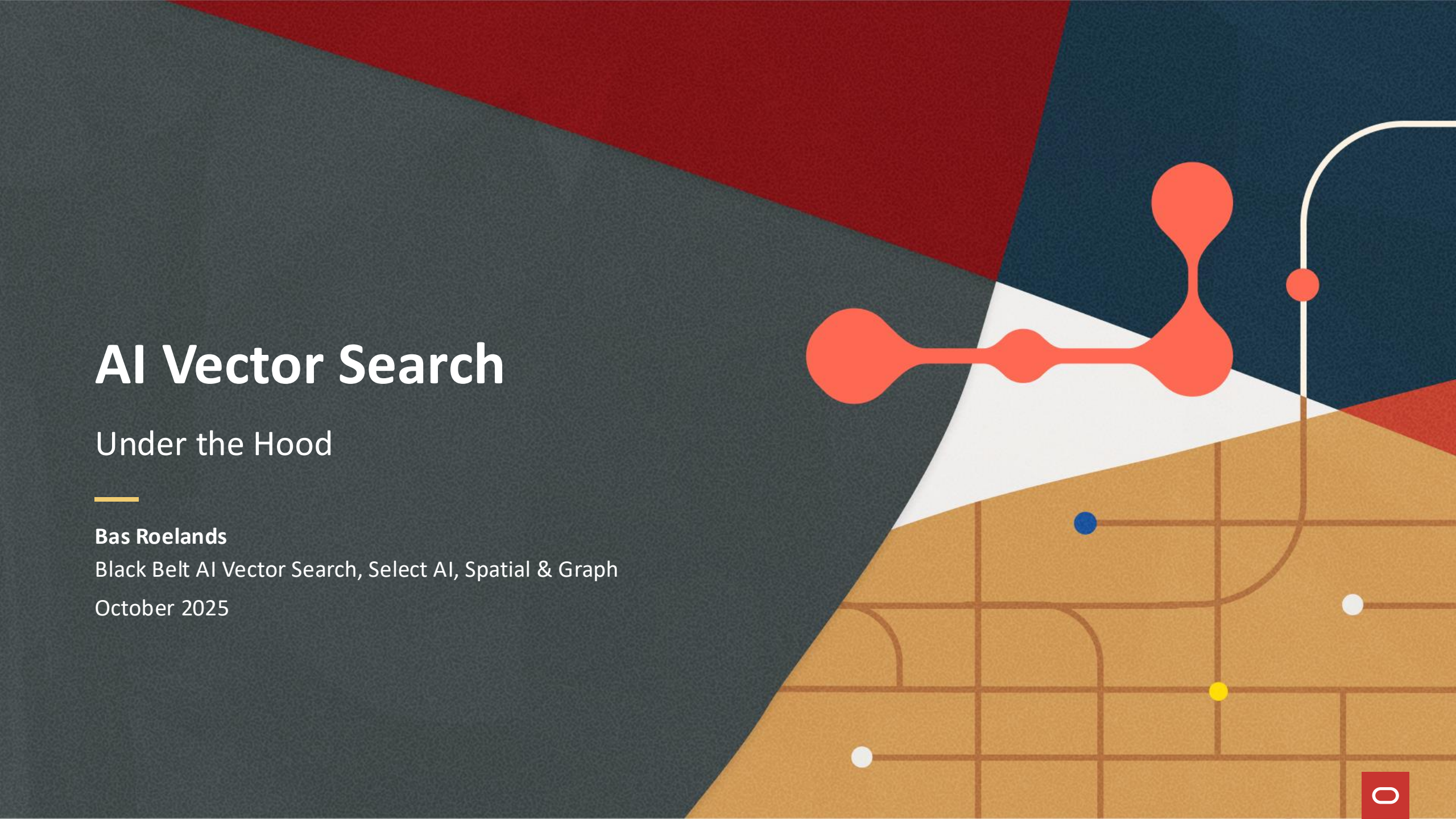
AI Vector Search

Under the Hood

Bas Roelands
Black Belt AI Vector Search, Select AI, Spatial & Graph

October 2025



Agenda

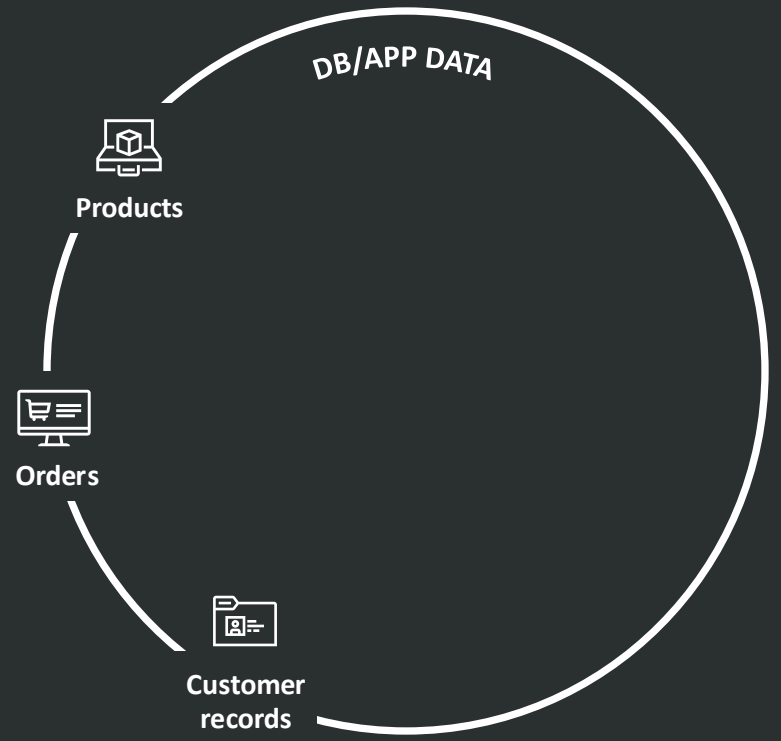
Overview

The [A, B, C] of Vectors

Vectors and RAG

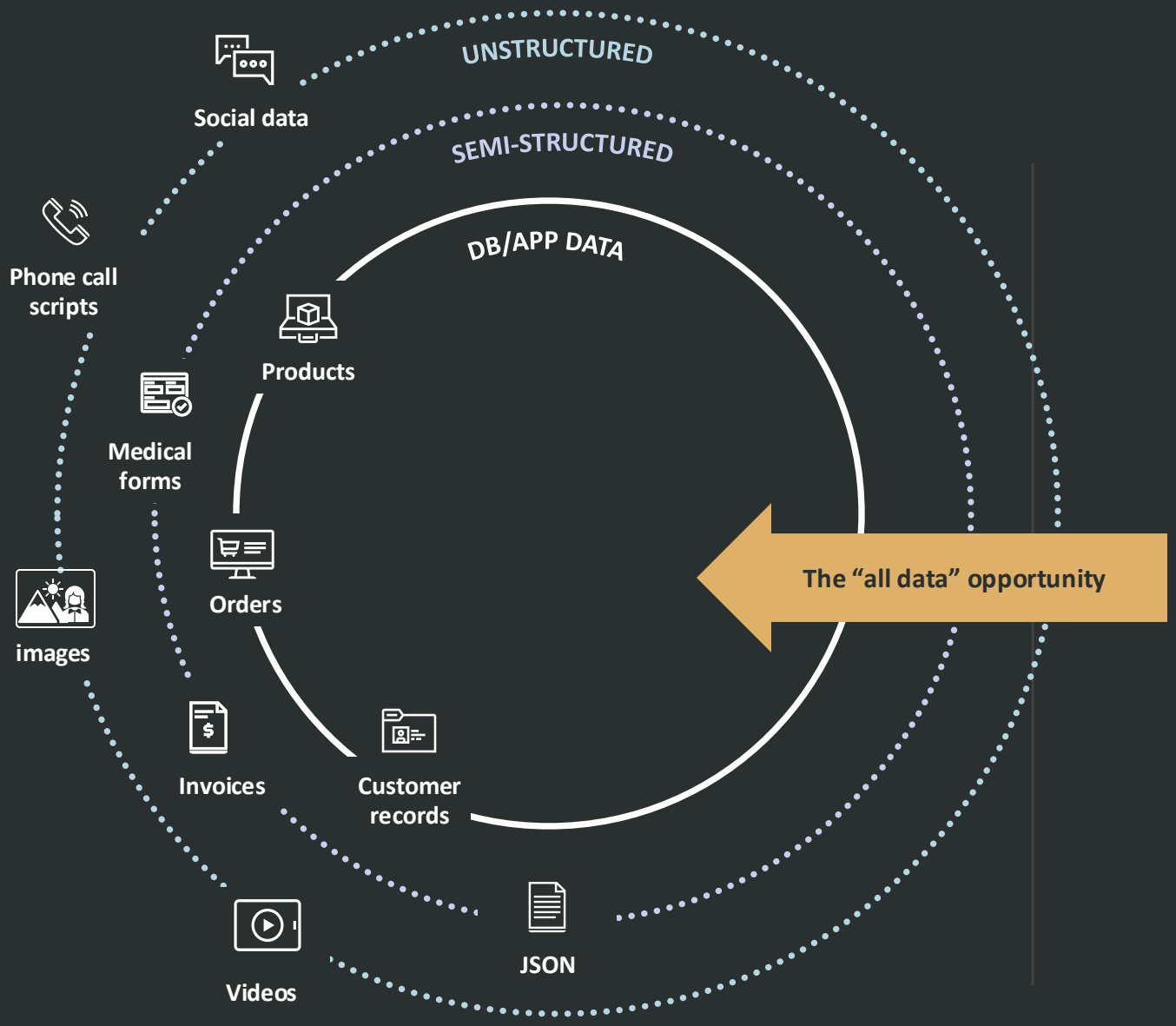
Takeaway

Overview



Today, databases are great at performing precise, value-based searches on structured business data

