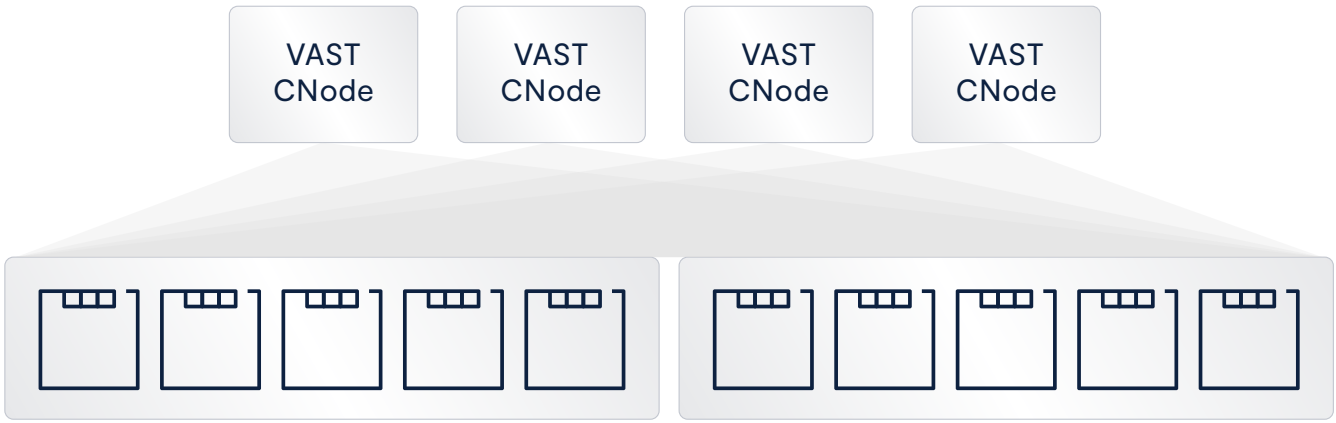
- Deliver efficiencies in infrastructure management, data center footprint, and power utilization.
- Scale compute and storage independently to avoid resource overprovisioning.
- Process smaller, more granular data chunks efficiently without wasting CPU cycles.
- Support transactional and real-time workloads, such as stream ingestion and change data capture.
- Optimize storage costs with advanced data protection techniques and data reduction capabilities.

Hadoop's inability to meet these requirements has made it a bottleneck for organizations looking to modernize their data infrastructure. The platform's outdated architecture not only inflates operational and capital expenses but also slows the pace of innovation in AI and analytics.



The VAST Data Platform Overview

Hadoop's initial design aimed to create large clusters using a 'shared nothing' architecture, where each node independently managed its data partition. This approach minimized network constraints by bringing computation closer to the data. However, as network speeds have dramatically increased, this architecture has become less efficient for modern workloads. These clusters were specifically designed for high capacity, low-performance HDDs to manage costs, and the east-west traffic created by the consistency enforcement services in a Hadoop cluster resulted in very poor transactional performance.



In 2016, VAST Data introduced a new system architecture that redefined the data platform landscape by addressing these fundamental challenges. Modern network speeds are now 800x faster than when Hadoop was created, allowing for a more flexible and efficient design. The VAST Data Platform introduces:

- **Disaggregated Compute from Storage:** Storage and compute scale independently using commodity NVMe fabrics, improving efficiency and flexibility.
- **Shared-Everything Data Structure:** A revolutionary "shared everything architecture that allows 100,000s of CPUs to access data simultaneously across a global shared namespace without east-west traffic bottlenecks, ensuring linear scalability and supporting native stream ingestion.
- **Cost-Effective NVMe Utilization:** New erasure coding and data reduction algorithms make NVMe storage as affordable as HDD, optimizing cost while maintaining performance.
- **Optimized Table Format:** The VAST DataBase table format is tailored for random-access flash media, delivering faster query performance than traditional data lakes. This approach provides significant advantages over Apache Iceberg, including superior performance, minimal maintenance requirements, and the elimination of complex partitioning.

