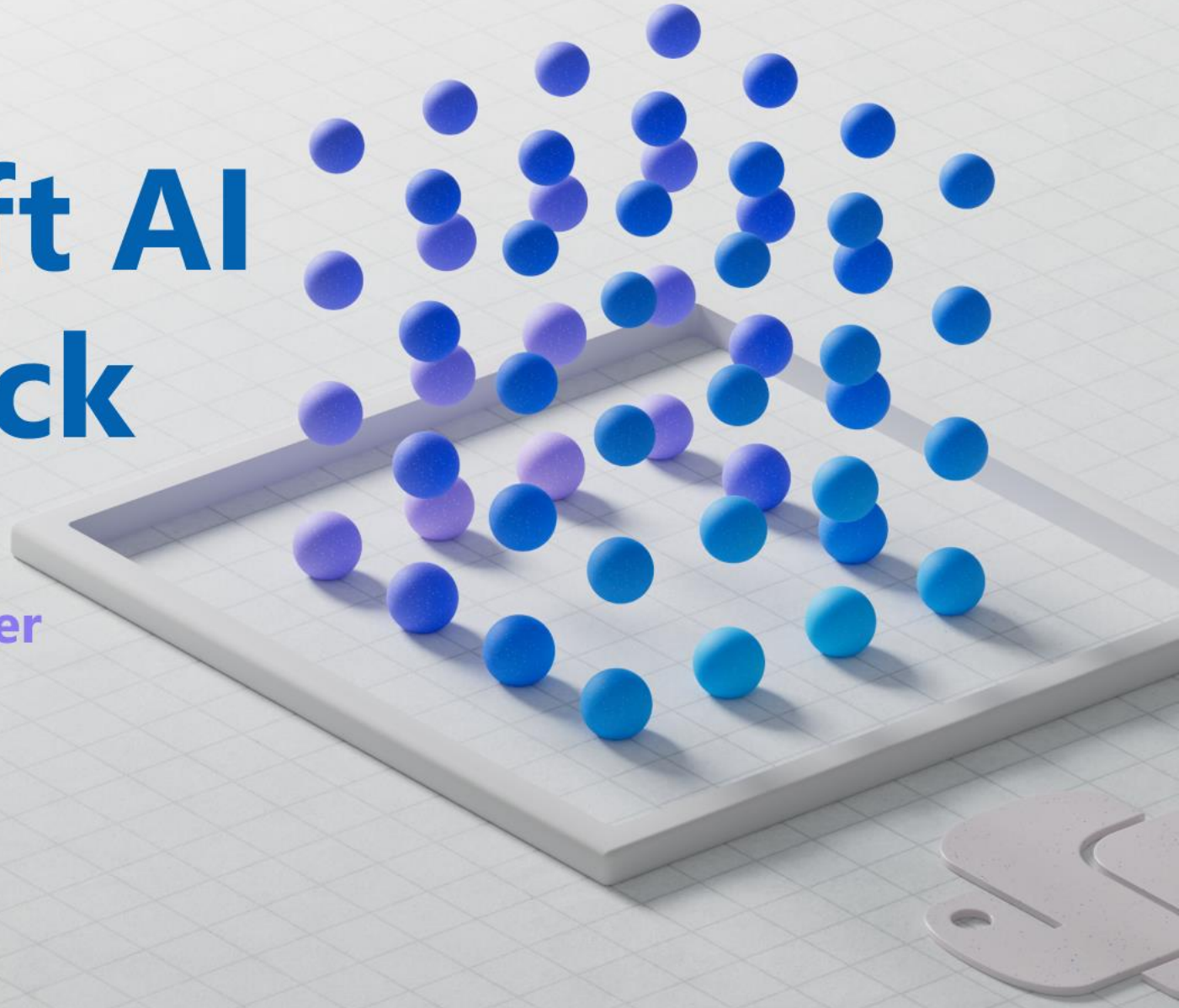
January 29th - February 12th

The Microsoft AI Chat App Hack



Build, innovate, and #HackTogether

aka.ms/hacktogether/chatapp

The AI Chat App Hack

Microsoft

January 29th - February 12th

29th
Building a RAG Chat App in Python

Connecting a RAG Chat App to Azure Cosmos

30th
Customizing your RAG Chat App

31st
Azure AI Search Best Practices

1st
GPT-4 with Vision

2nd
HACK HACK HACK

3rd
HACK HACK HACK

4th
HACK HACK HACK

5th
AM: RAG Chat Web Components

PM: Access Control in RAG Chat Apps

6th
Evaluating a RAG Chat App

7th
RAG Chat Special Topic

8th
Continuous Deployment of your Chat App

9th
HACK HACK HACK

10th
HACK

11th
