IBM Fusion + Atolio: Sovereign AI Context Layer

Deliver permission and context-aware enterprise knowledge retrieval on-premises with trusted AI access, governed data, and production-ready infrastructure.

June 2026

Table of Contents

Executive Summary

AI Cannot Scale Without Trusted Enterprise Knowledge

Enterprises are racing to apply AI to their most important asset: internal knowledge. But that knowledge is scattered across email, chat, documents, file repositories, ticketing systems, CRM platforms, knowledge bases, workflow tools, and cloud repositories. The business may already have the knowledge it needs, but employees and AI systems cannot reliably find, trust, or use it.

This creates the core barrier to AI value. Generative AI is only as useful as the context it can access. When that context is fragmented, stale, duplicated, poorly ranked, or permission-blind, AI produces incomplete answers, exposes sensitive information, and fails to earn user trust.

Current approaches do not fully solve the problem. Point-to-point agents, MCP tool calls, and live API retrieval can connect systems, but they often push raw, unranked information into models at runtime, increasing latency, token cost, and governance risk. Centralizing everything into a data lake creates a different issue: slow programs, stale repositories, and knowledge disconnected from daily workflows.

The result is shadow AI. When employees cannot find trusted answers internally, they take matters into their own hands, often copying sensitive information into public tools outside enterprise control.

Enterprise AI needs a governed knowledge layer that securely unifies information where it already lives, preserves permissions, keeps knowledge current, ranks results by relevance, understands collaboration context, and gives AI systems trusted retrieval. That is how organizations move from AI experimentation to faster decisions, safer adoption, and measurable business value.

The Solution: IBM Fusion HCI + Atolio

The joint solution combines IBM Fusion HCI and Atolio to deliver an intelligent, secure, and fully on-premises enterprise knowledge layer. Atolio provides the permission-aware retrieval engine that securely indexes scattered workplace data in real-time, while IBM Fusion HCI provides the hyper-converged, GPU-accelerated infrastructure that allows companies to run the platform entirely within their own secure compute boundary.

By bringing the AI application directly to where the data lives, this joint offering serves as a Sovereign AI Knowledge Layer. It fulfills the promise of greater data sovereignty, strict IT governance, and production-grade AI outcomes, allowing companies to safely scale AI without requiring sensitive data to leave their perimeter.

Introduction

The Business Problem: Why Traditional Approaches Fail to Scale

The challenge is not simply that enterprise knowledge is fragmented. The larger problem is that most AI initiatives hit an architectural wall when trying to access that knowledge safely, accurately, and cost-effectively.

Critical business context lives across email, chat, documents, file repositories, ticketing systems, customer records, knowledge bases, code repositories, and workflow tools. When that context remains locked in separate systems, employees waste time searching, teams make decisions from incomplete or outdated information, and institutional expertise is lost when experienced people leave.

Traditional approaches usually follow one of two paths: connect AI directly to every system at runtime or move everything into a centralized repository. Both approaches create new problems.

- **Runtime access creates inconsistent answers.**
  Live queries across multiple business systems may connect AI to more information, but every system has different search quality, ranking, metadata, freshness, and permission logic. AI is forced to process raw, uneven, and often unranked results at the moment of the request. This increases latency as AI must query multiple systems and process more content and makes answers less reliable when it struggles to determine which source is most relevant, trusted, or current.
- **AI costs become difficult to control.**
  When AI systems retrieve too much irrelevant, outdated or duplicated content, they push unnecessary context into the model. That drives up token usage, compute cost, and response time. Without a relevance layer that filters and ranks information before it reaches the model, enterprises pay more for lower-quality answers.
- **Security and governance become harder to enforce.**
  Direct connections into many systems make it difficult for IT and security teams to centrally govern access across enterprise knowledge sources, enforce permissions, monitor retrieval, and audit usage. Broad service accounts, fragile access connections, and after-the-fact permission checks can expose sensitive information to the wrong users or agents.
- **Centralizing everything is too slow and quickly becomes stale.**
  What enterprises need is a knowledge layer that resolves all four at once, without requiring them to choose between capability and control. Moving enterprise knowledge into a central data lake or repository requires teams to extract, clean, normalize, and maintain massive

volumes of constantly changing information. This is expensive, slow, and disconnected from how work actually happens. Chats, tickets, incidents, documents, customer updates, and project decisions change by the hour. A copied repository quickly falls behind.

What is Atolio?

Atolio is a self-hosted, permission-aware enterprise knowledge layer that helps employees and AI safely retrieve and trust information across the business. It is the context layer at the heart of the Sovereign AI Knowledge Layer that this brief describes.

