



Build governed, sovereign AI with IBM Fusion HCI and NeuralSeek

Move AI from pilot to production with a secure,
behind-the-firewall platform designed for regulated

Table of contents

- 1. Executive Summary 2**
- 2. Introduction 2**
 - 2.1 The Problem: AI Adoption in Regulated Industries2
 - 2.2 What NeuralSeek Provides3
 - 2.3 What IBM Fusion HCI Provides.....4
 - 2.4 Joint Value Proposition: NeuralSeek + IBM Fusion HCI5
- 3. Key Use Cases and Customer Benefits 5**
 - 3.1 Customer Benefits of IBM Fusion HCI + NeuralSeek5
 - 3.2 Use Cases Across Industries6
- 4. Architectural Overview and Diagram 8**
- 5. Guidelines and Sizing Recommendations 10**
- 6. Conclusion 11**

1. Executive Summary

Enterprises in regulated industries face a fundamental barrier to AI adoption: public-cloud LLMs and SaaS AI agents cannot guarantee data sovereignty, strict access control, or compliance with frameworks such as HIPAA, PCI-DSS, FINRA, GDPR and CJIS. Sending prompts or documents to an external endpoint risks exposure of sensitive information, while multi-tenant architectures introduce unacceptable uncertainty around data retention, cross-customer bleed and auditability.

NeuralSeek + IBM Fusion HCI solve this by delivering a fully governed, behind-the-firewall AI platform that turns private enterprise data into production-ready applications, without exposing any information to external networks. NeuralSeek provides deterministic, source-grounded responses, policy-driven orchestration, identity and governance; IBM Fusion HCI provides the resilient, sovereign hardware foundation. In effect, NeuralSeek is Docker for AI: the AI application framework, data connections and guardrails arrive pre-packaged as a governed base image, while LLMs remain pluggable and interchangeable. Teams customize only the last 10% and ship in weeks, running anywhere and swapping models underneath without a rebuild. Together they unlock AI productivity while preserving complete control over data, workflows and compliance.

2. Introduction

2.1 The Problem: AI Adoption in Regulated Industries

Boards in banking, insurance, healthcare, government, energy and other regulated sectors are under pressure to deliver AI productivity, yet the only readily available options are public, multi-tenant SaaS LLMs that were never designed for the controls these industries are legally required to enforce. The result is a widening gap: business units race ahead with shadow AI while risk, security and compliance teams must slow or block deployments that cannot prove what they did, with what data, on whose behalf. Specifically, regulated organizations struggle with:

Data sovereignty and zero-trust gaps — public-cloud LLMs route prompts, embeddings and documents through provider infrastructure, and the identity, RBAC and DLP that protect files and databases are bypassed the moment a user pastes content into an external chatbot.

Hallucinations and non-determinism — LLMs invent citations, names, dosages, dollar amounts and legal clauses; in regulated workflows a single fabricated fact can trigger fines, lawsuits or patient harm.

No audit trail or replay — most SaaS AI cannot reconstruct, on demand, the exact prompt, sources, model, parameters and response delivered to a specific user at a specific time — the baseline an examiner or auditor will ask for.

PII/PHI leakage and prompt injection — without inline detection and redaction, sensitive content flows into prompts, logs and caches, while external chat interfaces are increasingly weaponized to exfiltrate data or override instructions.

Vendor lock-in and layered compliance exposure — single-model bundles prevent swapping LLMs for cost or performance, while HIPAA, PCI-DSS, FINRA, SOX, GDPR, CJIS, ITAR, NYDFS, DORA and the EU AI Act each demand evidence that AI behavior is bounded, logged and reviewable.

Hardware and infrastructure gap — even when policy allows on-prem AI, most enterprises lack the GPU-optimized, OpenShift-ready hardware and tooling to run production LLM workloads behind their firewall.

The takeaway is straightforward: regulated industries do not need another general-purpose chatbot. They need an AI platform that operates inside their own security perimeter, on hardware they trust, with governance and evidence built in at every layer.

