

ORACLE

# HeatWave GenAI: A Peek Under the Hood

---

**Milos Vasic**

Consulting Member of Technical Staff

# Safe harbor statement

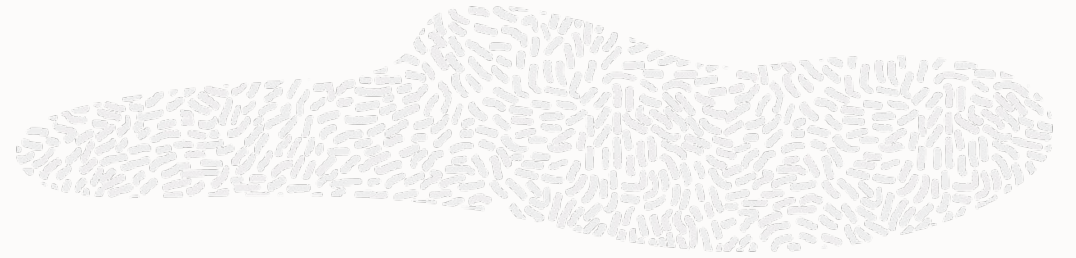


The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.

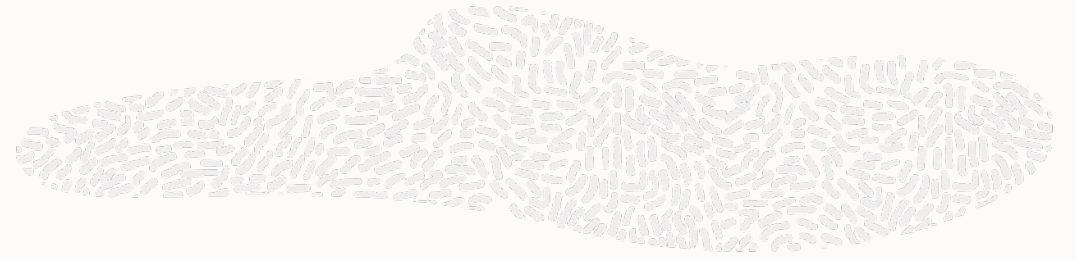
The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

# Who am I?

- PhD from EPFL in Lausanne, Switzerland
- Joined Oracle in 2018
  - Oracle Labs
  - MySQL HeatWave since 2024

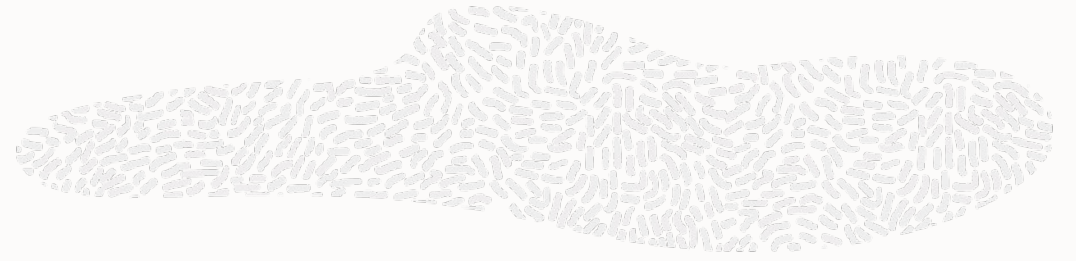


# Agenda



- GenAI in a Nutshell
- HeatWave GenAI Ecosystem
- Challenges
- Examples of Applications built with HeatWave GenAI

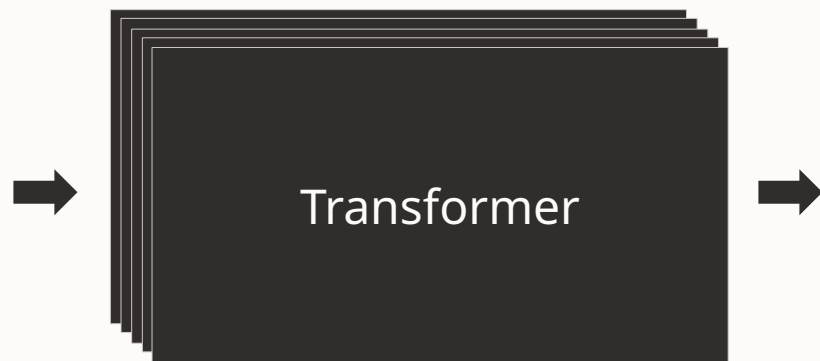
# Transformers: Heart of Generative AI



What follows is a conversation between a user and a helpful, very knowledgeable AI assistant.

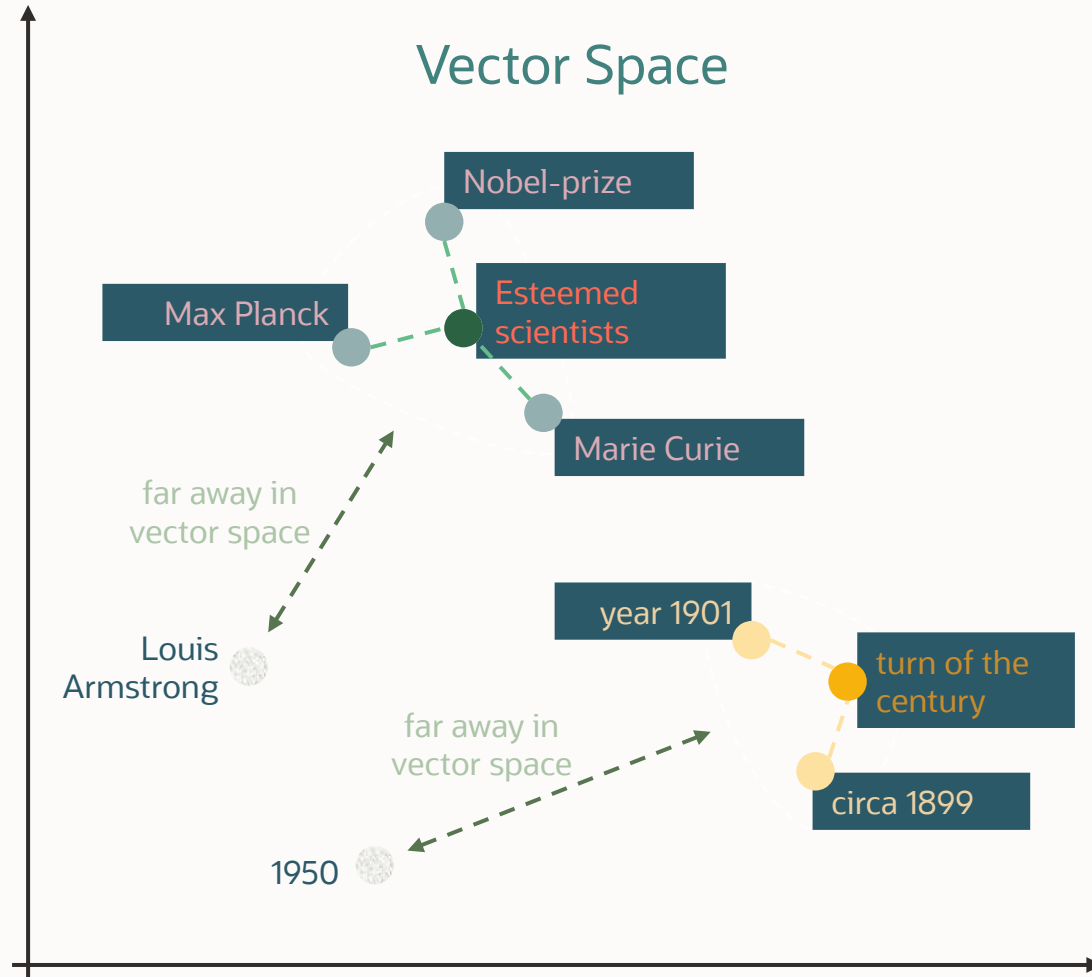
User: Give me some ideas for what to do when visiting Brussels.

AI Assistant: Sure, there are plenty of things to do in



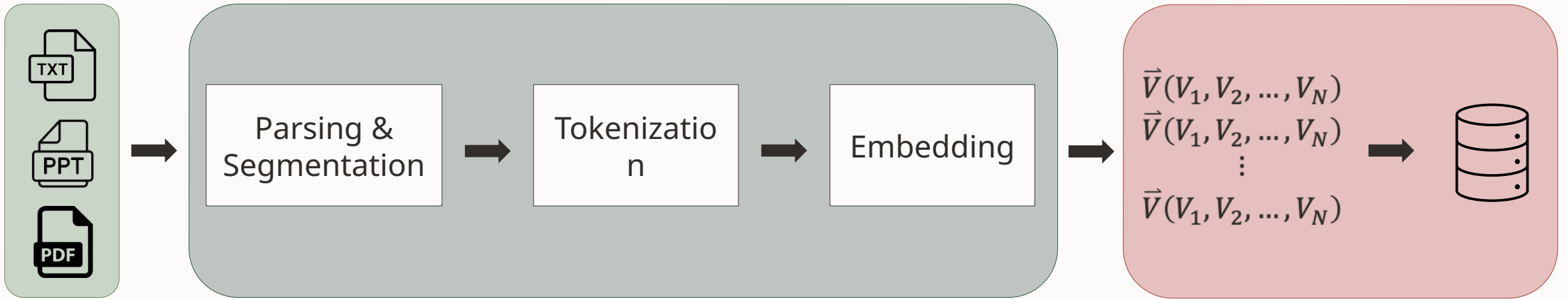
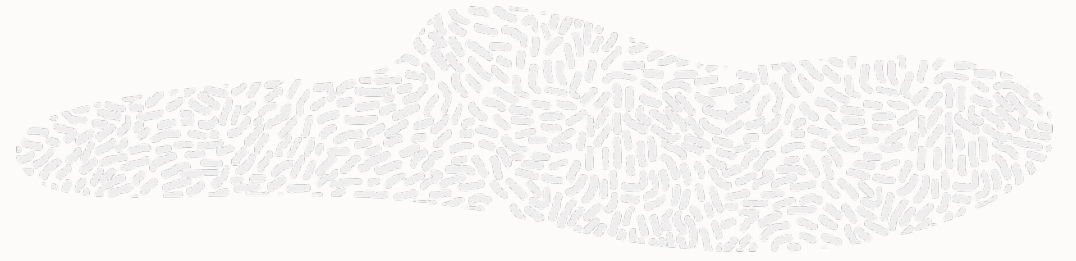
Brussels	95%
the	0%
beautiful	0%
this	1%
Belgium	4%
when	0%
a	0%
...	
once	0%

# Embeddings



- Each point represents vector embedding of data. Data similar in semantic meaning are closer in Vector Space
- Can be applied to various kinds of data (such as words, sentences or documents, but also images or audio)
- Enables searching data semantically (instead of keyword search)
- E.g. “Esteemed scientist” search key word is contextually close to Max Planck or Marie Curie

# Vector Store

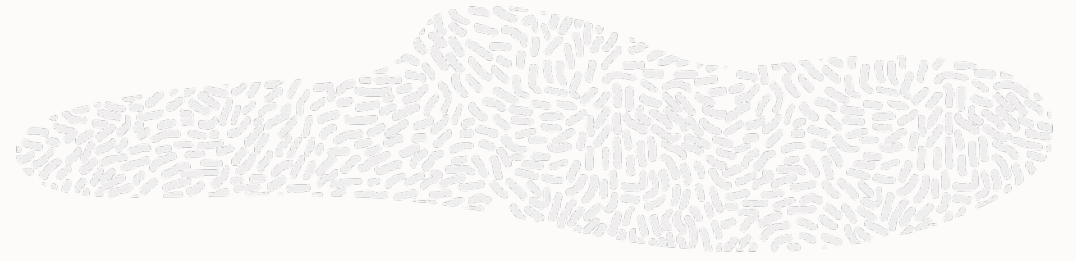


Proprietary Data

Vector Embedding Creation

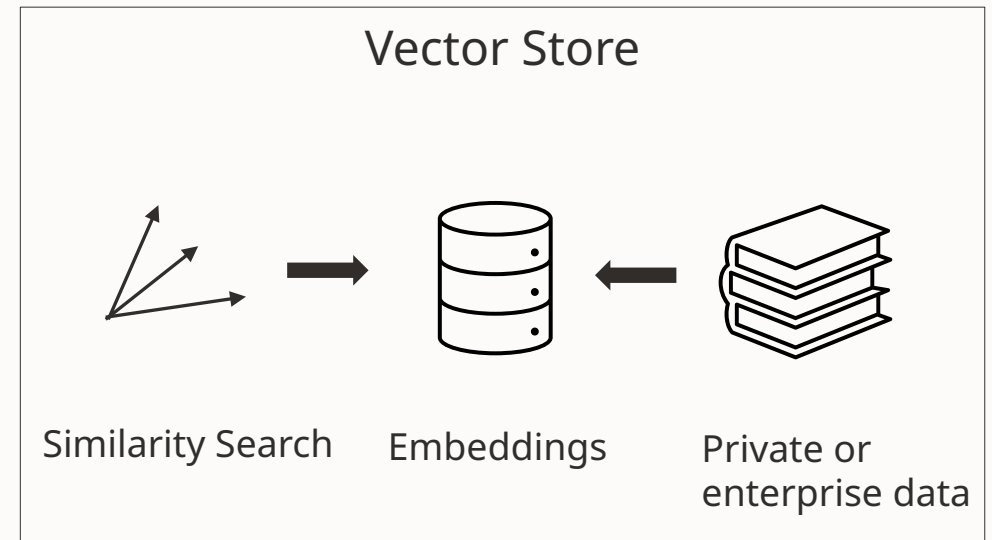
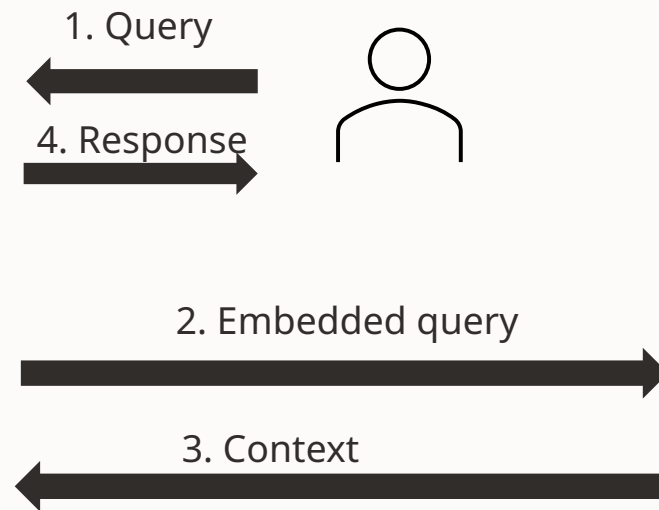
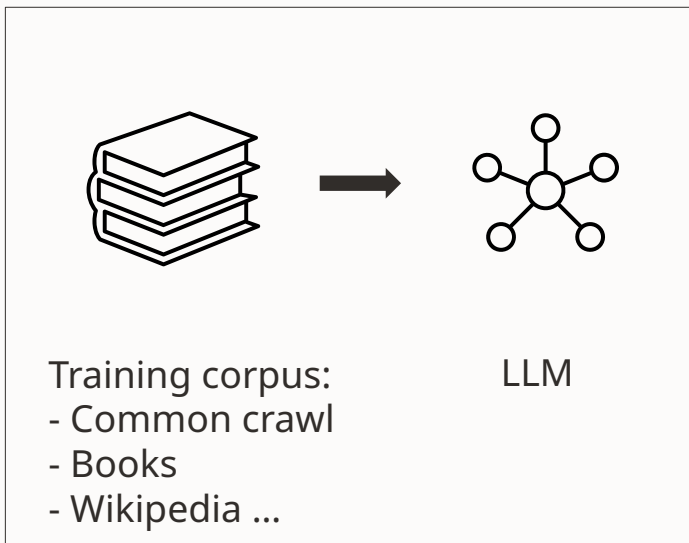
N-Dimensional Vectors inserted into Vector Database

