A Primer to
SAP HANA Cloud’s Vector Engine

Mathias Kemeter, SAP
June 03, 2024

Public

MANNHEIM
HANA Tech Night
#HANATechNight

Agenda

The Basics

- Vector, Embedding and Vector Database

SAP HANA Cloud

- Feature Scope
- Release Date

▶ ***Demo!***

Usage Patterns

- Retrieval Augmented Generation
- LangChain Integration

Summary & Outlook

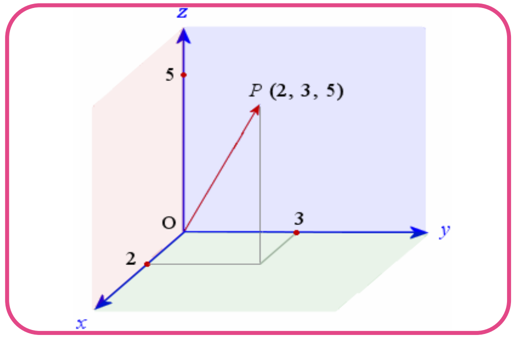
- Benefits
- Roadmap
- Further Resources

Photo by Tamas Tuzes-Katai on Unsplash

The Basics

What is a Vector?

Famous supervillain known for his nerdy personality and gadgets.

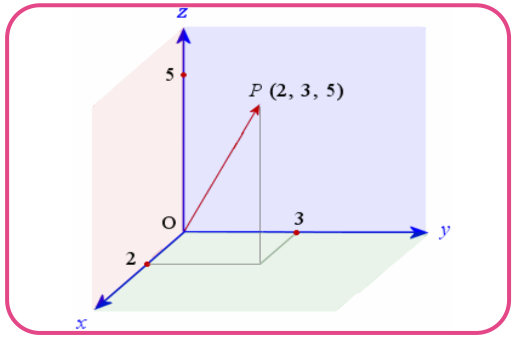


A quantity that has both magnitude and direction, often represented by an arrow.

A vehicle to transfer genetic material into a target cell.

What is a Vector?

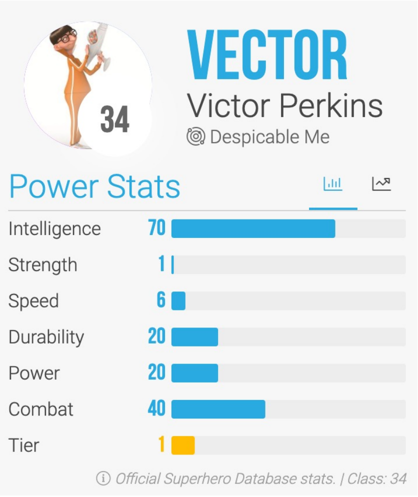
Famous supervillain known for his nerdy personality and gadgets.



A quantity that has both magnitude and direction, often represented by an arrow.

A vehicle to transfer genetic material into a target cell.

What is an Embedding?



