

REST APIs, AI and Vectors in SQL Server 2025

Ben Weissman (he/him)

Ben Weissman (he/him) @bweissman.bsky.social /in/weissmanben bweissman



Thank you, Sponsors!!!!

Data Community Austria

Platinum

ACP IT for innovators.

Gold

Lucient

b.telligent smart data. smart decisions.

HEDDA.IO

COHESITY

Bronze

paiqo The Platform & AI Company

redgate

devart

Power BI Camp www.linearis.at



Who am I?

Ben Weissman (he/him)

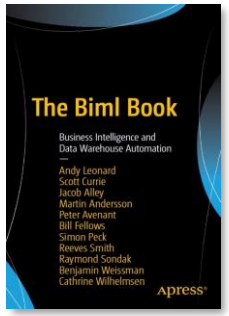
b.weissman@solisyon.de

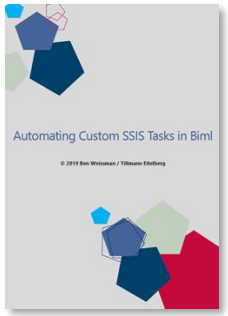
@bweissman.bsky.social

/in/weissmanben

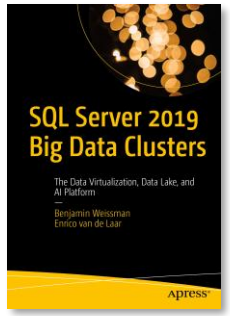
bweissman

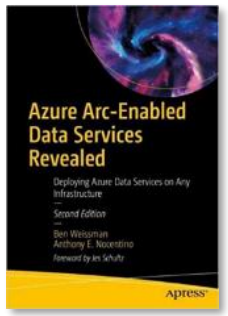
SOLISYON
www.solisyon.de

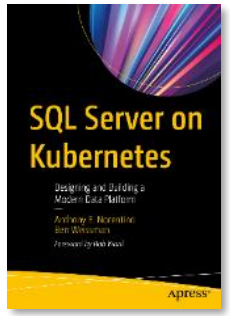


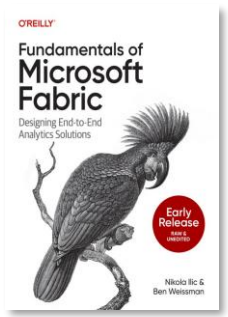


Apress®









Please Talk Data To Me





Finding Data Friends

Data uncut

MVP Microsoft® Most Valuable Professional

BimlHero CERTIFIED EXPERT

DataMinutes

Microsoft CERTIFIED Solutions Associate Machine Learning

Microsoft CERTIFIED Solutions Expert Data Management and Analytics

redgate COMMUNITY AMBASSADOR 2025

NEW STARS OF DATA

MICROSOFT PROFESSIONAL PROGRAM
Data Science
Big Data
Artificial Intelligence
Data Analysis

PLURALSIGHT

LEARNING

GENESEE ACADEMY World-Class Training
Certified Data Vault Modeler

Check on your HAPPY friends



Who am I?

Ben Weissman (he/him)



NEW STARS OF DATA

FindingData Friends



All ways to share your thoughts, communicate, and learn are welcome!





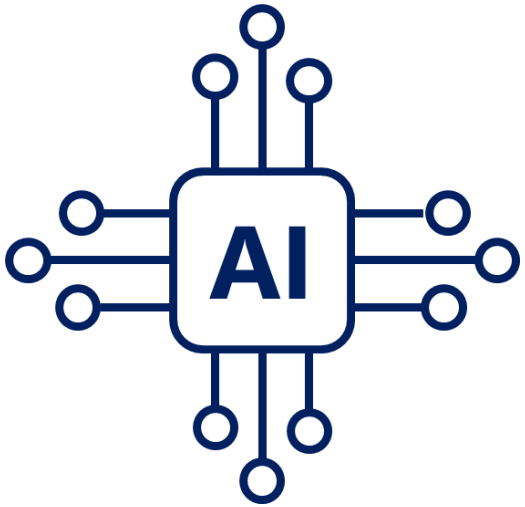
Created by DALL-E, inspired by Lisa Hoving and https://www.facebook.com/lilpenguinstudios/



