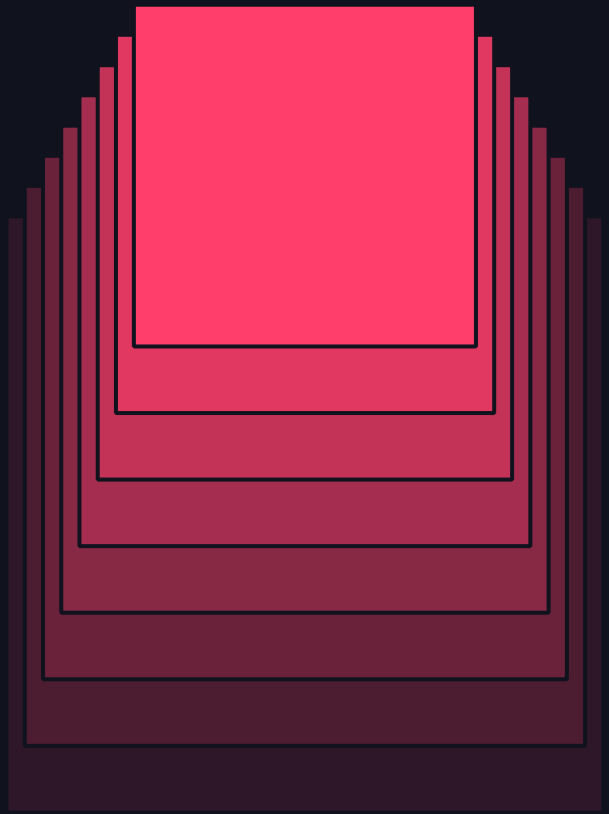
DATA+AI SUMMIT BY databricks

Introduction to Mosaic AI Vector Search



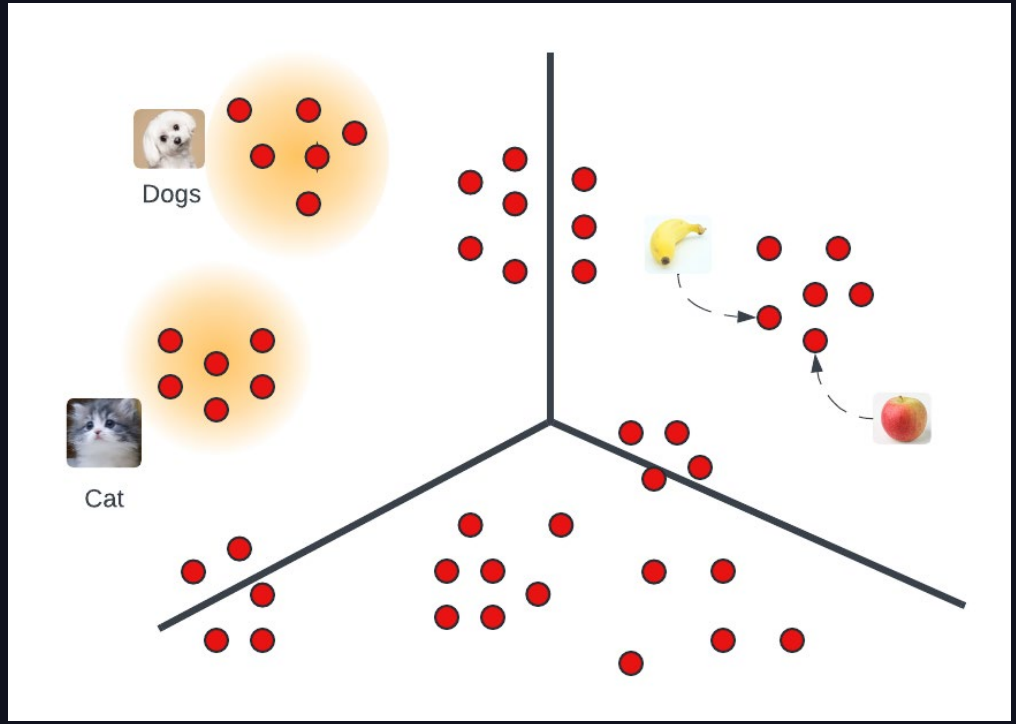
Akhil Gupta, VP of Engineering, AI Systems
Jun 13, 2024