2.2 What NeuralSeek Provides

NeuralSeek is Docker for AI. Roughly 90% of every enterprise AI app is the same plumbing: the model, security and guardrails, and the work of connecting private data; NeuralSeek pre-packages that 90% into one governed, LLM-agnostic base image, so business and IT teams customize only the 10% that is uniquely theirs. The result is production-ready AI agents and workflows shipped in weeks rather than a year, without exposing data to external networks or fine-tuning models on customer content, and with the freedom to run anywhere and swap any model underneath. It is the same platform trusted today by regulated organizations across banking, healthcare, government, higher education and the Fortune 500.

- **RAG and multi-agent orchestration** — connects to knowledge bases, SharePoint, Confluence, Salesforce, Workday, ServiceNow, vector stores, databases and APIs, delivering source-grounded answers through drag-and-drop agent design; agents call other agents, route between models and execute multi-step workflows.
- **80+ pre-built agent templates** — the pre-packaged base image in practice: a marketplace of ready-to-run agents (RAG patterns, guardrail enforcers, OCR enhancement, SQL-to-answer, document, image and video generation) spanning Core Toolkit, Insurance, IT, Rev Ops, Sales, Marketing and Legal; each deploys inside the customer's tenant against their own data, under their governance plane — compressing time-to-value from months to days.
- **Multi-LLM routing and BYO-model** — orchestrates across IBM Granite, watsonx, OpenAI, Azure OpenAI, Anthropic, Google, AWS Bedrock, open-source and self-hosted models, routing per task for accuracy, latency, cost or data-residency with no rewrites and no lock-in.
- **Deterministic, source-grounded answers** — every response ties back to approved sources with verifiable citations and semantic confidence scoring; low-confidence outputs are suppressed, escalated or routed to a human.
- **Three layers of built-in governance** — AI Guardrails (real-time PII/PHI detection and redaction, prompt-injection defense, profanity filtering, attribution protection), LLM Governance (token-level cost tracking, model benchmarking, confidence thresholds and fully replayable logs) and Platform Governance (a unified dashboard for intent analytics, documentation coverage, access logs and system health).
- **Enterprise identity and access control** — SSO via SAML 2.0 plus native Okta, Active Directory and LDAP; directory groups map to roles with real-time, per-document permission checks before any content reaches an LLM.

- **Deployment flexibility** — runs as containers on Red Hat OpenShift or Kubernetes and is engineered to run fully air-gapped behind the firewall, the mode required for FedRAMP, HIPAA, CJIS, ITAR and EU-sovereign workloads.
- **Safe use of private data without retraining** — retrieves and reasons over private data in real time; update the source and the agent's answers update with it — no fine-tuning of external models, no data shared with sub-processors for training.

2.3 What IBM Fusion HCI Provides

IBM Fusion HCI is an integrated, turnkey hyper-converged infrastructure platform that brings compute, storage, networking, Red Hat OpenShift, and integrated data services together in a single, pre-validated stack. It provides an enterprise-grade, Kubernetes-native foundation for running containerized applications, virtual machines, and AI/ML workloads on-premises or across hybrid cloud environments through a unified management experience.

- **Accelerates AI from pilot to production** — provides a pre-integrated, validated foundation with automation and self-service, reducing the integration, testing, and provisioning work required to move AI workloads into production.
- **Global namespace for distributed enterprise data** — consolidates files, documents, and object storage across sites into a single, accessible layer, giving distributed enterprises consistent access to large volumes of unstructured content.
- **High-performance data access for AI workloads** — supports high-throughput, low-latency access to enterprise data for AI inference and other data-intensive workloads, enabling applications to work against data in place.
- **Supports sovereignty and operational control** — helps organizations retain authority over where data resides, where workloads run, and how operating environments are managed, so applications and AI workloads can run where business, regulatory, security, and data-residency requirements demand.
- **Built-in data security and governance** — combines encryption, role-based access controls, and governance capabilities to help keep enterprise content protected, compliant, and auditable.
- **Resilience as a built-in data service** — integrates high availability, application-consistent backup, disaster recovery, and recovery capabilities to help maintain service continuity and protect enterprise data even if a server, rack, or entire data center fails.
- **Supports incremental modernization** — runs containerized and virtualized workloads side by side on a common Red Hat OpenShift-based foundation, allowing enterprises to modernize at their own pace rather than through a forced rip-and-replace.

