ORACLE

Acelere o desenvolvimento de recursos de Generative AI com Chat em Linguagem Natural e LLMs integrados sem o uso de GPU!

Narciso Oliveira Junior
narciso.junior@oracle.com
Oracle HeatWave LAD

HeatWave

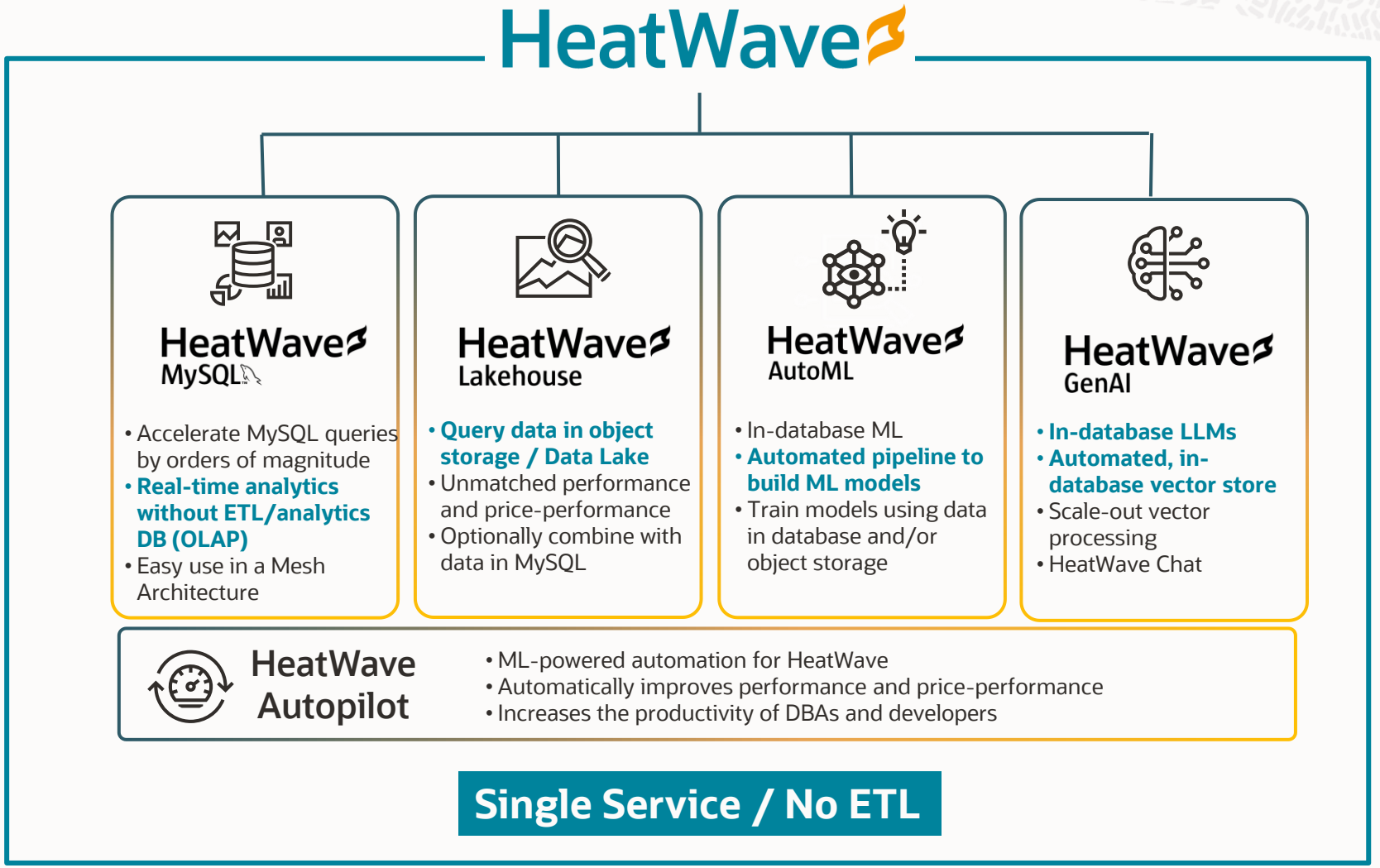
Narciso Oliveira Junior

- HeatWave Cloud Evangelist at Oracle
- 20 Years of experience in IT & Telecom
- Experience with SW development, Solution Integration and Sales
- MBA in Business Management
- Data & AI specialist

narciso.junior@oracle.com
/narcisooliveira

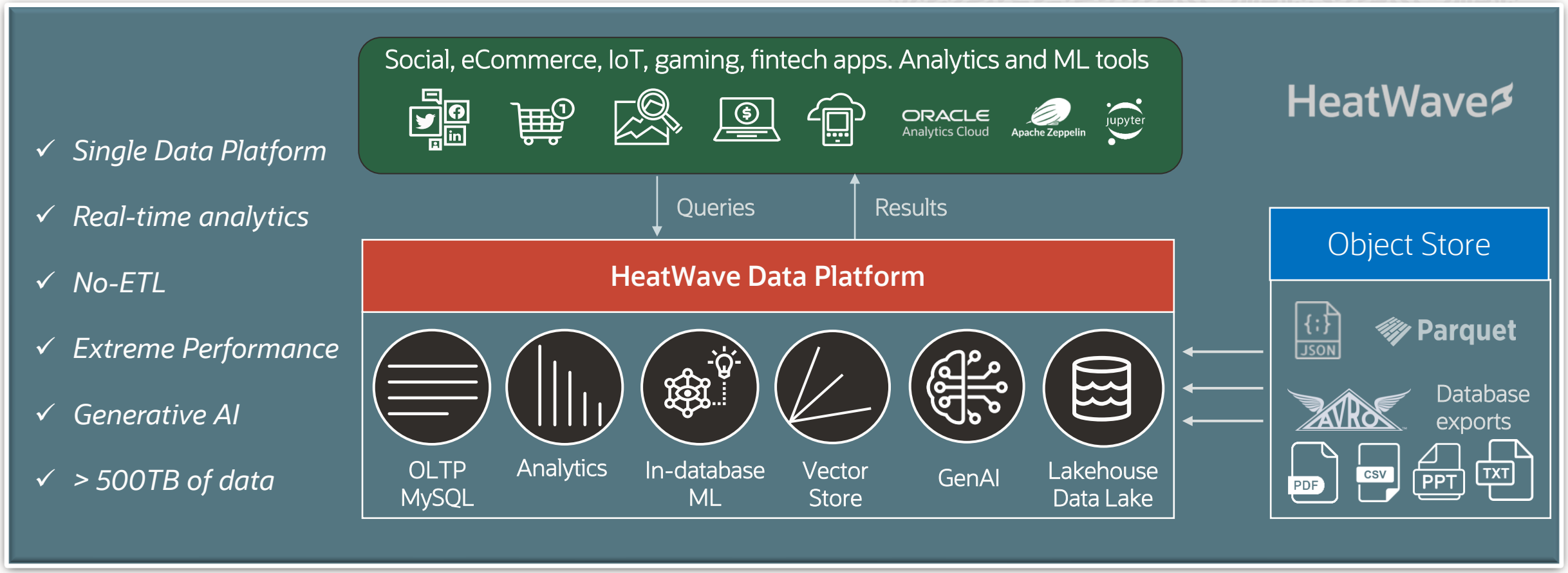
Addressing your challenges with HeatWave

Solutions for your different workloads

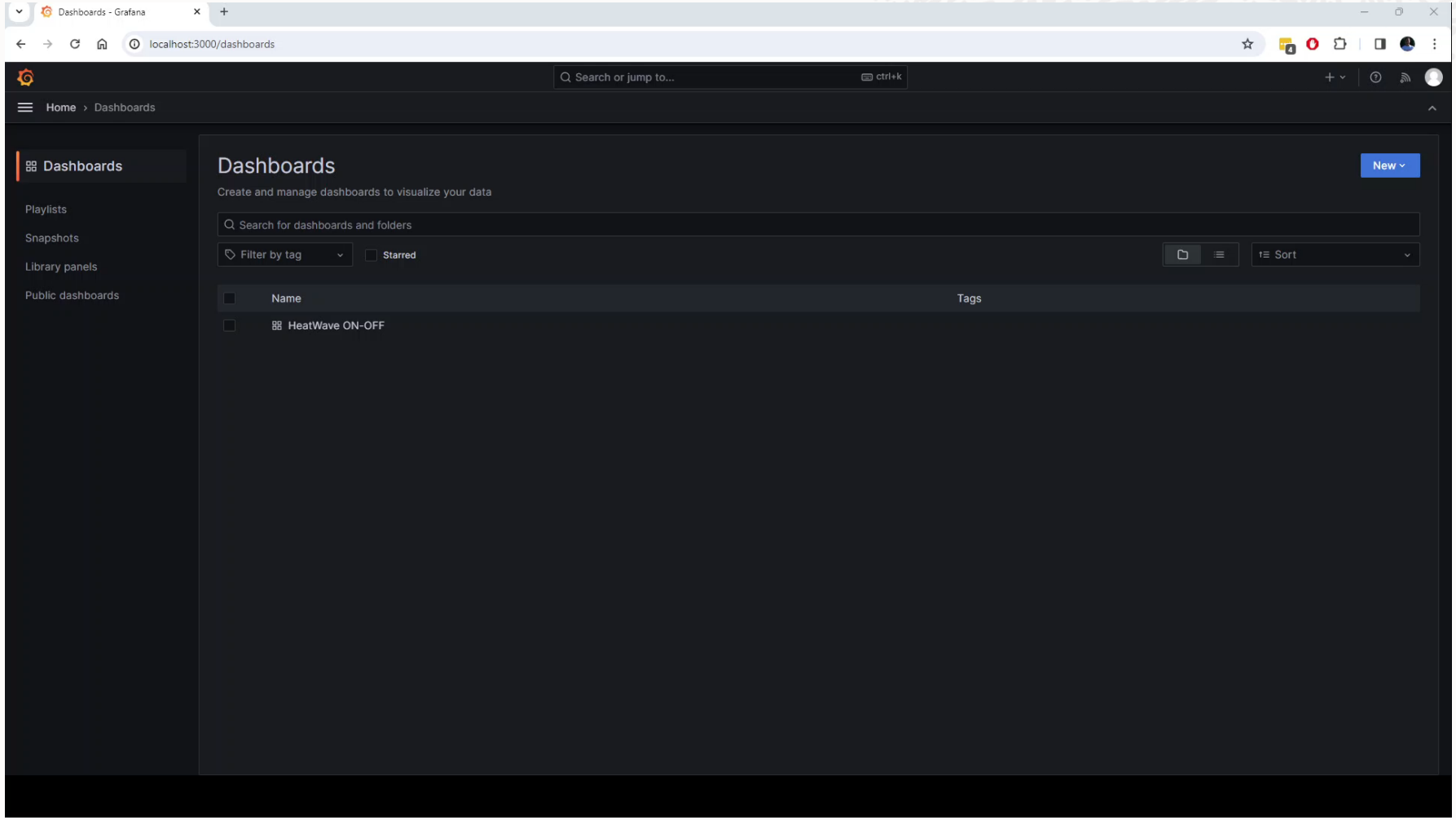


HeatWave Data Platform overview

Transactions, real-time analytics, machine learning and GenAI across data warehouse and data lake in one service



