

Velocity -V with VCF 9.0

Ensuring privacy (CIA triad)
for AI workload



Table of CONTENTS

VMware based Private Cloud ensuring privacy in CIA triad with VCF 9 Series	3
Shared responsibility principle in Private AI.....	3
VelocITy-V (VMware Cloud Foundation Platform).....	4
C. Confidentiality	4
Securing model access.....	4
User privacy controls.....	4
Platform controls.....	4
Access control.....	4
API Security.....	5
Model security.....	6
I. Integrity	7
Dynamic Nature of Gen-AI Applications.....	7
Data Integrity and Traceability.....	7
Tamper Detection and Mitigation.....	8
Continuous Monitoring and Security Compliance.....	8
A. Availability	9
Disaster Recovery for AI Workload.....	9
Resource Optimization for AI Workload.....	10
Data Availability.....	10
Live Migration & Network Availability.....	10
Air Gaped environment for Private AI.....	11
Conclusion	12
About the Author	13

VCF 9 series is a game changer for Private AI Practice, ensuring privacy in CIA triad for VMware based Cognitive Infrastructure part of HCLTech's AI Foundry

In the modern data-driven landscape, safeguarding sensitive information and maintaining privacy have become paramount for enterprises. Traditional AI solutions often rely on transmitting data to external servers, exacerbating concerns about data security and compliance. This challenge is further

To address these challenges, VMware based Private AI for Cognitive Infra emerges as a transformative architectural approach, designed to harmonize the organizational benefits of AI with the stringent requirements of privacy and compliance. By bringing AI and ML technologies in-house, organizations can reclaim control over their proprietary data, models, and intellectual property. Private AI for Cognitive Infra equips

compounded by stringent regulatory frameworks such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the Gramm-Leach-Bliley Act (GLBA).

Reliance on external AI-related APIs amplifies these risks, exposing organizations to vulnerabilities such

enterprises with the flexibility and tools necessary to securely and efficiently deploy AI workloads, fostering innovation driven by responsible AI practices.

as fluctuating pricing, reduced control, and potential deprecation of critical features. Additionally, the unpredictable nature of proprietary SaaS models—frequently subject to unannounced updates—can adversely affect operational performance and compliance readiness.



Shared responsibility principle in VMware based Cognitive Infrastructure part of HCLTech's AI Foundry

In the context of Private Cloud AI security and privacy are collaborative responsibilities that require active participation from both IT operations (IT Ops) and data science teams.

The IT Ops team is tasked with securing the platform and infrastructure services, implementing measures such as access controls, encryption, and intrusion detection to prevent unauthorized access or breaches. Meanwhile,

- The data science team focuses on safeguarding the application and its processes, ensuring the ethical and

confidential handling of data while protecting the infrastructure from misuse or malicious intent. While privacy and security are interconnected, they serve distinct purposes: privacy ensures the responsible use of data, protecting inputs, interactions, and outputs from unethical handling, whereas security prevents unauthorized access, data manipulation, and operational disruptions.

To guide these efforts

CIA triad—**confidentiality**, **integrity** and **availability**—serves as a foundational framework.

- **Confidentiality** emphasizes protecting sensitive data used to fine-tune and operate AI models, with IT Ops enforcing encryption and access controls while data scientists securely access and manage this data.
- **Integrity** ensures the accuracy and trustworthiness of both data and AI models, this requiring safeguards like intrusion detection and code verification to prevent data corruption or manipulation.
- **Availability** ensures that authorized users can access AI applications and their outputs

without interruption, a responsibility shared between IT Ops and DevOps teams to maintain robust infrastructure.

Organizations can systematically identify vulnerabilities, map them to key security principles, and

implement targeted controls by integrating the STRIDE threat modelling framework with the CIA triad. This approach ensures the confidentiality, integrity, and availability of AI workloads, providing a robust defence against potential threats.

The different threats are **Spoofing, Elevation of Privilege, Data Tampering, Info Disclosure and Denial of Service.**

VelocITy-V with VCF9.0 (VMware Cloud Foundation)

C Confidentiality

Confidentiality ensures data privacy by restricting access to sensitive information and AI models, protecting against threats like information disclosure. Techniques such as **data minimization, anonymization and strong encryption** help safeguard data, reducing risks and maintaining security even in the event of a breach.



