FAQs

VAST DataEnclave



Product Basics

What is VAST DataEnclave?

VAST DataEnclave is the confidential AI capability of VAST DataEngine, combining secure container execution, cryptographic attestation, and customer-controlled key management to protect sensitive data and proprietary models while AI is actively processing them.

How does VAST DataEnclave protect data while AI is running?

VAST DataEnclave runs AI workloads inside hardware-isolated confidential environments that prevent the underlying infrastructure and its administrators from inspecting protected CPU and GPU memory or the communication between them. Cryptographic attestation verifies the environment matches an approved baseline configuration before the keys or secrets required to access sensitive data and models are released.

What is confidential AI?

Confidential AI applies confidential computing to AI workloads to protect sensitive data and model assets while they are actively being processed. The entire AI inferencing pipeline, including its CPUs, GPUs, and data, can execute within a hardware-isolated environment that cannot be accessed by the operators managing the AI hardware infrastructure.

What is a “confidential AI factory”?

It describes an environment where state-of-the-art AI can operate on regulated and sensitive enterprise data while the data owner and model provider retain control of their respective protected assets.

Can VAST DataEnclave support proprietary AI models?

Yes. VAST DataEnclave creates a neutral, verified environment where proprietary models can operate against sensitive enterprise data while both remain protected from third parties and each other. This enables enterprises to use advanced proprietary models without surrendering control of their data, while model providers can securely serve those customers without exposing their model IP.

Is DataEnclave a new VAST engine?

No. VAST DataEnclave is the confidential AI capability of VAST DataEngine, combining secure container execution and cryptographic attestation with customer-controlled and third-party key management infrastructure. It extends the VAST AI Operating System with confidential execution for sensitive AI workloads without introducing another engine or infrastructure layer.

Who provides the compute runtime?

VAST provides and manages the secure container runtime as an integrated capability of VAST DataEngine. The runtime is compatible with the CNCF Confidential Containers (CoCo) ecosystem, including runtimes such as Kata Containers.

What is the core security benefit of confidential execution?

Confidential execution provides three fundamental protections: confidentiality, with workload memory protected from the host; isolation, with workloads separated from the host environment; and authenticity, with the execution environment cryptographically verified against approved baseline security requirements.

Does VAST DataEnclave work in air-gapped environments?

Yes. VAST DataEnclave supports both connected and fully air-gapped deployments. Connected environments use the VAST-managed DataEnclave attestation and key brokering services, while fully air-gapped environments support on-premises attestation and key brokering without requiring external network connectivity.

How can I audit confidential AI workloads?

VAST DataEnclave maintains a tamper-proof, queryable audit trail in VAST DataBase covering attestation events, key releases, policy decisions, and workload lifecycle activity. Security and compliance teams can verify what ran, where, and under what approved policy without exposing the sensitive data or proprietary models protected inside the confidential environment.

Technical Details & Architecture

How does VAST handle attestation?

For connected environments, VAST provides and manages the DataEnclave attestation service, which builds on the open-source components aligned with CNCF Confidential Containers and integrates with standard KMS providers.

How do completely air-gapped environments work?

For deployments with zero external connections, DataEnclave provides on-premises attestation and key brokering within the isolated environment, enabling the same verify-before-decrypt trust model without external network connectivity.

Who manages the Key Management System (KMS)?

The owners of protected assets retain control of their respective keys and key management infrastructure, using cloud KMS options (e.g., AWS KMS) or enterprise on-premises solutions (e.g., Thales, Fortanix). Customer data remains encrypted with customer-managed keys, while model-weight keys remain within the model provider's trust domain.

Security, Trust Model & Attack Surfaces

What is the CoCo Trust Model implementation?

The attestation report generated by the Attestation Agent measures hardware, guest firmware, kernel, filesystem, and workload configuration. An unsuccessful attestation denies secret release to the workload, preventing unverified or tampered workloads from being decrypted.

What does “verify before decrypt” mean?

Before protected data or models can be accessed, VAST DataEnclave cryptographically verifies that the hardware, software environment, and workload configuration meet approved baseline security requirements. If verification fails, the necessary keys and secrets remain unavailable, preventing the protected assets from being decrypted inside an untrusted environment.

Who controls the encryption keys?