Running entirely inside the customer's controlled environment (private cloud, VPC, on-premises, or platforms like Red Hat OpenShift), Atolio makes internal knowledge AI-ready while guaranteeing absolute data sovereignty, security, and operational governance.

By connecting fragmented SaaS and legacy data—such as documents, chats, emails, and tickets—Atolio bridges enterprise silos without requiring disruptive data migrations. Crucially, it strictly enforces source-level permissions (ACLs), ensuring users and AI models only retrieve information they are authorized to access.

What makes Atolio different is its focus on trusted relevance. Rather than simple keyword matching, it uses permission-aware indexing and rich collaboration context (projects, teams, and experts) to surface the knowledge that matters most.

Atolio helps organizations:

- **Deploy Securely Anywhere:** Run retrieval inside your own infrastructure to guarantee sensitive corporate knowledge never feeds external models.
- **Unify & Govern Access:** Replace fragile point-to-point AI connections with a single, auditable retrieval layer for accessing authorized knowledge across connected enterprise systems.
- **Mitigate Shadow AI Risk:** Give employees a sanctioned, highly accurate way to use enterprise knowledge without copying sensitive information into unauthorized public AI tools.
- **Elevate Answer Quality:** Deliver personalized answers driven by deep relevance ranking and collaboration signals.
- **Reduce Costs**: Minimize AI costs by ensuring models only read the exact information needed to answer a question, rather than scanning everything. Saves money by using more affordable models for simple tasks and remembering past answers to avoid repeating work.

Atolio turns scattered data into a secure, AI-ready intelligence layer that employees can use daily, AI systems can safely rely on, and executives can trust to protect company IP.

What Is IBM Fusion HCI?

IBM Fusion HCI is an integrated turnkey hyper-converged infrastructure platform that combines compute, storage, and networking into a single, pre-validated stack. Built on Red Hat OpenShift, it provides a Kubernetes-native foundation for running containerized applications, virtual machines, and AI/ML workloads on-premises or in hybrid cloud environments. Fusion HCI is managed through a unified management experience, reducing the operational complexity that typically comes with managing separate infrastructure layers. It is purpose-built for enterprises that require consistent performance, security, and scalability without custom integration work between components.

Fusion HCI's data management capabilities are particularly relevant for applications that need to access and act on large volumes of unstructured enterprise content. Its global namespace consolidates files, documents, and object storage across locations into a single, accessible layer. The platform supports high-throughput, low-latency data access tuned for AI inferencing workloads, and includes built-in high availability, disaster recovery, and application-consistent backup capabilities as standard data services. Built-in data governance, encryption, and role-based access controls ensure that enterprise content is handled in a compliant and auditable way.

Fusion HCI supports both containerized and virtualized workloads, enabling enterprises to modernize at their own pace. It also exposes open APIs that allow third-party applications to connect directly to the data and compute resources it manages, enabling external tools to index and retrieve content stored on the platform without additional data movement or custom middleware. For Atolio, this means the intelligence layer runs where the data already lives, with the performance, governance, and resilience that enterprise-wide knowledge retrieval demands.

How IBM Fusion HCI Work Together and Atolio

Atolio provides the intelligence and retrieval layer. IBM Fusion HCI provides the foundational infrastructure layer. Together, they allow organizations to move from fragmented, risky AI experiments to a trusted Sovereign AI Knowledge Layer.

Atolio’s role: Atolio connects directly to the systems where work already happens: all of a company’s documents, chats, tickets, wikis, CRM records, and file stores across all tools and platforms. It creates a unified, permission-aware index that allows employees and AI systems to securely retrieve trusted enterprise context.

IBM Fusion HCI’s role: IBM Fusion HCI provides the high-performance, integrated hardware foundation. Operating as a turnkey OpenShift in a box, it brings together the massive compute

(GPUs), storage, networking, and data services required to deploy and run Atolio natively within your own perimeter.
In practice, the joint solution works like this:

- **Faster Time to Value for Governed AI:** Fusion HCI gives Atolio a pre-validated OpenShift-based foundation, reducing infrastructure integration work and helping teams move from AI pilots to production knowledge retrieval faster.
- **Secure, Unified Knowledge Access with Performance:** Atolio connects to enterprise systems and builds a real-time, permission-aware index, while Fusion HCI enhances this with localized, S3-compatible object storage, reducing latency and accelerating search, retrieval, and AI-driven insights across both SaaS and legacy data.
- **Resilience, Governance & Production-Scale Reliability:** Fusion HCI strengthens Atolio with built-in high availability and disaster recovery (HA/DR) to ensure uninterrupted access to knowledge, while adding data cataloging and policy controls that make Atolio's insights not only fast, but also trusted, compliant, and auditable for large-scale AI deployments.