Find revenue by products for this fiscal year



Enterprises are facing a growing need to search both unstructured and structured business data, by their semantics or meaning

Find products that match a photo or a text description

AI
Vector Search

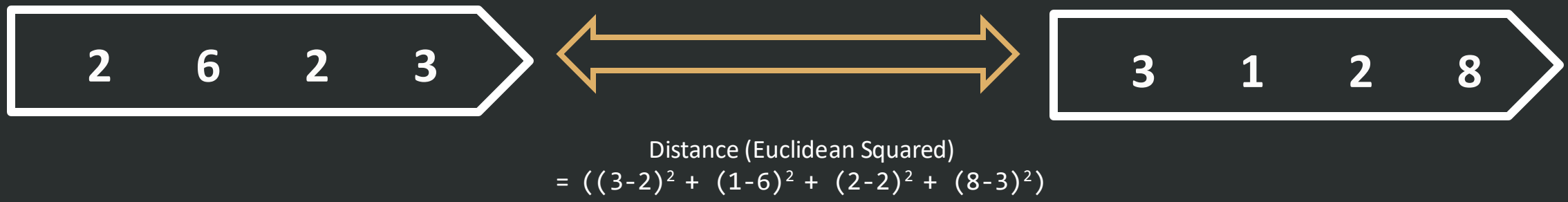
A new breakthrough technology to search documents, images, and other structured and unstructured data based on their semantic content, rather than their words or pixels

AI Vector Search works by representing the semantic content of a document, image, video, or even relational data as a sequence of numbers, called a vector

Developers create a vector for an object by just passing the object to a built-in vectorization function

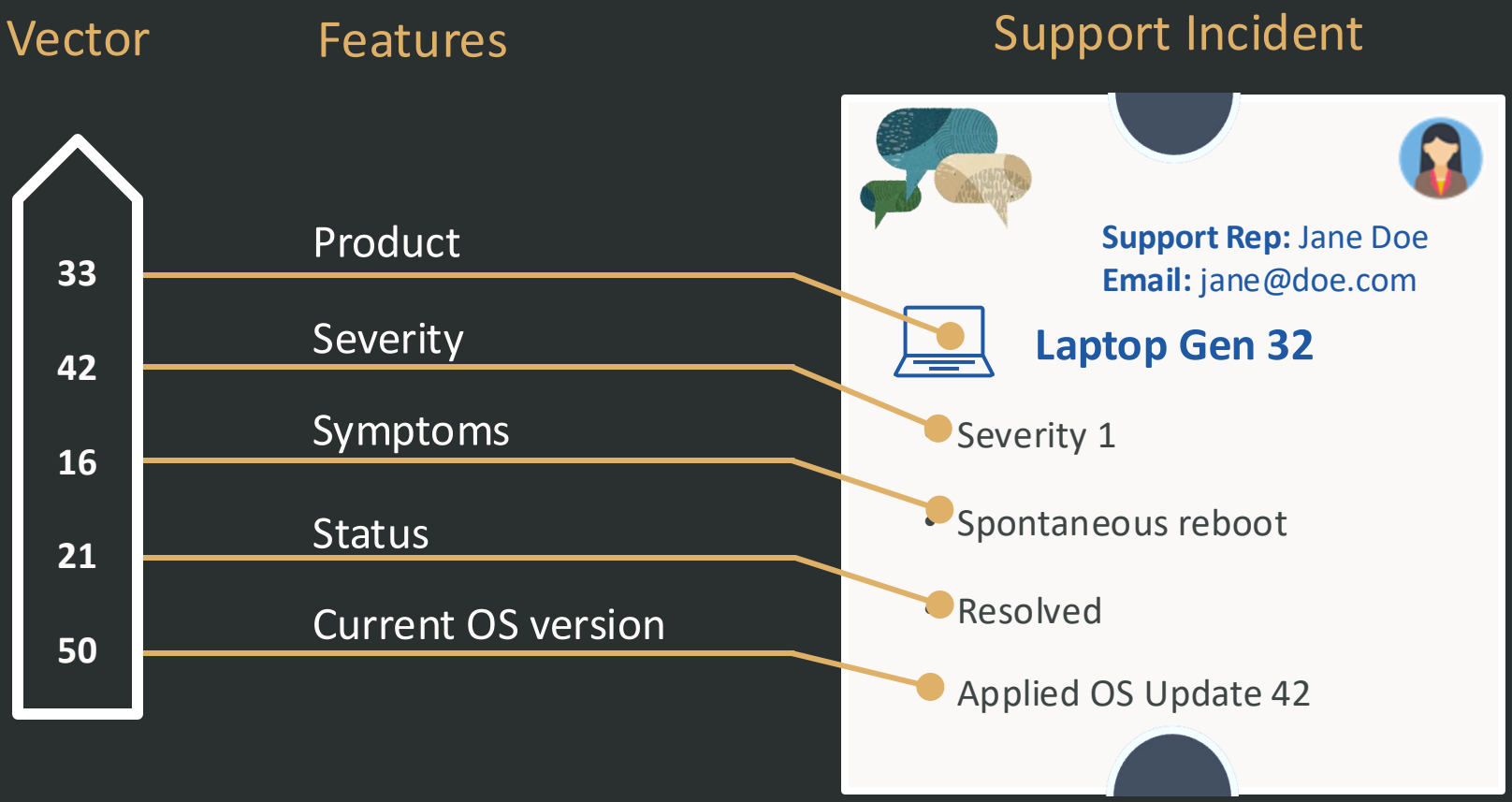
Oracle AI Vector Search natively stores vectors and compares vectors to find objects with similar semantic content

The main operation on vectors is the Mathematical Distance between them



There are many mathematical distance formulas (e.g., Euclidean, Cosine, Hamming)

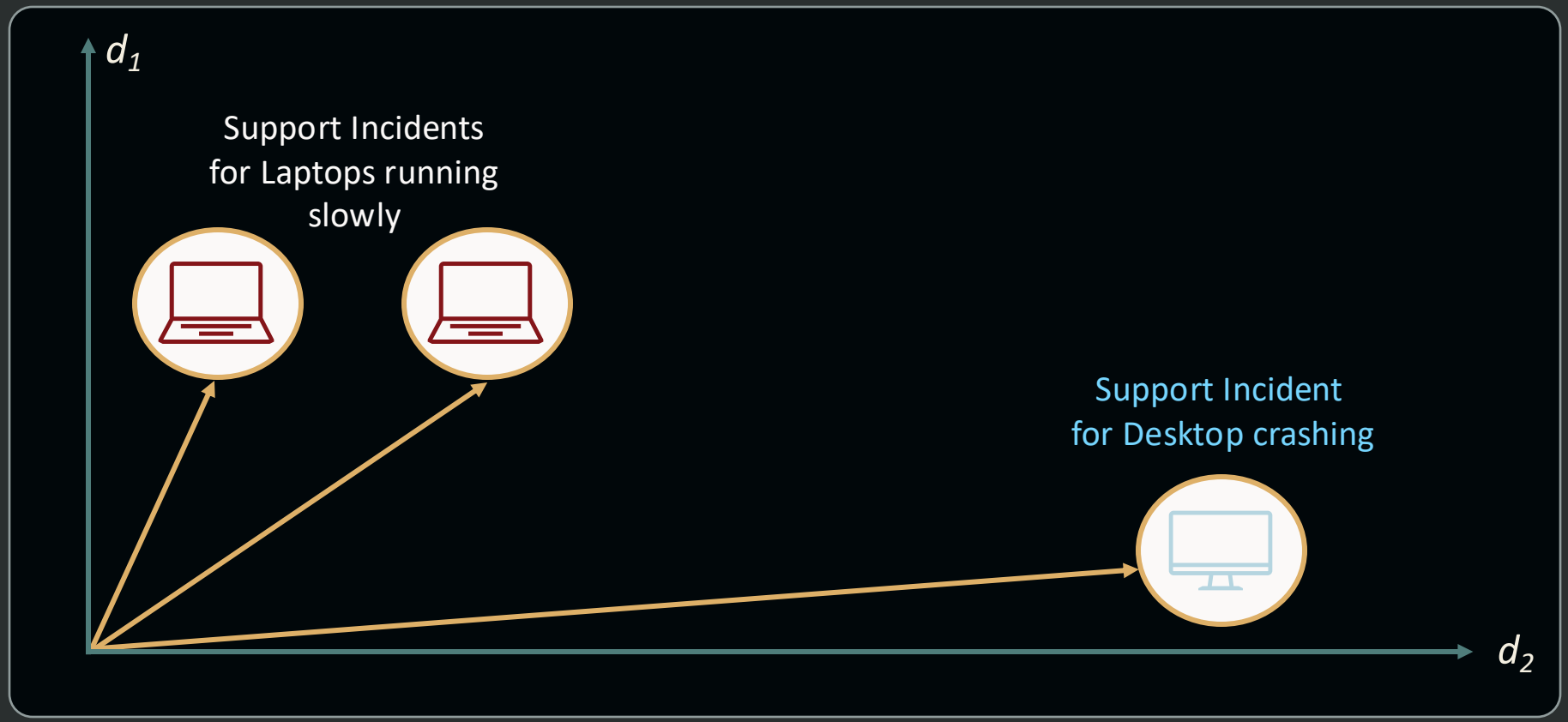
An Example Business Scenario:
The vector for a support incident could be ...



