



Vectors in Azure SQL Database: Bringing AI to Your Data



John Morehouse

Principal Consultant


Denny Cherry & Associates

✉ john@dcac.com

 [/in/johnmorehouse](https://www.linkedin.com/in/johnmorehouse)

 [@SQLRUS](https://twitter.com/SQLRUS)

 Sqlrus.com

 He/Him

Community Speaker

Blogger/Tweeter

Nerd

MVP – Data
Platform

Friend of Redgate

Denny Cherry & Associates



Certified IT professionals to help achieve IT goals

Clients ranging from small business to Fortune 10
corporations

Help save on costs while improving IT reliability and solving
challenges





<https://bit.ly/mypresentationfiles>



Public Preview (Vector)



<https://bit.ly/AzDbVectorPublicPreview>



Disclaimer



Not an AI Expert.

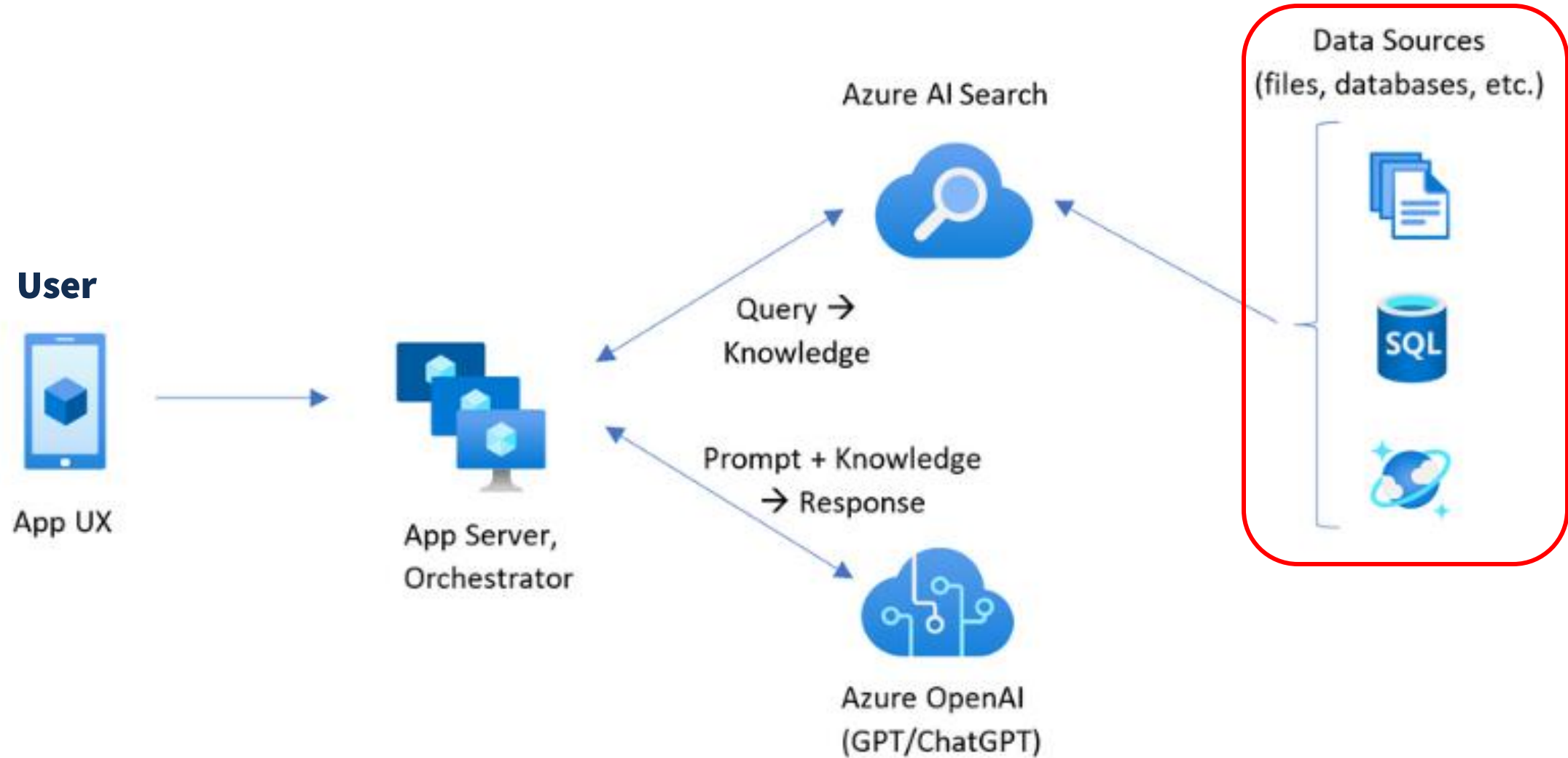
Did not stay in a Holiday Inn Express Last Night.