# Retrieval Augmented Generation (RAG)



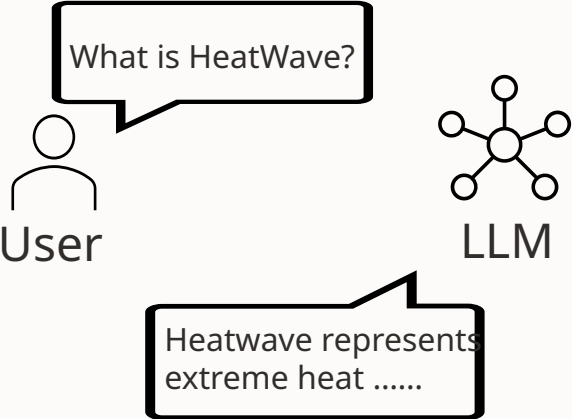
- Language models excel at processing information
- They lack the ability to interact with the outside world

- Empower LLM to interact with external data

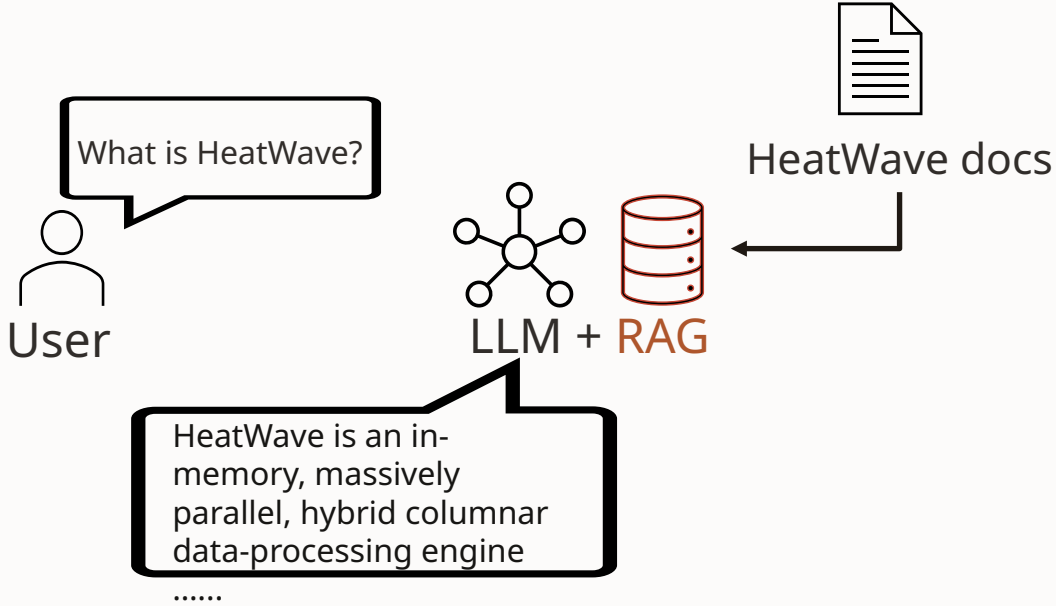




# Example of Using RAG

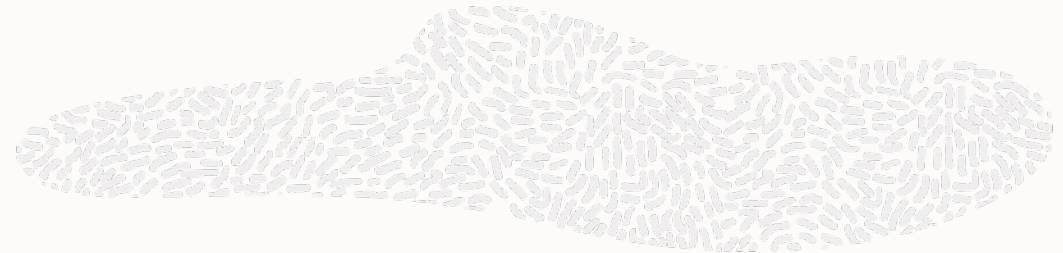


No domain-specific knowledge

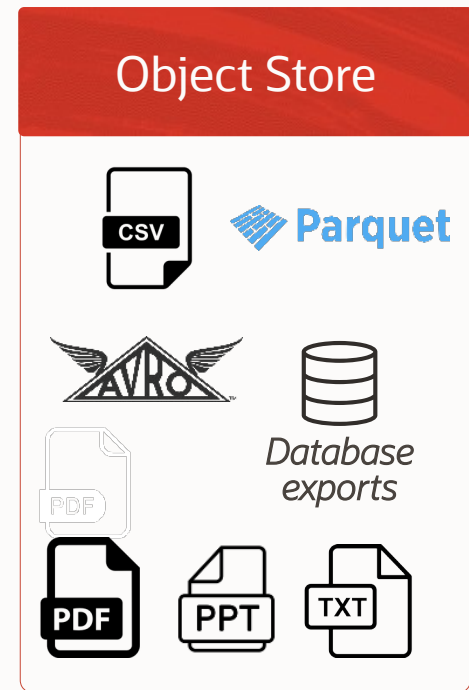
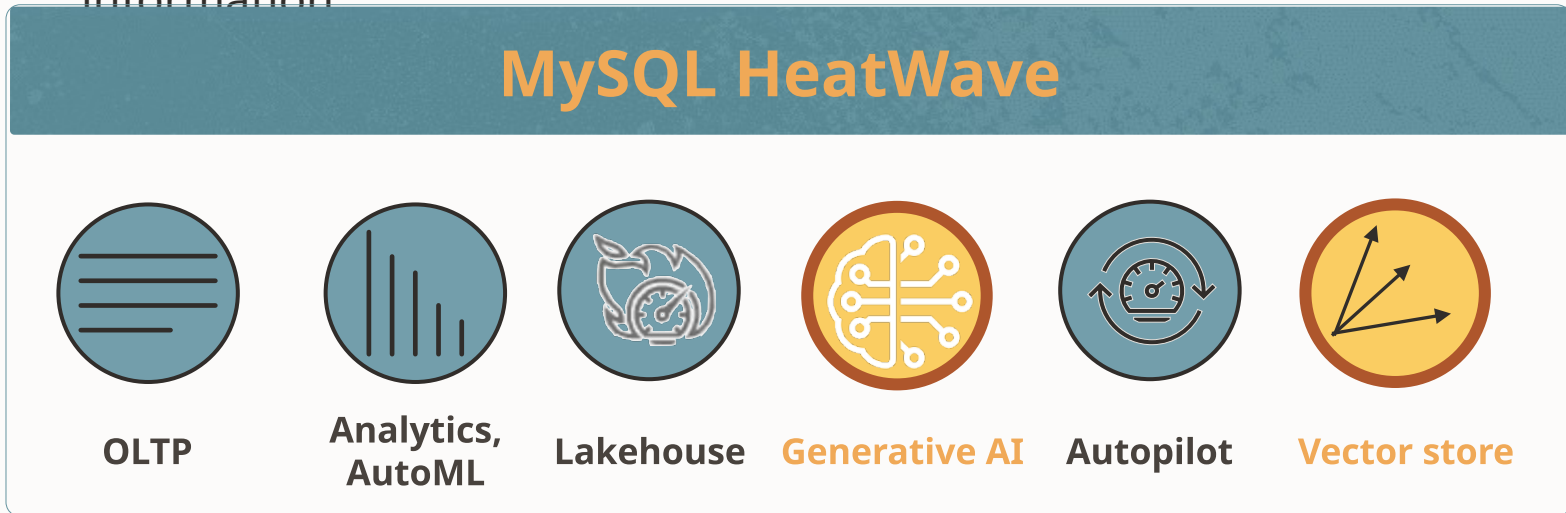


# HeatWave GenAI Ecosystem

# Generative AI with MySQL HeatWave



- A single system that takes advantage of Machine Learning and Generative AI
- Ask questions in natural language and get back answers
- Query user's documents stored in the object store and get relevant information



# HeatWave offers customers a choice to run LLMs

Select from a variety of pre-trained LLMs for different use cases

## In-HeatWave LLMs

- Native execution within the HeatWave database – data does not leave HeatWave
- Good tradeoff between price and generation quality by using smaller LLMs like Llama3-8B and Mistral-7B

## OCI Generative AI Service LLMs

- Excellent output quality by using large foundation models like Cohere-command-r and Llama3.1-70B
- Comparatively higher costs since they run on GPUs



# New Vector Datatype in HeatWave & MySQL

VECTOR is a first-class citizen now



Vector as first-class data type	<pre>mysql&gt; CREATE TABLE wikipedia (         title VARCHAR(1024),         page_data TEXT,         page_url TEXT,         page_embedding VECTOR(1024));</pre>
MySQL query syntax	<pre>mysql&gt; SELECT page_url,         DISTANCE(page_embedding,         @query_embedding, "COSINE")         as distance         FROM wikipedia         ORDER by distance DESC LIMIT 10;</pre>

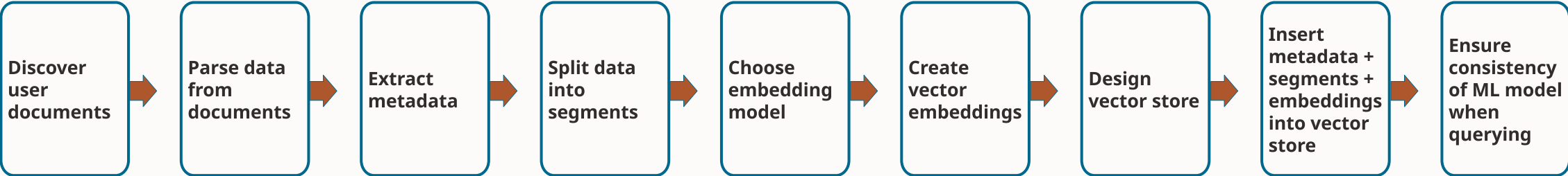
New distance function for similarity search

- EUCLIDEAN (L2)
- COSINE
- DOT

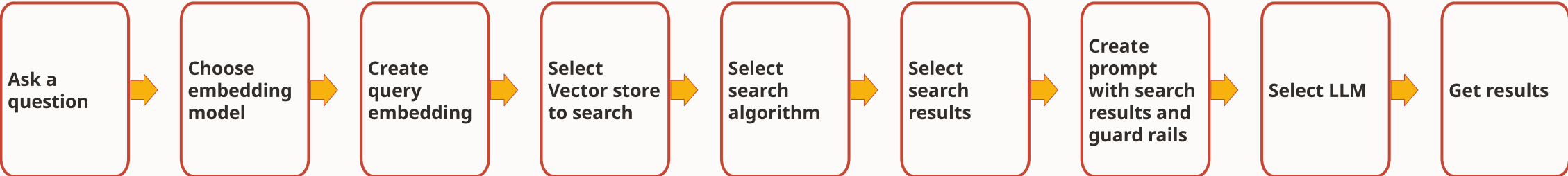
# Vector Store Pipeline



## Part 1 | Create a vector store

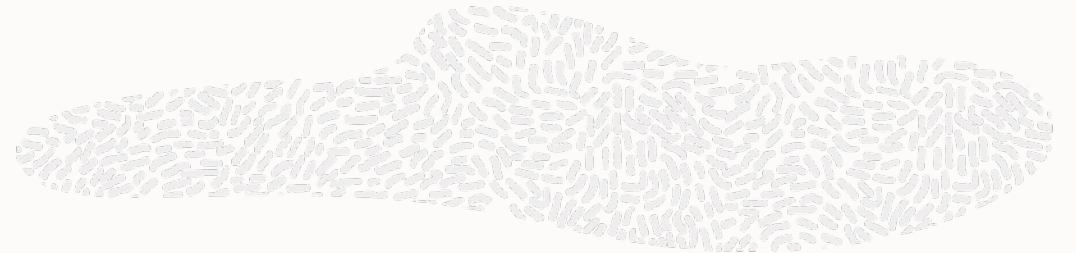


## Part 2 | Use vector store with LLMs



# HeatWave Vector Store

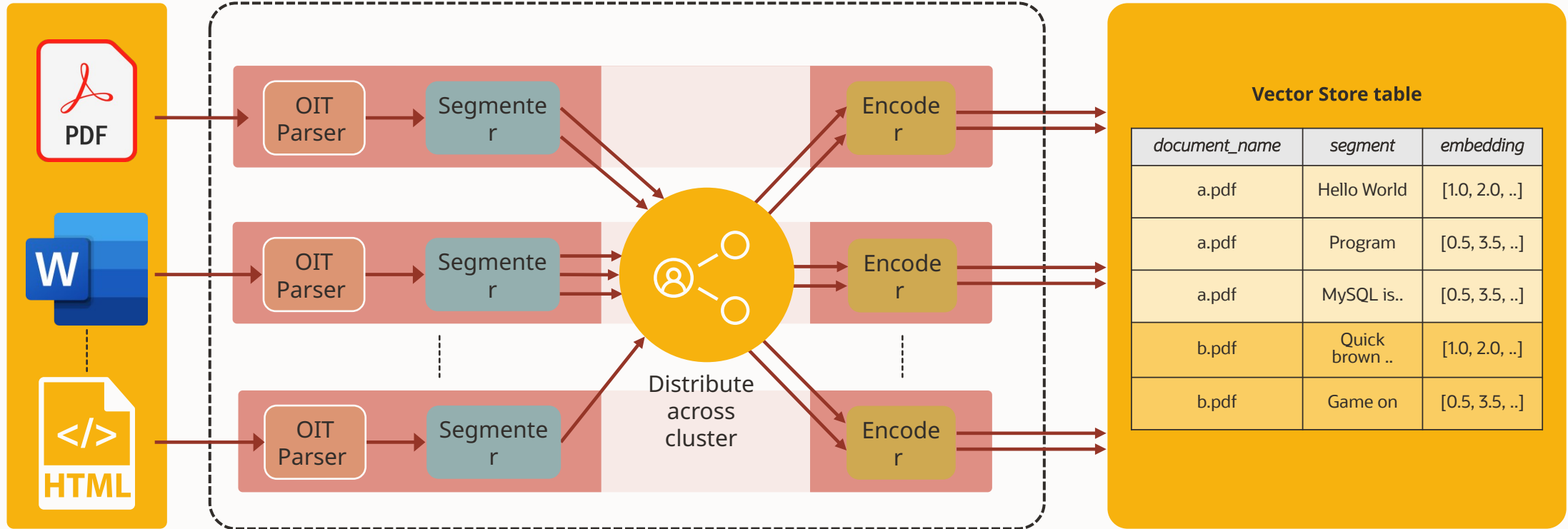
All system resources are optimized by HeatWave