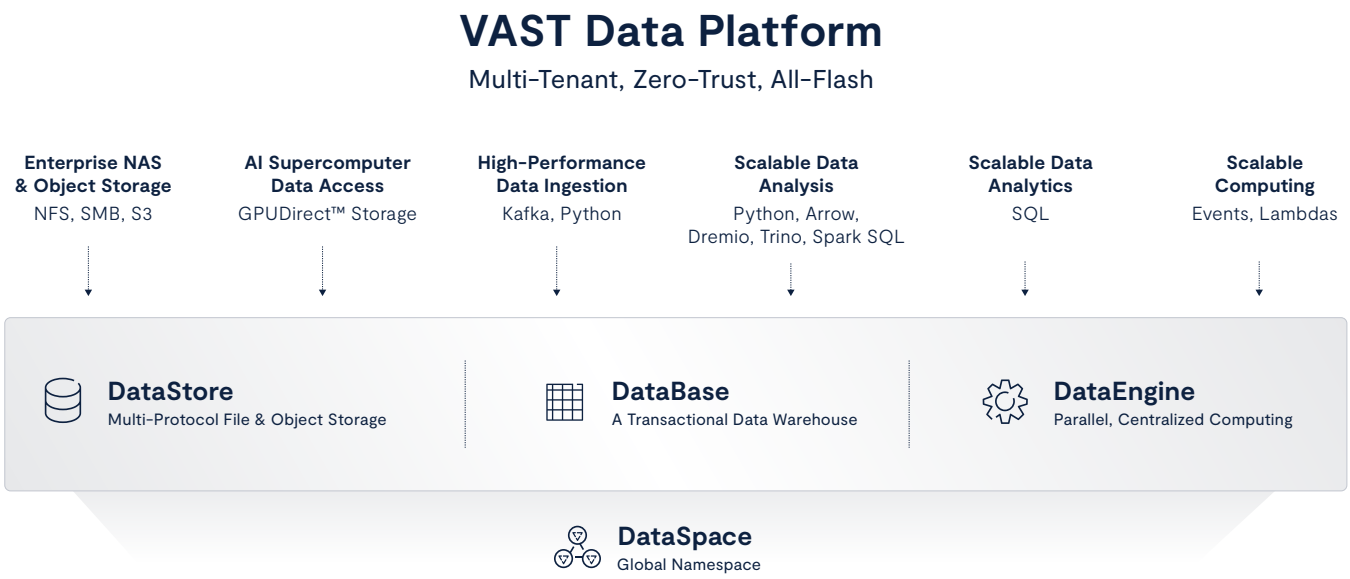
Past to VAST: A Single Data Platform for All Your Data and Workloads

The VAST Data Platform is a comprehensive solution for modern data needs, offering parallel and flash-optimized performance across file, object, and table storage. The VAST Data Platform transforms the data landscape with a powerful, unified foundation built for analytics and AI workloads. Designed to meet the unique demands of AI—real-time processing, massive scalability, and support for diverse data formats—it eliminates traditional trade-offs between cost, complexity, and performance.

The VAST DataBase redefines data architecture as the first transactional data warehouse tightly integrated with the data lake, offering true cross-table ACID compliance, exabyte scalability, and comprehensive access to all your data. Designed to deliver exceptional performance for both transactional (OLTP) and analytical (OLAP) workloads, the VAST DataBase stores newly written database rows in Storage Class Memory (SCM) SSDs and transforms them into small, columnar data chunks, optimizing query performance and minimizes overhead during updates.

The VAST Data Engine automates complex data tasks, enabling real-time insights and event-driven AI without the need for infrastructure management or coding workflows. It features a Kafka API-compatible event bus that simplifies real-time data streaming and event processing by treating Kafka as a protocol rather than a standalone application. This architecture ensures low-latency performance, eliminates data duplication, and enhances resource efficiency, making it ideal for high-speed analytics, AI model training, and inference directly on data at scale.

