

Solution Brief

Private RAG Appliance for Enterprise Knowledge Intelligence

Challenges

- Information is scattered across file servers and internal databases, making it difficult to search and utilize effectively
- Public LLMs generate answers without grounding in enterprise data, leading to unreliable or non-verifiable outputs
- Sending sensitive enterprise data to cloud-based AI services introduces security, compliance, and sovereignty risks

Solution

- Ingests unstructured data from documents, file shares, internal knowledge bases to converts data into vector embeddings for semantic search
- RAG combines semantic search and LLM reasoning to ensures responses are grounded in enterprise data rather than generic knowledge
- High-performance AI devices integrate computing and storage platforms, are optimized for low-latency retrieval, support concurrent users and high query throughput, and are suitable for enterprise-level deployments

Balancing AI Utility with Data Security

Enterprises today are rich in data but poor in accessible, trustworthy knowledge. Most critical information resides in unstructured and siloed repositories such as documents, PDFs, emails, and internal systems. This creates several challenges:

- **Data Privacy and Sovereignty:** Uploading proprietary documents (contracts, R&D records, employee data) to public cloud AI services poses significant compliance and security risks.
- **The "Hallucination" Problem:** Standard LLMs (Large Language Models) often generate confident but incorrect information when they lack access to specific, up-to-date company data.
- **Complex Data Integration:** Managing the pipeline from raw document storage to AI-ready vector databases is often technically fragmented and difficult to scale.
- **Performance Bottlenecks:** Processing thousands of internal documents requires high-speed data throughput between the storage repository and the GPU processing unit.

Altos GB10 F1 and QSAN XN1 for Private RAG

This integrated appliance combines high-density GPU compute with enterprise-grade storage to create a seamless, end-to-end RAG (Retrieval-Augmented Generation) pipeline.

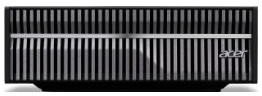
- **Secure On-Premise Data Repository:** The QSAN XN1 unified storage serves as the centralized "Source of Truth," storing all unstructured data (PDFs, Wikis, Docs). By keeping the data on-premise, enterprises ensure that no information leaves the corporate firewall.
- **High-Speed Vectorization and Embedding:** Powered by NVIDIA GPUs, the Altos GB10 F1 rapidly processes documents from the storage, converting them into high-dimensional vectors. Utilizing RDMA over 25 GbE, the system ensures zero-latency data movement between storage and compute.

- **Simplified Management with aiWorks:** The solution comes pre-integrated with Altos aiWorks 5.0, providing a user-friendly interface to manage document ingestion, model selection, and RAG workflow orchestration.
- **Auditability and Compliance:** Every AI response can be traced back to the specific source document stored on the QSAN XN1, meeting the "traceable and auditable" requirements of regulated industries like finance and legal.

Private RAG Appliance (Enterprise Knowledge Base)

Private RAG Appliance transforms enterprise data into trusted, explainable insights by combining secure on-prem AI with real-time knowledge retrieval.

Recommended ALTOS Models



Altos BrainSphere™
GB10 F1

- NVIDA® GB10 Grace Blackwell Superchip
- NVIDA® Blackwell GPU 128 GB Unified Memory
- 1 PetaFLOP FP4 AI Computing Performance
- Up to 4 TB NVMe Storage Capacity
- NVIDA® ConnectX®-7 SmartNIC
- One-click AI Developmnet and Deployment Platform

Recommended QSAN Models



QSAN XN1205-4A

- Desktop Tower NAS
- AMD® Ryzen™ 4-core
- 8 GB DDR5 DIMM up to 64 GB
- 3.5” Slot x 4 + 2.5” Slot x 1 + M.2 NVMe x 2
- Maximum Raw Capacity up to 153 TB



QSAN XN1209-4A

- Desktop Tower NAS
- AMD® Ryzen™ 4-core
- 8 GB DDR5 DIMM up to 64 GB
- 3.5” Slot x 8 + 2.5” Slot x 1 + M.2 NVMe x 2
- Maximum Raw Capacity up to 276 TB

© 2026 QSAN Inc. All Rights Reserved. QSAN, the QSAN logo, and the marks listed at www.qsan.com are trademarks of QSAN Inc. Other company and product names may be trademarks of their respective owners.
April 2026