HyperFRAME RESEARCH

Sponsored by: VAST

RESEARCH BRIEF

Confidential AI: Bringing Frontier Models to Regulated Enterprise Data

How VAST Data enables consistent security, policy, and audit across data at rest, in transit, and in use



SEPTEMBER 2026

Author:
Don Gentile
Analyst-in-Residence, Data Platforms and Resiliency

Executive Summary

Some of the most valuable enterprise data remains locked in organizationally controlled systems located on-premises, in colocation facilities, and sovereign environments where regulation, residency requirements, latency, or data-movement costs limit how information can be used. VAST estimates that roughly half of the world's data resides outside the environments where many frontier models run. That creates a constraint for enterprise AI: capable models may exist, but they cannot always reach the information that potentially holds the greatest business value. HyperFRAME Research Lens: State of the Enterprise AI Stack (1H 2026) shows that only 14% of enterprises describe their core data architecture as fully modernized for AI workloads.

Bringing models to regulated data creates a two-sided trust requirement. Enterprises need to preserve custody of sensitive information through processing. Model providers need to protect proprietary model IP when execution occurs on infrastructure they do not operate. Both requirements converge at the runtime.

Against this backdrop, VAST is positioning its AI Operating System to manage models alongside enterprise data. The company plans to extend its platform with a protected execution path using confidential virtual machines, hardware Trusted Execution Environments, GPU confidential computing, attestation, and independent key custody. In the DataEnclave approach, enterprises retain the keys to their data, including privately tuned models, while model providers retain the keys to their models. The architecture is designed so neither party needs to surrender control of the asset it protects.

The result is a larger usable AI surface. In our view, model placement, data access, security policy, and infrastructure choice are interconnected parts of the same decision. VAST's approach positions the AI Operating System as the layer that governs where protected models meet sensitive enterprise information.



HyperFRAME Research. All rights reserved. 2026



1. Enterprise AI Moves Toward the Data

Data location determines where enterprise AI can run. Some of the most valuable enterprise data remains in systems of record because regulations, residency requirements, third-party processing restrictions, latency, and data-movement costs limit where that data can go. Clinical records, transaction histories, design files, and production telemetry often cannot be centralized around a model endpoint. In these environments, enterprises may need to run the model closer to the data. A key challenge is that an organization may have access to capable models yet still be unable to use them against high-value information.

For example, a visual inspection model in a manufacturing plant processes image data and telemetry that can expose proprietary production steps. The decision may also need to occur too quickly for a round trip to a distant region. In addition, that model may come from a specialist provider whose weights represent valuable intellectual property. The enterprise protects its data while the model provider protects its model.

HyperFRAME Research sees the same dynamics in other verticals. For example, healthcare systems maintain imaging studies and clinical records under controls required by HIPAA in the USA. Banks retain transaction and counterparty data within environments subject to regulatory examination, and government agencies keep classified data inside sovereign or isolated environments. In each case, data location limits model use. The enterprise cannot simply move sensitive information to wherever the preferred model runs. Expanding AI access to that data requires execution closer to the data while preserving controls over both enterprise information and model IP.

2. The Trust Condition Created by Distributed Execution

In conventional AI execution, a model processes data after it has been decrypted. During execution, information is exposed in host and accelerator memory, beyond the protection provided by encryption at rest or in transit. The host operating system, hypervisor, and administrators with sufficient privileges may all be able to access it. Distributed AI expands that trust boundary when a model runs in a manufacturing plant, colocation facility, sovereign cloud, or service provider infrastructure controlled by another party.

Enterprises therefore need assurance that regulated information remains protected through processing and that they can show who accessed it, under which policy, where the workload ran, and which model performed the work. HyperFRAME Research Lens (1H 2026) found that 70% of enterprises rate compliance, security, and data sovereignty as very important drivers of data architecture modernization. Model providers require equivalent assurance for proprietary weights. Running a model on infrastructure the developer does not control can expose those weights to privileged administrators, host software, and other tenants. This risk has helped keep the strongest proprietary models inside vendor-operated endpoints and away from regulated enterprise data.