Vector Search

Basic Concepts

Embeddings

- A numerical representation of data as a point in N-dimensional space
- Vector of two data objects similar to each other will be close to each other.
- Generated using models

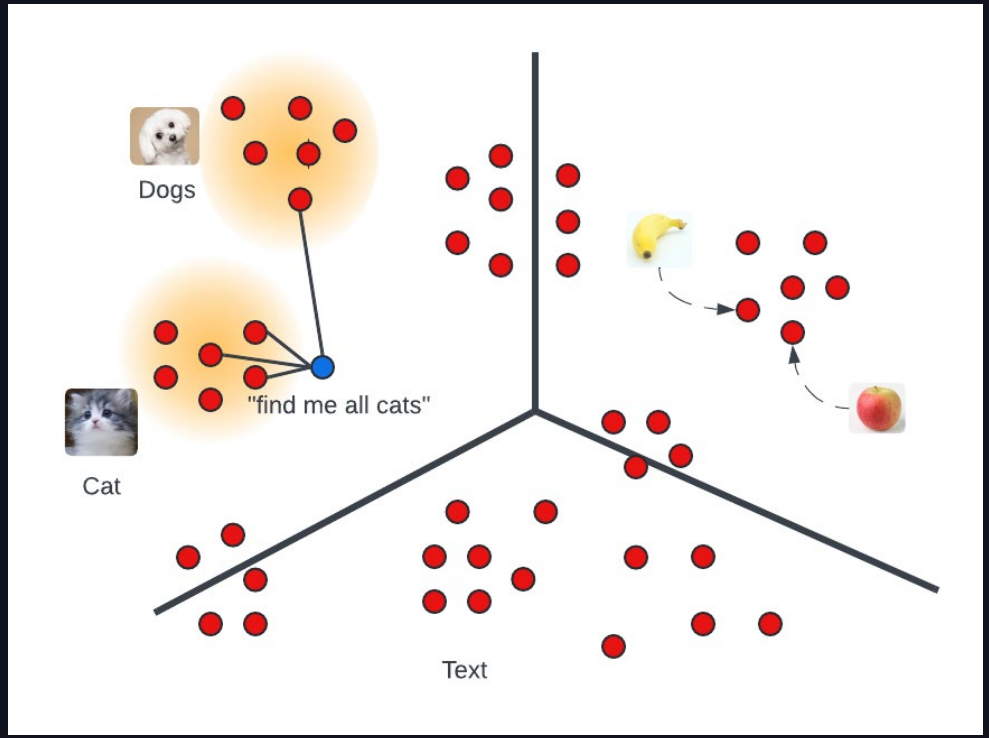


Vector Search

Basic Concepts

Retrieval Algorithm

- Nearest Neighbor
  - Approximate Nearest Neighbor (ANN) vs. Top-K Nearest Neighbor (KNN)
  - Hybrid Search
- Trade-off between latency vs. recall
- Indexing techniques matter.

