Talk + Demo

Db2 for LLM Apps

Vector Search, LLM Frameworks, & LLM Inferencing via SQL

Shaikh Quader
Db2 AI Architect

Marhaba Chariwala
Db2 Developer

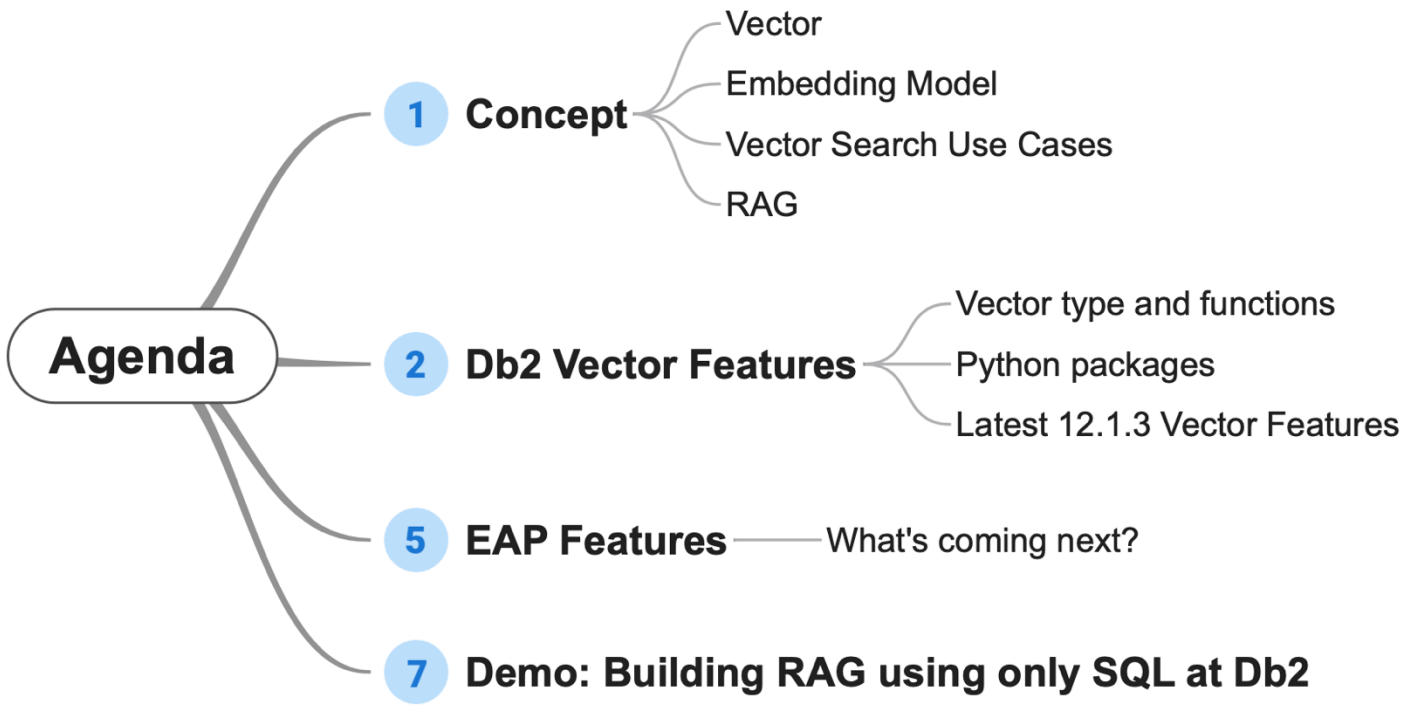
Thursday, Nov 13, 2025 @ 1:00 - 2:00 p.m

