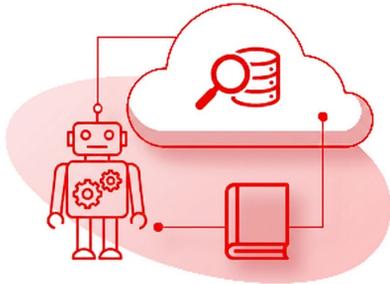# Build an Agentic RAG with Core AI Suite to improve system transparency and maximize productivity

Is your existing RAG system becoming a barrier to AI adoption?

As organizations scale AI-based knowledge management, they face limitations such as black-box responses, opaque reasoning, and inaccurate answers to complex questions. Existing RAG systems often cannot explain why a given answer was generated, making it difficult to operate AI reliably in enterprise environments.

Applied Technology

- Red Hat OpenShift AI
- Red Hat AI Inference Server (vLLM)
- Red Hat OpenShift
- Elasticsearch (Hybrid Search)
- Docling (Document Processing)

Core Values

- Complete Transparency
- Maximum Productivity
- High Response Accuracy
- Enhanced User Experience
- Flexible Architecture

Build a true Agentic RAG solution with S-Core and Red Hat

Core AI Suite combines Red Hat OpenShift AI and MCP (Model Context Protocol) to make the AI inference workflow fully transparent, reduce operational complexity, and ensure consistent performance across diverse environments.

- **Autonomous AI Agent**: Beyond simple search, the AI agent can reason, plan, and iteratively gather information until it finds the best answer. This enables more accurate responses, even for complex questions.

- **Complete Transparency**: Track the entire process in real time through visualizations that show which tools the agent called and why it made those decisions. AI no longer has to remain a black box.

- **Maximizing Productivity**: With our own LLM served through the Red Hat AI Inference Server, fast responses can be delivered without relying on external APIs.

- **Security and Regulatory Compliance**: By deploying in on-premises or hybrid environments, the solution supports data sovereignty and regulatory requirements, enabling safe use in finance, healthcare, and the public sector.

- **Streamlined Operations & Flexible Scalability**: The Red Hat OpenShift AI pipeline automates data preprocessing, chunking, and embedding generation, allowing developers to focus on higher-value tasks. From core to edge, the solution can be deployed close to where data resides, with the flexibility to scale as needed.

Traditional RAG vs Core AI Suite (Agentic RAG)

Traditional RAG	Core AI Suite (Agentic RAG)
Instant responses after a single search	The agent performs analysis, planning, and iterative retrieval
Black-box inference process	Transparent, real-time visibility into the entire MCP-based process
Inaccurate responses to complex questions	Breaks requests into sub-tasks and combines the results to generate more accurate answers
Requires manual pipeline management	Higher productivity through automated workflows
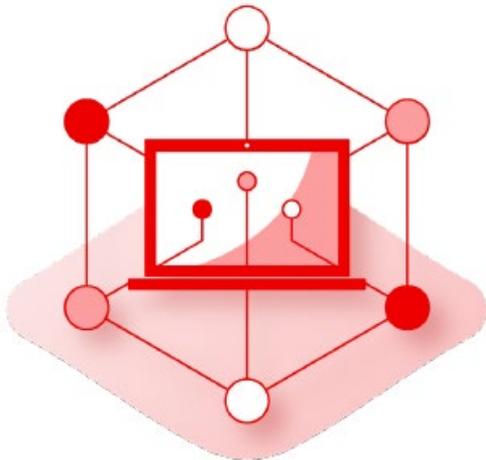
Key Features

- **MCP-based Real-Time Visualization:** Monitor all actions of the AI agent in real time. Use an intuitive dashboard to see which tools the agent called, what data was processed at each step, and how the final response was generated.

- **Red Hat AI Inference Server:** A high-performance inference engine based on vLLM that uses GPU memory efficiently, delivers high throughput under concurrent requests, and supports a wide range of open-source LLMs.

- **Red Hat OpenShift AI Pipeline:** It automates the end-to-end data workflow through an ML pipeline based on Kubeflow.

- **Multi-Format Support:** Powered by the open-source Docling engine, the solution can process a wide range of enterprise document formats, while preserving document structure to improve search quality.