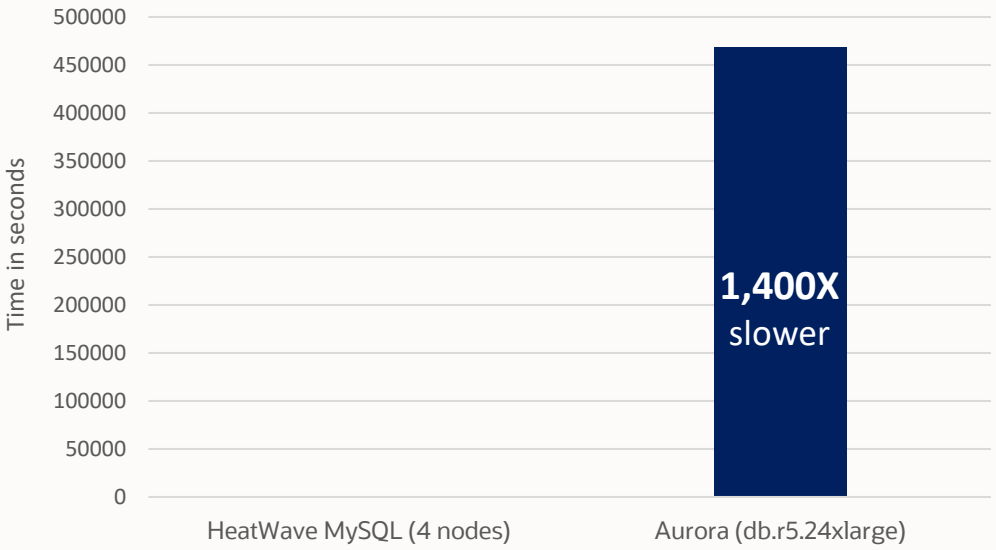
HeatWave MySQL in action



HeatWave MySQL vs. Amazon Aurora and RDS for MySQL

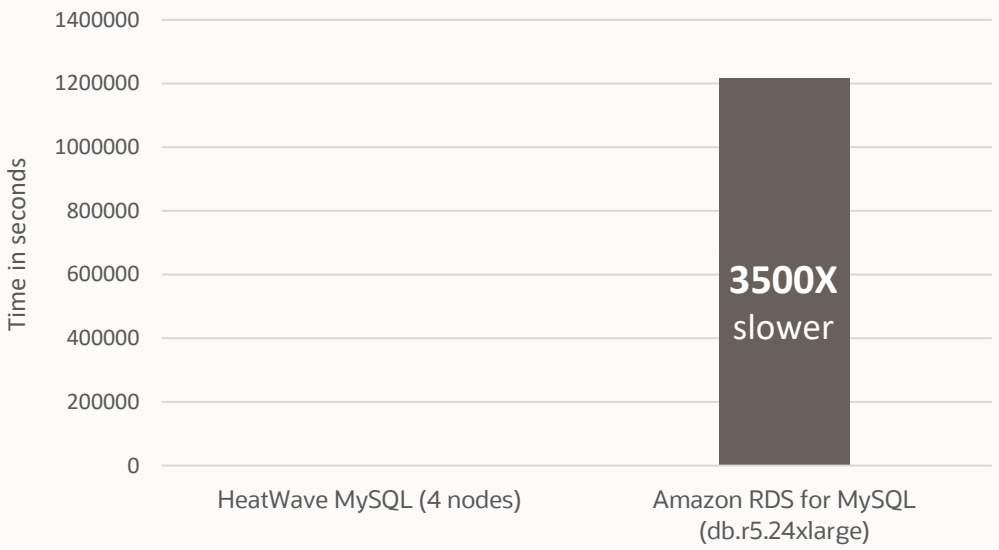
Complex and analytics queries

Amazon Aurora



2,200X worse price-performance

Amazon RDS



4,600X worse price-performance

*Benchmark queries are derived from the TPC-H benchmarks, but results are not comparable to published TPC-H benchmark results since these do not comply with the TPC-H specifications.

Best performance and price-performance for data warehouse

TPC-H 10TB

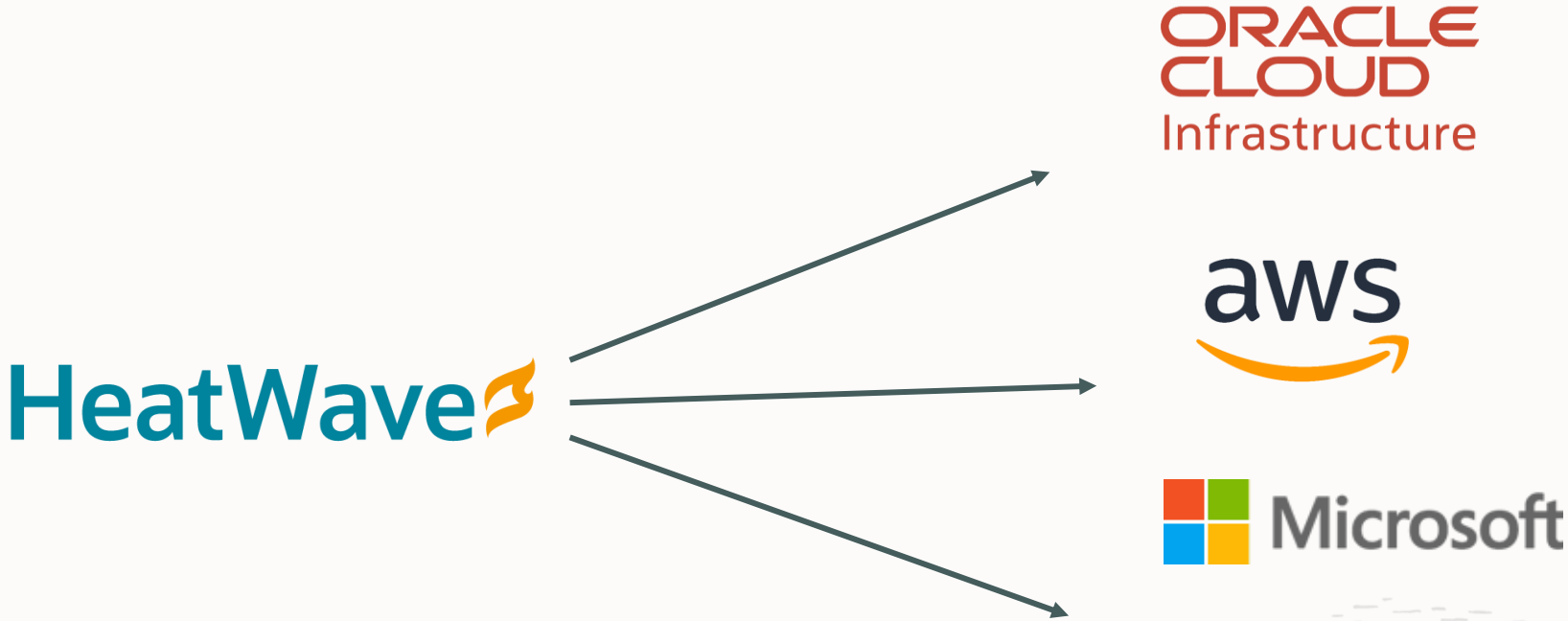
Faster time to insights = faster business response to market trends

4X	4X	9X	11X
faster than Redshift	faster than Snowflake	faster than BigQuery	faster than Databricks
10X better price-performance	15X better price-performance	20X better price-performance	37X better price-performance
10X ra3.4xlarge	X-Large Cluster	800 slots	Large Cluster

According to 10 TB TPC-H benchmarks as of May 23, 2023. Redshift, Snowflake, Databricks and BigQuery numbers for 10TB TPC-H numbers are provided by a third party.
Benchmark queries are derived from the TPC-H benchmarks, but results are not comparable to published TPC-H benchmark results since these do not comply with the TPC-H specifications.

HeatWave is multi-cloud

Deploy HeatWave MySQL-powered apps on multiple clouds



HeatWave GenAI

Generative AI is reshaping our world

86% of IT leaders expect generative AI to soon play a prominent role in their organizations

250% year-over-year growth for generative AI projects on GitHub in 2023

95% of developers are using generative AI tools to write new code

Sources: https://www.forbes.com/sites/bernardmarr/2024/01/29/10-mind-blowing-generative-ai-stats-everyone-should-know-about/?sh=7795815e1bdb; https://www.iotworldtoday.com/connectivity/generative-ai-projects-more-than-triple-on-github-in-2023; https://bloggingwizard.com/generative-ai-statistics/

Implementation challenges

Complexity

- External LLM integration
- Separate vector database
- Vector embedding generation
- Difficult to implement natural language capability

AI expertise

- Embedding model selection
- LLM selection
- Meaningfully apply LLMs, embeddings to domain problems
- Performance optimization

High costs

- Hiring AI experts
- Provisioning GPUs
- Storing vector embeddings
- Optimizing system resources

HeatWave GenAI: Integrated and automated Generative AI

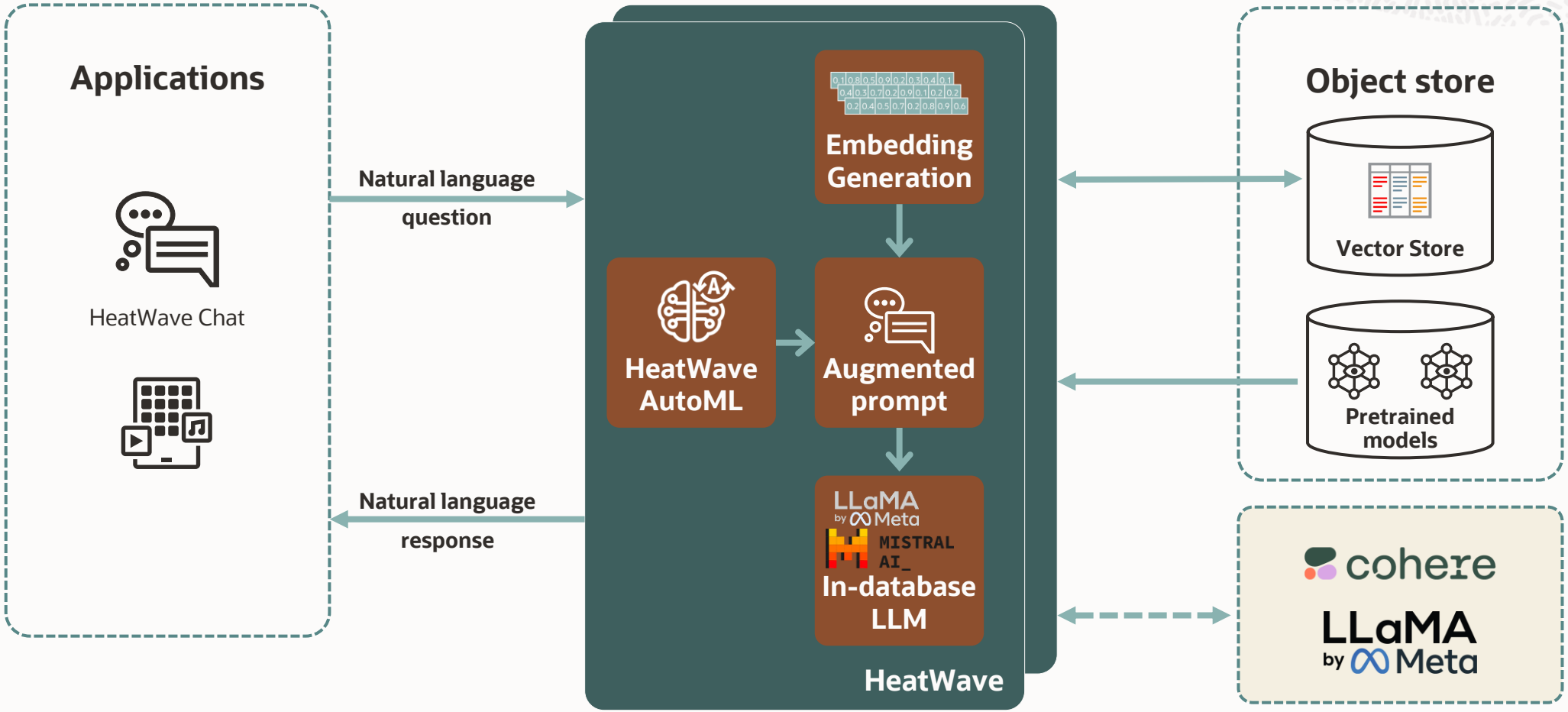
No AI expertise required, no data movement, and no additional cost