What Customers Gain from the Joint IBM Fusion HCI Solution + Atolio

By bringing your intelligence layer directly to your secure storage, you eliminate the trade-off between AI performance and data security. Fusion customers realize transformational benefits across speed, security, and operational simplicity.

1. **Accelerated Productivity & Elevated Performance:** Employees instantly find accurate answers and existing work materials across enterprise systems with unprecedented speed. This rapid access to the best internal knowledge not only eliminates wasted search time, but drives improved content creation, boosting both daily productivity and overall employee performance.
2. **Grounded & Trustworthy AI**: Responses are highly accurate and verifiable because they are grounded in approved internal knowledge. Users can see exactly where answers originate, building confidence in the system and ensuring decisions are based on facts.
3. **Secure, Permission-Enforced Governance:** Retrieval strictly respects existing security boundaries and user access rights. This lowers compliance risks and supports a strong hybrid strategy, keeping sensitive data completely secure and out of public clouds.
4. **Simplified Infrastructure & Reduced AI Costs:** The combination of Atolio and IBM Fusion accelerates time-to-value by eliminating the need to build and maintain custom connectors or complex RAG pipelines. Furthermore, by feeding LLMs only highly targeted, relevant data, organizations significantly minimize token consumption and the overall cost of AI.

Key Use Cases

Here are the high-value outcomes this architecture delivers, regardless of industry.

Decision Intelligence
A compliance officer needs to prepare a regulatory response in 48 hours. The evidence is spread across contracts, audit logs, policy documents, and email. Assembling it manually takes two days. With Atolio, one search returns everything she is authorized to see across every connected system in minutes. Sales leaders get live account context before a board review. Product managers get a complete customer signal before a roadmap call. Finance teams access source data before a forecast. Engineers surface incident history before a planning session. Every role gets the full picture in time to use it, not after the decision is already made.

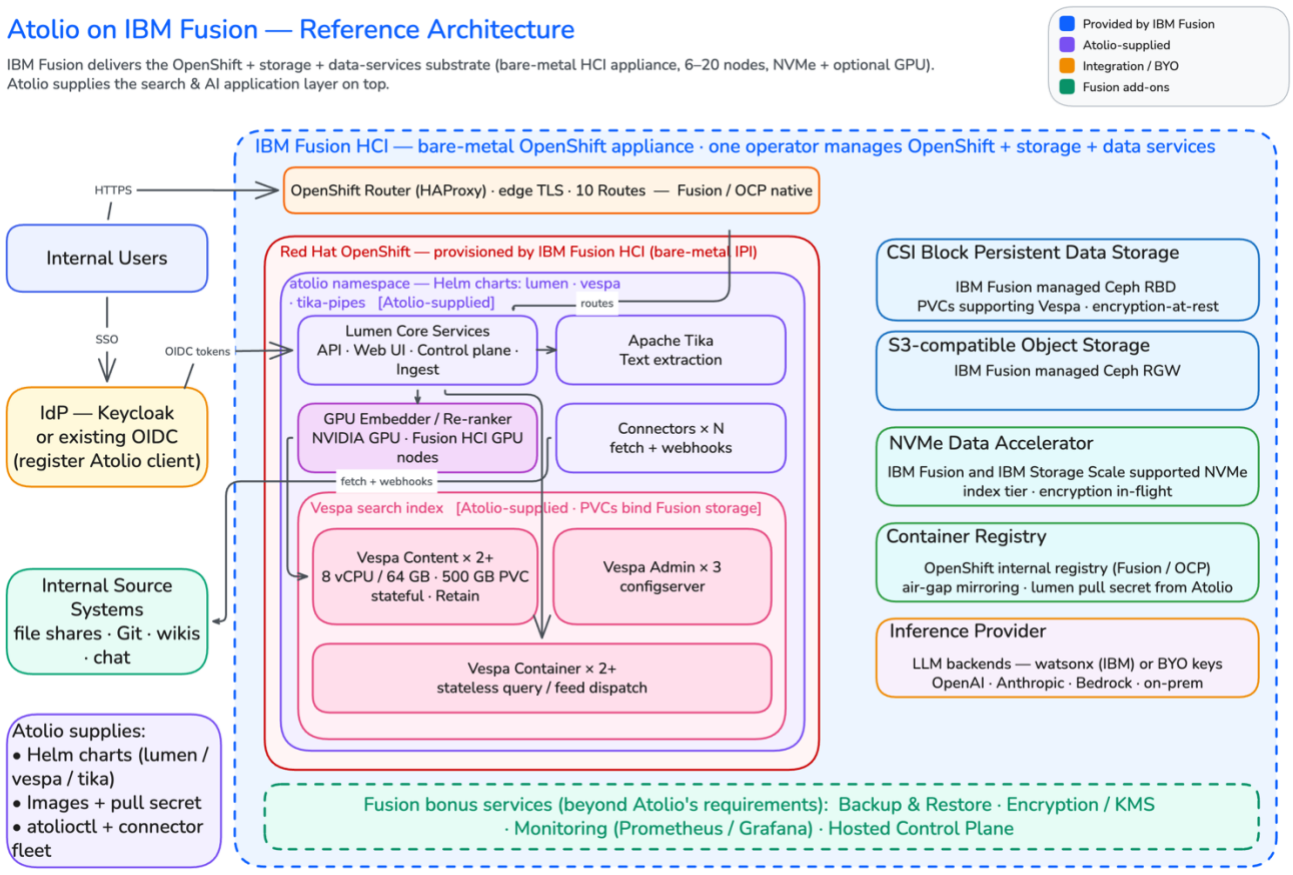
Governed Agent Access
Agents that can access everything are a liability. Agents that cannot access enough are ineffective.
An IT team is building an internal agent to answer HR policy questions and surface technical documentation. To be useful, it needs access to company knowledge. But broad access creates risk: a confidential compensation document surfaced to the wrong employee, a legal matter visible outside the matter team, restricted specs returned to someone in an unrelated department. Atolio acts as the agent's retrieval layer, returning only what each user is already authorized to see across every connected system. No access is granted. No permissions are overridden. Teams deploy customer-facing agents. Legal deploys matter research agents. HR deploys policy agents. All are governed by the same access controls already in place.