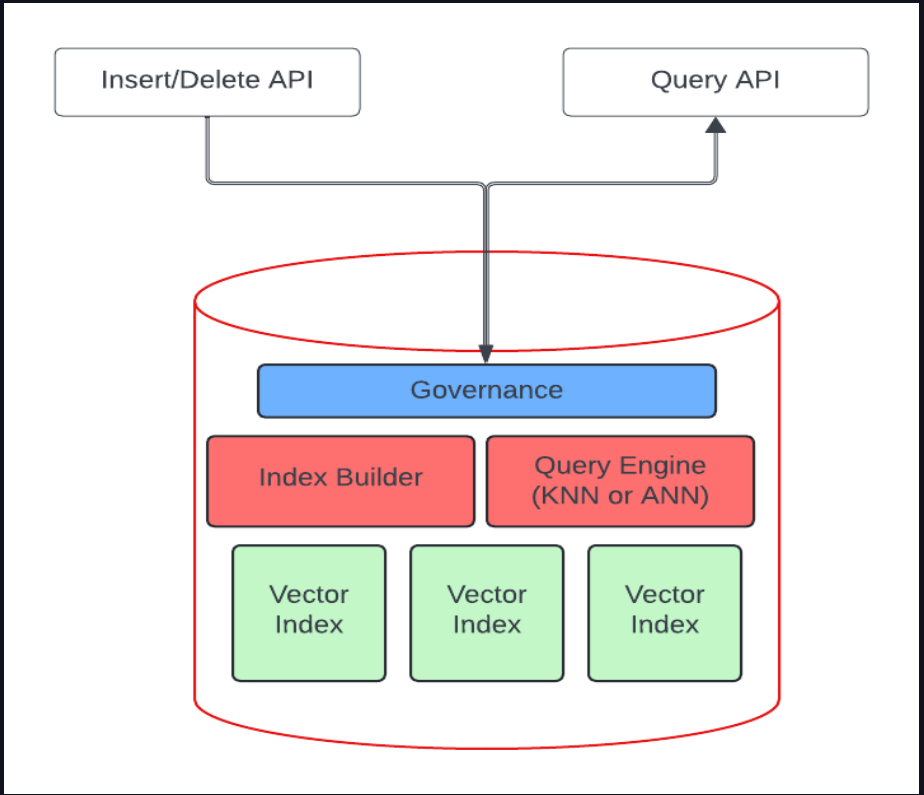


Vector Search

Basic Concepts

Database

- Indexing
- Scalability
- Performance
- Durability
- Governance



Vector Search

Basic Concepts

Embeddings

Retrieval Algorithm

Impacts Quality

Vector Search

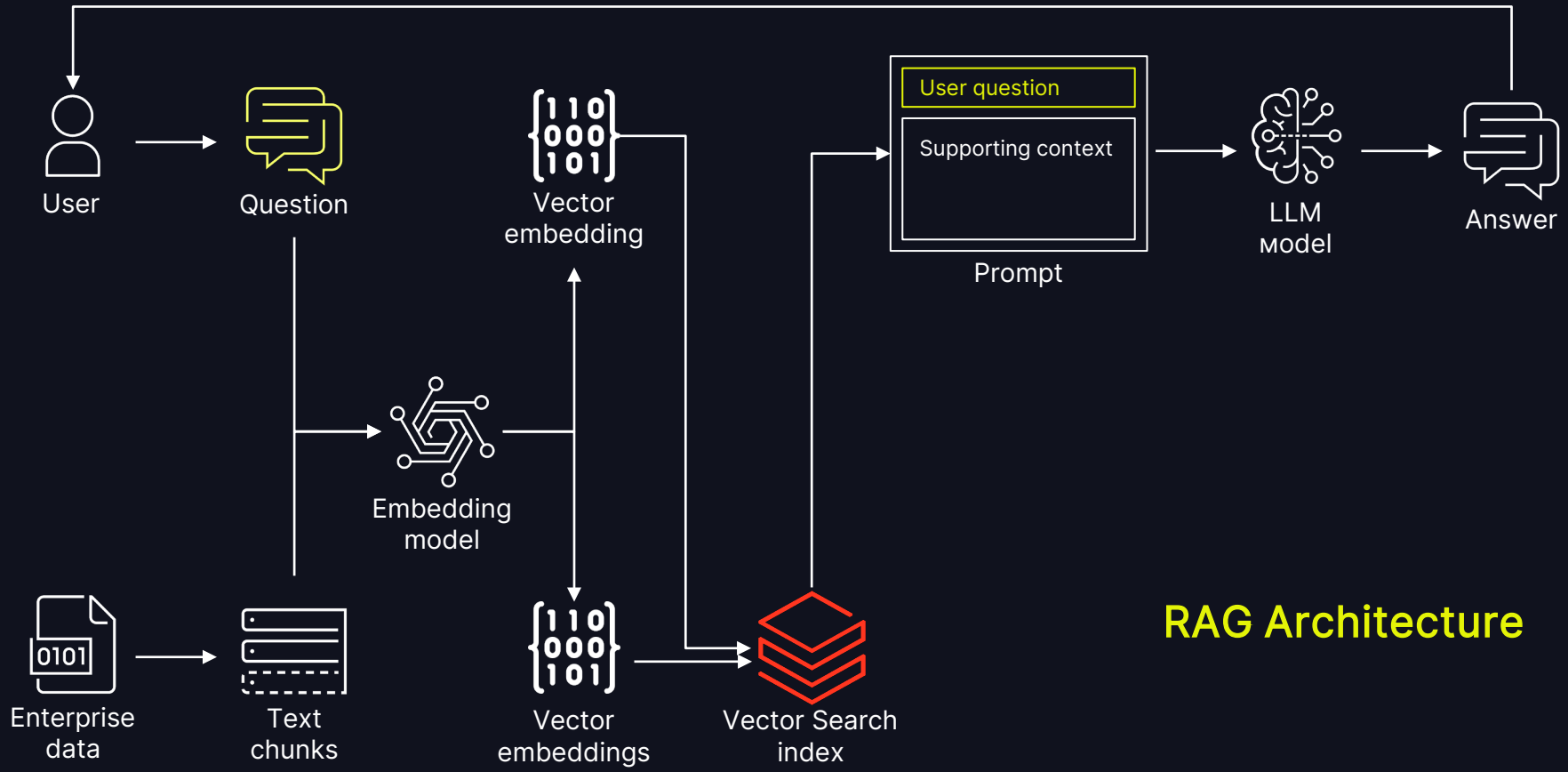
Basic Concepts

Database impacts performance, security and ease of use

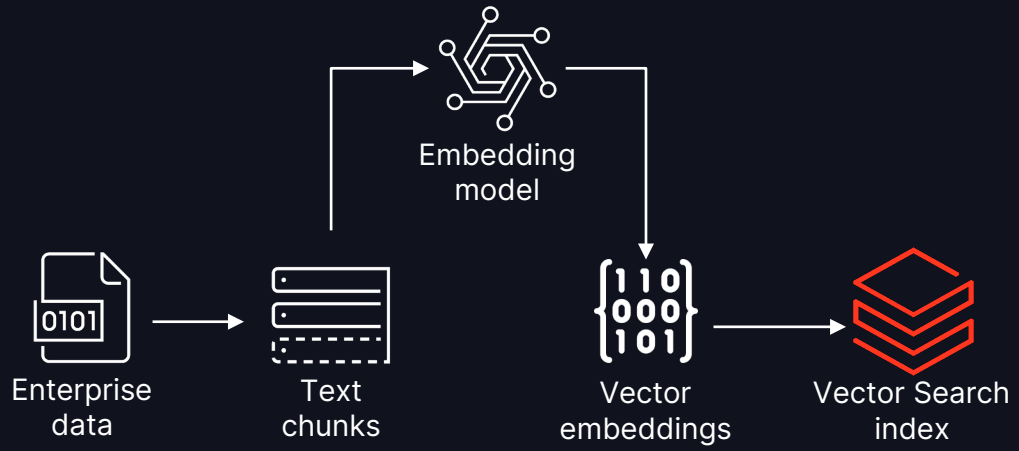
Vector Search

