ORACLE

HeatWave GenAI —Technical Overview

Integrated, automated, and secure.

Purpose statement

This document provides an overview of features and enhancements included in Oracle HeatWave GenAI. It is intended solely to help you assess the benefits of HeatWave GenAI and to plan your IT projects.

Disclaimer

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement, nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. 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, and timing of any features or functionality described in this document remains at the sole discretion of Oracle. Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

Executive Summary

Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) are groundbreaking technologies that have the potential to reshape how we interact with data and significantly expand insights into enterprise data. From super-charged semantic search of proprietary documents to summarizing deeply technical content, one can use GenAI to ask questions across languages and various areas of domain-expertise.

Imagine an e-commerce platform being able to display summarized product reviews derived automatically from hundreds of user review comments at scale; or a large bank being able to significantly reduce fraudulent transactions, without moving data and without AI expertise.

Oracle HeatWave GenAI provides the industry's first in-database large language models (LLMs); an automated, in-database vector store; scale-out vector processing; and the ability to have contextual conversations in natural language informed by unstructured content. These capabilities enable customers to bring the power of generative AI to their enterprise data in their native languages—without requiring AI expertise, having to move data to a separate vector database, or additional cost.

HeatWave uniquely integrates these capabilities in-database, enabling turnkey GenAI application development with enhanced performance and enterprise-grade data security. Unlike specialized vector databases, HeatWave users can combine GenAI with other HeatWave built-in capabilities, such as transaction processing, analytics across data warehouses and data lakes, and machine learning (ML) to create powerful and novel applications delivering more relevant and insightful responses—without the complexity, latency, security risks, and cost of extract, transform, and load (ETL) duplication.

HeatWave GenAI is available at no additional cost to HeatWave customers.

Existing GenAI Adoption Challenges

As powerful as they are, using GenAI technologies with enterprise data can remain a convoluted and expensive proposition, requiring AI expertise.

Proprietary data

While capable of semantic reasoning and language generation, most LLMs are trained on public data and have no knowledge of proprietary enterprise information. Enterprise data needs to be captured via a vector store to help LLMs generate more accurate and contextually relevant answers to users' questions.

Complexity

Building a GenAI application is a complex multi-step process requiring users to:

- Select the right LLM based on application domain, cost, and responsiveness.
- Capture proprietary data along with context in a vector store using the right embedding model.
- Ensure each input query is encoded using the same embedding model.
- Integrate all of the above spanning multiple cloud services.

The availability of these services is not universal, greatly hindering application portability and consistency of results across different cloud regions and cloud providers.

Expertise

Using GenAI to obtain relevant and useful results requires deep expertise in LLMs, embedding models, and similarity search algorithms. Each of these components must be carefully selected, integrated, and tuned to work well with each other to perform efficiently on a given compute infrastructure. The necessary expertise spans several domains and is rapidly evolving, making the learning curve very steep.

High Costs

The development of GenAI solutions requires a substantial investment in several areas. These include acquiring high-quality data for training, provisioning scarce and expensive GPUs and other computing resources, allocating enough storage space for vector embeddings, and hiring skilled professionals to design and manage these complex systems.

Thus, the enormous opportunity of using GenAI, recognized by IT leaders, developers, and creative professionals alike, can often be locked behind the complexity, required expertise, and high cost of understanding and using Generative AI tools and technologies.

"We believe that Generative AI can enhance the efficiency of our client-facing teams through use of semantic search and summarization of documents by using HeatWave GenAI with the HeatWave Vector, Store which offers unique capabilities. We are working on this potential use case and we hope to productize the benefits to our teams."

Ramesh Lakshminarayanan
CIO & Group Head-IT, HDFC Bank

HeatWave GenAI

With HeatWave GenAI, developers can create a vector store for enterprise unstructured content with a single SQL command, using built-in embedding models. Users can perform natural language searches in a single step using either in-database or external LLMs. Data does not leave the database, and, with HeatWave's very high performance and scalability, there is no need to provision GPUs. As a result, developers can reduce application complexity, increase performance, improve data security, and lower costs.