Together, the VAST Data Platform delivers ultra-fast performance, eliminates data movement, and empowers enterprises to unify data management, reduce costs, and scale AI-driven insights across all workloads.

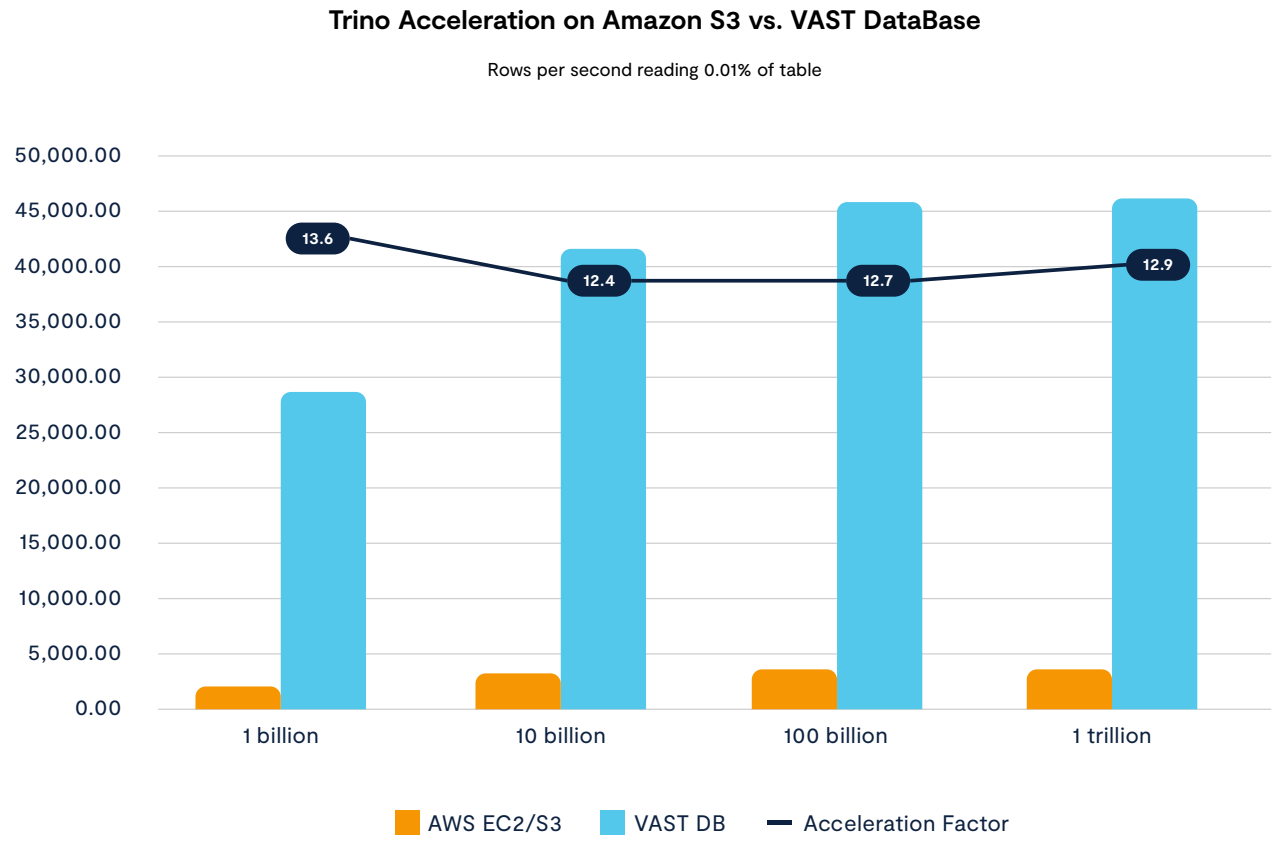


Hadoop Modernization with VAST Data by the numbers

Transforming your data infrastructure from Hadoop to the VAST Data Platform is not just about modernizing technology—it's about achieving tangible results that directly impact operational efficiency, cost savings, and scalability. The following metrics highlight the dramatic improvements organizations can achieve by migrating to the VAST Data Platform, from reducing power consumption to eliminating inefficiencies in data management and protection.

