VAST Data and the AI Operating System

Bringing Agentic AI to the Enterprise

Summary

There is a tension in modern business that is difficult to resolve. Business leaders, driven by the need to be faster, more responsive, and more efficient than their competitors, expect to realize the promise of AI immediately. Meanwhile, even as tech vendors and pundits tout the capabilities of generative and agentic AI, enterprise IT organizations are still in the throes of proofs of concept and pilots.

The challenge of activating AI is twofold. The first challenge is related to *data*. Building, deploying, and constantly tuning an AI pipeline is complex, resource-intensive, and time-consuming. The second challenge is related to *tooling*. While AI technology has quickly evolved from understanding large language models to agentic AI in just a couple of years, the toolchain that should remove the friction from operationalizing this technology has lagged.

This research brief will look at the AI activation challenges faced by enterprise IT and explore how companies like VAST Data act as the foundational operating system for the AI environment — the AI OS.

Agentic AI and the Self-Running Enterprise

Traditional or discriminative AI, which involves narrowly defined tasks such as quality control, visual recognition, or fraud detection, has proven its value in the enterprise for some time. Consulting giant McKinsey estimates that 78% of enterprise organizations have deployed AI to support at least one business function.

When Moor Insights & Strategy (MI&S) speaks with business leaders at large enterprise organizations, AI deployment to drive business outcomes is something far different than discriminative AI. Those leaders think of agentic AI, or AI that delivers business process automation through complex decision-making on queries made by understanding natural language. Autonomy, reasoning, learning and adaptability, and execution are all characteristics of agentic AI systems.

Planning a business trip is a good example scenario for the use of agentic AI. Through natural language, a business user could tell their AI assistant to book a trip to the

Boston area to visit a client. That assistant would break down the prompt into several subtasks (flight, rental car, hotel), along with context (client location) and learning (flight preferences, rental car types usually booked), then generate an itinerary for the user.

Draw a parallel between this basic example and a more complex business function, such as supply chain optimization. It is easy to see how a manufacturer can increase resilience and significantly reduce costs using agentic AI principles. This parallel can be drawn to virtually every business function, resulting in unprecedented operational and cost efficiencies. This is not hyperbolic language, nor does it require an exhaustive ROI or TCO study to prove.

If one is curious about the impact — and potential complexity — of agentic AI, consider the following: NVIDIA CEO Jensen Huang has publicly opined that his company could one day comprise 50,000 employees and deploy upward of 100 million AI agents (2,000 agents for every employee). NVIDIA is an interesting example. While some may look at the company as an outlier, it is, in many ways, just another enterprise that designs and manufactures products and services.

Agentic AI has a complementary relationship with generative AI. Whereas agentic AI focuses on autonomous execution, generative AI focuses on creating content that aids workflows. For instance, autonomous AI may require a generative AI function to create an email reserving a dinner reservation in support of that business user's trip. This relationship is important to understand, as many view agentic AI as an evolutionary step forward from generative AI.

It is easy to discuss agentic AI in the abstract and focus on its value. However, achieving a fully AI-optimized enterprise has been a hugely complex endeavor. As such, MI&S has seen enterprise organizations' activation efforts struggle and even stall.

Agentic AI and the Capability Maturity Model

Many organizations have developed AI capability maturity models (CMMs) based on the Carnegie Mellon CMM framework. A model from Deloitte captures the essence of AI transformation:

- **Foundational (Stage I):** Enterprise IT possesses a basic understanding of AI and has engaged in exploratory projects.
- **Skilled and Structured (Stage II):** Enterprise IT has developed some AI skills and has defined structured initiatives.

- **Integrated and Aligned (Stage III):** AI initiatives align with business objectives and are integrating with business functions.
- **Strategic and Transformational (Stage IV):** AI is core to business strategy.

While some IT organizations perceive themselves as further along this continuum, MI&S finds that many enterprises are still in the second stage of maturity. This is largely due to what we highlighted in the opening of this research brief — challenges in data management, data skills, and access to the tools that can deliver AI transformation.

The Agentic AI Toolchain Comes at a Cost

Activating agentic AI requires an organization to assemble an operating environment comprising tools from multiple vendors that span infrastructure, data management, security and governance, orchestration, observability, and optimization. In practice, numerous tools from several vendors connect an underlying data platform with agentic services, delivering synthesized data to potentially millions of services supporting a large enterprise.

For many enterprises early in their agentic AI journey, the result is an environment that has not yet reached its potential in terms of operational efficiency and cost savings. A data environment that can't adequately feed GPU clusters, on the one hand, and can't enable agents to drive true and accurate automation in the enterprise, on the other. In fact, the opposite effect is achieved in some cases, where the fully burdened cost of early agentic AI deployments pushes ROI targets further to the right as project budgets increasingly eat into IT costs.