VECTOR

Victor Perkins

Despicable Me

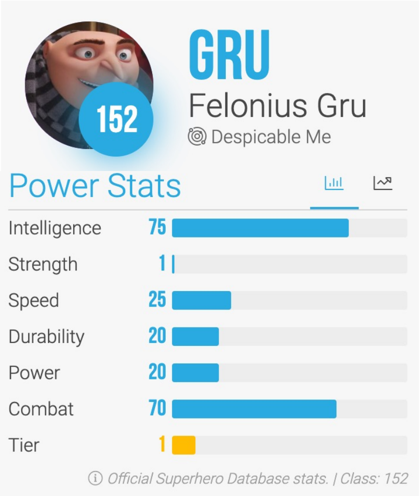
34

Power Stats

Intelligence	70
Strength	1
Speed	6
Durability	20
Power	20
Combat	40
Tier	1

Official Superhero Database stats. | Class: 34

Source: https://www.superherodb.com/



GRU

Felonius Gru

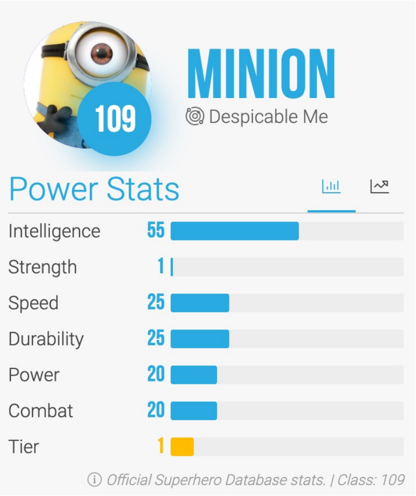
Despicable Me

152

Power Stats

Intelligence	75
Strength	1
Speed	25
Durability	20
Power	20
Combat	70
Tier	1

Official Superhero Database stats. | Class: 152



MINION

Despicable Me

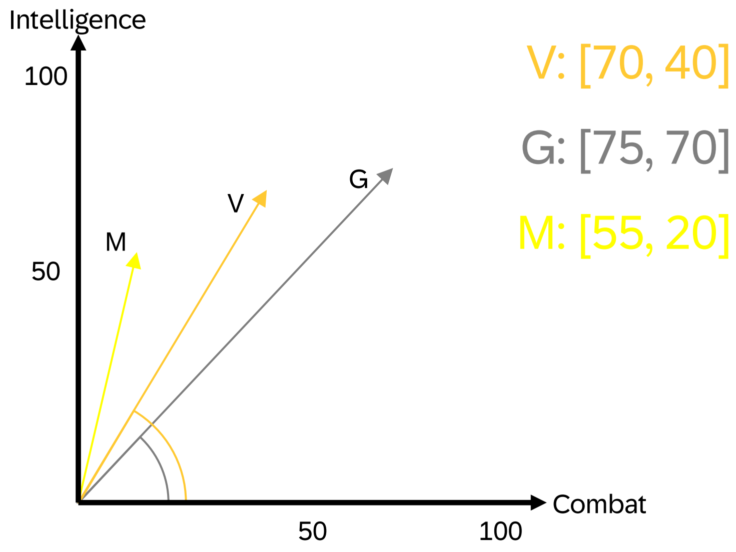
109

Power Stats

Intelligence	55
Strength	1
Speed	25
Durability	25
Power	20
Combat	20
Tier	1

Official Superhero Database stats. | Class: 109

Embed into 2d vector space



How to measure similarity?

- Distance: Look for character with similar (absolute) values.
- Angle: Look for character with similar ratio of values.

What is a Vector Database?

"A [..] vector database or vector store is a database that can store vectors (fixed-length lists of numbers) along with other data items. Vector databases typically implement one or more Approximate Nearest Neighbor (ANN) algorithms, so that one can search the database with a query vector to retrieve the closest matching database records."

