Couchbase

AI-powered Adaptive Applications and Vector Search

JUNE 2024

Agenda

New Opportunities with AI

Vectors and GenAI

Why Couchbase for Vectors and GenAI

AI Creates an Opportunity for a new Application Design

Everywhere

Interactions are mobile

Responsive

Latency is unacceptable

Contextually Proactive

Inputs: who, what, where, why, when, real-time data

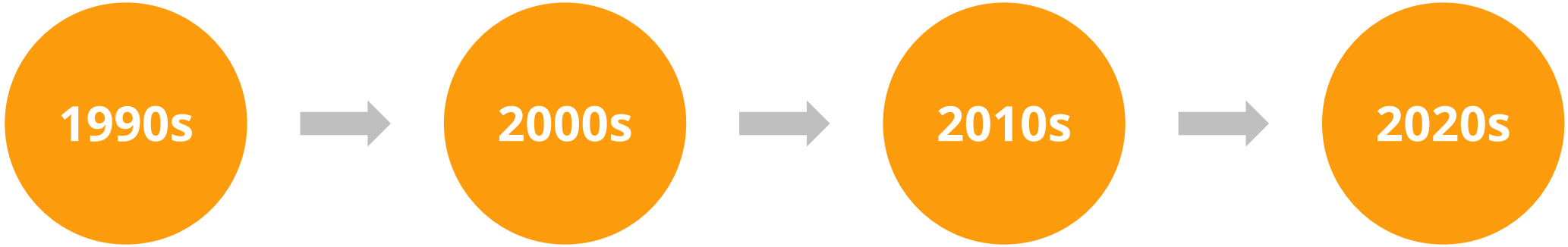
Hyper Personalized

Accurate responses that fit the situation perfectly

Innovative

The best adaptations are fueled by trustworthy data

The AI Shift is Happening



Client Server Applications

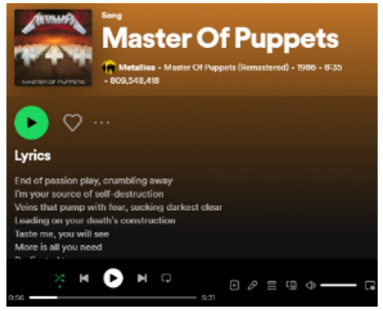
Web Applications

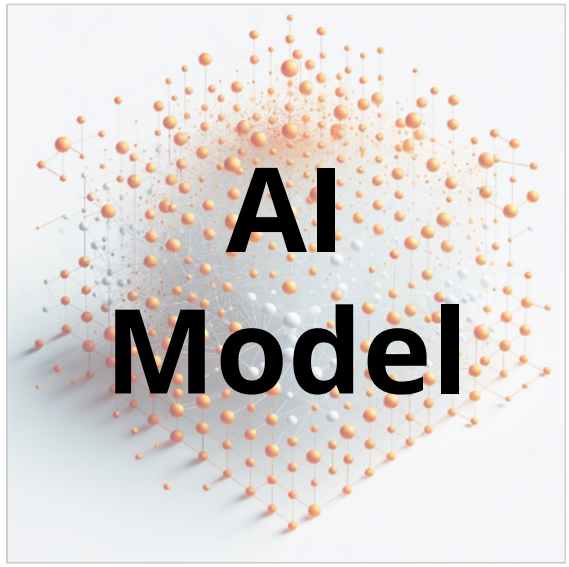
Smart Mobile Applications

AI-powered Applications

Vectors and GenAI

What is a Vector, Vector Search and Why is It Important?