All the elements of the pipeline necessary to use HeatWave GenAI with proprietary data are built-in, integrated, and optimized to work with each other, enabling turnkey generative AI application development.

HeatWave GenAI features include:

- **In-database LLMs:** simplify the development of generative AI applications at a lower cost. Customers can benefit from generative AI without the complexity of external LLM selection and integration,

and without worrying about the availability of LLMs in various cloud providers' data centers. The in-database LLMs enable customers to search data, generate or summarize content, and perform retrieval-augmented generation (RAG) with HeatWave Vector Store. In addition, they can combine generative AI with other built-in HeatWave capabilities such as AutoML to build richer applications. HeatWave GenAI is also integrated with the OCI Generative AI service and Amazon Bedrock to access pre-trained, foundation models from leading LLM providers.

- **Automated, In-database vector store:** lets customers to use generative AI with their business documents, without moving data to a separate vector database and without AI expertise. All the steps to create a vector store and vector embeddings are automated and executed inside the database, including discovering the documents in object storage, parsing them, generating embeddings in a highly parallel and optimized way, and inserting them into the vector store, making HeatWave Vector Store efficient and easy to use. Using a vector store for RAG helps solve the hallucination challenge of LLMs as the models can search proprietary data with appropriate context to provide more accurate and relevant answers.

- **Scale-out vector processing:** delivers very fast semantic search results without any loss of accuracy. HeatWave supports a new, native VECTOR data type and an optimized implementation of the distance function**,** enabling customers to perform semantic queries with standard SQL. In-memory hybrid columnar representation and the scale-out architecture of HeatWave enable vector processing to execute at near-memory bandwidth and parallelize across up to 512 HeatWave nodes. As a result, customers get their questions answered rapidly. Users can also combine semantic search with other SQL operators to, for example, join several tables with different documents and perform similarity searches across all documents.

- **HeatWave Chat:** lets users ask questions using natural language or SQL through a set of chatbot functionalities and a graphical interface integrated with the Visual Studio Code plugin for MySQL Shell. The integrated Lakehouse Navigator enables users to select files from object storage and create a vector store. Users can search across the entire database or restrict the search to a folder. HeatWave maintains context with the history of questions asked, citations of the source documents, and the prompt to the LLM. This facilitates a contextual conversation and allows users to verify the source of answers generated by the LLM. This context is maintained in HeatWave and is available to any application using HeatWave.

"For organizations looking to run highly accurate and blazingly fast similarity search queries on unstructured data in object storage at an affordable price, HeatWave has the answer today."

Marc Staimer
Senior Analyst
theCUBEresearch

HeatWave GenAI Use Cases

The addition of Generative AI capabilities in HeatWave enables a variety of new use-cases and applications. LLMs provide the ability for users to interact with data in the most natural way, i.e., natural language. Natural language

provides an intuitive way to interact with unstructured data. Use cases enabled by HeatWave GenAI include:

Retrieval Augmented Generation

HeatWave GenAI with HeatWave Vector Store enables semantic search on enterprise data and helps get more accurate and contextually relevant answers. This is achieved by performing similarity search on massive amounts of unstructured data. LLMs use the retrieved data as context to generate natural language responses to users' questions.

Content generation and summarization

HeatWave GenAI can generate insights and reports from enterprise data and documents. For example, it can help answer questions about content in PDF instruction manuals, discussion forums or blogs, as well as generate summaries of product review comments on e-commerce platforms.

Synergy of built-in GenAI and ML

The combination of AutoML, GenAI, and vector store, all within one database, delivers more value to customers. It helps reduce costs and get more accurate results faster. For instance, AutoML excels at rapidly identifying hidden patterns in structured data and can act as a filter for data that is then processed by GenAI.

Conversation in natural language

The new functionality enables conversations informed by unstructured documents using natural language. The conversations are possible since context and chat history are maintained in HeatWave to enable follow-up questions. Breaking down language barriers, these conversations are supported in multiple languages.