HACK HACK

12th
Hack Together Project Showcase

Build, innovate, and #HackTogether

Microsoft

Connecting a RAG Chat App to Azure Cosmos DB

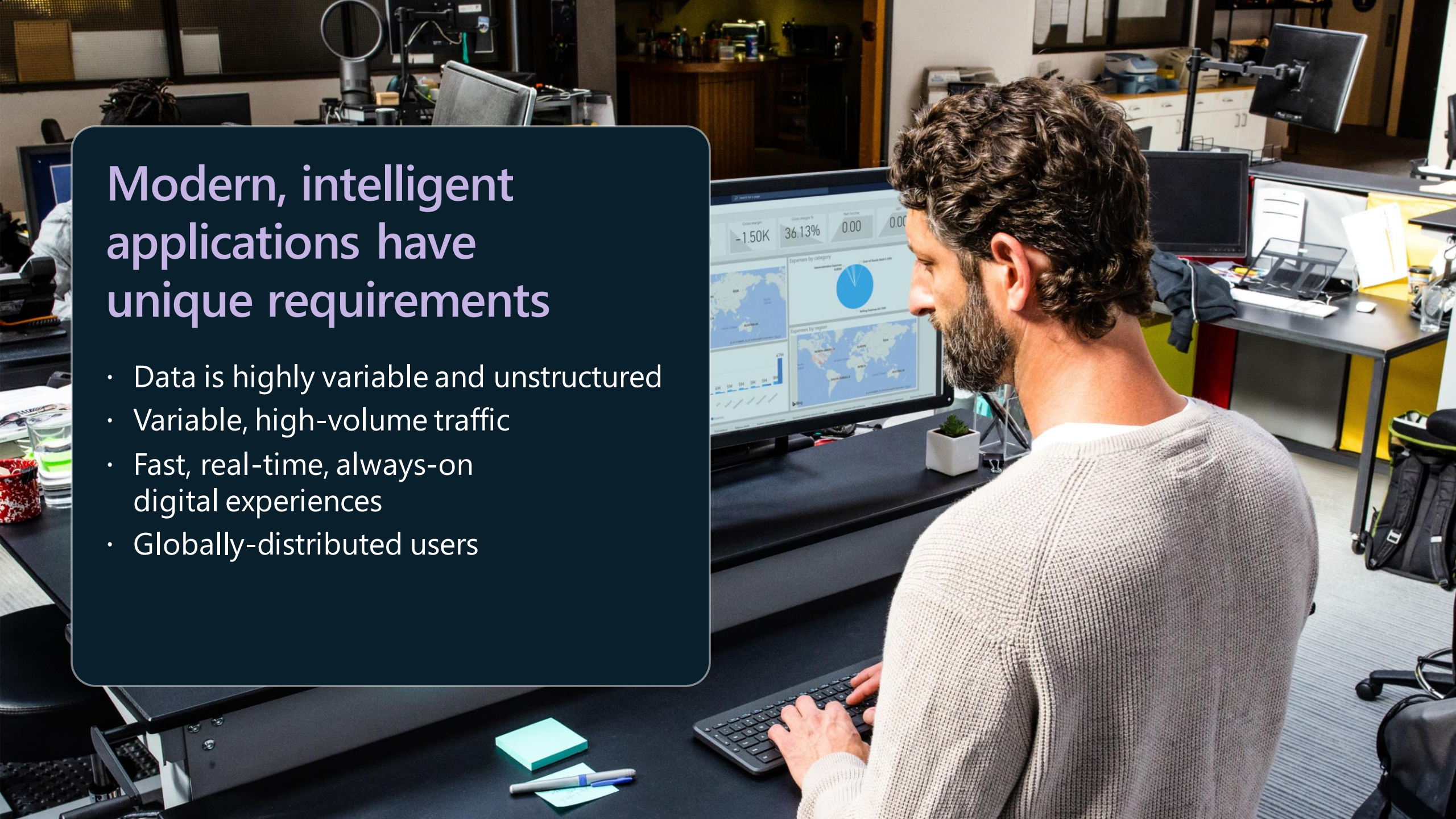
Khelan Modi
Product Manager

Agenda

- Why Azure Cosmos DB?
- Concepts
- Azure Cosmos DB for MongoDB vCore
- Demos
- Links
- Q&A

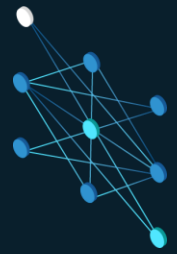
Modern, intelligent applications have unique requirements

- Data is highly variable and unstructured
- Variable, high-volume traffic
- Fast, real-time, always-on digital experiences
- Globally-distributed users



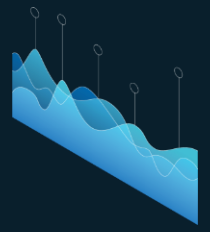
Azure Cosmos DB does it all

Build AI assistants and intelligent cloud-native apps with Azure Cosmos DB