- **Enterprise Deployment:** Its containerized architecture, built on Red Hat OpenShift AI, enables consistent deployment across environments.
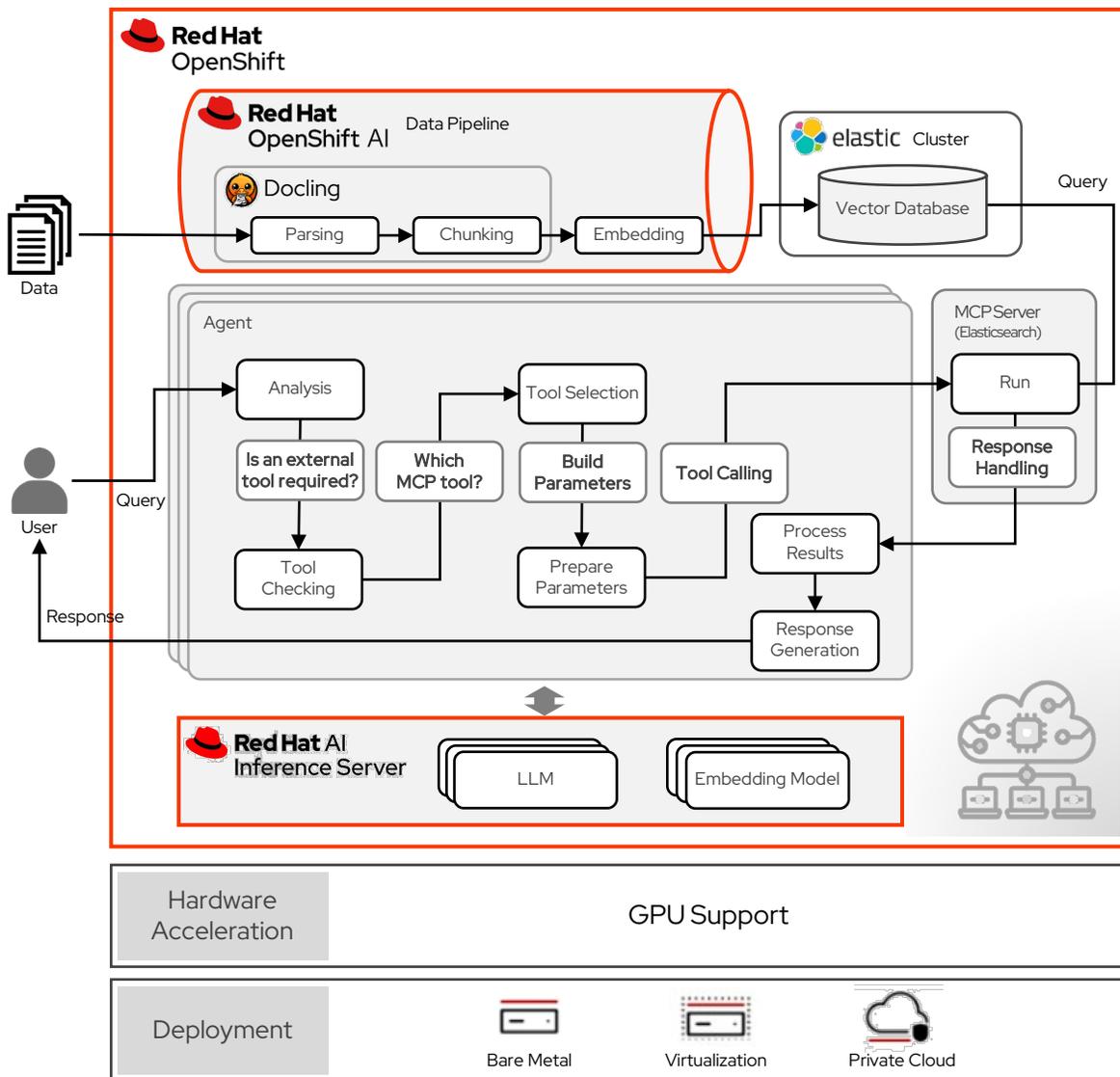
Core AI Suite built on Red Hat AI Products

Fig. 1 – The diagram shows how Red Hat OpenShift AI and MCP Server are integrated to form the complete Agentic RAG platform, Core AI Suite.

Red Hat OpenShift AI provides enterprise data pipelines, including document processing through Docling and embedding generation. The MCP Server manages agent tool calling and API orchestration, while Elasticsearch supports hybrid search and efficient data storage as the vector database. The LLM and embedding models are optimized and served through the Red Hat AI Inference Server, running on flexible, GPU-enabled infrastructure across bare metal, virtualized, and private cloud environments.

This enables enterprise-grade Agentic RAG with real-time transparency and automated workflows.

S-Core

S-Core is a company specializing in open-source solutions.

As a trusted Red Hat specialist, S-Core helps organizations build and operate stable, production-ready AI services.

With certified expertise across governance, licensing, security and technical support, S-Core provides end to end services grounded in deep open-source and Agentic AI capabilities.

Successful Agentic AI adoption depend on both the right technology and the right implementation partner. Unlock the full potential of AI with S-Core's open source expertise and a service model trusted by returning customers.

Getting Started

Core AI Suite enables rapid deployment through proven integration with Red Hat OpenShift AI. Leverage preconfigured pipeline templates without having to build complex infrastructure from scratch.

- Deployment Phase: Environment Setup and Model Selection → Platform Integration → Validation and Rollout

Value Creation Through AI Integration

As AI technology advances and organizational capabilities mature, AI integration is becoming a key driver of value creation and competitive advantage.

Take the lead by adopting Agentic RAG with Core AI Suite.