Solution Brief

Silk AI Enablement:
Microsoft SQL Server Retrieval Augmented Generation

Mission-critical SQL Server databases are at the heart of enterprise operations, and as businesses adopt advanced AI workflows, integrating SQL with vector retrieval augmented generation (RAG) systems has become essential. Whether leveraging traditional relational queries or hybrid approaches like LlamaIndex or Azure AI Search with Copilot, these systems demand low-latency, high-throughput storage to perform at scale.

Silk for High-Performance SQL Server RAG

The Silk Platform provides a high-performance, scalable storage solution designed to optimize SQL Server 2022/2025 and integrate seamlessly with AI-driven retrieval augmented generation workflows. Silk enables enterprises to deliver superior performance for transactional workloads, complex relational queries, and vectorized search pipelines—all within a unified storage platform.

Key Features and Benefits

1. Optimized SQL Server Performance

- **Ultra-Low Latency:** Sub-millisecond response times for transactional and analytical queries.
- **High Throughput:** Supports SQL workloads requiring up to 2m IOPS IOPS and upto 20GB/sec throughput, ensuring consistent performance under heavy loads.
- **Dynamic Scalability:** Scale storage dynamically as database sizes grow, with no impact on performance.

2. AI-Augmented Query Workflows

- **With LlamaIndex:** Enhance SQL Server by integrating LlamaIndex to retrieve and vectorize data for retrieval-augmented generation (RAG) workflows.
- **Without LlamaIndex: C**ombine SQL Server with Azure AI Search and Copilot for hybrid query pipelines, leveraging vectorized data searches alongside traditional SQL queries.
- **Real-Time Retrieval:** Silk ensures low-latency access to both relational and vectorized data, enabling seamless hybrid queries.

Silk is a Single Platform for All Your AI Workloads:

- Zero-footprint snapshots for instant backups without additional storage overhead.
- Active-active architecture for continuous data availability.
- Enhanced data privacy through Redgate data masking for secure AI model training and inference.
- Optimize costs while delivering peak performance for SQL and AI workloads.

3. Advanced Data Management

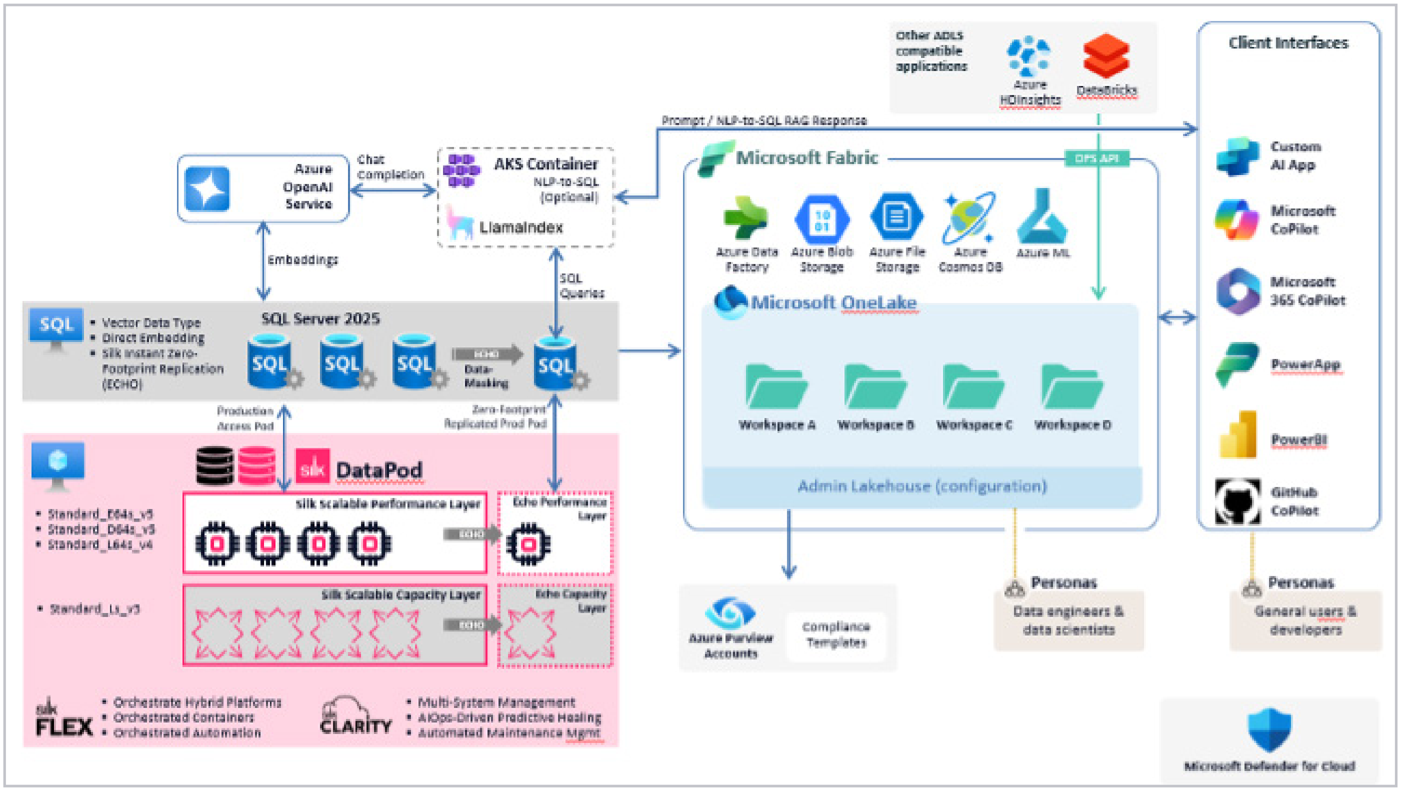
- **Zero-Footprint Snapshots:** Instantly create backups of SQL databases or vector indices without consuming additional storage.
- **Compression and Deduplication:** Reduce storage costs while maintaining fast access to large datasets and indices.
- **Resiliency:** Silk's symmetric active-active architecture ensures high availability for mission-critical workloads.