Source: https://en.wikipedia.org/wiki/Vector_database

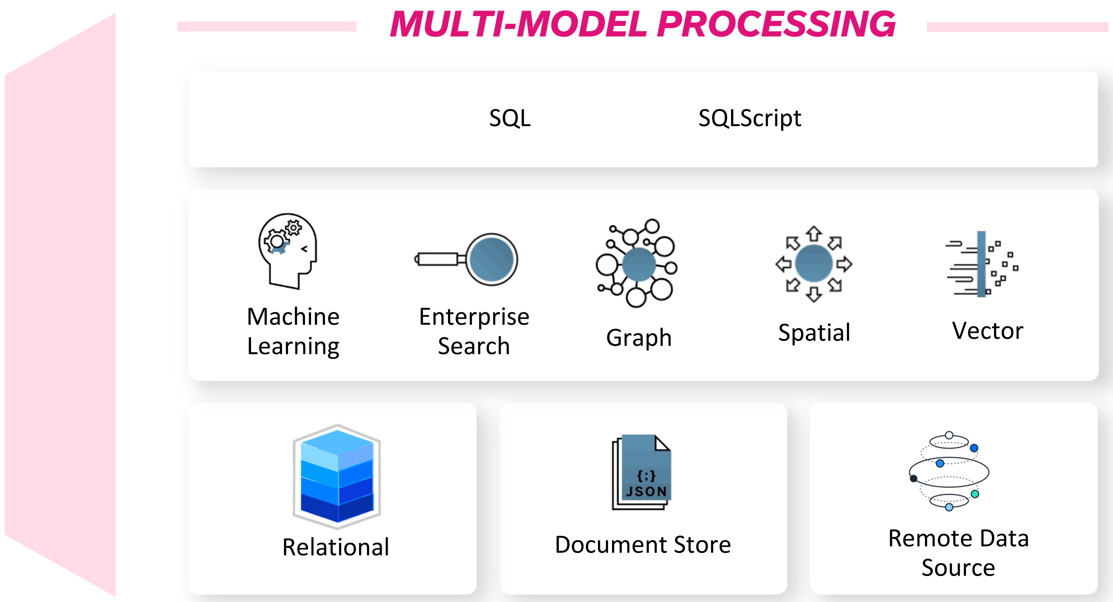
SAP HANA Cloud

SAP HANA Multi-model

Virtually blend data from remote sources with locally tiered data

Process all types of business data regardless of data model, type, or volume

Interact in real-time using an industry standard language



MULTI-MODEL PROCESSING

SQL SQLScript

Machine Learning

Enterprise Search

Graph

Spatial

Vector

Relational

Document Store

Remote Data Source

SAP HANA Cloud Vector Engine

Vector data type

- Native datatype for high-dimensional vector

Vector functions

- Vector distance functions
  – L2 distance and cosine similarity
- Full SQL integration
  – Conversion and tooling functions

Consumption

- SQL, Python (hana-ml), Langchain

Data Definition

```
CREATE TABLE "T1" (
 "ID" BIGINT,
 "TEXT" NCLOB,
 "VEC" REAL_VECTOR(1536)
)
```

Data Manipulation

```
INSERT INTO "T1" VALUES (
 1,
 'some text',
 TO_REAL_VECTOR('[0.1,...,0.9]')
)
```

Similarity Search

```
SELECT TOP 10 "TEXT"
FROM "T1"
ORDER BY COSINE_SIMILARITY(
 "VEC",
 TO_REAL_VECTOR('[..]') DESC
)

SELECT TOP 10 "TEXT"
FROM "T1"
ORDER BY L2DISTANCE(
 "VEC",
 TO_REAL_VECTOR('[..]’) ASC
)
```

Demo – SQL Interface

Release to Customer

The Vector Engine of SAP HANA Cloud has been introduced with

QRC01/2024

It can be tested free of charge using SAP HANA Cloud Trial.

Usage Patterns

Usage Patterns

Semantic / similarity search

- Retrieval Augmented Generation (RAG)
- Advanced search / Q&A
- Image similarity

Clustering

- Group documents to detect topic clusters

Recommendations

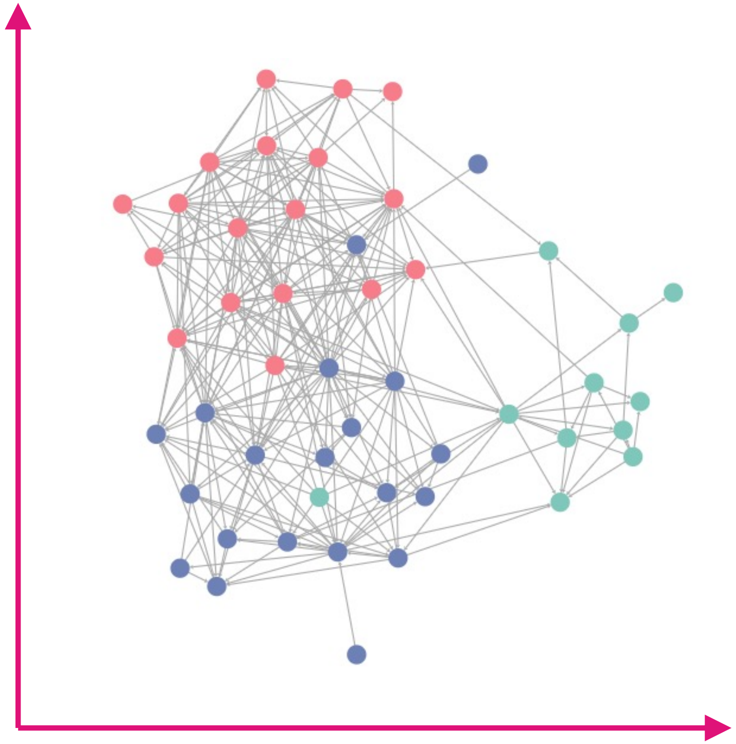
- Cold-start recommendations for service solutions