For NeuralSeek, this means RAG and multi-agent workflows can reason over governed enterprise data in place, while IBM Fusion HCI provides the performance, data services, security, resilience, sovereignty, and operational control required to run those workloads in production.

2.4 Joint Value Proposition: NeuralSeek + IBM Fusion HCI

Together, NeuralSeek and IBM Fusion HCI deliver one of the only architectures capable of meeting the full stack of requirements regulated industries demand. NeuralSeek solves the correctness, governance and developer-velocity problem; IBM Fusion HCI provides the infrastructure, data sovereignty, resilience, and data services required to run enterprise AI in production. The combination is greater than the sum of its parts. NeuralSeek accelerates that developer velocity through a no-code environment where teams build and deploy generative AI applications without writing code or managing infrastructure. Its drag-and-drop pipeline approach lets users assemble retrieval, generation and automation flows visually, turning what would be complex engineering work into configurable building blocks.

While NeuralSeek governs AI behavior at the content and model layer through guardrails, deterministic citations, and replayable audit trails, IBM Fusion HCI anchors that governance to resilient, behind-the-firewall infrastructure. The two combine into a single control plane that spans token cost, model behavior, and hardware health, with NeuralSeek enforcing content-level compliance and IBM Fusion HCI enforcing system-level compliance across HIPAA, PCI-DSS, FINRA, SOX, GDPR, CJIS, ITAR, NYDFS, DORA and the EU AI Act. Every prompt, retrieval, embedding, log and response stays inside the customer's data center, with no external API calls and no retention or training-reuse, and the architecture is auditable down to the hardware.

3. Key Use Cases and Customer Benefits

3.1 Customer Benefits of IBM Fusion HCI + NeuralSeek

By running your AI orchestration layer directly on your secure, governed infrastructure, you eliminate the trade-off between enterprise AI and data control. IBM Fusion HCI customers realize transformational benefits across speed, trust, governance, and cost.

Production AI Without the Year-Long Build: Governed agents and day-one applications ship in weeks on a turnkey platform, not months of custom build and a separate infrastructure project.

Grounded & Trustworthy AI: Every answer is grounded in approved enterprise knowledge with verifiable citations and real-time hallucination detection, so decisions rest on facts, not a copilot's confident guesses.

Governed, Sovereign, and Auditable: Guardrails, audit trails, and encryption are built in from day one, keeping sensitive data inside the customer's environment and satisfying what IT and legal require before approving AI.

Freedom from Lock-In and Predictable Cost: LLM-agnostic routing lets teams swap models without re-platforming, and one flat subscription on consolidated infrastructure avoids per-task billing surprises and cuts hardware cost.

One Platform Instead of Point Tools: A single governed platform replaces scattered notetakers, chatbots, and search tools, giving employees a sanctioned alternative to shadow AI under one vendor and one management plane.

Always Available, Always Protected: Built-in high availability and disaster recovery keep AI workflows running without interruption.

3.2 Use Cases Across Industries

NeuralSeek + IBM Fusion HCI is in production across the most regulated and complex sectors of the global economy. The use cases below reflect the workflows the joint stack is uniquely positioned to deliver behind the firewall: grounded, governed AI running inside the customer's own environment with role-based access, full audit logging and content filtering.