Venue: 8200 Warden Amphitheatre

Online: **ibm.biz/db2talk**

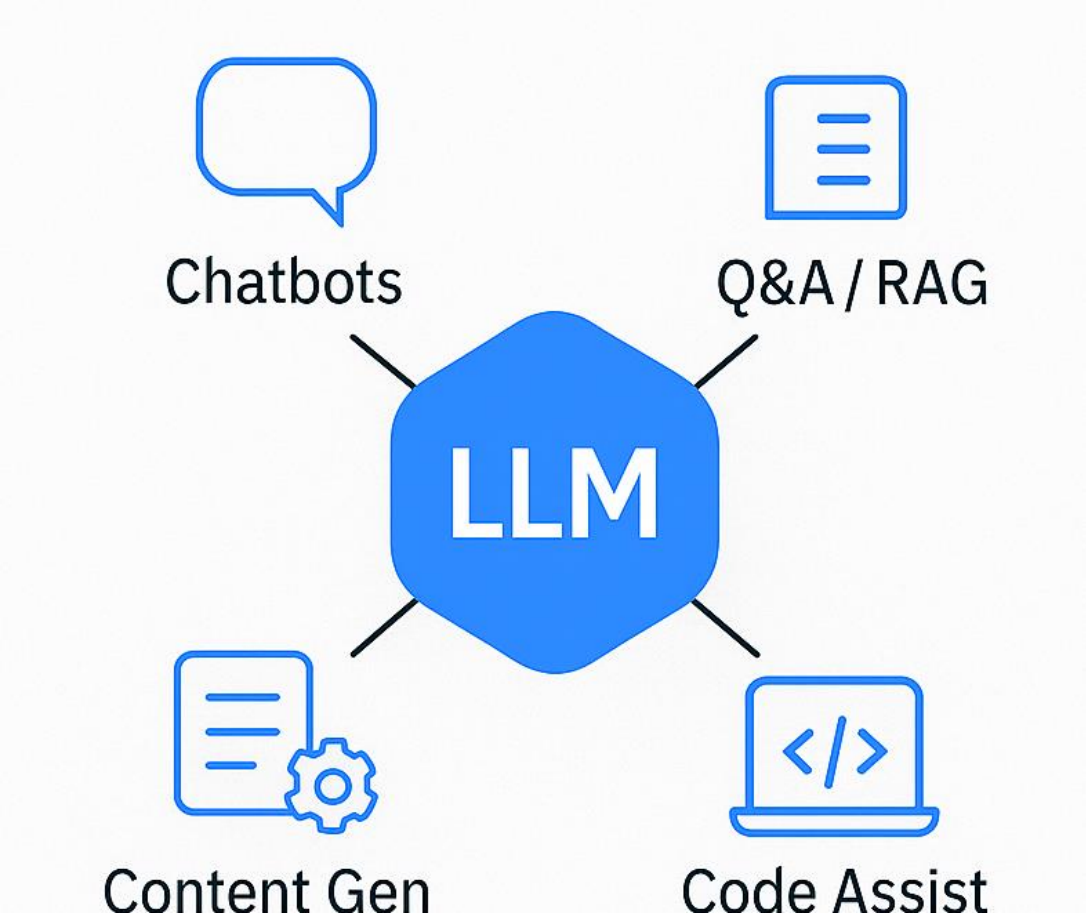
Agenda

Db2 for LLM Apps Talk @ Nov 13: Building Language Apps with Db2



LLM Applications

Apps that use **LLMs** to understand, generate, and interact with natural language.



What is a vector?

- A list of numbers, like (1, 2)
- Represents a point in space
- Like map coordinates for cities
- You can measure distance between two vectors

3 Common Retrieval Techniques

Keyword-based

Finds exact word matches such as IDs, emails

E.g., search query: "laptop repair"
Finds: "laptop repair guide", "repair@email" (no ranking)
Misses: "fixing laptops" (fixing ≠ repair), "repairing computers"

Full-text

Matches word variations, ranks results

e.g., search query: "laptop repair"
Ranked: "Fixing Laptops" > "Computer Tips" > "Support email"
Misses: "Device troubleshooting" (no word overlap)

