Microsoft

# Building generative AI applications with Azure Cosmos DB for MongoDB vCore and Java

Manish Sharma
Principal Program Manager

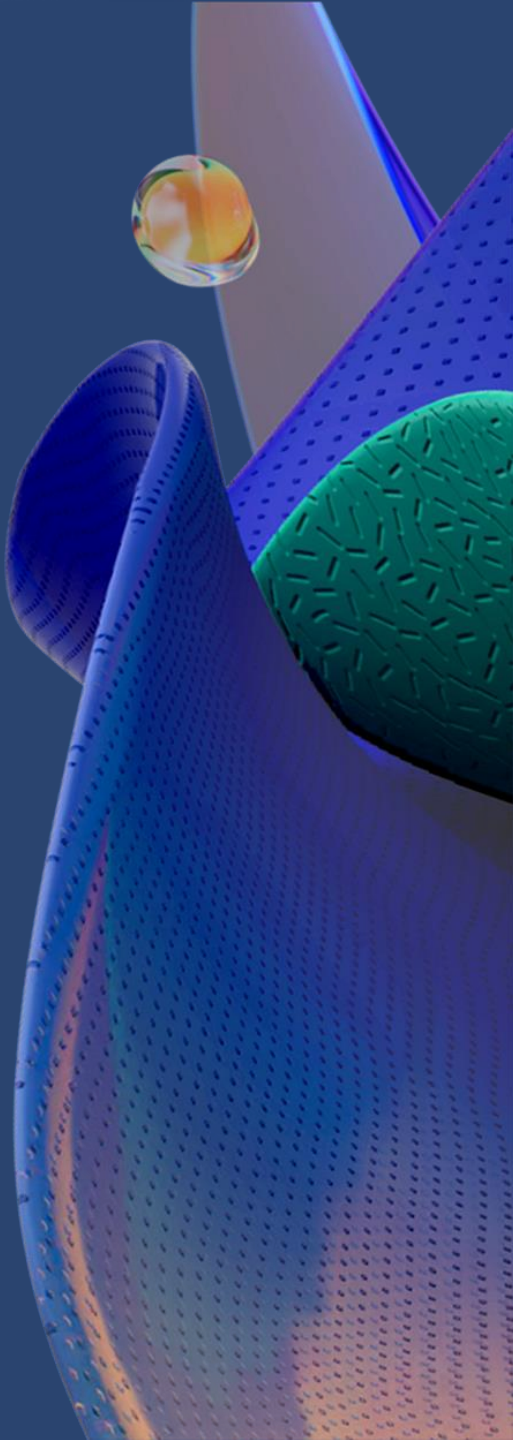
## Quick Intro

- Principal PM in Azure Cosmos DB Engg.
- Working in Azure Cosmos DB since 5+yrs
- Specializes in Graph Database (Analytical + Operational)

- Expertise in App (Java & .NET) & operational database (NoSQL & RDBMS).
  - *Still a nerd*

- AWS & Azure Certified Architect.
- Author of Azure Cosmos DB for Mongo DB developers published by Apress.
- Based in IDC Noida.
- 19+yrs of experience in Industry
- M.Tech from BITS Pilani,
- Operation & Analytical Graph database

Everyone is talking about AI.
What can you do with it?



# No matter where you are in your cloud journey, you can modernize and build new intelligent apps with Azure

1. Connected smart products
2. Transaction processing at scale
3. Real-time fraud detection
4. Service & support bot
5. Information & product discovery
6. Personalize & recommend

Build your own copilots

# Vector indexes supported by Azure Cosmos DB and Azure Databases for PostgreSQL today

## IVF (Inverted File Index)

- Partitions vectors into clusters and assigns each vector to one cluster.
- **Building the index is fast and memory-efficient**
- Requires a separate clustering step before indexing (slow)
- **Tuning parameters is important.** Can be very accurate if configured properly

## HNSW (Hierarchical Navigable Small World)

- Builds a multi-layer graph with long and short connections between the vectors.
- **Robust and accurate at scale**
- No-preprocessing step.
- **Can support many inserts/deletes efficiently.**
- **Larger memory footprint**
- It also has many parameters (such as the number of layers and neighbors) that need to be tuned carefully.

