Microsoft

Generative AI with JavaScript

Building Generative AI Apps with Azure Cosmos DB

Agenda

- Vector Indexing in Azure
- Azure Cosmos DB for AI apps
- GenAI use cases
- Demo

Choosing Vector Indexes in Azure

Vector Indexes in Azure databases is preferred when:

- You have structured or semi-structured **operational data** (e.g. chat history, customer profiles, etc) in the database
- Simplified architecture **for single source of truth** that combines **vector similarity search inline with database queries**. No need to synchronize separated database and vector index.
- The workload benefits from mission-critical **OLTP database characteristics** (e.g. stronger guarantees for availability, performance, and data durability)

↓

NoSQL / semi-structured data

Azure Cosmos DB

↓

PostgreSQL familiarity
Relational data

Azure Database for PostgreSQL (flex)

↓

Cassandra familiarity
Data is already in Apache Cassandra

Apache Cassandra Managed Instance

Azure AI Search is preferred when:

- You need to index structured/unstructured data (e.g. images, docx, PDFs) from **variety of internal and external data sources**
- Your application requires **state-of-the-art search technology** for higher search quality (e.g. hybrid full-text/vector search, fuzzy, autocomplete, semantic re-ranking, multi-language, metafiltering)
- The workload requires **multi-modal search** and/or multi-modal embeddings to perform OCR, image analysis, translation, etc.
- You're building a **Bing-like search experience** in a custom application

↓

Azure AI Search

Azure Cosmos DB is a set of highly-scalable & AI-ready databases

Azure Cosmos DB NoSQL

- Serverless or Provisioned Throughput
- High elasticity with instant autoscale
- Low latency, real-time data transactions
- Mission critical reliability (99.999%)
- Built-in vector index and search with DiskANN

Azure Cosmos DB for MongoDB

- Mongo DB compatible
- Provision dedicated compute + storage
- Store data + vectors together, keep consistent
- Built-in vector index and search (ft. IVF & HNSW)

AI scenarios with Azure Cosmos DB for NoSQL

Chat history

Retrieval Augmented Generation (RAG)

