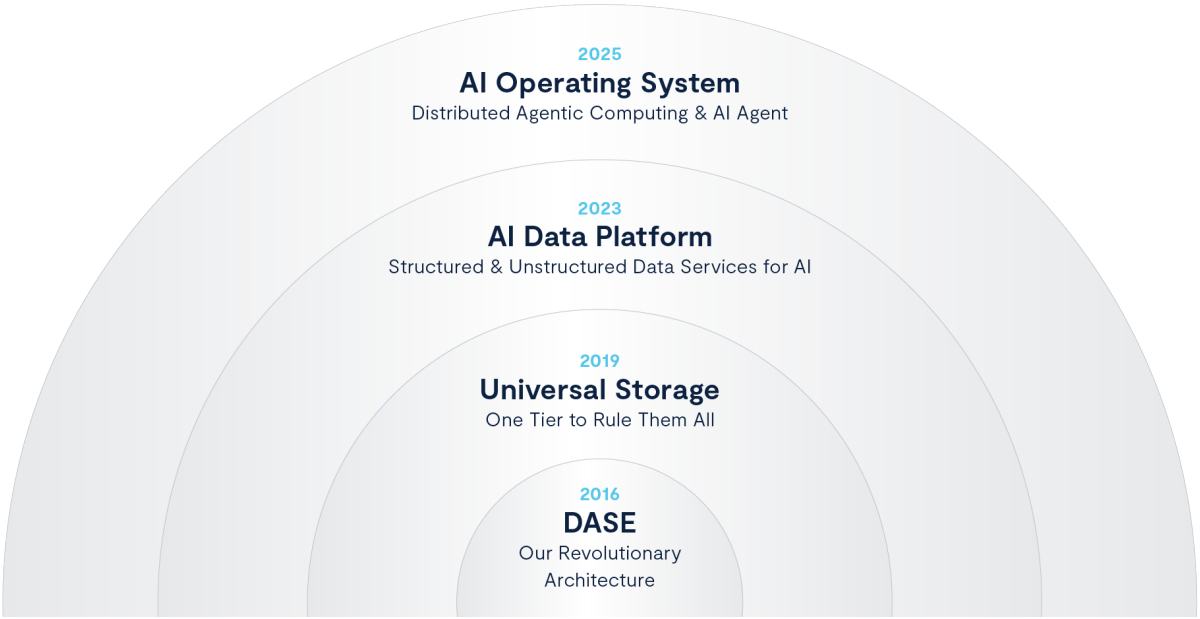
Solution Brief

Understanding the Software Stack of the VAST AI Operating System

A Newcomer's Guide

Welcome to VAST! If you're looking to understand our core product, the VAST AI Operating System, you're in the right place. This document will break down what it is, its key components, and how they work together to power the next generation of artificial intelligence applications.

The Evolution: From Revolutionary Architecture to an AI Operating System



VAST began its journey by revolutionizing data storage. As shown in the first image, our foundation was laid with **DASE (Disaggregated and Shared Everything architecture)** in 2016. This groundbreaking approach allowed us to build **Universal Storage** by 2019 – a single tier of storage designed to handle all data types with extreme performance and scalability.

Recognizing the growing needs of AI, we evolved our offering into an **AI Data Platform** by 2023, providing structured and unstructured data services specifically for AI workloads. Today, this has culminated in the **VAST AI Operating System**, launched in 2025. It's a comprehensive platform designed for **Distributed Agentic Computing & AI Agents**, aiming to provide a unified environment for data, an engine for insights, and a framework for intelligent automation.

What is the VAST AI Operating System?

At its core, the VAST AI Operating System is a software platform that simplifies and accelerates the entire AI pipeline, from raw data to intelligent action. It's designed to be the foundation upon which AI-driven applications and autonomous agents are built, deployed, and managed efficiently.

Think of it like the operating system on your computer, but built for the massive scale and unique demands of AI. It manages the underlying resources, provides essential services, and offers a consistent environment for AI applications to run.

