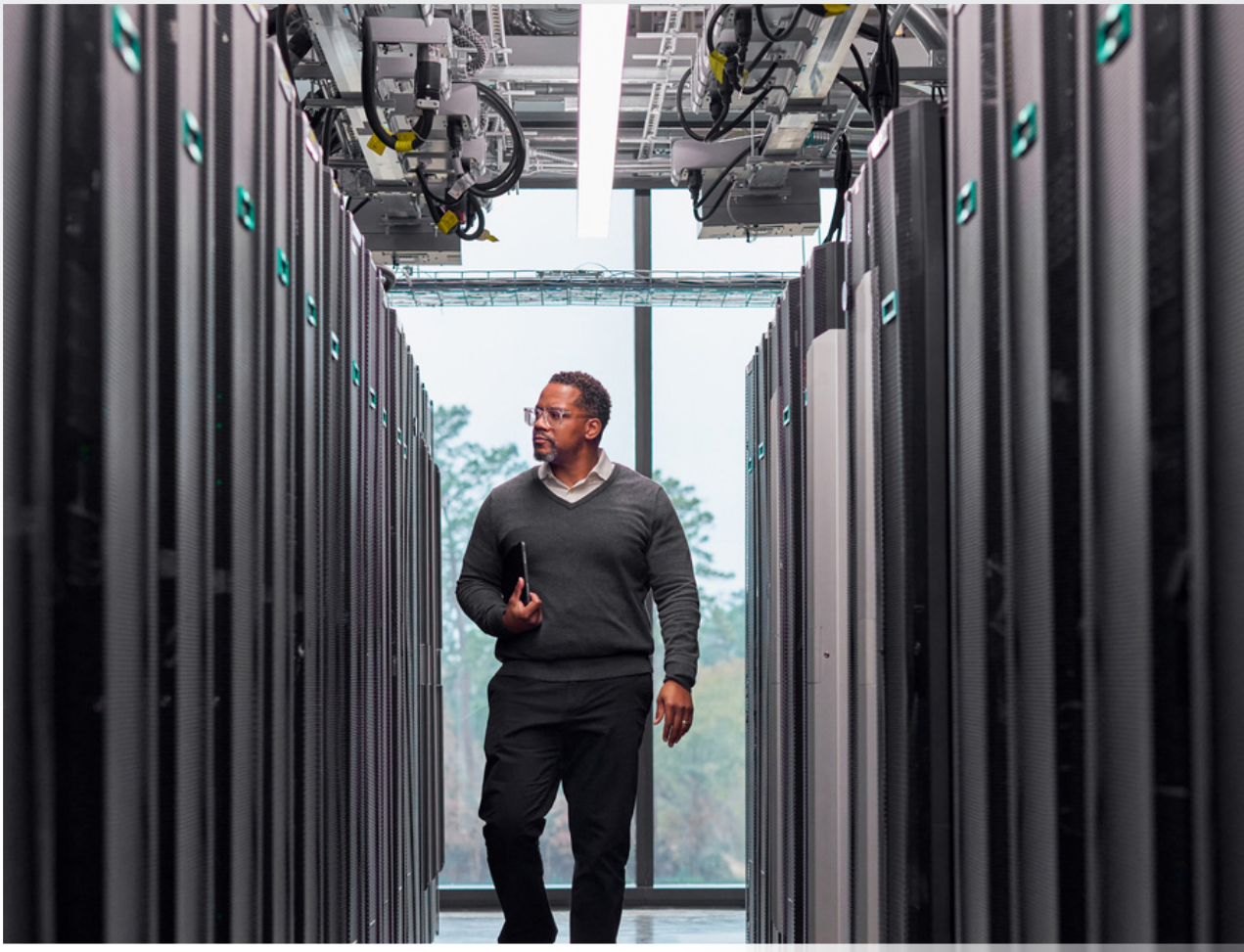
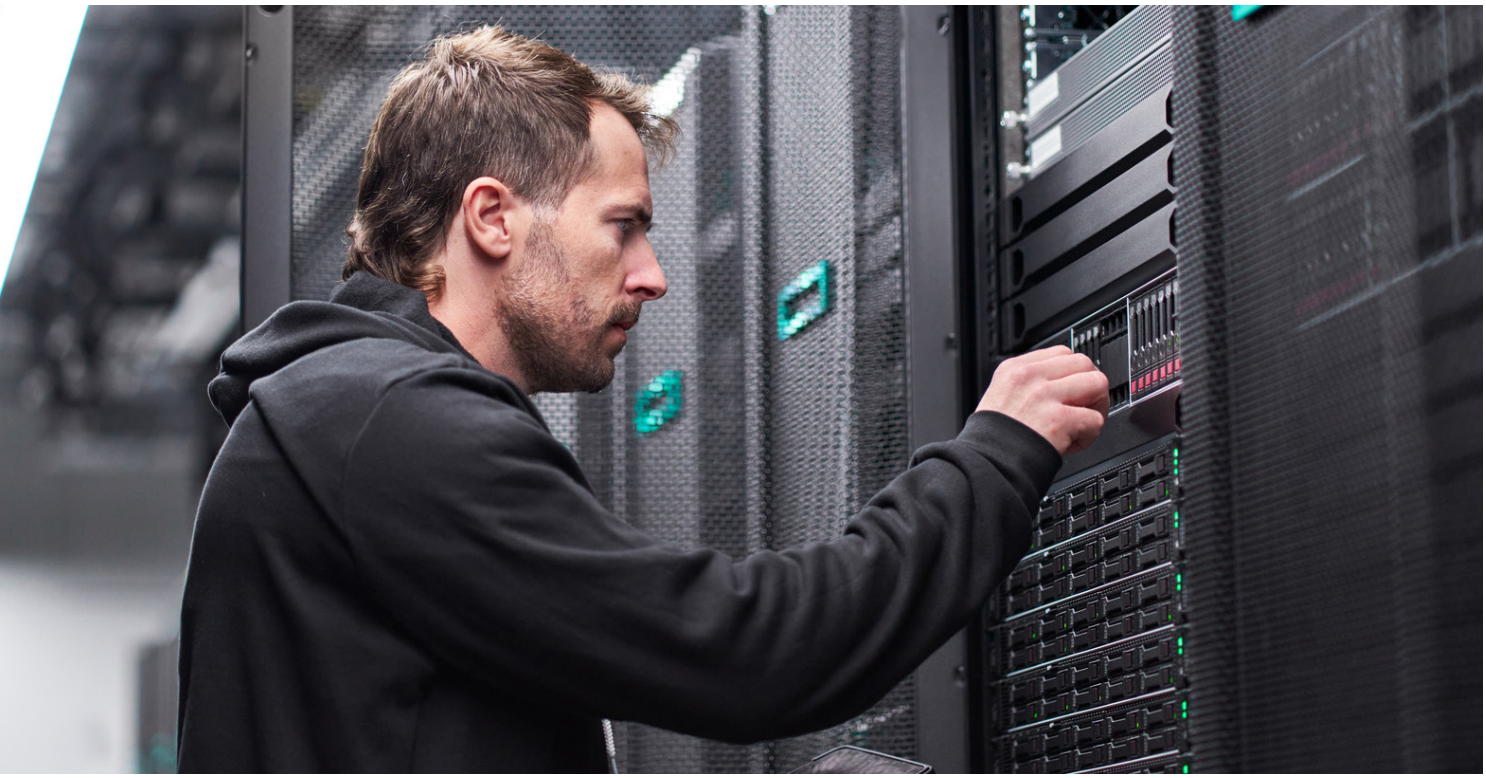
Business white paper



Deploying agentic AI in classified domains

Don’t debate—get your agent for classified information in just two weeks. With HPE Private Cloud AI, deployed in an air-gapped environment that is secured by secunet.