Vector-based

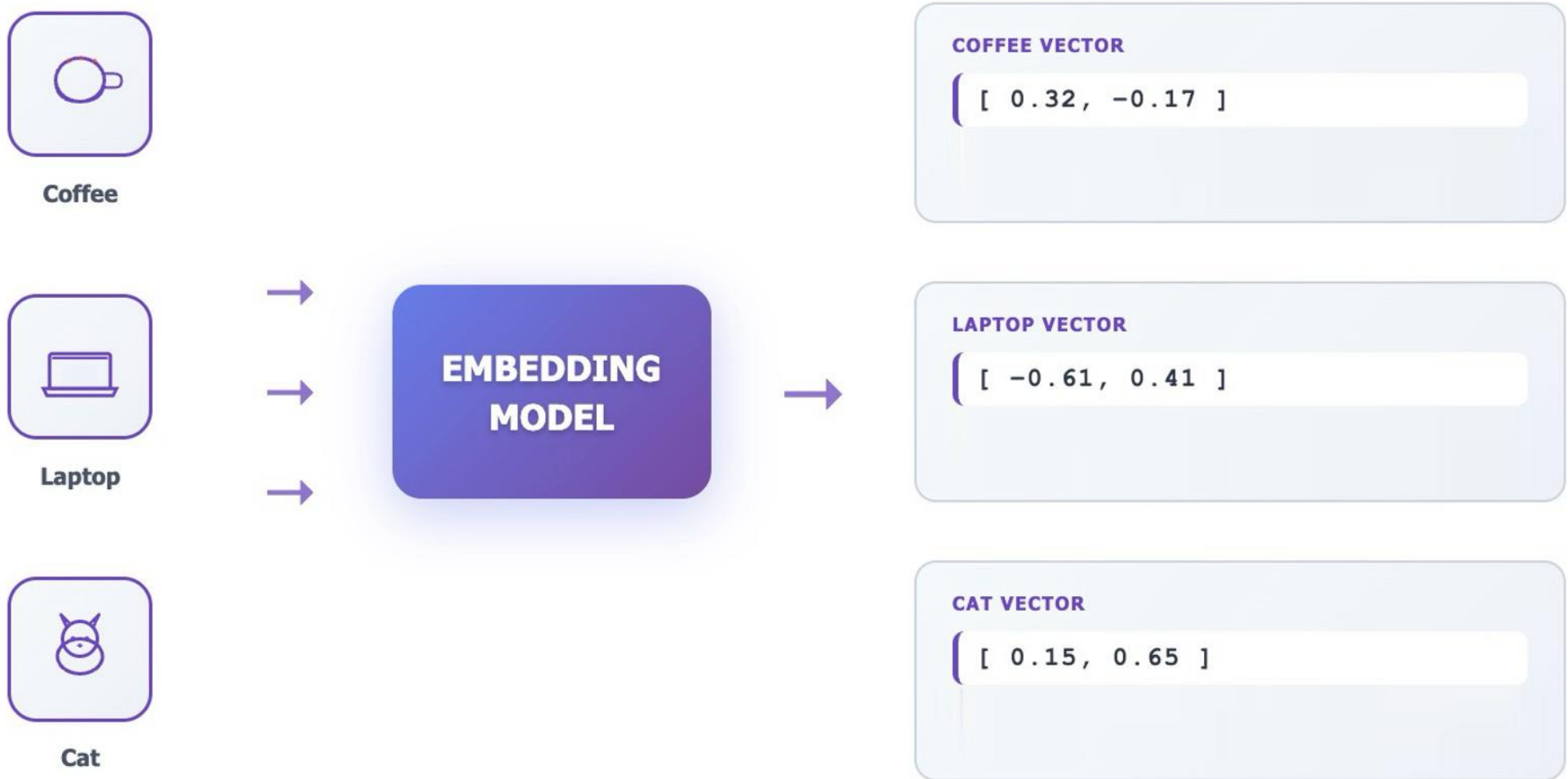
Find similar ideas, not just word matches

e.g., search query: "laptop repair"
Finds: "Fixing Laptops", "Device Troubleshooting"
✓ Matches meaning - works with any vocabulary

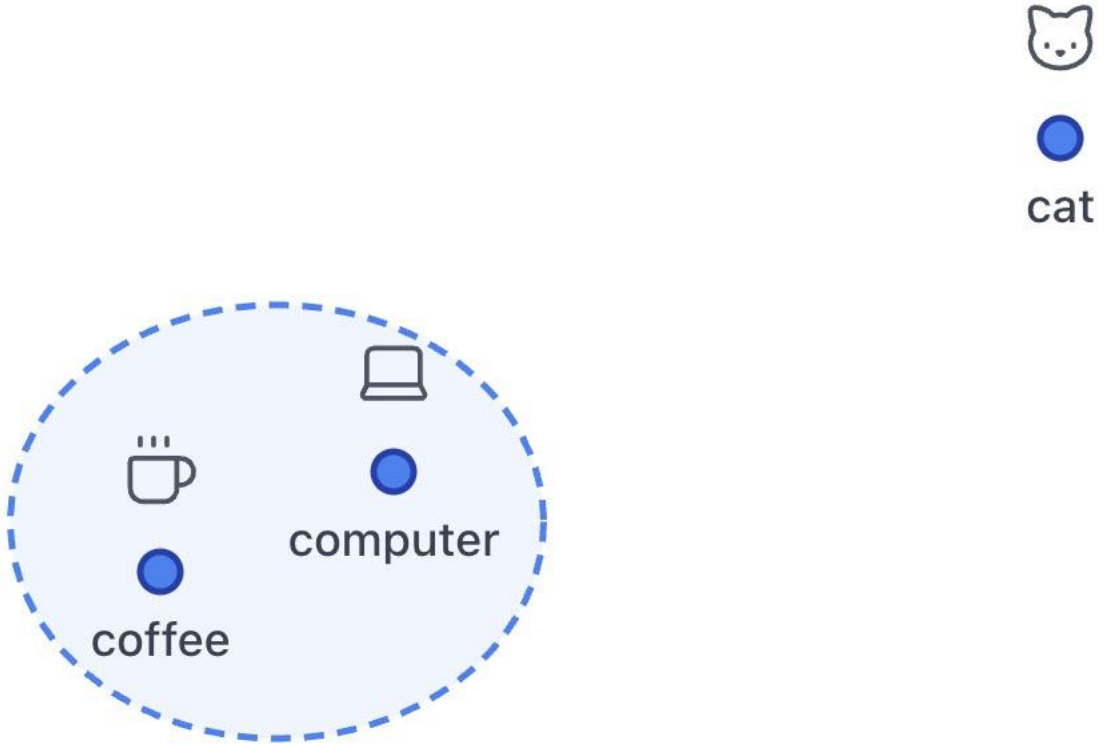
Which Two Are More Similar? (What is Similarity, BTW?)