## DiskANN

- Uses a combination of in-memory and on-disk storage to index vectors.
- Can handle very large datasets that don't fit in memory.
- Building the index can be slower than in-memory indexes.
- Can be very accurate if configured properly.
- It also has many parameters (such as the number of trees and neighbors) that need to be tuned carefully.

## Concepts – Vector Embeddings

- **Vector embeddings** are compact, semantically-rich representations of any data
- Vectors that are “close” are semantically similar
- Closeness is measured by distance (cosine, dot product, Euclidean, etc.)
- Easy to generate embeddings from your data via APIs (OpenAI, Hugging Face, etc.)

### Use cases

Answering Questions

Detecting anomalies

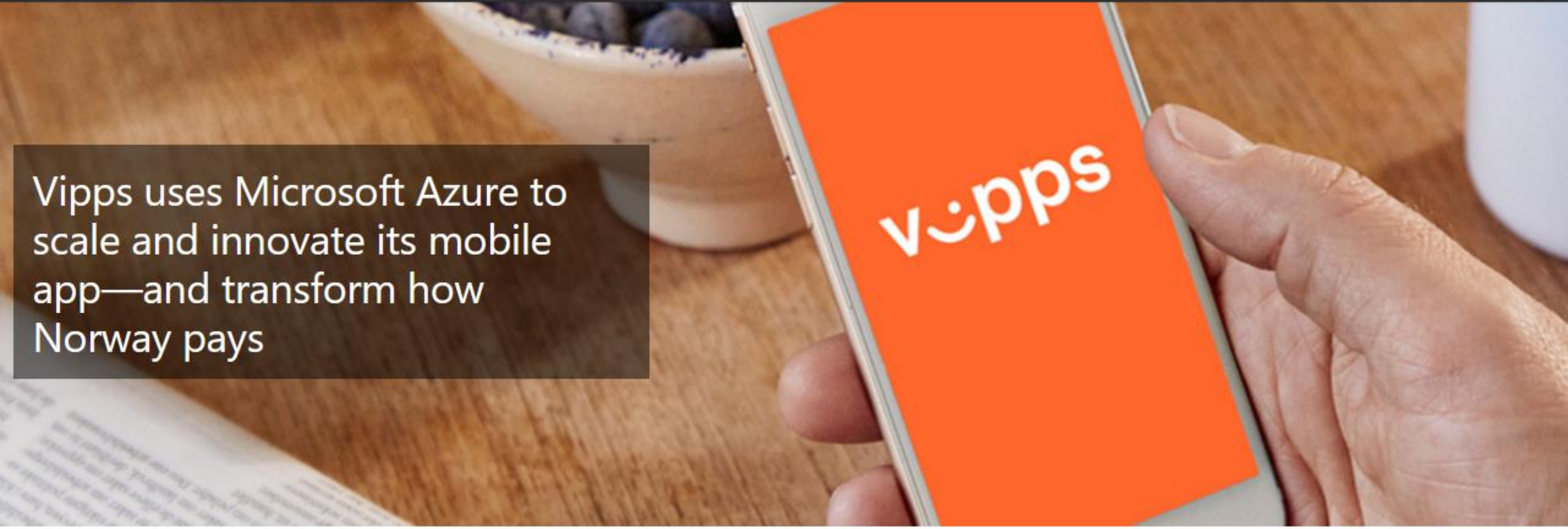
Making personalized recommendations

Searching for similar content

## Why should I care about Chat History?

- **Memory.** Short-term or Long-term memory for your LLM conversations to bring rich semantic context from chat histories.
- **User analysis.** Better understand your users and their interests
- **LLM improvement.** Meta-prompt engineering, fine-tuning, etc.
- **Semantic Caching.** Reduce cost at latency by reducing calls to LLMs
- **Auditing.** Maintain a historical account of LLM interactions

https://aka.ms/CDBSemanticCache


Vipps uses Microsoft Azure to scale and innovate its mobile app—and transform how Norway pays
vipps

## Customers have used Azure Cosmos DB for chat for years

Sign on, say hi

Collaborate with your coworkers through a Microsoft Teams meeting. Azure Cosmos DB enables consistent reliability on a massive scale so work can happen anywhere.