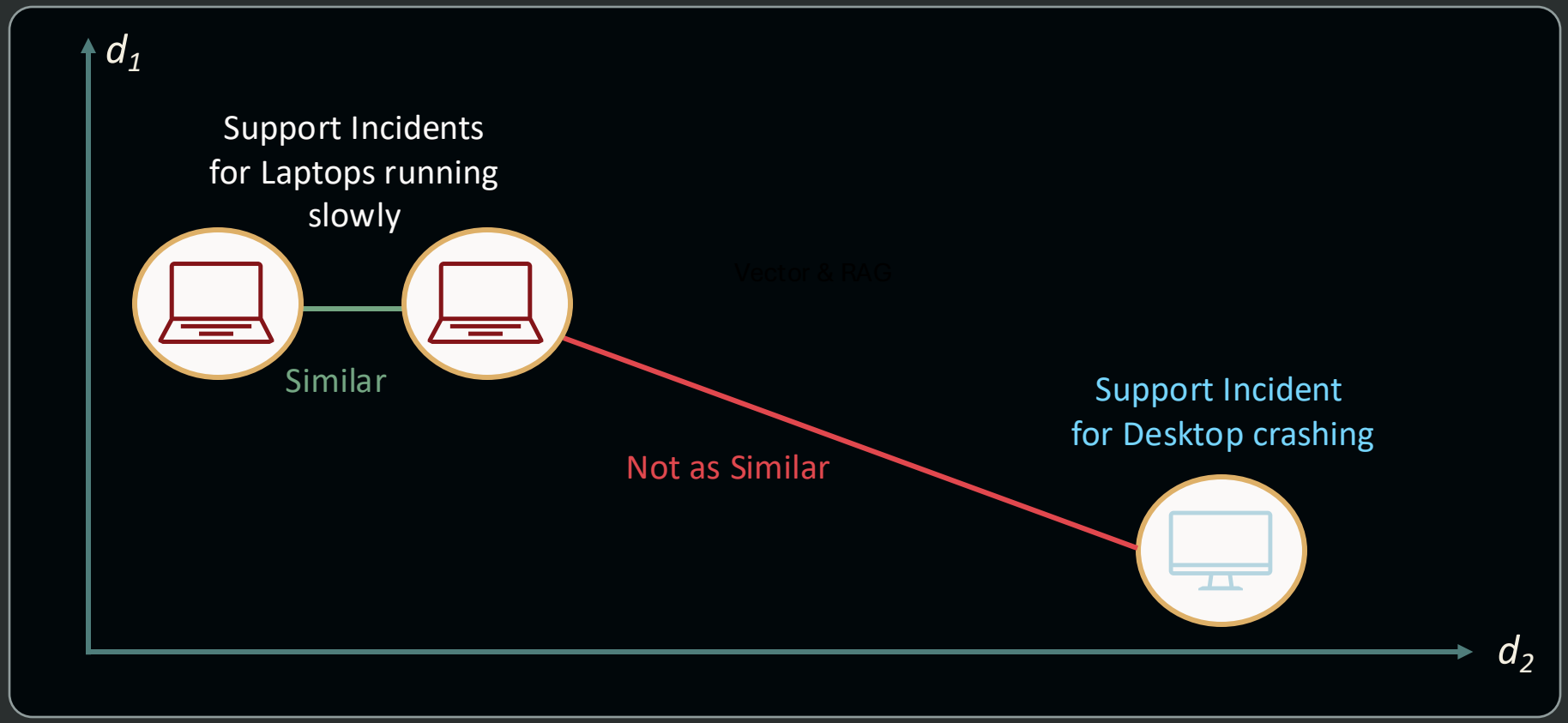
Each dimension (number), represents a different feature of the support incident

Note: Features are often chosen by ML algorithms and are not as simple as shown here

Support incident vectors when collapsed into 2 dimensions instead of hundreds could look like this



Similarity Property: Support Incidents that are more similar also produce vectors that are closer together



The more similar two entities are, the shorter the distance between their vectors

The Similarity Property powers AI Vector Search

User Query

Query Vector

ML Embedding Model

7 1 4 9

AI Vector Search

Top K matches

compare

...

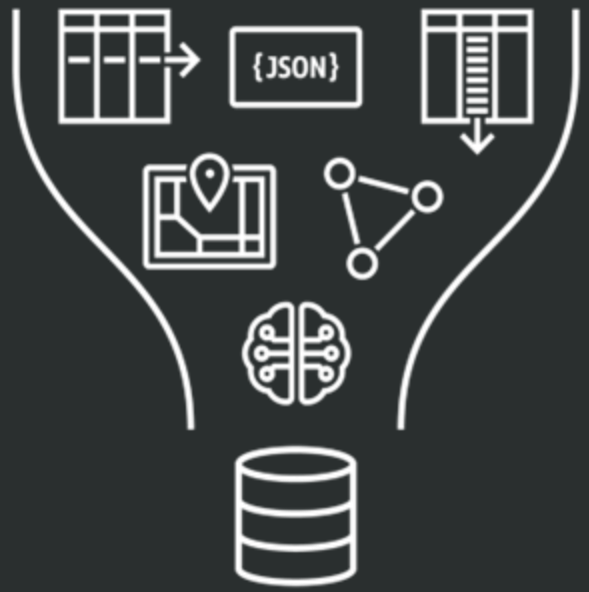
Data Corpus
(encoded with the
same embedding model)

Now that we know what vectors are, let's talk about how they are used in the enterprise

50 21 16 42 33

Vector Search works best when combined
with relational search to solve business problems

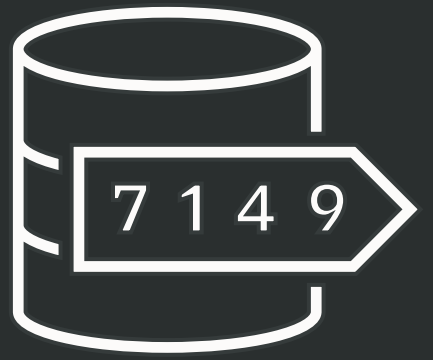
Searches on a combination of business and semantic data is more effective if both types of data are stored together



Business Database

Business Data

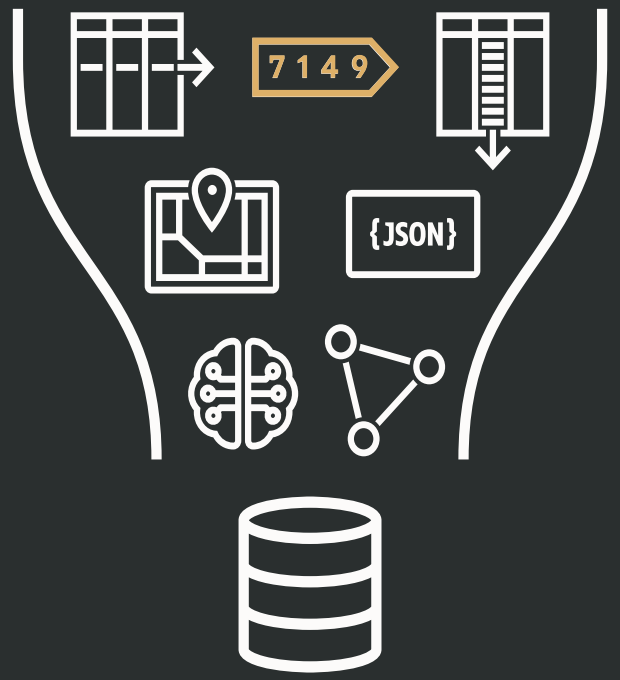
One solution is to continuously send your business data to a vector database



Vector Database

However, the business data that is relevant to a question varies widely
Plus, dedicated vector databases are not good at searching or securing business data

Better Together: Business Data and Business Vectors



Oracle's Converged Data Architecture

Uniquely combines sophisticated business data search with vector similarity search using simple SQL

There is no need to move and synchronize data, manage multiple products, etc.

