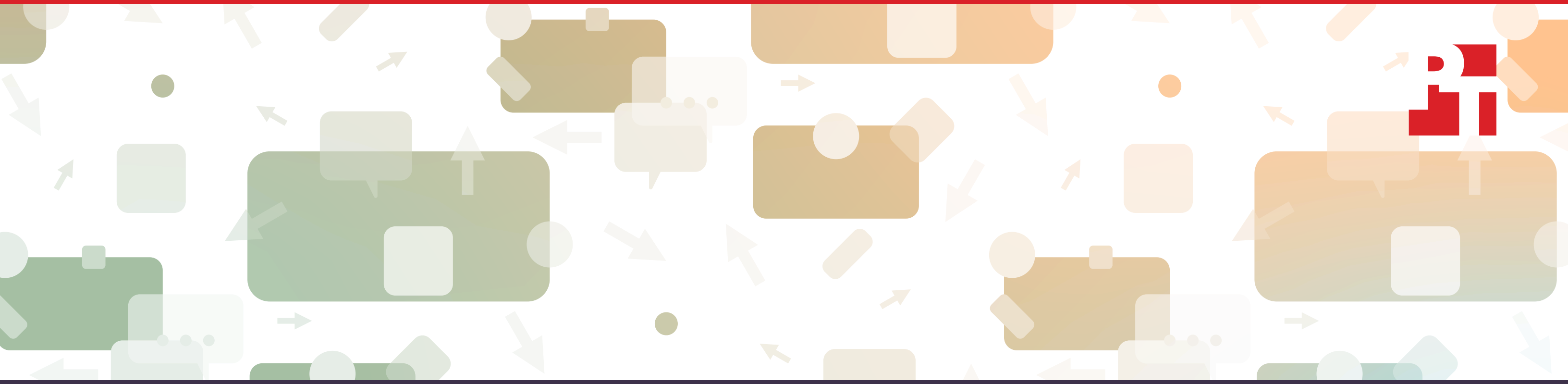
A Principled Technologies report: Hands-on testing. Real-world results.



Build an Azure OpenAI application using your own enterprise data

We built an intelligent AI chat application using Microsoft Azure Cosmos DB for NoSQL and Azure AI Services

Decision-makers across many industries and business sizes are excited about the potential power of a customized chat service using OpenAI to solve a wide range of business problems. Maybe your goal is reducing customer support costs or helping your employees access or generate information more easily. Many organizations in e-commerce, financial services, and beyond are also using Azure Cosmos DB to store corporate data and perform transactions on that data. Those groups can meet their AI needs by building an Azure AI application that draws from Azure Cosmos DB, uses Azure AI Search to categorize your data, and uses Azure's access to OpenAI generative pre-trained transformer (GPT) models to process prompts and return completions with responses relevant to your own data. Not only can Azure Cosmos DB store a large amount of unstructured data related to the primary application, it can store information related to an AI chatbot conversation history, helping to provide later business value.

Azure Cosmos DB is a NoSQL distributed database designed by Microsoft to be operational in nature, enabling low latency responses and powering transaction-driven workloads and also intelligent, AI-powered applications. Unlike a traditional relational model database, Azure Cosmos DB can store unstructured data. By using Azure Cosmos DB to host your data, utilizing Azure AI Search to search and retrieve data relevant to a query, and providing a chat interface to Azure OpenAI, it is possible to set up a retrieval augmented generation (RAG) architecture and avoid having to train the model itself, all while protecting your data.

To demonstrate this, Microsoft engineering developed a sample Vector Search & AI Assistant application using a bicycle shop website as a sample business. Using its web-based chat interface, hypothetical customers of the bicycle shop could gather information on products by asking a sequence of connected questions that the solution uses to deliver better responses.[1] This method improves on the usual search function, which answers one question at a time and cannot link questions. This web-based application stores a customer's questions during a session, uses the questions to provide better prompts to Azure OpenAI, and uses Azure AI Search to act as a retrieval system for the business data stored in Azure Cosmos DB. This application was a good start, but it used a very small database. While Azure Cosmos DB is designed for production use in processing orders or other transactional workloads, this sample application did not have those e-commerce features.