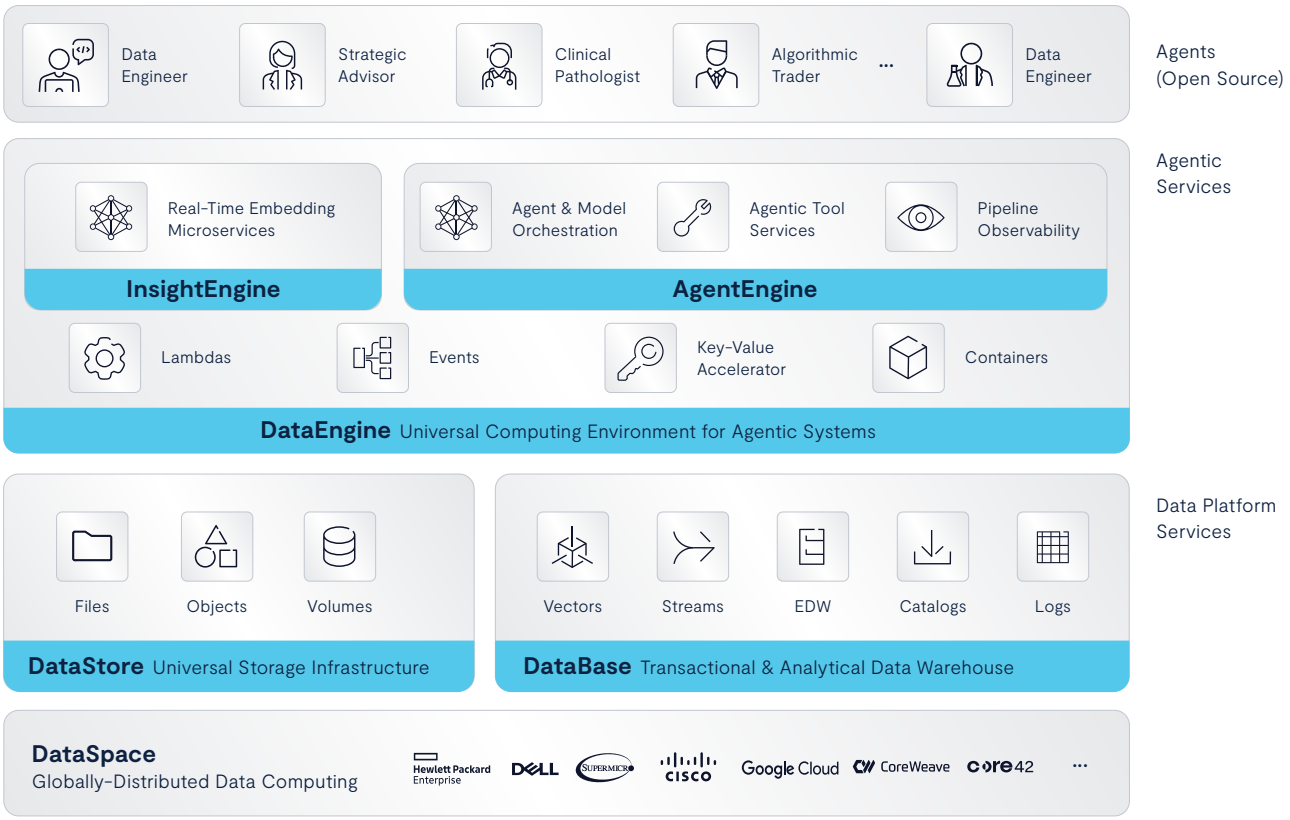
Deconstructing the VAST AI Operating System: Key Components and How They Interact

The Architecture: Disaggregated and Shared Everything (DASE)

At the heart of VAST's technological innovation is its Disaggregated and Shared Everything (DASE) architecture, first introduced in 2016. This architecture was a fundamental rethinking of how distributed systems should be built to overcome the limitations of traditional designs, particularly as the scale of data and computing continued to explode.

DASE is built on the principle of disaggregating (separating) compute logic from the physical storage media, while allowing all compute nodes to access all storage devices in parallel. Using high-performance NVMe-over-Fabrics, CPUs can access remote SSDs as if they were local. This "shared everything" approach, combined with novel transactional data structures, eliminates the need for data partitioning and complex inter-server coordination typical in "shared-nothing" architectures. The result is a highly scalable and resilient system that breaks traditional tradeoffs between performance and capacity. DASE enables terabyte-to-exabyte linear scalability, simplifies data management, and allows for unprecedented efficiency in utilizing all-flash storage. This architectural foundation was crucial for VAST's evolution from Universal Storage to a comprehensive AI Data Platform and ultimately, the VAST AI Operating System.

Let's look at the key layers and components of the VAST AI Operating System built upon this DASE foundation, as illustrated in the provided architecture diagrams:



1. The Foundation Data Platform Services: DataSpace, DataStore, and DataBase

VAST DataSpace

This is the globally-distributed data computing layer. It allows organizations to manage and access data that might be spread across different locations, whether on-premises, in the cloud, or at the edge. It provides a unified view of all your data.