Anomaly detection

- Detect outliers comparing product descriptions and attributes

Classification

- Use distance metrics for classification



Retrieval Augmented Generation

Common issues when using LLMs

Cut-off date for knowledge base

Hallucinations

Only public data

Solution approach

Lookup relevant knowledge for a given user prompt

Augment the prompt with this knowledge before sending it to the LLM

Retrieval Augmented Generation

Common issues when using LLMs

Cut-off date for knowledge base

Hallucinations

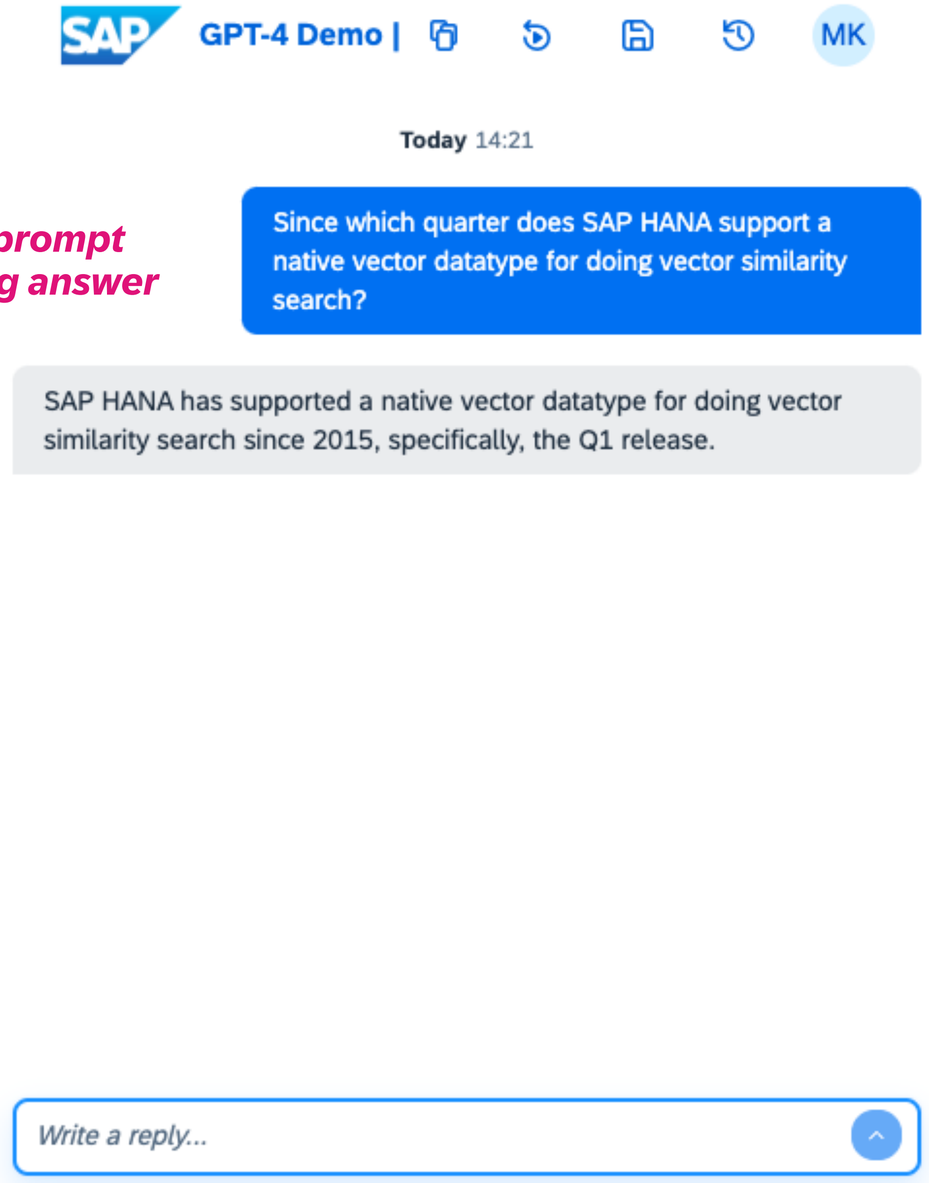
Only public data

Solution approach

Lookup relevant knowledge for a given user prompt

Augment the prompt with this knowledge before sending it to the LLM

Direct prompt = wrong answer



Today 14:21

Since which quarter does SAP HANA support a native vector datatype for doing vector similarity search?

SAP HANA has supported a native vector datatype for doing vector similarity search since 2015, specifically, the Q1 release.