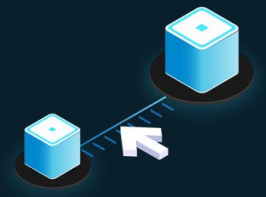


AI ready

Guaranteed performance and scale



Flexibility and efficiency



Mission-critical

Azure Cosmos DB is AI ready

- All-in-one Solution
- Save cost and complexity
- Real-time AI
- Highest fidelity with Azure Services
- Built-in vector search
  - Native support for MongoDB vCore and PostgreSQL APIs
  - Integrated with Azure Cognitive Search for core NoSQL API

Coming soon: native vector search for core NoSQL API

High performance and elasticity, great for multi-tenant apps

OpenAI is built on Azure Cosmos DB

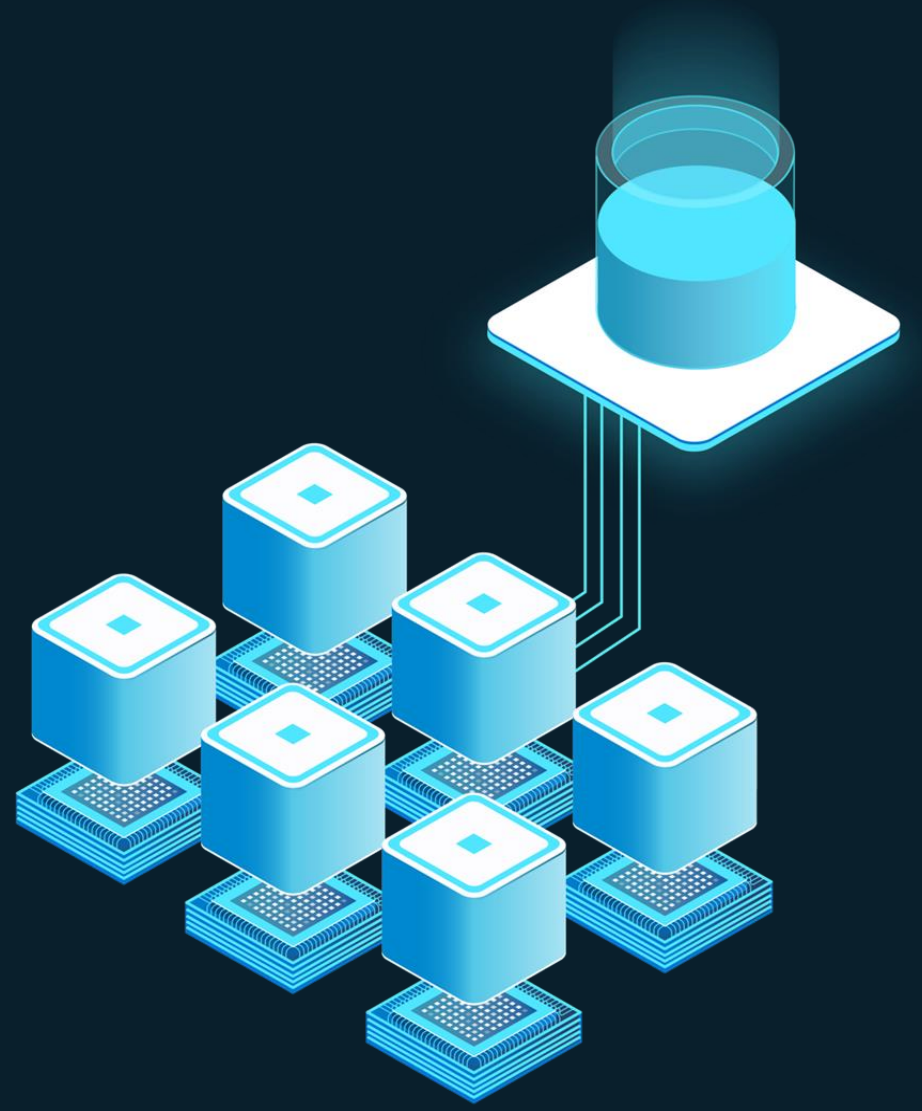
Your AI-powered apps can be too!