Vector Embedding: Turning Data Into Vectors

Transforming real-world objects into numerical representations

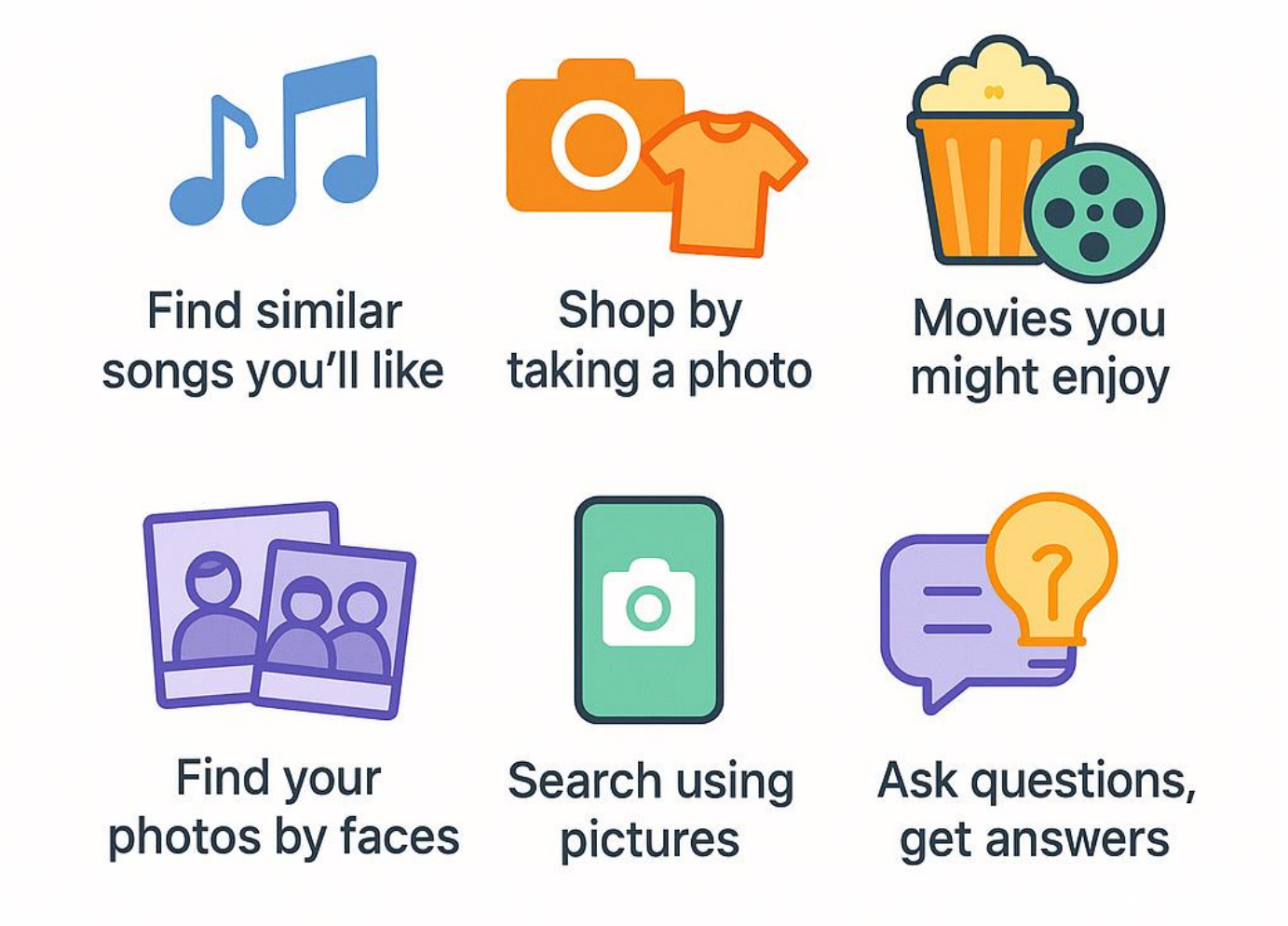