## Even some really big “Teams” use Cosmos DB for chat

# OpenAI is built on Azure Cosmos DB

Your AI-powered apps can be too

## OpenAI use Java and Python

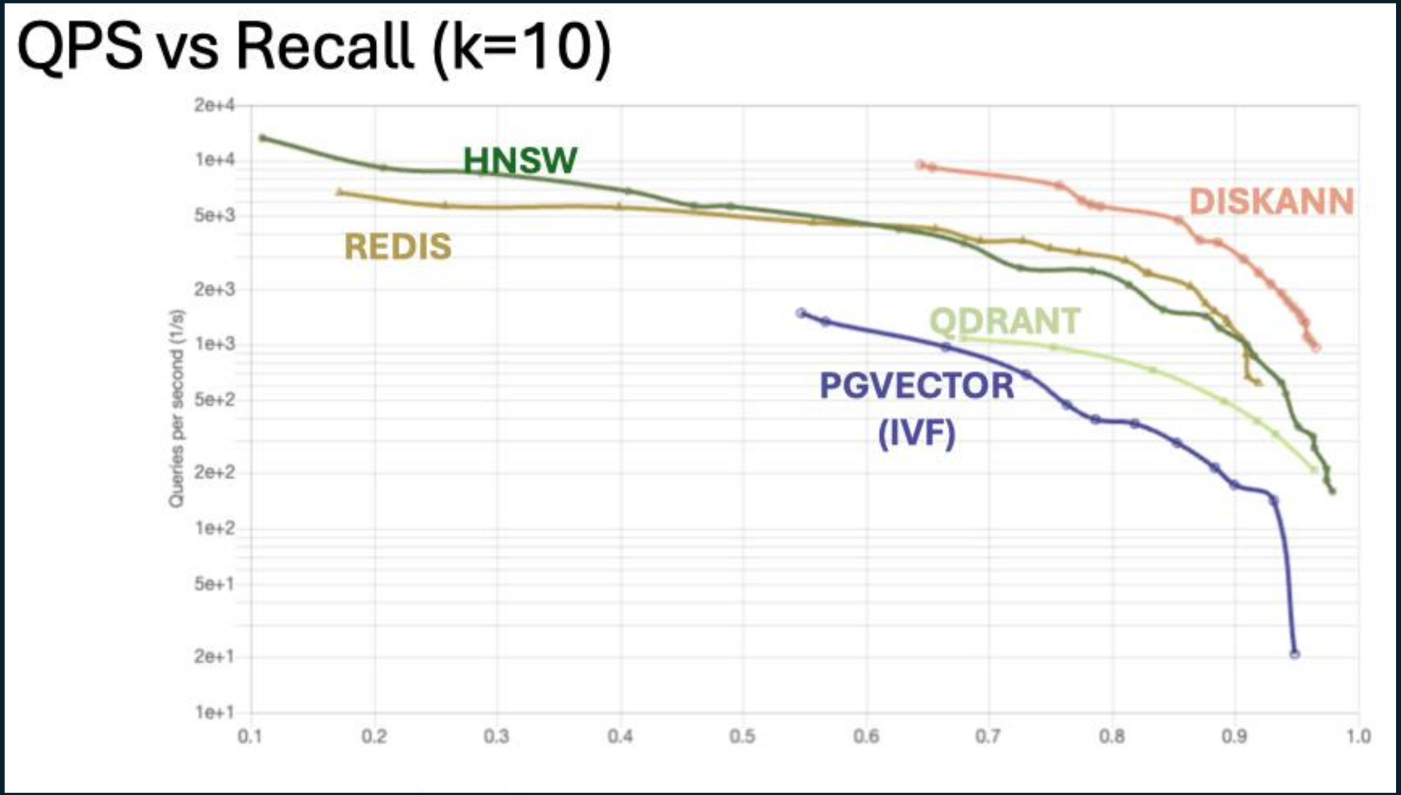
# Vector Search

**Today:**

- Native vector search using Azure Cosmos DB for MongoDB vCore, PostgreSQL, and Cassandra Managed Instance
- Integrated with Azure Cognitive Search for Azure Cosmos DB NoSQL