Customer Bucket

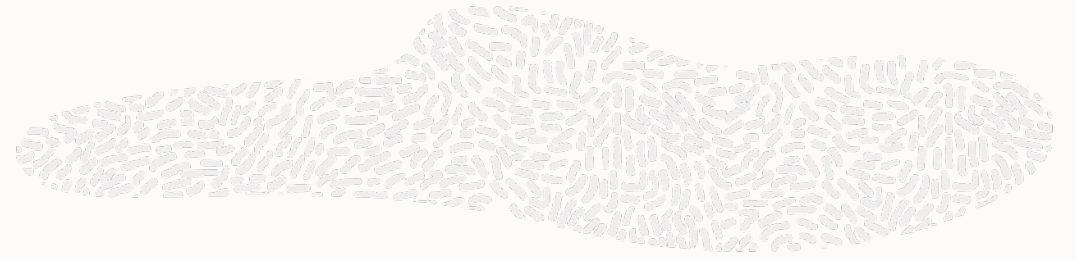
HeatWave Cluster

HeatWave Storage



Faster than generating vector store at the application layer

# High Level API



## Part 1 Create a vector store

```
SQL> CALL sys.vector_store_load("oci://bucket@namespace/path/",  
@optional_params)
```

## Part 2 Query the vector store

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

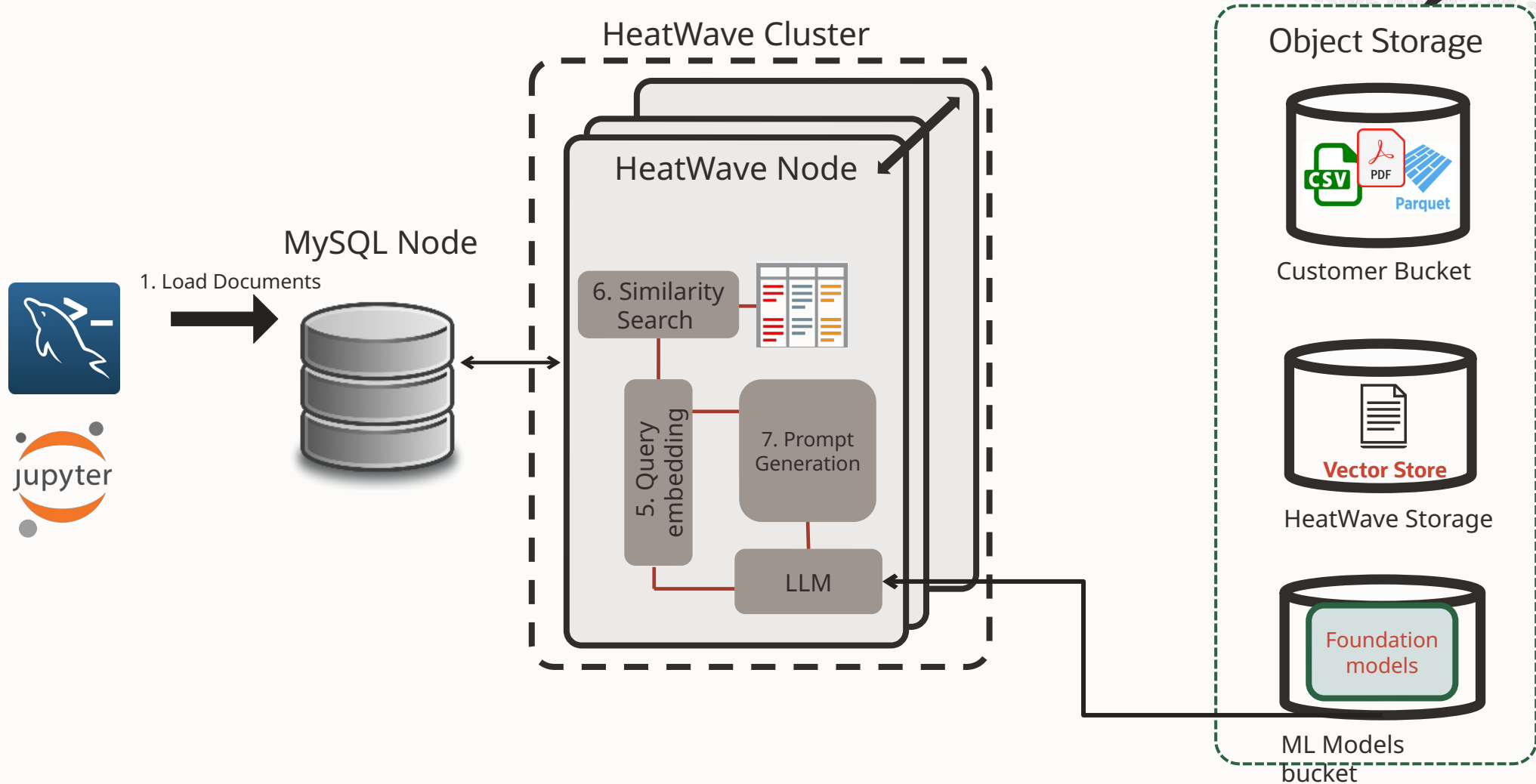
**Simpler and faster**

**No additional cost**



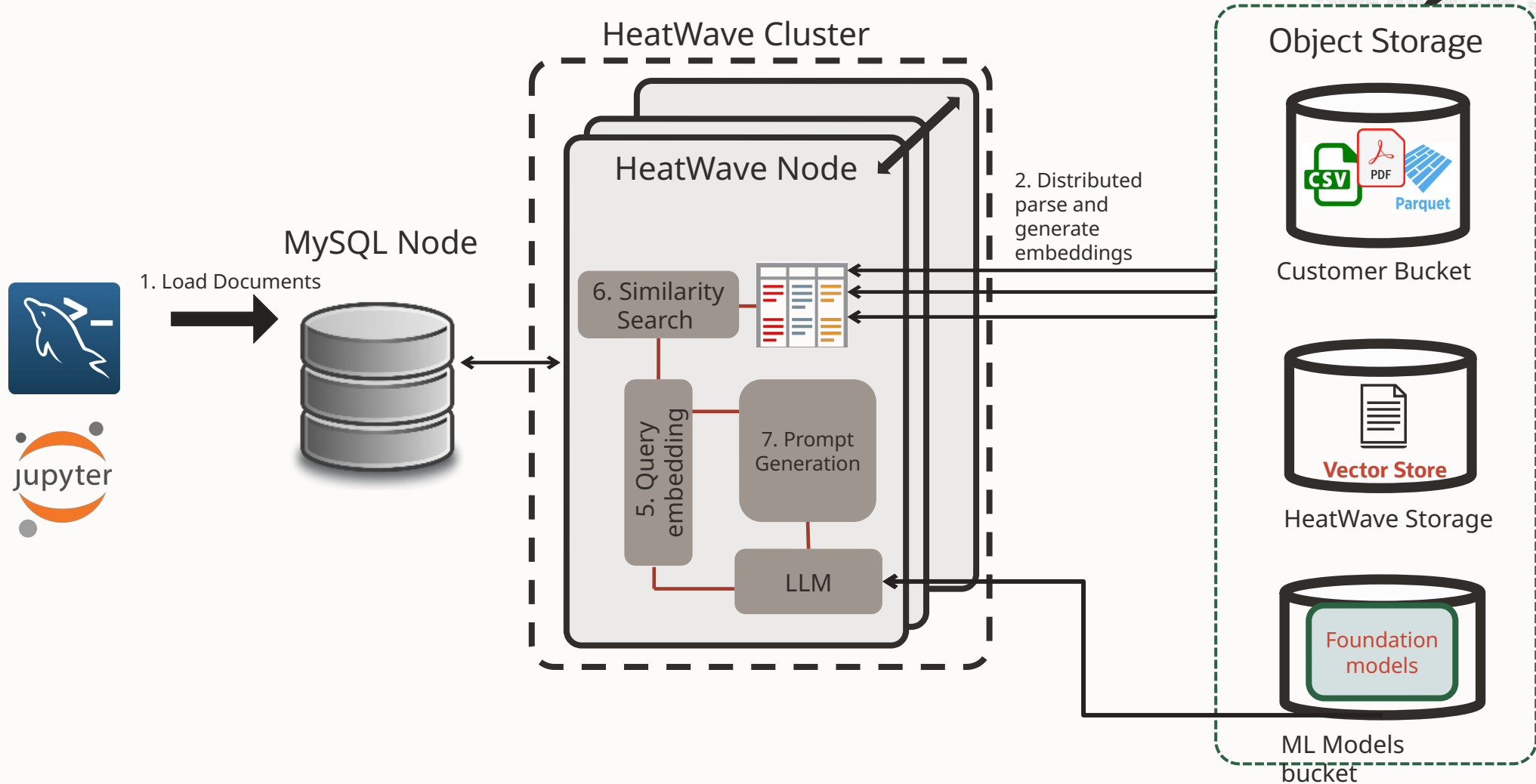
# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



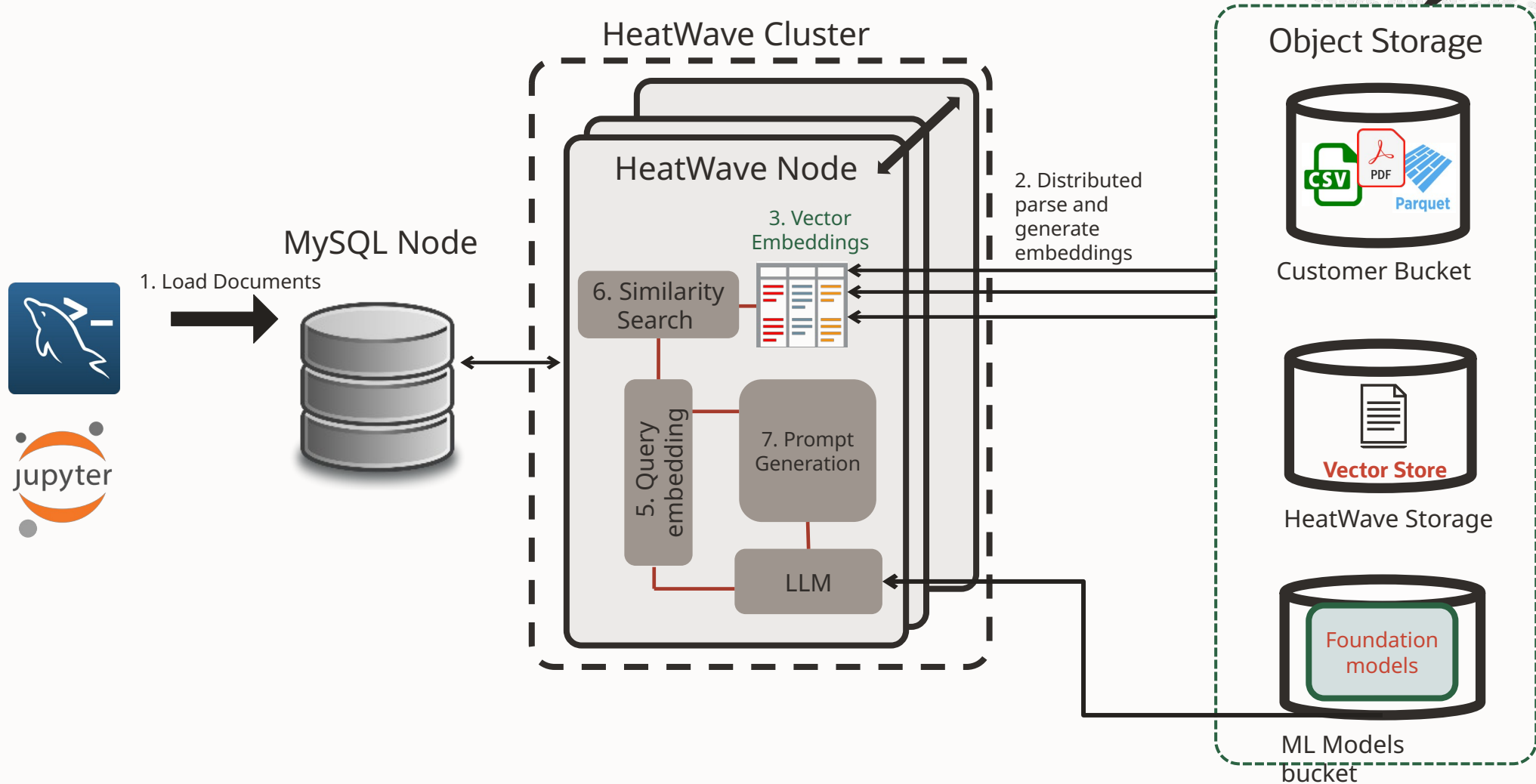
# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