To address those limitations, Principled Technologies (PT) extended this application by using a custom, much larger Azure Cosmos DB dataset. See the science behind the report for more details on how we extended the application, and read on to learn how you could benefit from building such an application yourself.

From OpenAI ChatGPT to a private Azure OpenAI application with your own Azure Cosmos DB data

Perhaps the most widely discussed AI tool of 2023 is ChatGPT, the free large multimodal chatbot from the OpenAI organization. Most Americans have heard of ChatGPT,[2] and as of November, it had over one hundred million weekly users.[3] People have utilized it for everything from job applications, essay outlines, and research projects to personal letters and Dungeons & Dragons campaigns. OpenAI currently offers two versions of ChatGPT: GPT-3.5 and a paid upgrade to GPT-4, which, according to OpenAI, "exhibits human-level performance on various professional and academic benchmarks."[4]

GPT holds countless possibilities for business, especially for organizations that choose to build an application using their own data. Companies such as Morgan Stanley and Khan Academy have benefited from putting GPT to work.[5] By training an AI application with your organization's Azure Cosmos DB data, you can create a powerful chat, search, and research tool tailored to your specific history, culture, and needs. In addition to these use cases, which mostly assume an existing data lake paradigm, the advantage of using Azure Cosmos DB in a scenario such as the one we tested is that it can work with your operational and transactional data in near real time.

One such real-world example of this use case is KPMG's use of Azure Cosmos DB for MongoDB vCore with vector search and Azure OpenAI (ChatGPT) to build KymChat, its internal conversational AI assistant. The company used this combination of services to speed response times and increase user satisfaction, and KymChat search quality jumped from 50 to 91 percent, with results being delivered in under a second.[6]

Consider a large and growing retail chain with physical stores as well as an e-commerce presence. With this kind of AI application, the chain could offer friendly, realistic customer support via chatbot across time zones and outside normal business hours. This has the potential to reduce customer support costs and enable the company to gather more data about customer complaints and concerns. An AI application could also improve the customer experience, not just through better support but also with a more comprehensive, accurate search function online. These improvements are scalable as the business continues to expand their offerings, open more stores, and experience busy periods during holidays or seasonal sales.

Retail is far from the only industry that can benefit from a private AI implementation using Azure services. Let's consider a manufacturing organization—perhaps the company building the products that our imaginary retail chain sells. An Azure OpenAI application personalized to the manufacturing company could allow employees to search for the status of parts and equipment more easily, potentially allowing for easier troubleshooting and time savings. And with access to the vast amount of data generated by Internet of Things (IoT) applications on the factory floor, engineers and product managers could have near-human chat experiences with the company's data.

To take just one more example, imagine a medical research facility. Whether the facility has teams developing new drugs or running research studies, its doctors and employees could use the AI application to easily search through data from past papers and research studies, quickly summarize findings, and even brainstorm solutions to difficult scientific problems.

If you're interested in moving forward with such a solution for your Azure Cosmos DB data, you may be wondering where to start. The possibilities are enormous—and we take a deeper dive into the ways that organizations across different industries can harness the power of this tool in our companion report, Utilizing Azure Cosmos DB for intelligent AI-powered applications. In that companion report, we also discuss specifics of how to tie Azure Cosmos DB data and the various Azure services and application elements together.

In this report, we discuss our proof of concept an intelligent AI-powered application combining services entirely within Azure: Microsoft Azure Cosmos DB, Azure AI Search, and Azure OpenAI.

