BRIEF

SEPTEMBER 2025

VMware Private AI Services in VCF: Addressing Challenges in Enterprise AI Deployment

Torsten Volk, Principal Analyst

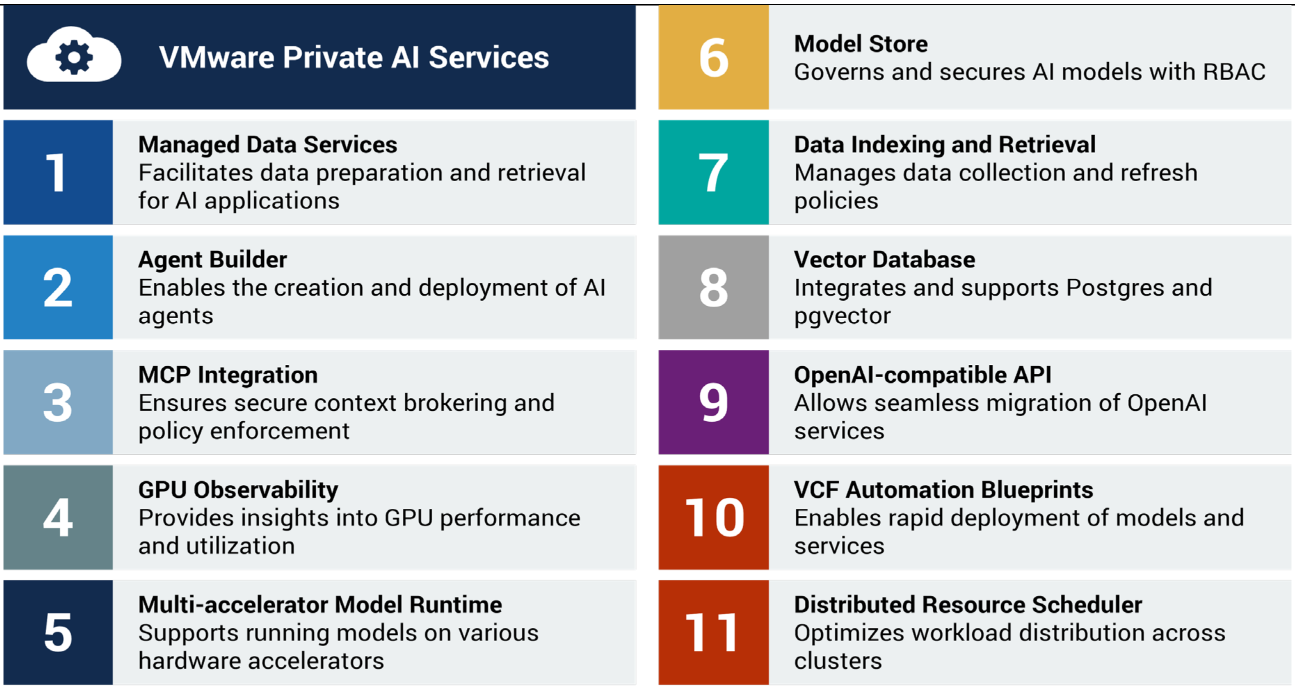
Overview

The integration of VMware Private AI Services with the company's private cloud platform, VMware Cloud Foundation (VCF), aims to significantly enhance VCF by offering developers turnkey AI capabilities. At the same time, it enables virtualization administrators and other operational roles to provision infrastructure resources and establish guidelines for compliance, security, cost, and performance. This approach restores control to virtualization administrators while letting developers easily deploy AI stacks through declarative APIs, including Terraform templates. This addresses key challenges hindering AI project success, such as skill shortages, compliance issues, and cost concerns.