Did not pass Go and did not collect \$200.

I think this is cool stuff and shows what you can
maybe do with this tech.

Desired Pattern

<https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview>



Retrieval Augmented Generation (RAG)



RAG – the combination of retrieval & generative AI models. LLMs can access information outside of their training, in near real time, when generating a response.

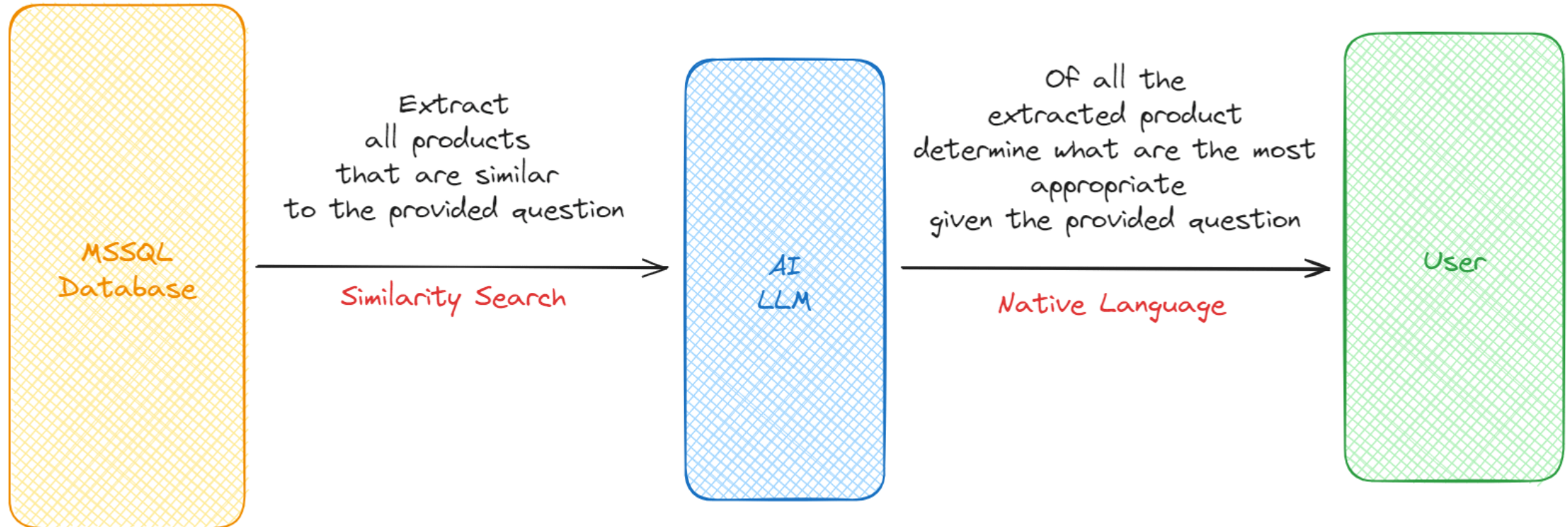
Embeddings – the process of representing data mathematically in a vector.

Vectors – list of floating-point numbers that reflect a multi-dimensional similarity map.

Retrieval Augmented Generation (RAG)



<https://github.com/Azure-Samples/azure-sql-db-chatbot>





Embeddings

Embeddings



"THE FOUNDATION SERIES
by ISAAC ASIMOV"



[-0.31456 , 0.0089,
0.99123 , 0.4412,
0.7411 ,
... 0.1437]

Embeddings



```
1 # Azure OpenAI metadata variables
2 $openai = @{
3     api_key    = 'cca05f0748e841f6b0e730163809b2f8'
4     api_base   = 'https://myembeddings.openai.azure.com/' # your endpoint should look like the following https://YOUR_RESOURCE_NAME.openai.azure.com/
5     api_version = '2023-05-15' # this may change in the future
6     name       = 'text-embedding-ada-002' #This will correspond to the custom name you chose for your deployment when you deployed a model.
7 }
8
9 $headers = [ordered]@{
10     'api-key' = $openai.api_key
11 }
12
13 $text = 'My name is John'
14
15 $body = [ordered]@{
16     input = $text
17 } | ConvertTo-Json
18
19 $url = "$($openai.api_base)/openai/deployments/$($openai.name)/embeddings?api-version=$($openai.api_version)"
20
21 $response = Invoke-RestMethod -Uri $url -Headers $headers -Body $body -Method Post -ContentType 'application/json'
22 return $response.data.embedding
```