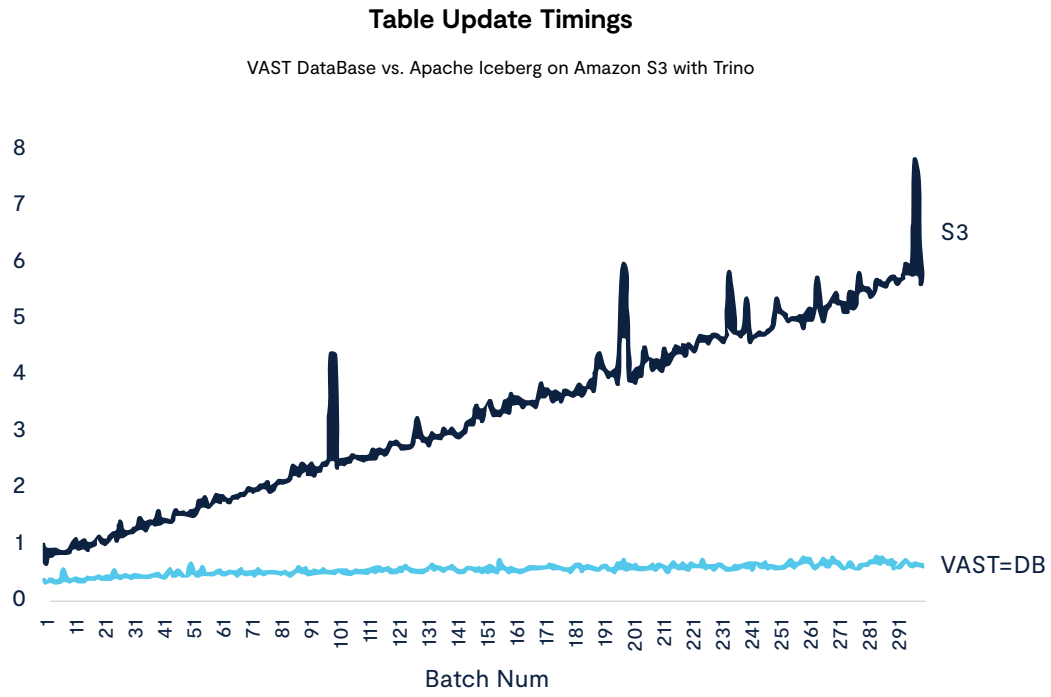
VAST Data Acceleration and Performance

VAST DataBase accelerates compute environments like Trino by leveraging small-granular columnar data objects, reducing sifting time and maximizing NVMe-native performance. For selective operations, VAST accelerates Trino by a factor of 10x+ when compared to Amazon S3.



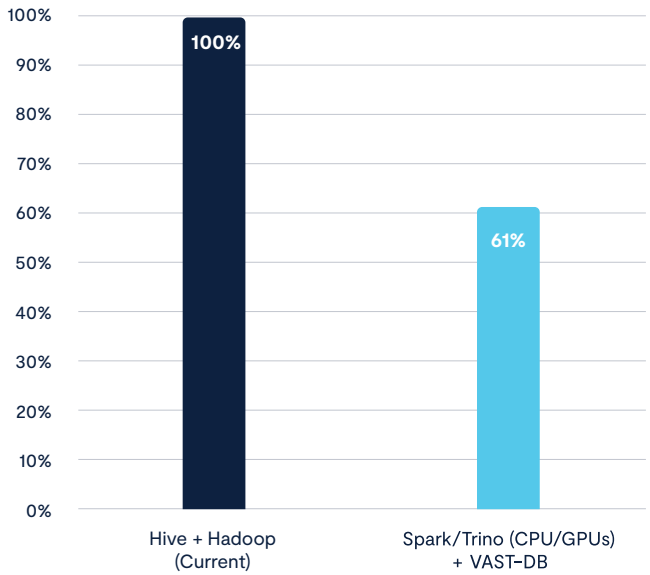
VAST Data Table Update Timing

VAST DataBase delivers consistent, sub-second table update performance, even as the number of tables scales, outperforming Amazon S3-backed Trino, where update timings increase significantly with scale.