Agenda



Rest APIs



Talking to AI Models



Using AI Models
to query your data

REST APIs in SQL Server 2025



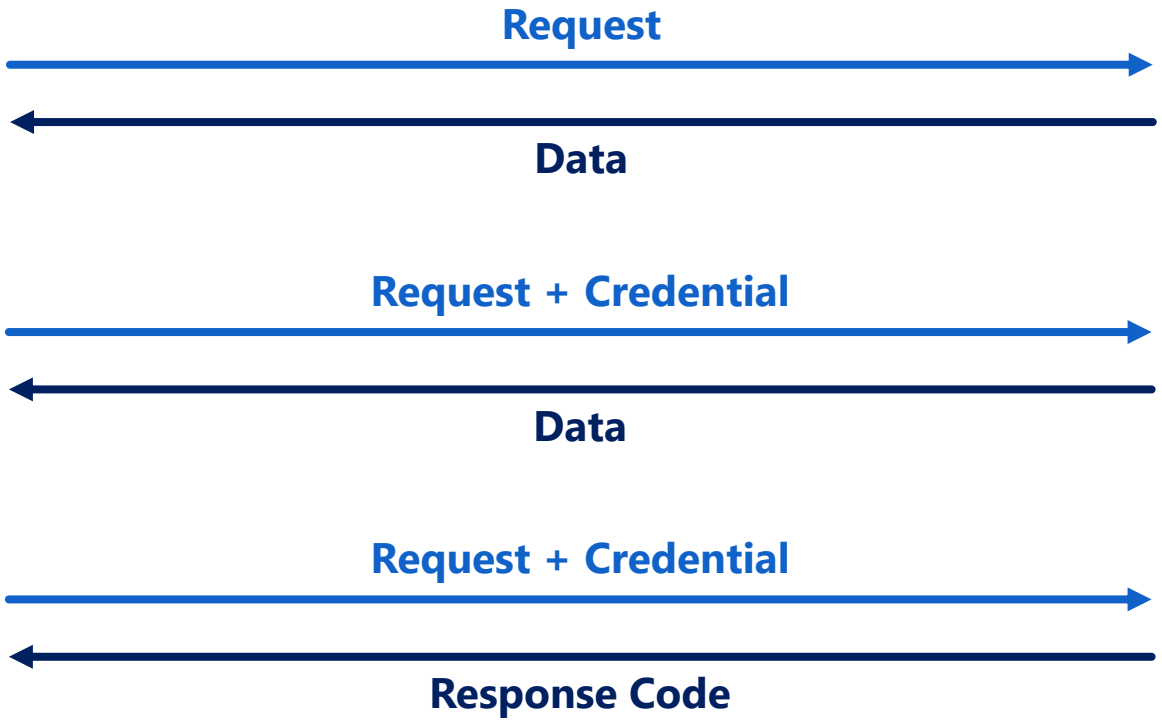


REST APIs in SQL Server 2025



REST API














```
EXEC @returnValue = sp_invoke_external_rest_endpoint
  [ @url = ] N'url'
  [ , [ @payload = ] N'request_payload' ]
  [ , [ @headers = ] N'http_headers_as_json_array' ]
  [ , [ @method = ] 'GET' | 'POST' | 'PUT' | 'PATCH' | 'DELETE' | 'HEAD' ]
  [ , [ @timeout = ] seconds ]
  [ , [ @credential = ] credential ]
  [ , @response OUTPUT ]
```



AI Integration

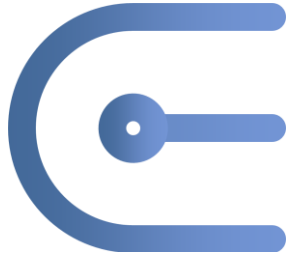


REST API



Request

Data



AI Endpoint


Demo
REST APIs + Talking to AI


Vectors/Embeddings



What's an Embedding?

A vector embedding is a way to represent data - like words, images, or documents - as numerical vectors in a multi-dimensional space, capturing their meaning or characteristics.