Retrieval Augmented Generation

Common issues when using LLMs

Cut-off date for knowledge base

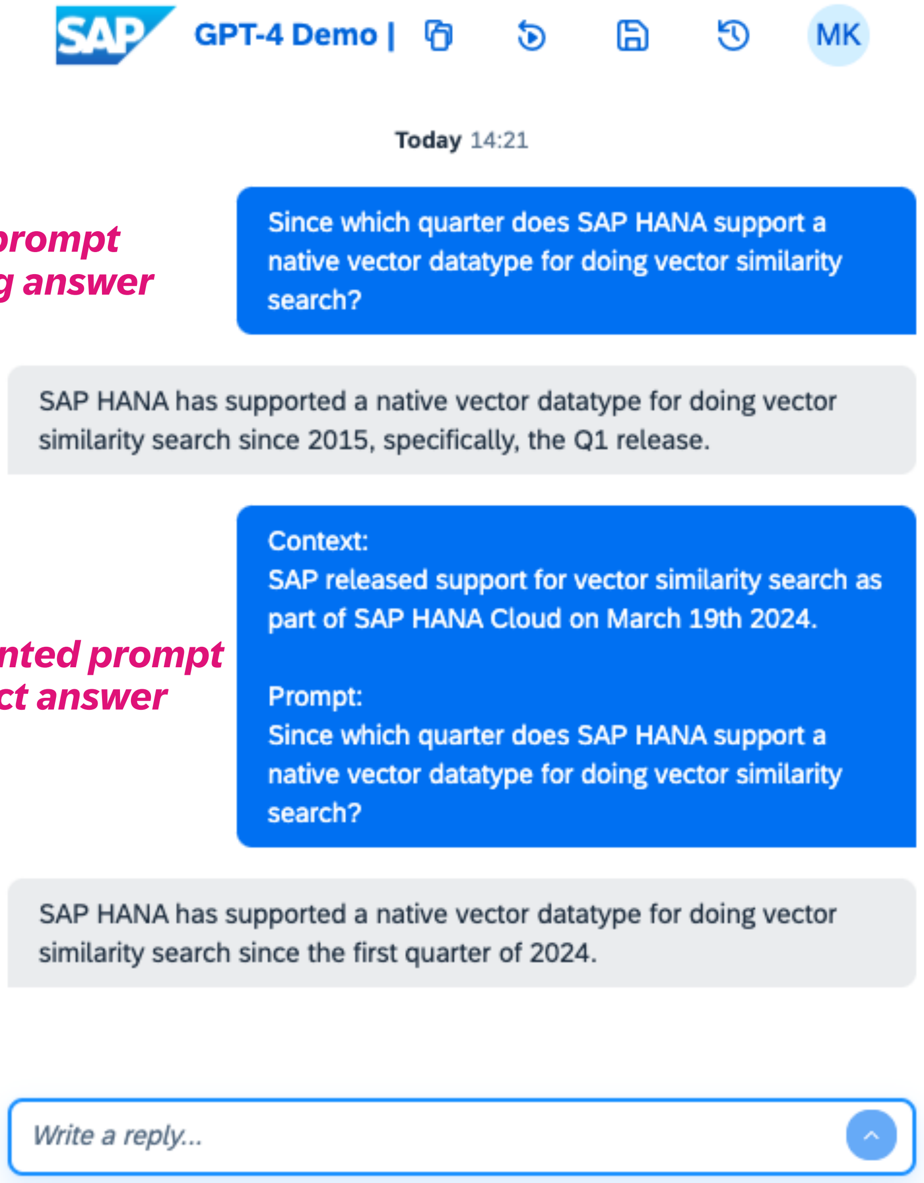
Hallucinations

Only public data

Solution approach

Lookup relevant knowledge for a given user prompt

Augment the prompt with this knowledge before sending it to the LLM



Today 14:21

Direct prompt = wrong answer

Since which quarter does SAP HANA support a native vector datatype for doing vector similarity search?

SAP HANA has supported a native vector datatype for doing vector similarity search since 2015, specifically, the Q1 release.