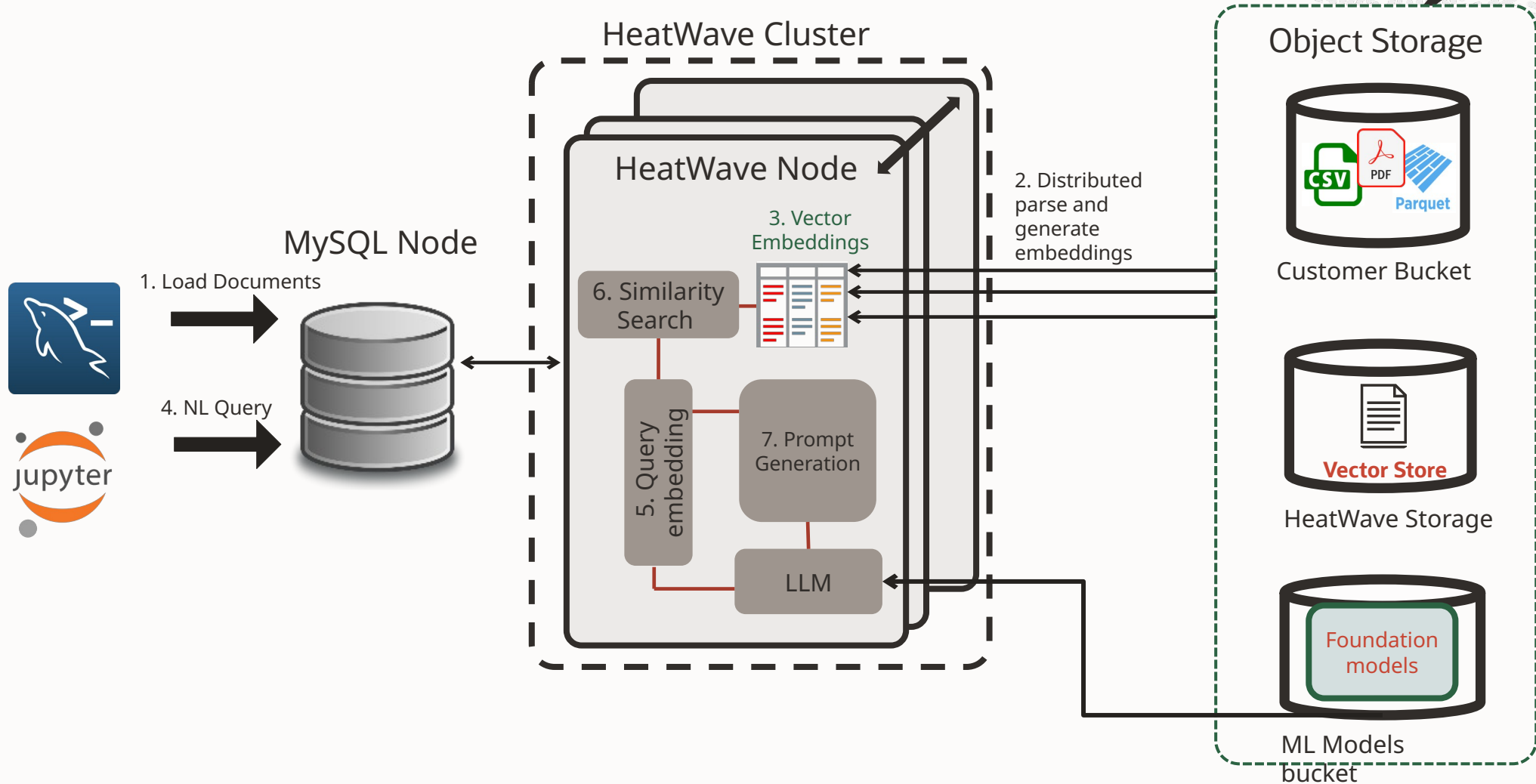
# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



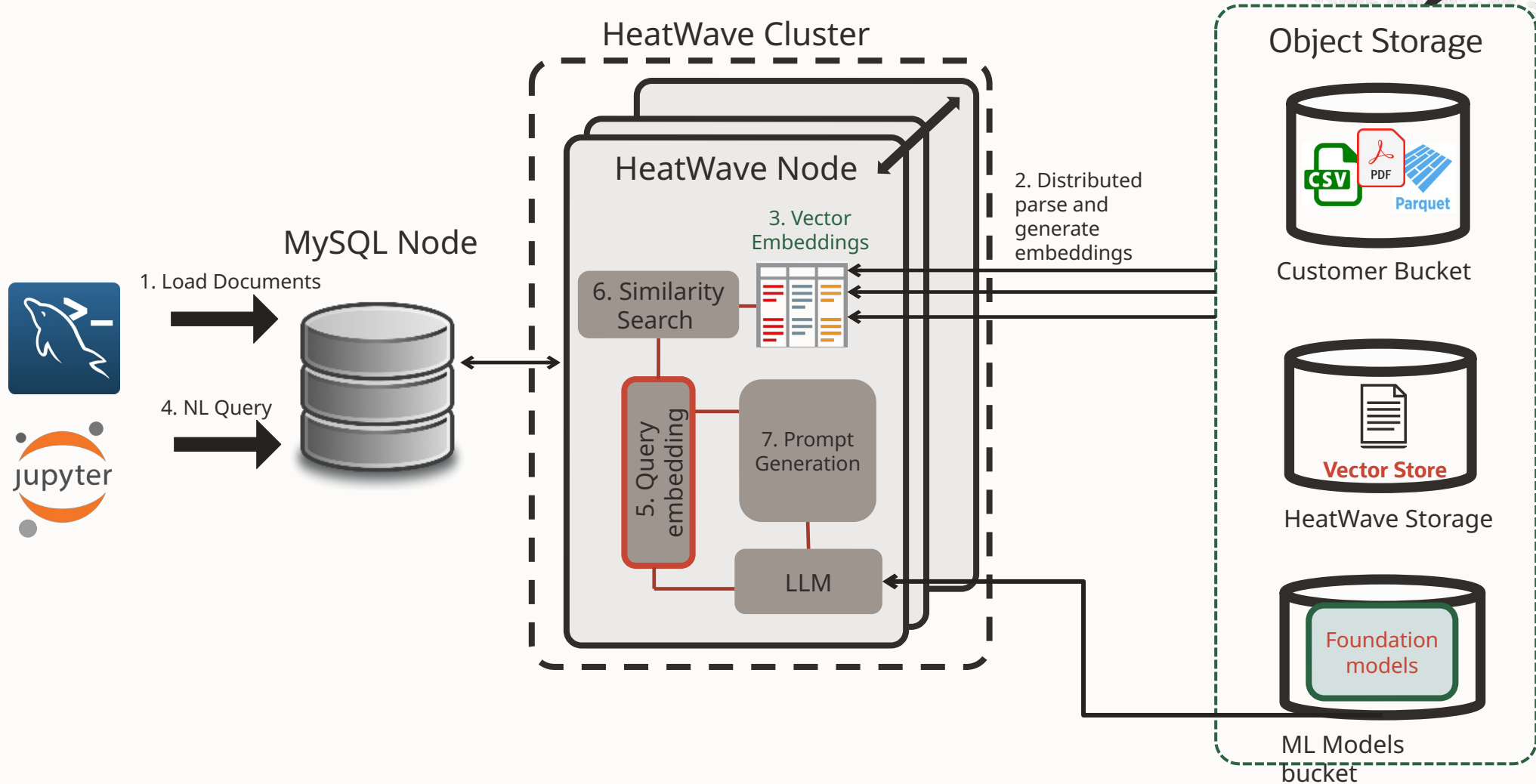
# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



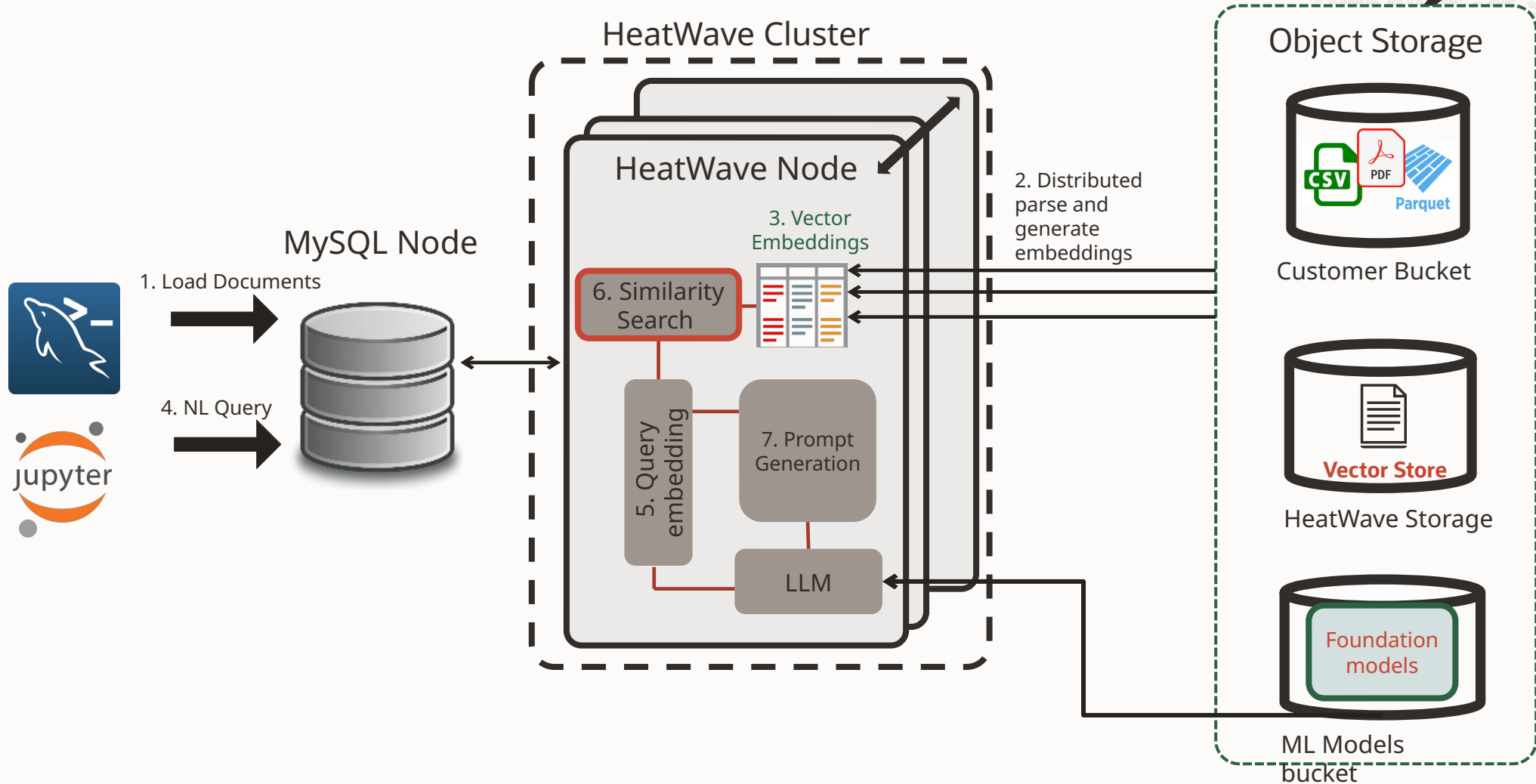
# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



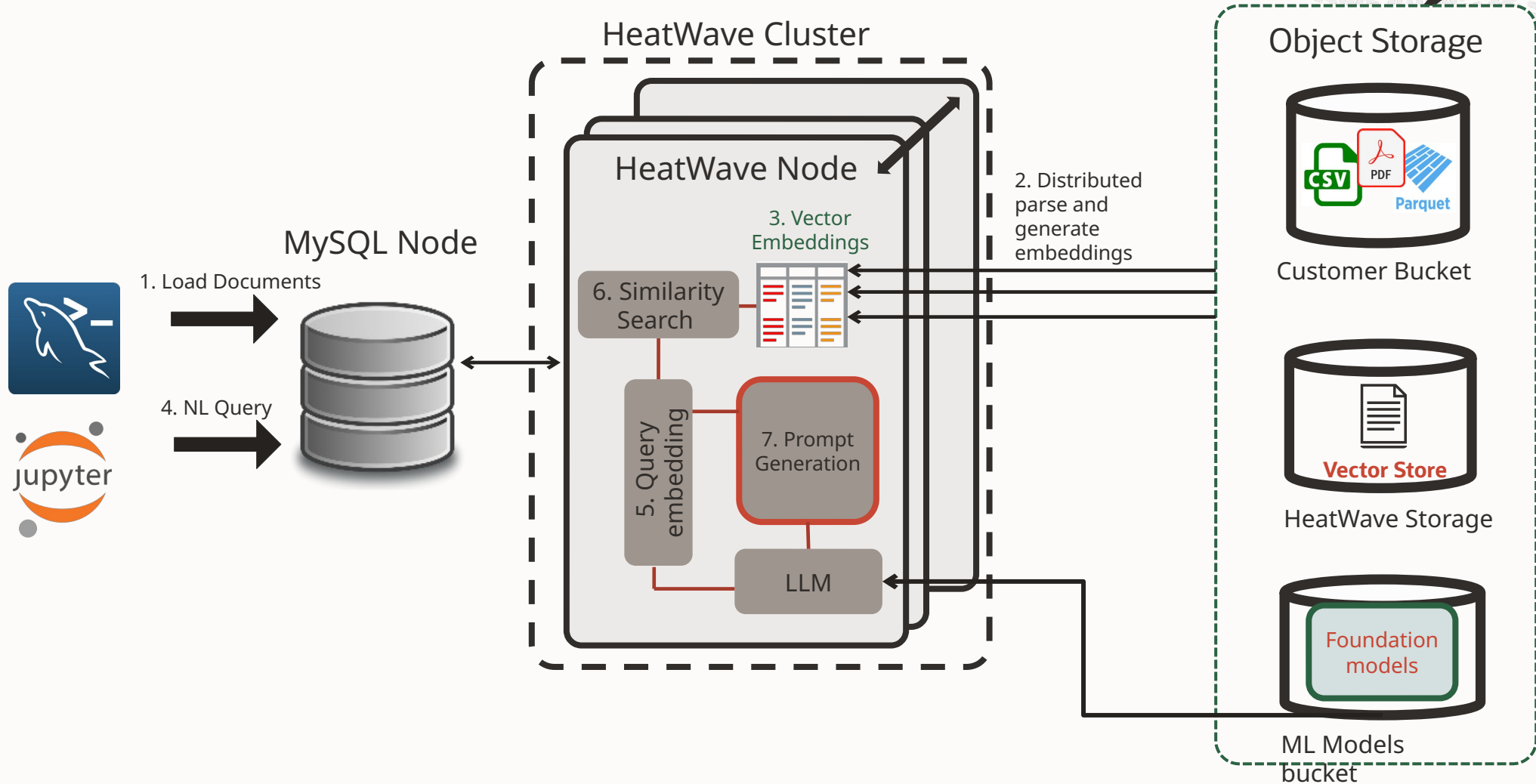
# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