Contents



Abstract

Agentic Artificial Intelligence (AI) and Retrieval-Augmented Generation (RAG) are enabling organizations operating in classified environments to automate complex analysis, reason across classified and sensitive data sources and orchestrate workflows that previously required manual synthesis across siloed systems. These organizations increasingly depend on such capabilities to maintain operational effectiveness but face a difficult trade-off between performance and data sovereignty.

Deploying these capabilities against classified data introduces three key requirements that neither public cloud nor traditional on-premises deployments can satisfy.

1. Infrastructure must be air gapped.

2. Security must be certified to national standards, not merely configured.

3. And the application layer must remain fully governed by the operator, from model selection to pipeline-level access control.

This paper presents a reference architecture that addresses all three requirements: integrating certified security, sovereign infrastructure, optimized inference, and AI orchestration into a unified stack. The architecture enables compliance with the German classified information directives, with central security functions certified for processing classified information up to SECRET and can rapidly accelerate deployment timelines. The final section demonstrates a real-world example of the architecture in practice and the advanced capabilities it enables.

Introduction

AI is providing organizations operating in classified environments with new capabilities across use cases such as classified document analysis, operational reporting, and cross-domain data synthesis. Deploying these capabilities against classified data requires more than air-gapped infrastructure. It demands certified security, hardware-enforced isolation, and full application-level governance.

> Maintaining control over one's own AI in defence will be crucial for one's digital future. This is a legitimate and desirable aim for all (NATO) states. That is because, without national control over data spaces, models and operating environments, there is a risk of capability loss—especially in the context of multinational cooperations.
>
> – NATO Science and Technology Organisation[1]

RAG enables AI systems to retrieve and reason over domain-specific documents, databases, and knowledge bases before generating a response. Rather than relying solely on pretrained model knowledge, RAG grounds outputs in the operator's own data, improving accuracy and relevance for tasks like classified document analysis and operational reporting.

Agentic AI extends this further. Agents can plan multistep tasks, invoke external tools, query multiple data sources, and run workflows with human oversight. In classified environments, for example, analysts can task an agent to correlate across classified document sets, sensor data, and operational reports rather than manually synthesizing across siloed systems. As agentic capabilities mature, the range of automatable workflows continues to expand.

Air-gapped infrastructure addresses external exposure but does not mitigate internal threats, administrative misconfigurations, or insufficient access separation between operators and classified data. For classified environments, security must be certified to national standards and enforced at the hardware level not assumed through network isolation alone.

[1] "Towards Responsible Military Use of Artificial Intelligence through Consistent Strategy, National Governance, International Norms and Ethical Principles." NATO Science and Technology Organisation, February 2026.

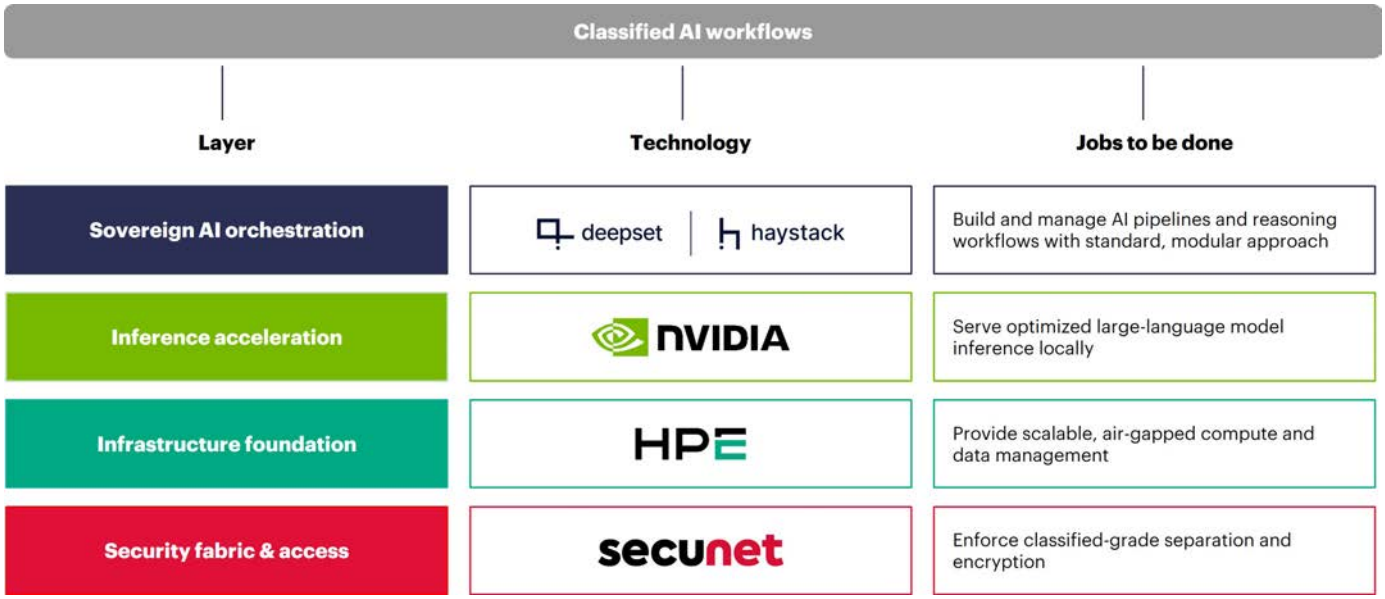


Figure 1. Integrated sovereign reference architecture by secunet, HPE, NVIDIA and deepset—based on air-gapped HPE Private Cloud AI

To address the challenges of developing AI applications in secure environments on classified data quickly and reliably, we present a full stack solution integrating four purpose-built components:

— **secunet SINA Cloud**—BSI-certified security layer enforcing cryptographic domain separation through NVIDIA® BlueField-3 data processing units (DPUs). It protects against both external and internal threats and separates security administration from classified data access.

— **HPE Private Cloud AI**—turnkey AI factory solution, co-engineered with NVIDIA, that provides air-gapped, prevalidated compute infrastructure with cloud-like usability. It provides lifecycle management, role-based access control, and integrated AI tooling out of the box, part of the NVIDIA AI Computing by HPE portfolio.

— **NVIDIA NIM microservices**—containerized inference microservices for deploying optimized open-source large language models (LLMs) on-premises. It runs within the secured infrastructure without dependency on external cloud services.

— **deepset Haystack**—modular, vendor-agnostic AI orchestration platform for building, deploying, and governing agentic and RAG applications. It provides pipeline-level access control and supports both low-code and pro-code development.

The reference architecture in detail

This section describes the components of the reference architecture and their functional role in the proposed stack for developing AI applications in secure environments.