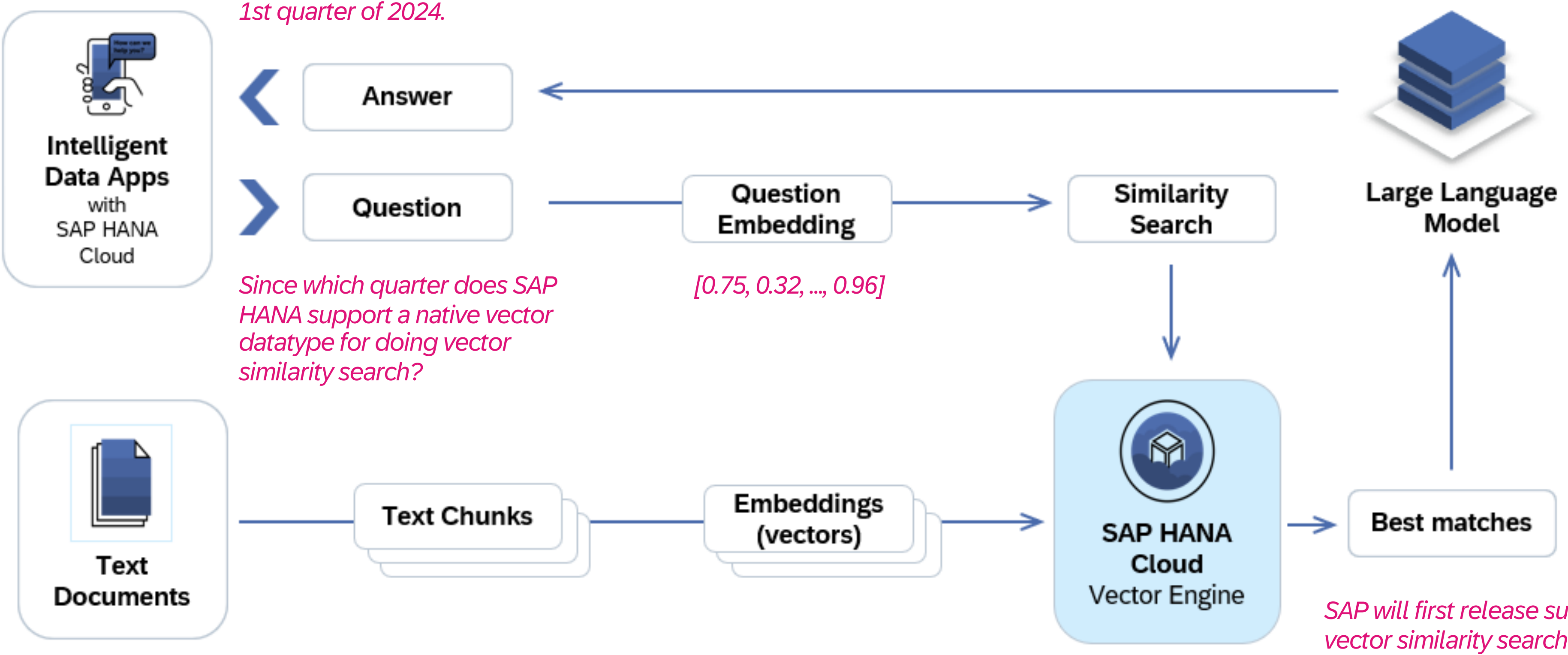
Augmented prompt = correct answer

Context:
SAP released support for vector similarity search as part of SAP HANA Cloud on March 19th 2024.

Prompt:
Since which quarter does SAP HANA support a native vector datatype for doing vector similarity search?

SAP HANA has supported a native vector datatype for doing vector similarity search since the first quarter of 2024.

Retrieval Augmented Generation



SAP HANA will start supporting a native vector datatype for doing similarity search starting from the 1st quarter of 2024.

Context: < >

Prompt: < >

Intelligent Data Apps with SAP HANA Cloud

Answer

Question

Question Embedding

Similarity Search

Large Language Model

Since which quarter does SAP HANA support a native vector datatype for doing vector similarity search?

[0.75, 0.32, ..., 0.96]

Text Documents

Text Chunks

Embeddings (vectors)

SAP HANA Cloud Vector Engine

Best matches

[0.75, 0.32, ..., 0.96]
[0.71, 0.33, ..., 0.94]