C.1 Securing model access

Inference server frameworks often expose APIs (via HTTP/HTTPS or gRPC) to allow external applications—such as chatbots, services, or other applications—to query the hosted models. However, these APIs typically lack built-in authentication and authorization, creating a potential security vulnerability.

To address this, VMware recommends placing an API gateway in front of the inference server's API. The API gateway acts as a security layer, managing authentication and authorization for applications accessing the model API. It supports various authentication methods, with OAuth 2.0 being the recommended standard due to its robust security features.

C.2 User privacy controls

To protect data, organizations should prioritize privacy and security measures like anonymization and encryption based on data sensitivity. Data science teams should focus on minimizing and anonymizing data, while IT Ops ensures encryption both in transit and at rest. Data integrity is maintained through

cleansing, validation, and tools like checksums or cryptographic hashes to detect tampering. Reducing data collection and removing personal information, such as redacting PII (Personal Identifiable Information) before indexing or embedding in vector databases, helps comply with regulations like GDPR and HIPAA and minimizes security risks.

To address this, Libraries like Private AI from VMware, integrated into frameworks like LangChain, streamline PII redaction. Combining access control mechanisms with robust encryption enhances data security by restricting unauthorized access and limiting the attack surface.

C.3 Platform controls

To enhance security, indexing and retrieval workflows, as well as data transfers, should be encrypted. If data sources contain PII, encrypting data at rest is also recommended to protect against breaches. The strength of encryption should match the sensitivity of the data, with highly sensitive information requiring more robust encryption algorithms.

Platform control involves control of Data Encryption, Data Access control and Model Security.

C.4 Access Control

VCF 9.0 & upcoming 9.0 ensures robust access control by restricting unauthorized access to data and models through user authentication and authorization mechanisms. A key component of this is Identity and Access Management (IAM), which plays a vital role in securing workloads, applications, and data within the Private AI for Cognitive Infrastructure Foundation.

The platform includes native Role-Based Access Control (RBAC) capabilities, enabling organizations to assign roles and responsibilities to users based on their personas and job functions. This targeted access minimizes the risk of unauthorized interactions with resources.

VCF also supports identity federation, allowing administrators and operators to use existing corporate credentials for seamless access to resources. The platform integrates with external identity providers such as Active Directory Federation Services (AD FS), Microsoft Entra ID, and Okta. This streamlined approach reduces the need for additional usernames and passwords, simplifying access management while enhancing overall security.

C.5 API security

VCF 9.0 & upcoming 9.0 enhances API security to ensure secure communication between application components, protecting user data and preventing unauthorized access. The platform mandates the use of secure communication protocols, with HTTPS as the baseline standard, and

implements TLS (Transport Layer Security) for advanced protection. TLS encrypts the communication between clients and Gen-AI applications, safeguarding data from interception by scrambling it into an unreadable format. It also uses digital certificates to authenticate application endpoints, ensuring users are connecting to legitimate services and not malicious decoys.

Furthermore, TLS verifies the integrity of transmitted data, detecting unauthorized modifications during transit. By combining encryption, authentication, and integrity checks, TLS significantly reduces the risk of man-in-the-middle attacks, ensuring secure and reliable communication for Gen-AI applications within the VCF ecosystem.

The more complex a system is, the higher the risk of data leaks.

In an LLM system that uses an orchestration layer and a vector database to store custom or private data, skilled attackers can manipulate prompts to find and exploit weaknesses, gaining access to information they shouldn't have.



Risk

Data Leakage in complex systems

- **Cause:** Systems with orchestration layers and vector databases storing proprietary data are vulnerable to attackers manipulating prompts or accessing the database.
- **Impact:** Attackers could extract sensitive information, such as financial data or internal documents, with high accuracy, especially if the system continuously accesses additional data sources.



Solutions

Securing vector databases and systems

- **Access control:** Use Access Control Lists (ACLs) and Role-Based Access Control (RBAC) to restrict access to sensitive data.
- **Data Isolation:** Ensure data in the vector database is isolated based on classification levels.
- **Encryption:** Encrypt data at rest and in transit to protect it from unauthorized access.
- **Monitoring:** Implement secure audit logs and user activity monitoring to detect and respond to unauthorized actions.

