Solution Brief

VAST Vector Search

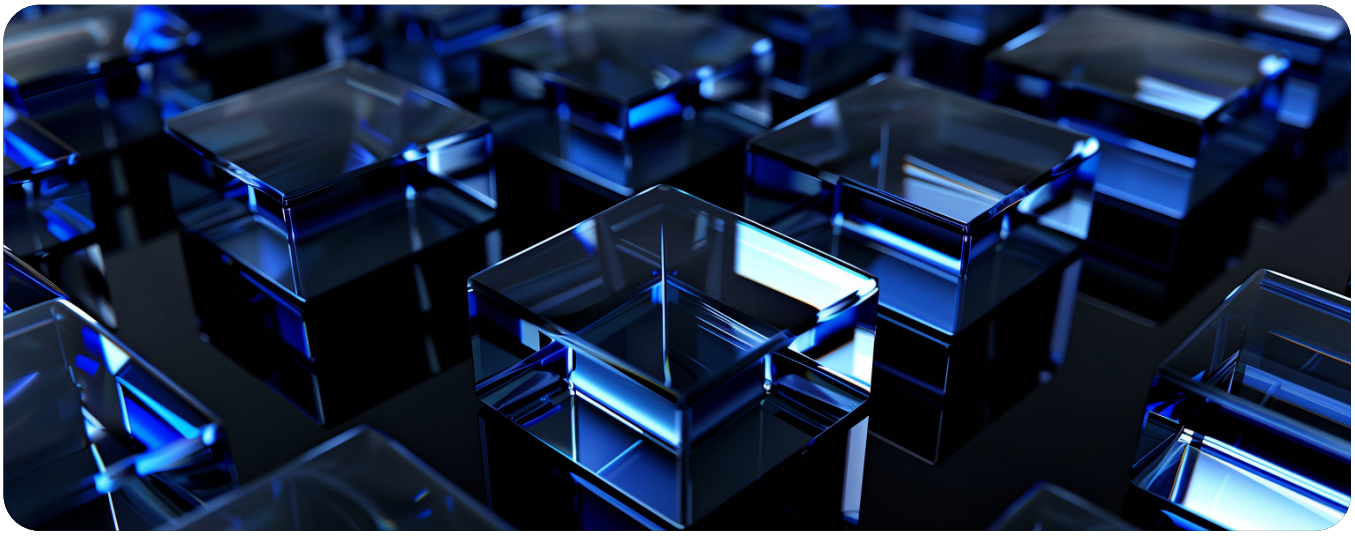
Real-Time AI Retrieval Without Limits

VAST Vector Search is the foundation for scalable, real-time context retrieval in AI systems—built into the VAST AI Operating System. Unlike legacy architectures that bolt on vector support as a silo, VAST natively integrates vector capabilities into its high-performance, unified data platform. With VAST, you can power intelligent agents, enhance recommendations, and scale retrieval-augmented generation (RAG) pipelines—without separate infrastructure, external indexes, or sharding overhead.

Accelerate AI Pipelines with Unified Vector Infrastructure

Native Vector Capabilities Without Silos or Sharding

VAST treats vectors as a first-class data type stored alongside structured data in the VAST DataBase. With no orchestration layers or GPU memory constraints, VAST enables hybrid queries that combine vector similarity with SQL-style filters—such as retrieving nearest neighbors where metadata matches specific criteria—all executed through a single query path. This dramatically simplifies AI infrastructure while unlocking new possibilities for intelligent agents.



Real-time Ingest, Indexing, and High-Performance Search

Every vector is indexed at ingest using flash-native zone maps and vector indexes. Compact columnar data chunks allow for early block pruning and fast, selective retrieval—even at trillion-vector scale. With compute nodes accessing storage directly via NVMe, VAST delivers sub-second query latency without memory-bound limits or costly caching.

Built for Scale, Security, and Intelligent Applications

Powered by the VAST DASE architecture, vector search scales linearly by adding compute—no rebalancing or re-indexing required. Governance is built-in, with row/column-level permissions, audit trails, and policy enforcement across all modalities. Whether for secure RAG, multimodal agents, or fraud detection, VAST delivers the governance, scale, and speed that real-time AI applications demand.

Simplified Operations with Memory-Free Architecture

Unlike traditional vector databases that rely on complex sharding, manual rebalancing, or memory-bound GPU indexes, VAST's DASE architecture eliminates these operational burdens. Its unified, memory-free design ensures consistent performance and linear scalability by simply adding compute nodes, dramatically simplifying management and freeing up valuable engineering resources.

Enable Advanced AI with Cross-Modal Context

VAST stores vector embeddings, structured metadata, and full unstructured content in a single place, enabling rich cross-modal retrieval without external indexes or complex ETL pipelines. This native integration provides AI agents with instant, full-context retrieval, allowing them to reason over diverse data types and power more sophisticated, accurate, and real-time AI applications.

Why VAST

Unified Data Management

Consolidate structured and unstructured data and all analytics workloads for seamless access and faster insights.

Scalability

Easily scale to exabyte-level datasets with high performance as data grows.

Unified Security & Governance

Manage all data, from object to vector, under a single, unified governance model.

AI-Ready Infrastructure

Unified, GPU-accelerated architecture delivers low-latency I/O and integrated vector search to power real-time AI workloads at scale.

Automated AI Pipelines

Trigger real-time vector search and retrieval with InsightEngine's event-driven workflows.

About VAST Data

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.

Looking to simplify your data center and unlock insights from all of your data?
Contact us at hello@vastdata.com

Trusted By Industry Leaders

PIXAR FOX CoreWeave NHL servicenow agoda

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

06/2025