**Coming soon:**

- Native vector search for Azure Cosmos DB NoSQL using DiskANN.
- High performance and elasticity, great for multi-tenant apps



# Azure Cosmos DB for MongoDB vCore

## New Additions

- Free tier w/ 32GB storage
- Burstable SKUs
- New cluster tiers & storage SKUs
- Private link
- Migration from MongoDB

## AI Ready

- Native Vector Search, including HNSW
- Plugins: LangChain, Semantic Kernel, and LlamaIndex (like Hibernate but for LLMs and embedding stores)
- Integration with Azure OpenAI Studio https://aka.ms/CosmosMongoVcoreAIStudio

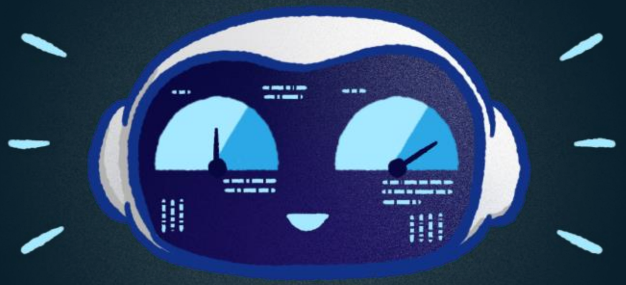
**Learn more:** aka.ms/tryvcore

## KPMG KymChat

AI agent to streamline KPMG employee operational tasks.

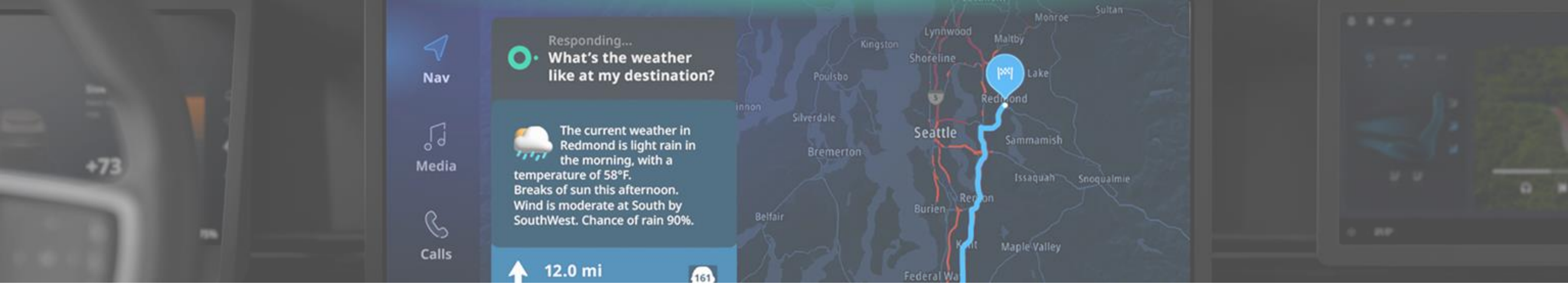
Leveraging Vector Search in Azure Cosmos DB for MongoDB vCore enabled KPMG to provide value to their employees at scale.

**Accurate**
PCI, a key relevancy metric increased from **50% to 90%+**



**Performance**
7,000+ employee issuing 120,000+ requests for up to 50% productivity gain

**Scalable**
Performance improvements enabled rollout to all KPMG member firms



Responding...
What's the weather like at my destination?