Azure Cosmos DB at work in retail: ASOS

Azure Cosmos DB supports a wide range of worldwide organizations, including the clothing retailer ASOS. A 2017 customer case study explains that Azure Cosmos DB has helped ASOS offer product recommendations quickly—no small feat when one considers the more than 85,000 items available for purchase on the ASOS website. ASOS has also put Azure Cosmos DB to work on order processing, refunds, and returns.[8]

The intelligent Azure application we extended

Our approach

We set out to build a proof-of-concept of an intelligent Azure chat application, based on the bicycle shop application that Microsoft created, but with a much larger database. To represent the work of a real-world team creating such an application for their business and using custom data, we brought in a new dataset—in our case, a publicly available dataset containing Airbnb website data (property rental listings).

To begin, we engaged with Microsoft and studied their Vector Search & AI Assistant reference application that used a fictional bicycle shop as a sample business. Then, we populated Azure Cosmos DB containers using our Airbnb data.

Next, we set up a user-friendly web chat application and prompted it with questions about the data and the kind of answers we wanted from OpenAI so that the application would behave as a retail assistant. With this RAG architecture, our goal was to incorporate more context into the answers the application supplies to users' questions. The application initiated calls to Microsoft Azure OpenAI and Azure AI Search.

Please see the companion report to this one, Utilizing Azure Cosmos DB for intelligent AI-powered applications, for a more detailed description of how your company can use the reference applications and toolsets.

Solution overview

Our solution comprises four primary components working in concert: Azure Cosmos DB for NoSQL, the web application with which the end user interacts to pose questions and receive answers, Azure AI Search, and Azure OpenAI using the GPT-3.5-Turbo and ADA text embedding models. Figure 1 shows at a high level how the data flows through the solution when a customer makes a request and gets a response.