1. SINA Cloud by secunet brings unique security properties certified by BSI

All AI conversations in classified environments start with security. The HPE Private Cloud AI is a private cloud platform that already offers a high level of security, especially in the air-gapped version. However, in particularly security-sensitive environments, strict security measures—from users to workloads—are of particular importance. This additional security is provided by the SINA product family of secunet. With the help of the SINA Cloud network encryption, the production network of the HPE Private Cloud AI is secured with an IT security product that is approved by the Federal Office for Information Security in Germany, BSI, for the processing of classified data.[2]

SINA Cloud enforces its security properties using NVIDIA BlueField-3 DPUs.[3] The secunet software is run on the DPU in zero trust DPU mode.[4] This represents the only entry and exit point for productive traffic (carrying the Haystack application workload) on the HPE Private Cloud AI compute nodes in use. This further increases the level of protection provided, for example, as even:

— Internal attackers have no effective means of extracting data after compromising servers.

— Operating personnel for the fabric network infrastructure cannot gain access to plain text data.

In addition, the well-proven SINA VPN encryption is available to secure user connections to the private cloud infrastructure across public networks. An exemplary, secured network path from the end user to the HPE Private Cloud AI compute nodes is shown in the following figure.

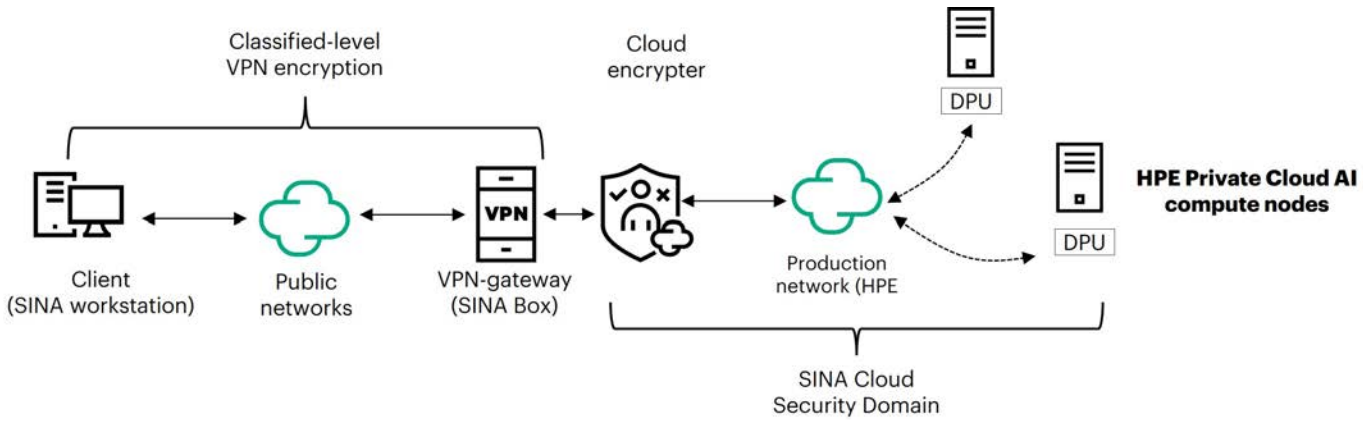


Figure 2. Exemplary, classified-level secured network path from the user to the HPE Private Cloud AI compute nodes

[2] Secure Inter-Network Architecture | BSI

[3] NVIDIA BlueField-3 Data Processing Units—Programmable Data Centre Infrastructure on a Chip | NVIDIA

[4] The NVIDIA BlueField card is operated in a way that the Arm® chip takes control of the PCI switch and is therefore able to not only perform host-independent computations but also control all packets entering and leaving the host. The specific mode of operation is called zero trust DPU mode and enhances security by preventing the host system administrator from accessing BlueField from the host side.

SINA Cloud security layer

With the SINA Cloud, secunet offers a platform that enables the automatic provision of infrastructure as a service (IaaS) while complying with the strict security requirements of the German classified information directive.[5]

In order to enforce the required high level of security and meet the strict requirements, SINA Cloud decouples security-relevant components from the cloud software. This separation makes it possible to expand and update the functional scope of the cloud stack without compromising compliance with the VSA requirements.

Security-related mechanisms in terms of the VSA, in particular for the secure separation of so-called security domains, have been implemented into a logical architecture layer, which can be referred to as the SINA Cloud Security Layer. Secunet's long-term experience in secure networking in classified environments with national and international certifications is also being utilized for SINA Cloud's Security Layer. Most notably, with the help of an implementation of secunet's own high-performance encryption method High-Speed Encryption Acceleration Track (HEAT) on NVIDIA BlueField-3 DPUs, all communication over the cloud fabric is cryptographically secured, enforcing a strong separation between security domains. The mentioned DPUs enforce network security isolation, offload and accelerate network processing.

The implementation on the DPUs ensures that the cloud platform remains trustworthy even if a single Security Domain gets compromised. Moreover, in combination with the SINA Cloud security management, the network encryption supports a separation of concerns, that is the administration of the Security Layer is by design separated from the administration within single security domains (for example, an isolated HPE Private Cloud AI deployment). This implies that Security Layer administrators can assign nodes to security domains and initiate the provisioning of cloud software, but do not have access to classified data.

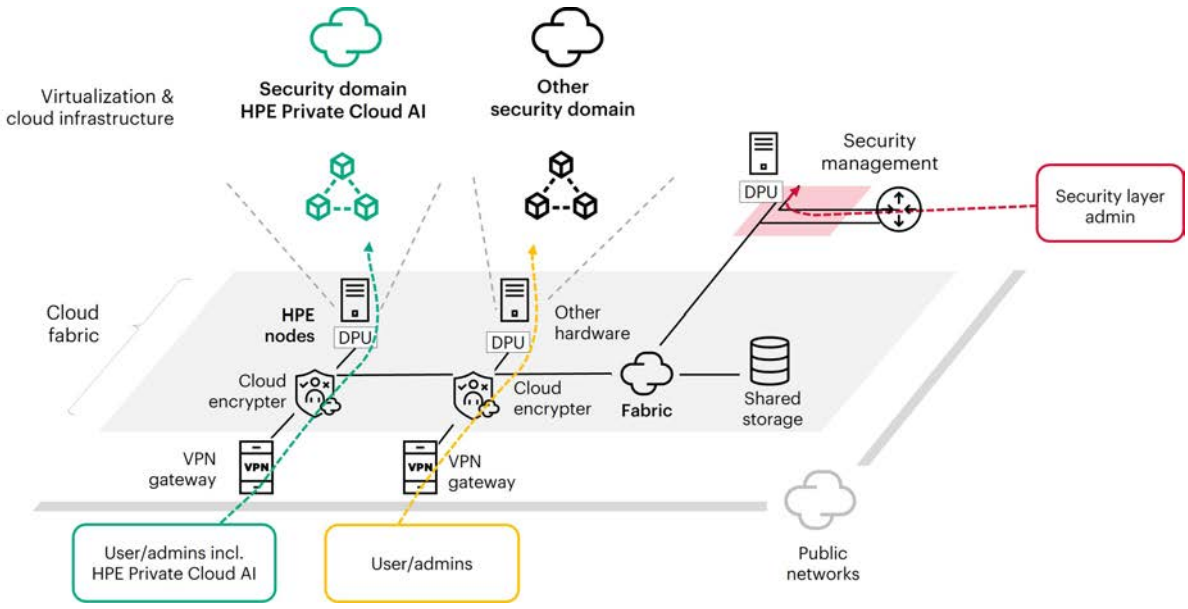


Figure 3. High-level architecture of SINA Cloud with two security domains

[5] Allgemeine Verwaltungsvorschrift zum materiellen Geheimschutz (Verschlusssachenanweisung—VSA) | BSI

2. Simplify AI deployments with HPE Private Cloud AI

HPE Private Cloud AI,[6] co-engineered with NVIDIA, is a fully integrated, turnkey private cloud platform designed to simplify the deployment and operation of AI at enterprise scale. It brings together optimized infrastructure, NVIDIA AI software, lifecycle management, and enterprise-grade security into a single, prevalidated solution, dramatically reducing integration complexity and accelerating production readiness.