In-database LLMs

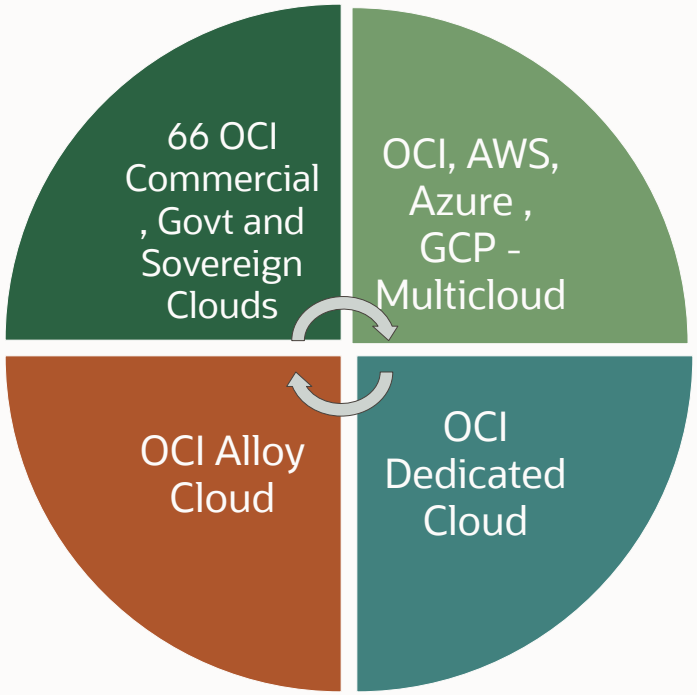
- Quickly benefit from GenAI anywhere without integration hassle
- Help reduce infrastructure costs
- Use external LLMs via integration with OCI Generative AI

In-database LLMs and in-database embedding generation

Also integrated with OCI Generative AI service



Available wherever you need it



Use the same LLMs, embeddings, and vector store across OCI regions and other public clouds

Benefits of in-database LLMs

Simplicity

- No need to select and integrate external LLMs
- Develop turnkey GenAI apps, ready out-of-the-box
- Choose external LLMs if needed for your use case

Lower cost

- No additional cost to use LLMs
- System resources are optimized

Flexibility

- Use HeatWave GenAI across regions and clouds, with consistent results across deployments
- Integration with HeatWave AutoML enables new applications and higher quality results

Security and performance

- Data doesn't leave the database - data isolation
- Not a shared service - performance isolation

HeatWave GenAI: Integrated and automated Generative AI

No AI expertise required, no data movement, and no additional cost

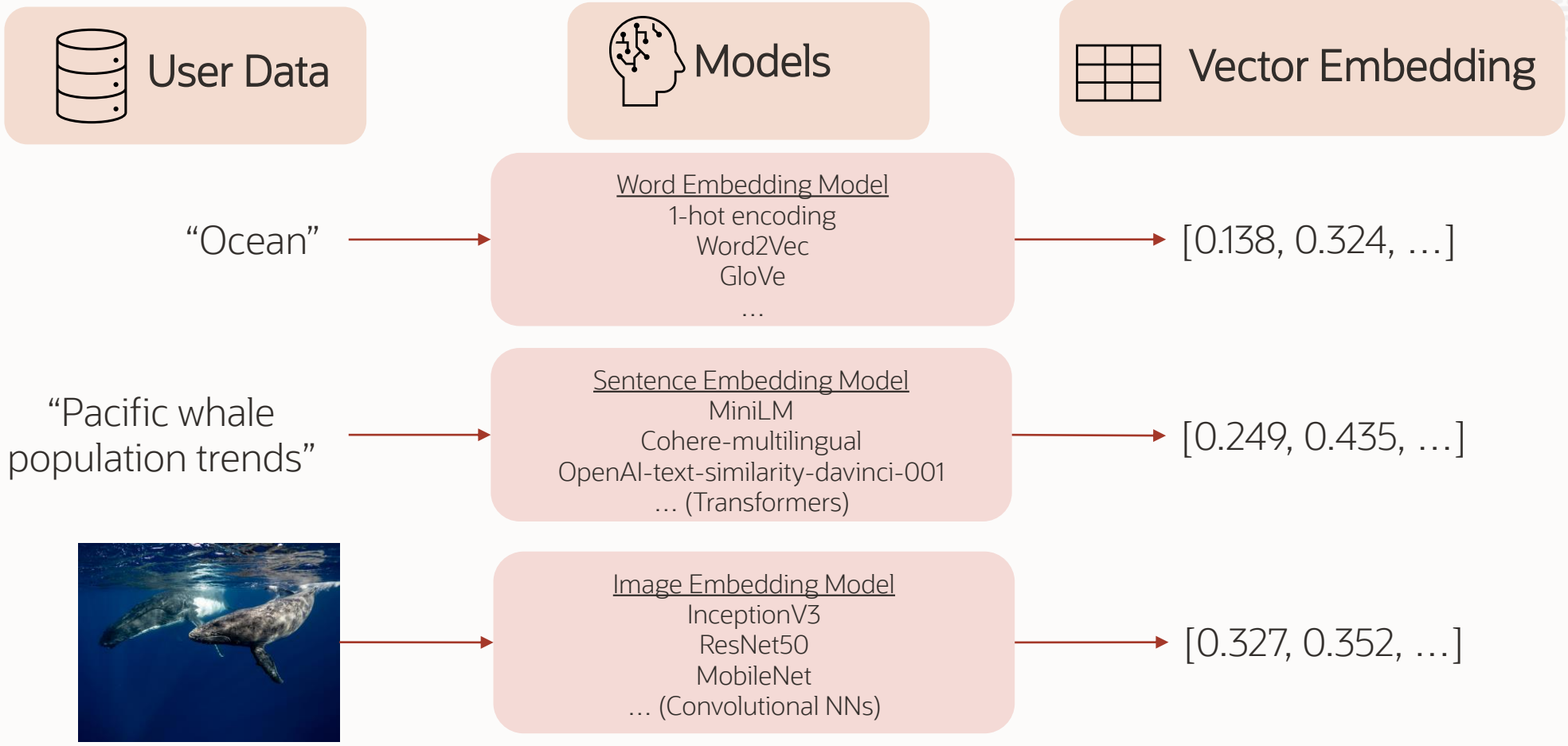
In-database LLMs

- Quickly benefit from GenAI anywhere without integration hassle
- Help reduce infrastructure costs
- Use external LLMs via integration with OCI Generative AI

Automated, in-database vector store

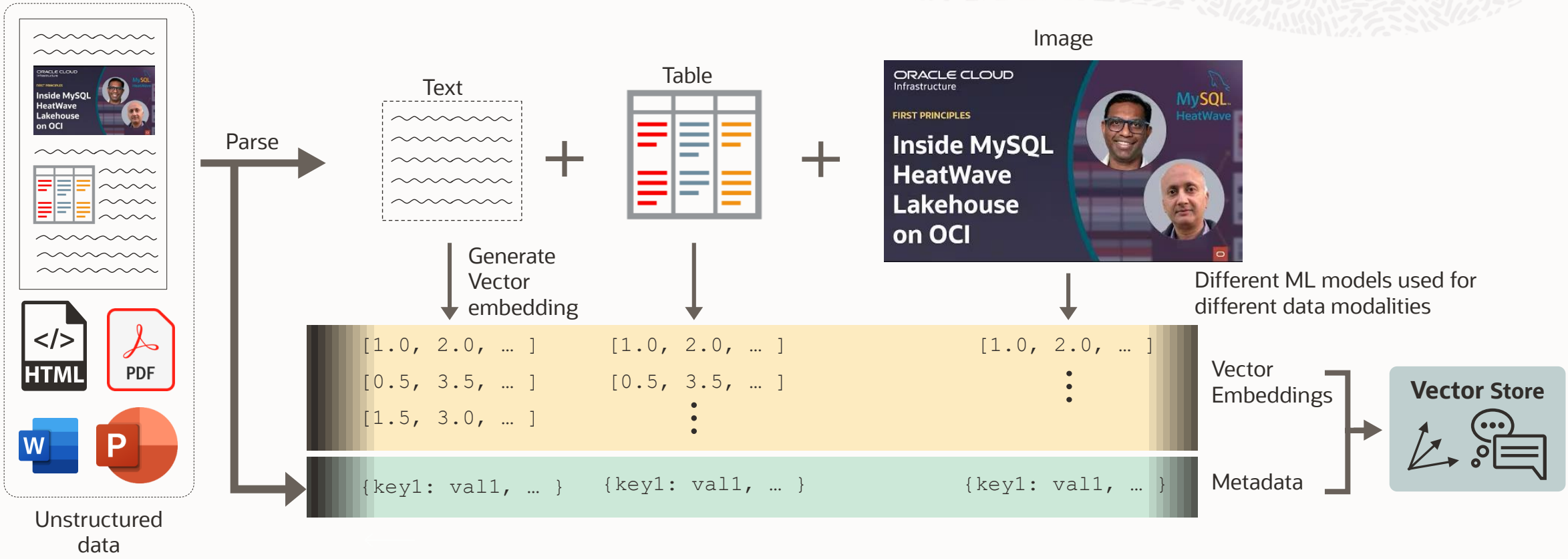
- Use GenAI with your business data without moving data to a separate vector database
- Automate vector embedding generation without AI expertise
- Combine GenAI with in-database ML

Vector is a compressed representation of data



Entities that are similar/related will be closer in the latent space

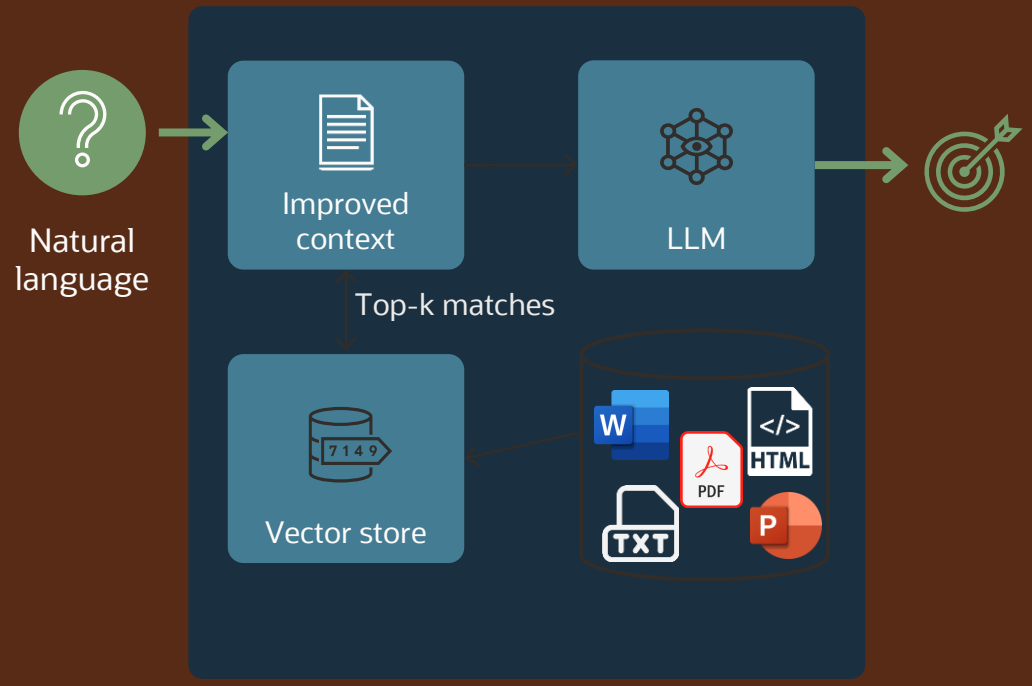
Unstructured data is transformed in vector store