> "HeatWave GenAI exemplifies elegant and synergistic engineering design, demonstrating that such a combination of performance, efficiency, and security is impossible to achieve by just randomly connecting individual cloud services."
>
> **Alexei Balaganski**
> Lead Analyst and CTO
> KuppingerCole Analysts

Automated, In-database Vector Store

HeatWave Vector Store provides simple, fast, scalable, fully integrated, and automated vector store creation. Users only need to execute one simple and familiar SQL command. Under the hood, this involves a multi-step process including reading unstructured data in PDF, HTML, Word, TXT, and PowerPoint formats from object storage, parsing the text out of these documents, partitioning it into smaller semantically sound segments, generating vector embeddings from them, and finally, storing the embeddings in a standard HeatWave Lakehouse table:

- In addition to data embeddings, the document metadata is also stored in HeatWave Vector Store, enabling users to combine similarity search with filtering based on document attributes like name, author, or creation date.
- The unstructured data itself, which is typically much larger than the parsed text and its embeddings, is not copied to the vector store, which lowers cost.
- Documents in more than two dozen languages are supported, with HeatWave automatically selecting the most appropriate embedding model to create vector embeddings.

HeatWave Vector Store is:

Simple

The creation of HeatWave Vector Store is fully automated and integrated within the database, triggered with one simple and familiar SQL command:

```
mysql> CALL sys.heatwave_load(@db_and_doc_path, @options);
```

- Document parsing (using Oracle OutsideIn), and text segmentation functionalities are built-in and used without additional user input.
- No AI expertise is needed: the embedding model is in-database and automatically chosen.
- Updates to documents in object storage are incrementally applied to the vector store.
- One interface: this is the same interface used for any HeatWave Lakehouse table across structured, semi-structured, and unstructured data.

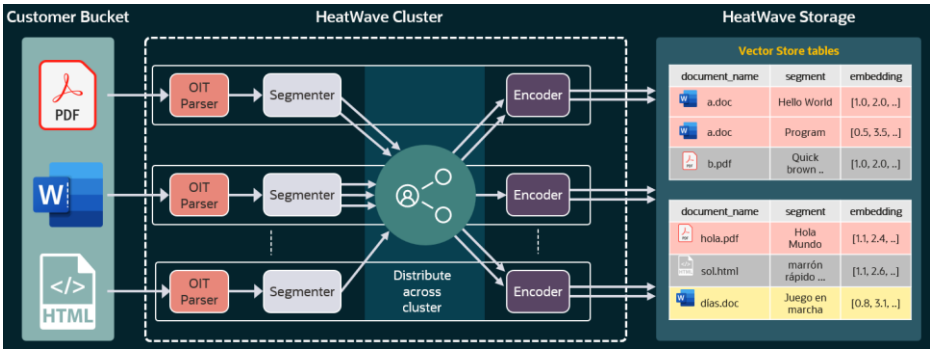
Secure

User data remains in HeatWave:

- All data transformations, including vector embedding creation, are executed in the database.
- VECTOR is a native data type in HeatWave, and similarity search on vectors is completed within HeatWave, using the highly performant HeatWave processing engine.
- HeatWave Vector Store is integrated with the in-database LLMs, thereby keeping user data in HeatWave.

Better Performance

Each stage of HeatWave Vector Store creation has been optimized for performance and scalability. Different stages from document parsing to vector embedding generation have wildly different characteristics, resource requirements, and parallelism opportunities. These have been integrated and tuned to work efficiently with each other, enabling HeatWave Vector Store creation to scale to thousands of cores.