The result is a gap between model capability and data access. Enterprises restrict access to sensitive information, while model providers restrict where their weights can execute. One approach is to use a runtime that both parties can verify without requiring either to trust the other or the underlying host. For organizations that already have the data platform, accelerators, and production AI infrastructure, the remaining constraint is trusted execution against data they already control.

HyperFRAME Research. All rights reserved. 2026

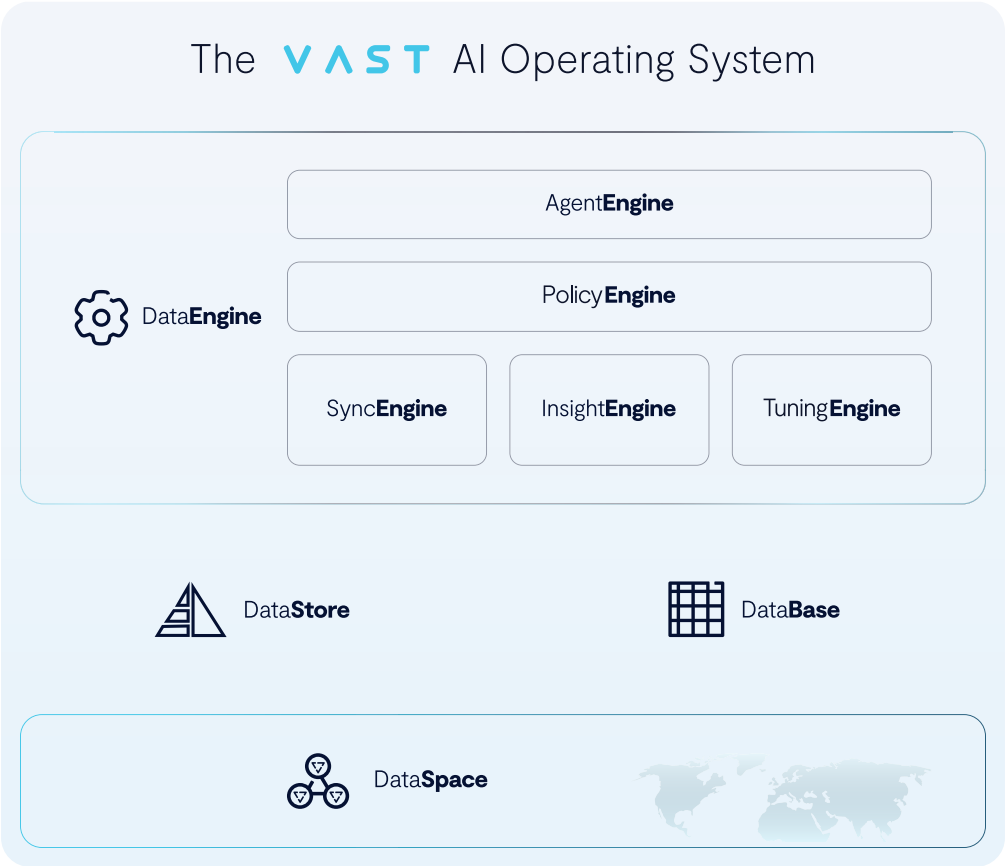


3. The AI Operating System as the Layer That Manages Models

VAST Data provides enterprise data infrastructure for AI and large-scale data environments, with deployments spanning data storage, processing, governance, and security. VAST positions its AI Operating System above the hardware layer and beneath models and applications. The company is now extending that role to model management, treating models as logical resources that the operating system can place, govern, and control alongside enterprise data. In that expanded role, the VAST platform aims to determine which model can run against which information, under which policy, and in which location. Provider terms, enterprise policy, and the trust relationship between the parties can all limit those choices. Confidential execution expands the set of workloads that can satisfy both sides.

The AI Operating System combines persistent storage with data processing and automation. The platform supports block, file, object, tabular, streaming, and vector data. DataEngine, InsightEngine, AgentEngine, and related services act on that information as it moves through the platform. DataSpace extends access across distributed locations, allowing governed enterprise information to remain available as workloads move between sites. VAST intends model management to extend those capabilities by governing where models execute and what information they can access.