Similar items have embeddings that are close together, enabling tasks like search, clustering, or recommendation.

(Source: ChatGPT)



What's an Embedding?





Vector embedding in JSON format (representing the word "cat" as a 5-dimensional vector)

```
{
  "item": "cat",
  "embedding": [0.12, -0.85, 0.33, 0.44, -0.27]
}
```



Again, similar contexts get similar vectors:

```
[
  {
    "item": "cat",
    "embedding": [0.12, -0.85, 0.33, 0.44, -0.27]
  },
  {
    "item": "dog",
    "embedding": [0.08, -0.80, 0.30, 0.50, -0.25]
  },
  {
    "item": "house",
    "embedding": [-0.45, 0.60, -0.10, 0.05, 0.90]
  }
]
```





How does that help me?

Similar content results in similar (=close) vectors.
You can find what you are looking for through context.

"I want a not so expensive bike."

vs.

```
SELECT * FROM PRODUCTS where type = 'bike' and description like '%cheap%'
```

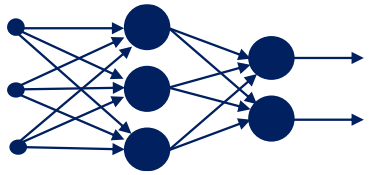


How do I get there?

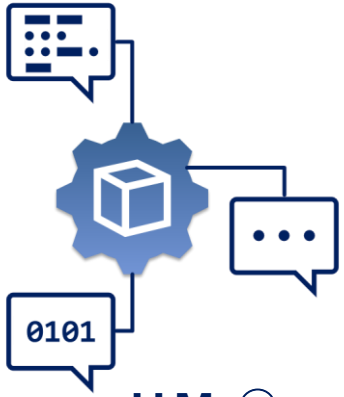


dbo.ConferenceSessions

Title	Description	Speaker	Embedding (Vector(768))
			[0.12, -0.85, 0.33, 0.44, -0.27]



Model ③



LLM ②

⑤ ⑤ ⑤ ⑧

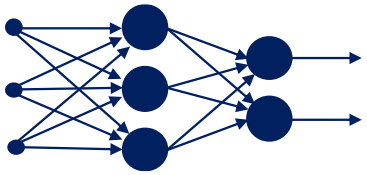
AI_GENERATE_EMBEDDINGS

④ ⑥ ⑦



How do I query that?

"I look for a session on Embeddings?"



Model

AI_GENERATE_EMBEDDINGS → [0.12, -0.85, 0.33, 0.44, ...]