Real-world Entities on Vector Space

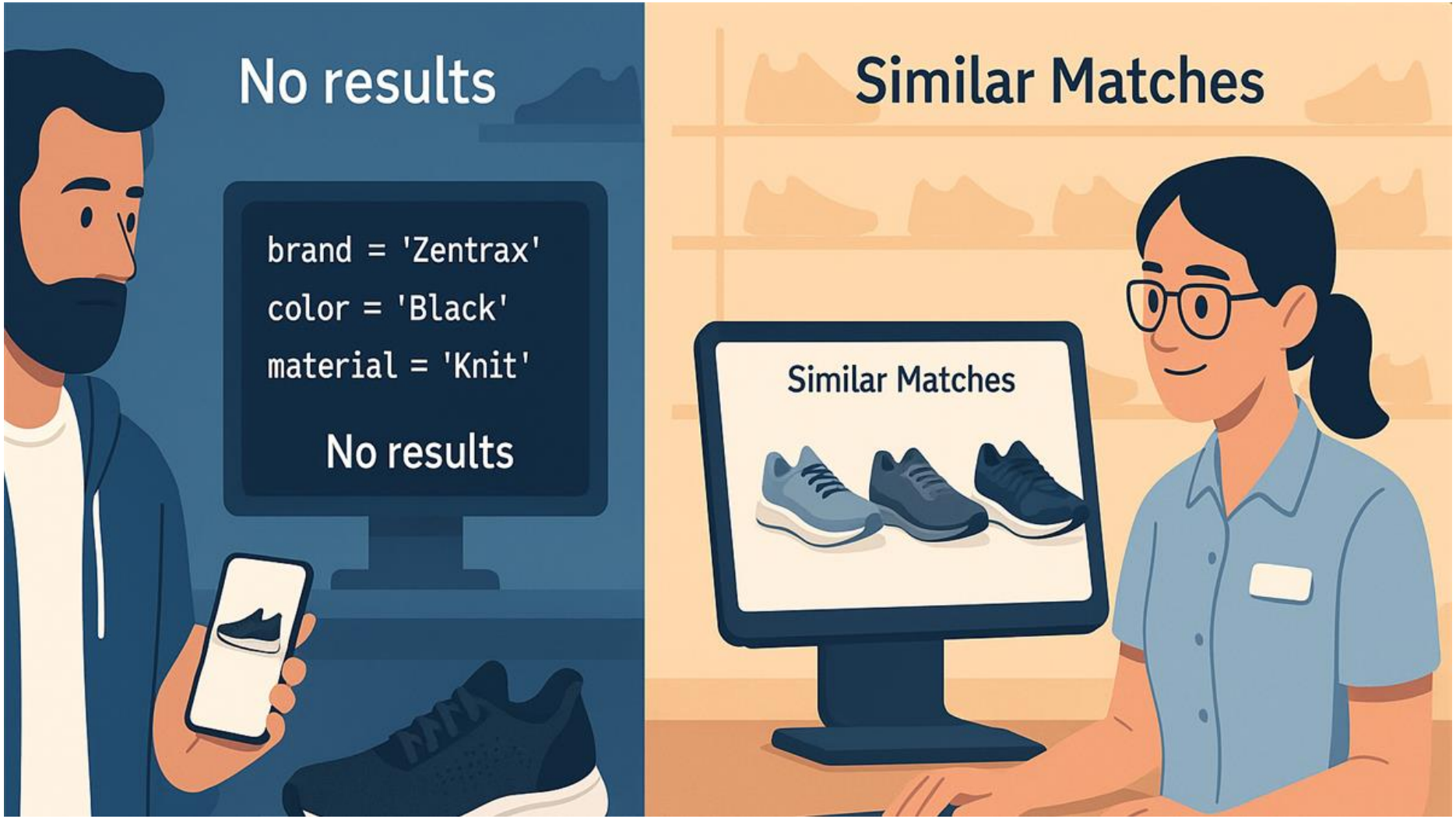


So, what? What can Vectors do for you?