Multi-tenant AI apps

Real-time Recommendations

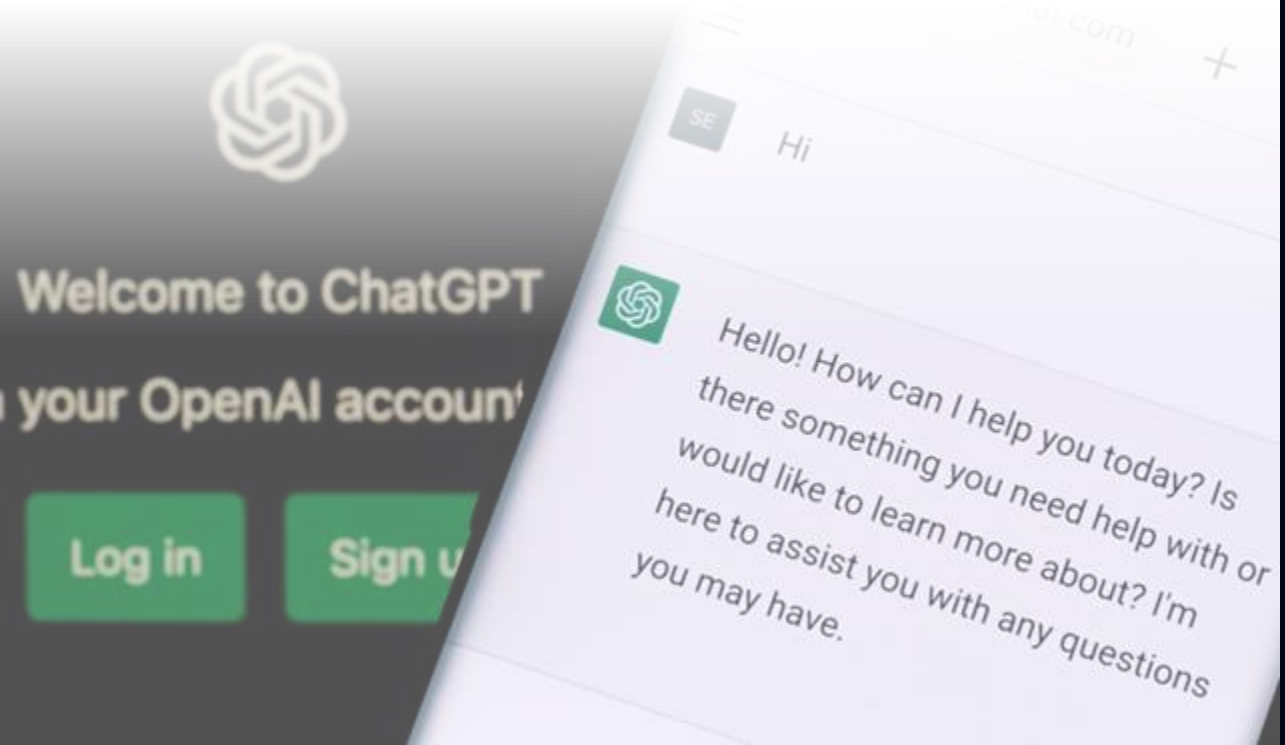
Real-time Anomaly Detection

Multi-Agent AI

tomtom KPMG asos MasterCard

ChatGPT scales with Azure Cosmos DB

Open AI stores ChatGPT conversations and all other user interactions in Azure Cosmos DB, 40+ workloads



Hi

Hello! How can I help you today? Is there something you need help with or would like to learn more about? I'm here to assist you with any questions you may have.

OpenAI

Challenge

- Meet incredible demand from traffic spikes, without having to worry about database operations

Outcomes

- Rapidly and seamlessly scaled as service grew, with zero downtime
- Able to iterate fast on data shapes thanks to schemaless flexibility
- Maintained high performance and availability

Key Azure products used

Azure Cosmos DB

Azure Kubernetes Service

Azure AI Search

GenAI use cases with Azure Cosmos DB

What	Why	When

GenAI use cases with Azure Cosmos DB

What	Why	When
Vector + Operational Database	No ETL Consistent data Reduce complexity & costs	Data & vectors together Cosmos DB scale & performance

GenAI use cases with Azure Cosmos DB

What	Why	When
Retrieval Augmented Generation (RAG)	Personalize LLM on your data Cheaper than fine tuning Faster iteration on new data	Any workload for GenAI apps
Vector + Operational Database	No ETL Consistent data Reduce complexity & costs	Data & vectors together Cosmos DB scale & performance

GenAI use cases with Azure Cosmos DB

What	Why	When
Chat History	Conversational context UX improvements LLM optimizations Auditing	A MUST for Chat sessions Improving cost & performance
Retrieval Augmented Generation (RAG)	Personalize LLM on your data Cheaper than fine tuning Faster iteration on new data	Any workload for GenAI apps
Vector + Operational Database	No ETL Consistent data Reduce complexity & costs	Data & vectors together Cosmos DB scale & performance

GenAI use cases with Azure Cosmos DB

What	Why	When
Semantic Caching	Drastically reduces latency Saves on Token consumption Reduces costs and latency for LLM	Slow moving / static content FAQs, Policies…
Chat History	Conversational context UX improvements LLM optimizations Auditing	A MUST for Chat sessions Improving cost & performance
Retrieval Augmented Generation (RAG)	Personalize LLM on your data Cheaper than fine tuning Faster iteration on new data	Any workload for GenAI apps
Vector + Operational Database	No ETL Consistent data Reduce complexity & costs	Data & vectors together Cosmos DB scale & performance

Vector Search in Azure Cosmos DB for NoSQL

Vector Search in Azure Cosmos DB for NoSQL

Store data + vectors together

Reduced Complexity & Cost
Transactional Data & Vectors
Optimized for App Developers

Vector Search in Azure Cosmos DB for NoSQL

Store data + vectors together

Reduced Complexity & Cost
Transactional Data & Vectors
Optimized for App Developers

Vector Search + Query Filters

Combine with equality, range & spatial filters
Optimize query focus

Vector Search in Azure Cosmos DB for NoSQL

Store data + vectors together

Reduced Complexity & Cost
Transactional Data & Vectors
Optimized for App Developers

Vector Search + Query Filters

Combine with equality, range & spatial filters
Optimize query focus

Flexible Indexing

Flat, quantized flat, and DiskANN indexing available

Vector Search in Azure Cosmos DB for NoSQL

Store data + vectors together

Reduced Complexity & Cost
Transactional Data & Vectors
Optimized for App Developers