Automatically generate embedding for text from multiple file formats

Enables new search capabilities for unstructured enterprise data

Brings generative capabilities of LLMs to *enterprise content*



Enables *semantic* search on unstructured data

Esteemed scientists from the turn of the century

Document 1

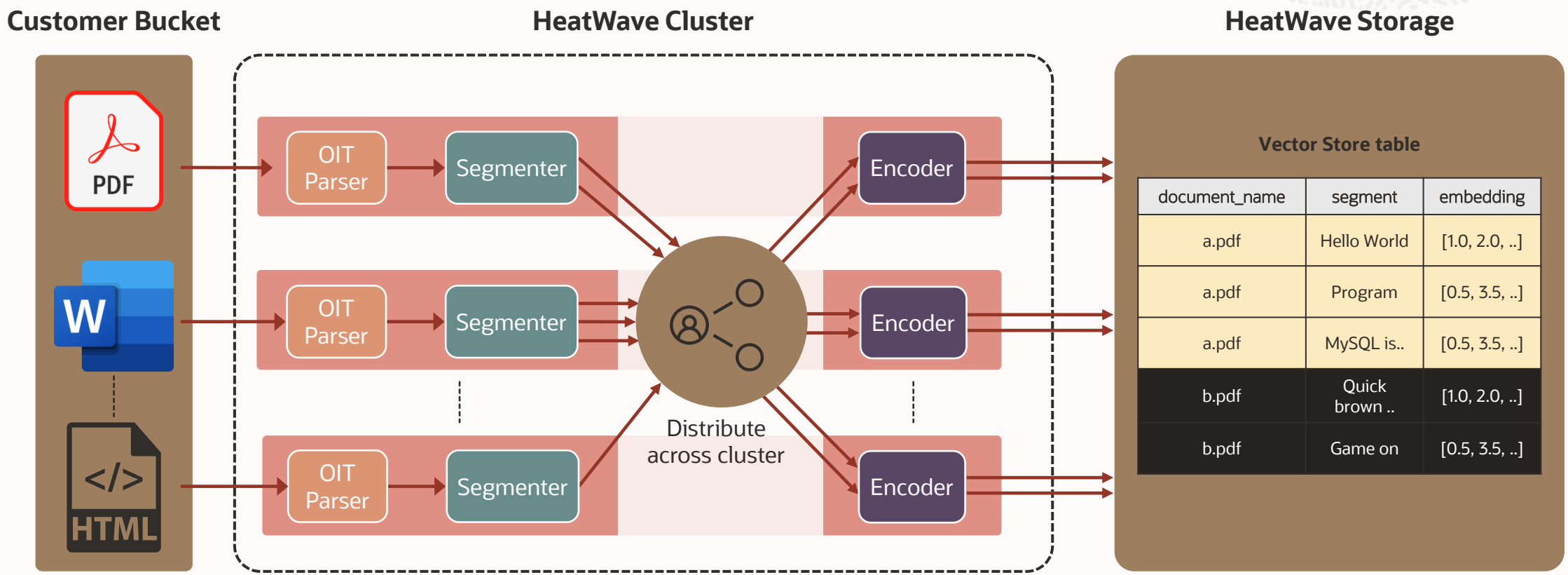
Nobel-prize winning scientists like Marie Curie, Max Planck,…

Document 2

…Wilhelm Rontgen, who won the inaugural Nobel Prize in Physics in 1901…

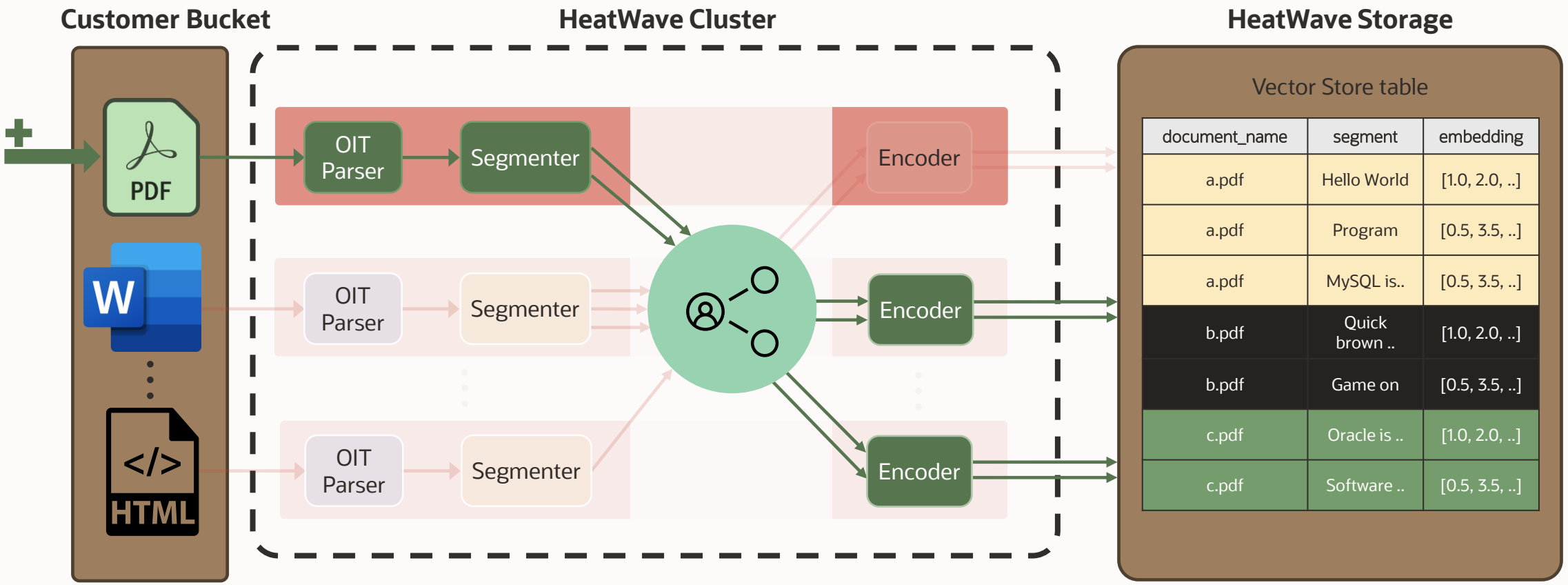
All the steps for vector store creation are completed inside HeatWave

All system resources are optimized by HeatWave



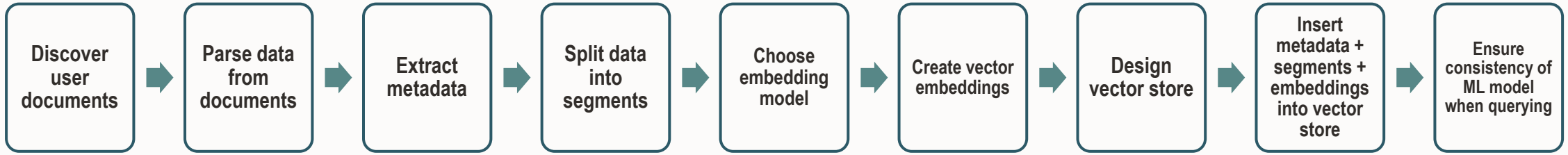
Faster than generating a vector store at the application layer

Changes to data are incrementally updated in the vector store



Building GenAI applications with most databases is complex

Part 1 - Create a vector store



Part 2 – Use the vector store with LLMs

Only one step with HeatWave

Part 1 - create a vector store

```
SQL> call sys.heatwave_load(schema_name, @source_location)
```

Part 2 – query the vector store

```
SQL> sys.ML_RAG("What is HeatWave?", @NL_response, @optional_search_params)
```

- Simpler and faster
- No additional cost

Benefits of HeatWave Vector Store

Simplicity

- Reduced application complexity: in-database and single-step process
- No AI expertise required
- Data changes are incrementally updated to the vector store

Lower cost

- No additional resources needed
- Vector embeddings are persisted in object storage
- No charge for using the embedding function

Security and performance

- Data transformation completed inside the database
- Processing parallelized and tuned inside the database

HeatWave GenAI: Integrated and automated Generative AI

No AI expertise required, no data movement, and no additional cost

In-database LLMs

- Quickly benefit from GenAI anywhere without integration hassle
- Help reduce infrastructure costs
- Use external LLMs via integration with OCI Generative AI

Automated, in-database vector store

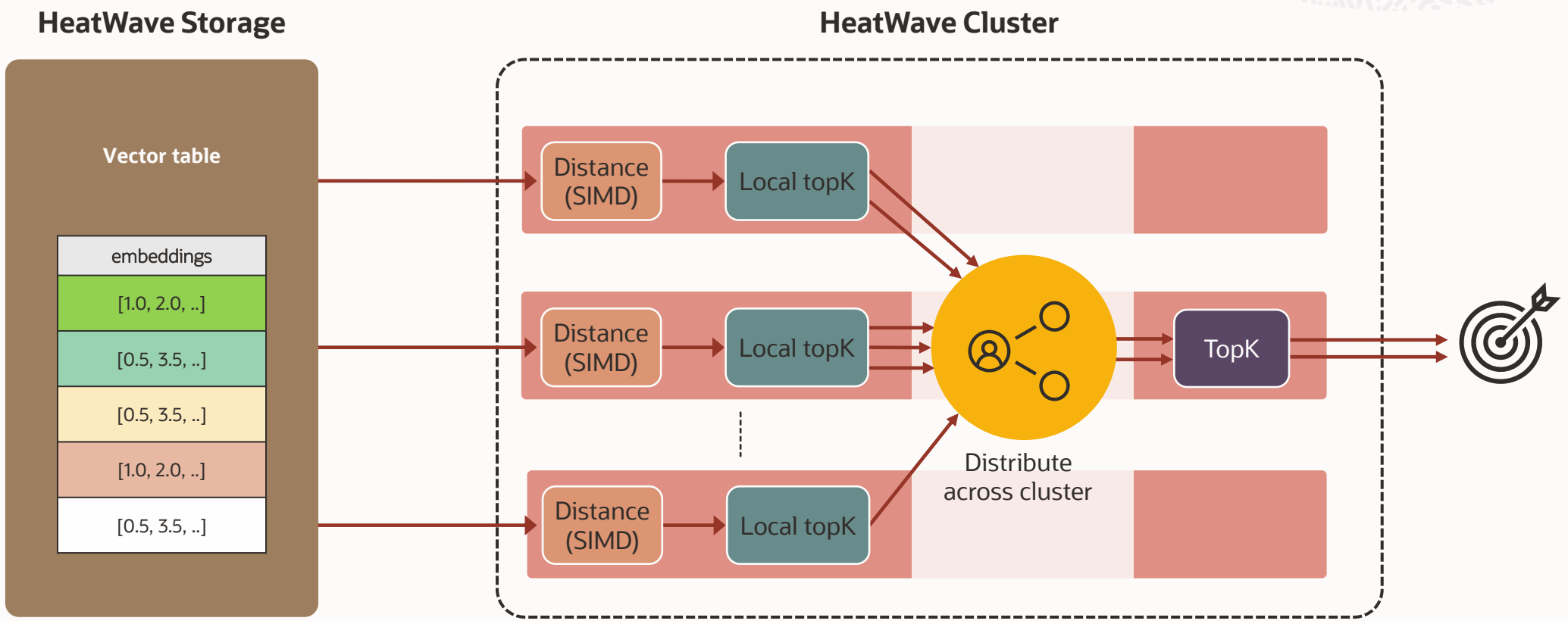
- Use GenAI with your business data without moving data to a separate vector database
- Automate vector embedding generation without AI expertise
- Combine GenAI with in-database ML

Scale-out vector processing

- In-memory, scale-out architecture
- Perform fast semantic searches
- 15X faster than Databricks, 18X faster than Google BigQuery, and 30X faster than Snowflake.