Healthcare and Life Sciences

- **HIPAA-aware clinician and staff chatbots:** ChatGPT-style assistants running inside the hospital tenant with SSO, RBAC scoped to each user's role and unit, PHI-aware content filtering, and a full audit trail on every prompt and response.
- **Clinical policy and pathway chat with cited answers:** pre-loaded with clinical pathways, medication protocols (including weight-based dosing), infection-prevention standards and regulatory guidance; clinicians get direct citations to the source policy, auto-refreshed when policy is republished. OCR brings outside records, faxed referrals, handwritten intake forms and legacy scanned charts into chat for inline summarization and Q&A.
- **Self-service cohort analytics:** quality, operations and research teams ask plain-English questions of governed datasets (readmissions, length of stay, throughput, wait times, outcomes by cohort) with the underlying query and source data surfaced for audit.
- **Pharmaceutical R&D and regulatory drafting:** grounded summarization of trial protocols, IND/NDA submissions and medical literature with verifiable citations, critical for FDA and EMA submissions.

Banking, Financial Services and Investment Research

- **Public-facing AI banking assistants under GDPR:** RAG, AI Guardrails, replay and granular governance deliver conversational banking across web, mobile and online channels with full GDPR-grade traceability. On-prem AI also powers internal knowledge search across processes and policies, with audit logging suitable for central-bank scrutiny.
- **Hybrid-RAG financial intelligence:** a unified retrieval layer blending SEC EDGAR filings, live web research and private documents (BM25 plus dense vectors), orchestrated through graphs of specialized agents, with XBRL financial dashboards, cited multi-source analyst chat and deterministic PDF/PPTX report generation.
- **Multi-channel customer service for pensions and payments:** consistent, policy-compliant answers across IVR, WhatsApp, web and internal support, handling thousands of concurrent interactions while keeping regulators satisfied.
- **AML, fraud and KYC support agents:** policy-aligned agents route sensitive inquiries through OCC, FINRA, FDIC and FFIEC-aware guardrails with full decision auditability.

Insurance

- **Underwriting and claims acceleration:** LLM-enhanced OCR ingests scanned policies, claims forms and medical records, extracts entities with contextual meaning, and routes them through governed downstream workflows.
- **Agent-and-broker enablement:** grounded answers to product, eligibility and regulatory questions, sourced from approved internal knowledge with full audit trail.

Government and Public Sector

- **Air-gapped, sovereign AI for federal, state and local agencies:** NeuralSeek runs fully on-premises on IBM Fusion HCI to satisfy FedRAMP, GovCloud, CJIS and ITAR controls, with no external dependency on public LLM providers.
- **Citizen services and benefits navigation:** grounded, policy-compliant assistants help residents find programs, complete forms and understand eligibility, with PII redaction enforced inline.
- **Defense and intelligence knowledge augmentation:** multi-model agents reason over classified knowledge bases inside isolated enclaves, with full prompt/response replay for oversight.

Higher Education

- **University-scale digital concierges:** personalized access to academic, health, wellness and financial-aid resources, grounded in institutional data and protected by harmful-language and sensitive-data filtering.
- **Research administration and grants:** grounded assistants for research compliance, IRB workflows and grant documentation, with citations to institutional policy.