VCF 9.0 for Private AI practice

Securing containerized applications

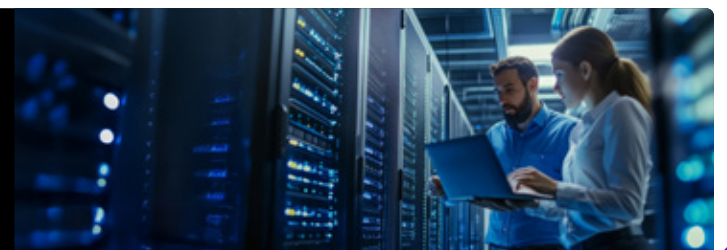
VMware Harbor Registry:

Provides RBAC to control access to container images and a curated catalogue of images to maintain a gold standard for AI application development.

RBAC Benefits:

Ensures only authorized users or applications can access, push, or pull images, reducing security risks.

By combining these measures, organizations can mitigate data leakage risks and secure their AI systems effectively.



C.6 Model security

Fine-tuned foundation models often contain sensitive information from the training data, which cannot be "unlearned" once incorporated.

Unauthorized users can potentially extract this data using specific prompts if they gain access via the model's UI or API. To protect this information, it's critical to restrict access rigorously. Additionally,

organizations has to address other potential attack vectors beyond the UI and API to ensure comprehensive security.



Risk

To keep traditional ML models and foundation models secure, trusted execution environments (TEEs) and secure enclaves create isolated areas that protect the model and its data from the rest of the system, reducing the chances of unauthorized access.

A big challenge in serving confidential models, like traditional models or LLMs, is ensuring that

data stays encrypted even during processing, which is when it's most vulnerable to attacks like memory dumps.

Even though VMware's hypervisor is highly secure, as proven by its Common Criteria certification, application secrets (like sensitive data or encryption keys) are still stored in system memory, CPU cache, or GPU memory.



Solutions

VCF, powered by vSphere, provides robust capabilities for creating secure enclaves on modern CPUs, leveraging technologies such as Intel Software Guard Extensions (SGX) and AMD Secure Encrypted

Virtualization-Encrypted State (SEV-ES). These secure enclaves within virtual machines isolate the fine-tuning process and model inference, safeguarding the confidentiality of both the data and model parameters.



Gen-AI applications are dynamic and continuously evolving, making it challenging to track their components and identify vulnerabilities. To enhance security, Software Bill of Materials (SBOMs) provide a detailed inventory of all components used to build these applications, including open-source libraries, frameworks, pretrained models, datasets, data sources, and custom code. SBOMs improve transparency and traceability, enabling better management and protection



I.1 Dynamic nature of Gen-AI applications



Risk

Gen-AI applications evolve over time, making it difficult to track their components and vulnerabilities. Missing transparency introduces

risks from changing data sources, pretrained models, and third-party code.



VCF Solution

- **Software Bill of Materials (SBOM):** Offers comprehensive inventories of all software components, datasets, pretrained models, and custom code. Linked with vulnerability databases, SBOMs provide alerts for flaws, enabling timely patches.
- **VCF Tamper Detection:** Content libraries ensure that only approved VM images and Tanzu Kubernetes Grid distributions are

- used. OVF template security policies enforce strict validation, while trusted CA certificates enhance security.
- **Harbor Registry:** Signs and verifies container images, scanning for vulnerabilities and preventing malicious code injection. Supports third-party tools like Cosign for enhanced image verification.

I.2 Data integrity and traceability



Risk

Ensuring the authenticity and reliability of model-generated data is challenging, especially during fine-tuning, updates, or migrations.



VCF Solution

- **Secure Boot and Live Migration:** Ensures only authorized OSs load and minimizes downtime during VM migrations.
- **VM Snapshots with Version Control:** Provides rollback points to address model

- corruption or errors while maintaining traceability.
- **Template Security:** Ensures the integrity of deployment templates with strict OVF validation and signing certificates.