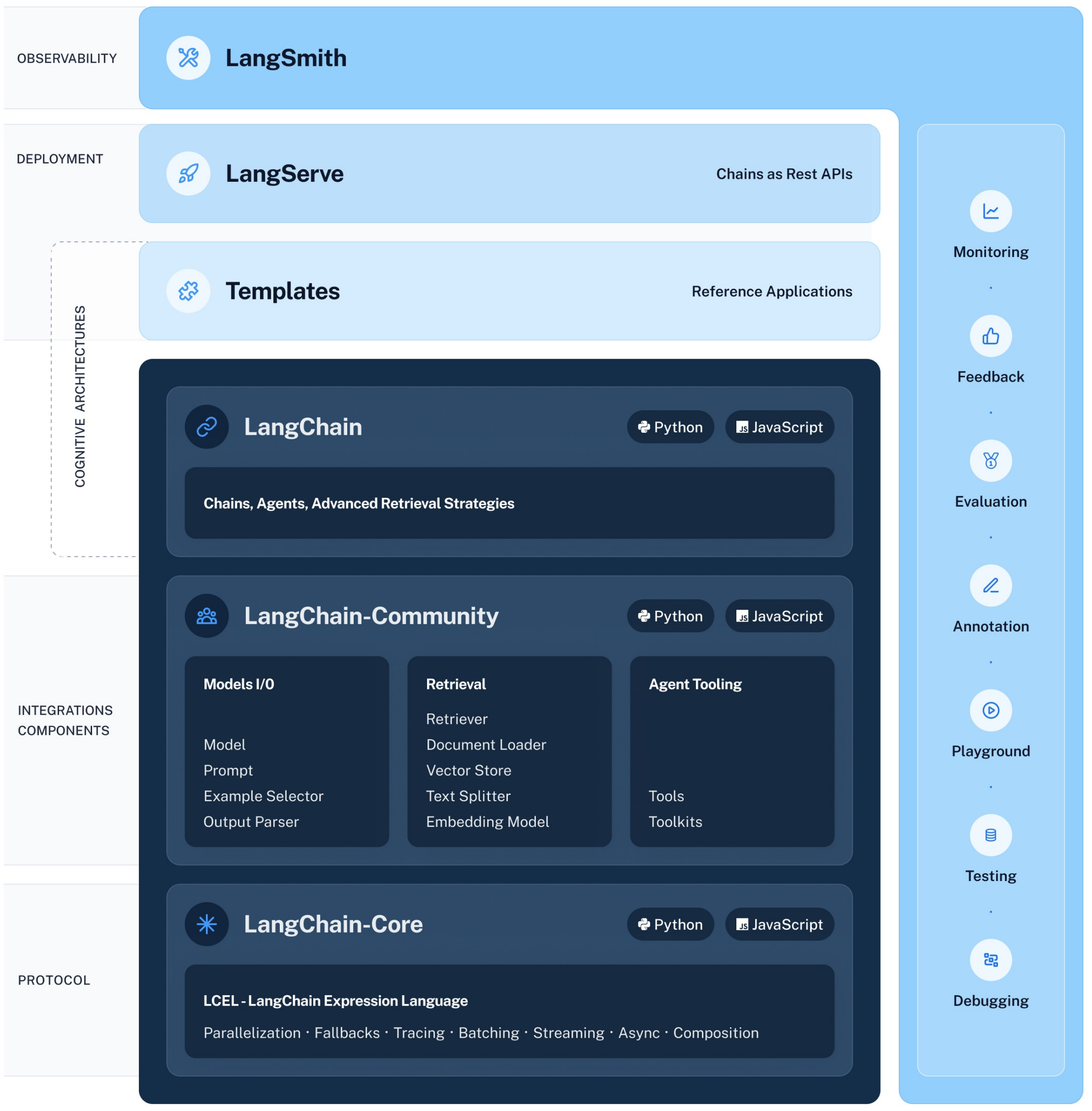
SAP will first release support for vector similarity search as part of SAP HANA on upcoming March 19th 2024.

LangChain 🦜🔗

LangChain is a framework for developing applications powered by language models. It enables applications that:

Are context-aware: connect a language model to sources of context (prompt instructions, few shot examples, content to ground its response in, etc.)

Reason: rely on a language model to reason (about how to answer based on provided context, what actions to take, etc.)



OBSERVABILITY

LangSmith

DEPLOYMENT

LangServe

Chains as Rest APIs

COGNITIVE ARCHITECTURES

Templates

Reference Applications

LangChain

Python

JavaScript

Chains, Agents, Advanced Retrieval Strategies

INTEGRATIONS COMPONENTS

LangChain-Community

Python

JavaScript

Models I/O

Model
Prompt
Example Selector
Output Parser

Retrieval

Retriever
Document Loader
Vector Store
Text Splitter
Embedding Model

Agent Tooling

Tools
Toolkits

PROTOCOL

LangChain-Core

Python

JavaScript

LCEL - LangChain Expression Language

Parallelization · Fallbacks · Tracing · Batching · Streaming · Async · Composition