Enterprise AI initiatives often stall due to fragmented infrastructure, complex security requirements, and manual integration across hardware and software stacks. The traditional AI deployment journey often feels like assembling a complex puzzle with pieces that don't quite fit together. Teams spend countless hours on infrastructure configuration, security setup, and tool integration before they can even begin their actual AI work.

HPE Private Cloud AI removes these barriers by providing a standardized, factory-integrated foundation that enables organizations to focus on building and scaling AI applications rather than assembling the underlying platform. Key capabilities include:

— **Out-of-the-box readiness**: Preintegrated hardware, AI software, models, and tooling enable rapid deployment without custom engineering.

— **Prevalidated architecture**: Jointly tested and certified configurations from HPE and NVIDIA deliver predictable performance and reliability for demanding AI workloads right out of the box.

— **Cost-efficient operations**: Optimized resource utilization and economics help reduce total cost of ownership compared to public cloud or do-it-yourself (DIY) on prem approaches.

— **Built-in governance and security**: Recognizing the importance of safeguarding sensitive information, native role-based access control, data and model governance, and zero trust security controls supports enterprise and regulated environments.

— **Integrated AI tooling**: Support for NVIDIA AI Enterprise, Jupyter environments, application programming interfaces (APIs), and an extensible collection of frameworks enables collaboration across IT, data engineering, and data science teams.

[6] HPE Private Cloud AI—Deploy AI in days, not months, with a turnkey private cloud platform.

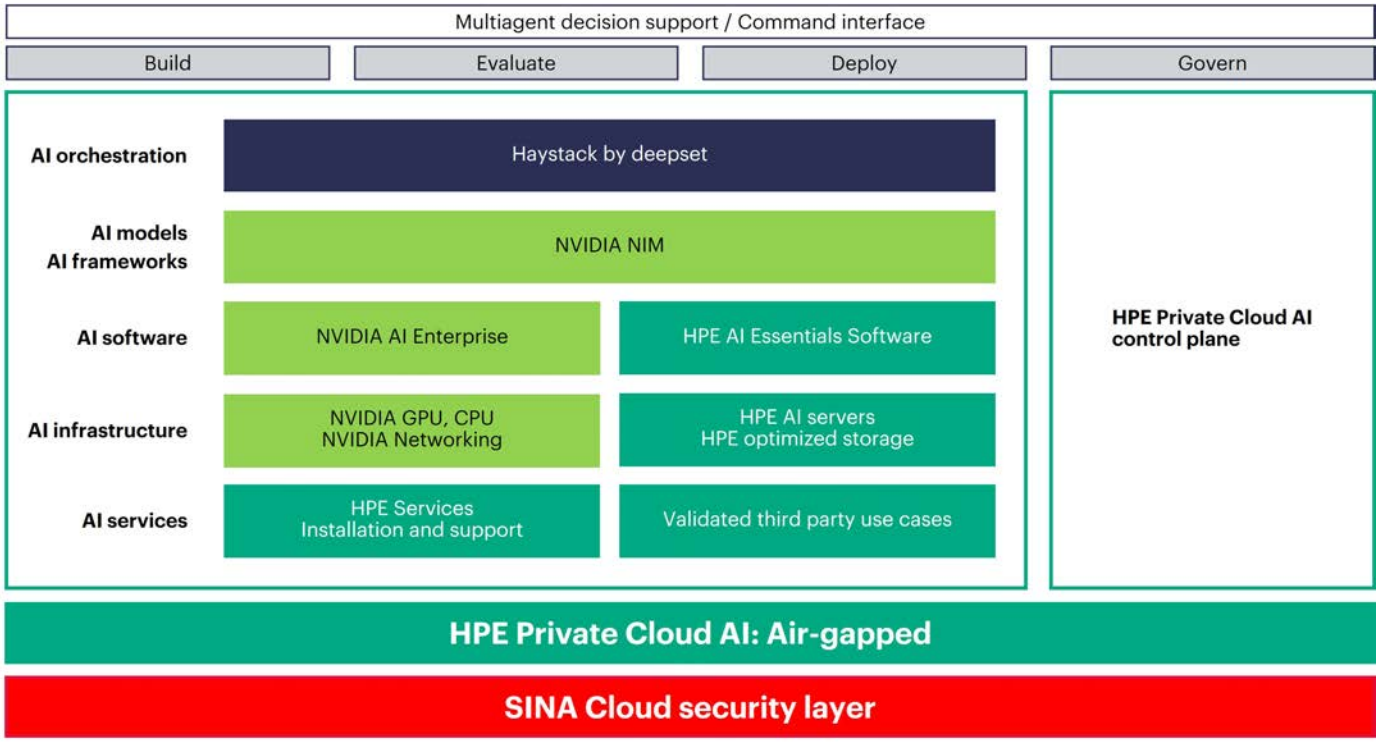


Figure 4. HPE Private Cloud AI, a turnkey AI factory

Together, these capabilities provide a consistent, enterprise ready platform for deploying and scaling AI while maintaining full control over data, models, cost, and compliance.

HPE Private Cloud AI disconnected

For classified, public sector, and highly regulated environments, HPE Private Cloud AI is available in a fully air gapped configuration, enabling advanced AI capabilities without external network dependencies.[7] Key characteristics include:

— **Security and compliance by design** supports private model customization and inference within isolated environments, meeting sovereignty and regulatory requirements.

— **Digital sovereignty and operational resilience**, fully disconnected operation, ensures availability and control in high security or remote deployments.

— **Consistent AI experience** delivers the same AI software stack, tooling, and workflows as standard HPE Private Cloud AI, optimized for classified use cases.

— **End-to-end lifecycle support by HPE** provides deployment, operations, and lifecycle services adapted for air gapped and sovereign environments.

— **Trusted supply chain** helps ensure hardware, software, and documentation are assembled in TAA compliant facilities.

This offering enables organizations to deploy mission critical AI with confidence—combining performance, security, and operational autonomy in a single, turnkey platform.

[7] HPE Private Cloud Air-Gapped—fully isolated, on premises cloud solutions designed specifically for organizations with the highest security, compliance, and sovereignty requirements.