Every mission-critical feature of Oracle Database works transparently with AI Vector Search

Oracle Database's robust security controls ensure compliance with corporate security standards

Allowing AI Vectors to be used immediately in enterprise apps of any scale or criticality

Enterprise Similarity Search Use-Cases

Find Similar Support Tickets

Biometric pattern recognition

Find Similar Products

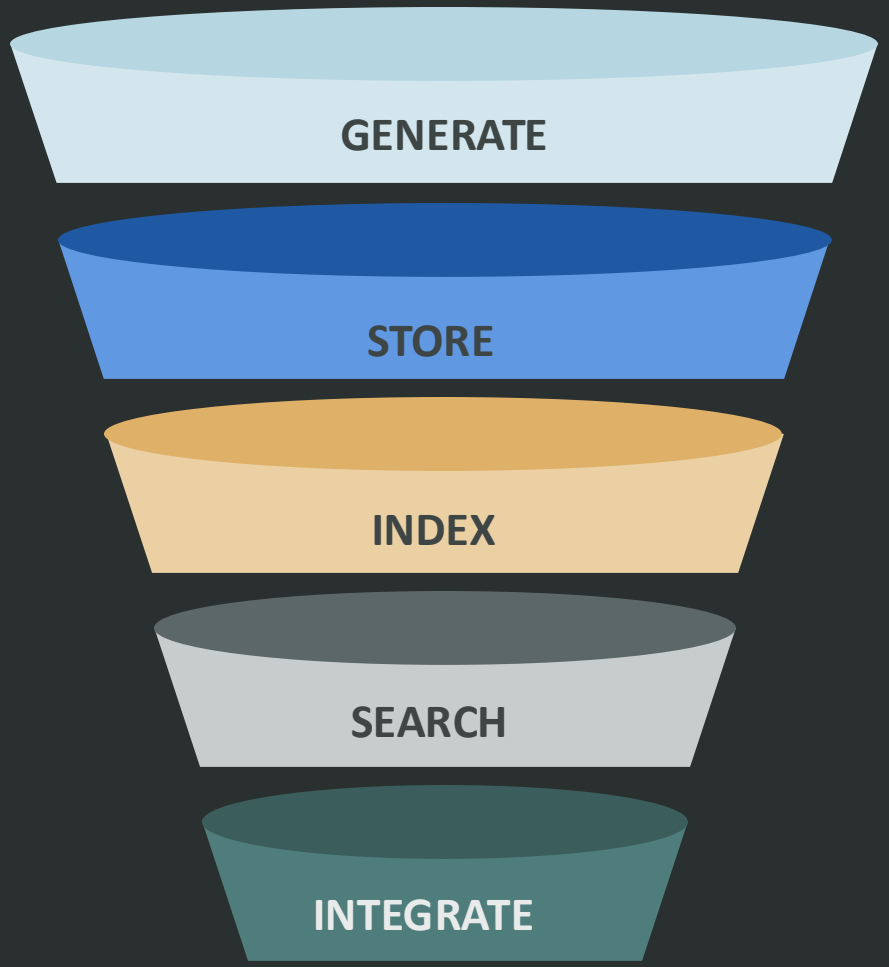
Detect manufacturing anomalies

Product Recommendation

Natural language catalog search

The [A, B, C] of Vectors

AI Vector Search Highlights



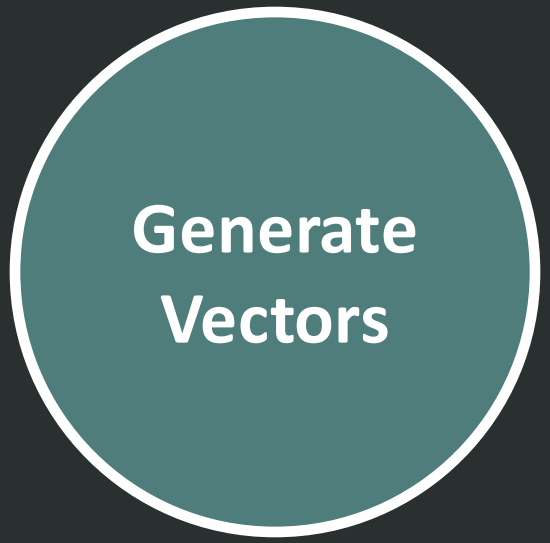
Generate vector embeddings from unstructured data

Store vectors in table columns using new VECTOR type

Build approximate vector indexes on VECTOR columns

Perform AI Vector Search on VECTOR columns using SQL

Integrate with Mission-Critical Enterprise Capabilities



Vector Embedding Generation | Your Way

AI Vector Search offers 4 alternatives for vector embedding generation

1

Use Pre-created embeddings

2

Use an external embedding cloud-service

3

Use an external embedding library

4

Use a database resident embedding model

Vector Embedding Generation | Your Way

Generate vector embeddings inside the database

4 Use a database resident embedding model

Generate embeddings using the `VECTOR_EMBEDDING()` SQL function using an imported **ONNX** embedding model, so that no data leaves the database

```
-- import onnx embedding model
DBMS_VECTOR.load_onnx_model(
  directory => <database directory>
  file_name => 'my_embed_model.onnx'
  model_name => 'embed-model',
  metadata =>  <source>
);

-- Generate vectors from support incident descriptions
SELECT
VECTOR_EMBEDDING(embed-model USING incident_text)
FROM Support_Incidents;
```

Supports Text, Image, Multi-Lingual models

VECTOR Datatype to Store and Process Vectors

```
CREATE TABLE Support_Incidents(
      id              NUMBER,
      incident_text   CLOB,
      incident_vector VECTOR(128, FLOAT32, DENSE));
```

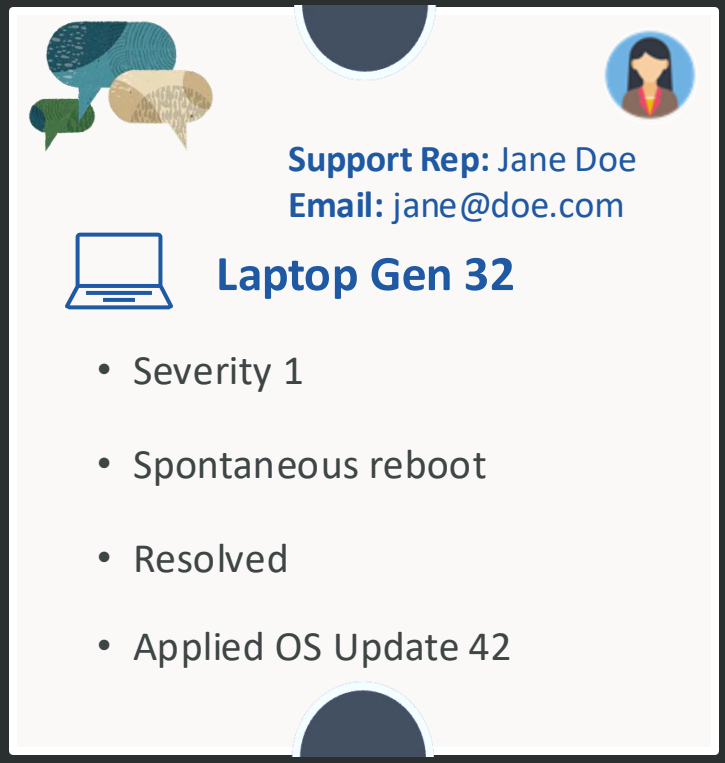
Dimension Count (optional)

Dimension Format (optional)

FLOAT32, FLOAT64, INT8, BINARY

Storage Format (optional)

SPARSE, DENSE




```
INSERT INTO Support_Incidents(1, 'Problem...',
                              TO_VECTOR('[1.1, 2.2, ..]'));

SELECT FROM_VECTOR(incident_vector) FROM Support_Incidents;
```

Native VECTOR support available in all major client drivers (e.g., Python, node.js, JDBC etc.)

An exhaustive search for top-K matches will be 100% accurate but slow as data volumes grows

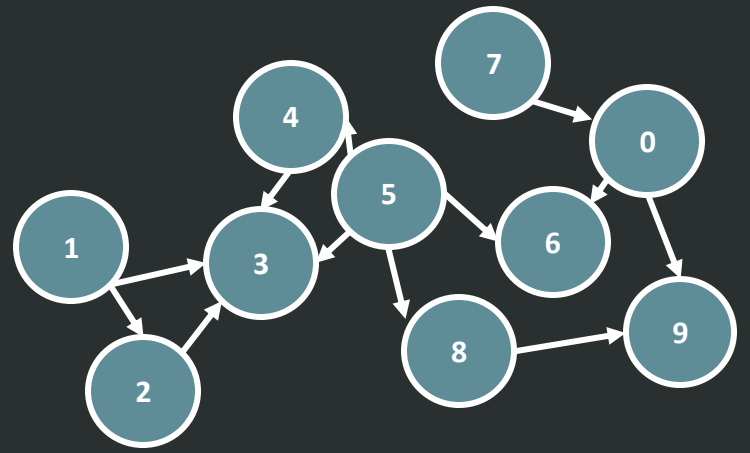
New vector indexes trade-off some search accuracy for up-to 100x speed up

Vector Indexes | Neighbor Graph Vector Index

Graph-based index where vertices represent vectors and edges between vertices represent *similarity*

In-Memory only index - highly efficient for both accuracy and speed

```
CREATE VECTOR INDEX incident_idx
ON SUPPORT_INCIDENTS(incident_vector)
ORGANIZATION INMEMORY NEIGHBOR GRAPH;
```