Enterprise HR, IT and Internal Operations

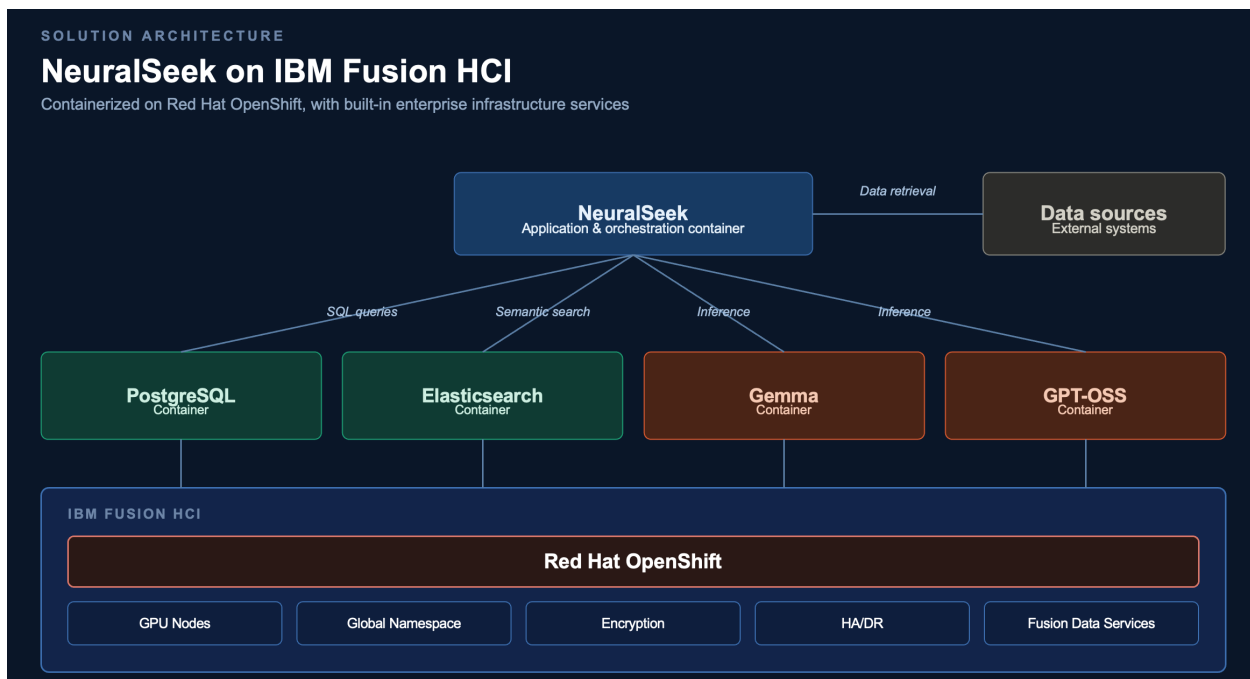
- **AI-powered HR assistants:** centralized HR bots with strict guardrails, contextual awareness by country/city, and workflow integration (Workday, Zendesk, onboarding nudges) for globally distributed workforces.
- **Automated HR inbox triage and Agentic-AI Center of Excellence:** triage HR inbox traffic, unify knowledge across emails/messages/documents, and run a hub-and-spoke CoE where business users build agents from pre-packaged templates while IT governs centrally.
- **Multi-LLM employee chat with operational grounding:** a secure, authenticated, multi-LLM chat experience spanning HR policy Q&A and call-center or line-of-business workflows in a single governed platform inside the customer's own cloud tenant.
- **IT helpdesk, ITSM and developer enablement:** grounded answers over Confluence, SharePoint, ServiceNow and code repositories with real-time, per-document permissions enforcement.

Cross-Industry Production Patterns

- **AI-powered PDF research and reports:** parallel retrieval from internal knowledge and approved external sources synthesized into citation-rich PDFs. SQL-to-PowerPoint automation extends this: NeuralSeek queries the warehouse, drafts charts and slides, and assembles a complete deck with source attribution on every claim.
- **Client-ready communication and briefings:** Slack-triggered email and message drafting with confidence checks and human-in-the-loop escalation, plus daily meeting and account briefings generated from calendar, CRM and approved internal sources, tailored per executive and replayable on audit.
- **Secure call notetaking:** meetings captured, transcribed and summarized entirely behind the firewall, with PII/PHI redacted in real time, grounded and cited summaries, a full audit trail, and no audio or transcript sent to an external provider.
- **Air-tight virtual assistants behind the firewall:** self-hosted AI-agent factories built on NeuralSeek plus open-source LLMs running on IBM Fusion HCI, delivering a ChatGPT-like experience without the data-leakage risk.

4. Architectural Overview and Diagram

4.1 Reference architecture diagram



4.2 Core components

- **NeuralSeek runtime** — Seek, mAIstro, KnowledgeBase, NeuralEdit, Run Agents and Governance, deployed as containers on Red Hat OpenShift.
- **Pre-assembled stack** — GPT-OSS 120B, PostgreSQL with pgvector, Elasticsearch and Red Hat OpenShift, pre-integrated and ready for AI workloads.
- **IBM Fusion HCI** — provides the underlying data fabric and operational services that support NeuralSeek workloads at scale. Its global namespace, integrated object and file storage, encryption, backup, disaster recovery, and lifecycle management services ensure that enterprise knowledge repositories remain continuously available, protected, and accessible to AI workflows. By exposing data through Kubernetes-native services on OpenShift, IBM Fusion HCI enables NeuralSeek to retrieve, process, and govern enterprise content without requiring additional storage integration layers.