# Generative AI in MySQL HeatWave

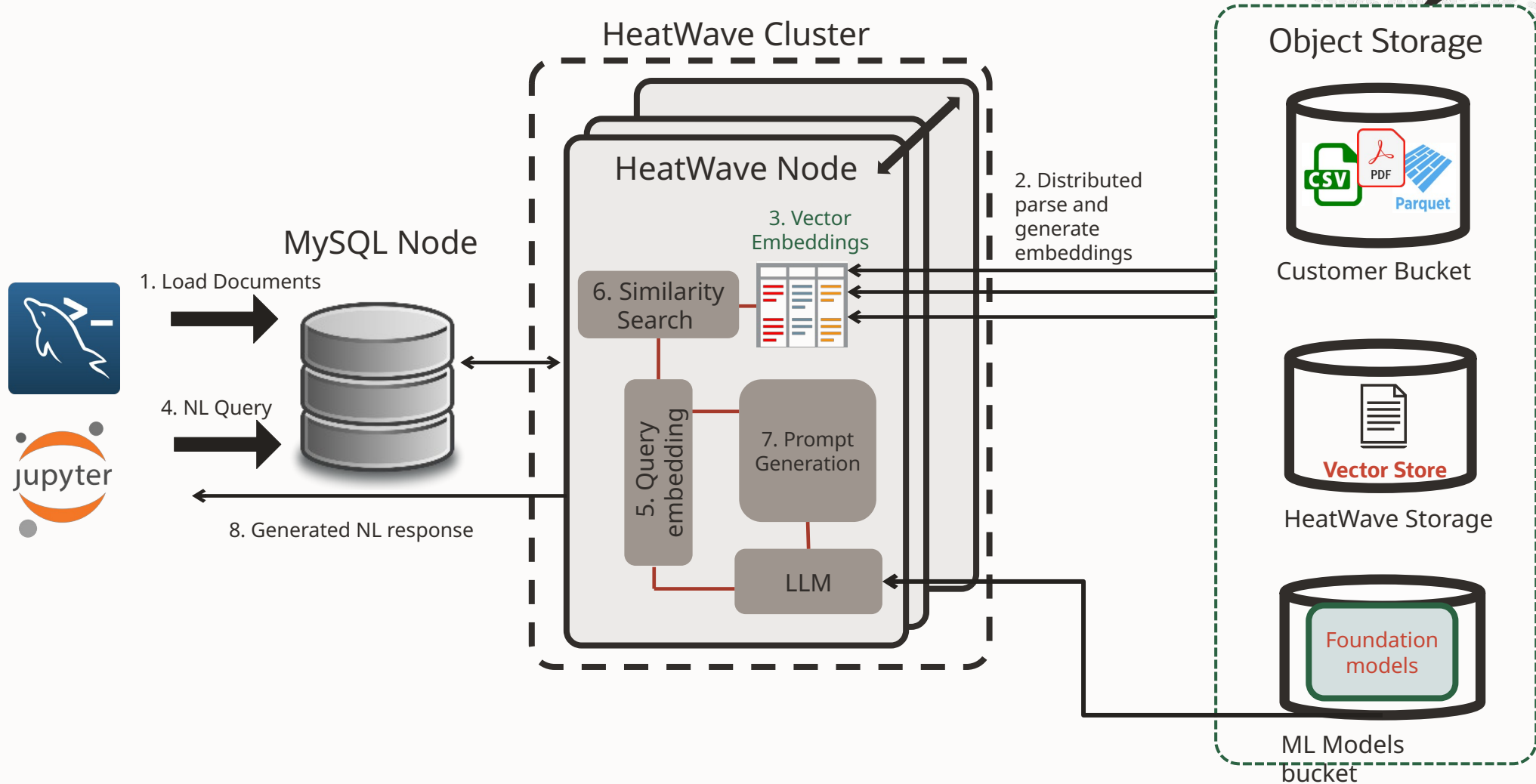
Combines LLMs, Lakehouse Ingest and Vector Store





# Generative AI in MySQL HeatWave

Combines LLMs, Lakehouse Ingest and Vector Store



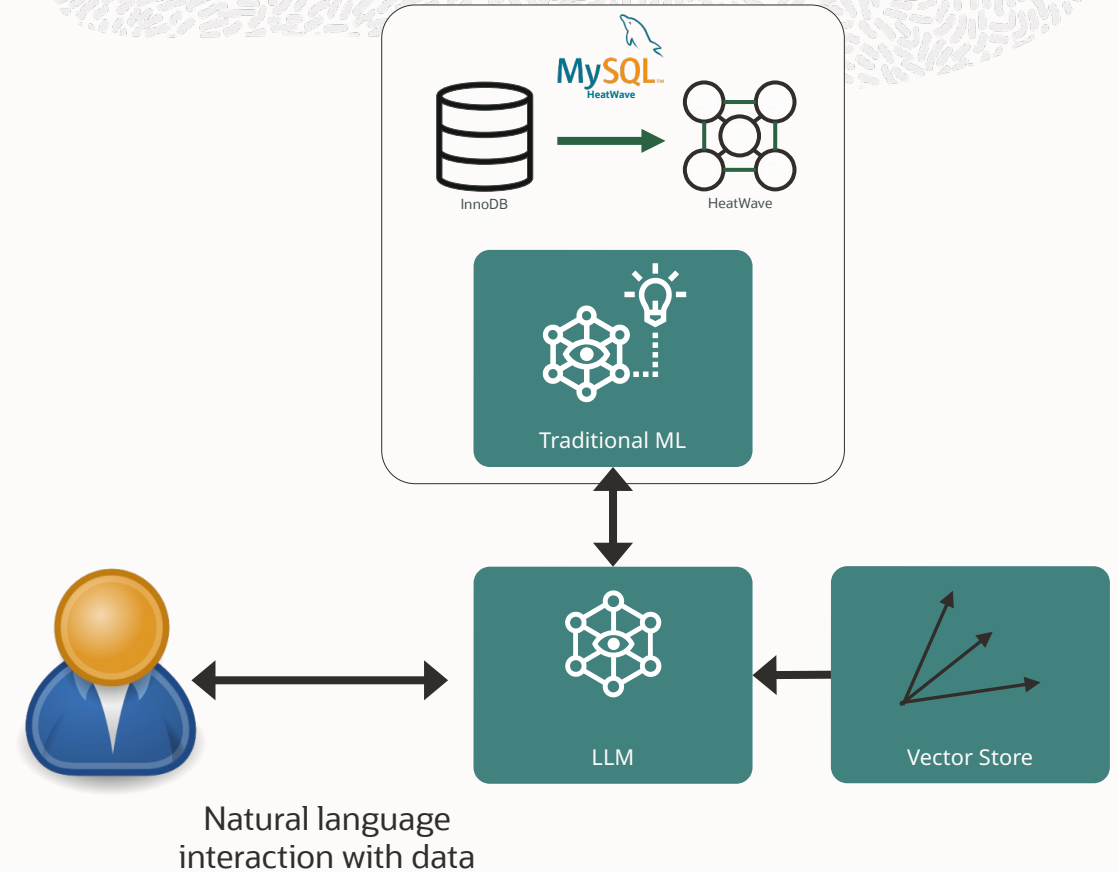


# Synergy of Generative AI and AutoML in HeatWave

## A differentiator in HeatWave

### Advantages:

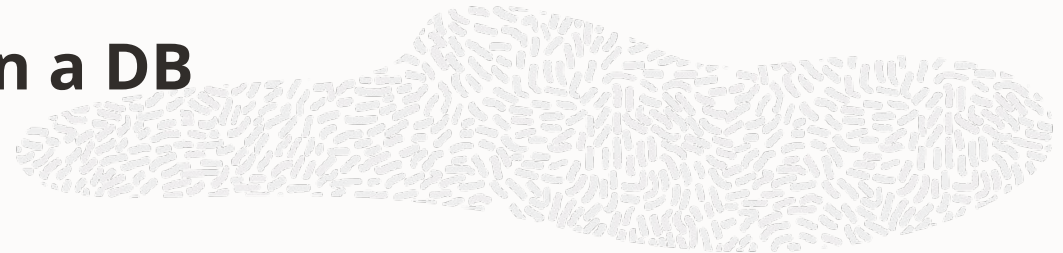
- Use NL to interact with data coming from AutoML
- More accurate LLM results by filtering irrelevant data
- Faster LLM inference due to smaller search space



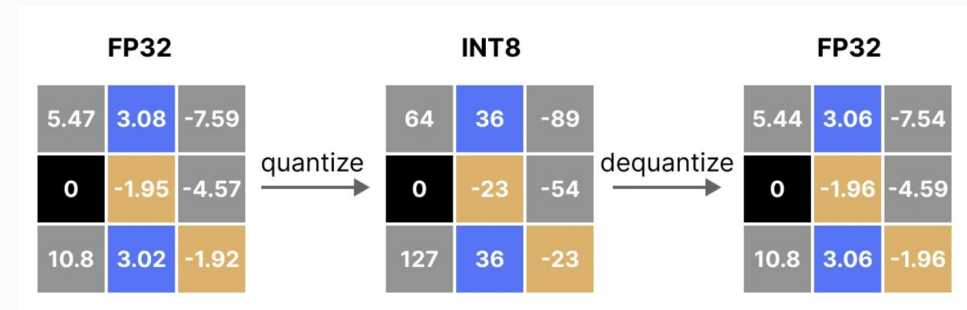
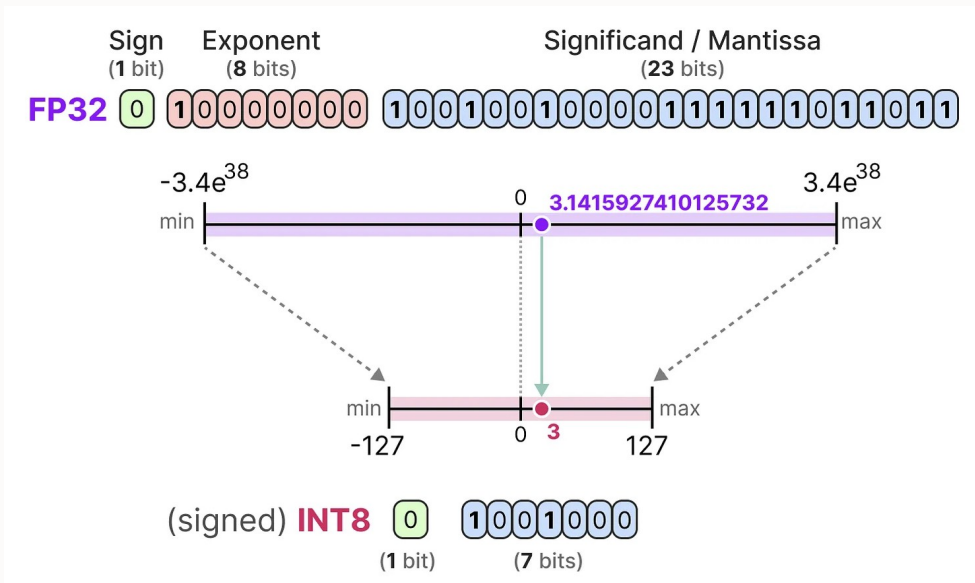
The background of the slide is a repeating pattern of overlapping circles. Each circle is filled with numerous concentric, slightly irregular lines that create a textured, wood-grain-like effect. The circles are arranged in a staggered, brick-like pattern, with some overlapping others. The overall color palette is a range of light greys and off-whites.

# Challenges

# Quantization: Running LLM Inference in a DB



- Reduce the memory size of LLMs
- Maintain an acceptable level of performance and accuracy
- Converting model parameters from high precision to lower-precision data types
- Computations are generally faster when using fewer bits

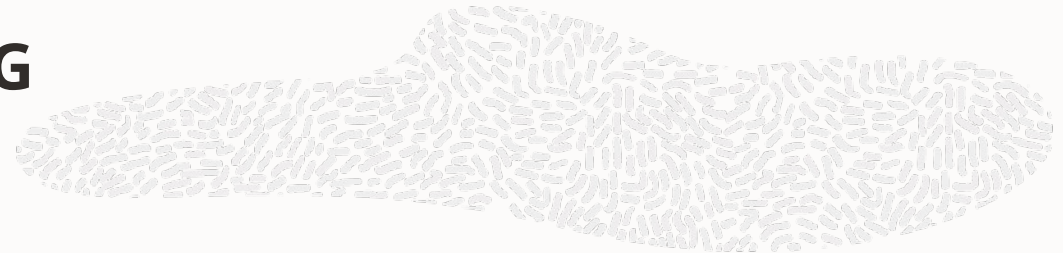


LLM with 8b params

FP16	INT8	INT4
16 GB	8 GB	4 GB

Source: <https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization>

# Precise and Complete Segments for RAG



*“How does indexing improve database performance, and what are the trade-offs of using indexes?”*

n\_citations=3

id	segment	Cosine distance
1	Indexes improve query speed by...	0.36
2	For example, a B-tree index allows binary searches...	0.38
3	Common types of indexes include...	0.35
4	Best practices for indexing include...	0.40
5	While indexes improve query performance, they introduce trade-offs...	0.40
6	For example, adding an index...	0.42

...

Indexes improve query speed by

...

For example, a B-tree index allows binary searches... Common types of indexes include... Best practices for indexing include... While indexes improve query performance, they introduce trade-offs... For example, adding an index...