Embeddings = Arrays of Floats

[0.9229, 0.7824, 0.194, 0.0402, 0.206, 0.4446, 0.5875, 0.4214, 0.2371, 0.4038, 0.9229, 0.7824, 0.194, 0.0402, 0.206, 0.4446, 0.5875, 0.4214, 0.2371, 0.4038]

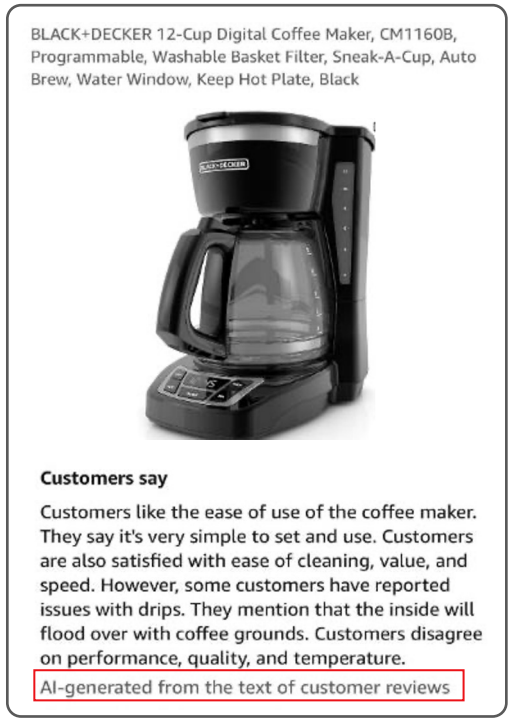
[0.0509, 0.8862, 0.179, 0.4116, 0.5995, 0.769, 0.5282, 0.8079, 0.0468, 0.272, 0.0509, 0.8862, 0.179, 0.4116, 0.5995, 0.769, 0.5282, 0.8079, 0.0468, 0.272]

[0.3819, 0.1375, 0.0828, 0.0856, 0.9081, 0.1556, 0.6251, 0.7356, 0.2964, 0.2198, 0.3819, 0.1375, 0.0828, 0.0856, 0.9081, 0.1556, 0.6251, 0.7356, 0.2964, 0.2198]