Monitoring

Feedback

Evaluation

Annotation

Playground

Testing

Debugging

https://github.com/langchain-ai/langchain

https://python.langchain.com/docs/integrations/vectorstores/sap_hanavector

Summary & Outlook

Benefits of using SAP HANA Cloud

Store high-dimensional vector data ***alongside business data***

Run fast vector-based ***similarity search*** using SQL

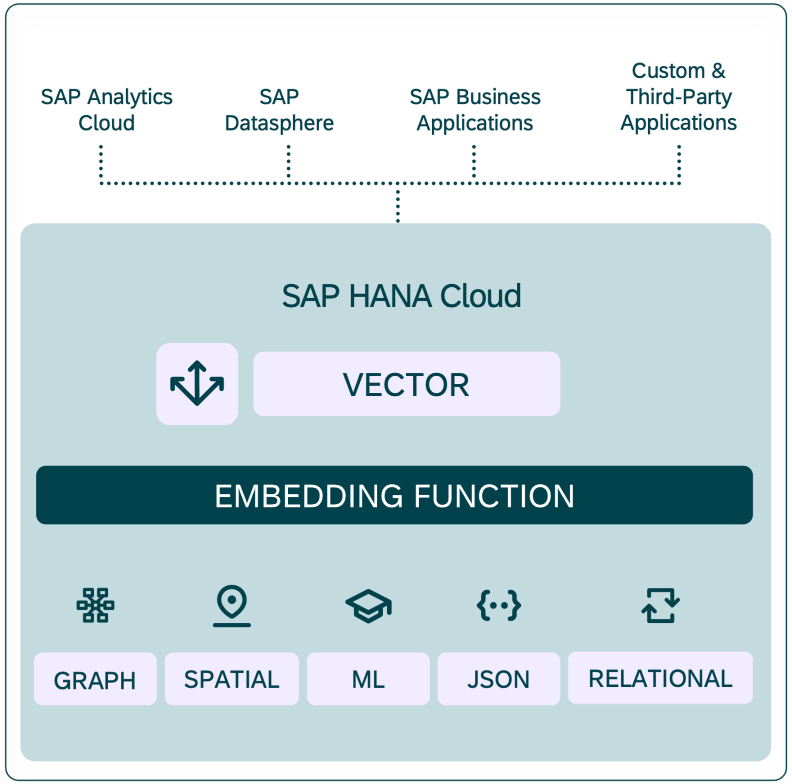
Combine vector-based search with SQL operations like ***joins, filters and aggregations***

Leverage ***advanced multi-model processing*** by integrating spatial, graph, JSON, machine learning and vector work loads.

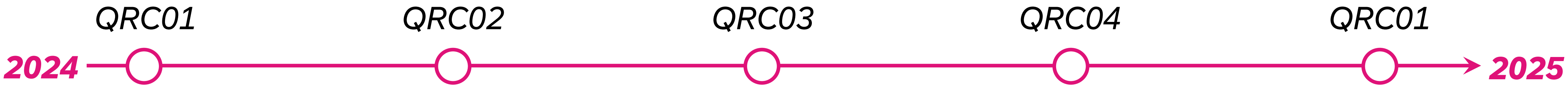
Re-use existing ***authorization concept***

Simplified architecture and operations

Integration with SAP and Open Source ecosystem



Preliminary Outlook 2024



2024

QRC01

- RTC of Vector Engine
- Langchain (Python) Integration
- CAP Support

QRC02

- Functions MEMBER_AT and CARDINALITY
- Parquet Import/Export
- LangChain (Typescript) Integration

QRC03

- ANN-search with vector index support
- Vector Clustering via PAL

QRC04

QRC01

- NSE for vectors

2025

Slide states preliminary planning and is for illustrational purpose only. Some deliverables are not bound to the release cycles (QRC) of SAP HANA Cloud.

Further Resources

Our overview blog is featuring further helpful resources:

- Official Documentation
  https://help.sap.com/docs/hana-cloud-database/sap-hana-cloud-sap-hana-database-vector-engine-guide/sap-hana-cloud-sap-hana-database-vector-engine-guide

- SAP HANA Cloud Guided Experience
  https://www.sap.com/products/technology-platform/hana/trial.html

- LangChain plug-in
  https://python.langchain.com/docs/integrations/vectorstores/sap_hanavector

- LangChain.js plug-in
  https://js.langchain.com/v0.1/docs/integrations/vectorstores/hanavector/

Thank you.

Contact information:

Mathias Kemeter

mathias.kemeter@sap.com
https://www.linkedin.com/in/mathiaskemeter/