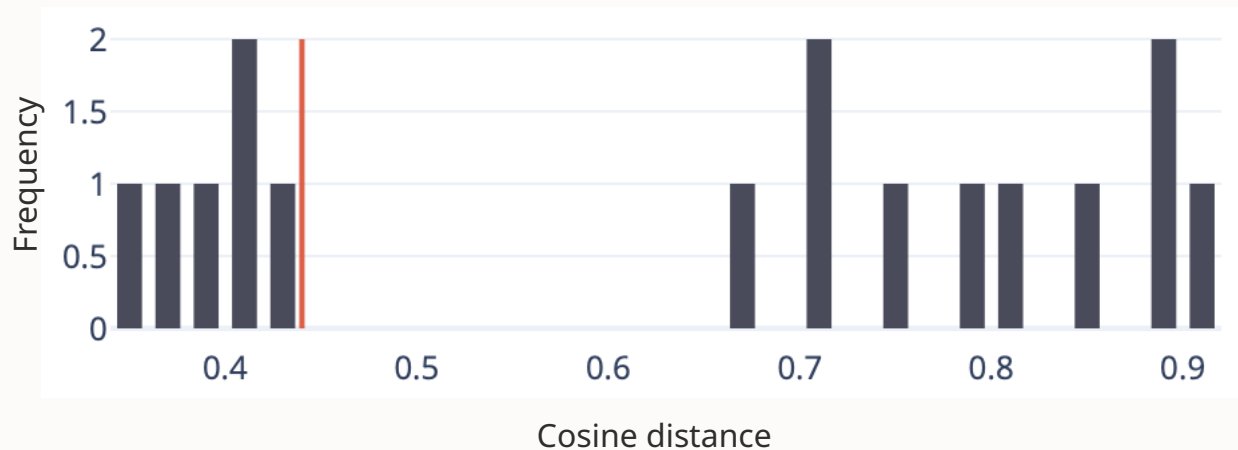


# Additional Retrieval Options



- max\_distance
- segment\_overlap
- **percentage\_distance**

n\_citations=15, *percentage\_distance=20*

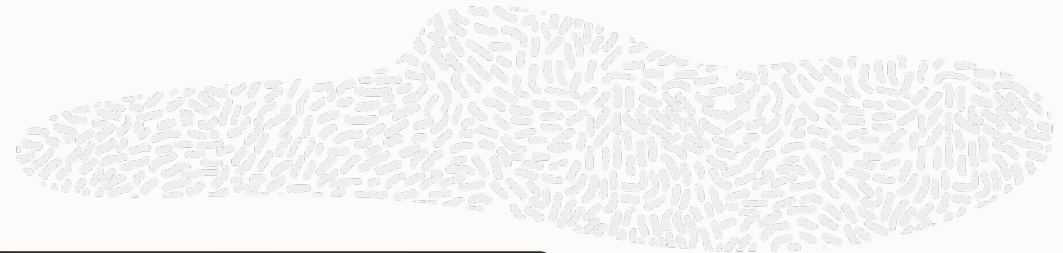


id	segment	Cosine distance
1	Indexes improve query speed by...	0.36
2	For example, a B-tree index allows binary searches...	0.38
3	Common types of indexes include...	0.35
4	Best practices for indexing include...	0.40
5	While indexes improve query performance, they introduce trade-offs...	0.40
6	For example, adding an index...	0.42

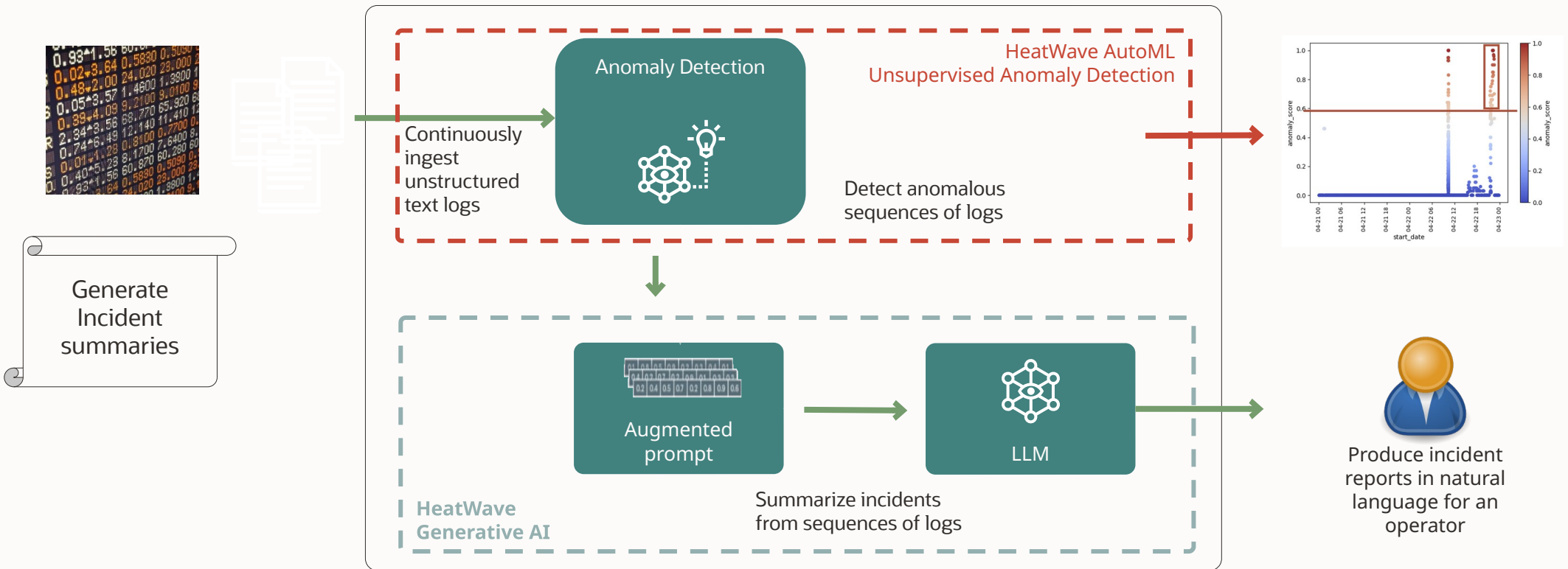
# Example Applications

# Example 1: Report Generation

## Content generation/summary



### Report Generation: HW AutoML + Summarization of anomalous logs





# Example 1: Report Generation

## Incident Report



```
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777883620,"msg":["8145:140643603412800][fit_transform:140][1.939] fit_transform: Remove non-categorical columns with > 20% missing values. Dropping 0 columns"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777883805,"msg":["8145:140643603412800][fit_transform:147][1.939] fit_transform: Remove constant columns, Dropping 0 columns"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777964868,"msg":["8145:140643603412800][fit_transform:227][173.397] fit_transform: After feature engineering and transformations, Updated shape : (157776, 72976)"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777964868,"msg":["8145:140643603412800][fit_transform:231][173.397] fit_transform: col_types_: ['text', 'text']}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777964879,"msg":["8145:140643603412800][fit_transform:247][173.397] fit_transform: class counts: 0 157494\n1 282\nName: label, dtype: int64"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975619,"msg":["8145:140641670100736][run:272][249.292](249.29190063476562, 249.29190063476562)(current, max) GB for process 8145 is NOT within the threshold 248.94811630249023 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975619,"msg":["8145:140641670100736][run:279][249.292] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975725,"msg":["8145:140641670100736][run:272][250.038](250.03799438476562, 250.03799438476562)(current, max) GB for process 8145 is NOT within the threshold 248.19862747192383 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975725,"msg":["8145:140641670100736][run:279][250.038] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975830,"msg":["8145:140641670100736][run:272][250.782](250.78213500976562, 250.78213500976562)(current, max) GB for process 8145 is NOT within the threshold 247.45304489135742 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975830,"msg":["8145:140641670100736][run:279][250.782] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975936,"msg":["8145:140641670100736][run:272][251.528](251.52822875976562, 251.52822875976562)(current, max) GB for process 8145 is NOT within the threshold 246.70550918579102 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777975936,"msg":["8145:140641670100736][run:279][251.528] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976042,"msg":["8145:140641670100736][run:272][252.274](252.27432250976562, 252.27432250976562)(current, max) GB for process 8145 is NOT within the threshold 245.95797729492188 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976042,"msg":["8145:140641670100736][run:279][252.274] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976148,"msg":["8145:140641670100736][run:272][253.02](253.02041625976562, 253.02041625976562)(current, max) GB for process 8145 is NOT within the threshold 245.21044158935547 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976148,"msg":["8145:140641670100736][run:279][253.02] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976253,"msg":["8145:140641670100736][run:272][253.767](253.76651000976562, 253.76651000976562)(current, max) GB for process 8145 is NOT within the threshold 244.46290588378906 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976253,"msg":["8145:140641670100736][run:279][253.767] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976359,"msg":["8145:140641670100736][run:272][254.513](254.51260375976562, 254.51260375976562)(current, max) GB for process 8145 is NOT within the threshold 243.71537017822266 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976359,"msg":["8145:140641670100736][run:279][254.513] Sending SIGTERM to the main thread"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976465,"msg":["8145:140641670100736][run:272][255.261](255.26055013476562, 255.26055013476562)(current, max) GB for process 8145 is NOT within the threshold 242.96588134765625 GB memory usage"}
{"analyticsClusterid":"","component":"mi_driver","level":"INFO","ts":1688777976465,"msg":["8145:140641670100736][run:279][255.261] Sending SIGTERM to the main thread"}
```

Ask the question ('What is the main problem in the following collection of logs. Provide a two-sentence summary.')

The main problem in this collection of logs is that the memory usage of a particular process (with ID 8145) is consistently exceeding the defined memory threshold. This issue is repeatedly highlighted in the logs, and SIGTERM signals are sent to the main thread to terminate the process when it exceeds the memory threshold. Would you like me to help you with anything else regarding this collection of logs?"



# Example 2: MySQL Shell for VS Code

## HeatWave Lakehouse Navigator and HeatWave Chat

**HeatWave Lakehouse Navigator**

**On-Premise**  
Data to be analyzed (on your local storage)

**Object Storage**  
Data to be analyzed (on OCI Object Storage)

**Lakehouse**  
Data ready for analysis (held in memory)

Structured & Semi-structured Data (CSV, Avro, Parquet, PDF, etc.)

Data Warehouse, Mobile, Web, CRM, Sensors

Data Science, Reports, BI, Machine Learning

MySQL HeatWave Lakehouse

**Upload Data to Object Storage**  
Select this option to upload files from your local file system or local company network to the Object Storage.  
Upload Files >>

**Load Data into Lakehouse**  
If your files have been uploaded to the OCI Object Storage select this option to load the data into Lakehouse.  
Start Load >>

**Manage Lakehouse Tables**  
This option allows you to monitor running your data loading task and manage your data inside Lakehouse.  
Manage Lakehouse

**HeatWave Chat**

Execute \chat for natural language chat mode.  
Execute \sql to switch to SQL, \js to JS and \ts to TypeScript.  
Execute \help or \? for help;

chat>

**HeatWave Chat AI Profile Editor**

Schema Scope: All Database Schemas

History

Database Tables

Matched Documents

Model Options  
Model: Default  
Language: English

Advanced Model Options  
Review Generated Prompt  
Language Settings

Save Load Start New Chat

# Loading into Lakehouse

The screenshot displays the Lakehouse Navigator interface with the following components:

- Editor:** Lakehouse Navigator
- Navigation Tabs:** Overview, Upload to Object Storage, Load into Lakehouse (active), Lakehouse Tables
- Object Storage Browser:**
  - OCI Profile: heatwavedev
  - Object Storage Item: HeatwaveAutoML
  - Selected folder: formula1
  - Files in formula1:
    - Cloud-based infrastructure runs complex engineering simulations.html (60.06 KB, 05/16/24, 17:36)
    - Formula-One-Championships.html (1.04 KB, 05/16/24, 19:02)
    - How Oracle Red Bull Racing's chief engineer fuels racing strategy with data.html (87.69 KB, 05/16/24, 17:36)
    - Oracle Red Bull Racing achieves record-breaking F1 season with Oracle Cloud.html (89.11 KB, 05/16/24, 17:36)
    - Oracle Red Bull Racing energizes fans, increasing loyalty program membership more than 950%.html (94.66 KB, 05/16/24, 17:36)
    - Oracle Red Bull Racing increases simulations and sharpens decision-making on the track with Oracle Cloud Infrastructure.html (92.32 KB, 05/16/24, 17:36)

**New Loading Task Configuration:**

- Vector Store Table Name: formula1
- Table Description: Data from Bucket genai-shell-test
- Target Database Schema: car\_racing
- Formats: HTML (HyperText Markup Language Files)
- URI: oci://genai-shell-test/formula1
- Start Loading Task button

# Manage Lakehouse Tables

The screenshot shows the Lakehouse Navigator interface with the following components:

- Navigation Bar:** Includes 'Overview', 'Upload to Object Storage', 'Load into Lakehouse', and 'Lakehouse Tables' tabs.
- HeatWave Memory:** Displays '36.91 GB free'.
- Database Schemas:** A list on the left includes ML\_SCHEMA\_mike, ML\_SCHEMA\_milos, ML\_SCHEMA\_root, car\_racing, dsm, e2e\_tests, mysqlsh, test, vector\_store, and vs.
- Lakehouse Table List:**

Database Schemas	Lakehouse Table	Loaded	Rows	Size	Date	Comment
ML_SCHEMA_mike	formula1	40%	-	-	2024-06-02 10:18	Data from Bucket genai-shell-test
ML_SCHEMA_milos	legal_documents	Yes	93	206.42 KB	2024-06-02 08:37	Data from Bucket genai-shell-test
ML_SCHEMA_root	performance_data	Yes	37	14.40 KB	2024-06-02 08:35	Data from Bucket genai-shell-test
- Current Task List:**

Task	Id	Status	Start Time	End Time	Message
Loading formula1	14	RUNNING	2024-06-02 10:18	~ 2024-06-02 10:18 (23s)	Loading in progress...
Loading mysql_ref_man	13	COMPLETED	2024-06-02 09:45	2024-06-02 09:57	Task completed.
Loading formula1	12	COMPLETED	2024-06-02 09:17	2024-06-02 09:17	Task completed.
Loading formula1	11	COMPLETED	2024-06-02 09:06	2024-06-02 09:06	Task completed.

# HeatWave Chat

DB Notebook (GenAI Dev Instance) ×

Editor: DB Notebook

sql> \about

## HeatWave Chat

Querying your HeatWave Data using SQL, natural language or TypeScript/JavaScript.

Execute \chat for natural language chat mode.  
Execute \sql to switch to SQL, \js to JS and \ts to TypeScript.  
Execute \help or \? for help;

chat> Which team won the Formula 1 championship in 2023?

The Formula 1 Constructors' Championship 2023 was won by Red Bull Racing.

Would you like to know more about any specific year?

chat> Did Red Bull Racing use AI and ML to achieve that?

Yes, according to Christian Horner, Team Principal and CEO of Oracle Red Bull Racing, the team uses AI technologies to help them make strategic decisions and deal with the pressure of high-stakes situations. Specifically, he mentions how AI is being used to provide young drivers with valuable feedback on their performance that Red Bull Racing is actively using AI and machine learning to analyze video and digital data.

It is not explicitly stated whether these technologies were used to gain an edge in the Formula 1 World Championship in 2023; however, Horner's comments suggest that the team's use of technology, including AI, has contributed to their success and consistency.

Is there anything specific you would like to know regarding the use of AI and machine learning in sports?

chat> Which role does Oracle Cloud play?

Oracle Cloud Infrastructure plays a vital role in the AI and ML capabilities of Red Bull Racing. The data collected from the car's sensors during races (about 400GB per race) is stored and processed using Oracle Cloud. This enables the team to use data science and analytics to gain insights and make informed decisions.