Common Use Cases for Vector Search



Semantic Search Over Tabular Data



Basic LLM vs RAG

Closed-Book: Vanilla LLM

Relies only on internal knowledge
— may forget or guess

Open-Book: RAG
(Retrieval-Augmented Generation)

Retrieves supporting information
before answering

Retrieval Augmented Generation (RAG)

Non-RAG LLM

Traditional Large Language Model

User Query
Direct input to LLM

↓

Static Training Data
Fixed knowledge cutoff

↓

Generation
Response based on training only

RAG System

Retrieval-Augmented Generation

User Query
Input converted to vector

↓

Retrieval
Search external knowledge base

↓

Augmentation
Combine query + retrieved context

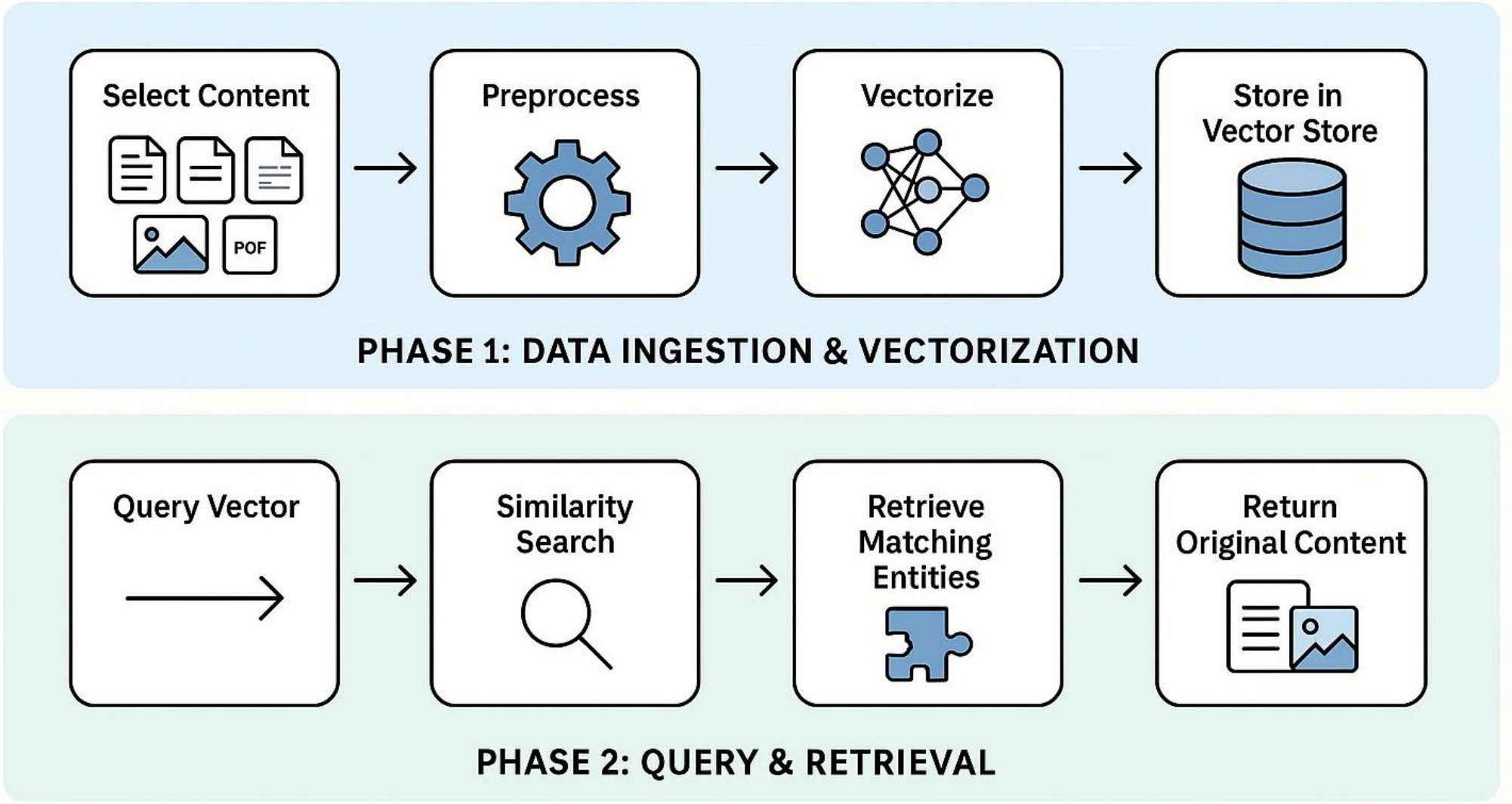
↓

Generation