What it is good for and what it is not good for

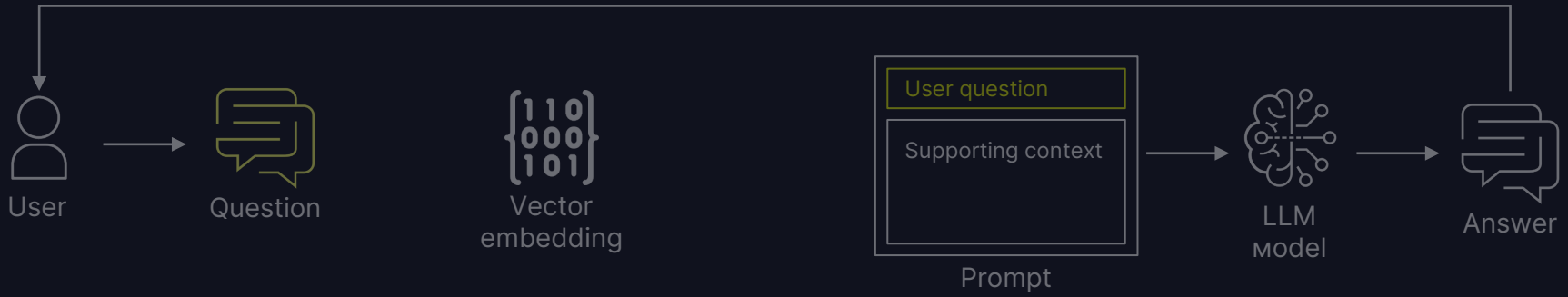
- Use it for semantic search over unstructured data
  - Text, Video, Audio
- Not good for typical database SQL-style queries
  - Aggregation, Joins
- A must-have component when building GenAI applications
  - Critical to reduces hallucinations and provide better context to LLMs
  - Example GenAI Applications - RAG, Sentiment Analysis/Classification