With Oracle Cloud's high-performance infrastructure, Red Bull Racing can accelerate its data analysis. This enables the team to extract valuable information from the data in real time, predicting race outcomes, strategies, and making adjustments to car performance and handling.

Overall, Oracle Cloud is a powerful tool that helps Red Bull Racing achieve maximum performance on an advanced analytics and data-driven decisions.

Would you like to know more about Oracle Cloud's capabilities in the world of Formula One?

chat>

## HeatWave Chat AI Profile Editor

Schema Scope: car\_racing

History:

User	Chatbot
Which team won the Formula 1 championship in 2023?	The Formula 1 Constructors' ...
Did Red Bull Racing use AI and ML to achieve that?	Yes, according to Christian Ho...
Which role does Oracle Cloud play?	Oracle Cloud Infrastructure pla...

Database Tables:

- car\_racing.formula1
- car\_racing.legal\_documents
- car\_racing.performance\_data

Lock table list:

Matched Documents:

Title	Segment
Oracle Red Bull R...	Oracle Cloud played a critical role in helping our team deliver the...
Oracle Red Bull R...	/customers/red-bull-racing/ <!DOCTYPE html> <p>On the way t...
How Oracle Red B...	Increasingly, F1 teams are looking to give comprehensive data a...

Model Options:

Model: Cohere

> Advanced Model Options

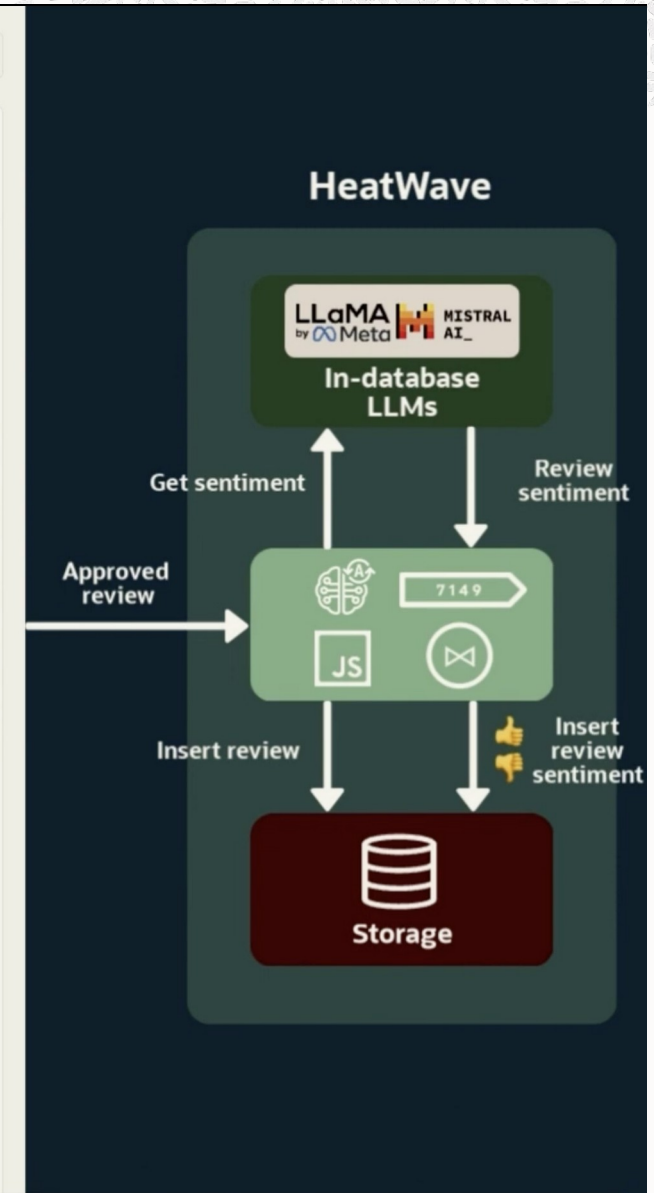
> Review Generated Prompt

> Language Settings

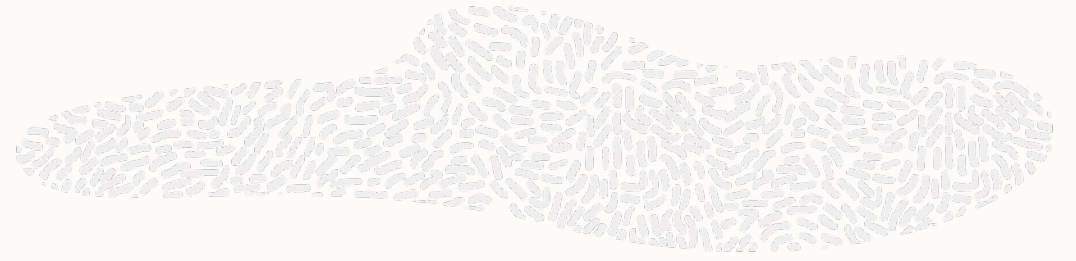


# Example 3: E-Commerce Platform

The screenshot shows a product page for a black T-shirt. At the top, there are navigation links for 'Accessories', 'Clothing', and 'Kitchenware', along with a search bar and language selection icons. The product title is 'T-Shirt' with a description: 'A high-quality T-shirt manufactured from breathable cotton is perfect for casual outings. It comes in a variety of colors and sizes.' The price is listed as '24.99 \$ USD' and there is an 'Add To Cart' button. Below the product image, there is a 'Customer Reviews' section with a 'Tell us what you think' form. The form includes a star rating (5 stars), a text input field for the review, and a 'Submit review' button. A feedback message says 'Thank you for your feedback!'. The 'Customer Reviews' section also features a 'Customers say' summary with a bar chart showing 'Positive reviews: 60%' and 'Negative reviews: 40%'. Below the chart, there are two columns of review snippets. The first column shows a thumbs-up icon and a positive review snippet. The second column shows a thumbs-down icon and a negative review snippet. At the bottom, there is a 'Top reviews' section with two review snippets, each with a star rating and a text preview.



# Sentiment Analysis JS Program



```
CREATE PROCEDURE SENTIMENT_ANALYSIS(  
  IN review TEXT,  
  IN review_id INT  
) LANGUAGE JAVASCRIPT AS $$  
  
  let prompt = `Classify the review into NEGATIVE or POSITIVE \n${review}.  
  Please provide a single word to describe the sentiment: "POSITIVE" or "NEGATIVE". \nSentiment:`;  
  
  let sentiment = ml.generate(prompt);  
  let processed_sentiment = sentiment.toUpperCase().search("POSITIVE") ? "POSITIVE" : "NEGATIVE";  
  
  let sql = session.prepare(`UPDATE reviews SET sentiment = ? WHERE id = ?`);  
  sql.bind(processed_sentiment, review_id).execute();  
$$;
```

# Review Summary

## Customers say Generated by HeatWave GenAI

Positive reviews: 67%

Negative reviews: 33%



The T-shirt is made from **high-quality cotton material** and suits various occasions like **casual outings**. One review mentions how it **is comfortable to wear** and another notes that it is made from sustainable material, which is good for the environment. However, the lack of bleach instructions and difficult-to-follow washing instructions are concerns for some customers. Overall, the reviews give the T-shirt an average rating of 3 out of 5 stars.



Two customers review a t-shirt unfavourably. The first customer **complains that the material is not durable enough** and that the **washing instructions are not clear**. The second customer states that the **t-shirt is not breathable** and that **the size was too small**. Both customers advise others against buying the product.

## Tell us what you think

Vince Roberts



Write a review...

Submit review

## Top reviews

Vince Roberts



The T-Shirt is fantastic as it is not only comfortable but it is also made of sustainable material. The organic cotton feels like normal cotton while being environment-friendly. However, the washing instruction is difficult to follow.

David Johnson



I highly recommend this T-Shirt. The softness of the cotton material and the **comfortable fit** make it perfect for everyday wear. The washing instructions are clear and easy to follow, ensuring that the T-Shirt retains its quality.

John Brown



The T-Shirt is made **from high-quality cotton** and is perfect for **casual outings**. However, the lack of bleach instructions may be a concern for some customers. Overall, I would rate it 3 out of 5 stars.

Fred Smith



This T-Shirt is not up to the mark. **The material is not durable enough** and the **washing instructions are not clear**. I would not recommend this product to anyone looking for a long-lasting and easy-to-care-for T-Shirt.

Chris Davis

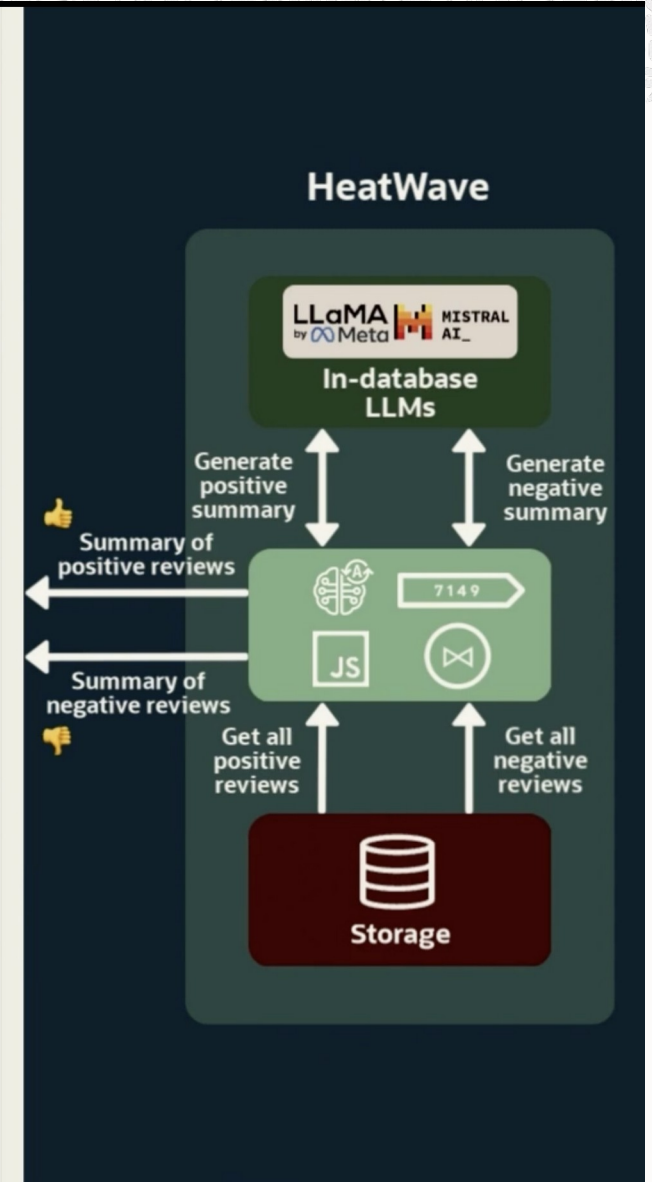


This T-Shirt is not up to my expectations. **The material is not breathable enough** and the **size is too small**. I would not recommend this product to others.