The current weather in Redmond is light rain in the morning, with a temperature of 58°F. Breaks of sun this afternoon. Wind is moderate at South by SouthWest. Chance of rain 90%.

12.0 mi

# TomTom brings AI-powered, talking cars to life with Azure

## Key Azure products used:

- Azure Cosmos DB
- Azure Kubernetes Service
- Azure Open AI Service

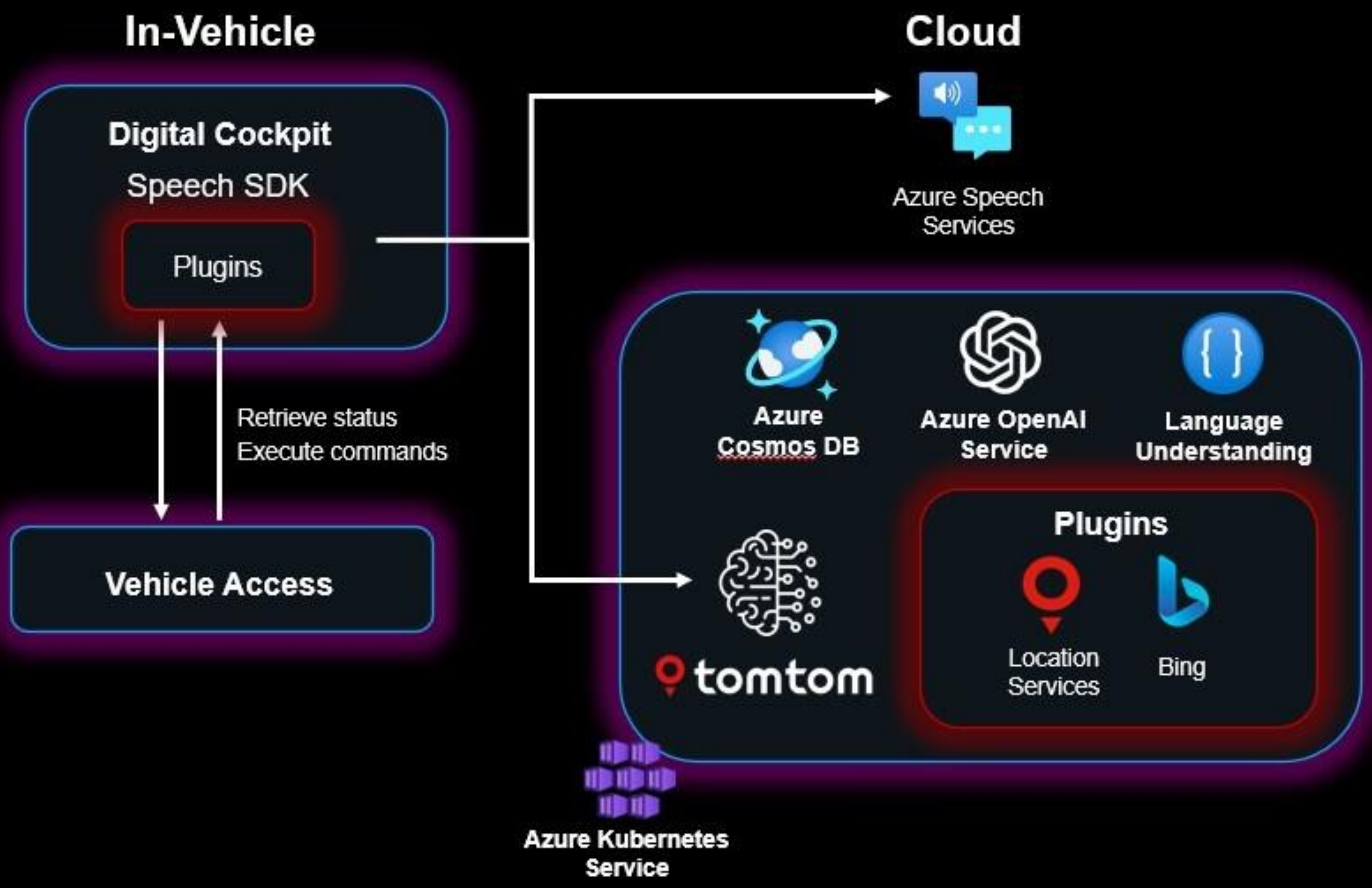
## Challenge

- TomTom wanted to transform the everyday in-car driving experience. They sought to enable drivers to communicate in a natural way with their vehicles to help them seamlessly achieve tasks like opening windows, finding routes, making reservations, and more.