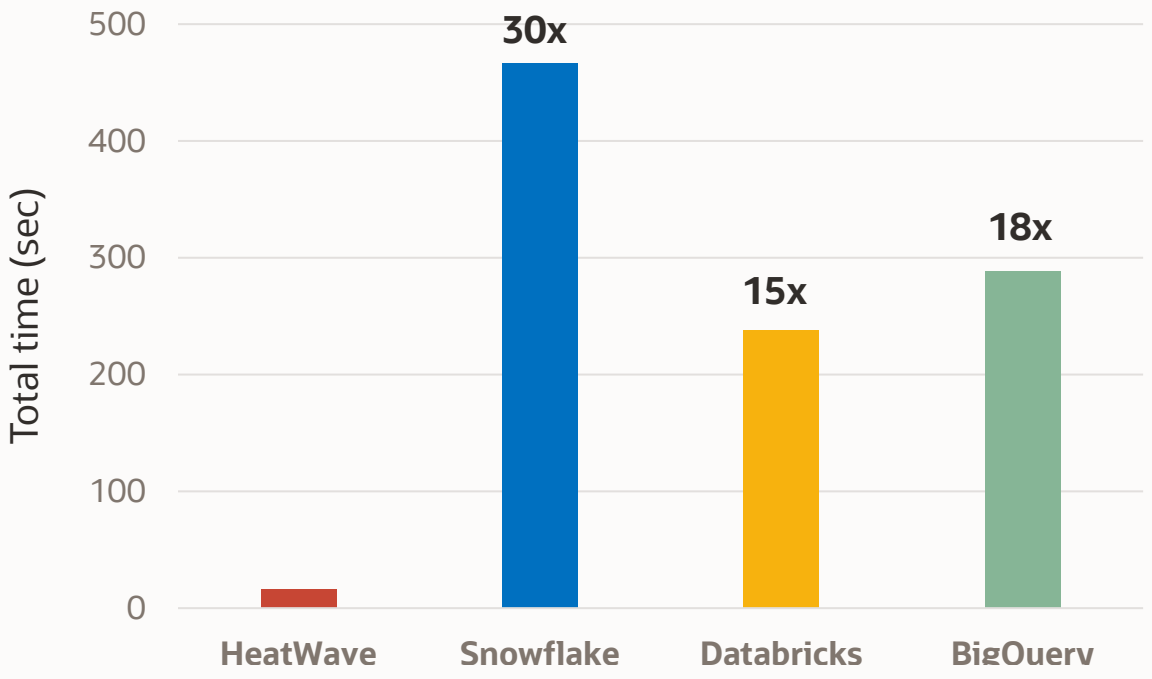
Similarity search in HeatWave is exact and done at near memory bandwidth

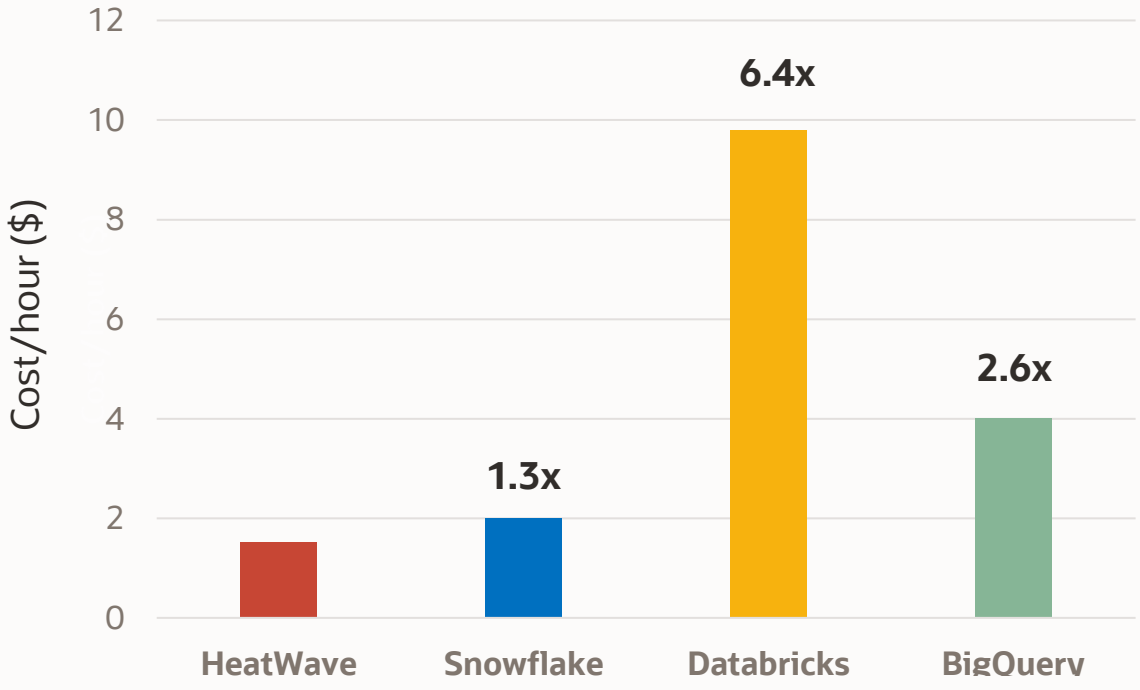
Scales to 512 nodes



Similarity search

HeatWave GenAI is 15X-30X faster and less expensive





https://www.oracle.com/heatwave/performance-benchmarks/#heatwave-genai

HeatWave GenAI: Integrated and automated Generative AI

No AI expertise required, no data movement, and no additional cost

In-database LLMs

- Quickly benefit from GenAI anywhere without integration hassle
- Help reduce infrastructure costs
- Use external LLMs via integration with OCI Generative AI

Automated, in-database vector store

- Use GenAI with your business data without moving data to a separate vector database
- Automate vector embedding generation without AI expertise
- Combine GenAI with in-database ML

Scale-out vector processing

- In-memory, scale-out architecture
- Perform fast semantic searches
- 15X faster than Databricks, 18X faster than Google BigQuery, and 30X faster than Snowflake.

HeatWave Chat

- Engage in natural language conversations informed by unstructured documents
- Ask follow-up questions; chat context preserved
- Guide LLMs to retrieve information from specific data sets to help increase speed and accuracy

HeatWave Chat

Chat

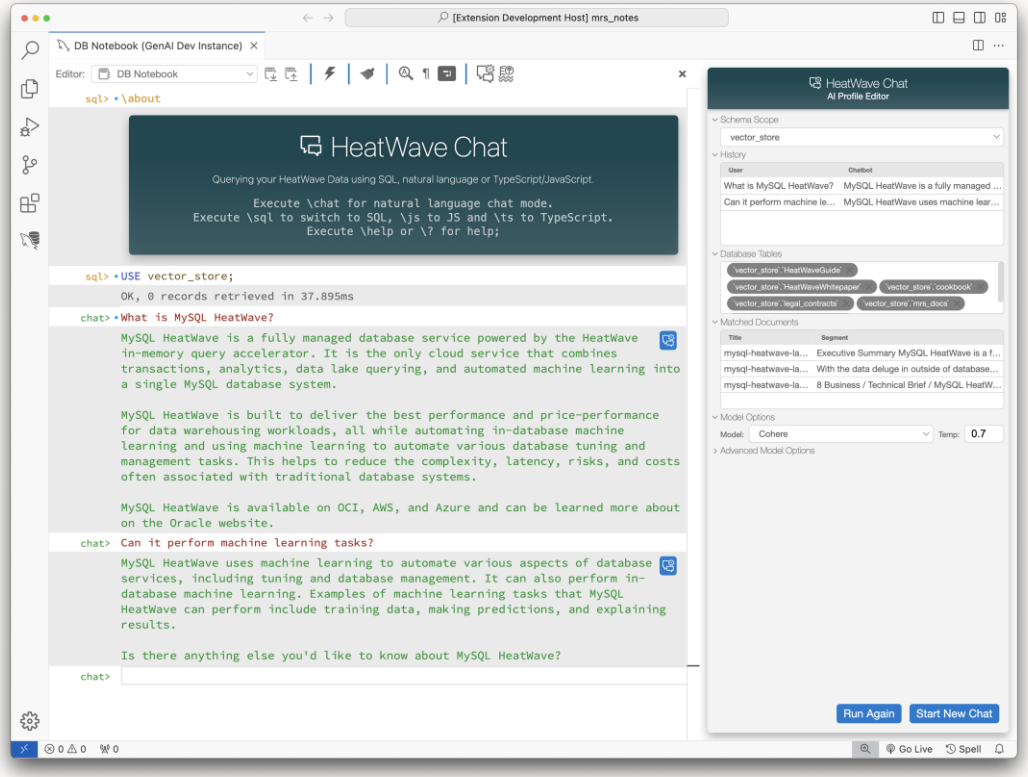
Interact with your documents using natural language. Context is preserved to enable conversations with follow-up questions.

Lakehouse Navigator

Guide LLMs to retrieve information from specific datasets across the database, HeatWave Lakehouse, and HeatWave Vector Store to increase speed and accuracy.

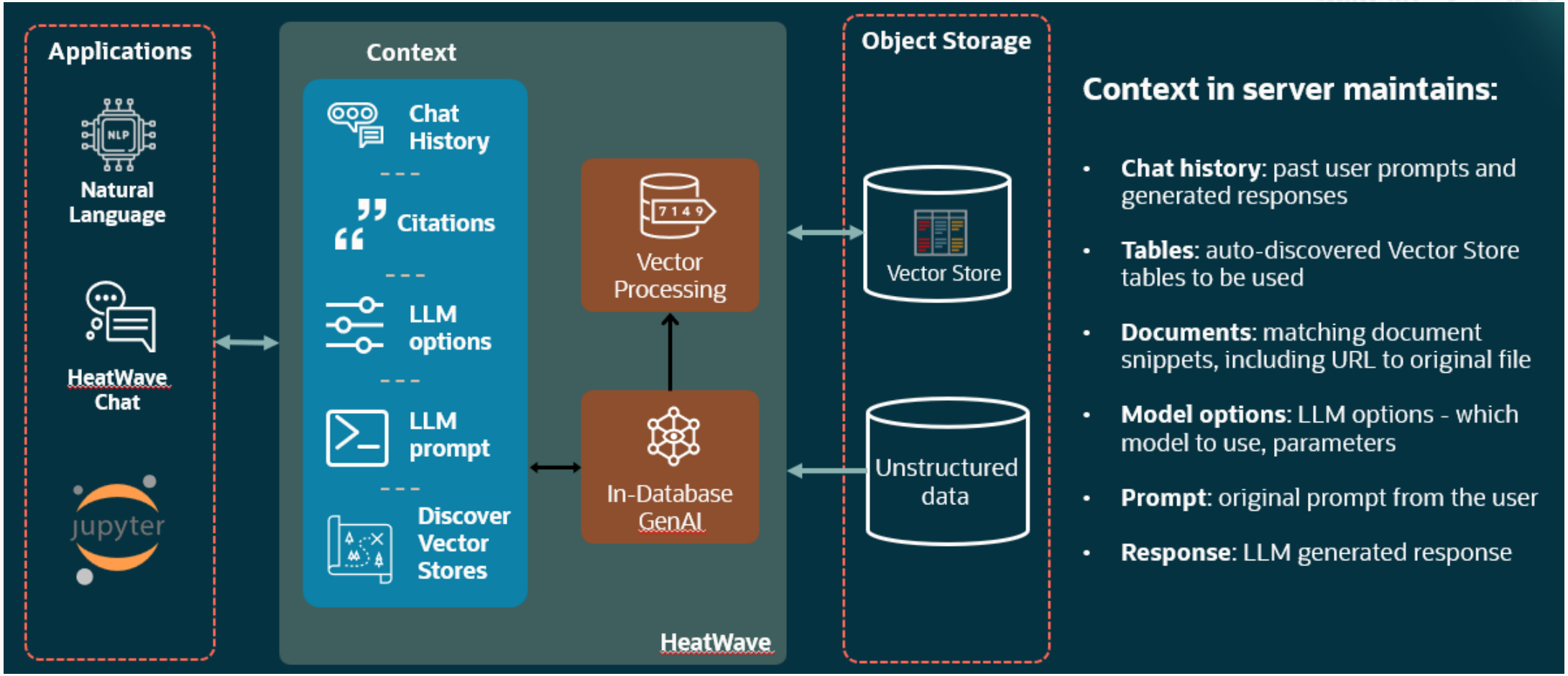
Global and refined search

Query all the vector stores or limit the scope of the search to a particular schema.