3. Accelerating the deployment of AI models with NVIDIA NIM

In addition to supplying the necessary computational power to the infrastructure/cloud layer of the reference architecture, NVIDIA also provides a standardized, high-performance inference layer specifically engineered to operate within the most secure, air-gapped environments. NVIDIA NIM microservices provide accelerated, optimized inference containers of state-of-the-art AI models, covering all modalities including the latest LLMs. The containers automate a variety of techniques to deploy and serve various AI models in an optimized fashion on the available graphics processing unit (GPU) infrastructure. The deployed models are exposed through standard APIs that any AI engineer is familiar with developing AI applications.

NVIDIA NIM allows enterprise teams to deploy state-of-the-art open-source LLMs and other pretrained AI models as containers within a private cloud or on-premises environment, ensuring sensitive data never leaves secure infrastructure:

— **No external dependencies**: Once models are pulled into the secure trust domain, inference operates independently of external connections or services.

— **Local processing**: All computations are performed locally, maintaining the integrity of the air-gapped environment.

— **Optimized runtimes**: NIM integrates TensorRT-LLM to deliver accelerated, hardware-optimized performance tailored for high-tempo operational intelligence.

Unlike traditional DIY deployments, which often result in complex, fragmented architectures that take weeks to secure and tune, NIM offers a standardized, turnkey approach. While DIY methods force teams to manually integrate and validate disparate components, NIM provides prevalidated, production-ready containers that lower the total cost of ownership. This eliminates the risk of untested configurations and allows organizations to leverage day 0 support for the latest generative AI models without the operational overhead of building custom inference stacks from scratch.

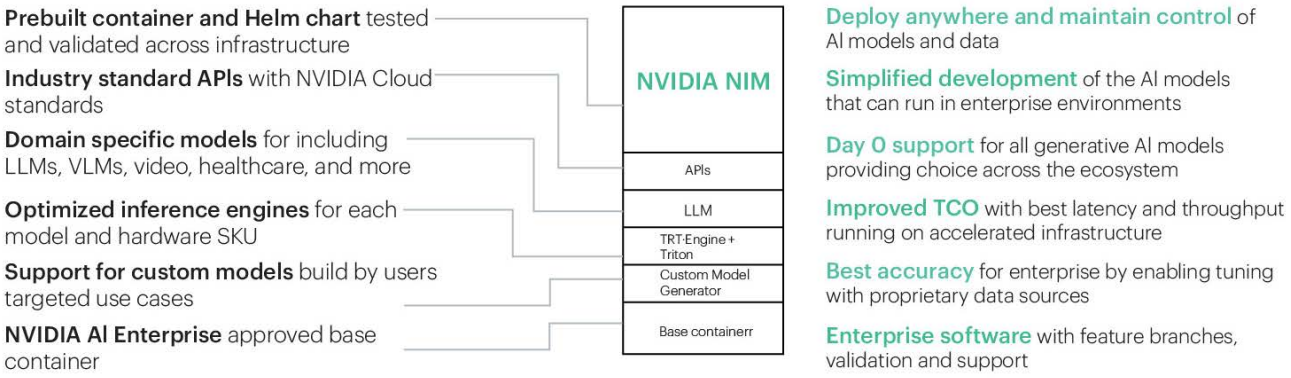


Figure 5. Component breakdown of a NIM

In domains where security must be certified to national standards, NIM provides a hardened application layer designed to mitigate both external and internal threats.[8, 9] Each model is cryptographically signed and audited for backdoors, enabling customers to verify model integrity before deployment. Containers undergo continuous scanning; no critical or high CVEs are permitted in published images. NIM containers utilize least-privilege principles and receive regular security-only patches to minimize attack vectors in classified networks.

[8] "Securely Deploy AI Models with NVIDIA NIM." NVIDIA Developer, June 2025.

[9] "NVIDIA AI Enterprise Security White Paper." NVIDIA, March 2026.

4. Build sovereign AI applications with Haystack by deepset

Haystack serves as the orchestration control plane for building, governing, and operating AI applications over classified data. The framework and enterprise platform manage the full AI application lifecycle, including knowledge ingestion, pipeline orchestration, and production governance, across air-gapped and on-premises deployments. Organizations including the European Commission, the German Ministry of Research, Technology, and Space (BMFTR), and Airbus Defence and Space rely on Haystack for AI orchestration within sovereign environments.

Each component of this reference architecture addresses a layer of the classified deployment challenge. Haystack addresses the application layer: how agents and RAG pipelines are built, governed, and operated over classified data.

The Haystack open-source framework

Sovereign AI orchestration enables resilient, modular pipelines. Haystack is an open-source AI orchestration framework designed for modularity developed by deepset.[10] Agentic, RAG, and other AI workflows are constructed as pipelines of interchangeable components (Figure 6): models, data sources, guardrails, tools, and techniques such as Model Context Protocol (MCP). Each pipeline defines how information is retrieved, processed, reasoned over, and acted on within a single auditable flow. Pipelines deploy instantly as REST APIs and can be updated without operational interruption.

Because every component is modular and open, pipelines remain resilient as models, tools, and threats evolve. Organizations can swap models, update embedding strategies, and integrate new data sources without rebuilding workflows. On HPE infrastructure within secunet secured environments, accelerated by NVIDIA GPUs and DPUs, this means classified AI systems stay current without compromising security posture.

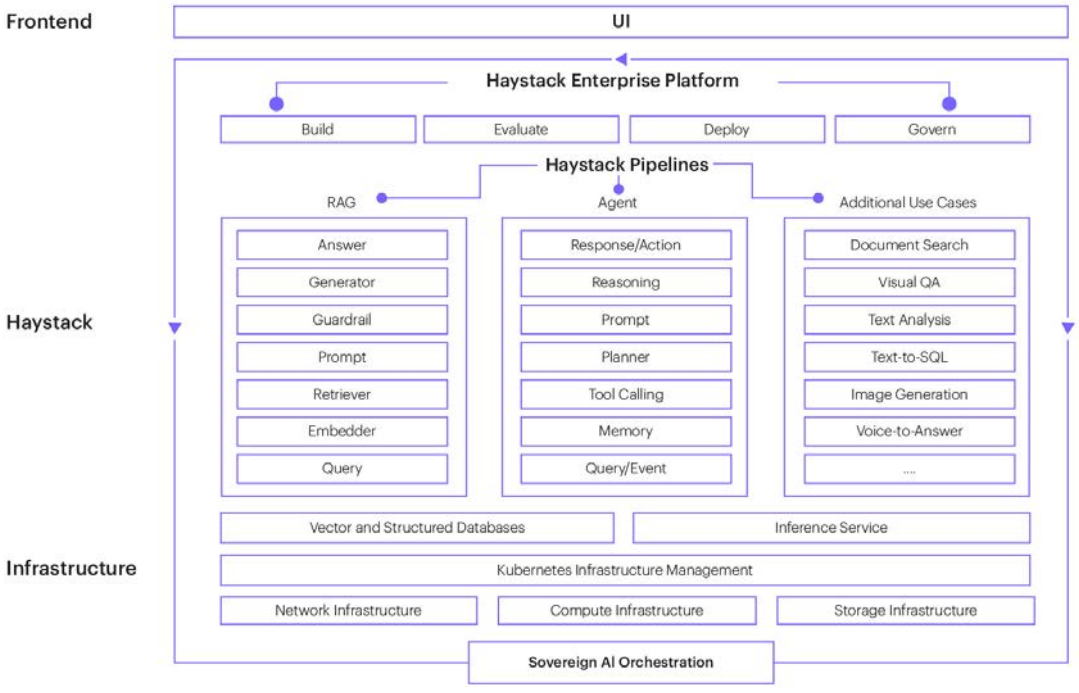


Figure 6. Haystack Enterprise Platform architecture showing pipeline components, orchestration layers, and lifecycle governance

[10] Deepset.ai

The Haystack Enterprise Platform. The Haystack Enterprise Platform extends the open-source framework into a production-grade control plane for managing AI applications across classified environments. This includes building and iterating on pipelines, evaluating their performance, deploying them into production, and ensuring governance aligned with internal policies and external regulatory frameworks such as the EU AI Act.