VAST Data Total Cost of Ownership Savings

VAST significantly reduces analytics total cost of ownership by more than 60%, as shown in a six-year analysis, leveraging advanced data reduction, efficient erasure coding, and all-flash scalability to drive down storage, compute, and power costs while delivering unmatched performance.



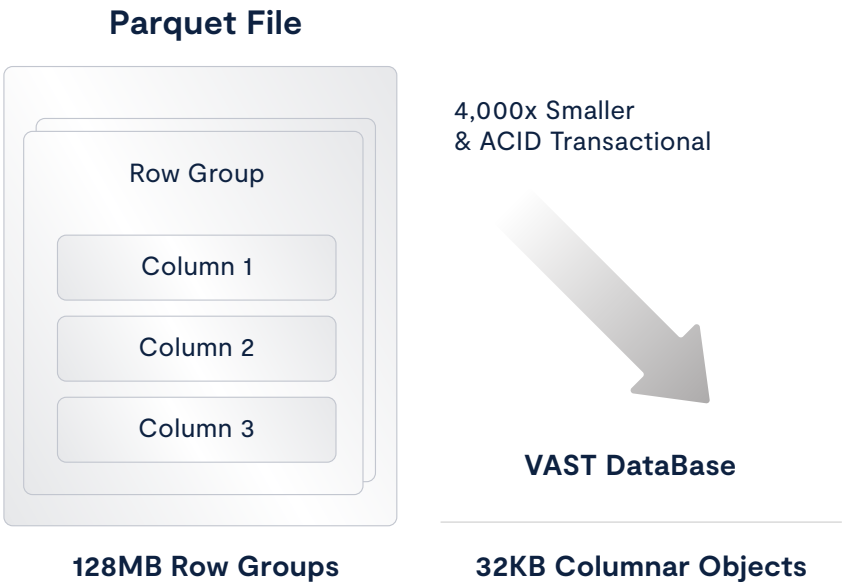
Benefits of Hadoop Modernization with VAST Data

One Platform for All Data and Workloads

The VAST Data Platform delivers a single platform for all data types and workloads from archives to analytics to AI. VAST does this at exabyte scalability with all-flash performance at hard drive economics, providing a unified namespace that seamlessly supports files, objects, and tables with strict ACID consistency for data and metadata updates. The VAST Data Platform dramatically minimizes infrastructure and data pipeline complexity as the single platform for analytics and AI.