4. Seamless Azure Integration

- **Azure Ecosystem Support:** Fully compatible with Azure SQL Database, Azure AI Search, and OpenAI services.
- **Flexible Storage Access:** Expose Silk storage via NFS/SMB to compute resources for AI workflows or directly to Azure SQL VMs.
- **Data Orchestration:** Integrates with Azure Machine Learning, Azure Data Factory, and Azure Synapse for end-to-end data management

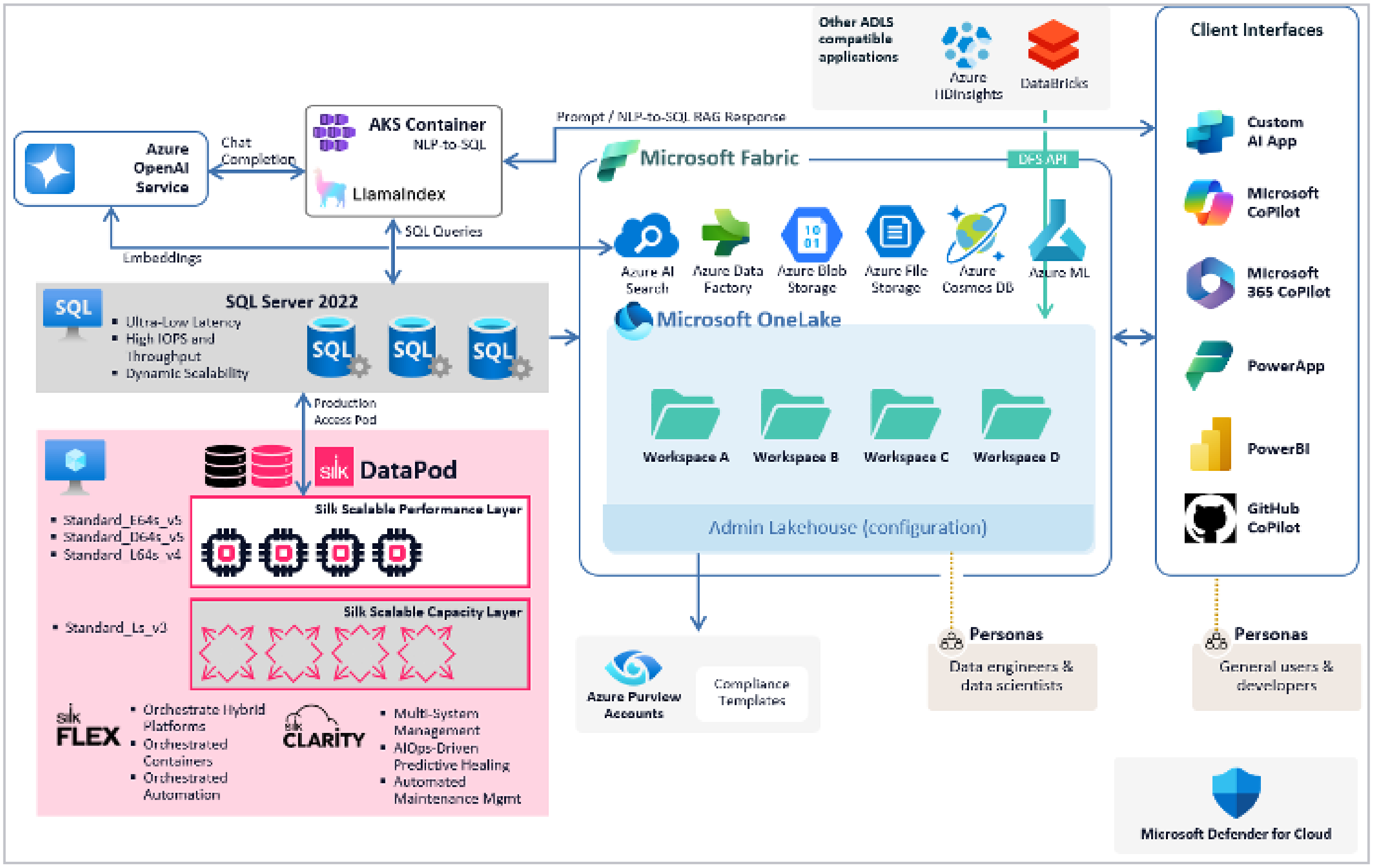
Reference Architecture

Silk-Optimized SQL Server 2025 RAG with Optional LlamaIndex



- **Native Embedding Support:** SQL Server 2025 integrates in-database embeddings and vector data types for seamless AI workflows.
- **Hybrid Query Optimization:** Leverage Silk's ultra-low latency for high-performance relational and vector similarity queries.
- **AI-Ready Scalability:** Silk's storage ensures efficient growth for large embedding datasets with real-time compression and snapshots.
- **Built-In AI Functions:** Accelerate workloads with SQL 2025's cosine similarity and vector operations powered by Silk's storage performance.

Silk-Optimized SQL Server 2022 RAG with LlamaIndex



These two reference architecture diagrams describes typical deployments:

1. SQL + LlamaIndex Pipeline

- **SQL Server (2022/2025):** Store transactional and relational data in Silk-backed databases.
- **LlamaIndex:** Query SQL Server for raw data, convert results into embeddings, and retrieve relevant information for RAG workflows.
- **Silk Storage:** Host both SQL databases and vector embeddings, ensuring low-latency access to relational and vectorized data.

2. SQL + Azure AI Search Pipeline

- **SQL Server (2022/2025):** Store structured data in Silk-backed databases.
- **Azure AI Search:** Use Silk to store pre-indexed vectorized embeddings for unstructured data search.
- **Copilot Integration:** Combine SQL queries with Copilot for hybrid responses that integrate structured relational results with AI-driven vector searches.

Technical Advantages of Silk in SQL Server RAG:

- **Unified Storage:** Host both SQL databases and vector embeddings on a single platform.