Figure 1: The VAST AI Operating System Architecture



HyperFRAME Research. All rights reserved. 2026



VAST already applies a common security architecture to data at rest and in transit. The software runs in multi-tenant service provider deployments as well as isolated and air-gapped facilities. Encryption at rest uses a validated cryptographic library. Encryption in transit protects internal cluster traffic and client protocols. VAST's existing tenant-level key management supports external providers through KMIP, including Thales CipherTrust, HashiCorp Vault, Fortanix DSM, and Fornetix VaultCore, alongside enterprise identity and platform audit controls.

Built on the company's Disaggregated Shared Everything (DASE) architecture, the VAST stack has expanded in stages, with each capability broadening where it can participate. HyperFRAME Research has observed the company establish a presence with model providers and neoclouds, expand into hyperscalers, and more recently gain traction with enterprise customers. DataStore provides persistent access to enterprise information. DataBase adds structure and governance. DataEngine brings computation closer to the data. Model management extends the same control plane to AI models.

4. Model Management Becomes a Placement Problem

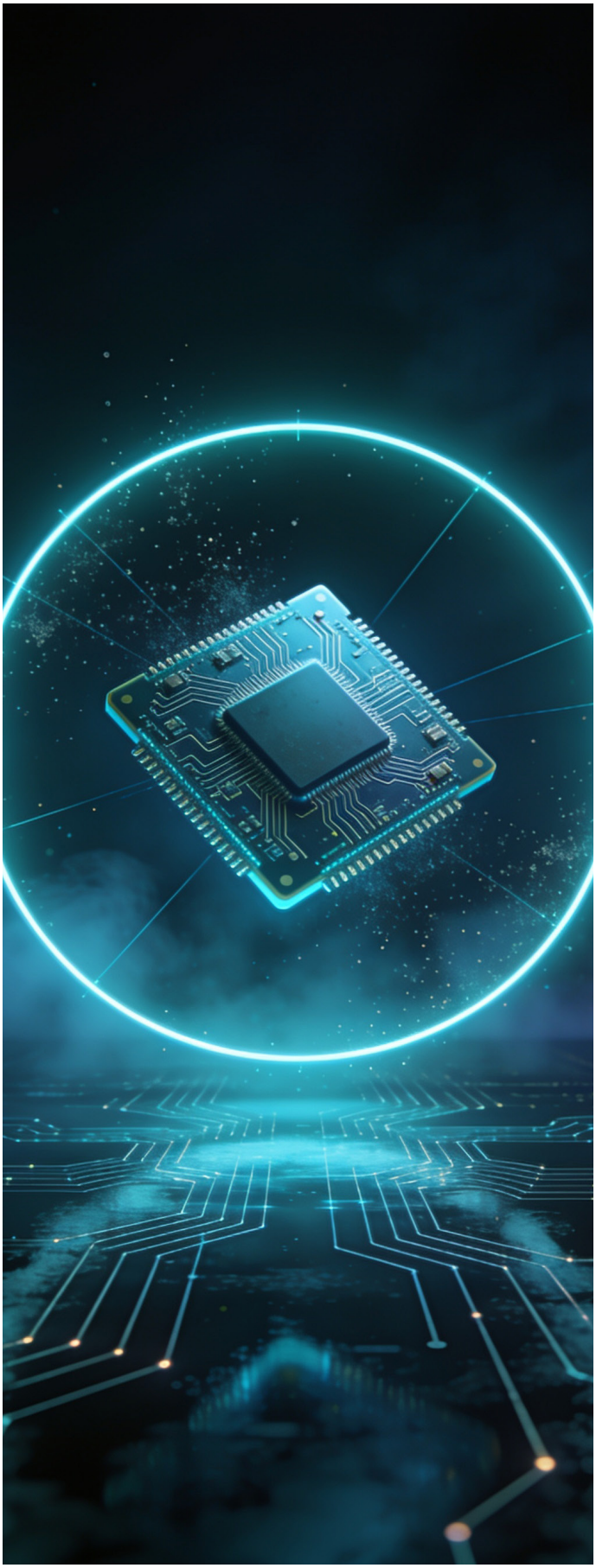
Once models move to regulated data, the enterprise must decide which model can run, where it can run, what information it can access, and under which policy. Those decisions vary by workload. Capability, cost, latency, security requirements, data sensitivity, and execution location can all affect model selection. Fine-tuned and enterprise-owned models add another requirement because their weights become governed intellectual property.