John Allen



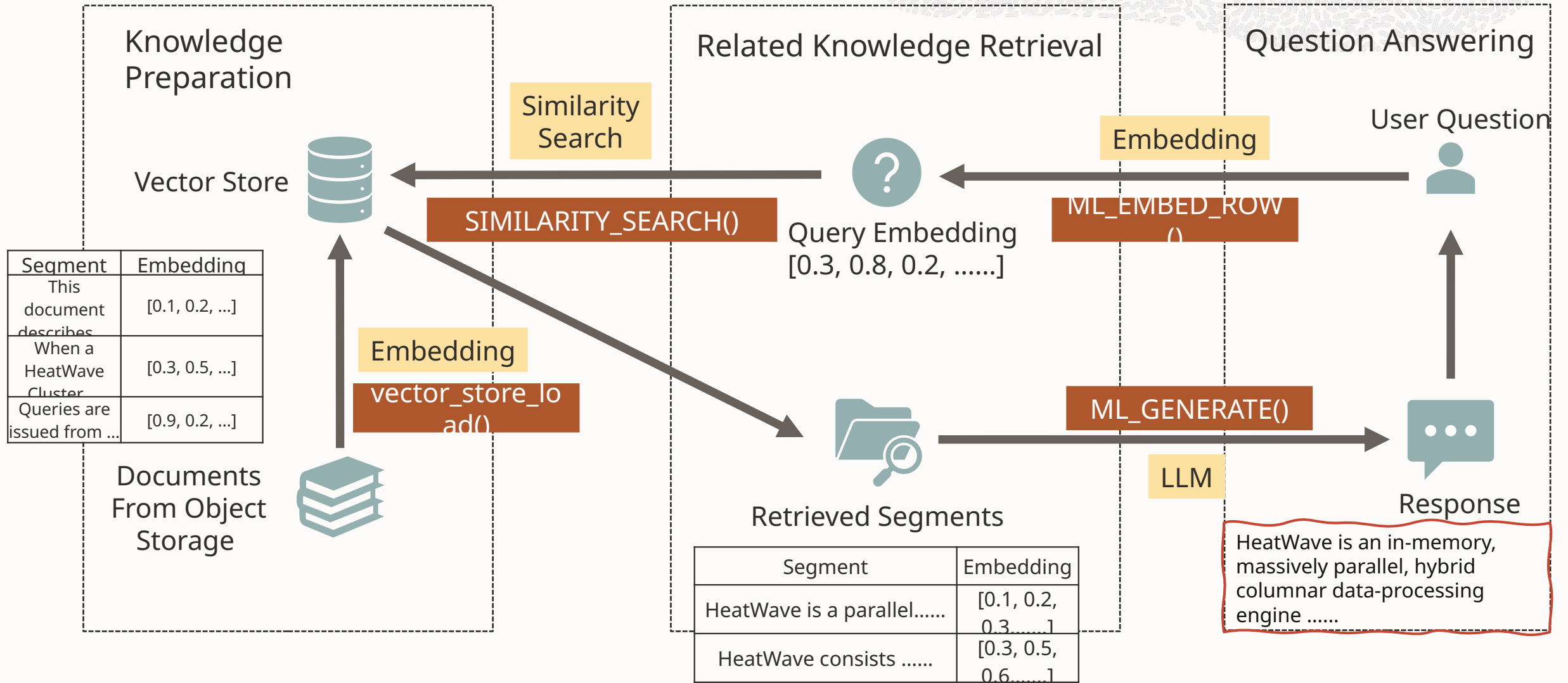
1/31/25



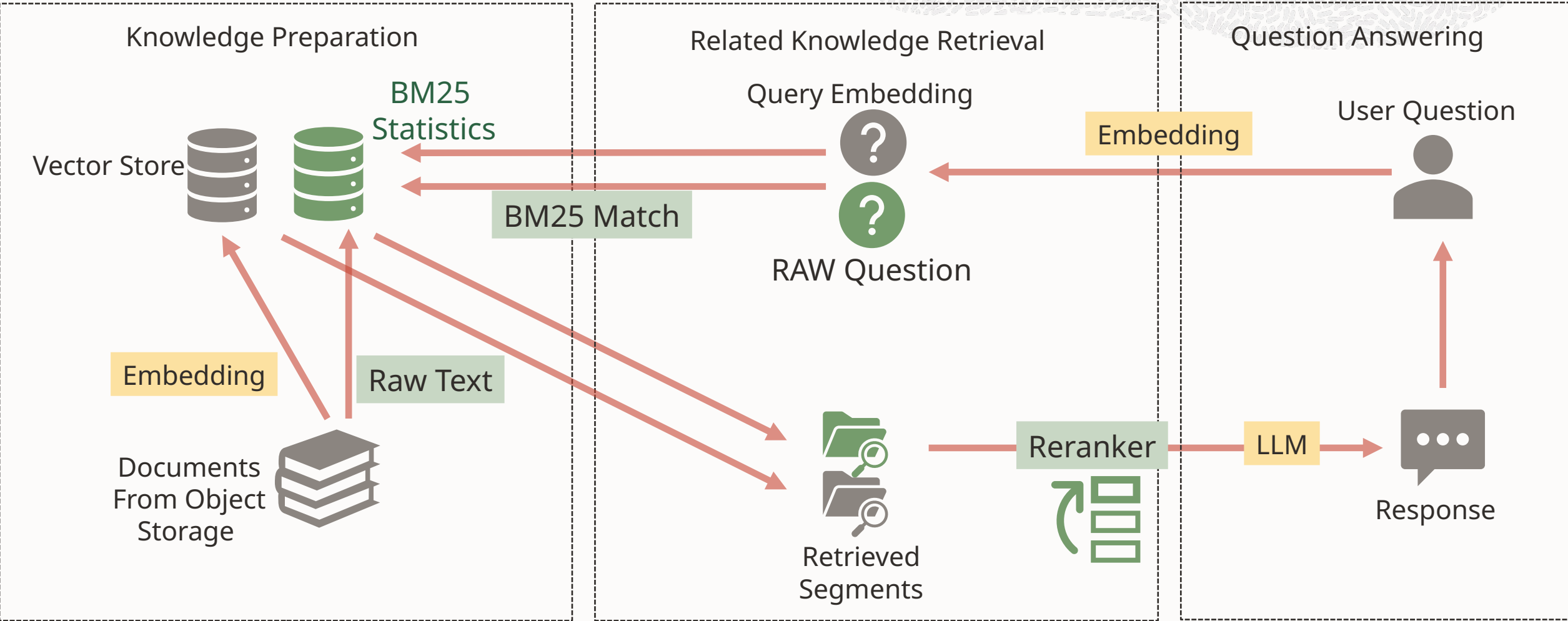
ORACLE



# Low Level RAG API



# Hybrid Search with Multi Retrievers (Full Text Search + Vector)

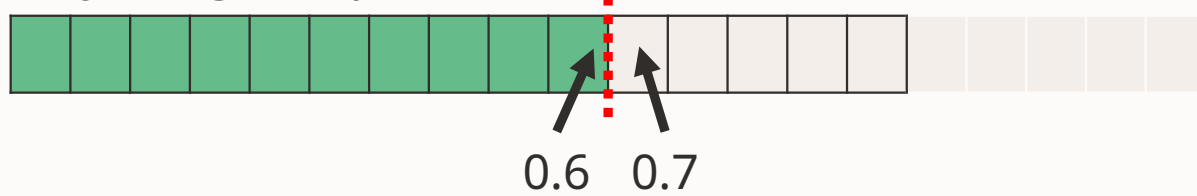


# Additional Retrieval Options



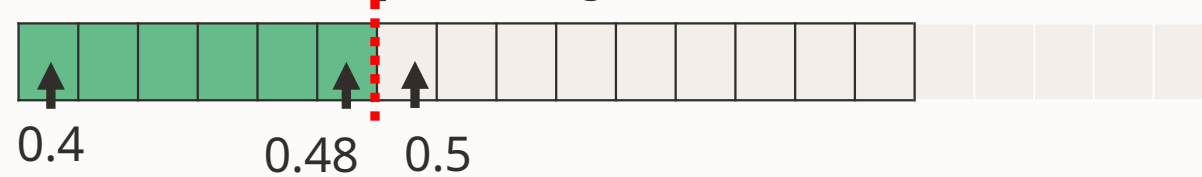
- max\_distance

n\_citations=15, **max\_distance=0.6**,  
only the green part will be returned



- percentage\_distance

n\_citations=15, **percentage\_distance=20**



pct\_distance → threshold =  $0.4 + 0.4 * 20\% = 0.48$   
Then use 0.48 to filter the segments.

- segment\_overlap

Retrieves segments adjacent to the nearest ones

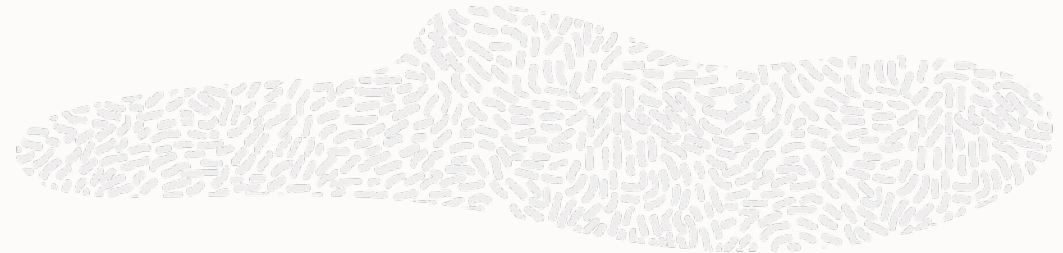
n\_citations = 2, segment\_overlap = 1

→ retrieved segments: 1 2 3 4

id	segment	Cosine distance
1	Super Bowl LVIII was an American football ...	0.4
2	In a rematch of Super Bowl LIV from four years earlier..	0.6
3	<b>The game was played on February 11, 2024, at Allegiant Stadium in Paradise, Nevada.</b>	0.7
4	This was the first Super Bowl to be held in the state of Nevada...	0.5

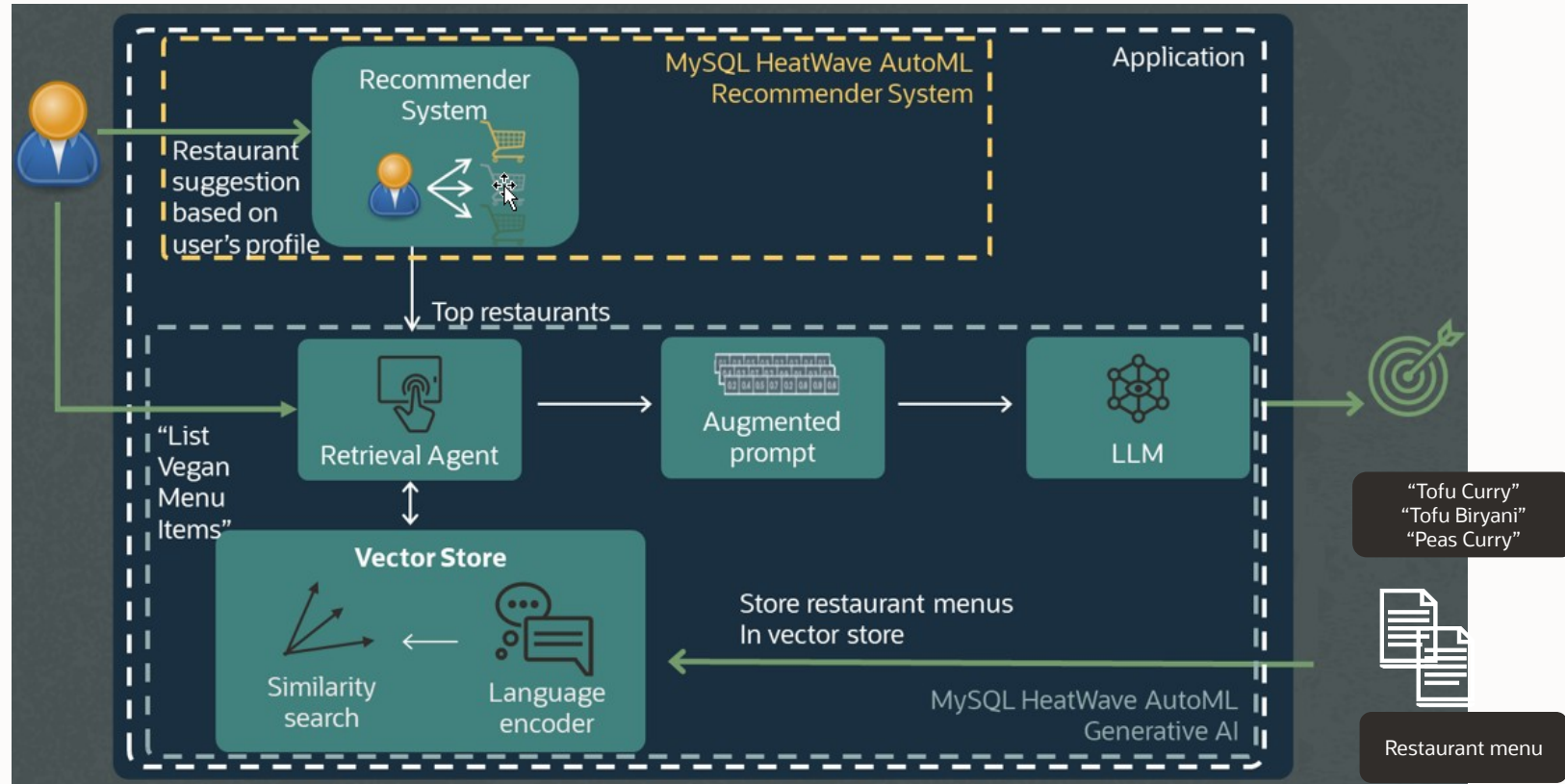
# Example : Personalization

## RAG

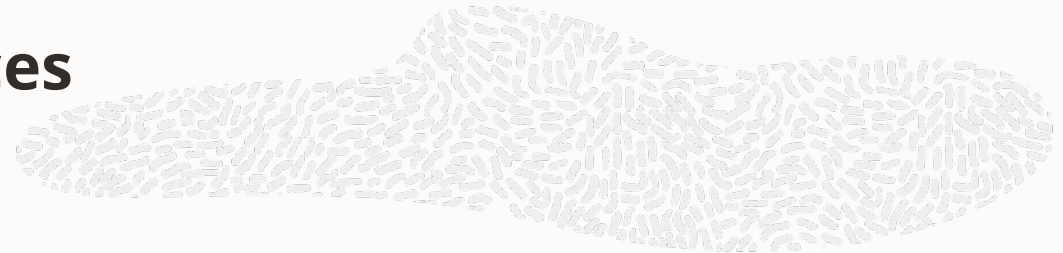


Recommend, Retrieve, and Generate descriptions of dishes based on user preference


### Personalized Menu: HW AutoML + Retrieval Augmented Generation




# Recommend dishes based on preferences




Home > Dubai > Al Barsha 1 Restaurants > **China Chowk**



**Sweet Spicy Chili Beef**  
Mains  
"Fiery Beef Fusion: Succulent beef stir-fried with a tantalizing blend of sweet chili sauce, aromatic garlic, and punch of dry chili, creating a harmonious balance of sweetness and heat"  
**48 AED**




**Tofu Stir Fry**  
Mains  
Indulge in our delightful **Sweet Chili Tofu Stir-Fry!** This vibrant dish features tender tofu cubes stir-fried with green beans, crunchy bean sprouts, and julienne-cut carrots, all perfectly coated in a tantalizing sweet chili sauce. The harmony of textures and flavors creates a savory-sweet sensation that's both satisfying and wholesome.  
**44 AED**




**Sweet And Sour Prawns**  
Mains  
"Golden tempura-coated shrimp tossed with a trio of colorful capsicums, sweet pineapple, and white onion—a fusion of crispy, savory, and sweet flavors in every bite."  
**45 AED**


Home > Dubai > Al Barsha 1 Restaurants > **Geppetto Pizzeria**




**Margherita Vegan Pizza**  
Pizza Vegan  
Italian Tomato, Vegan Mozzarella, Basil And Oregano  
**49 AED** +  
Customize




**Pepperoni Vegan Pizza**  
Pizza Vegan  
Italian Tomato, Vegan Mozzarella, Vegan Sausage, Basil And Oregano  
**96 AED** +  
Customize




**Quattro Formaggi Vegann Pizza**  
Pizza Vegan  
Italian Tomato, Vegan Mozzarella, Vegan Parmesan, Vegan Cheddar, Vegan Gouda And Basil  
**75 AED** +  
Customize



**Verde Vegan Pizza**  
Pizza Vegan  
Italian Tomato, Fresh Vegan Pesto, Vegan Mozzarella, Red Onion, Pepper, Zucchini, Eggplant... [more](#)  
**65 AED** +  
Customize



**Funghi Vegan Pizza**  
Pizza Vegan  
Italian Tomato, Vegan Mozzarella, Mushrooms, Basil And Oregano  
**53 AED** +  
Customize



**Tartufo Vegan Pizza**  
Pizza Vegan  
Italian Vegan Truffle Cream, Vegan Cream Cheese, Mushrooms And Fresh Italian Summer Truffl... [more](#)  
**99 AED** +  
Customize



```
ask_question('Recommend three vegan dishes with tofu', color='b
```

1. Sweet Chili Tofu Stir-Fry - a savory and sweet dish featuring tender tofu cubes stir-fried with vegetables and coated in a delicious sweet chili sauce.
  2. CustomizeTofu Stir Fry - a customizable dish where you can choose various vegetables and sauces to accompany your tofu stir-fry.
  3. Vegan Pizza - Italian Tomato, Vegan Mozzarella, Tofu, and Basil - a delicious and satisfying pizza option by customizing with tofu as your protein choice.
- Would you like me to assist you with more vegan dish recommendations that incorporate tofu?



---

*The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.*



# What is the goal ?

- Provide Generative AI capabilities inside HeatWave
  - Be able to ask questions in natural language and get back answers
  - Be able to query user's documents stored in the object store and get relevant information
- Give users the option to try out different models from functionality/cost perspective
- Enable a single system that takes advantage of Machine Learning and Generative AI
- Three components:
  - Document Ingest
  - Vector Store
  - LLMs