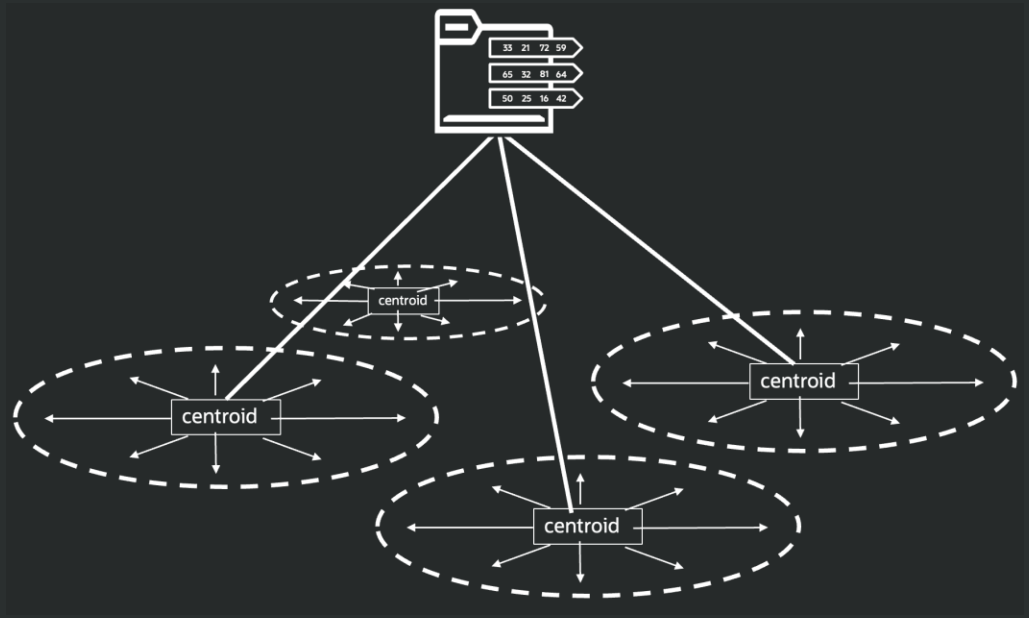
Graph Vector Index (e.g., **HNSW** Index)

Vector Indexes | Neighbor Partition Vector Index

Partition-based index with vectors clustered into table partitions based on *similarity*

Efficient scale-out index for unlimited data size

```
CREATE VECTOR INDEX incident_idx
ON SUPPORT_INCIDENTS(incident_vector)
ORGANIZATION NEIGHBOR PARTITIONS;
```



Partition Vector Index (e.g., **IVF_FLAT** index)

Search Vectors

Vector Search SQL | Distance Function

The main operation on vectors is to find how similar they are

```
VECTOR_DISTANCE(VECTOR1, VECTOR2, <distance metric>)
```

Different embedding models can use different distance metrics like Euclidean, cosine similarity, dot product, etc.

All embedding models must obey the same similarity property

e.g., `VECTOR_DISTANCE(<Tiger Vec>, <Lion Vec>) < VECTOR_DISTANCE(<Tiger Vec>, <Apple Vec>)`

Vector Search SQL | Specifying Similarity Search

Support Incident Search Example

Find the top 10 matching support incidents


```
SELECT ...
FROM   Support_Incidents
ORDER BY VECTOR_DISTANCE(incident_vector, :search_vector)
FETCH FIRST 10 ROWS ONLY
TARGET ACCURACY [<percent> | <Low-level parameters>]
```

Accuracy specification is optional

Vector Search SQL | Similarity Search w/ HNSW Vector Index

```
SELECT id
FROM   Support_Incidents
ORDER BY VECTOR_DISTANCE(incident_vector, :search_vector)
FETCH FIRST 10 ROWS ONLY;
```

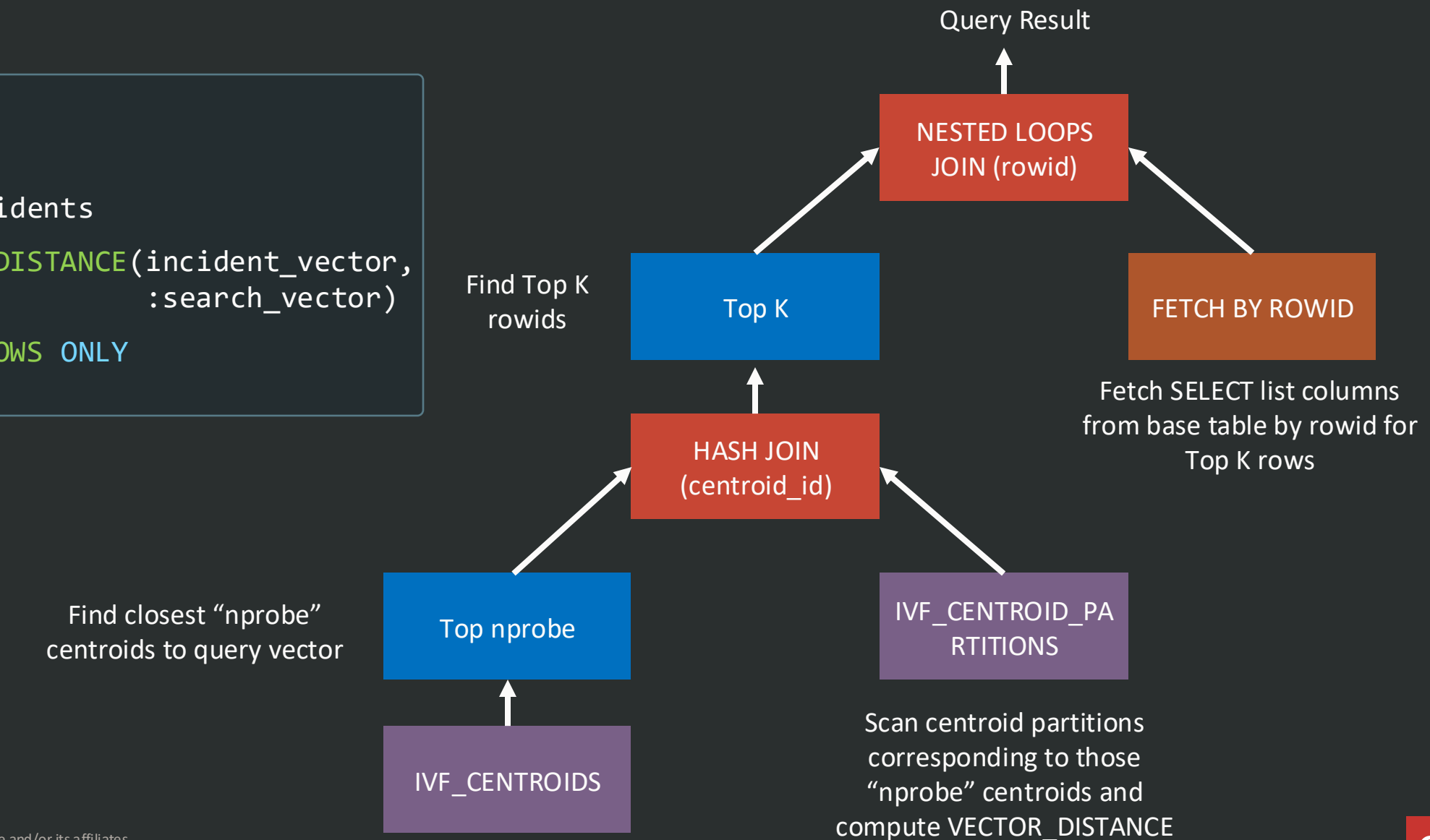
```
-------------------------------------------------------------------------------------
| Id  | Operation                      | Name              | ...
-------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT               |                   |
|*  1 |  COUNT STOPKEY                 |                   |
|   2 |   VIEW                         |                   |
|*  3 |    SORT ORDER BY STOPKEY       |                   |
|   4 |     TABLE ACCESS BY INDEX ROWID| SUPPORT_INCIDENTS |
|   5 |      VECTOR INDEX HNSW SCAN    | INCIDENT_IDX      |
-------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------
1 - filter (ROWNUM <= 10)
3 - filter (ROWNUM <= 10)
```

Obtain Top K vector matches and return rowids

Vector Search SQL | Similarity Search w/ IVF Vector Index

```
SELECT ...
FROM Support_Incidents
ORDER BY VECTOR_DISTANCE(incident_vector,
                         :search_vector)
FETCH FIRST 10 ROWS ONLY
```



Vector Search SQL | Combining Similarity Search with Relational Search

Value-based Attribute Filters can be combined seamlessly with Vector Search in SQL

Optimizer picks the best access plan based on filter selectivity

Three possible techniques:

- PRE-FILTER: Filter by attributes, then rank by vector distance
- IN-FILTER: Filter by attributes and compute vector distance in one step
- POST-FILTER: Rank Top K by vector distance, then filter by attributes

Find the top 10 matching support incidents that were **filed within the last 7 days**


```
SELECT ...
FROM   Support_Incidents
WHERE  Incident_Date > SYSDATE - 7
ORDER BY VECTOR_DISTANCE(incident_vector, :search_vector)
FETCH FIRST 10 ROWS ONLY;
```

