Whitepaper: Sovereign RAG Architecture Guide - w38

Singularity IOSovereign Agentic AI Platform Zurich, Switzerland I www.singularityio.ch

---

SOVEREIGN RAG ARCHITECTURE GUIDE

Secure Sovereign RAG Architecture on Exoscale SKS
Whitepaper for Cross-Industry Use
Version 1.0 I May 2026Singularity IO – Zurich, Switzerland

Executive Summary

Many organizations need powerful Retrieval-Augmented Generation (RAG) capabilities but cannot accept the security, compliance, and sovereignty risks of sending sensitive data to foreign cloud providers.

The **Sovereign RAG Architecture on Exoscale SKS** is a production-grade, fully Swiss-hosted solution that combines local LLM inference (Ollama), a high-performance vector database (Qdrant), and secure orchestration — all running inside your dedicated tenant on Exoscale's Swiss data centers.

This **Standard Agent**-ready architecture delivers accurate, context-rich responses while guaranteeing 100% data residency, DSG/GDPR compliance, and EU AI Act readiness.

Key Benefits

- Full Swiss data sovereignty — nothing leaves Switzerland
- High-accuracy RAG with low latency and strong privacy
- Seamless integration with agentic workflows (LangGraph, Dify, OpenClaw)
- Enterprise-grade security and auditability
- Rapid deployment and strong ROI through reduced external API costs and risk

---

1. The Challenge: Foreign Cloud RAG Risks

Organizations across industries face critical barriers when implementing RAG:

- Data residency and sovereignty requirements (especially in Switzerland and the EU)
- Strict DSG/GDPR and upcoming EU AI Act obligations
- High costs and latency of sending documents to foreign LLM providers
- Security and confidentiality risks with sensitive or regulated data
- Lack of transparency and control over retrieval and generation processes
- Difficulty integrating RAG into production agentic systems

Traditional cloud RAG solutions often violate sovereignty or compliance needs.

---

2. The Solution: Sovereign RAG on Exoscale SKS

The Singularity **Sovereign RAG Architecture** is a complete, secure blueprint built natively on the Singularity Agentic Platform.

It combines local inference, sovereign vector storage, and agent-ready orchestration — giving you full control while delivering production-grade performance.

Core Components

- **Ollama** — GPU-accelerated local LLM inference (Llama 3.1, Mistral, etc.)
- **Qdrant** — High-performance, open-source vector database
- **LangGraph / Dify** — Orchestration and agent integration
- **n8n** — Secure document ingestion and workflow automation
- **LangSmith** — Full observability and audit trails

All components run inside your isolated Kubernetes namespace on Exoscale SKS.

---

3. How the Sovereign RAG Architecture Works (Technical Blueprint)

Standard Agent Architecture on Singularity Platform

1. **Ingestion & Embedding Layer** (n8n + Qdrant)
    - Secure document upload, chunking, and embedding using local models.
    - Metadata tagging and access-control enforcement.
1. **Retrieval Layer** (Qdrant)
    - Hybrid search (semantic + keyword + metadata filtering).
    - Advanced reranking and context compression.
1. **Generation Layer** (Ollama)
    - GPU-accelerated inference with prompt engineering and guardrails.
    - Context-aware response generation.
1. **Agentic Orchestration Layer** (LangGraph)
    - RAG as a reusable tool inside larger autonomous agents.
    - Stateful memory and multi-step reasoning.
1. **Governance Layer** (LangSmith + Kyverno)
    - Complete traceability, confidence scoring, and audit logging.

Deployment Tier: Standard (with dedicated GPU slice) — sufficient for most production RAG workloads.

---

4. Key Architecture Features

- **100% Swiss Data Residency** on Exoscale SKS (ch-gva-2 or ch-dk-2)
- **Multi-tenancy with strong isolation** (Kubernetes namespaces + NetworkPolicy)
- **Hybrid Search & Advanced Retrieval** techniques
- **Fine-grained Access Control** at document and chunk level
- **Human-in-the-Loop & Guardrails** configurable per use case
- **Scalable & Cost-Efficient** — pay only for your dedicated resources

---

5. Proven Business Outcomes

Typical Results

- High retrieval accuracy (often >85–92% with proper chunking)
- Dramatic reduction in external LLM API costs
- Full compliance and audit readiness for regulated industries
- Faster development of agentic applications using sovereign RAG as a core tool
- Enhanced trust and adoption of AI solutions internally and with clients

ROI Example:

- Monthly Standard Tenant: CHF 2,990
- Significant savings on foreign APIs + risk mitigation → **ROI 171–192%**

---

6. Why Deploy on the Singularity Agentic Platform

- **True Sovereignty**: Built from the ground up on Exoscale SKS in Switzerland
- **Production-Ready Stack**: Ollama, Qdrant, LangGraph, Dify, n8n, LangSmith
- **Agent-Native**: RAG is seamlessly integrated into your autonomous agents
- **Compliance by Design**: DSG/GDPR + EU AI Act ready
- **Rapid Time-to-Value**: Start with Dify prototypes, move to LangGraph for production

Standard Tier

- CHF 2,990 monthly base fee (12 agents included)
- Extra agents: CHF 299/month

---

7. Implementation Roadmap

- **Phase 1 (Week 1–2)**: Provision tenant, deploy Qdrant + Ollama, and set up secure ingestion pipelines using Dify
- **Phase 2 (Week 3–4)**: Implement chunking strategy, embeddings, and hybrid retrieval with LangGraph
- **Phase 3 (Week 5–6)**: Add guardrails, human-in-the-loop, and test with real documents
- **Phase 4 (Week 7–8)**: Integrate into production agents, enable monitoring, and optimize performance

---

8. Compliance & Security Framework

- Full DSG/GDPR compliance and data residency guarantees
- EU AI Act alignment (transparency, risk management)
- Kyverno policies, RBAC, NetworkPolicy, and sealed secrets
- Comprehensive LangSmith tracing for every query and response
- All data and inference stay exclusively in Swiss data centers

---

Conclusion

The **Sovereign RAG Architecture Guide** provides a complete, secure, and production-proven blueprint for organizations that refuse to compromise on data sovereignty.

By deploying this architecture on the Singularity Agentic Platform, you gain powerful, accurate, and trustworthy RAG capabilities — fully hosted in Switzerland — that serve as the foundation for secure agentic AI across your entire organization.

Singularity IO
www.singularityio.ch
Zurich, Switzerland