I.3 Tamper detection and mitigation



Risk

Unauthorized modifications to infrastructure or data can compromise application security.



VCF Solution

- **Content Libraries:** Centralized repositories for approved VM images.
- **Harbor Registry:** Prevents use of compromised container images.
- **vSphere Lifecycle Manager (vLCM):** Enforces consistent, secure configurations and flags non-compliance, ensuring tamper resistance across ESXi hosts.



I.4 Continuous monitoring and security compliance



Risk

Detecting and addressing anomalies or vulnerabilities requires consistent monitoring and auditing.



VCF Solution

- **VMware Aria Operations:** Monitors infrastructure health, performance and security, detecting threats or anomalies in real-time.
- **vLCM Auditing:** Ensures secure provisioning, configuration, and baseline compliance for ESXi hosts, with image health checks and automated remediation for deviations.
- **Cluster-Aware Updating:** Coordinates updates across ESXi clusters with minimal downtime, ensuring security and workload availability.

By integrating these robust features, VCF provides a comprehensive solution to address the dynamic challenges of Gen-AI applications, ensuring transparency, security, and operational efficiency.

A Availability

Availability refers to ensuring that platforms, networks, and applications operate reliably and consistently, providing uninterrupted access to information and services



VCF enhances availability by delivering a robust set of features designed to keep AI applications running smoothly, even in the event of outages or disruptions

vSphere High Availability (HA)

- Clusters ESXi hosts and monitors their health and the VMs running on them.
- Automatically restarts VMs on healthy hosts in the cluster in case of host failure, ensuring minimal downtime for AI applications.

Dynamic DirectPath I/O

- Provides flexible assignment of hardware accelerators to workloads based on attributes rather than hardware addresses.
- Ensures compatibility with vSphere HA and Distributed Resource Scheduler (DRS) by allowing seamless workload placement and recovery on hosts with similar hardware.

These features collectively enable robust high availability for AI workloads, ensuring consistent performance and minimizing service interruptions.



A.1 Disaster Recovery for AI workload

VCF 9.0 and upcoming 9.0 ensures disaster recovery for Gen-AI applications through Site Recovery Manager (SRM) and vSphere Replication, which work together to protect AI workloads and their associated data.

Site Recovery Manager (SRM)

- Provides automated orchestration for disaster recovery.
- Helps IT teams identify critical components and dependencies of Gen-AI applications, ensuring they are prioritized during recovery.
- Manages failover and failback processes between primary and recovery sites.

vSphere Replication

- Replicates essential data, including AI models, datasets, and application configurations, to the designated recovery site.
- Supports point-in-time recovery options, allowing flexibility to restore the most accurate and relevant data.

These tools ensure that even after a major outage, Gen-AI applications can be swiftly restored with minimal disruption, maintaining their operational integrity and availability.

A.2 Resource Optimization for AI workload

Resource optimization is a crucial benefit of the VMware platform, ensuring that AI applications run efficiently by managing workloads and resources effectively.

Distributed Resource Scheduler (DRS)

- **Intelligent VM distribution:** Works with vSphere HA to balance workloads across ESXi hosts in a cluster.
- **Optimizes resource utilization:** Prevents overloading of any single host, ensuring that all resources are used efficiently, and AI applications maintain optimal performance.
- **Minimizes workload interference:** Reduces the impact of demanding workloads on AI applications by intelligently balancing resources.
- **Host failure mitigation:** In case of a host failure, DRS automatically selects the most appropriate healthy host to minimize disruption to AI workloads.

These features help optimize resources, improve the performance of AI applications, and maintain stability even during system failures or increased demand.

A.3 Data availability

Support for shared storage solutions

- VCF supports various shared storage options, including block-level, file-level and object-based storage access for VMs and containers.

Data mirroring and replication

- Configurable mirroring and replication of data ensure availability even during storage or host failures.

These features help maintain data availability by ensuring that data is accessible and protected across multiple storage environments.

A.4 Live migration and network availability

vMotion for Live Migration

- Enables live migration of running AI applications between ESXi hosts within a cluster, ensuring near-zero downtime during infrastructure maintenance.

Network redundancy for availability

- ESXi hosts are configured with multiple redundant network paths, ensuring network availability even if one path fails.