Vector Search SQL | Combining Similarity Search with Relational Search

IN-FILTER w/ HNSW Vector Index

```
-------------------------------------------------------------------------------
| Id  | Operation                          | Name                  | ...
-------------------------------------------------------------------------------
|   0 | SELECT STATEMENT                   |                       |
|*  1 |  COUNT STOPKEY                     |                       |
|   2 |   VIEW                             |                       |
|*  3 |    SORT ORDER BY STOPKEY           |                       |
|*  4 |     TABLE ACCESS BY INDEX ROWID    | SUPPORT_INCIDENTS     |
|   5 |      VECTOR INDEX HNSW SCAN IN-FILTER| INCIDENT_IDX        |
|   6 |       VIEW                         | VW_HIF_86A2           |
|*  7 |        TABLE ACCESS BY USER ROWID  | SUPPORT_INCIDENTS     |
-------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------
1 - filter (ROWNUM <= 10)
3 - filter (ROWNUM <= 10)
7 - filter ("INCIDENT_DATE" > SYSDATE - 7)
```

For each vector found during HNSW graph exploration, apply the relational filter via the ROWID

Vector Search SQL | Combining Similarity Search with Joins

Combines customer and product data, and AI search in a few lines of SQL

Essential capability as enterprise data is normalized

Any developer or DBA can learn to use it in ~10 minutes

Find the top 10 matching support incidents for a **Laptop** reported by customers in **Las Vegas**


```
SELECT ...
FROM   Support_Incidents SI
JOIN   Products  P ON SI.product_id = P.id
JOIN   Customers C ON SI.customer_id = C.id
WHERE  P.Type = 'Laptop'
AND    C.City = 'Las Vegas'
ORDER BY VECTOR_DISTANCE(SI.incident_vector, :search_vector)
FETCH FIRST 10 ROWS ONLY;
```

Integrate Vectors

Every mission-critical feature of Oracle Database works transparently with AI Vectors

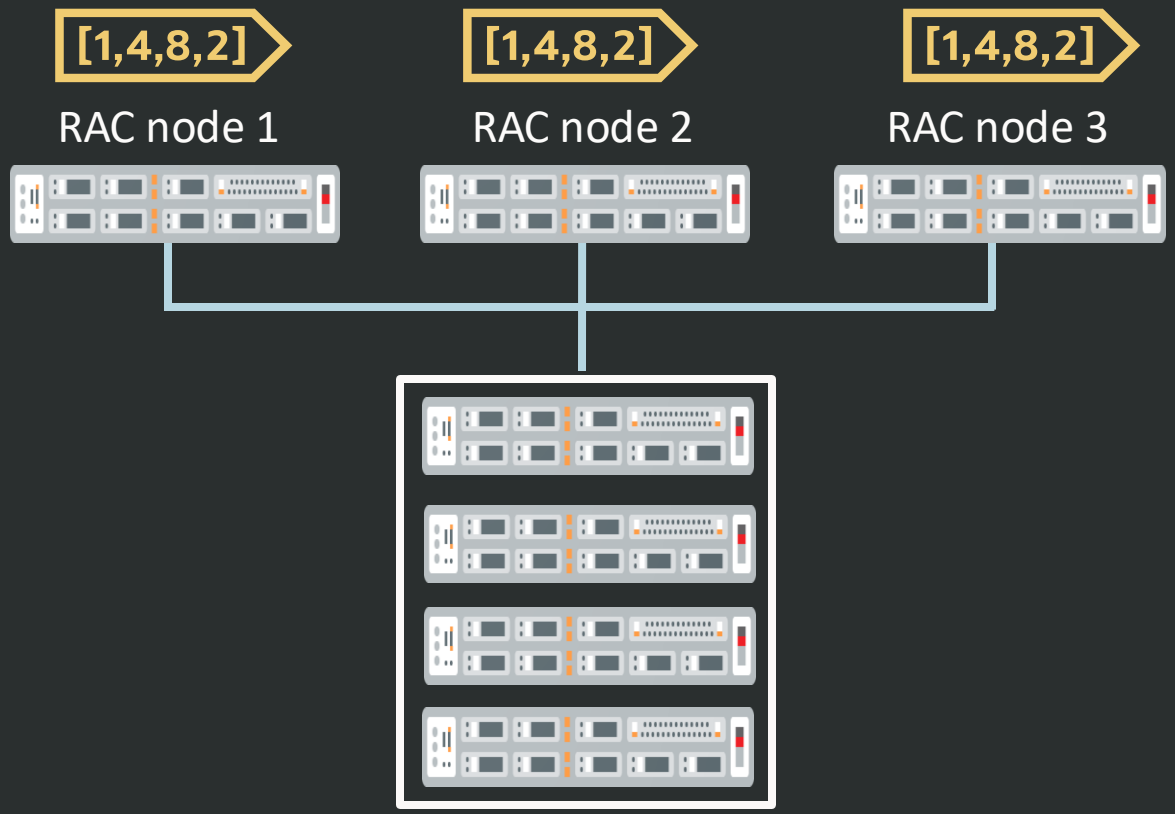
Allowing AI Vectors to be used immediately in enterprise apps of any scale or criticality

- Real-Application Cluster
- Parallel SQL
- Transactions
- Security
- Analytics
- Disaster Recovery

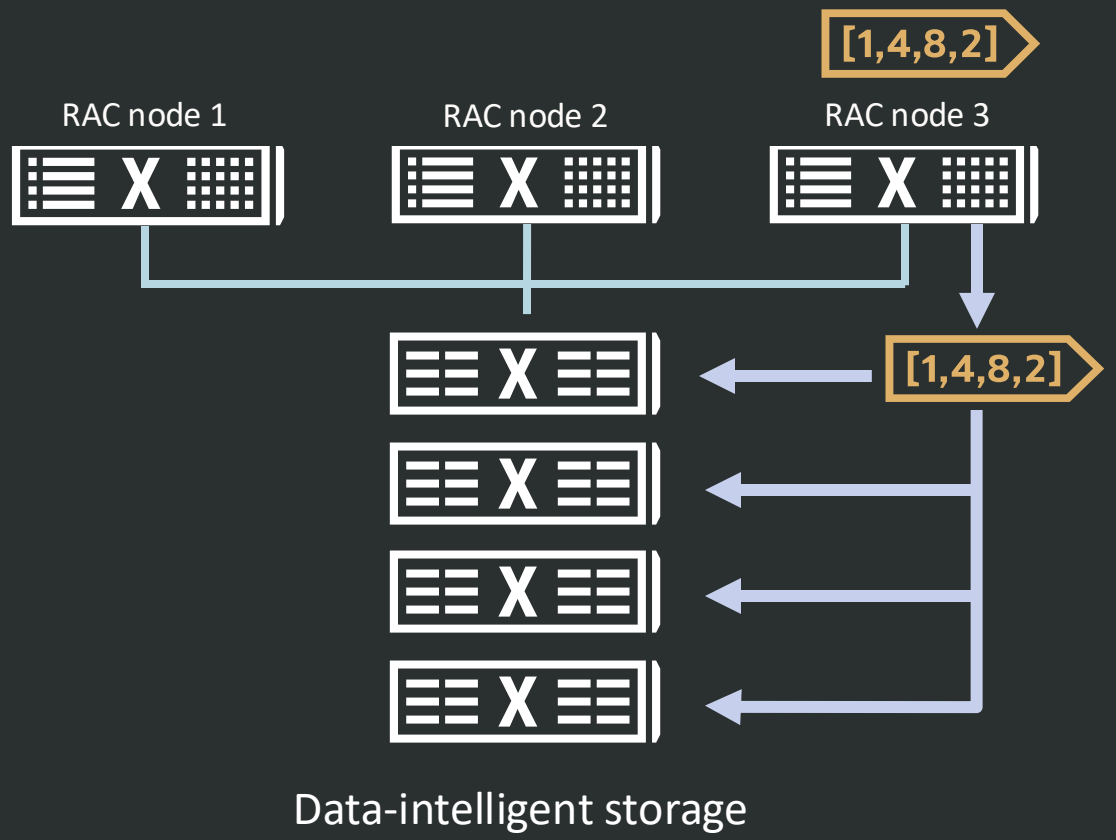
AI Vector Search | Scale-Out with Real Application Clusters

AI Vector search transparently scales vector processing across the compute nodes in a RAC cluster

With full data consistency



AI Vector Search | Scale-Out with offload to Exadata Intelligent Storage



Oracle's AI vector search can be transparently offloaded to intelligent Exadata storage for up to 30X faster AI Vector queries