Key Capabilities

VMware Private AI Services, integrated with VCF, provides core features for developers to build and deploy applications, and for IT operators to manage policies and lifecycle activities:

Figure 1. VMware Private AI Services



VMware Private AI Services

1	**Managed Data Services** Facilitates data preparation and retrieval for AI applications
2	**Agent Builder** Enables the creation and deployment of AI agents
3	**MCP Integration** Ensures secure context brokering and policy enforcement
4	**GPU Observability** Provides insights into GPU performance and utilization
5	**Multi-accelerator Model Runtime** Supports running models on various hardware accelerators
6	**Model Store** Governs and secures AI models with RBAC
7	**Data Indexing and Retrieval** Manages data collection and refresh policies
8	**Vector Database** Integrates and supports Postgres and pgvector
9	**OpenAI-compatible API** Allows seamless migration of OpenAI services
10	**VCF Automation Blueprints** Enables rapid deployment of models and services
11	**Distributed Resource Scheduler** Optimizes workload distribution across clusters

Source: Enterprise Strategy Group, now part of Omdia

Managed Data Services

- **Operators** deliver native data ingestion, chunking, and indexing pipelines to prepare enterprise content for semantic search and retrieval-augmented generation (RAG).
- **Developers** utilize managed vector stores and retrieval APIs (e.g., embeddings, top-k queries) to enable private search and grounding without the burden of operating databases. This also includes a fully supported vector database (Postgres with pgvector).

Agent Builder

- **Operators** oversee agent templates, tool permissions, and publishing, ensuring approved agent workflows are instantly accessible across tenants.
- **Developers** assemble tools, models, and knowledge bases into agentic workflows and invoke them via the built-in API gateway for synchronous or asynchronous use.

MCP (Model Context Protocol) Integration

- **Operators** enforce secure context brokering, policy, and auditing for MCP tool and data connections across environments.
- **Developers** register MCP tools and data sources so agents can safely fetch relevant application context using a standardized protocol.

GPU Observability

- Operators gain comprehensive visibility into GPU utilization, memory, thermal status, and saturation to optimize placement, capacity, and costs.
- Developers access workload-level telemetry and alerts to fine-tune inference performance for batch and online tasks without manual instrumentation.

In essence, integrating VMware Private AI Services with VCF aims to maximize developer productivity while providing operators centralized, policy-based control over security, compliance, performance, and costs.

Competitive Analysis

Hyperscalers compete by offering a wide array of capabilities for easy integration, including speech-to-text, document AI, computer vision, natural language processing analysis, forecasting, recommendations, and enterprise search. These capabilities are seamlessly connected through hyperscalers' event frameworks and serverless runtimes. They also provide ARM-based GPUs for model training and inference, along with extensive marketplaces featuring ready-made agents and tools from third parties and first-party vendors. Various AI models are available, some tailored for specific use cases or agent types. These services aim to make it simple for developers to leverage the hyperscaler ecosystem, maximizing customer loyalty.

VMware's value proposition centers around delivering a core set of standardized private RAG and agent patterns, with lifecycle governance, staged rollouts, and full control over data pipelines. VCF hosts all AI services within VCF-defined infrastructure, ensuring strict data locality, multi-tenant isolation, and integrated controls for identity, networking, and GPU resources. This enables operators to govern end-to-end workflows independently of external cloud services.

Business Impact and Conclusion

Empowering traditional virtualization administrators to evolve into platform engineers by offering complete AI services within their existing infrastructure is a key advantage of this new solution. Instead of creating separate platform teams for on-premises and cloud services, VCF with integrated Private AI Services lets existing virtualization teams centrally manage these tasks, forming the backbone of the entire organization.

Organizations must weigh these benefits against the convenience of enabling developers to rapidly assemble applications using the wide range of public cloud APIs available today. Ultimately, most will adopt a hybrid approach: keeping sensitive, high-value workloads on VCF for governance, cost control, and data locality, while utilizing public cloud services for access to specialized APIs and elastic scaling.

About Enterprise Strategy Group
Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

✉ contact@esg-global.com
🖥 www.esg-global.com