Vector Database Scalability Analysis

Test Configuration

• Document corpus: 1M+ technical documents
• Query load: 100 concurrent users
• Embedding model: text-embedding-3-large (1536 dimensions)
• Test duration: 72 hours continuous operation

Performance Results

Azure AI Search demonstrates superior performance for hybrid search scenarios, combining semantic similarity with traditional keyword matching. The system maintains sub-2-second response times even under heavy load conditions.

Scalability Metrics
• Index creation time: 45 minutes for 1M documents
• Query latency P95: Under 2 seconds
• Memory usage: 8GB for 1M vectors (1536-dim)
• Concurrent query capacity: 200+ simultaneous requests

Recommendations
For production deployments processing large document collections, Azure AI Search provides the optimal balance of performance, cost, and feature richness. The hybrid search capabilities are particularly valuable for RAG applications requiring both semantic and exact match capabilities.