Enterprise knowledge infrastructure derisks classified AI operations. In classified environments, knowledge ingestion requires the same level of governance as the infrastructure it runs on. Decisions made on outdated or inconsistent knowledge carry operational consequences. The Haystack Enterprise Platform enforces this through governed ingestion pipelines, versioned knowledge indexes, and controlled context assembly. Knowledge is managed as independent infrastructure, decoupled from the applications that consume it, preventing drift and duplication across classified systems.

Mission-critical applications and APIs separate users from sensitive layers. Haystack pipelines surface as secure APIs and applications that classified end users interact with through dedicated interfaces. This separation ensures mission-critical users access trusted intelligence safely and reliably while models and data layers remain isolated and governed independently.

Sovereign control and compliance are enforced by design. The enterprise platform manages the full AI application lifecycle with built-in governance:

— **Dev/Test/Prod separation** with evaluation gating ensures no untested pipeline reaches production.

— **Role-based access control** down to the pipeline and component level governs who can build, modify, and deploy across the orchestration layer (deepset/Haystack), connected databases (MongoDB, Elastic), and inference services (NVIDIA NIM).

— **CI/CD integration and audit trails** provide traceability across every change in classified systems.

— **Self-hosted deployment** on Kubernetes infrastructure (NVIDIA NIM Operator, GPU Operator, Container Runtime, Enterprise Linux®) ensures classified data never leaves the operator's control.

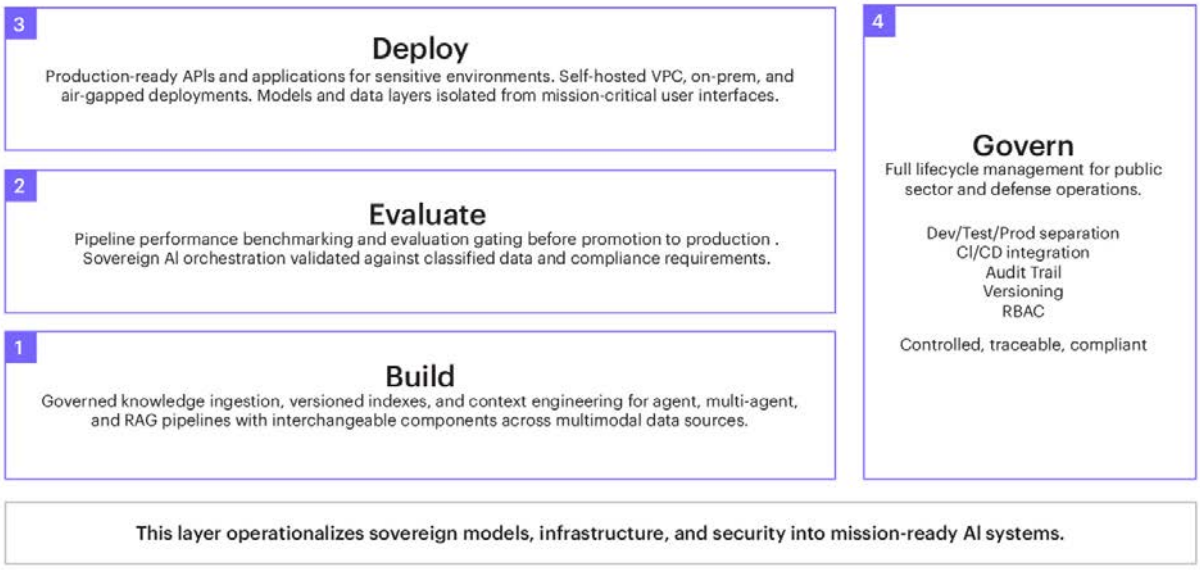
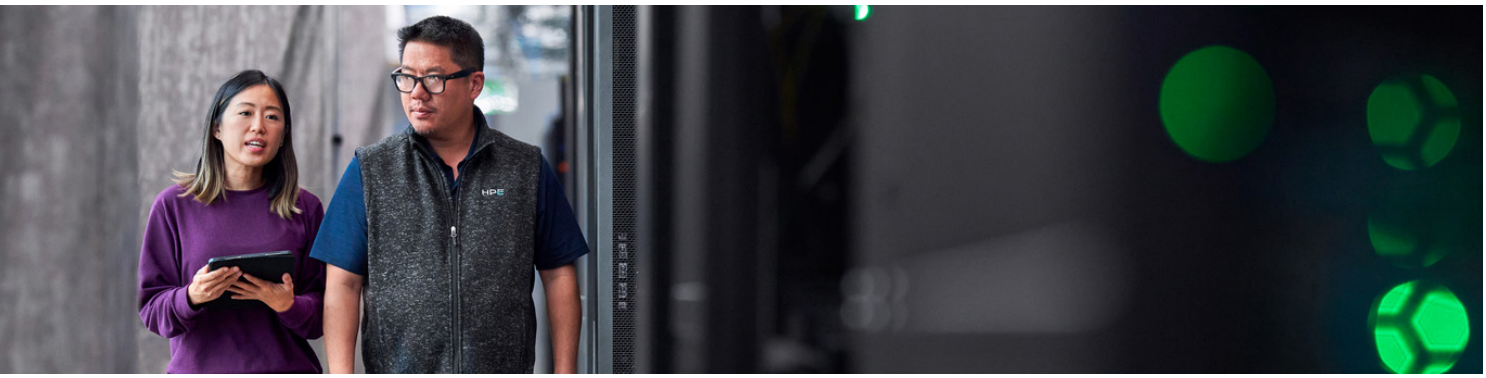


Figure 7. Functional layers of the Haystack Enterprise Platform by deepset.[11]

[11] Haystack sets the Standard for Agentic AI Across Industries | deepset.ai



Real-world application: The reference architecture in operation

This section demonstrates what the reference architecture delivers under classified requirements, including security certification, data sovereignty, and real-time performance. An emerging situation or escalating event use case illustrates how the four components of the architecture integrate to enable multiagent decision support. The pattern is applicable to a broader class of public sector and classified knowledge use cases, including classified document analysis, crisis response, and cross-domain planning.

Exemplary use case: Multiagent decision support in a classified environment

When an emerging situation or escalating event requires a rapid, informed response, organizations handling classified data must synthesize large volumes of information under time pressure: policy documents, regulatory frameworks, operational directives, real-time sensor feeds, status data, and situational reports. Today, this synthesis is largely manual and constrained by the cognitive limits of individual analysts.