Concepts

- Retrieval Augmented Generation (RAG)
- Vector Embeddings & Vector Search
- Vector Indexes: IVF & HNSW

Concepts – Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) intelligently retrieves a subset of data from data stores to provide specific, contextual knowledge to the large language model to support how it answers a user's prompt.



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

Vector indexes supported by Azure Cosmos DB

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.

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
- Integration with Azure OpenAI Studio

Learn more: aka.ms/tryvcore

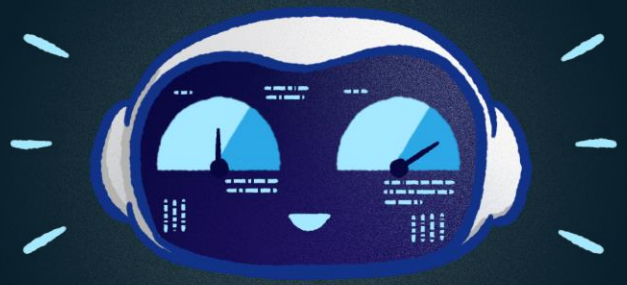
KPMG KymChat

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%+**

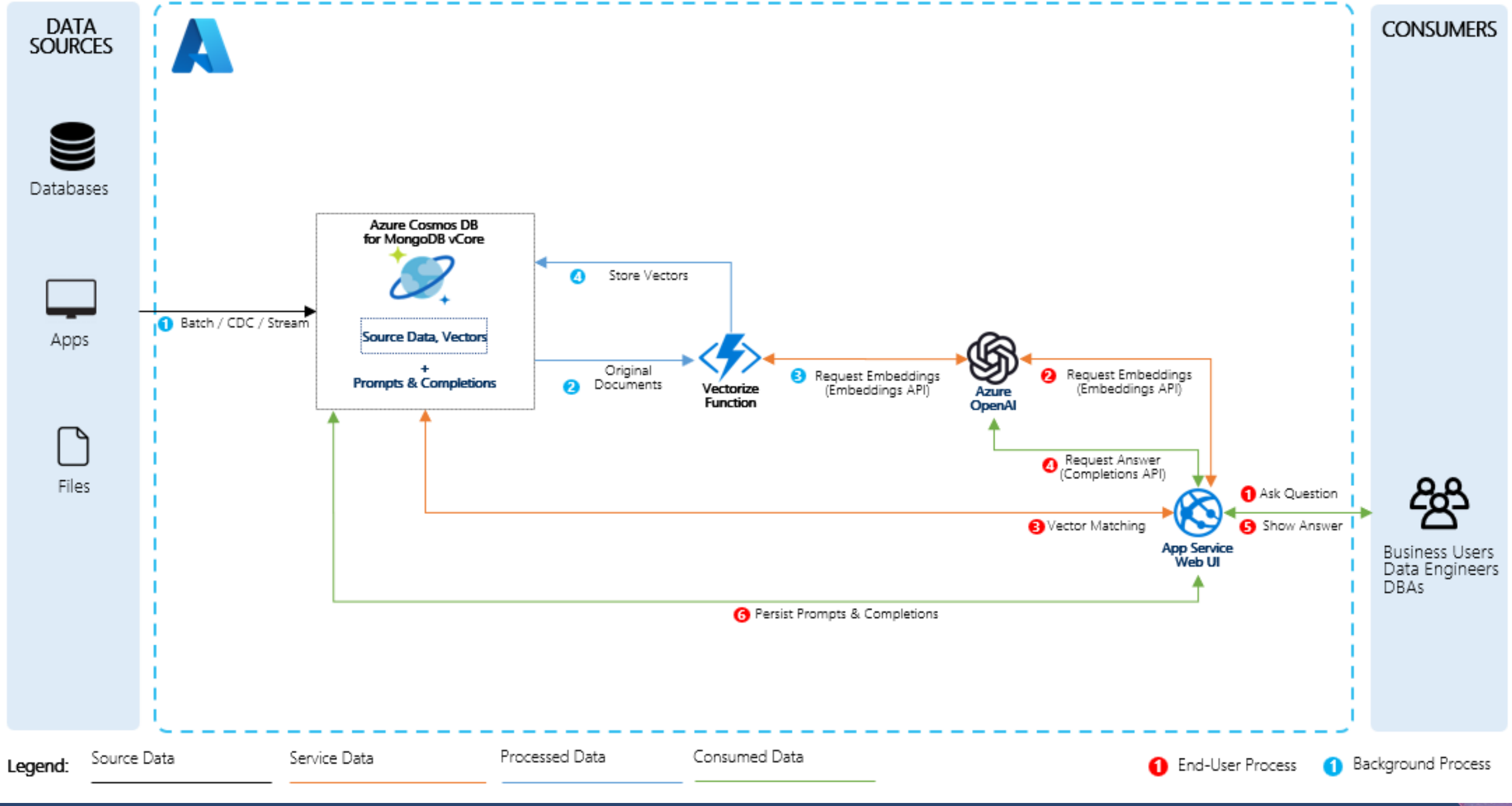


Performance

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

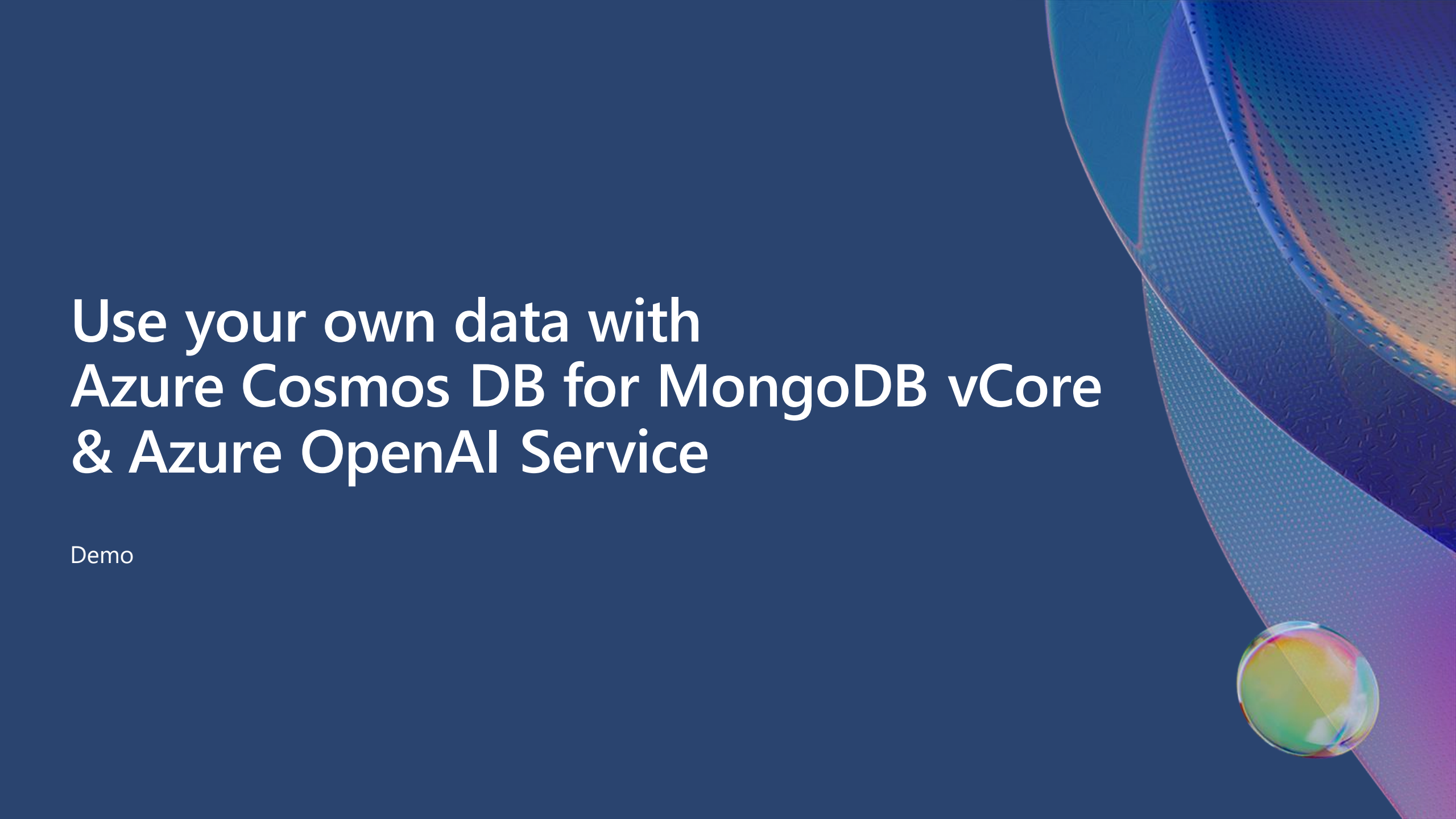
Scalable

Performance improvements enabled rollout to all KPMG member firms