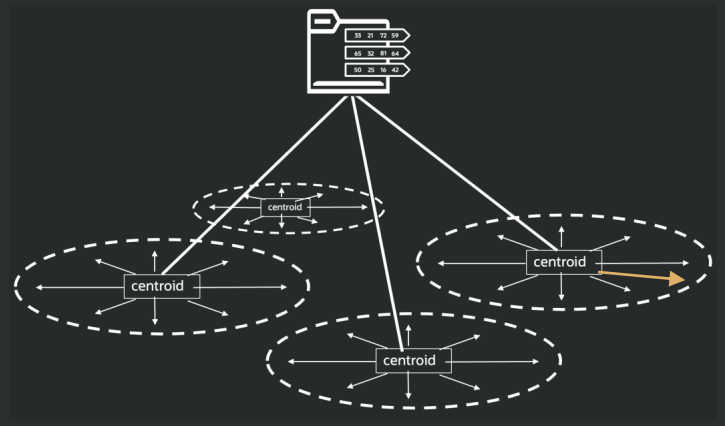
Vector search queries are automatically parallelized across the storage servers

Each storage server independently computes the top-K matches enabling faster Top-K processing

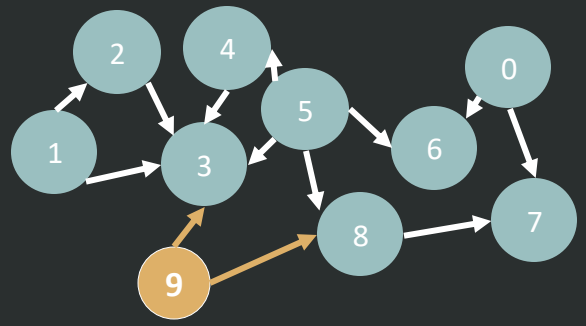
Supports extreme scale environments with thousands of concurrent AI vector searches

AI Vector Search | Transactionally Consistent Vector Indexes

Neighbor Partition Vector Index



Neighbor Graph Vector Index



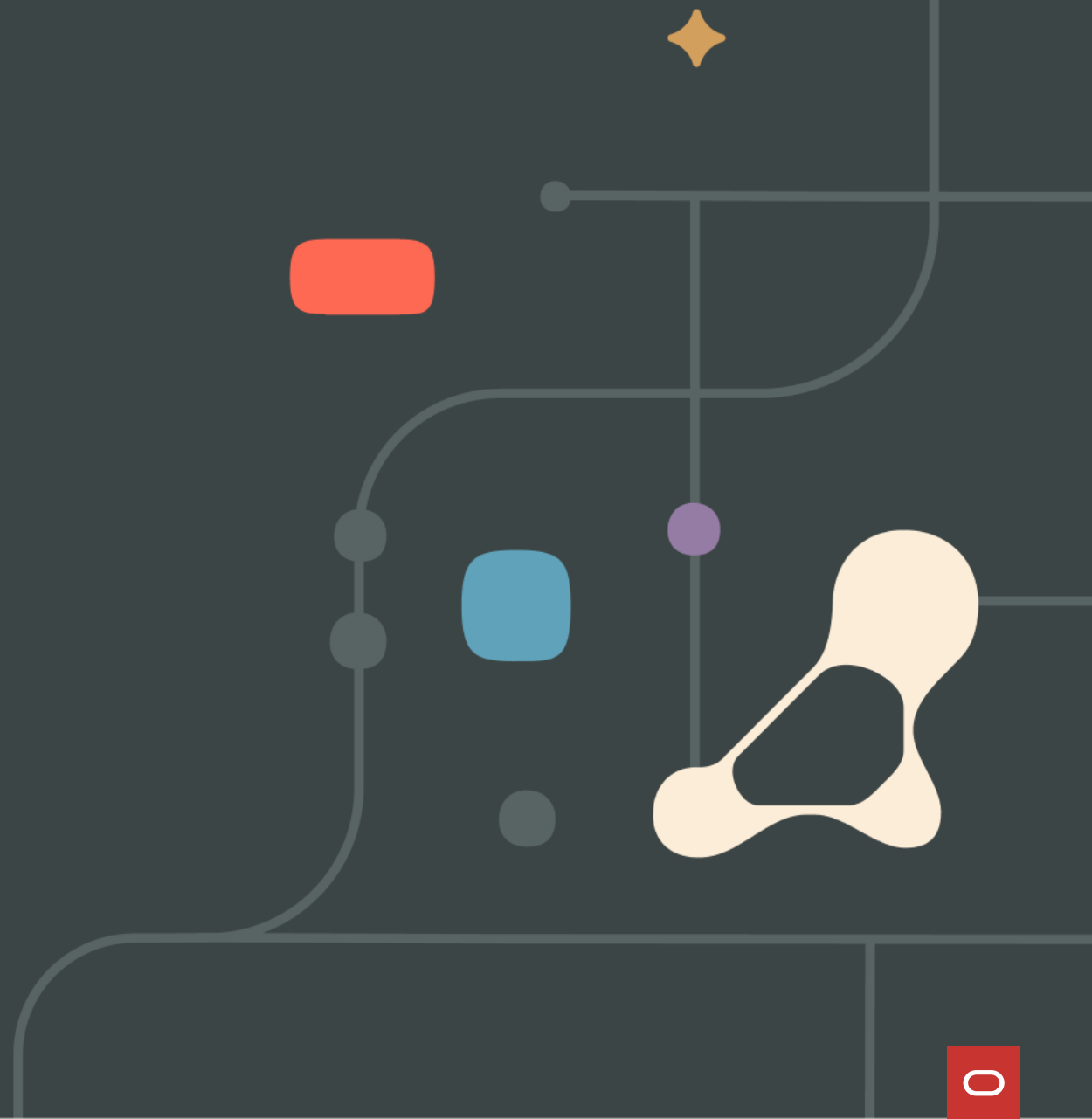
Oracle's AI Vector Search Indexes maintain transactional consistency with DML activity

INCIDENT_ID	INCIDENT_TEXT	INCIDENT_VECTOR
…	…	…
…	…	…
…	…	…

33
42
16
21
50

A new vector is inserted

Vectors and RAG

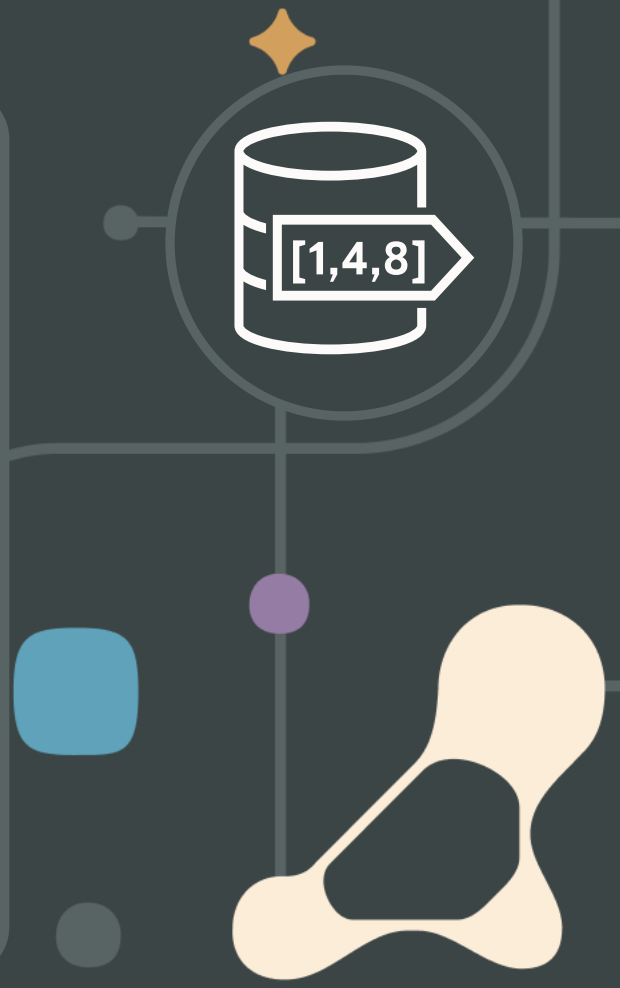


Adding Generative AI to AI Vector and business data search enables a new era of data and app dev productivity

Oracle 23ai improves Generative AI by **augmenting LLM prompts** with **private database content** that is found using any combination of data and AI Vector Search

Enables LLMs to use business data to produce better and more contextually relevant answers to user questions while keeping business data secure

Called: Retrieval Augmented Generation (RAG)



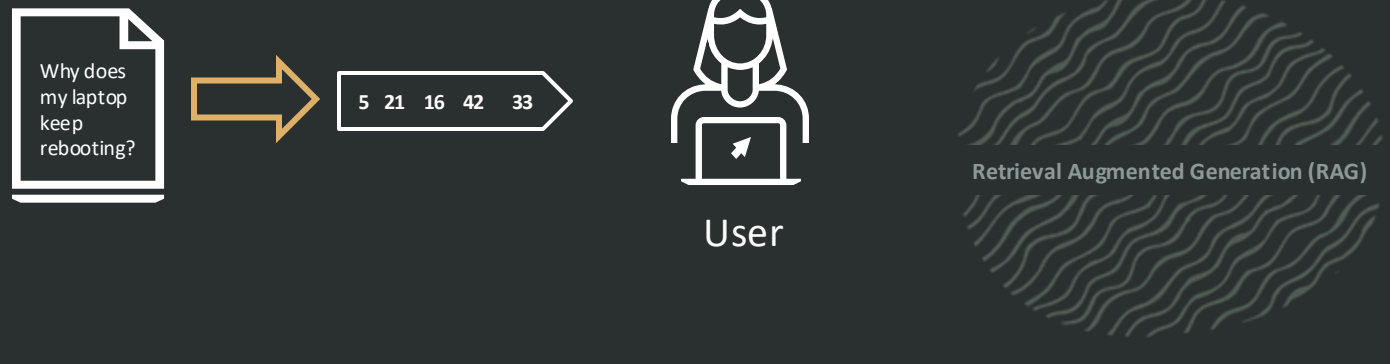
AI Vector Search in Oracle Database Powers Complete Gen AI Pipeline