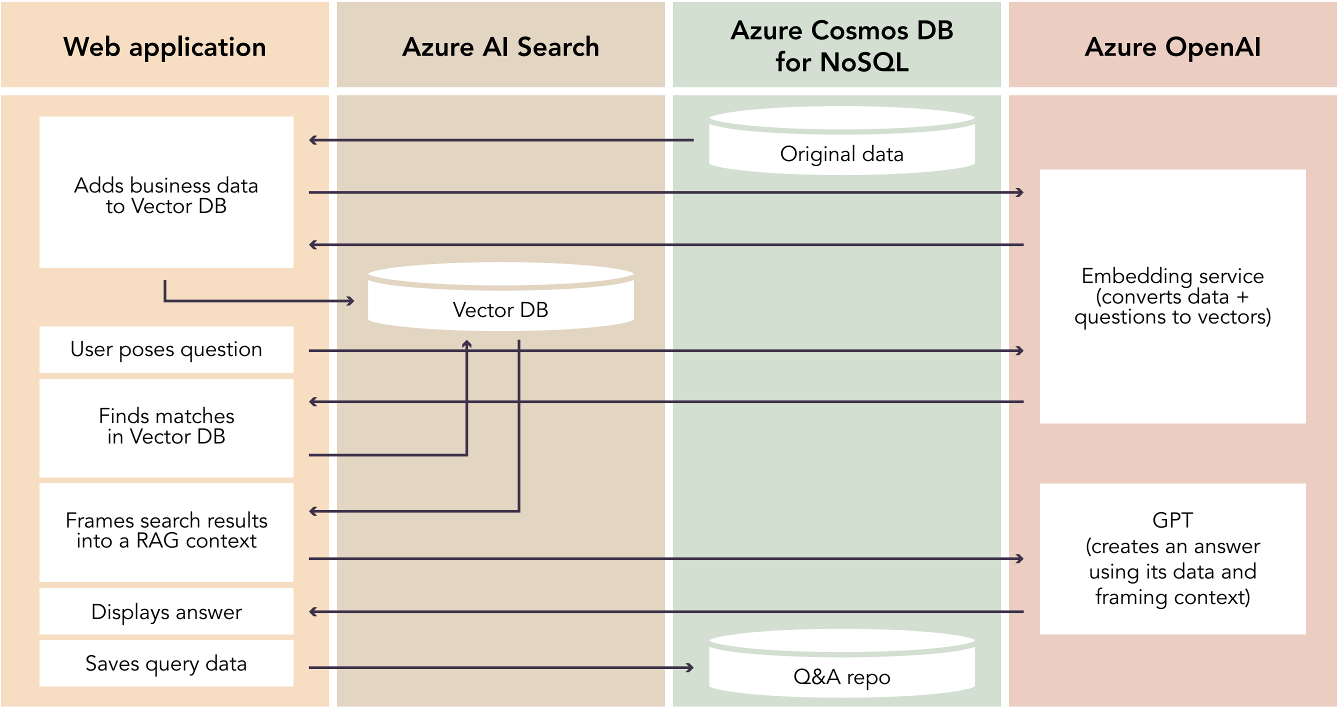

Web application
Azure AI Search
Azure Cosmos DB for NoSQL
Azure OpenAI
Original data
Adds business data to Vector DB
Embedding service (converts data + questions to vectors)
Vector DB
User poses question
Finds matches in Vector DB
Frames search results into a RAG context
GPT (creates an answer using its data and framing context)
Displays answer
Saves query data
Q&A repo

Figure 1: How the data flows through our Azure Cosmos DB-based Azure and GPT-3.5-Turbo solution when a customer makes a request and gets a response. Source: Principled Technologies.

The steps are as follows:

- The app
  - lets the user pose a question
- Azure OpenAI Service
  - receives the raw question
  - converts it to vectors
- Azure AI Search
  - uses vector search to retrieve relevant business data pertaining to the question
  - uses this result to frame the context for answers from Azure OpenAI.
- Azure Cosmos DB for NoSQL
  - uses one container for the Airbnb data set
  - uses one container for a Q&A repository, which receives vectored questions
  - supplies the web app with the customer's previous questions with the current one to create an improved prompt for Azure OpenAI
- The Azure OpenAI GPT-3.5-Turbo model
  - receives the prompt
  - translates the results into an answer in language the user can understand
  - sends the answer to the app
- The app
  - displays the answer for the user
  - sends the answer to the Azure Cosmos DB Q&A repository for storage

About the components of the solution

About Azure Cosmos DB for NoSQL

Azure Cosmos DB is a fully managed and distributed NoSQL relational database for modern app development. Azure Cosmos DB offers "single-digit millisecond response times, automatic and instant scalability, along with guaranteed speed at any scale. Business continuity is assured with SLA-backed availability and enterprise-grade security."[7] Azure Cosmos DB for NoSQL supports retrieval augmented generation (RAG) for use in AI-powered applications built with Azure OpenAI models such as GPT-3.5 and GPT-4.[9] RAG is valuable in that it helps to optimize large language models (LLMs) for better responses based on more up-to-date data, without the high costs of retraining the model.

About Azure AI Search

Azure AI Search is a full-featured vector database with integrated state-of-the-art search ranking technology, so your application delivers the highest quality experiences for every user question and interaction. With its deep data & platform integrations, proprietary re-ranking machine learning, built-in vectorization, and hybrid search, customers can run optimal RAG in their enterprise AI applications.[10,11]

About Azure OpenAI Service

The Microsoft Azure OpenAI Service lets organizations access the OpenAI API through the Azure platform. Through the service, customers get to use OpenAI GPT-3.5 and GPT-4 models and enjoy the "security, reliability, compliance, data privacy and other enterprise-grade capabilities that are built into Microsoft Azure."[12] We chose to use GPT-3.5-Turbo, which OpenAI calls "our most capable and cost effective model in the GPT-3.5 family."[13]

According to Microsoft, "Customers of all sizes across industries are using Azure OpenAI Service to do more with less, improve experiences for end-users, and streamline operational efficiencies internally."[14] The use cases to which organizations are applying the capabilities of Azure OpenAI Service include "customer support, customization, and gaining insights from data using search, data extraction, and classification."[15]

About the reference application we used: Vector Search & AI Assistant

Vector Search & AI Assistant is part of the Microsoft Official Build & Modernize AI Applications reference solutions library that Microsoft has developed to help users build AI-enabled applications and services in Azure. Organizations can use these solutions as a starting point for their own bespoke solutions.[16]

According to Microsoft, the Vector Search & AI Assistant solution focuses on "a consumer retail 'Intelligent Agent' that allows users to ask questions (RAG Pattern) on vectorized product, customer and sales order data stored in the database."[17]

Learn more about these solutions at https://github.com/Azure/Build-Modern-AI-Apps#readme and access the Vector Search & AI Assistant solution at https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector.