Use Cases

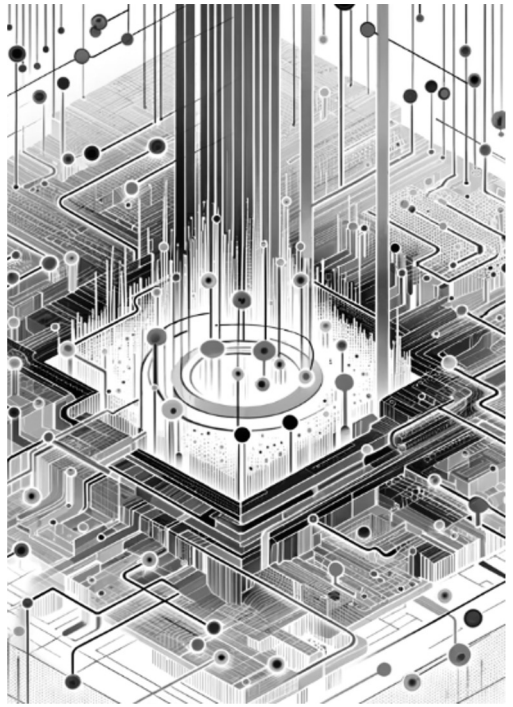
AI-powered Chatbots and Applications

Content Generation

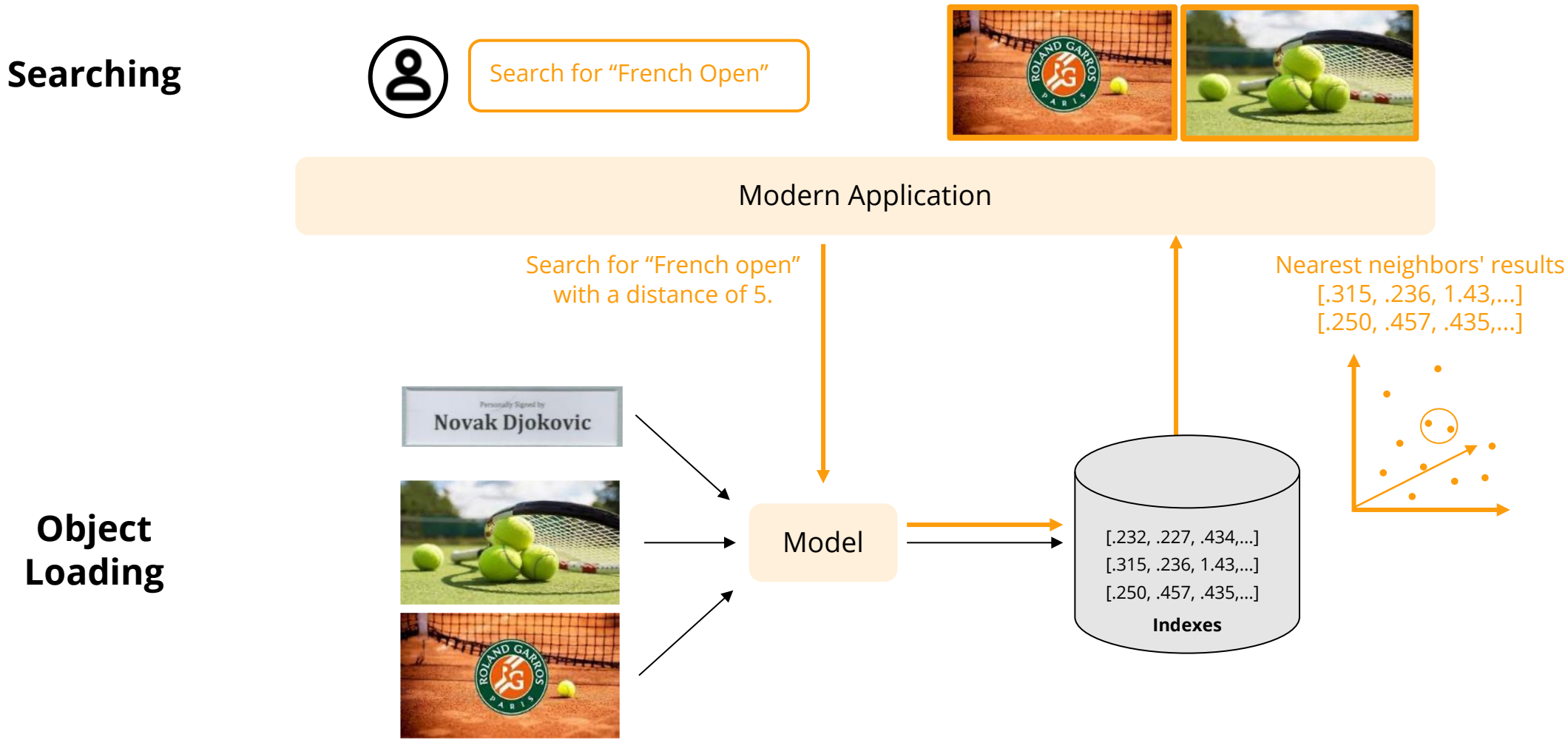


Advanced Semantic / Hybrid Search

Data Analysis: Classification / Anomalies



Vector Similarity Search Example



Key Risks with Apps that Share Data with AI Models

These are C-level showstoppers if they are not addressed

Sharing proprietary and sensitive data

Sharing data that induces hallucinations

Retrieval-augmented generation (RAG) enhances the accuracy and reliability of GenAI-powered applications

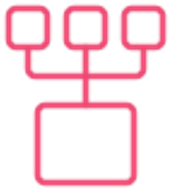
Why Couchbase for Vectors and GenAI

Couchbase

Extending Couchbase Search for Vectors

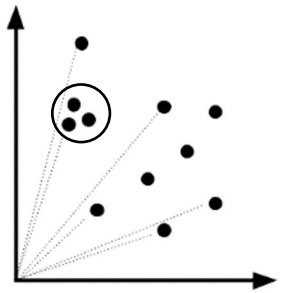
No Separate Database Needed

```
my:     Doc 1, Doc 2, Doc 3
dog:    Doc 1, Doc 2, Doc 81
has:    Doc 1, Doc 2, Doc 3
fleas:  Doc 1, Doc 81
…
```