## Solution

- TomTom used GenAI to create their Digital Cockpit, a conversational automotive assistant enabling voice interaction with infotainment, location search, and vehicle commands.
- Using Azure OpenAI Service, Azure Cosmos DB, and AKS, they built an intelligent, fast, and highly scalable AI chatbot copilot that can be integrated into other automotive systems.
- Digital Cockpit can respond to 95% of complex driver requests, helping drivers accomplish more.

## TomTom Digital Cockpit Architecture



# Azure Cosmos DB for MongoDB vCore integrations

**Semantic Kernel** – Vector database semantic-kernel

**LlamaIndex** – Vector database LlamaIndex

**LangChain** – Vector database & semantic caching LangChain

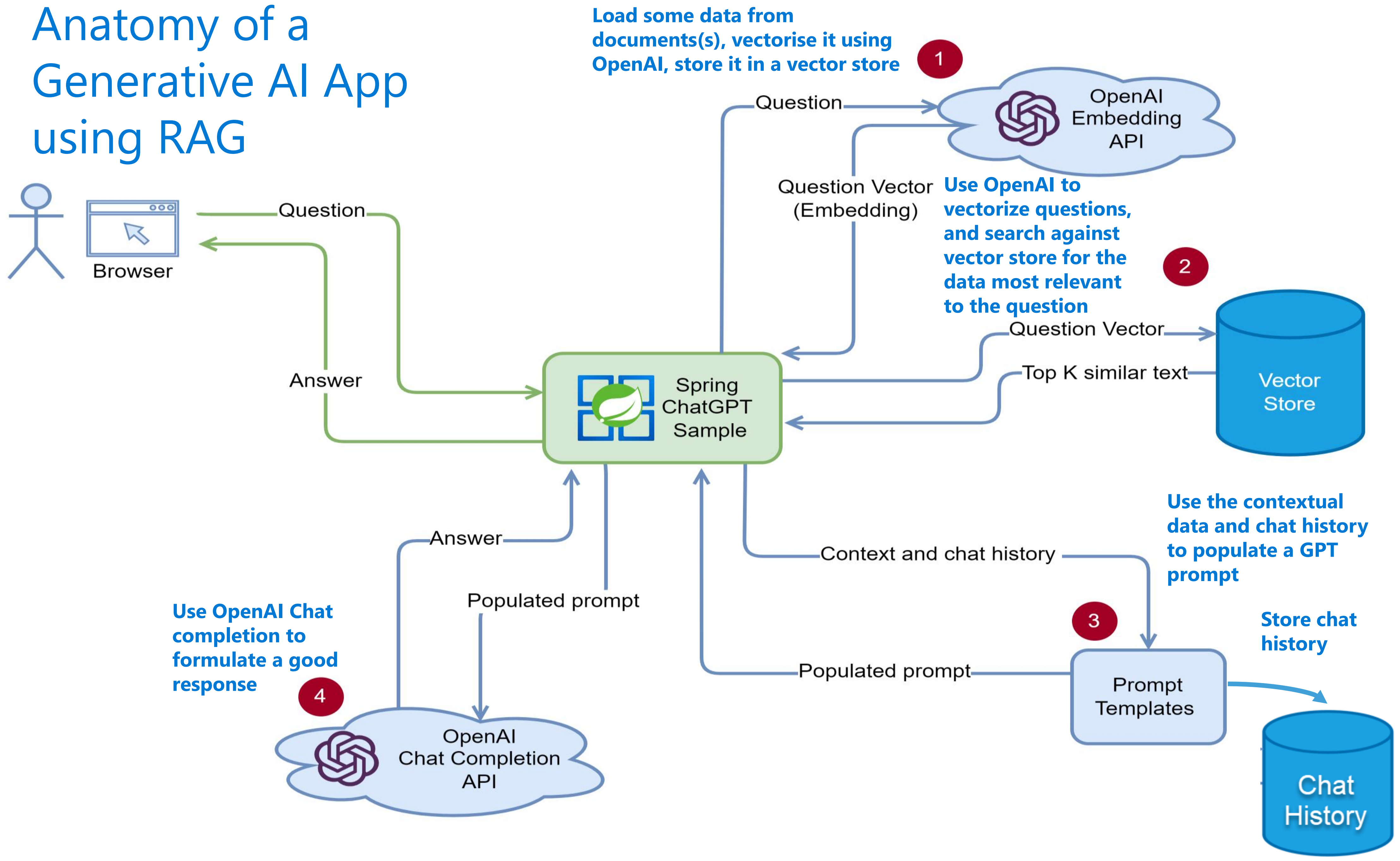
**Azure OpenAI** "on your data" – Vector database Azure