Artificial Intelligence Decision Agents (AIDA) propose an alternative to this manual approach: a multiagent system that distributes the cognitive workload across specialized AI agents, each operating over classified data in parallel to support faster, more informed decision-making during emerging situations and escalating events.

This use case deploys three specialized agents, each running as an independent pipeline on the reference architecture, orchestrated hierarchically through a custom interface:

— **Interaction agent,** the primary interface for the analyst, responds to direct queries by reasoning across regulatory frameworks, situational parameters, and real-time data to deliver grounded, referenced answers as events evolve.

— **Assessment agent** operates continuously in the background. It monitors the current situation against the expected state by comparing live status data against directives and standing instructions. Assessment agent detects and escalates deviations to the interaction agent.

— **Observation agent** collects and merges situational data (environmental conditions, sensor inputs, and external feeds) as an additional source, operating continuously and informing the other two agents as emerging situations develop.

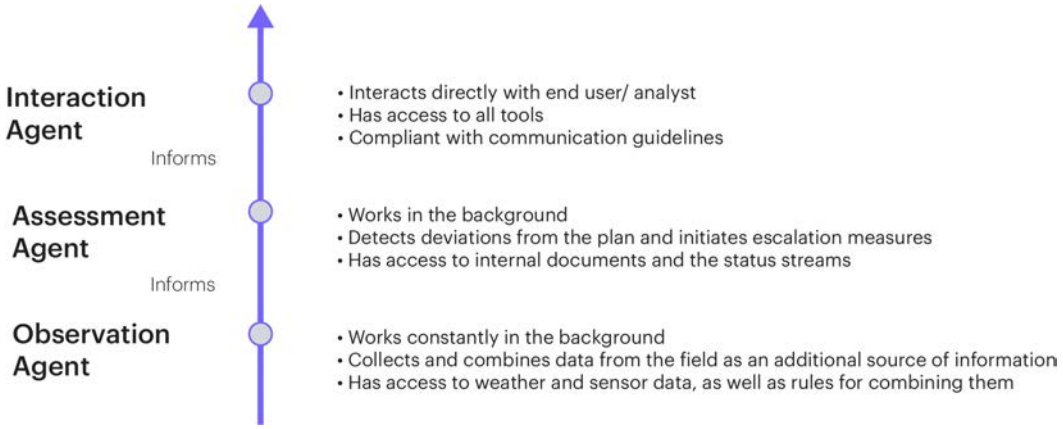


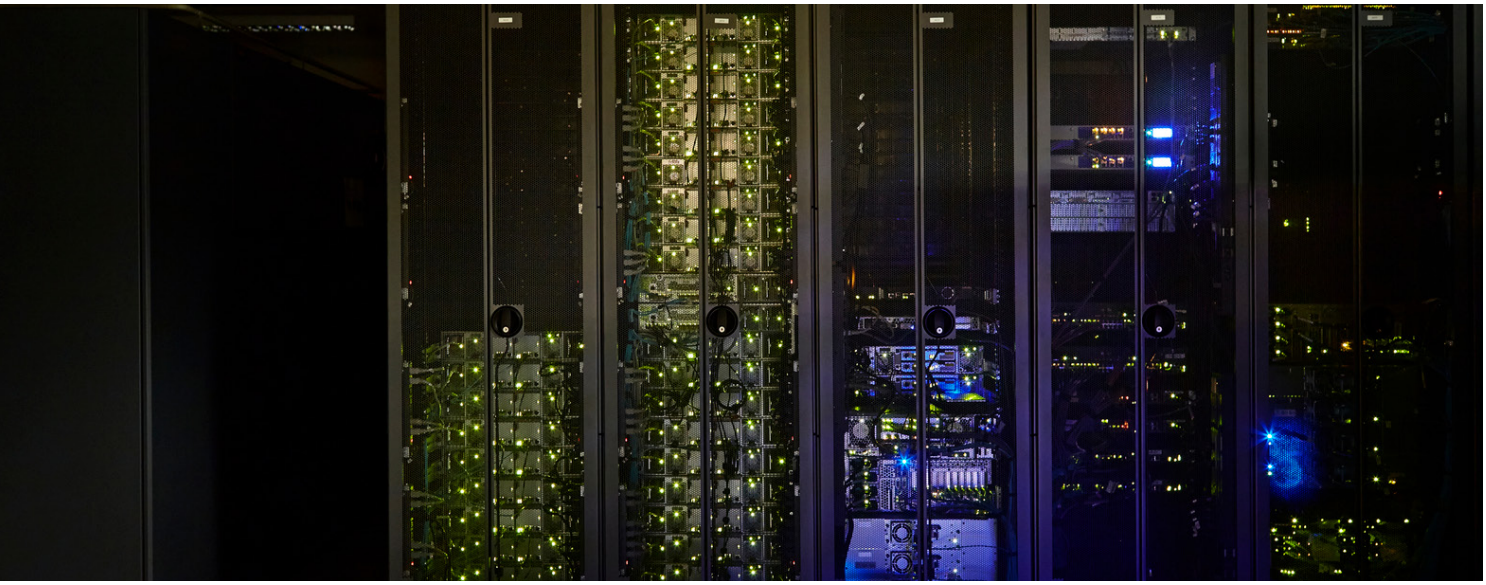
Figure 8. Multiagent hierarchy showing interaction, assessment, and observation agents and their data flows

How the four layers operate in Unison. This use case depends on all four layers operating together. The agents reason over classified documents, regulatory frameworks, and real-time sensor data. None of this data can leave the security domain. Using BSI-certified components of the secunet SINA ecosystem, strong cryptographic mechanisms ensure that classified traffic is secured both within the cloud fabric network and during transport through public networks.

The three agents run concurrently and require GPU-accelerated inference with low latency as situations escalate. HPE Private Cloud AI provides the air-gapped compute environment, prevalidated with NVIDIA hardware and operational without extended integration cycles.

Each agent runs foundational model inference for reasoning, classification, and structured output generation across many modalities (for example, video sensor streams, documents with pictures, tables and infographics, and databases). NVIDIA NIM packages the various models needed for that task as containerized microservices with optimized throughput on local GPU infrastructure, eliminating any dependency on external endpoints.

At the application layer, the three agent pipelines share knowledge indexes, expose uniform REST APIs for orchestration by the custom interface, log full pipeline trajectories (every tool call, intermediate results, and final output), and return structured data for integration into downstream systems and decision-making. deepset Haystack manages this orchestration, including agent state, context engineering, and observability.