	Full-Text Search	Vector Search
Pre-query work	• Documents scanned for words • **Inverted indexes created** linking word to document id and text positions.	• Text, image, audio, other scanned into vectors for “nearest-neighbor” similarities • These multidimensional **embeddings are indexed** as a special “vector” type.
Query Process	• **User** wants docs contains “Big Cat” • Text analysis performed, to find same or similar misspelled words, term frequency, other criteria • Documents are scored	• **User** sends query sentence “Big Cat” that is converted to its own vector embedding • Analysis performed, to find top N nearest matches with other • Results returned to AI-based Application • App sends top N result (context) plus the query (prompt) to AI model
Results	• Documents returned to **User** • Sorted by highest ranking score **Results**: Docs w/ “Cat in a Hat, Big Hat”	• **Retrieval Augmented Generation** • AI model generates output • AI app delivers result to **User** **Results**: Text discussing tigers, leopards, lions and pumas

.2758, .2637, .4315,
.1372, .6194, .2164,
.1654, .2856, .2785,
.4615, .7516, .3194,
.5456, .5229, .9567,
.2342, .5234, .6345,
.3634, .3153, .9502,
.9124, .2856, .2785,
.9634, .7516, .3194,
.5456, .5229, .9567

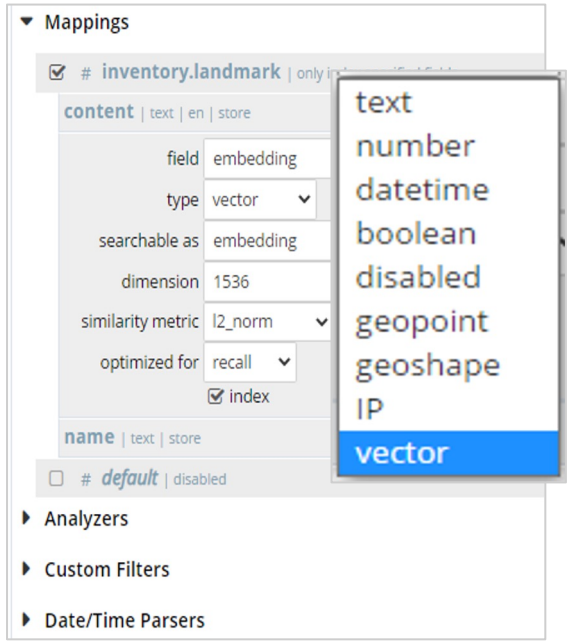


Vector Storage, Indexing and Query

JSON Storage

{ "type": “shoes",
"productId": "CP123456",
"category": “Gym Shoes",
"name": “Beach Sneakers",
"brand": “Ultimate Surf",
},
"description": “The ultimate companion for beach adventurers, designed to seamlessly transition from sandy shores to urban landscapes. This innovative sneaker features a water-resistant, quick-drying mesh upper, allowing your feet to breathe while keeping them dry.",
“descriptionVector": [0.131, 0.339, -0.611,0 .981,...]