LLM generates grounded response

Knowledge-Cutoff / Static Knowledge, Hallucinations

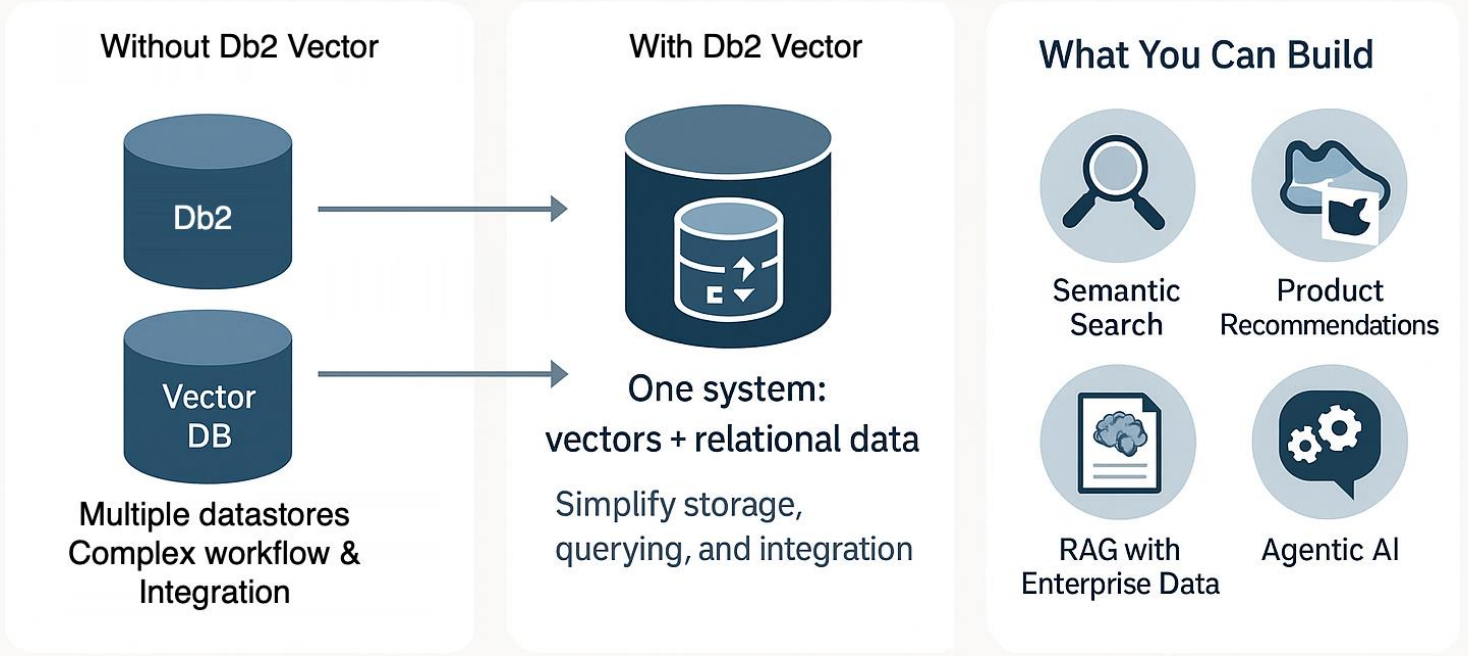
Dynamic Knowledge, Domain-Knowledge, Source citation

Meta Workflow of Vector Search Use Cases



Db2 Unlocks Semantic Search and GenAI Workflows

Built-in Vector Support for Modern AI Use Cases



Your data. Your vectors. One Db2 platform.

Vector Support in Db2

VECTOR type:

- FLOAT32 vectors
- INT8 vectors
- Max dimension:
  - 8168 (FLOAT 32)
  - 32672 (INT8)

VECTOR Functions:

- VECTOR_DISTANCE – between vectors
- VECTOR constructor: string -> vector
- VECTOR_SERIALIZE – vector -> string
- *VECTOR_NORM* – vector magnitude
- VECTOR_DIMENSION_COUNT – number of dimensions

LLM Apps Dev Support:

LangChain Connector

🎉 NEW RELEASE

Db2 12.1.3
Production-Ready Vector Support

Data movement utilities • SQL routines • LLM framework integration

LangChain & LlamaIndex

Native Python packages for RAG apps

SQL Routines

Custom AI logic at database layer

Data Movement

LOAD, backup, schema migration

Columnar Engine

VECTOR Data Type

Enterprise Scale

Build LLM Apps with Db2

LangChain

LlamaIndex

Native Python

Quick Start

RAG Apps

AI Agents

Db2 VECTOR Data Type

Cosine • Dot Product • Euclidean

✓ Enterprise Reliability ✓ Python-Native APIs ✓ No Low-Level SQL

Custom AI Logic in SQL Routines

Encapsulate AI application logic at the database layer where your data lives

Domain Logic
Custom metrics

Reusable Operations
Normalization, validation

Hybrid Queries
Vector + structured filters

Batch Processing
Compute statistics

Consistency
Same logic across apps

Columnar Engine
Native optimization

✓ Write logic once in SQL • Execute where data resides • No network overhead

UDFs & Stored Procedures with VECTOR data types

Production-Grade Data Movement

Vector support across Db2's data movement utilities

Db2 12.1.2
IMPORT • EXPORT

→

Db2 12.1.3
+ 6 More Utilities

LOAD
High-volume insertion

ADMIN_COPY_SCHEMA
Clone schemas

ADMIN_MOVE_TABLE
Relocate tables

db2move
Migrate databases

External Tables
Query in data lakes

Logical Backup
Schema-level backups

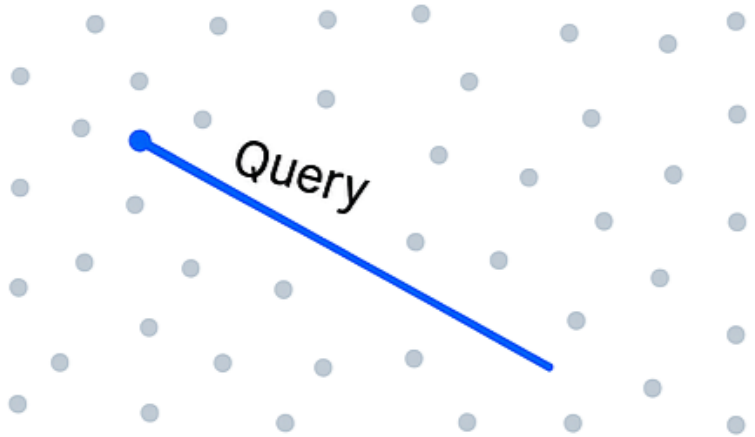
Key Capabilities

Full & Incremental · Table-Level · External File · Production-Ready

✓ Move vector data using the same tools as traditional data types

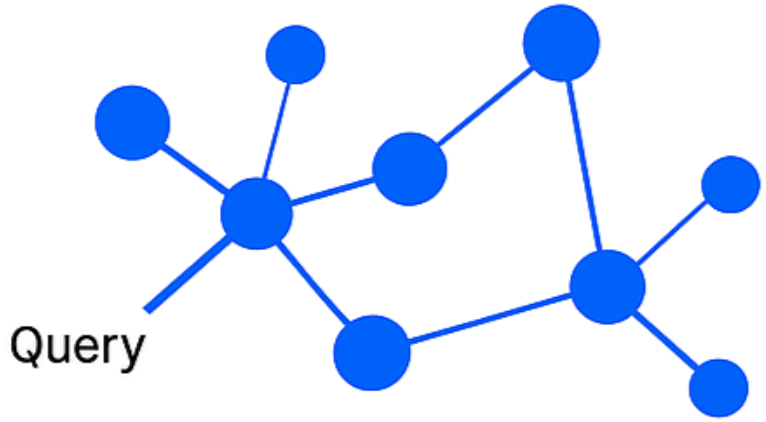
EAP Feature: Db2 Vector Index – Scaling Vector Search

Exact Search



Exact Search
– slower for large collections

Vector Index



Indexed search
– faster at million / billion scale

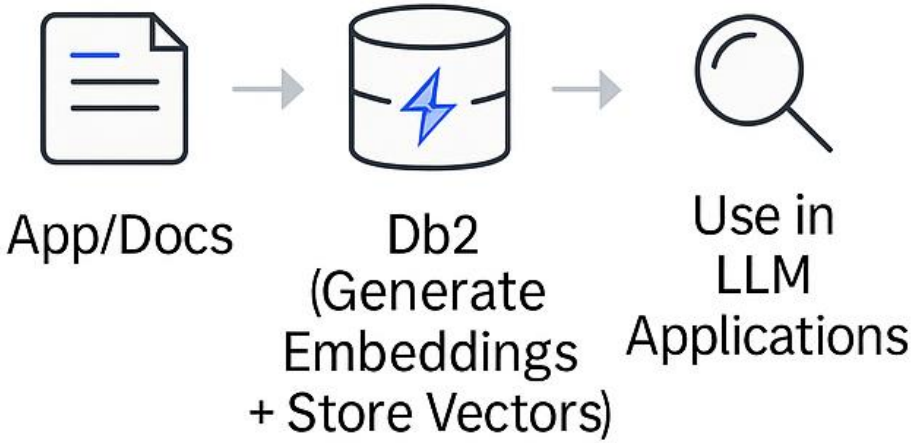
Faster vector search with Db2 Vector Index

EAP: SQL-Native Embedding Generation

The Problem:
Too Much Glue Code

Challenges:
Time • Fragility • Cost

The Solution
SQL-Native Embeddings in Db2



With Db2, register your embedding endpoint once. Then generate, store, and search embeddings SQL.

Demo

SQL-Native Embeddings in Db2

Simplifying LLM application development with fewer moving parts

Ibm.biz/learndb2ai

Connect with Me

Shaikh Quader

AI Architect @ IBM Db2 | Ph.D. Candidate in AI | IBM Master Inventor | Sharing lessons from building & d...