VMware NSX for Network Availability

- VMware NSX, paired with vSphere HA, enhances network failover times for AI applications.
- NSX automatically replicates its configuration across multiple managers to minimize network traffic disruptions.

VCF (VMware Cloud Foundation) offers a comprehensive set of features that help maintain the availability and performance of AI applications. These features ensure that AI applications remain operational, even during maintenance or unexpected failures,

by offering solutions like vMotion for live migration, vSphere High Availability (HA) for automatic VM restarts, and redundant network paths for uninterrupted connectivity. VCF's ability to optimize resources and provide data availability ensures that AI applications can scale and

perform effectively, while network virtualization with NSX enhances failover capabilities. Together, these features guarantee that AI applications continue running smoothly and provide uninterrupted access to critical information.

A.5 Air Gaped environment for VMware based Private AI under Cognitive Infrastructure with HCLTech's AI Foundry

For AI workloads, especially those involving sensitive data or proprietary models, maintaining privacy and security is crucial. AI systems often process vast amounts of personal, confidential, or regulated data, and any breach or unauthorized access could have severe consequences, including legal

repercussions, financial loss, and reputational damage.

An air-gapped environment, or a more secure "electronic air-gap," ensures that these AI workloads are isolated from potentially vulnerable external networks, reducing the risk of cyberattacks, data leaks, or

unauthorized access. Since AI models and their underlying datasets can be valuable targets for attackers, securing these systems with an air-gap architecture adds an extra layer of protection by limiting the avenues through which malicious actors can gain access.

Features of Private AI practice on VCF with NAIE for Air-Gapped Environment for AI Workload:

1. VMware Harbor Registry

Used as a private container registry to provide images and Helm charts from both internal and external sources, including NVIDIA GPU Cloud (NGC), in an air-gapped environment.

2. NVIDIA Delegated License Service (DLS)

Can be hosted on-premises to store NVIDIA AI Enterprise (NVAIE) licenses offline, allowing license management without external connectivity.

3. NVIDIA Licensing Portal

Licenses for NVIDIA AI Enterprise can be downloaded from the NVIDIA licensing portal for offline use in an air-gapped setup.



4. VMware Deep Learning VM Templates

Provides pre-configured virtual machine templates to quickly start AI projects and experiments, tailored for AI workloads.