Vector Search + Query Filters

Combine with equality, range & spatial filters
Optimize query focus

Flexible Indexing

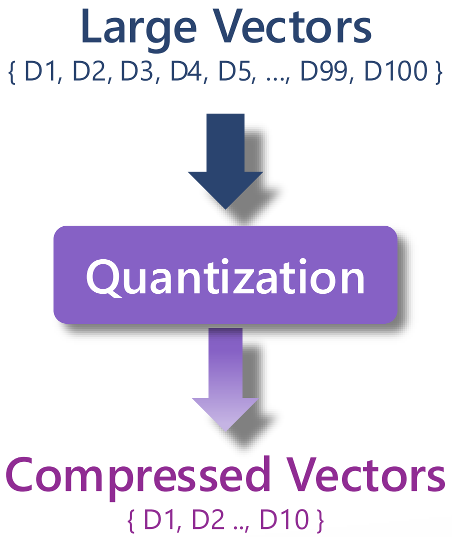
Flat, quantized flat, and DiskANN indexing available

Azure Cosmos DB for NoSQL Capabilities

Serverless or provisioned throughput
Built-in multitenancy
Instant & dynamic autoscale
<10ms point-reads
Globally-replicated
Industry-leading 99.999% SLA

Multi-modal database with built-in DiskANN*

Vector compression



Large Vectors

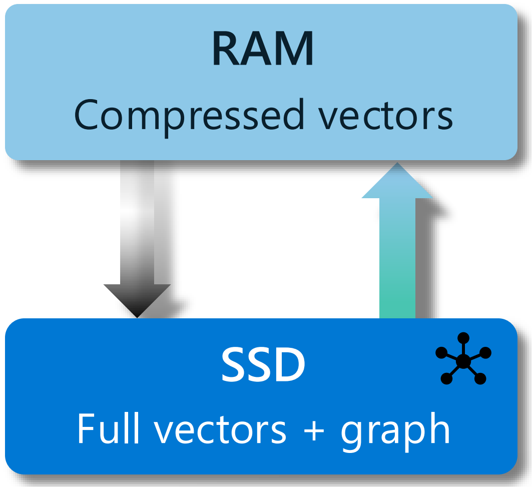
{ D1, D2, D3, D4, D5, ..., D99, D100 }

Quantization

Compressed Vectors

{ D1, D2 .., D10 }

Storage and graph construction



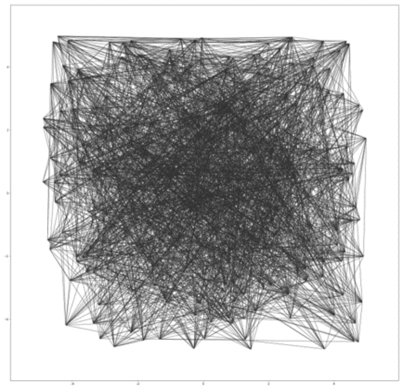
RAM

Compressed vectors

SSD

Full vectors + graph

Algorithms



Unlimited scale | Low latency | Robust to data changes | Serverless

RAG with Azure Cosmos DB & LangChain.js

Demo – aka.ms/ai/js/chat

Azure AI Advantage free offer

Up to $6,000 Azure Cosmos DB free for 90 days[1]

Eligibility: customers using Azure AI Services or GitHub Copilot

Why Azure Cosmos DB for Era of AI

- AI ready
- Guaranteed performance and scale
- Flexibility and efficiency
- Mission critical

Learn more: Aka.ms/AzureAIAdvantageBlog

**Terms & conditions: Offer available to all Azure customers who are using or will use Azure AI Services (including Azure OpenAI Service), Azure Machine Learning, Azure AI Search) and/or GitHub Copilot. Offer provides 40,000 RU/s for free for a 90 day period. Offer only available where Azure Cosmos DB available. Customers must create a new Azure Cosmos DB account and register via Microsoft Form at https://aka.ms/AzureAIAdvantageSignupForm. Time period: the last day for customers to opt-in for the offer via the signup form is June 30, 2024.*

Resources

• Vector search with Cosmos DB	aka.ms/CosmosVectorSearch
• Docs, demos, videos and more	aka.ms/CosmosAISamples
• Serverless AI chat demo	aka.ms/ai/js/chat

Microsoft