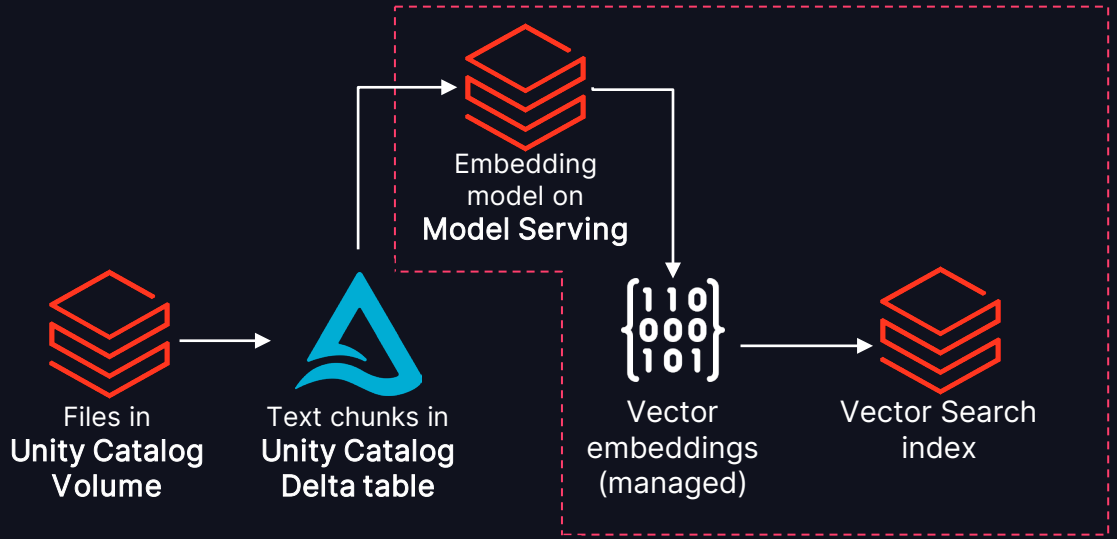


RAG Architecture



Vector Search





Mosaic AI Vector Search

Mosaic AI Vector Search

Some Stats

1000+	200+	400%+
Weekly Active Customers	Large Scale Deployments	YoY Growth

Evaluating Vector Search

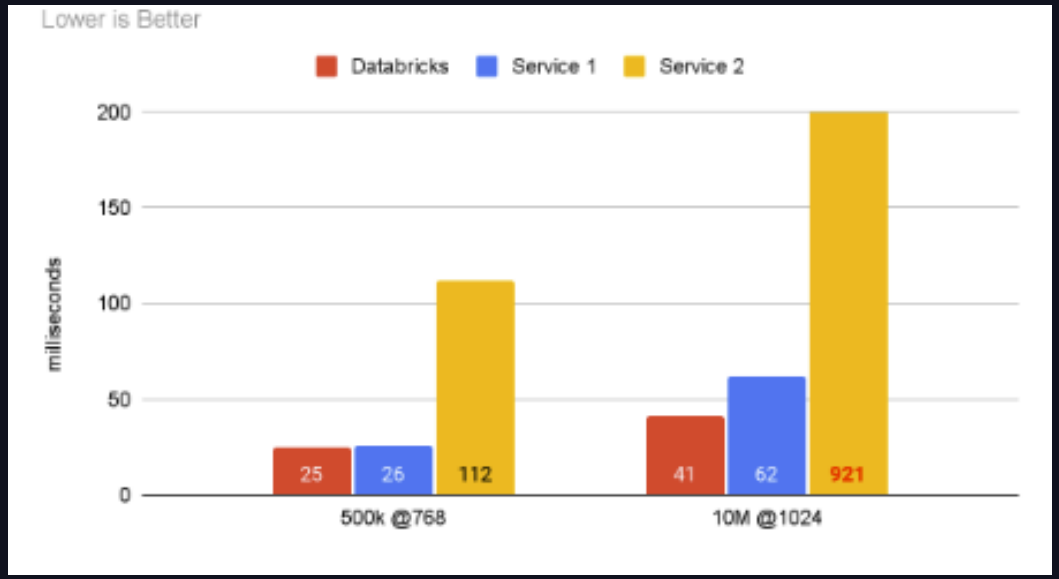
Key criterion for evaluating vector search

- Scalability and Performance
  - *How many embeddings do you need to support?*
  - *What are your latency requirements?*
- Ease of use and management
  - *How easy is it to set up and operate?*
- Governance
  - *Does it respect your existing security and governance policies?*
- Retrieval Quality
  - *Does it provide you with all the knobs you need to improve the quality of your retrieval?*

Mosaic AI Vector Search

Scalability and Performance

- Autoscales with zero downtime
- Optimized for high performance at low cost
- Scales to hundreds of millions of vectors

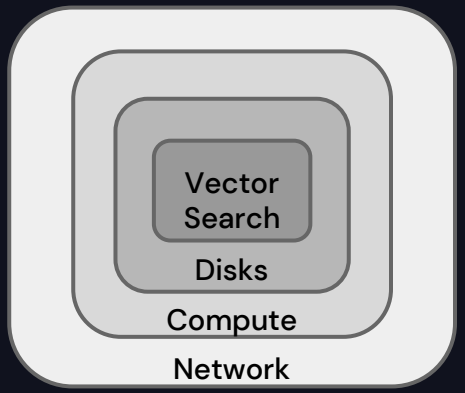