Indexes

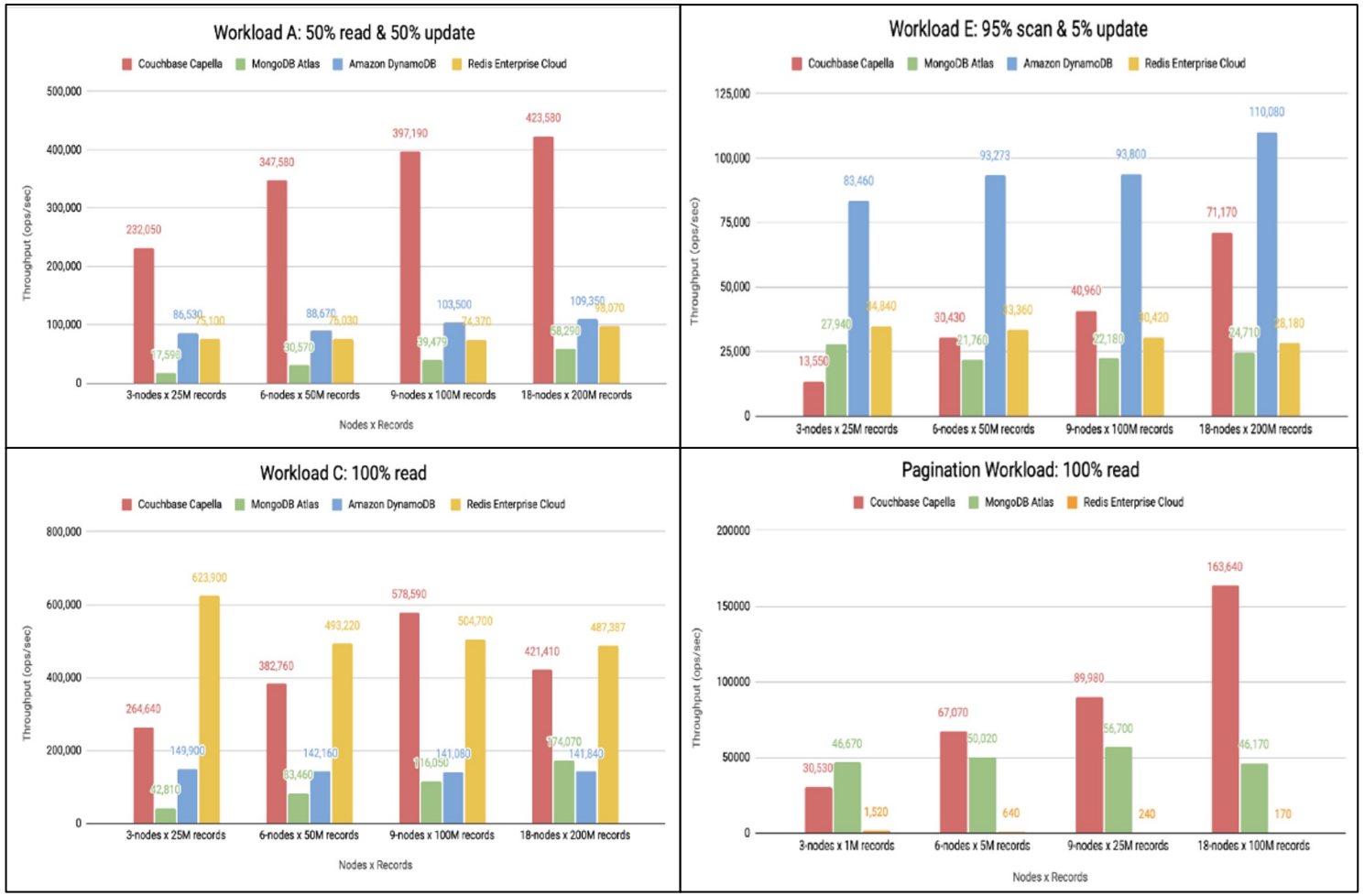


SQL++

```
SELECT *
FROM product
WHERE LOWER(product.type) = ‘shoes’
AND product.size = 11
AND product.price between 50 and 80
/* desc SIMILAR TO  ‘blue running
shoes’ */
ORDER BY
GSI_VECTOR_ORDER(desc_embedding,
{
  "knn": [{
    "field": "desc_embedding",
    "vector": [0.1, 0.334, -0.604, 0.985]
}]}
LIMIT 4
```

Performance Proven via YCSB Tests

Proven Speed and Flexibility



Capella

- Excels at Read and Write
- Scales effectively for multiple workloads
- Best price-performance

MongoDB

- Struggles to scale, strongest at 3 nodes
- Price performance makes it worse
- Weakest performer

DynamoDB

- Excels at scans, but throws excessive errors
- Challenged across other workloads