HeatWave provides support for chat capabilities

Chat context maintained in the server for applications to use



HeatWave GenAI – Use cases

HeatWave GenAI enables new use cases and apps

Conversations in natural language

- Conversations informed by your unstructured documents using natural language
- HeatWave Chat preserves context for follow-up questions

Content generation and summarization

- Generate insights/reports from enterprise documents
- Generate blogs from PDF instruction manuals
- Summarize content

RAG and similarity search

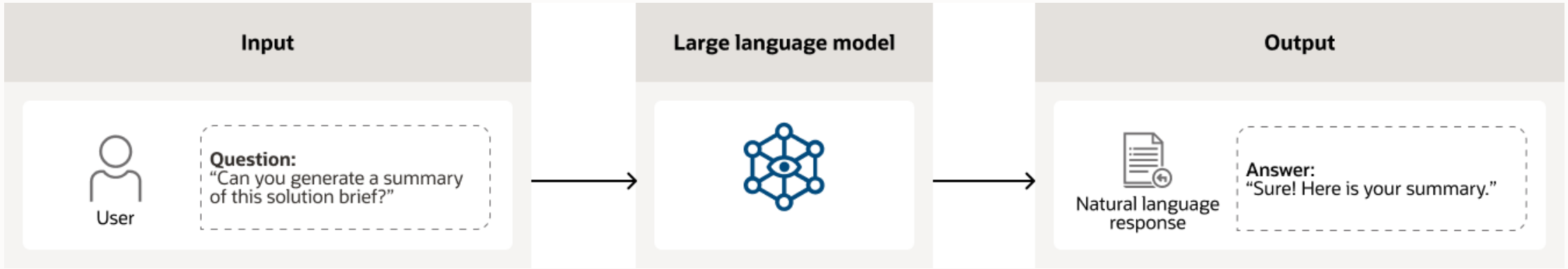
- Use GenAI with your organization's data (Retrieval Augmented Generation) to get more accurate and contextually relevant answers
- Perform similarity searches on unstructured data

Synergy of integrated GenAI and ML

- Save time and deliver more value to customers by combining ML and GenAI
- Help reduce costs and get more accurate results faster by using GenAI on data filtered by AutoML

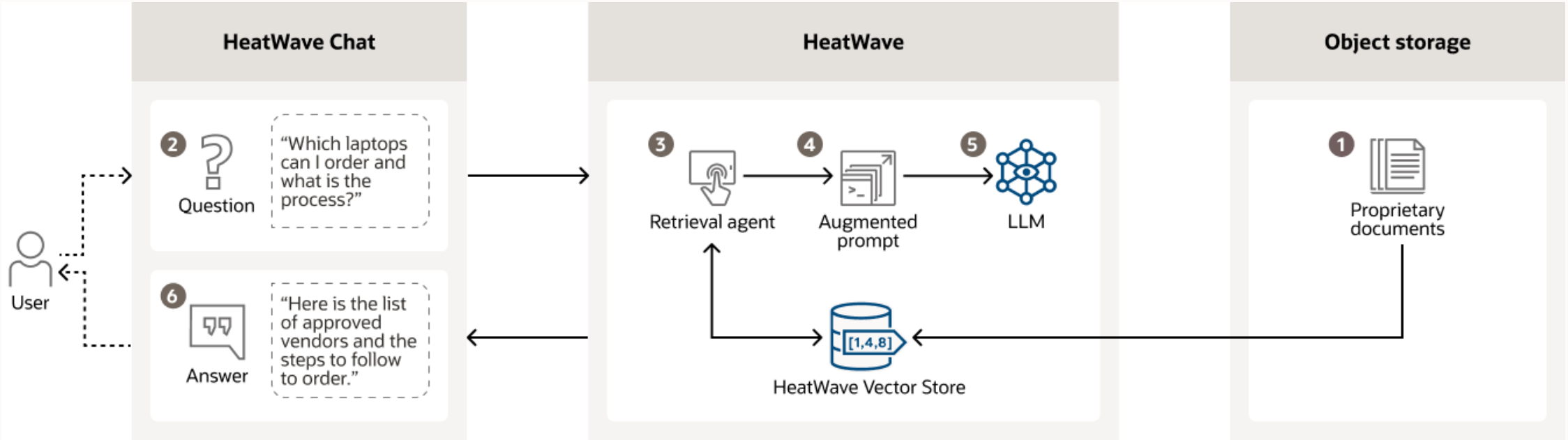
USE CASE 1: Content generation

Summarization



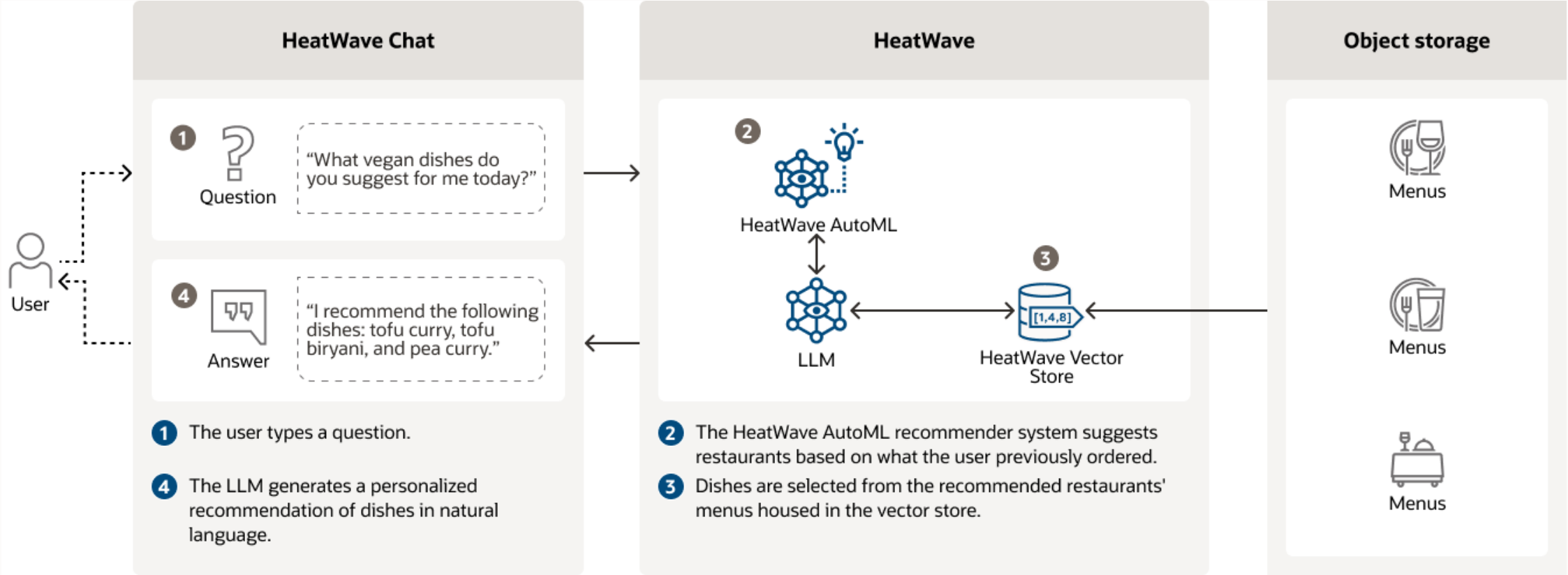
USE CASE 2: Retrieval Augmented Generation (RAG)

Accessing internal policy documents to get fast answers



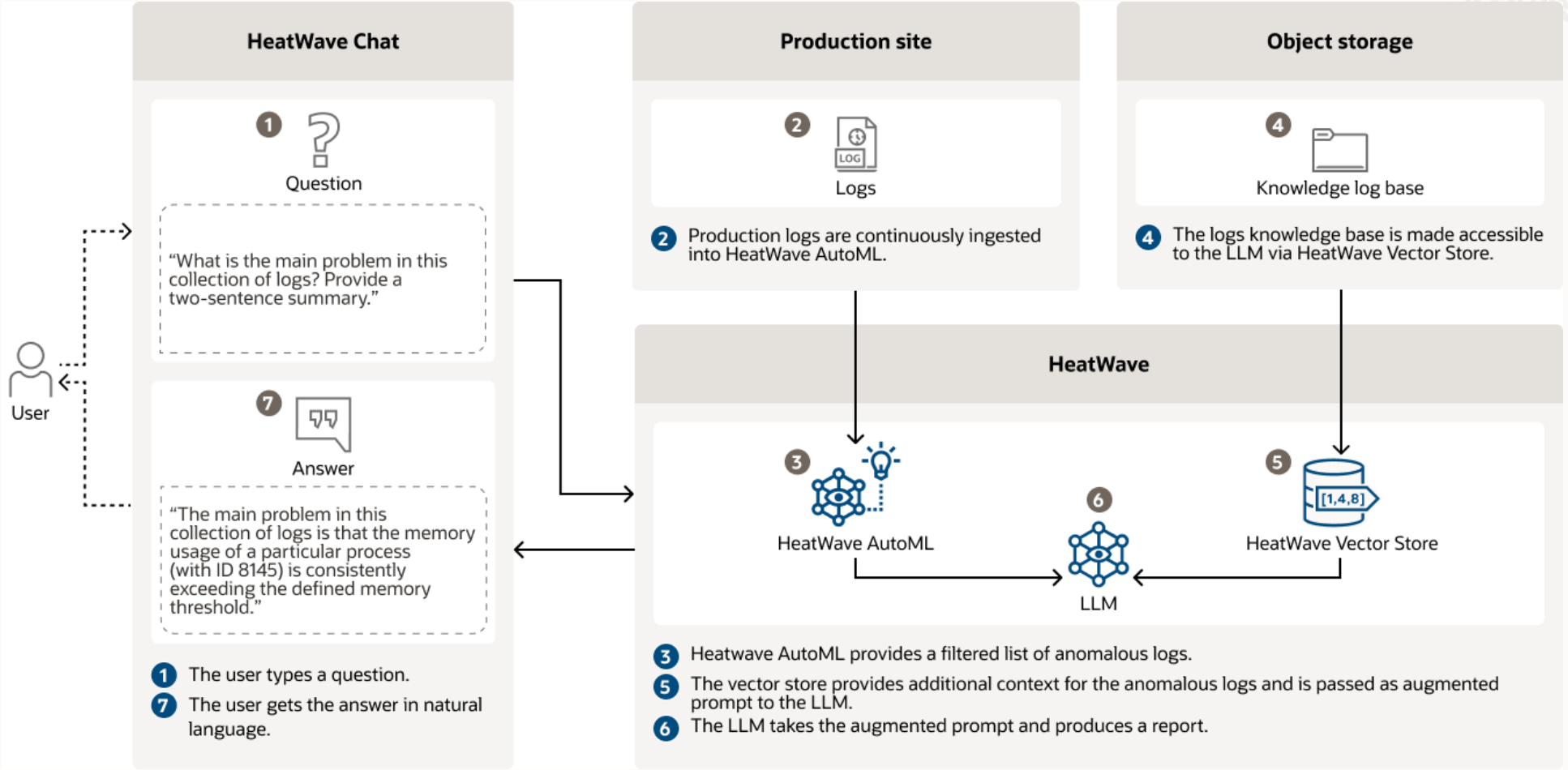
USE CASE 3: Personalized recommendations