4.3 End-to-end workflow

A request enters through a governed channel, is checked against the user's permissions, retrieved against the KnowledgeBase using hybrid search, routed to the chosen model and filtered through guardrails before being returned with citations, with the full prompt, sources, model and response logged for replay.

Throughout the workflow, IBM Fusion HCI provides the infrastructure services that enable reliable and secure execution. Enterprise content is stored and managed within IBM Fusion HCI's data services and global namespace, allowing NeuralSeek to access governed data sources with low-latency retrieval. As prompts, embeddings, retrieval operations, and model inferences are processed, IBM Fusion HCI supplies the compute, GPU acceleration, storage performance, encryption, and high-availability capabilities required to maintain consistent service levels. All artifacts, logs, and audit records remain within the customer's environment, ensuring data sovereignty, resilience, and compliance from ingestion through response delivery.

5. Guidelines and Sizing Recommendations

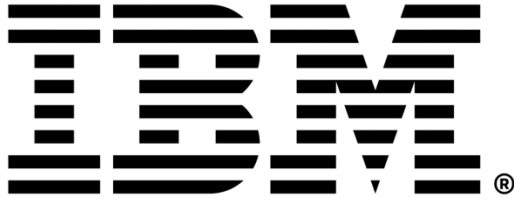
IBM Fusion HCI Resources	Small evaluation (non-HA)	Small production	Medium production	Large production
Control Nodes (2 x 16-core Intel Processors, 256 GB RAM, 2 x 7.68 TB NVMe drives)	3 (can be virtual control nodes)	3 (can be virtual control nodes)	3	3
Compute-Only Nodes (2 x 16-core Intel Processors, 256 GB RAM)	1	1	2 (HA)	3 (HA)
Optional GPU Nodes (2-8 RTX PRO 6000 per node, 256 vCPUs, 1536 GB RAM)	0	1 (1x RTX PRO)	1 (2x RTX PRO)	2 (4 x RTX PRO)

NeuralSeek does not require GPUs. If we want to add an LLM to the IBM Fusion HCI appliance, the LLM will require GPUs.

6. Conclusion

NeuralSeek + IBM Fusion HCI is one of the few architectures capable of meeting the full stack of requirements demanded by regulated industries: security, governance, accuracy, auditability and operational control. The two halves are complementary by design. IBM Fusion HCI consolidates an enterprise's unstructured content into a single governed namespace and keeps it protected, sovereign, and highly available on-premises or in hybrid cloud. NeuralSeek reasons over that data in place, turning it into source-grounded answers and multi-agent workflows without moving it off-platform or fine-tuning models on it. NeuralSeek solves the correctness and explainability problem; IBM Fusion HCI solves the data-sovereignty and infrastructure problem.

The governance reinforces itself at both layers: IBM Fusion HCI enforces encryption and role-based access to the content, while NeuralSeek enforces PII and PHI redaction, verifiable citations, and confidence thresholds on every response. When the underlying data changes, the answers change with it, so the system stays current without retraining. Together, they enable enterprises to move from prototypes to production-grade AI systems that regulators, auditors and risk committees can trust, packaged to run anywhere and to swap any model underneath, so adoption is fast without locking the enterprise to a single vendor or model.



IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
September 2026

IBM, the IBM logo, and the names of IBM products and services referenced herein are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries. A current list of IBM trademarks is available at ibm.com/trademark.

Red Hat and OpenShift are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

NeuralSeek and the NeuralSeek logo are trademarks or registered trademarks of their respective owner.

Other product and service names might be trademarks of IBM or other companies. References in this document to third-party products or services do not imply endorsement or affiliation.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.