VAST's introduction of DataEnclave seeks to address protected model placement. At launch, the company is partnering with NVIDIA, Cohere, TwelveLabs, Factory, CrowdStrike, Deepgram, and Fundamental to deliver a curated list of content-specific models into customer data centers, cloud deployments, and edge sites. The immediate capability extends selected partner models into locations they may not have previously reached while protecting the provider's weights.

VAST intends to expand model management beyond placement. Its longer-term scope includes model choice and routing across models with different capabilities, costs, and deployment requirements. Those functions are not part of the initial launch. The September 2026 announcement establishes the architectural foundation VAST intends to use for broader policy-based model selection.

Enterprise-owned models expand the management challenge. Fine-tuning produces model artifacts that encode proprietary knowledge and business context. Agents may also accumulate specialized behavior and experience over time. As the number of models grows, enterprises will need to track versions, provenance, training and tuning inputs, access rights, and approved execution locations. Those assets require protection for many of the same reasons as proprietary frontier models. Model placement therefore becomes part of enterprise security policy. An operating system can select and route models only if it has a trusted location to run them when sensitive or regulated information is involved.

HyperFRAME Research. All rights reserved. 2026



5. Extending Protection Into Data in Use

VAST DataEnclave aims to extend the AI Operating System's security controls into active processing. The capability builds on DataEngine and sits within the pipeline and automation layer, where VAST already executes services against enterprise information.

DataEnclave uses Kata Containers and Confidential Containers (CoCo) to place workloads inside confidential virtual machines protected by a hardware Trusted Execution Environment (TEE). Guest memory remains encrypted and isolated from the host operating system, hypervisor, and administrators with privileged access. Processor-level confidential computing from Intel and AMD establishes the host-side isolation boundary. Protection continues into NVIDIA GPUs operating in confidential compute mode. Accelerator memory remains isolated, while protected communication extends across the GPU interconnect. VAST plans to supply the secure runtime as part of each DataEnclave deployment. The NVLink domain defines the enclave boundary. A single GPU is the minimum configuration, with support extending through four- and eight-GPU systems to rack-scale NVL72. Workloads requiring accelerators beyond a single NVLink domain currently fall outside that limit.

Confidential computing can affect performance, with the impact varying by workload and model. VAST expects to characterize those effects further before general availability. Some workloads will not require protected processing, and model providers may set different latency and throughput requirements than enterprise customers.

Attestation verifies the hardware, guest software, and workload configuration before regulated data or proprietary model IP is released into the runtime, while keys are released only when those checks satisfy policy. If verification fails or the attestation service is unavailable, the keys remain withheld. VAST plans to provide an attestation service with DataEnclave, while customers and model providers can operate their own. This gives each party the option to verify the execution boundary independently before granting access to protected assets.

HyperFRAME Research. All rights reserved. 2026

Independent trust domains

In the VAST approach, each party retains control of the keys protecting its own assets. Enterprises keep the keys to regulated data, while model providers keep the keys to proprietary weights. VAST separates key custody from execution so that no single party controls both the protected workload and the credentials needed to unlock it. That emphasis aligns with HyperFRAME Research Lens (1H 2026) findings that 63% of enterprises prioritize access controls and encryption in their AI governance and security practices.