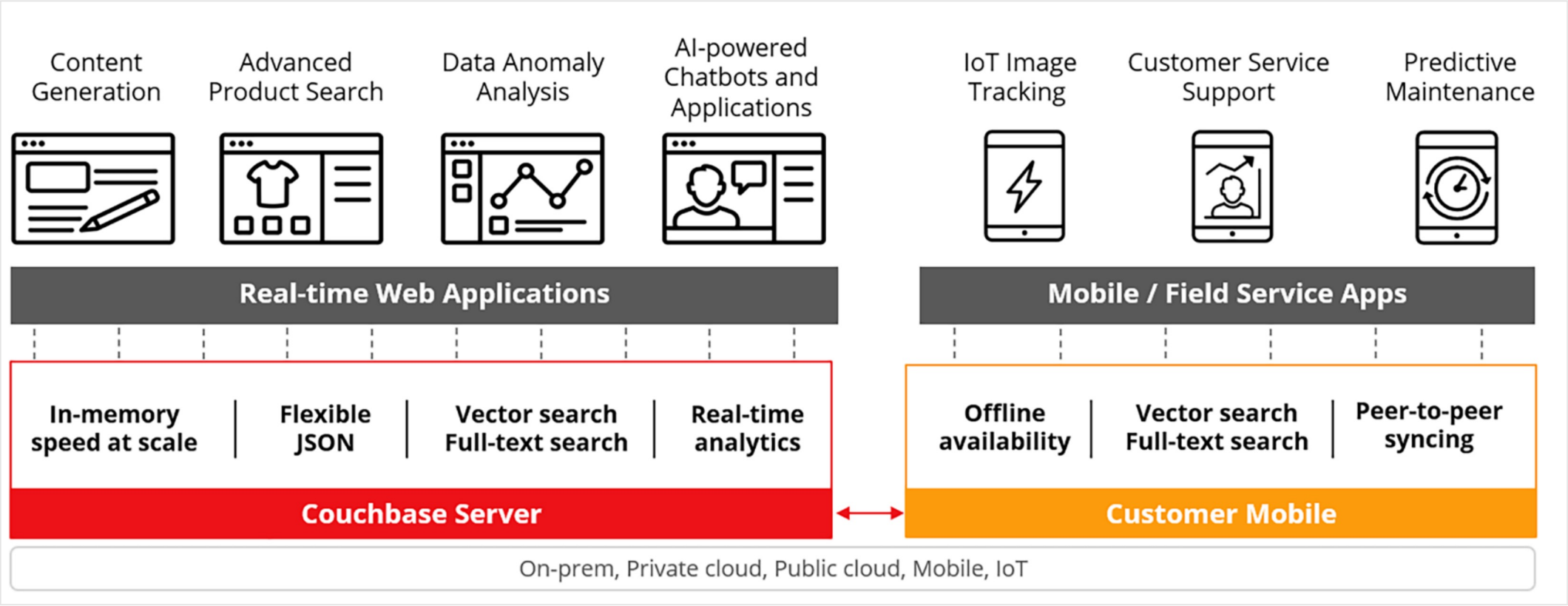
Redis

- Excels at read-only, Capella meets at scale
- Regularly used as cache for Atlas or DynamoDB
- “Fails” Pagination workload