DATA SOURCES
Databases
Apps
Files
Azure Cosmos DB for MongoDB vCore
Source Data, Vectors
+
Prompts & Completions
1 Batch / CDC / Stream
4 Store Vectors
2 Original Documents
Vectorize Function
3 Request Embeddings (Embeddings API)
Azure OpenAI
2 Request Embeddings (Embeddings API)
4 Request Answer (Completions API)
1 Ask Question
5 Show Answer
3 Vector Matching
App Service Web UI
6 Persist Prompts & Completions
CONSUMERS
Business Users
Data Engineers
DBAs
Legend:
Source Data
Service Data
Processed Data
Consumed Data
1 End-User Process
1 Background Process

Use your own data with Azure Cosmos DB for MongoDB vCore & Azure OpenAI Service

Demo



'R' of RAG using Azure Cosmos DB for MongoDB vCore

Demo

Scenario guidance: Azure AI Search vs. Azure Cosmos DB

Azure AI Search

Key value proposition

Highest quality results out of the box.

Example scenario

Search and Knowledge Management for enterprise data across SharePoint, Data Lake, blob storage and databases*.

When to use

Relevance: offers **most relevant results** (via ranking and hybrid search) and **highest, most premium capabilities**

Distributed: Azure AI Search may be ideal when data is distributed across multiple databases

Azure Cosmos DB

Key value proposition

Operational efficiency: no data movement required. Native, built-in vector search capabilities at scale.

Example scenario

Transactional applications, such us an eCommerce app with real-time inventory data. Chat history for conversational context and prompt engineering.

When to use

Operational impact: if you **do not want to move your data outside of your operational database**

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.*

Learn More

Azure Cosmos DB for Mongo vCore Free tier:	**Aka.ms/tryvcore**
Chatbot (Wheelie) Demo:	**Aka.ms/MongovCoreAzureAIsample**
AI-advertisement:	**Aka.ms/adgen**
RAG Jupyter Notebook	**Aka.ms/RAGwithCosmosDB**
Azure AI Advantage:	**Aka.ms/AzureAIAdvantageBlog**