RAG enhanced with ML



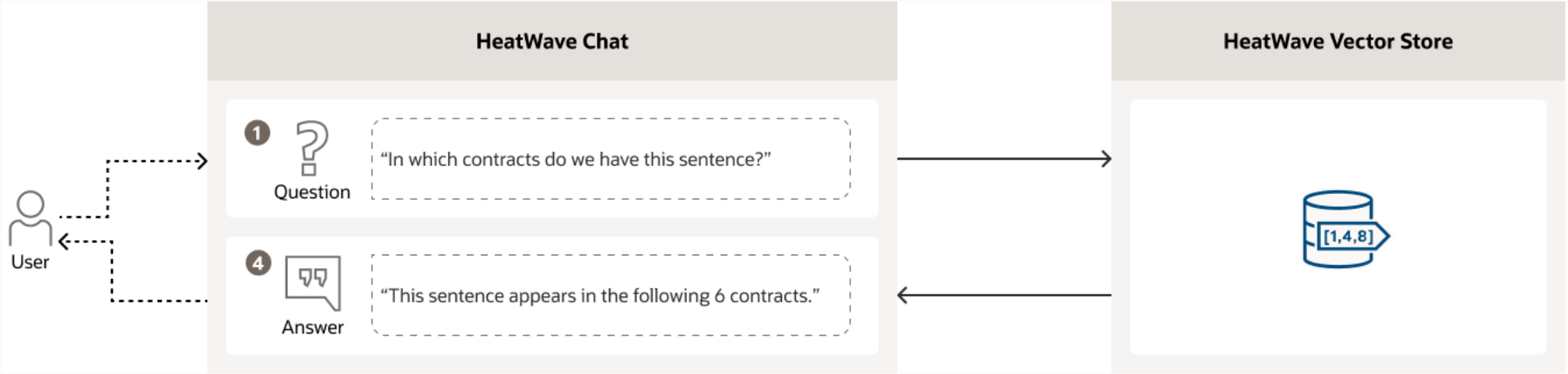
USE CASE 4: Predictive maintenance

Analysis Generation



USE CASE 5: Similarity search

Identifying a potentially problematic clause in contracts



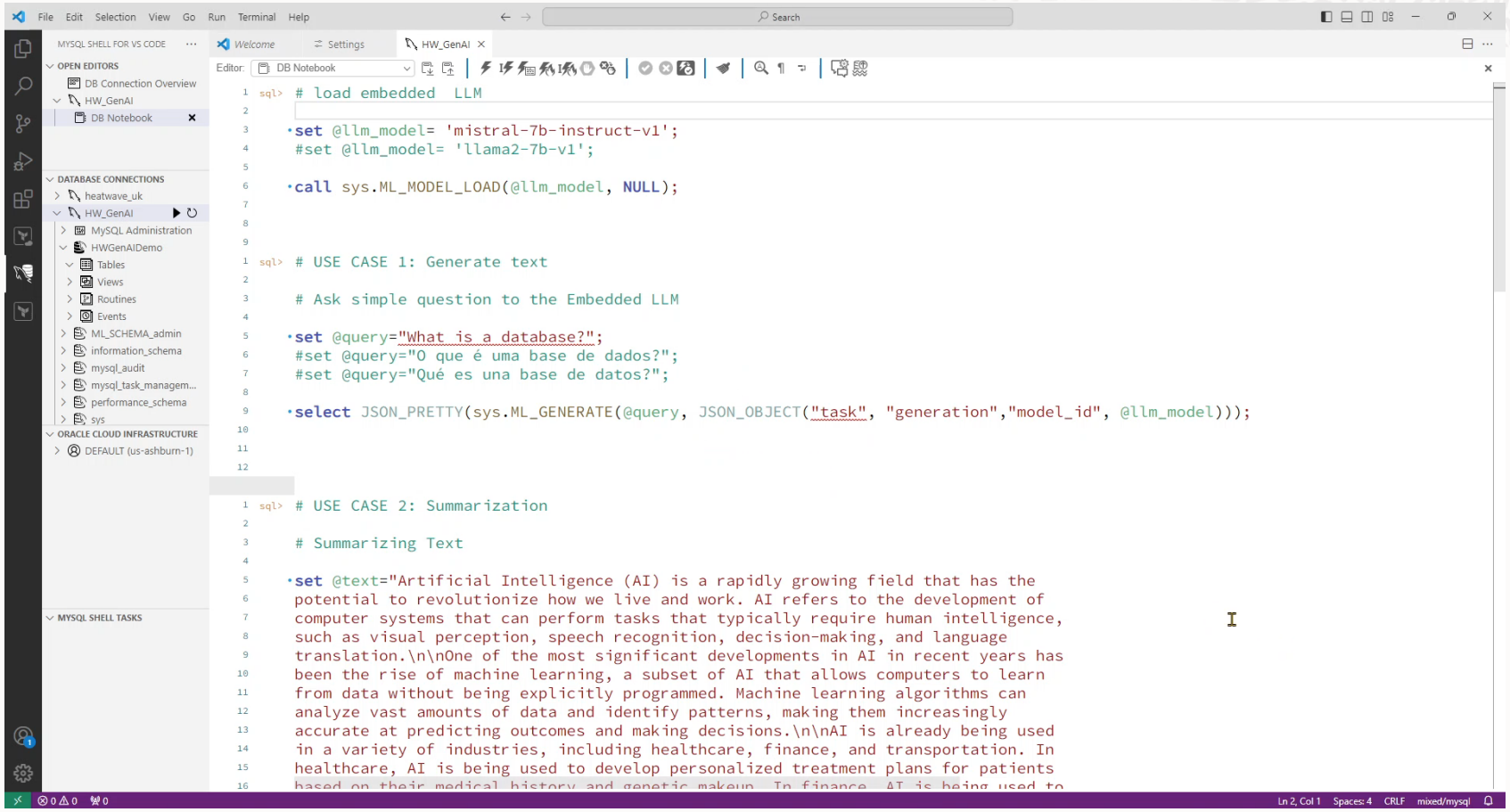
"HeatWave GenAI makes it **extremely simple to take advantage of generative AI**. The support for in-database LLMs and in-database vector creation leads to **significant reduction in application complexity**, predictable inference latency, and most of all **no additional cost to us to use the LLMs or create the embeddings**. This is truly the democratization of generative AI and we believe it will result in building richer applications with HeatWave GenAI, and **significant gains in productivity** for our customers."

Vijay Sundhar
CEO, SMARTERD

"We heavily use the in-database HeatWave AutoML for making various recommendations to our customers. HeatWave's support for in-database LLMs and in-database vector store is differentiated and **the ability to integrate generative AI with AutoML provides further differentiation for HeatWave** in the industry, enabling us to offer **new kinds of capabilities** to our customers. The synergy with AutoML also **improves the performance and quality** of the LLM results."

Safarath Shafi
CEO, Eat Easy

Demo

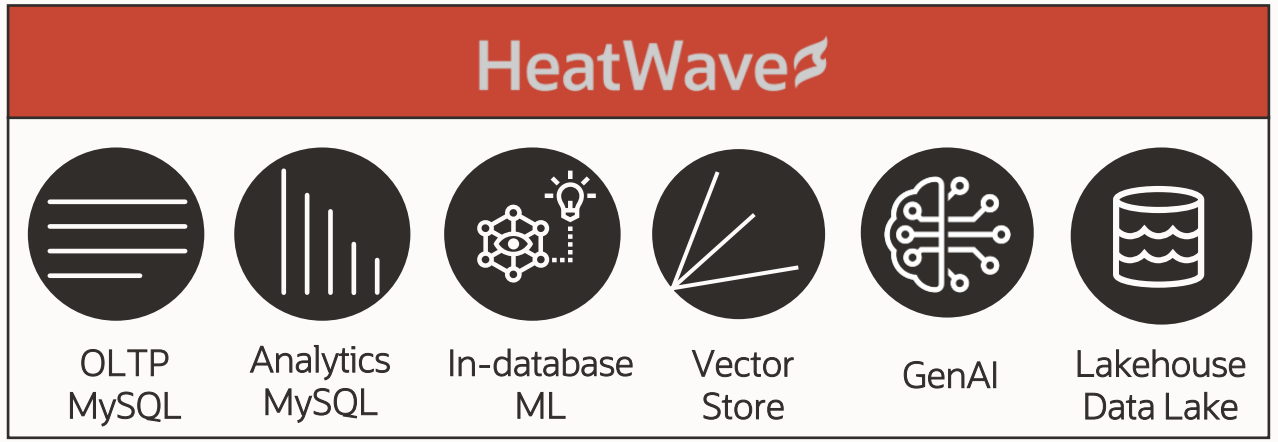


HeatWave GenAI

https://www.oracle.com/heatwave/genai/

Integrated **Automated** **Lower cost**

- Enables querying of unstructured documents
- Allows semantic search of content
- Users can query unstructured documents in natural language



Success Stories

Estuda.com achieves real-time insights

- Brasil's leading ed-tech serves over 8 million students from more than 500 K-12 schools. It needed a data platform to deliver real-time insights for 3 million users per month.
- Real-time analytics enable faster development to improve app usability and adoption
- Scales queries to any data size for more flexibility growth to impact more students

300x

300X faster performance from migrating from BigQuery to MySQL HeatWave with no code changes and low-latency

85%

85% cost reduction by eliminating ETL processes and pay-for-use consumption model

Read Estuda's story

62 Success cases and growing...

MySQL HeatWave: Case Studies

eD-Online Maximizes Uptime for e-Learning LMS with MySQL HeatWave

eD-Online is a Malaysian SaaS ISV that produces creative content for learning, communication, and entertainment. In the e-learning space, eD-Online began working with MySQL for open source LMS platforms such as Moodle, and other customized solutions. By implementing MySQL HeatWave, eD-Online launched a high availability platform for the business-critical onboarding content delivered through the client's inhouse LMS.

Read more »

EatEasy Transforms Food Delivery Services using MySQL HeatWave AI and ML

The EatEasy mobile application allows users to order from a variety of cuisines across more than 10,000 restaurants, and offers order tracking, contactless delivery, and the ability to schedule orders in advance. EatEasy deployed MySQL HeatWave on Oracle Cloud Infrastructure (OCI) for its food delivery app, leveraging MySQL-native AI and ML technologies to boost operational efficiencies, drive sales and customer loyalty, and streamline the user experience across order placement and delivery processes.

Read more »

Sinopay, SaaS Payments Gateway Provider, Expands Globally with MySQL HeatWave on Oracle Cloud Infrastructure

NVIDIA

[Click => MySQL HeatWave: Case Studies]

Learn more and stay in touch!

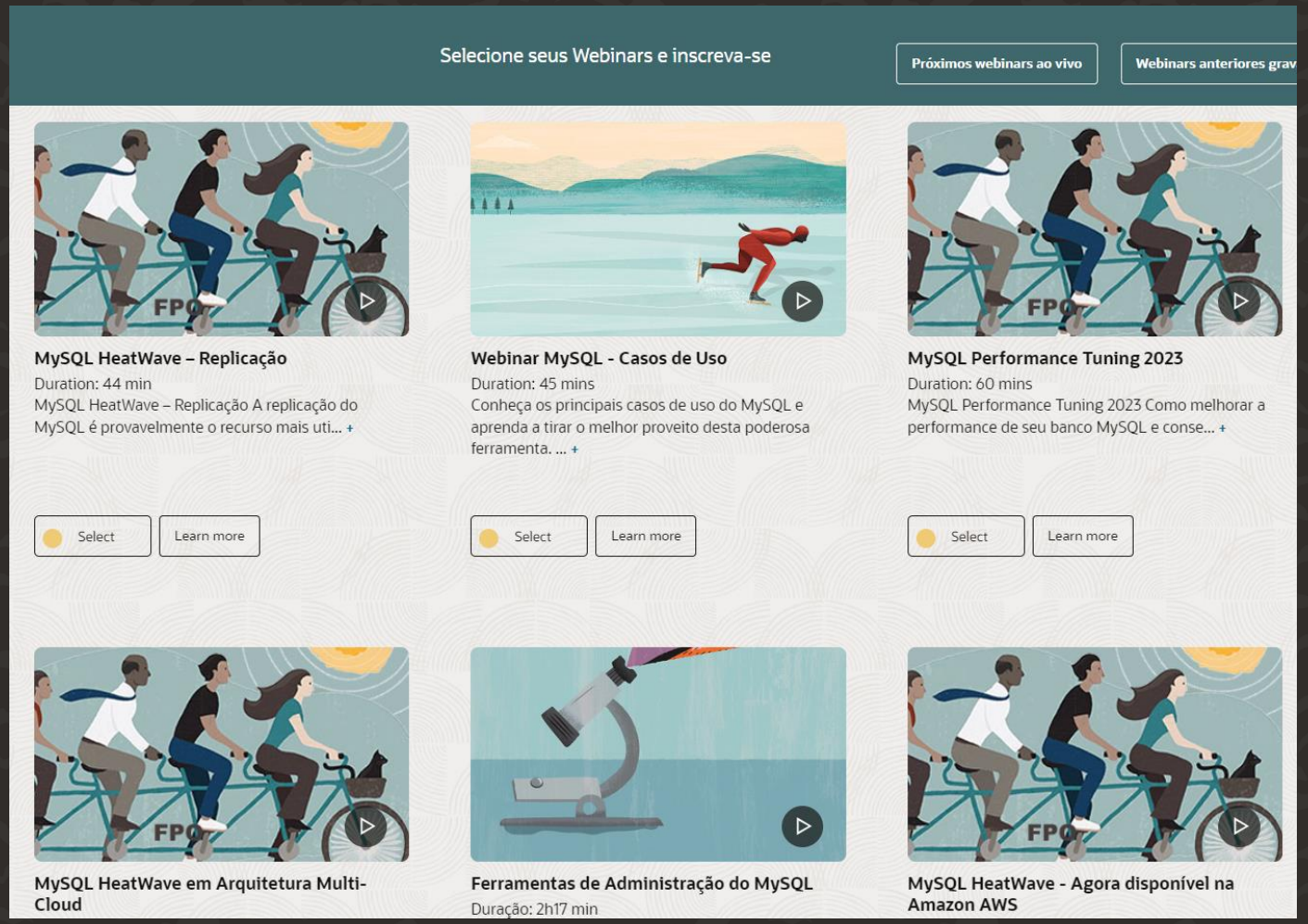
Oracle Cloud Infrastructure (OCI)