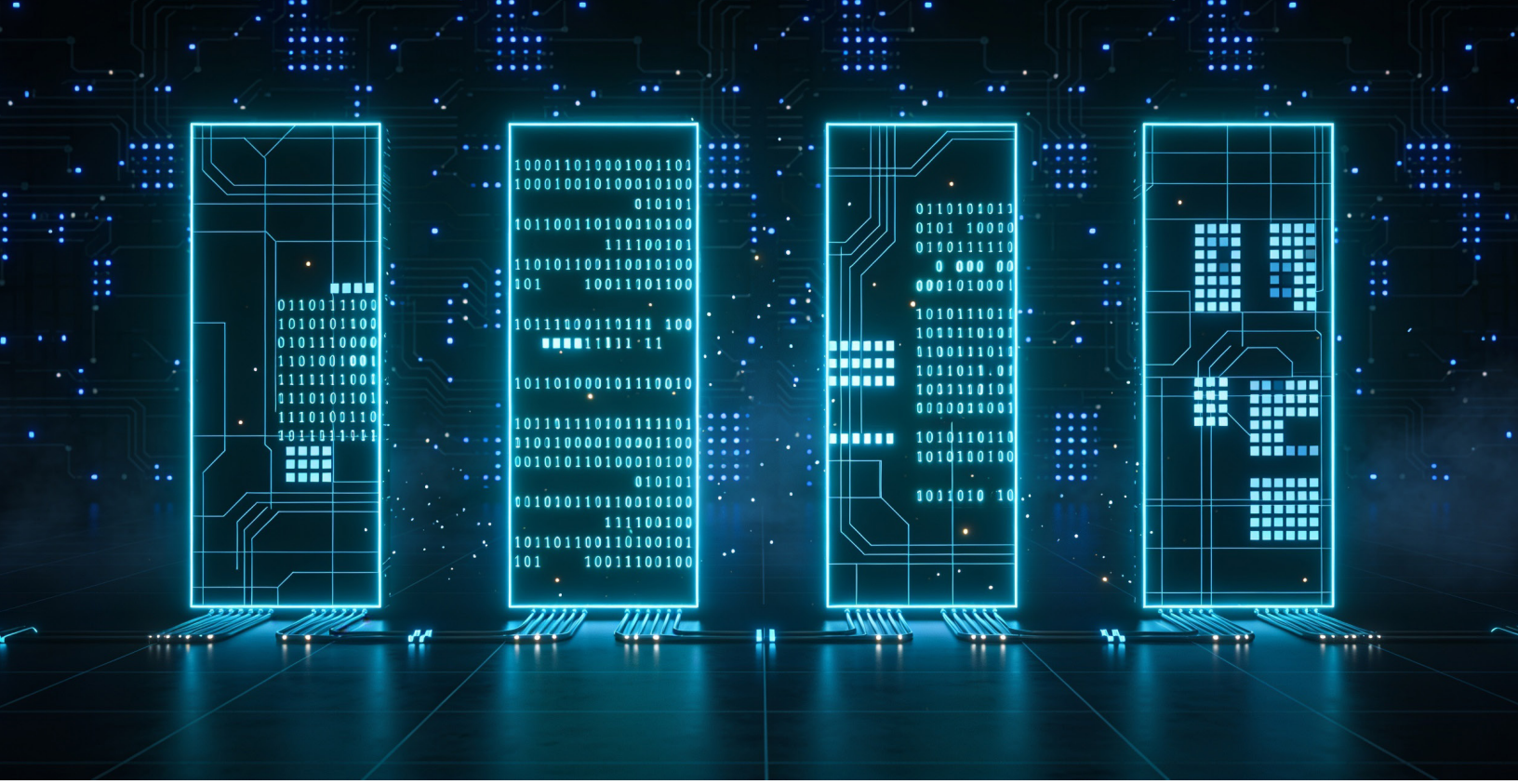
A confidential run depends on two independent approvals: the enterprise releases access to its data under its own policy; the model provider does the same for its model. Both sets of keys are released only into an attested environment that satisfies the requirements of each party.

The other participants contribute infrastructure without assuming ownership of those assets. Neocloud providers supply GPU capacity and hosting. OEM partners including Cisco and Supermicro can package deployment configurations, from four- to eight-GPU enterprise systems to rack-scale platforms for frontier model providers. Delivery models will vary by workload, infrastructure provider, and deployment site. Each participant controls a different part of the deployment, while trust is established through attestation and independent key custody instead of operator assurance.

Trusted execution extends governance into data in use, but the evidence may span multiple components. Attestation services, key managers, and execution platforms can come from different providers, each producing part of the record needed to verify what ran, under which policy, and whether access was authorized. VAST plans to record enclave activity in a tamper-evident audit trail stored in the VAST DataBase, where it can be queried alongside the information processed by the workload. Enterprises may still need to correlate those records with external attestation results and key-release events to reconstruct the full chain of trust.