AI Grounding
A legal team asks their AI assistant to summarize contract risk. It returns generic answers that could apply to any company. It has not seen the actual agreements, so they revert to manual work. This happens repeatedly across functions. Atolio connects AI assistants to the full body of company knowledge across systems, while respecting existing permissions. Legal gets summaries grounded in executed agreements. Engineers get triage guidance from real incident history. HR gets policy answers from internal documentation. Finance gets analysis from historical data. GTM teams get account insights from CRM, tickets, and conversations.

Institutional Knowledge
A production incident hits Friday afternoon. The engineer who built the service left six months ago. Without Atolio, searching returns partial answers. With Atolio, one search surfaces the design doc, original decision thread, and prior postmortem, all within existing permissions. Legal teams find prior rulings without waiting on a senior attorney. Support teams resolve escalations without routing to a specialist. GTM teams find account history without tracking down prior owners. New hires ramp faster because answers are easy to find, not dependent on others’ availability.

Architecture Overview

Reference Architecture Diagram



Atolio Core Components

Connector Fleet: 48+ enterprise connectors built and maintained in-house by Atolio's engineering team in Go. All connectors run inside the Kubernetes cluster and connect outbound to enterprise systems. API tokens and credentials never leave the customer's environment. Each connector handles user directory resolution, permission structure extraction, historical backfill, and real-time streaming of new content and permission changes.

Document Extraction Pipeline (Apache Tika): Inbound content from all connected systems passes through Atolio's extraction pipeline, which handles multimodal document types including

PDFs, Word documents, spreadsheets, presentations, images, code files, and nested layouts via built-in OCR. Extracted text, tables, metadata, and structure are normalized before indexing.

Vespa Search and Vector Database: The core of Atolio's retrieval engine is a self-hosted Vespa deployment running on IBM Fusion HCI's NVMe-backed IBM Storage Scale. Vespa stores full document text, dense vector embeddings calculated on IBM GPU nodes, object-level ACLs, structured metadata, and Collaboration Graph weightings for every indexed document. Vespa provides horizontal scaling, replication, and high availability inside the cluster.

Collaboration Graph Engine: A proprietary, real-time weighted graph built for every user that continuously maps who they collaborate with, what systems and projects they are active in, and how those relationships evolve. The graph is stored in Vespa and applied at query time to personalize results, resolve ambiguous queries, surface subject matter experts, and understand cross-team context references in natural language.

Fusion HCI Core Components

IBM Fusion HCI underpins each core component of the Atolio platform by providing a tightly integrated OpenShift-based substrate for compute, storage, and data services. The connector fleet, extraction pipeline, and Vespa search layer all run within a unified Kubernetes environment, while Fusion’s NVMe-backed Storage Scale and S3-compatible object storage ensure high-throughput ingestion, low-latency retrieval, and durable persistence of indexed enterprise knowledge. GPU nodes embedded within the Fusion HCI stack accelerate embedding generation, reranking, and model inference, enabling Atolio to operate at enterprise scale without compromising performance or data locality.