Visual Studio Code interface showing the script in a file named 'Untitled-2'. The interface includes a sidebar with icons for Explorer, Search, Source Control, Run and Debug, Extensions, Testing, Docker, Remote Explorer, Accounts, and Settings. The bottom status bar shows 0 errors, 0 warnings, 4 info messages, and 0 debug messages.



Vectors

Vector Data Type



New for Azure SQL Database

Stored as VARBINARY(X), exposed as JSON array

Each element in the array is a single-precision 4 bytes floating point value

Must have 1 dimensions, 1998 maximum dimensions

```
DECLARE @v VECTOR(3) = '[1.0, -0.2, 30]';
```

Vector Data Type



Constraints are not honored except for NULL/NOT NULL

Normal operations do not work with vectors

Vector columns cannot be used with in-memory optimized tables

ALTER COLUMN away from VECTOR is not allowed

CAST/CONVERT to VARCHAR/NVARCHAR do work

Native Vector Functions



`VECTOR_NORM` - Takes a vector as an input and returns the norm of the vector.

`VECTOR_NORMALIZE` - Takes a vector as an input and returns the normalized vector, which is a vector scaled to have a length of 1.

`VECTOR_DISTANCE` (cosine/Euclidean/dot)- Calculates the distance between two vectors using a specified distance metric.

Demo Requirements



Azure SQL Database

<https://bit.ly/AzDBChatbot>

Azure OpenAI Resource

Data



Two Models:

gpt-4

test-embedding-ada-002



DEMO

OpenAI - Costs



Standard – Consumption based, pay as you go for input & output tokens

Provisioned (PTUs) – predictable costs, with monthly or annual reservations available to further reduce costs

Batch API – LLM that returns completions within 24 hours for a 50% discount on Global Standard Pricing

OpenAI – Deployment Types



These works for Standard or Provisioned deployments

Global Deployment – Global SKU

Data Zone Deployment – Geographic based (EU or US)

Regional Deployment - Local Region (up to 27 regions)

OpenAI – Cost Example



Model	Pricing (1M Tokens)
GPT-4o-2024-1120 Global	Input: \$2.50 Cached Input: \$1.25 Output: \$10
GPT-4o-2024-1120 US/EU – Data Zones	Input: \$2.75 Cached Input: \$1.375 Output: \$11
GPT-4o-2024-1120 Regional	Input: \$2.75 Cached Input: \$1.375 Output: \$11

Costs



Tokens are represented as pieces of words.

Tokens are not precise

1 token = ~4 characters in the English language

100 tokens = ~75 words

Note: Tokens can include trailing spaces.

Known Issues



Tools such as SSMS/ADS could possibly script the table definition wrong or reflect **VARBINARY** vs **VECTOR**.

BCP / BULK INSERT don't currently work if tables contain the **VECTOR** type.

AE & Column encryption doesn't currently support the **VECTOR** type.

Known Issues



Data Masking currently shows **vector** data as **varbinary** data type in the portal.

LEN/DATALENGTH throws error 8116 is returned if you pass it a **VECTOR** column.

In some cases, you may get error 42211 (Truncation of vector is not allowed during the conversion) when used in Procs. Cast the column to NVARCHAR(max).

Resources



<https://learn.microsoft.com/en-us/azure/ai-services/openai/overview>

<https://learn.microsoft.com/en-us/azure/azure-sql/database/ai-artificial-intelligence-intelligent-applications?view=azuresql#vectors-1>

<https://learn.microsoft.com/en-us/azure/azure-sql/database/ai-artificial-intelligence-intelligent-applications?view=azuresql#retrieval-augmented-generation>

<https://learn.microsoft.com/en-us/sql/t-sql/data-types/vector-data-type?view=azuresqlldb-current&viewFallbackFrom=sql-server-ver16&tabs=csharp-sample>

<https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/#pricing>



Questions?
Answers!

**Got
Questions?**

**Follow Me on
Twitter X!**

**Check out
my blog!**

John Morehouse

Denny Cherry & Associates Consulting

DCAC*



john@dcac.com



Sqlrus.com



@SQLRUS



/in/johnmorehouse



Slides & Demos