## Demos…

## Use your own data with Azure Cosmos DB for MongoDB vCore & Azure OpenAI Service in a simple app

**Aka.ms/CosmosMongoVectorSearchDemoJava**

# Anatomy of a Generative AI App using RAG


Load some data from documents(s), vectorise it using OpenAI, store it in a vector store
1
Question
OpenAI Embedding API
Question Vector (Embedding)
Use OpenAI to vectorize questions, and search against vector store for the data most relevant to the question
2
Question Vector
Top K similar text
Vector Store
Browser
Question
Answer
Spring ChatGPT Sample
Answer
Populated prompt
Context and chat history
Use the contextual data and chat history to populate a GPT prompt
3
Store chat history
Populated prompt
Prompt Templates
Chat History
Use OpenAI Chat completion to formulate a good response
4
OpenAI Chat Completion API

# 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

**Azure AI Advantage Offer entitles customers to up to 40,000 Request Units per second for free for 90 days. This is the equivalent of up to $6,000 in savings.*

## Get started skilling with AI on Microsoft Learn

Build AI skills, connect with the community, earn Microsoft Credentials, learn from experts, and take the Cloud Skills Challenge.

aka.ms/LearnAtAITour

## Learn More

| | |
|---|---|
| Azure Cosmos DB for Mongo vCore Free tier: | **Aka.ms/tryvcore** |
| Cosmos Mongo Vector Search Java demo: | **Aka.ms/CosmosMongoVectorSearchDemoJava** |
| ACME Fitness Store demo: | **Aka.ms/FitnessStoreDemo** |
| Azure MI for Apache Cassandra Vector Search: | **Aka.ms/CassandraMIVectorSearch** |
| Microsoft Copilot for Azure in Cosmos DB: | **Aka.ms/CopilotForAzureInAzureCDBBlog** |
| Azure AI Advantage: | **Aka.ms/AzureAIAdvantageBlog** |

## Learn More

- Vector Databases on Cosmos DB - https://aka.ms/CosmosDBVectorSearch
- RAG Pattern Samples (Python) - https://aka.ms/RAGwithCosmosDB
- Langchain Semantic Caching  (Python) - https://aka.ms/CDBSemanticCache
- Langchain Vector Store (Python) – https://aka.ms/CosmosDBLangChain
- Semantic Kernel Vector Store (Python) – https://aka.ms/CosmosDBSemanticKernel
- Azure OpenAI Studio Integration - https://aka.ms/bring-cosmosdb-to-openai
- Solution Accelerator (C#) - https://aka.ms/BuildModernAiAppsSolution
- Chat Application Demo - https://aka.ms/cosmos-chatgpt-sample
- RAG with Cosmos DB + Langchain - https://aka.ms/RAGwithCosmosDB
- DiskANN GitHub repo

Microsoft

## Thank You