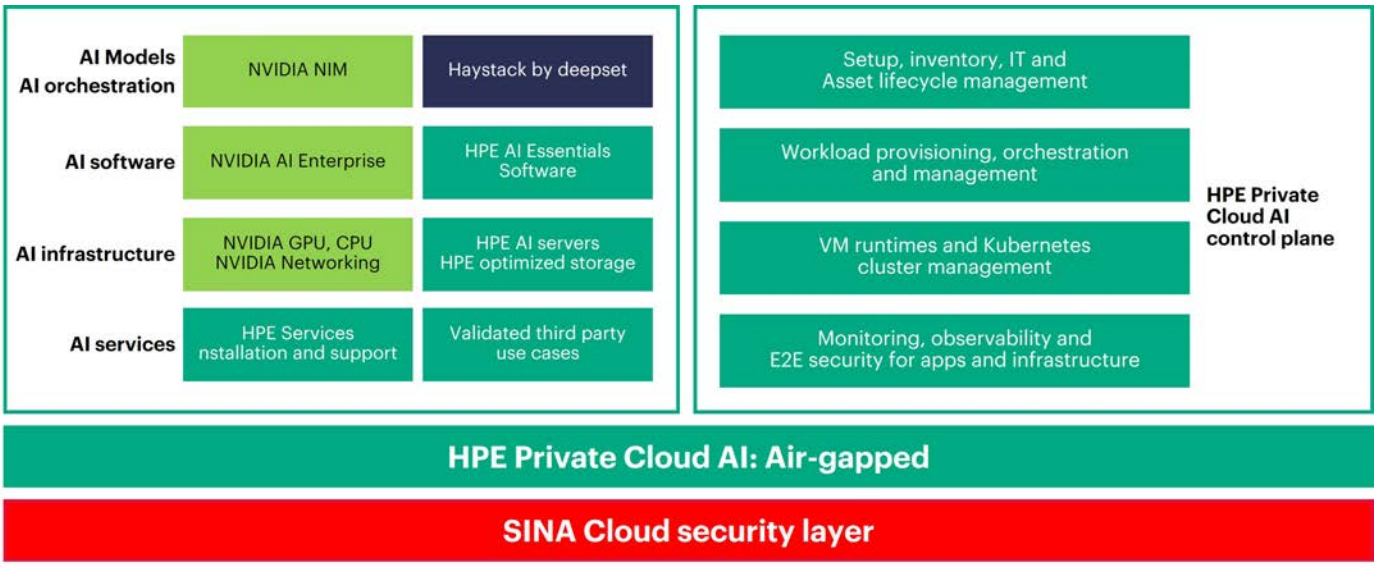


Figure 9. Reference architecture overview showing how the four partner components integrate across this use case

The architecture can support a wide range of secure operational environments. A few representative examples include:

— **Operational environment monitoring**: The multiagent system consolidates reports, sensor feeds, and status updates, highlights deviations from expected conditions, and provides operators with a real-time view of evolving situations.

— **Critical infrastructure monitoring**: Operators overseeing energy, transportation, or telecommunications systems use the architecture to detect anomalies across telemetry, maintenance reports, and incident alerts, supporting rapid response to potential disruptions.

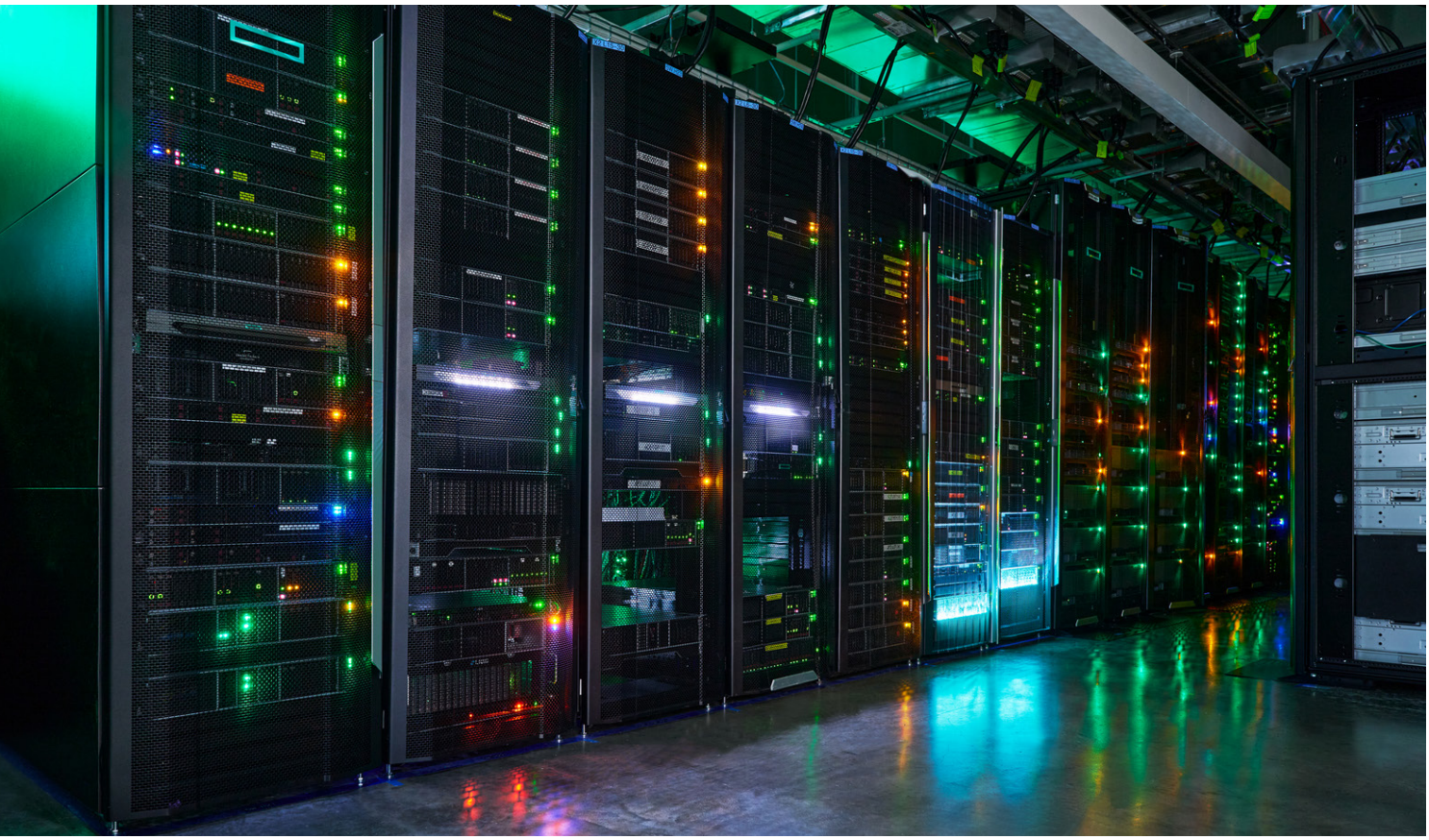
— **Cybersecurity incident response**: The architecture continuously analyzes system alerts, vulnerability reports, and network telemetry, identifying abnormal behavior and providing analysts with contextual explanations and recommended actions.

— **Crisis coordination**: AI agents track response progress, consolidate situational reports, logistics data, and environmental information, assisting decision-makers during rapidly evolving events.

— **Strategic supply chain risk monitoring**: The architecture analyses trade flows, production signals, and supply chain dependencies to identify emerging disruptions, assess risk exposure, and generate structured situational briefings.

Conclusion

The reference architecture presented in this paper addresses a persistent gap in AI adoption within classified environments: deploying production-grade agentic AI against classified data without compromising security certification, operational control, or speed of deployment. As requirements scale beyond single-agent applications toward multiagent systems operating across domains, the modularity, extensibility, and governance built into this architecture provide a foundation for sustained operational capability.



Visit HPE.com

Chat now

a00159136ENW

HEWLETT PACKARD ENTERPRISE

hpe.com