Altoros, April 2023

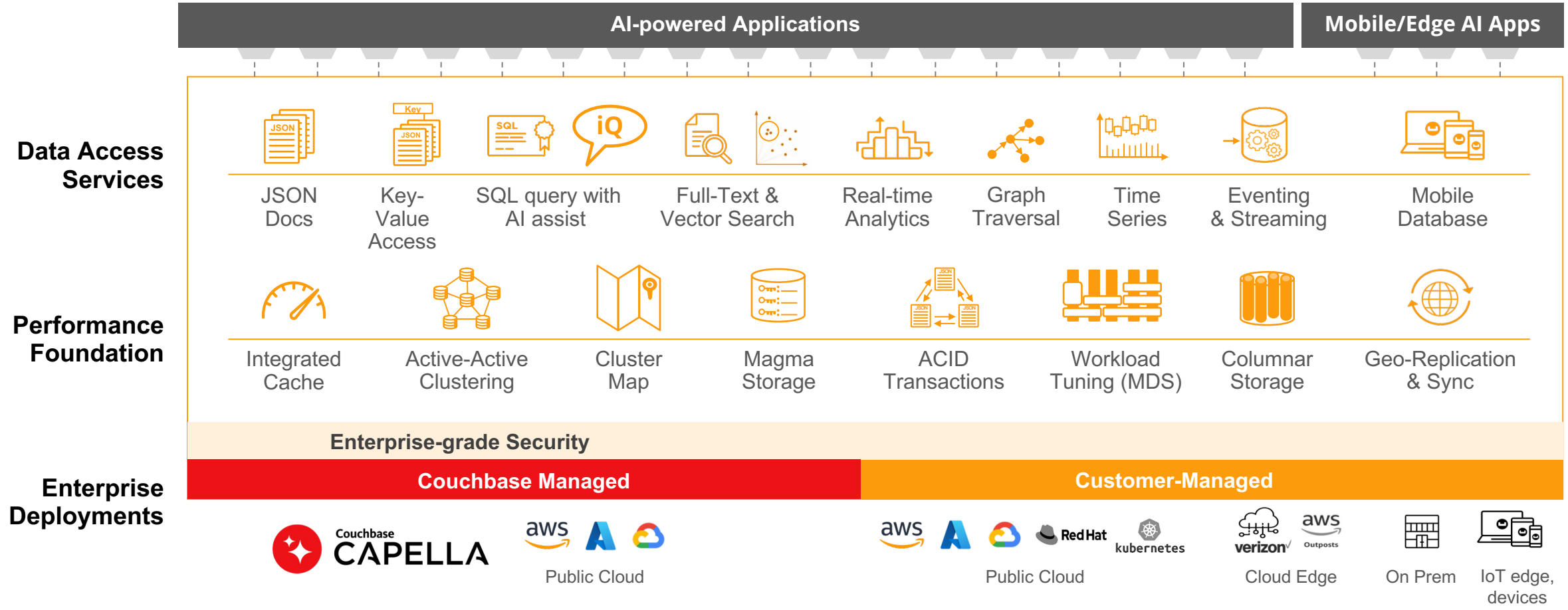
Vector Across our Products

First in the industry to announce support for all 3 deployments: Cloud, on-prem, mobile