Vector

VECTOR_DISTANCE



dbo.ConferenceSessions → 

Search Result

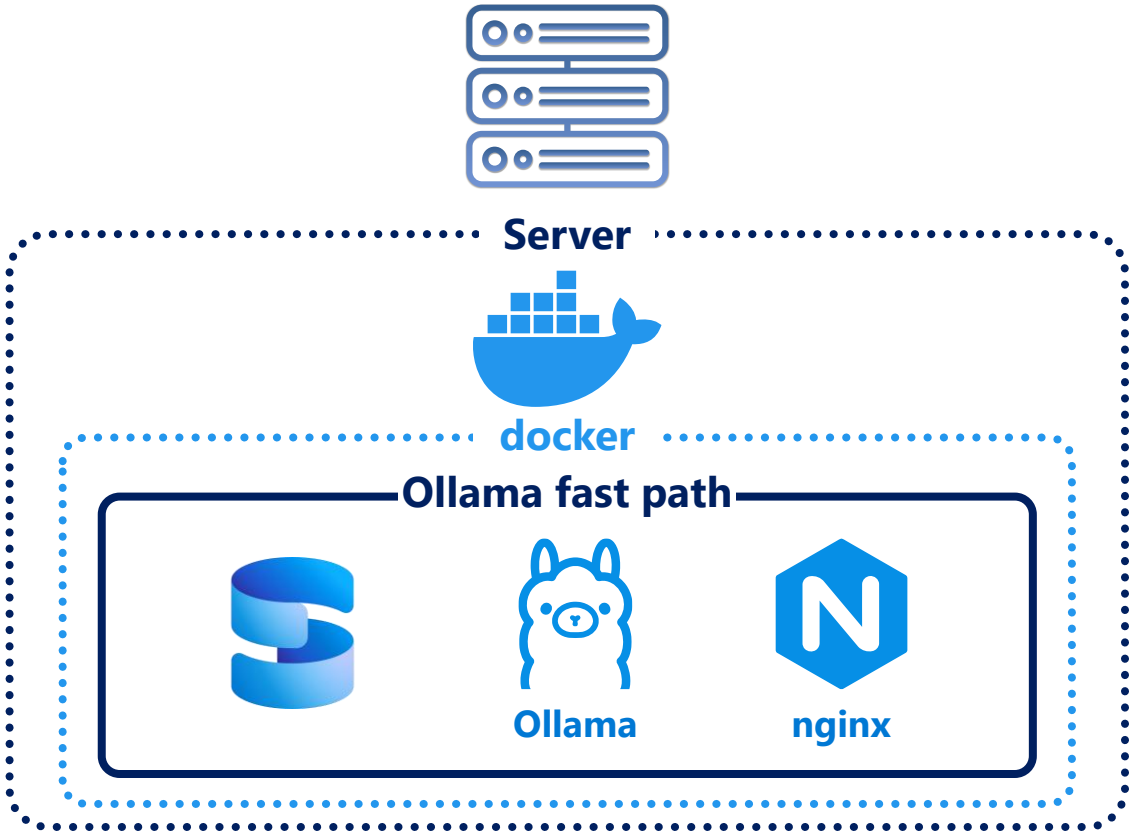


Caveats and Gotchas

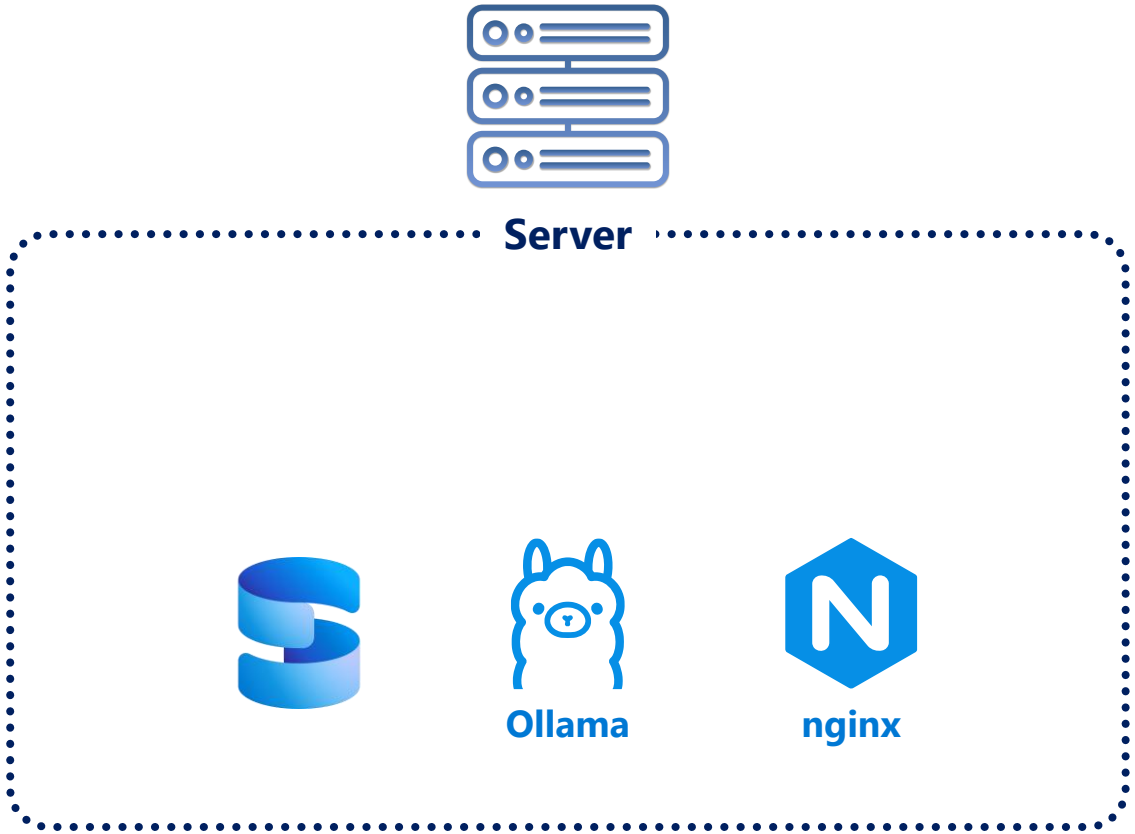
- Embeddings take time – use GPUs
  (GPUs are “all or nothing”!)
- Be clear what goes into your embedding for context
  (Name + Bio vs. Name only)
  → consider multiple embeddings
- Same input results in same embedding (with same model)
- Take the time to find the right model