Glossary of terms

Retrieval augmented generation (RAG)

RAG involves retrieving supplementary data that the LLM can use when it generates responses. When it receives a question or prompt from a user, RAG searches for the most current and relevant knowledge from articles, documents, and other external sources. When generating responses, RAG uses the information it has retrieved. Together with prompt engineering, RAG improves the quality of responses by including more contextual information in the model.[18]

Prompts and completions

With prompt-based models such as GPT-3.5, the user enters a text prompt, and the model responds with a text completion, which is the model's continuation of the input text.[19]

A prompt is specific text or information that an LLM can use as an instruction or build on as contextual data. Questions, statements, or pieces of code can all be prompts. Prompt engineering is the process of creating effective prompts for a given scenario.[20] A user typing a question or request into the application would also entail a prompt.

Tokens

Tokens are short text strings that OpenAI creates by taking the input text and breaking it into shorter components. Tokens can be words, but do not have to be, and can be as short as a single character. Azure OpenAI takes text ingested by the API and turns it into tokens (or tokenizes it). A variety of factors affect the number of tokens the application may demand, and thus the number of tokens that OpenAI must process. The number of tokens in turn affects the response time and throughput of the models, in addition to other factors.[21]

Vectors and vector search

Vectors are arrays of numbers that represent information about data. When you vectorize a photo, for example, it becomes an array of numbers that represent pixel values. As another example, vectorizing text changes it to a set of numbers representing ASCII values.[22]

A vector search identifies all the vectors that are similar in meaning to a query vector. The greater the number of data points in the repository you're searching, the more intensive the search task becomes. Typically, it becomes necessary to balance factors such as latency, throughput, accuracy, and cost. The specific requirements of an application determine which of these an organization will prioritize.[23]

Vector search is currently supported in both Azure Cosmos DB for MongoDB vCore and Azure Cosmos DB for PostgreSQL. Per Microsoft, "Instead of adding a separate vector database, you can use our vector database extensions when working with multi-modal data. By doing so, you avoid the extra cost of moving data to a separate database. Moreover, this keeps your vector embeddings and original data together, and you can better achieve data consistency, scale, and performance. The latter reason is why OpenAI built its ChatGPT service on top of Azure Cosmos DB."[24]

Embeddings

Vectors representing important features of data are called embeddings. Models in Azure OpenAI divide text into tokens and create embeddings from text data. Words that are similar in meaning will have similar embeddings.[25]

Speed time to value by building an intelligent application with Azure Cosmos DB

Intelligent AI-powered applications drawing on your organizations’ own data have the potential to expand your capabilities and unlock myriad efficiencies. By storing your data in Azure Cosmos DB and starting with the Microsoft reference solution Vector Search & AI Assistant, like we did, you can start bringing value to your organization quickly.

In this proof of concept, a small team at Principled Technologies sought to create a functional chat application with a new, large sample set of data stored in Azure Cosmos DB, using the Vector Search & AI Assistant as our starting point. In the science behind the report, we detail exactly what we did and how we did it, so you can use our work as a starting point for your own application.

If you have chosen to develop an intelligent Azure application that uses large language models with your own data, an effective approach is to store your unstructured or document-based data in Azure Cosmos DB, use Azure AI Search as a search and retrieval system, and utilize Azure OpenAI Service for your LLM access. We implemented this application with a custom dataset, proving that it is not only possible but straightforward to built an intelligent AI-powered application using Azure Cosmos DB and other tools from Azure. While our application is specifically a chat application based on the Vector Search & AI Assistant application, Azure Cosmos DB is also suitable for intelligent AI-powered applications beyond the chat realm. Depending on your organization’s specific needs, you may wish to build an application that performs recommendations, one that allows for speedier transactions, or something entirely unique to you.