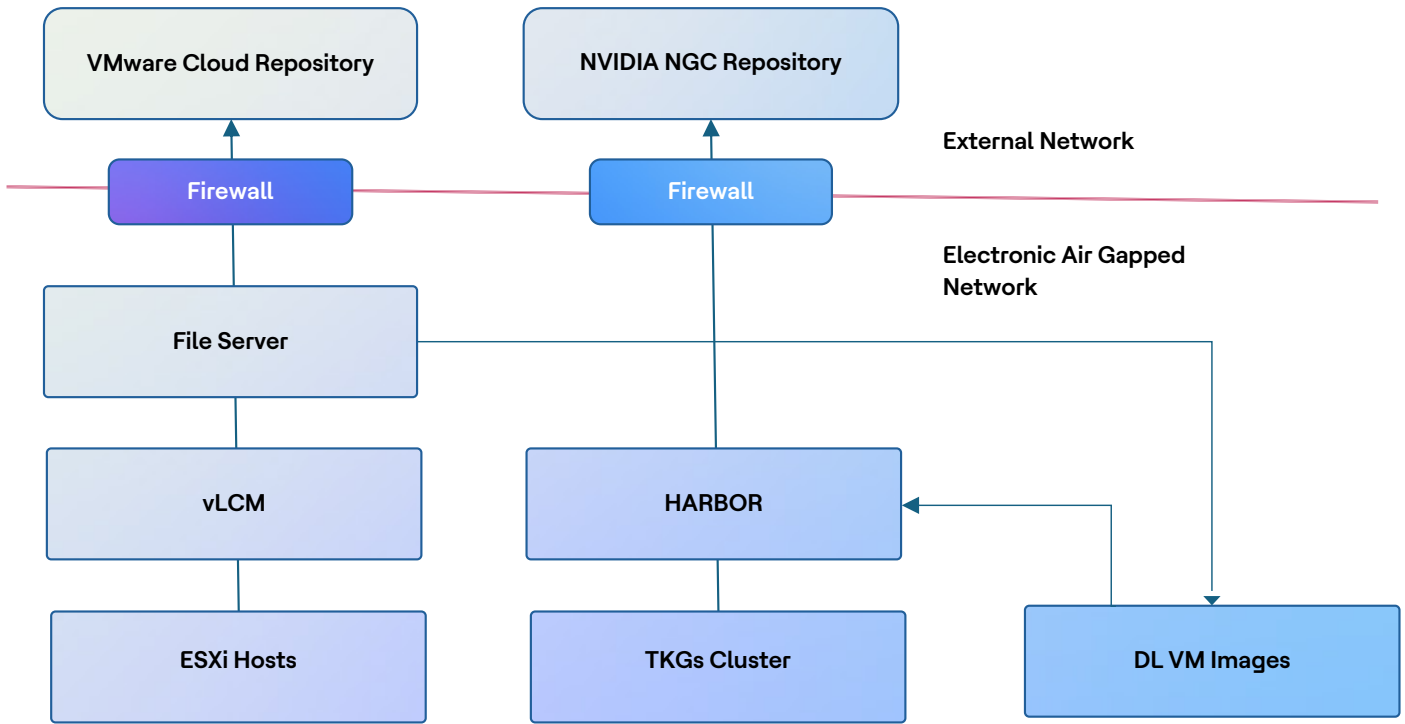
5. NVIDIA NGC Containers

VMware deep learning VM templates support pulling NVIDIA NGC containers for AI workloads, even in an air-gapped environment.

6. GPU Driver File Share

A local file share must be available containing the GPU driver for the template installation in the air-gapped environment.

Block diagram of VCF 9.0 based Air-Gaped with NVIDIA NGC



7. Offline Software Repositories

Local repositories and offline licensing functionality minimize external connectivity by ensuring all required software and licenses are accessible within the air-gapped system.

Conclusion

HCLTech's VelocTy-V offering now integrated with VCF 9 delivers a robust, secure, and agile platform for hybrid cloud Infra for heterogeneous environment, addressing critical needs across security, observability, and cost efficiency. Its Zero-Trust architecture, unified management, and Kubernetes-native capabilities empower enterprises to modernize infrastructure while maintaining industry specific governance & compliance.

HCLTech help organizations to navigate initial cost outlays and skill gaps to fully leverage its potential. For industries like Financial Services, Life Sciences, Manufacturing and

Retail, VCF 9 offers a future-proof foundation, enabling scalable, compliant, and optimized Multi cloud operations.

From a technical standpoint, VCF 9.0's integration with HCLTech's VelocTy framework accelerates hybrid cloud adoption by streamlining people, processes, and operations. By combining validate automation and security architecture with HCLTech's expertise, enterprises can overcome deployment hurdles and achieve faster time-to-value. This synergy not only reduces operational overhead but also ensures a seamless transition to a modern,

cost-effective hybrid cloud model—positioning businesses for sustained growth in an evolving AI enabled digital landscape.





Manish Chauhan

Group Manager,
Product Management Group,
Hybrid Cloud Business Unit,
HCLTech

About the Author

Manish Chauhan brings 18 years of industry experience in digital transformation to his role as Product Owner for HCLTech's offering 'VelocITy' and 'Private AI as a Service.' He empowers customers to leverage Generative AI on VMware-based private cloud infrastructure, ensuring security, data privacy, and ethical AI practices with robust guardrails.

HCLTech | Supercharging
Progress™

HCLTech is a global technology company, home to more than 223,000 people across 60 countries, delivering industry-leading capabilities centered around digital, engineering, cloud and AI, powered by a broad portfolio of technology services and products. We work with clients across all major verticals, providing industry solutions for Financial Services, Manufacturing, Life Sciences and Healthcare, High Tech, Semiconductor, Telecom and Media, Retail and CPG, and Public Services. Consolidated revenues as of 12 months ending June 2025 totaled \$14 billion. To learn how we can supercharge progress for you, visit hcltech.com.

hcltech.com