End to End Workflows

Query and LLM Orchestration Layer: At query time, Atolio determines the execution path: a simple keyword query triggers a fast hybrid retrieval; a complex natural language question triggers a multi-step reasoning loop. The orchestration layer builds a prompt enriched with Collaboration Graph context, issues targeted searches against Vespa, applies an LLM-as-judge evaluation step to assess result quality, and generates a grounded, cited, hallucination-minimized answer using the customer's chosen model. All inference runs inside the customer's Fusion HCI environment.

Platform API and Agent SDK: All platform functionality is exposed via a gRPC API running inside the cluster, providing high-performance service-to-service communication. The Agent SDK provides orchestration primitives for building autonomous AI research agents that decompose queries, retrieve across multiple systems in parallel, evaluate results, and synthesize structured, cited outputs. The Platform API supports the Model Context Protocol (MCP) for integration with downstream agent frameworks. All agentic workflows run entirely within the Fusion HCI perimeter.

Guidelines and Sizing Recommendations for deploying Atolio on IBM Fusion HCI

For a brand new deployment

Fusion HCI Resources	Small (1000 users)	Medium (5000 users)	Large (20,000 users)	POC (50-100 users)
Control Nodes (2 x 16-core X86 Processors, 256 GB RAM, 2 x 7.68 TB NVMe drives)	3	3	3	3
Compute-Only Nodes (2 x 16-core X86 Processors, 256 GB RAM)	2	1	3	1
GPU Nodes (1 RTX Pro 6000/node, 256 vCPUs, 1536 GB RAM)	1	2	3	0

Each 32-core Control and 32-core Compute Node comes with:

- 1x NVIDIA ConnectX-6 DX dual-port 100GbE PCIe NIC
- 1x NVIDIA ConnectX-6 LX dual-port 25GbE PCIe NIC
- 1x 1GbE RJ45 4-port OCP Ethernet Adapter
- Additionally, a Fusion HCI rack comes with:
- 2 x Ethernet 1 Gb management switches
- 2 x Ethernet 200 Gb network switches
- Support/service node server (one required per cluster)

Each 128-core GPU Node comes with:

- 1x NVIDIA ConnectX-7 dual-port 200GbE PCIe NIC
- 1x NVIDIA ConnectX-6 LX dual-port 25GbE PCIe NIC
- 1x 1GbE RJ45 4-port OCP Ethernet Adapter

Conclusion

Enterprise knowledge is not missing. It is scattered, hard to reach, and disconnected from the AI systems that need it most.

Atolio provides the permission-aware intelligence layer: connecting to knowledge where it already lives, ranking what matters most, and ensuring users and AI systems only ever access what they are authorized to see. IBM Fusion HCI provides the infrastructure that makes it possible to run all of this entirely within the enterprise perimeter, with the performance, resilience, and governance that production-scale AI demands.

The use cases in this brief show what that looks like in practice. Compliance officers assemble regulatory responses in minutes instead of days. Legal teams get AI summaries grounded in actual executed agreements. Engineers find the institutional knowledge they need without waiting on the one person who holds it. AI agents retrieve company knowledge without creating the security exposure that makes IT and compliance teams say no.

By combining these strengths, organizations can move from hopeful AI pilots to real business outcomes:

- **Sovereign AI Knowledge Layer:** A jointly validated enterprise knowledge solution deployed natively on IBM Fusion, offering complete data sovereignty across on-premises HCI, VPC, or hybrid cloud environments.
- **One Knowledge Layer, Every System:** Replaces fragile point-to-point connections and shadow AI workarounds with a single, governed retrieval layer across every enterprise system. Less complexity, more control.
- **Faster Decisions, Lower Costs:** Employees spend less time searching and more time acting. AI systems retrieve only what is relevant, reducing token consumption and the cost of every AI interaction.
- **Expand Without Starting Over:** New teams, new systems, and new AI use cases connect to the same knowledge layer. No additional infrastructure projects, no data migrations, no disruption to what is already running.

In every case, the value comes from the same foundation: a knowledge layer that is trusted, governed, and sovereign. One that makes AI useful without asking the enterprise to surrender control of its most sensitive information. With Atolio on IBM Fusion HCI, organizations do not have to choose between AI capability and data sovereignty. They get both, running securely inside their own environment.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
June 2026