P90 Latency of Mosaic AI Vector Search

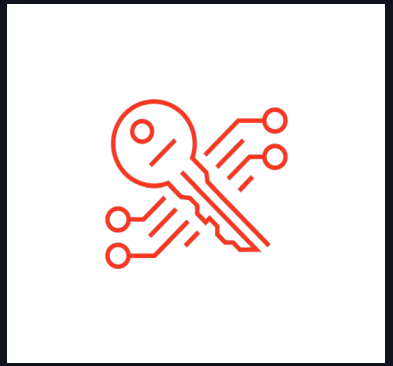
Mosaic AI Vector Search

Governance

Secure by Design



Complete Data Control



Integrated with Unity Catalog

Mosaic AI Vector Search

Ease of Use

Serverless

- No infrastructure to maintain
- Autoscales to workloads

Integration with Lakehouse

- Delta Sync API makes it trivial to create and update vector indexes on database
- Integration with Unity Catalog provides out-of-box governance

Integration with Databricks Platform

- Integration with Model Serving makes embedding generation easy
- Integration with Agent Serving and Custom Apps makes building GenAI applications easy

Mosaic AI Vector Search

Retrieval Quality

Data Pre-Processing

- Parsing content (pdf, html)
- Chunking

Embedding Model

- Use off-the-shelf model
- Fine tuned model
- Train custom model

Retrieval Algorithm

- LSH vs. IVF vs. HNSW
- Hybrid Search