Nuvem da Oracle!

Free Trial

Visit us on-line too!

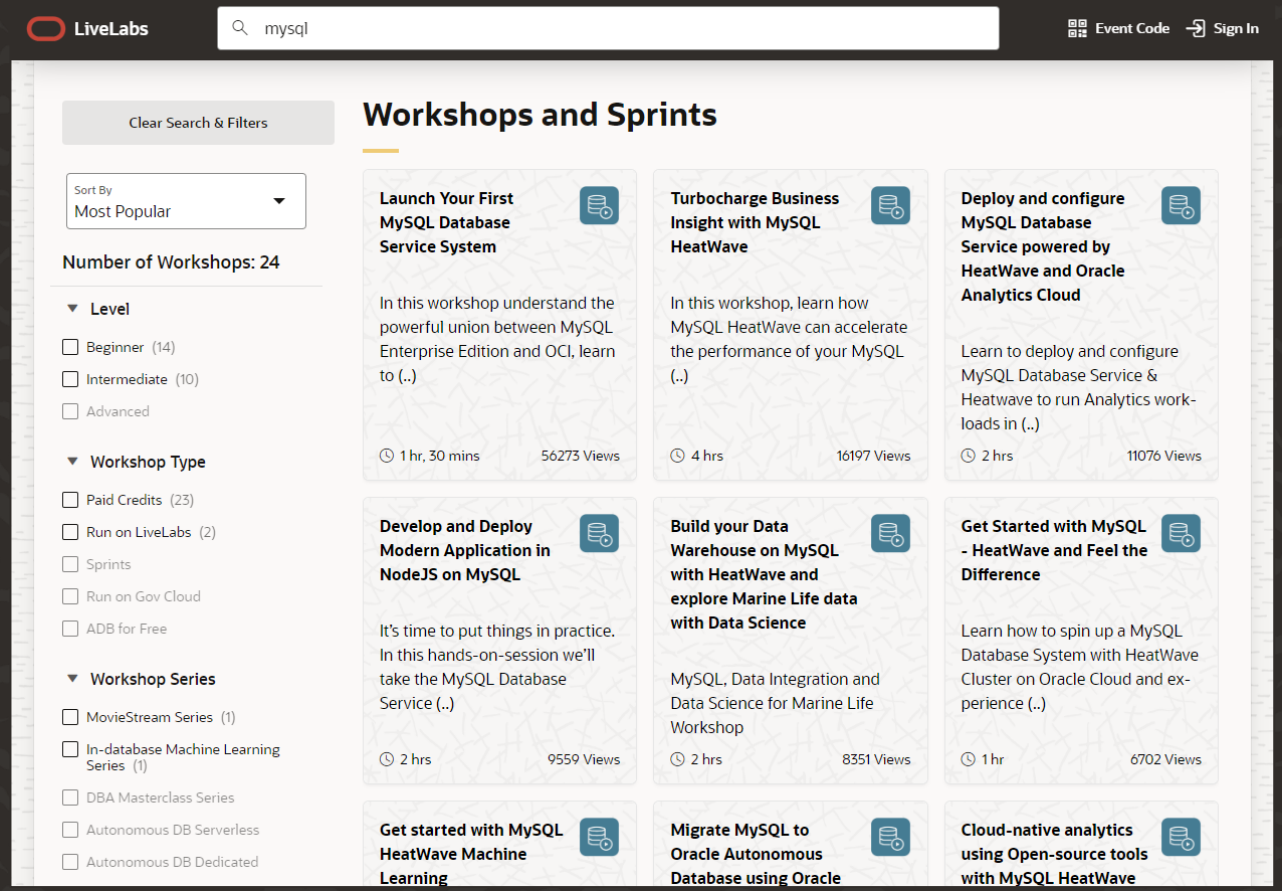
Webinars MySQL em Português



Link => Webinars MySQL em Português

Labs and free courses

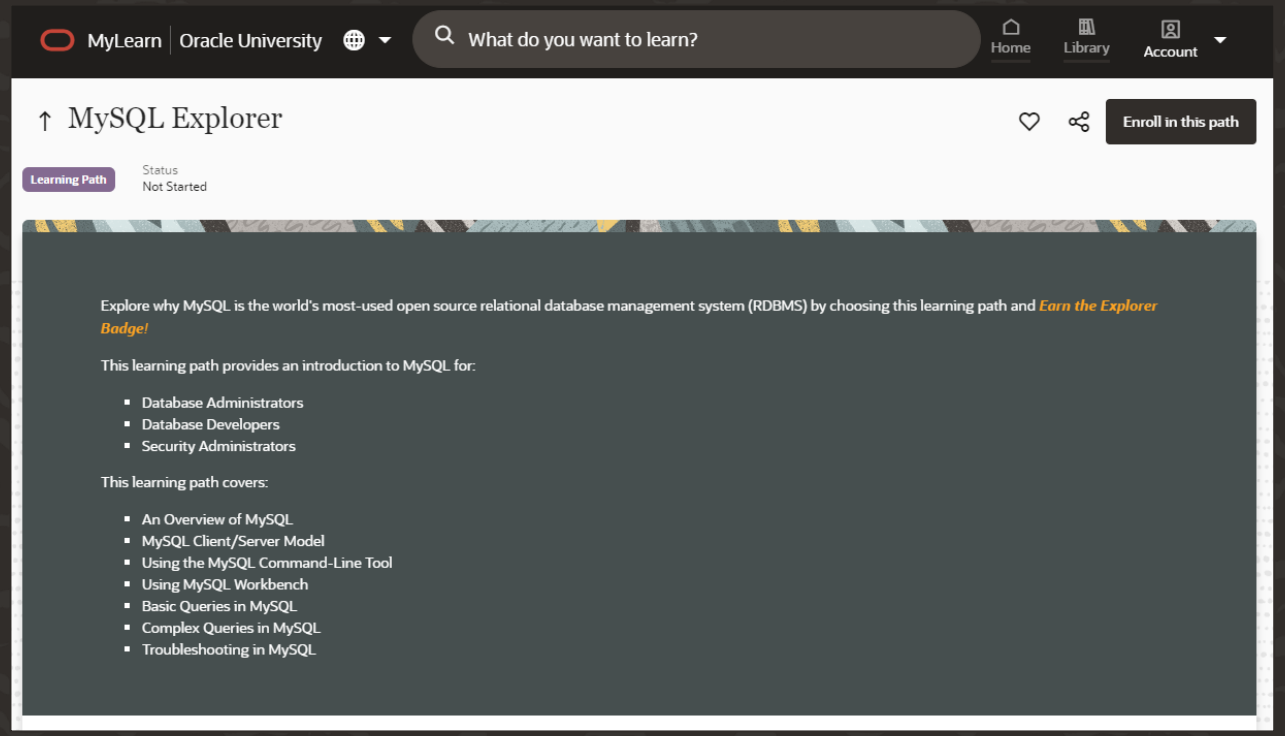
Live Labs



Link => Live Labs

Labs and free courses

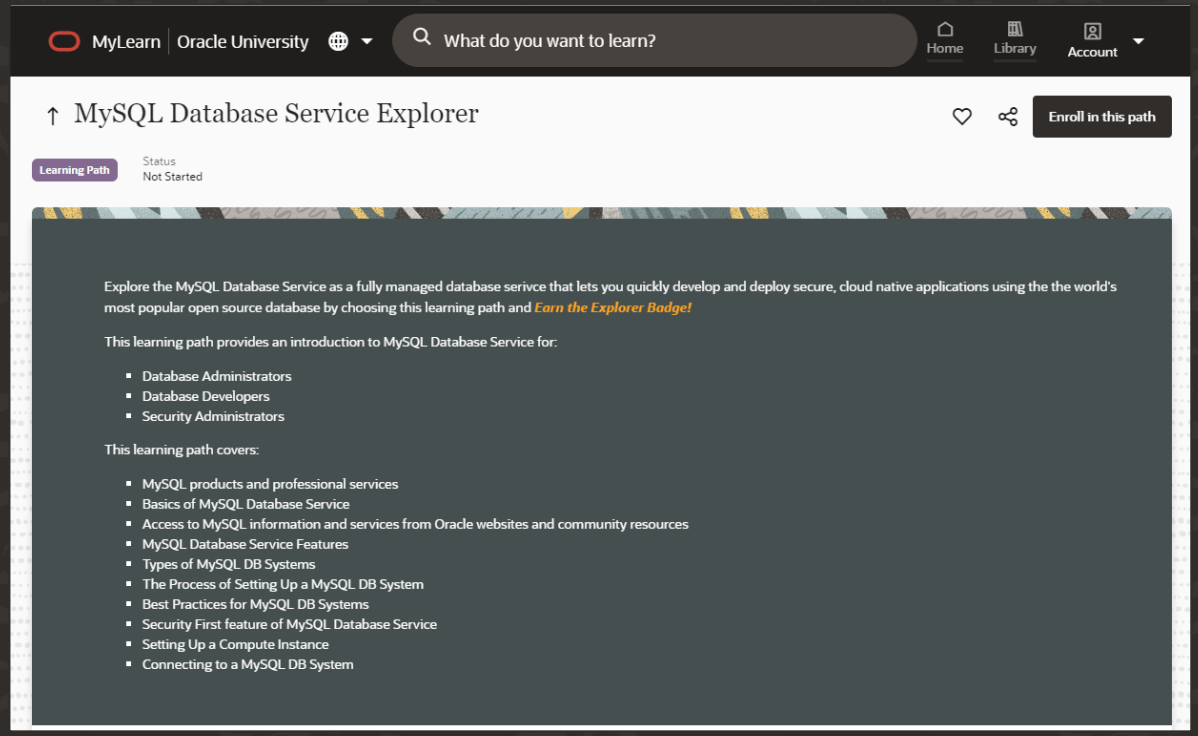
Free MySQL Courses
MySQL Explorer



Link => MySQL Explorer

Labs and free courses

Free MySQL Courses
MySQL Database Service Explorer



Link => MySQL Database Service Explorer

Q&A

Narciso Oliveira Junior
HeatWave Cloud Evangelist
narciso.junior@oracle.com

Thank YOU!