Retrieval Augmented-Generation (RAG) with your enterprise data

Vectorize Question

An end-user's human language question is encoded as a vector

1



AI Vector Search in Oracle Database Powers Complete Gen AI Pipeline

Retrieval Augmented-Generation (RAG) with your enterprise data

Vectorize Question

An end-user's human language question is encoded as a vector

1

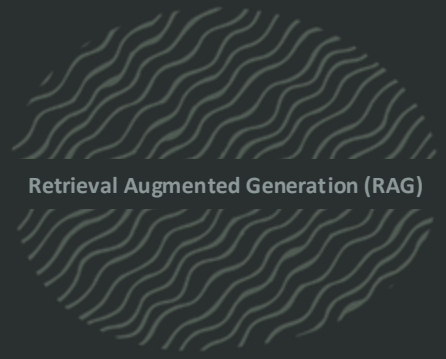
AI Vector

2

Find Related Data

AI Vector Search finds private database data that matches the user's vector including product info and other support tickets for the same laptop

User



Product Info

Support Tickets

GenAI

AI Vector Search in Oracle Database Powers Complete Gen AI Pipeline

Retrieval Augmented-Generation (RAG) with your enterprise data

Vectorize Question

An end-user's human language question is encoded as a vector

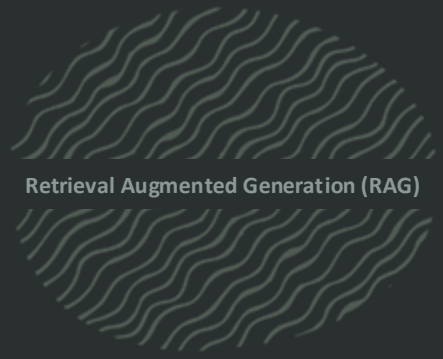
1

2

Find Related Data

AI Vector Search finds private database data that matches the user's vector including product info and other support tickets for the same laptop


User




Product Info


Support Tickets


GenAI

3

Augment Prompt

The user's question is augmented with this
private data

AI Vector Search in Oracle Database Powers Complete Gen AI Pipeline

Retrieval Augmented-Generation (RAG) with your enterprise data

Vectorize Question

An end-user's human language question is encoded as a vector

1

[1,4,8]

AI Vector

2

Find Related Data

AI Vector Search finds private database data that matches the user's vector including product info and other support tickets for the same laptop

The issue is with the firmware controlling the fan. Apply OS update 42 while plugged in, in a cool air-conditioned environment to prevent overheating

User

Retrieval Augmented Generation (RAG)

Product Info

Support Tickets

Ask LLM

The combination is sent to an LLM to provide an informed answer to the question

4

GenAI

3

Augment Prompt

The user's question is augmented with this private data

The entire Retrieval Augmented Generation (RAG) pipeline can be executed directly from SQL

1. Generate the question vector using SQL

2. Perform Vector Search using SQL

3. Compose and run the LLM prompt using SQL

4. Retrieve LLM response using SQL

```
WITH TOP10 AS (SELECT incident_text FROM SUPPORT_INCIDENTS
               ORDER BY VECTOR_DISTANCE(incident_vector,
                                        SELECT VECTOR_EMBEDDING(embedding_model
                                               USING :question_text AS data))
               FETCH FIRST 10 ROWS ONLY),

 LLM_PROMPT AS (SELECT ('Answer this question using the following context,\
                        QUESTION: ' ||:question_text || ' ,CONTEXT: '
                        || LISTAGG(incident_text, CHR(10))) AS prompt_text
                FROM TOP10)

SELECT DBMS_VECTOR_CHAIN.UTL_TO_GENERATE_TEXT(prompt_text, json(:LLM_params))
       AS Answer
FROM   LLM_PROMPT;
```

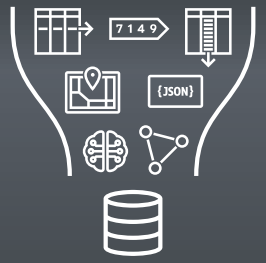
LLMs, Data, and SQL Engineered to Work Together

Takeaway

Oracle AI Vector Search

Architected together: unified business data and vector search

Enterprise-grade similarity search



Seamlessly combines vectors & biz data

Secure Agentic Retrieval Augmented Generation

Mission-Critical Reliability & Scalability

Oracle AI Vector Search supports all the leading AI models and frameworks


OCI Generative AI


OpenAI


Google


Grok


Meta


Anthropic


Cohere


Hugging Face


Microsoft


Amazon Bedrock


Mistral AI


Jina AI


CrewAI


Agno


LangChain


ONNX


LangGraph


Pytorch


Ollama

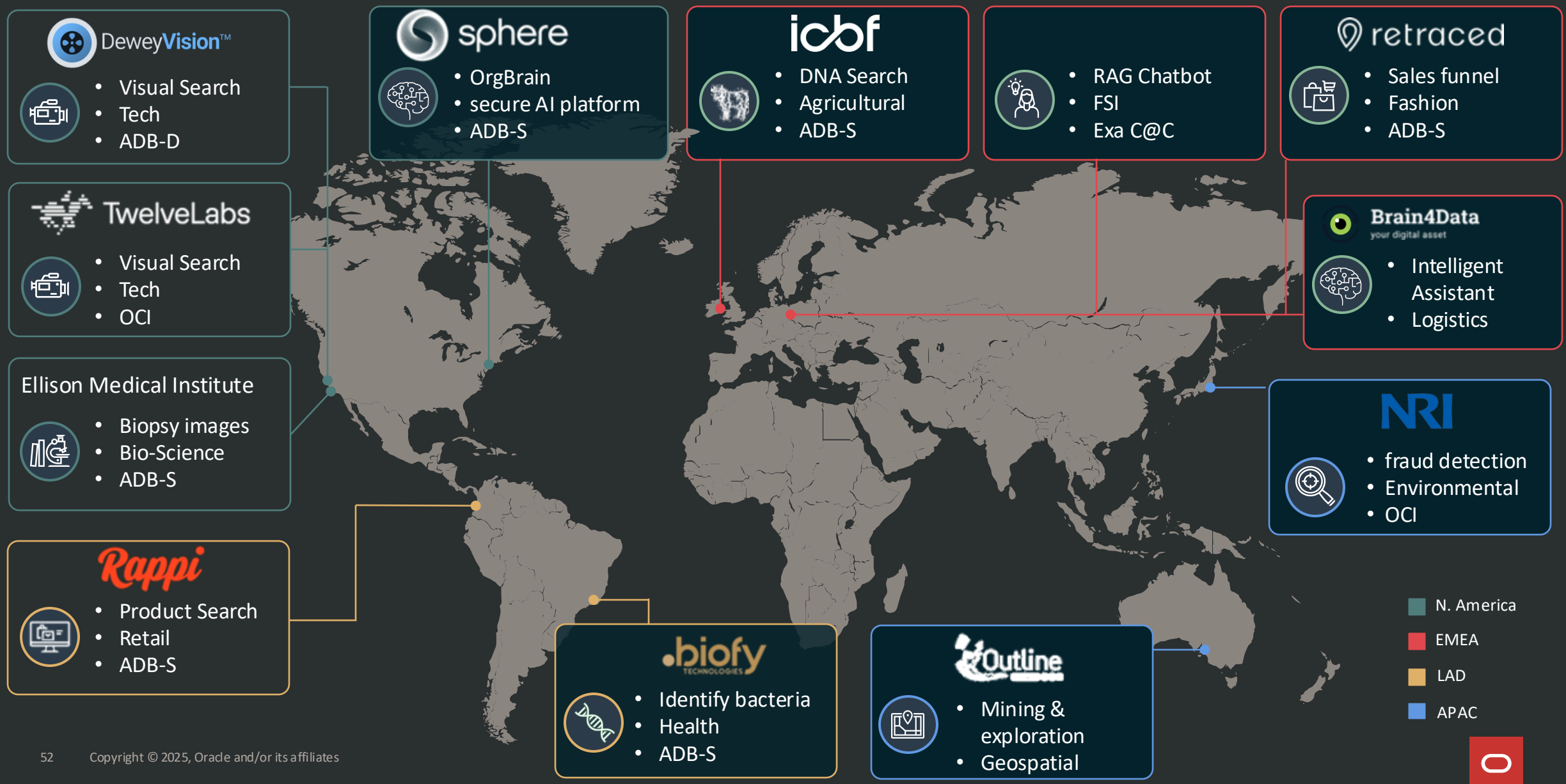

MCP Server

A2A

A2A Protocol

Customers can call them via APIs, or deploy them as private instances for added security

AI Vector Search is used across the globe in a variety of industries




ORACLE