VAST DataStore (Universal Storage Infrastructure)

Building on our Universal Storage roots, the DataStore is where your data lives. It's a high-performance, scalable, and resilient file and object storage system. It's designed to ingest, store, and provide rapid access to the massive datasets required for AI, including files, objects, and volumes.

VAST DataBase (Universal Database Infrastructure)

This isn't just a traditional database. It's a database built for AI, capable of handling diverse data types including structured data (like tables in an Enterprise Data Warehouse - EDW), contextual meta data about unstructured data (like text and images), vectors (critical for many AI algorithms like similarity search), streams (for real-time data processing), catalogs, and logs.

How they work together:

The DataSpace provides the overarching fabric. Within it, the DataStore efficiently holds all types of raw and processed data, while the DataBase organizes and makes this data queryable and usable for AI model training and inference.

2. The Engines: DataEngine and InsightEngine

VAST DataEngine (Universal Computing Environment for Agentic Systems)

This layer sits on top of your data and provides the computational power. It's a runtime and scheduler that can execute various workloads, including those based on Lambdas (serverless functions), events, key-value acceleration, and containers. It's the engine that processes data, runs transformations, and prepares data for AI models.

VAST InsightEngine

This component is focused on extracting intelligence from your data. It provides services like real-time embedding microservices (transforming data into a format AI can understand), agent and model orchestration, agentic tool services, and pipeline observability. This is where data begins to turn into actionable insights.

How they work together:

The DataEngine provides the raw power and execution environment. The InsightEngine leverages this power to run more specialized AI-related tasks, generate embeddings, and manage the complex workflows of AI models and agents. For example, the DataEngine might run a data cleaning pipeline, and then the InsightEngine would use that cleaned data to generate vector embeddings for a similarity search application.

3. The Intelligence Layer: VAST AgentEngine and AI Agents/Applications

VAST AgentEngine

This is a crucial component for managing and orchestrating AI agents. These agents can be thought of as autonomous programs that can perform tasks, make decisions, and interact with their environment. The AgentEngine provides the message bus and infrastructure for these agents to communicate and collaborate.

AI Agents & Applications

This is the top layer where the value of the AI Operating System is realized. These are the actual AI models and applications your organization builds or uses. Examples include:

- **AI Agents:** Data Engineers, Strategic Advisors, Clinical Pathologists, Algorithmic Traders, Biochemists (as depicted, these can be open source or custom-built).
- **Applications:** Video Editors, Quant Traders, Personal Assistants, and other data-driven applications.

How they work together:

The underlying layers (DataStore, DataBase, DataEngine, InsightEngine) provide the data, processing power, and foundational AI services. The AgentEngine then empowers developers to build and deploy sophisticated AI agents that leverage these capabilities. These agents can then be embedded into various applications or operate autonomously to achieve specific goals. For instance, an "Algorithmic Trader" agent would use market data from the DataBase, processed by the DataEngine, with insights generated by the InsightEngine, all orchestrated by the AgentEngine to make trading decisions.

4. Security & Observability:

Integrated throughout the entire stack, ensuring that data is secure, access is controlled, and the system's performance and behavior can be monitored and understood.

Putting It All Together: A Unified Platform for AI

The VAST AI Operating System is designed to break down silos that often exist between data storage, data processing, and AI model deployment.

- Data is ingested and stored efficiently in the **DataStore** and made AI-ready by the **DataBase**
- The **DataEngine** provides the power to process this data at scale
- The **InsightEngine** helps extract meaningful patterns and insights
- The **AgentEngine** allows for the creation and management of intelligent agents that can act on these insights
- All of this is accessible through the **DataSpace** and governed by robust **Security and Observability**

This integrated approach means:

Faster time to insights

Less friction in moving data and getting it ready for AI.

Simplified AI MLOps

A more streamlined process for developing, deploying, and managing AI models and agents.

Scalability

The architecture is built to handle the growing demands of AI workloads.

Flexibility

Support for various data types, AI frameworks, and deployment models.

Global Operations

Execute AI workloads seamlessly across geographies while adhering to data sovereignty requirements through a unified, policy-driven framework.

By providing a comprehensive and unified platform, the VAST AI Operating System empowers organizations to unlock the full potential of their data and accelerate their AI initiatives, moving from complex, fragmented toolchains to a cohesive environment for building and running intelligent systems.

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 Data Platform 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