Because of this, MI&S sees the ideal agentic AI environment as one that begins with what is essentially an operating system. Such an AI operating environment abstracts underlying infrastructure — storage, powerful GPUs, and accelerators — making it easily accessible to the data and tools that enable agentic AI. Further, MI&S views VAST Data and its AI Operating System as uniquely positioned to deliver the agentic AI operating environment.

VAST Data — the AI OS

The AI OS is one in which hardware and software resources are managed for optimal performance, cost, security, and governance. When we think of this from a traditional server perspective, we think of Microsoft Windows Server or, perhaps more aptly, the more open Linux operating system.

FIGURE 1: THE VAST AI OPERATING SYSTEM

The VAST AI OS architectural block diagram
Source: VAST Data

Below is an overview of the VAST architecture and its role in the AI Operating System:

DASE Architecture

Because of the data requirements associated with AI, scalability is critical. More precisely, independent scalability, where compute and storage resources are untethered from one another, is the basis for enabling what amounts to unlimited growth of globally accessible data across the enterprise.

The foundation of the VAST AI OS lies in its DASE (Disaggregated, Shared Everything) architecture. DASE architecture is specifically designed to enable GPU-equipped systems to be fed with extremely large datasets, scaling seamlessly to meet the demands of modern AI workloads. With exabyte-level storage capacity and linear scale to feed GPUs, DASE supports the most extreme use cases.

Part of what makes DASE so scalable is the use of VAST DataSpace, a global data management and access platform that provides a single, unified namespace that spans edge, core, and cloud environments. Because it decentralizes lock management to the element level (file, object, table), DataSpace resolves the storage system tension that

organizations must contend with related to absolute best read performance and absolute best write consistency.

Because VAST re-architected storage around embarrassingly parallel principles, its performance scales linearly as capacity grows. Many other storage environments will experience latency and performance challenges as (storage) nodes are added. However, VAST's embrace of independent task execution and stateless data services on disaggregated resources — using the VAST DataSpace — enables the highest levels of throughput as storage environments scale to exabyte capacity.

VAST DataStore

Sitting atop DASE is the VAST DataStore, the foundational storage layer for the AI OS — or, as one would think in OS terms, the AI file system. DataStore unifies all enterprise data — structured, unstructured, and streaming — under a global namespace and is accessible over any protocol.

DataStore builds on the principle of no compromise in how it uses flash storage to optimize data access. Metadata such as file lookups and directory traversals is stored in low-latency flash storage, whereas bulk data is stored in more affordable flash. This is all fully accessible over the storage protocols utilized in the enterprise — NFS, SMB, S3, and NVMe/TCP

While performance is a key focus of VAST's story and relevance in AI, DataStore should also be recognized for contributing to significant TCO savings. For example, the use of low-latency flash to store metadata (what makes retrieval fast) and more affordable flash for bulk data will drive significant capital expenses. The centralization of storage eliminates complexity, thereby reducing the resources (human and equipment) associated with maintenance and operational costs.

VAST DataBase

If DataStore is the AI OS file system, VAST DataBase is the OS index, the data structure that organizes data and makes it easily retrievable. Just as DataStore unifies data element types, DataBase stands as a single repository that brings together data warehousing (OLAP), transaction processing (OLTP), and data lakes (real-time analytics).

DataBase is optimized for AI and deep learning workloads, allowing applications to query, refine, and analyze data directly where it is stored without the need for external data lakes or third-party platforms.

In plain terms, DataBase is a unified, multiprotocol data platform for modern analytics, AI, and enterprise workloads. Breaking well-established datacenter norms, DataBase supports deep learning and real-time analytics while driving down costs by potentially providing an alternative to the dozens of database distributions that exist across the enterprise.

One powerful capability of the VAST DataBase is native support for vectors, which are mathematical representations that allow algorithms to understand and process complex information. This is particularly crucial for unstructured data, such as text, images, or audio, as vectors bring context to this data by capturing its semantic meaning and relationships. This contextual understanding is important because it enables AI agents to reason, act, and learn more effectively in real time. By storing vector embeddings directly alongside the data, systems can support AI applications at a massive scale, allowing them to quickly find relevant information, identify patterns, and make more accurate predictions from vast amounts of unstructured content.

Further, VAST's recent open-sourcing of its key distributed cache, called VUA (VAST Undivided Attention), extends cache across an entire GPU cluster, making data retrieval far faster and cache more efficient. Another architectural benefit to VUA is its tiered memory hierarchy, which extends cache from GPU and CPU memory to a shared pool of NVMe flash, delivering fast context storage for AI models.