[1] GitHub, “Azure/Vector Search AI Assistant,” accessed December 1, 2023, https://github.com/Azure/Vector-Search-AI-Assistant/tree/cognitive-search-vector.

[2] Emily A. Vogels, “A majority of Americans have heard of ChatGPT, but few have tried it themselves,” accessed December 3, 2023, https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/.

[3] Aisha Malik,” OpenAI’s ChatGPT now has 100 million weekly active users,” accessed December 1, 2023, https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/.

[4] OpenAI, “GPT-4,” accessed December 1, 2023, https://openai.com/research/gpt-4.

[5] David Ingram, “These 5 companies say GPT-4 has dramatically changed their priorities at work,” accessed December 1, 2023, https://www.nbcnews.com/tech/innovation/chatgpt-gpt-4-gpt4-openai-access-microsoft-how-to-rcna75116.

[6] Microsoft, “KPMG delivers huge productivity gains and drives new business with generative AI,” accessed December 6, 2023, https://customers.microsoft.com/en-us/story/1700854587537724864-kpmg-australia-azure-ai-professional-services.

[7] Microsoft, “Azure Cosmos DB – Unified AI database,” accessed December 1, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/introduction.

[8] Microsoft, “Online retailer uses cloud database to deliver world-class shopping experiences,” accessed December 19, 2023, https://customers.microsoft.com/fr-fr/story/asos-retail-and-consumer-goods-azure.

[9] Microsoft, “Azure Cosmos DB – Unified AI database.”

[10] Microsoft, “Azure AI Search,” accessed December 1, 2023, https://azure.microsoft.com/en-us/products/ai-services/cognitive-search/.

[11] “Semantic ranking in Azure AI Search,” accessed December 16, 2023, https://learn.microsoft.com/en-us/azure/search/semantic-search-overview.

[12] Microsoft, “New Azure OpenAI Service combines access to powerful GPT-3 language models with Azure’s enterprise capabilities,” accessed December 1, 2023, https://news.microsoft.com/source/features/ai/new-azure-openai-service/.

[13] OpenAI, “Models,” accessed December 1, 2023, https://platform.openai.com/docs/models/gpt-3-5.

[14] Eric Boyd, “General availability of Azure OpenAI Service expands access to large, advanced AI models with added enterprise benefits,” accessed December 1, 2023, https://azure.microsoft.com/en-us/blog/general-availability-of-azure-openai-service-expands-access-to-large-advanced-ai-models-with-added-enterprise-benefits/.

[15] Eric Boyd, “General availability of Azure OpenAI Service expands access to large, advanced AI models with added enterprise benefits.”

[16] GitHub, “Azure/Build-Modern-AI-Apps,” accessed November 30, 2023, https://github.com/Azure/Build-Modern-AI-Apps#readme.

[17] GitHub, “Azure/Build-Modern-AI-Apps.”

[18] Microsoft, “Vector Database,” accessed December 1, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/vector-search.

[19] Microsoft, “Introduction to prompt engineering,” accessed November 30, 2023, https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/prompt-engineering.

[20] Microsoft, “Vector Database,” accessed December 1, 2023, https://learn.microsoft.com/en-us/azure/cosmos-db/vector-search.

[21] Microsoft, “Vector Database.”

[22] Microsoft, “Vector Database.”

[23] Microsoft, “Vector Database.”

[24] Microsoft, “Vector Database.”

[25] Microsoft, “Vector Database.”

Read the science behind this report at https://facts.pt/J5inR81 ▶

Principled Technologies®

Facts matter.®

For additional information, review the science behind this report.

This project was commissioned by Microsoft.