- **Low Latency:** Optimize both SQL queries and AI-driven vector searches for sub-millisecond performance.
- **Scalable and Reliable:** Silk's architecture ensures seamless scalability and high availability.
- **Cost Efficiency:** Minimize storage costs with compression and deduplication.

Use Cases

1. Hybrid Relational and Vectorized Search

Challenge: Enterprises need to combine relational SQL queries with vector-based searches for RAG or hybrid pipelines.

Solution: Silk supports simultaneous high-speed SQL queries and vectorized searches, enabling seamless integration of relational and unstructured data workflows.

2. Transactional and Analytical Queries

Challenge: Traditional SQL workloads often face latency and throughput issues during peak usage.

Solution: Silk accelerates transactional and analytical queries by ensuring ultra-low latency and high IOPS, even under heavy loads.

3. AI-Enhanced Applications with SQL

Challenge: AI tools like LlamaIndex or Copilot require rapid access to SQL datasets and embeddings for hybrid responses.

Solution: Silk provides the high-performance storage foundation needed for real-time retrieval and AI-powered query augmentation.

Integration Workflow Example

1. SQL + LlamaIndex (RAG Workflow)

- Query SQL Server for structured data stored on Silk.
- LlamaIndex converts results into vector embeddings stored in Silk.
- Perform vector similarity searches for relevant content.
- Combine SQL and RAG results into hybrid responses for downstream applications.

2. SQL + Azure AI Search + Copilot

- Store structured SQL datasets and unstructured vectorized embeddings on Silk.
- Azure AI Search indexes data for hybrid query execution.
- Copilot combines SQL query results with vectorized insights for user-facing applications.

Conclusion

The Silk Platform is the ideal solution for hosting high-performance vector databases in Azure environments. Its unmatched speed, scalability, and cost efficiency empower businesses to deploy AI-driven applications with confidence. Whether you're building RAG pipelines, powering recommendation systems, or optimizing AI search, Silk ensures your workloads perform at their best.