The owners of protected assets retain control of their keys. Enterprise data can remain protected by customer-controlled keys while proprietary model weights remain under the model provider’s control. VAST DataEnclave integrates with existing key management infrastructure, including AWS KMS, Google Cloud KMS, Microsoft Azure Key Vault, HashiCorp Vault, Thales, and Fortanix, rather than requiring VAST to own the keys.

Can a cluster administrator access confidential workload data?

No. CPU memory is encrypted, while GPU memory is protected from unauthorized access when GPUs operate in Confidential Computing mode. Communication within the confidential environment is also protected from the host. In the Vera Rubin platform, NVLink and NVLink-C2C traffic is encrypted. The host administrator cannot inspect protected workload memory or execute unauthorized processes inside the confidential environment.

How is workload authenticity verified?

Workload authenticity is verified through signed or encrypted container images. Upon successful attestation, the Key Broker Service within DataEnclave releases the necessary key to the confidential environment: either a public key for signature verification or a decryption key for an encrypted image. Separately, policy authorizes pulling only images that match an approved digest.

How are Man-in-the-Middle (MITM) attacks prevented during attestation and secret release?

- Attestation report forging: The attested evidence report generated by the enclave’s attestation agent is signed by a chip-unique hardware secret to protect against tampering. The report includes the enclave’s configuration, a public key, and a fresh nonce to eliminate replay.
- Secret interception: After verification, DataEnclave’s Key Broker Service releases keys encrypted specifically with the enclave’s public key. Decryption occurs strictly inside the enclave using the enclave’s private key.

What are the primary attack surfaces in this architecture?

Assuming the Attestation Verification Service runs in a trusted environment, attack surfaces are limited to:

- Compromised guest components (e.g., DataEnclave agent that enforces policies within the enclave) or buggy guest code inside the enclave. Attestation verifies these components are authentic and unmodified, but cannot guarantee they are free of exploitable flaws.
- Compromised hardware roots of trust within nodes, the foundational trust assumption on which the entire model depends.
- Vulnerabilities within the user workload application itself (e.g., unsafe APIs). DataEnclave governs what containers can run, but it does not govern what happens inside of them.

GPU & Workload Execution

How are GPUs used inside a Confidential Virtual Machine (CVM)?

GPUs running in Confidential Computing mode are assigned to an enclave, and GPU hardware measurements are validated during attestation. CPU memory is encrypted, while GPU memory is protected from unauthorized access and communication between the CPU and GPUs is secured. This enables standard CUDA workloads to execute within the same confidential trust domain.

Operations & Debugging

How are Confidential Computing environments debugged and investigated?

Traditional forensic methods do not work. CPU guest memory is encrypted, and host-side memory dumps return ciphertext, while GPU memory is protected from host access when operating in Confidential Computing mode. Debugger attach is blocked by policy bound into the attestation measurement. Debugging and operational monitoring are performed using approved in-guest tools, such as secure logging, metrics, and health probes, unless guest policy explicitly allows data to be emitted to the host.

Can administrators monitor a confidential workload without accessing its contents?

Yes. DataEnclave separates operational visibility from access to protected workload contents. Security and operations teams can monitor attestation status, workload lifecycle activity, policy decisions, and approved in-guest telemetry without gaining access to protected memory, enterprise data, model weights, or cryptographic secrets.

Bottom Line

Does software need to be repackaged to run in DataEnclave, or do existing containers work as-is?

There is no need to rebuild the application components or containers. However, deployment automation will have to be tailored to VAST DataEnclave for consistent deployments. DataEnclave will provide its own Infrastructure as Code (IaC) service analogous to CloudFormation.

How does a customer actually buy and deploy proprietary AI models through DataEnclave?

VAST expects customers will bring their own licenses for deploying models and apps through DataEnclave through existing relationships with model providers and other ISVs.

About VAST Data

VAST Data is the AI Operating System company – powering the next generation of intelligent systems with a unified software infrastructure stack that was purpose-built to unlock the full potential of AI. The VAST AI OS consolidates foundational data and compute services and agentic execution into one scalable platform, enabling organizations to deploy and facilitate communication between AI agents, reason over real-time data, and automate complex workflows at global scale. Built on VAST's breakthrough DASE architecture – the world's first true parallel distributed system architecture that eliminates tradeoffs between performance, scale, simplicity, and resilience – VAST has transformed its modern infrastructure into a global fabric for reasoning AI. Learn more at **vastdata.com** and follow VAST Data on **LinkedIn**, **YouTube** and **X**.

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

09/26 I 14790