HyperFRAME Research. All rights reserved. 2026



6. An Enterprise Evaluation Framework

Enterprises should evaluate confidential AI execution from initial deployment through day-two operation.

Trust model and key custody. Define the protected boundary across the confidential virtual machine, accelerators, and interconnect. Confirm that the enterprise retains keys for its data and the model provider retains keys for its model, each under its own release policy. Verify who operates attestation, what evidence it produces, and what happens when verification fails, or the service is unavailable.

Deployment model. Choose where the workload will run and understand the dependencies each option creates. On-prem deployment leaves facilities, power, cooling, and accelerated infrastructure with the enterprise. Neocloud deployment adds an infrastructure operator to the trust model. OEM-integrated systems provide validated hardware configurations with less deployment flexibility.

Data path and governance continuity. Trace how information enters and leaves the enclave, where decryption occurs, and which identity the workload presents. Confirm that existing access policy governs what the workload can read and that execution records can be correlated with the underlying data-access trail. Firmware, driver, or container changes alter attested measurements, so re-attestation and approval of new measurements become ongoing operating requirements. External tool calls and writes also cross the enclave boundary and need equivalent governance.

Workload fit and performance. Use confidential execution where the sensitivity of the data or model warrants it. Establish acceptable latency and throughput, then benchmark the actual workload and data path. Confirm that required accelerator capacity fits within a single NVLink domain.

Platform choice and maturity. Separate current capability from roadmap. Confirm supported model providers, validated hardware configurations, and available deployment options at general availability. Assess whether the architecture preserves model and infrastructure choice as requirements change.

Responsibility and operating model. Assign ownership for attestation, keys, policy, audit, incident response, upgrades, and maintenance. Define responsibilities across the enterprise, model provider, infrastructure operator, platform provider, OEM, and systems integrator. Validate the complete architecture at the intended deployment site and establish who responds when a control fails.

HyperFRAME Research. All rights reserved. 2026

Looking Ahead

DataEnclave is not yet generally available, but the architectural implication is already clear. Protected execution can open regulated enterprise data to models that previously could not reach it, while giving model providers a way to deploy proprietary IP into customer-controlled environments.

For enterprises, that expands the usable AI surface. Healthcare providers can bring stronger models to clinical data without moving records outside governed boundaries. Financial institutions can apply advanced models to transaction and risk data under existing custody requirements. Manufacturers can run proprietary models against production data at the plant. Government agencies can extend model choice into sovereign and isolated deployments.

The same architecture can also protect enterprise-owned models. As fine-tuned models and agents encode more proprietary knowledge, organizations will need to govern where those assets run and what information they can access. VAST-powered neocloud environments can give enterprises additional deployment options for privately tuned models and regulated data, extending model placement beyond infrastructure they operate directly. Model placement, data access, and security policy begin to converge as one infrastructure decision.
That requirement is already moving into formal governance: HyperFRAME Research Lens (1H 2026) found that 91% of enterprises have established or plan to establish formal AI oversight.

VAST also gains a broader role in the AI ecosystem. Its presence with model providers, neoclouds, hyperscalers, OEMs, and enterprises gives both model builders and customers more locations in which to place protected models and sensitive data according to workload requirements. If DataEnclave performs as intended, VAST can extend its platform from managing enterprise data and computation to governing where proprietary models meet sensitive information.



HyperFRAME Research. All rights reserved. 2026

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers in-depth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research

Email Address:

steven.dickens@hyperframeresearch.com

Telephone Number:

+1 845 505 1678

X: @StevenDickens3

LinkedIn: Steven Dickens

BlueSky: Steven Dickens

CONTRIBUTORS

Don Gentile
Analyst-in-Residence,
Data Platforms and Resiliency

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations.

LICENSING

This document, including any supporting materials, is owned by HyperFRAME Research. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of HyperFRAME Research.

DISCLOSURES

HyperFRAME Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.


HyperFRAME
RESEARCH