Installation



https://github.com/nocentino/ollama-sql-faststart/

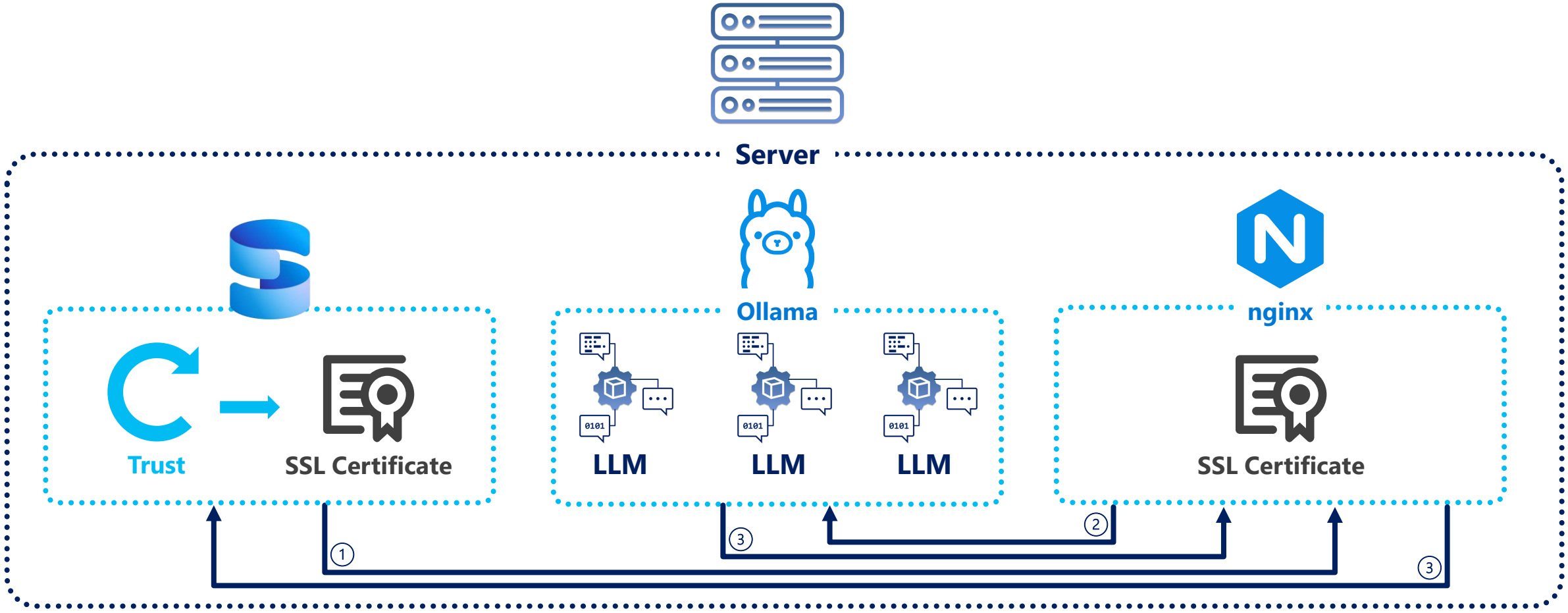


https://bookmark.ws/sql25

⚠ GPUs don't work with Docker except on WSL2



Installation

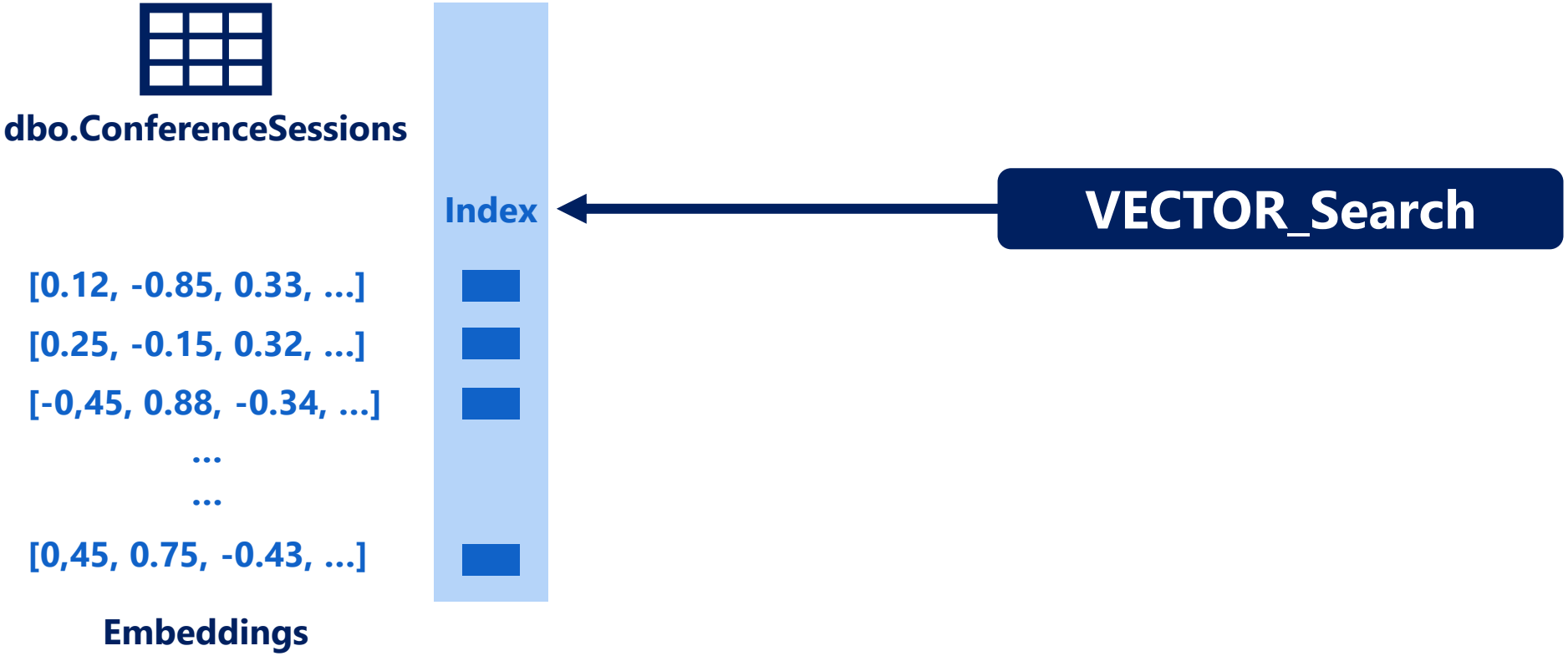


⚠ You can use alternative tools & solutions


Demo
Deploying an LLM
Query using Vector Distances



Vector Indexes – ANN vs. ENN



⚠ Vector indexes make your table read only!


Demo
Vector Search



Thank you! Questions? Feedback, please!

Ben Weissman (he/him)

b.weissman@solisyon.de

@bweissman.bsky.social

/in/weissmanben

bweissman



YouTube

Ben Weissman

@benweissman

www.youtube.com/@benweissman/playlists



LEARNING

Ben Weissman

/in/weissmanben

www.linkedin.com/learning/rest-apis-vectors-and-ai-in-sql-server-2025