VAST DataEngine

As we move up the AI OS stack, the VAST DataEngine plays the role of scheduler and messaging bus. As VAST positions it, DataEngine is the universal computing engine for agentic systems. In layperson's terms, the automation and intelligence layer makes operationalizing AI far more effective through orchestration.

When considering data movement and sharing between systems, one thinks of the time consuming (and not always accurate) extract, transform, and load (ETL) process. However, in data pipelines that feed AI and analytics, tools require access to fresh data immediately. The VAST DataEngine enables real-time, event-driven data processing by ingesting, transforming, enriching, and presenting data immediately.

This real-time capability is the result of DataEngine's Event Broker, a Kafka-compatible message bus that allows native event streaming directly into the DataBase. This eliminates the need for an external Kafka cluster, allowing for real-time access to streaming data alongside all historical data to support agentic AI or analytics.

Real-time streaming is critical to VAST's **InsightEngine** and the new VAST **AgentEngine**. InsightEngine leverages DataEngine's events, triggers, and lambda functions to transform raw data into AI-ready insights through real-time RAG and AI searches.

VAST positions AgentEngine as its specialized agent runtime environment designed to build, deploy, manage, and optimize AI agents through observability and orchestration tools. Included in this runtime environment is a developer studio that enables organizations to build, connect, and govern AI agents, giving them access to data pipelines and related services via an MCP toolkit. AgentEngine leverages DataEngine's real-time event streaming to deliver richer observability and more effective orchestration by placing agents in the closest proximity to the data they rely on.

Recalling Jensen Huang's vision of someday having 100 million AI agents supporting NVIDIA's business operations, one can envision the challenges of observing and managing the data pipeline and agent (and agent-to-agent) environment. The aim of AgentEngine is to remove this complexity through automation and easy-to-use tools.

FIGURE 2: THE VAST AGENTENGINE TOOLBOX

The AgentEngine Toolbox provides the tools that are vital for agent operations
Source: VAST Data

While MI&S has not been able to evaluate AgentEngine firsthand, the demonstration provided by VAST shows a tool that is both rich in capability and simple in usability — a difficult feat to achieve in emerging technologies. Features such as distributed tracing, where data and requests between agents are visualized onscreen and used to deliver debugging and root cause analysis, should resonate with IT professionals tasked with managing infrastructure and environments.

AGENTS ARE YOUR APPLICATIONS

As demonstrated above, VAST has truly built what can be described as the operating system for the AI era, the value of which can be immeasurable to an enterprise. In addition to the performance aspect of VAST environments, there are the less obvious but perhaps more important IT factors tied to manageability, simplicity, and time-to-value that come from working with a single source for your enterprise AI operations.

Consider the alternative where a dozen different tools with various functions must be stitched together and continuously tuned and updated to deliver a similar function. The resources required to support such an environment would be prohibitive, and the staffing requirements would be impossible to resource.

VAST is delivering on the vision of deploying an infrastructure environment to the enterprise that simply requires IT organizations to deploy their agents (or VAST's) as applications. It is the only such real-world environment MI&S has seen.

CALL TO ACTION

While the agentic AI hype cycle continues toward its apex, enterprise organizations are struggling with how to operationalize AI to its fullest extent. Part of this is due to the velocity of innovation. In just a few short years, we've seen the discussion shift from GenAI-powered chatbots to autonomous business driven by agentic AI.

The challenges around achieving the autonomous enterprise are rooted in data and tooling. That is, enterprises need a universal data platform capable of making all enterprise data AI-ready in real time and the tooling to connect data with infrastructure and deliver it to agents for activation. These challenges are reinforced in the vast majority of engagements MI&S has with enterprise organizations.

While organizations can manually craft an AI operating environment, the ongoing effort and costs of deploying, integrating, optimizing, and continually tuning such an

environment drive down the solution's value. Such tedious work, which is complex and prone to errors, is likely to result in an AI environment that never reaches its full potential.

VAST and its AI Operating System simplify the process of creating the optimal AI environment. With DASE as its design principle for the new AgentEngine to manage the enterprise agent landscape, VAST has delivered a truly unique agentic AI and data platform. Because of this, MI&S believes organizations beginning their agentic AI journey should strongly consider the VAST Data AI Operating System.

For more information, visit: https://www.vastdata.com/platform/ai-os.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR
Matt Kimball, Vice President and Principal Analyst, Datacenter Compute and Storage

PUBLISHER
Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES
Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS
Accredited press and analysts can cite this paper, but it must be cited in context, displaying the author's name, title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission from Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES
VAST Data commissioned this paper. Moor Insights & Strategy provides research, analysis, advice, and consulting to many of the high-tech companies mentioned in this paper. No employees at the firm hold equity positions with any of the companies cited in this document.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators, not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of this document's publication date. Please remember that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.