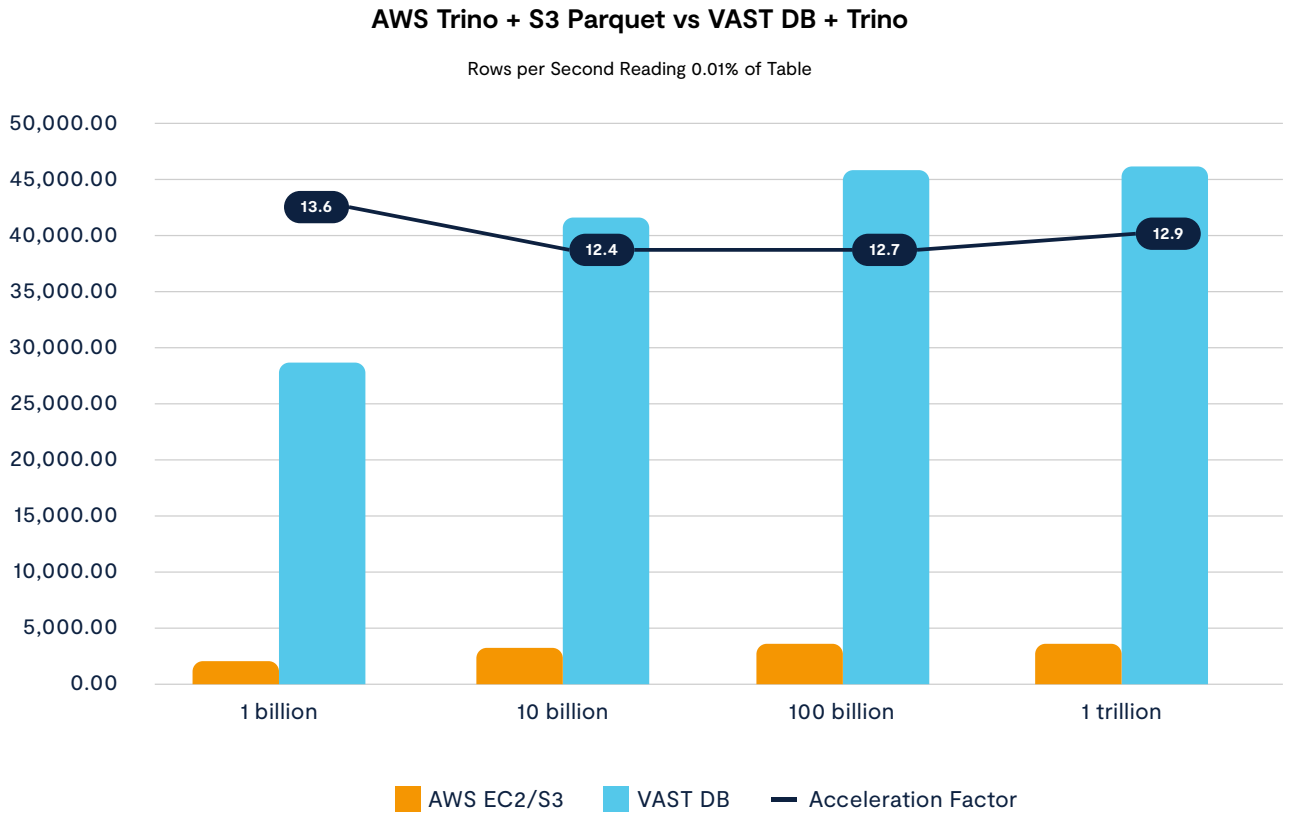
Superior Data Storage Efficiency

The VAST DataStore enhances data storage efficiency by organizing tables into finely grained columnar data chunks, similar to Parquet files but with significantly smaller segments—up to 4,000 times smaller than standard Parquet row groups. This design minimizes the need for complex partitioning and enables metadata access without scanning large files, thereby improving query performance and reducing maintenance overhead.



Enhanced Performance

VAST's flash-optimized architecture accelerates data access for AI and analytics workloads. The VAST DataBase supports transactional (OLTP) and analytical (OLAP) workloads. By storing newly written database rows in Storage Class Memory (SCM) SSDs and transforming them into small, columnar data chunks, it optimizes query performance and minimizes overhead during updates. This design allows the system to efficiently handle high-frequency updates and complex queries, providing a unified platform capable of managing diverse data workloads at massive scale.



For selective operations, VAST accelerates Trino by a factor of 10x+ when compared to AWS S3.

Storage Cost Efficiency

VAST also delivers reduced storage overhead through erasure coding and data reduction. VAST Similarity - a similarity-based data reduction technology - enhances cost efficiency by identifying patterns across the entire storage namespace at a fine granularity—up to 4x times smaller than traditional deduplication methods. This approach enables superior data reduction, thereby reducing costs associated with data storage and management.

Power Cost Efficiency

Another cost-savings dimension can be realized through reduction in power consumption. VAST Data Platform significantly reduces power consumption, delivering up to 90% energy cost savings compared to legacy systems. This efficiency not only lowers operational expenses but also supports sustainability initiatives by minimizing energy usage and reducing carbon footprints.

Future-Ready

Designed for modern data architectures, including support for AI workloads and real-time analytics at massive scale. VAST eliminates the tradeoffs between storage at scale and real-time analytics, offering a single platform designed to position organizations to use all of their data to drive business insight and AI.