Couchbase's Differentiated Architecture

Innovate Faster with Couchbase



Beyond Vector Search: Hybrid Search

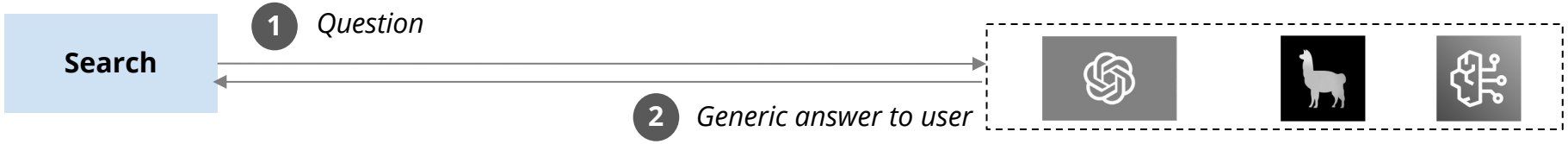
Building real-world use cases

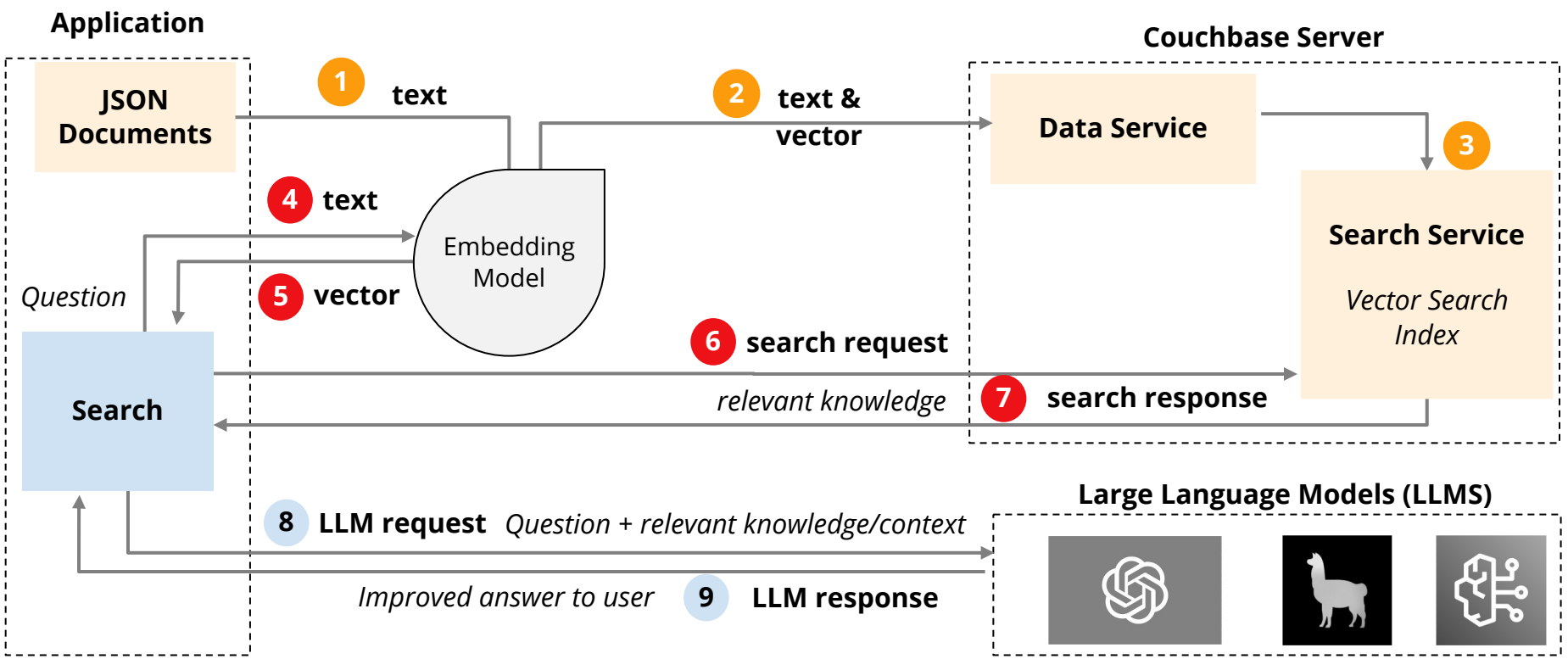
<u>Customer wants new shoes</u>

- Match color and style of an object (semantic/vector)
- Description to mention “casual and fun” (text/fuzzy)
- Price between $100 and $200 (range)
- With rating over 4.5 stars (range)
- Within 15 miles (geospatial)
- Available today in stores (inventory)

RAG Application: Chatbot for Auto Parts Supplier Chain

***Question**: What is the best way to reduce muffler noise on my 1967 Ford Bronco?*





1. Application sends the relevant knowledge base or corpus of documents to Embedding Model adds vectors to JSON documents
2. Uploads JSON documents (text with vector) to the Data Service to load corpus.
3. Search Service receives mutations (upsert or delete) JSON document. maintains indexes across partitions & nodes.
4. Application sends “question” to Embedding Model
5. Embedding Model returns embedding Vector
6. Application calls Search Service with a Vector search with a “question”.
7. Search Service returns top k-NN results of query with requested fields typical text or semantically similar documents. This is the relevant knowledge base.
8. Application combines the initial question with the relevant knowledge and sends the combined information as a prompt to a LLM
9. Foundation Models create factual answer based on relevant knowledge for the original question.

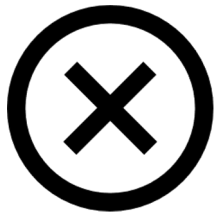
Ecosystem Integration

1. Drive developer productivity
2. Optimize AI processing
3. Enable AI-driven apps anywhere, including the Edge
4. Part of a vibrant AI partner ecosystem

Today

Roadmap

Couchbase: Architected for AI-based Applications



No Separate Vector Database Needed

Simplify | Lower latency | Save

Proven Speed and Flexibility

In-memory architecture | Flexible JSON | Powering Indexing

Vector Across our Products

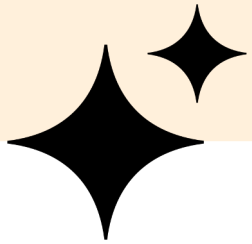
On-prem | Cloud | Offline mobile

Broad Platform Capabilities

Profiles & Catalogs | Vector Search |Text Search | Graph | Real-time Analytics| Powerful SQL

Ecosystem Integrations

LangChain | LlamaIndex



Thank you!

Couchbase