Solution Brief

What's New in VAST AI OS 5.5

Faster Retrieval. Smarter Analytics. Unified Pipeline Compute. All in One Release.

The next era of enterprise AI demands more than faster components. It demands a platform where data, vectors, events, and execution operate as one system without the fragmentation, data movement, and operational sprawl that has defined the AI infrastructure stack for the past decade. VAST AI OS 5.5 is generally available today, delivering major advances across the platform: hyperscale vector retrieval, native analytical execution, managed pipeline compute inside the cluster, and zero-copy data operations. Whether you are scaling RAG pipelines to trillion-vector workloads, accelerating enterprise data warehouse queries, or reducing the operational burden of separate Kubernetes infrastructure, 5.5 gives you the foundation to move faster with fewer systems to manage.

What's New: A Platform Built for Real-Time AI at Scale



Hyperscale Vector Index: Trillion-Vector Search Without the Overhead

Vector databases have long required a painful tradeoff: performance at modest scale, or operational complexity at enterprise scale. Memory-resident indexes work until they don't. Sharding buys headroom but introduces rebalancing overhead, governance gaps, and unpredictable latency as workloads grow. External vector stores add another system to integrate, secure, and maintain.

VAST 5.5 introduces the VAST Hyperscale Vector Index, a proprietary hierarchical clustering architecture embedded directly in the VAST DataBase. Vectors are organized into a multi-level hierarchy of distance-based clusters. At query time, search descends only into the most promising candidates and loading targeted subsets of quantized vectors into memory. Query latency plateaus rather than scaling with size, and the entire index lives on flash alongside structured and semi-structured data under a single governance model.

The result is vector search that is memory-bounded, highly parallel, and predictable at trillion-vector scale, with no external vector store, no sharding, and no permissions drift between source data and its vector representations. Use cases such as video analytics, semantic search, fraud detection, and large-scale RAG all benefit from retrieval that is faster and more contextually precise.



VAST Native Query Engine: Analytics Where the Data Lives

The conventional approach to enterprise analytics separates the execution engine from the storage platform. Data moves outward to be processed, introducing latency, pipeline complexity, and architectural sprawl. VAST is taking the opposite approach: bring the execution to the data.

The VAST Native Query Engine has powered vector retrieval inside the platform since its introduction. With 5.5, it expands significantly into broader analytical workloads. More than 50 new aggregation and statistical functions now execute natively inside VAST.

These include core operations such as SUM, MIN, MAX, and COUNT; statistical measures including variance, standard deviation, skewness, kurtosis, and covariance; regression and correlation functions; quantile and percentile estimation; and conditional aggregation via the FILTER clause. Boolean and bitwise aggregations are also supported.

Complex statistical analysis, machine learning feature preparation, and exploratory data analysis that previously required exporting data to external engines can now run directly on the platform. Analytical logic executes where the data lives, under the same governance model, without additional infrastructure or data movement. Now, with this native VAST Engine, you can also conduct hybrid vector + SQL filter analytics in the same query path.



Partitioning for Large Tables: Performance at Scale Without Operational Cost

Query performance on large tables has traditionally required a tradeoff: either manage complex sharding schemes and rebalancing operations, or accept slower query execution as data grows. VAST 5.5 eliminates that tradeoff.

New SQL-native partitioning delivers meaningful performance gains for large-scale analytical workloads through a declarative model that places no operational burden on the user. Simply declare high-level partitioning intent — by time range, category, or any logical dimension — and the VAST engine automatically applies partition pruning, join optimization, and efficient deletes at query time. There are no shards to manage, no data to rebalance, no node ownership to reason about.



VAST Native Compute: Pipeline Compute Inside the Cluster

Running data and AI pipelines has meant operating two environments in parallel: a data platform for storage and retrieval, and a separate Kubernetes environment for compute-intensive pipeline workloads. Managing that boundary adds cost, latency, and operational friction. Data must move between systems simply to keep pipelines running.

VAST 5.5 introduces VAST Native Compute, bringing container orchestration directly into the VAST AI Operating System for DataEngine serverless Python functions. Administrators can provision managed Kubernetes clusters on CNodes within the VAST cluster, allocating dedicated compute capacity for containerized AI applications, data processing workloads, event-driven services, and serverless functions, while supporting external AI pipelines for vectorization, chunking, and embedding generation. Provisioning, scaling, upgrades, and lifecycle operations are handled entirely by the platform.

Pipeline functions run inside the cluster, adjacent to the data they operate on, with direct access to VAST file, object, database, and event services. The result is a unified execution model where data and pipeline compute share the same operational foundation — with fewer moving parts, stronger governance, and lower latency than fragmented “data here, compute there” architectures.



Efficient Data Operations and Pipeline Automation

Two additional capabilities in 5.5 extend automation and eliminate unnecessary data movement across the platform.

File-Triggered Pipelines for NFS: Extend VAST DataEngine’s event-driven automation beyond S3 to native NFSv4 workflows. File creation, deletion, and metadata changes can now trigger pipelines directly — without protocol translation or object storage re-architecture. For organizations that rely on NFS for application data, media processing, scientific

workloads, or legacy file-based environments, this closes a meaningful gap in real-time automation and removes friction from modernization efforts.

Instant Data Cloning for Block: Convert block-level copy operations from data movement tasks into lightweight metadata operations. Rather than reading from a source range and rewriting to a target, the platform allows the target to reference the same underlying flash data. Large copy operations such as VM full clones, Storage vMotion, and SMB server-side copy complete almost instantly without consuming storage bandwidth or write buffer resources. Source and target remain logically independent, so subsequent writes or deletions do not affect the cloned data.

VAST 5.5: Performance That Changes the Conversation

Capability	Result	Compared To
Vector Retrieval Speed	10x faster	OSS vector (AISAQ Index, 1B vectors, 128-dim)
Cost per 1K Vector Searches	91% lower	Leading OSS vector database, 50B vectors
Selective Query Performance	>60% faster	Optimized Iceberg
Table Update Speed	200x faster	Optimized Iceberg
Query Performance (EDW)	20% faster	Iceberg-based tables, real-world manufacturing

Benchmark results represent VAST internal testing. Results may vary by workload and configuration.

Why Upgrade to VAST AI OS 5.5

Trillion-Vector Scale:
The Hyperscale Vector Index delivers fast, memory-bounded vector search at scales that break conventional vector databases, with no external store and no sharding.

Analytics Without Egress:
50+ new native aggregation and statistical functions bring complex analytical workloads inside the platform, eliminating the need to export data to external engines.

Faster Queries on Large Tables:
SQL-native partitioning accelerates scans, joins, and updates on large datasets with no operational overhead and no manual data management.

Unified Pipeline Compute:
VAST Native Kubernetes eliminates the separate Kubernetes cluster for DataEngine functions. Pipelines run inside the cluster, adjacent to the data.

NFS Pipeline Automation:
File-triggered pipelines extend real-time event-driven automation to NFSv4 workflows.

About VAST

VAST Data is the leading data platform software company bringing businesses into the AI era. By accelerating time-to-insight for workload-intensive applications, the VAST AI Operating System delivers scalable performance, simplifies data management, and enhances productivity. Launched in 2019, VAST has become the fastest-growing data infrastructure startup in history.

Ready to see what VAST AI OS can do for your environment?
Contact us at hello@vastdata.com or reach out to your VAST account team.

©2026 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.
6/26