Upgrading Beyond Hadoop is Easy

Transitioning from Hadoop to the VAST Data Platform, with VAST DataBase and VAST Storage, offers a modern, efficient, and proven approach to data infrastructure. This migration provides high-performance, consolidated access to both structured and unstructured data, streamlining operations and enhancing data accessibility.

Many organizations running demanding analytics and AI-driven workloads have successfully migrated by leveraging VAST's extensive partner ecosystem and advanced toolset. The migration process is designed to support a seamless transition with minimal disruption.

1. Strategic Planning and Architectural Design

- **Comprehensive Assessment:** Begin with a thorough evaluation of your current Hadoop environment, focusing on data usage, access patterns, performance metrics, and security requirements.
- **Phased Migration Plan:** Collaborate with VAST's partners to develop a phased migration strategy that addresses immediate objectives and long-term growth, ensuring minimal disruption to daily operations.

2. Deploying VAST DataBase as the Core Data Platform

- **Unified Data Layer:** Implement VAST DataBase to consolidate structured and unstructured data into a single, high-performance layer, eliminating the silos that hinder data access and insight.
- **Accelerated Query Performance:** Leverage VAST DataBase's architecture to expedite complex queries across petabyte-scale datasets, enhancing data agility for analytics, machine learning, and AI applications.
- **Centralized Management:** Provide applications with consistent, high-performance data access through a centrally managed interface, simplifying operations and boosting performance at scale.

3. Migrating Legacy Data to VAST Storage for Long-Term Storage

- **Cost-Effective Storage:** Migrate less frequently accessed data to VAST's S3-compatible object storage, optimizing storage costs while maintaining accessibility.
- **Seamless Integration:** Ensure VAST's S3 interfaces comply with AWS S3 API standards, facilitating straightforward migration and access for applications requiring legacy data without disrupting active data processes.

This structured approach ensures a seamless transition from Hadoop to the VAST Data Platform, empowering your organization to leverage a unified, high-performance data platform tailored for modern analytics and AI workloads.

Modernize for the AI-Era with VAST Today

Migrating beyond Hadoop is a strategic move for modern enterprises seeking to enhance performance, reduce costs, and prepare for next-generation data analytics and AI workloads.

- **Optimized Performance:** By using the VAST DataBase for structured, performance-sensitive data and VAST Storage for legacy or infrequently accessed data, organizations can streamline data access across file, object, and database storage layers, enhancing overall application efficiency, while positioning the organization to apply all of its data for analytics and AI.
- **Streamlined Operations and Real-Time Monitoring:** VAST's unified platform enables efficient data management, optimizing real-time analytics and capacity planning across both VAST DataBase and VAST Storage, simplifying operations and reducing administrative overhead. VAST's partner ecosystem provides ongoing support, troubleshooting, and optimization to ensure your infrastructure remains high-performing and future-ready.
- **Data as a Strategic Asset:** Centralizing structured data in VAST DataBase and VAST Storage for legacy storage eliminates data silos and streamlines data workflows, improving query performance and reducing storage and compute costs. The platform is engineered to support large-scale, data-driven applications, offering the agility needed for advanced analytics and AI/ML workloads in a single, unified data environment.
- **Proven Path to Transformation:** Organizations have successfully modernized from Hadoop to the VAST Data Platform, realizing immediate improvements in data access, simplicity, and cost-effectiveness. VAST's network of experienced partners and our advanced tooling ecosystem provides a reliable and efficient path to modernize data infrastructure, harnessing VAST DataBase for high-performance workloads and VAST Storage for comprehensive data management across the enterprise.

The VAST Data Platform provides a proven, cost-effective path to modernization with a future-ready architecture optimized for both structured and unstructured data. Learn more today at vastdata.com.



For more information on the VAST Data Platform and how it can help you modernize, simplify, and reduce costs for your modern data architecture, reach out to us at hello@vastdata.com.

©2025 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.