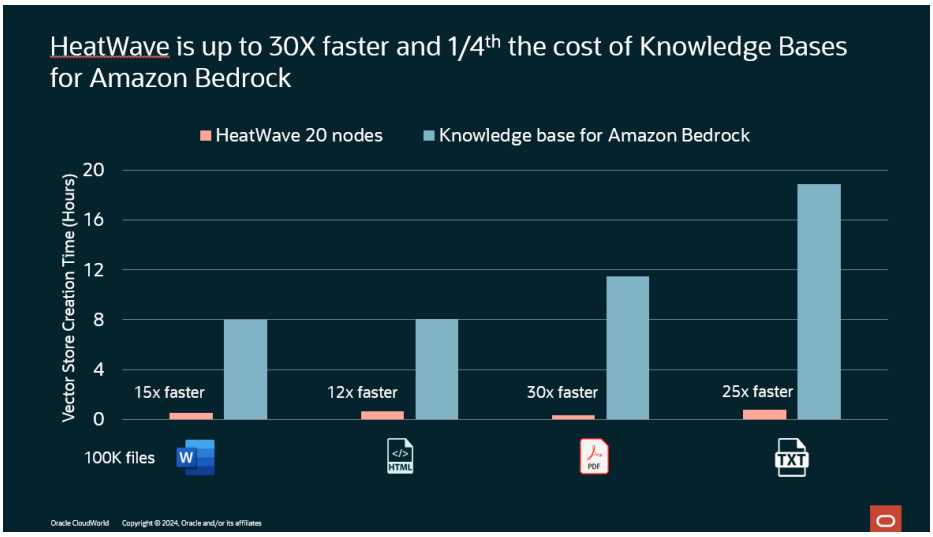
"HeatWave GenAI makes it extremely simple to take advantage of generative AI. The support for in-database LLMs and in-database vector creation leads to significant reduction in application complexity, predictable inference latency, and, most of all, no additional cost to us to use the LLMs or create the embeddings."

Vijay Sundhar
CEO
SmarterD

Lower Cost

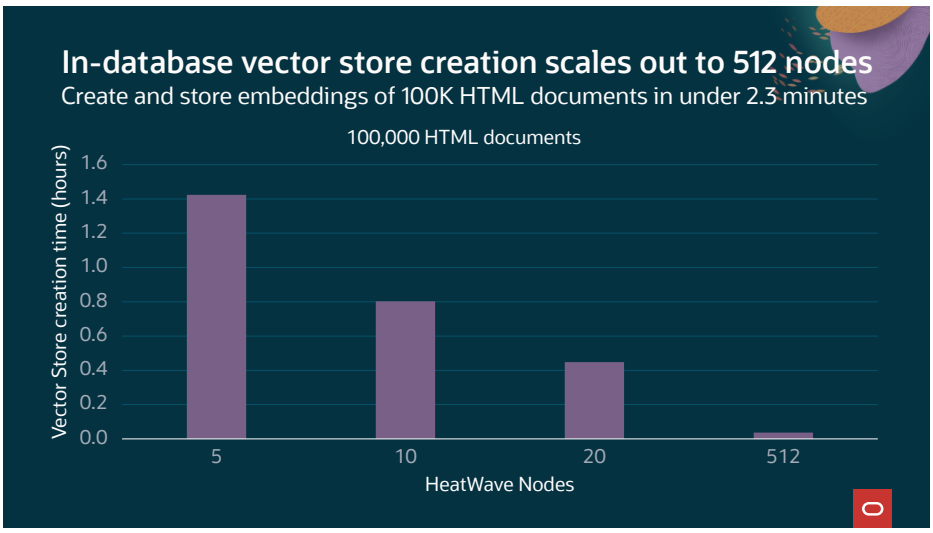
The tight integration of all vector creation stages offers optimization and tuning opportunities unavailable to a fragmented vector store creation pipeline. No additional services are needed, removing the need to move data across various AI/ML and database services. All of this can be achieved using an existing HeatWave cluster.

HeatWave Vector Store creation is up to 30X faster than Knowledge Bases for Amazon Bedrock at 1/4th the cost.



Flexibility in Cost and Performance

HeatWave Vector Store creation performance scales with larger cluster sizes, giving users a simple way to reduce their vector store creation time. Users have the flexibility to choose the required performance and associated cost according to their needs for each workload.



"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
EatEasy

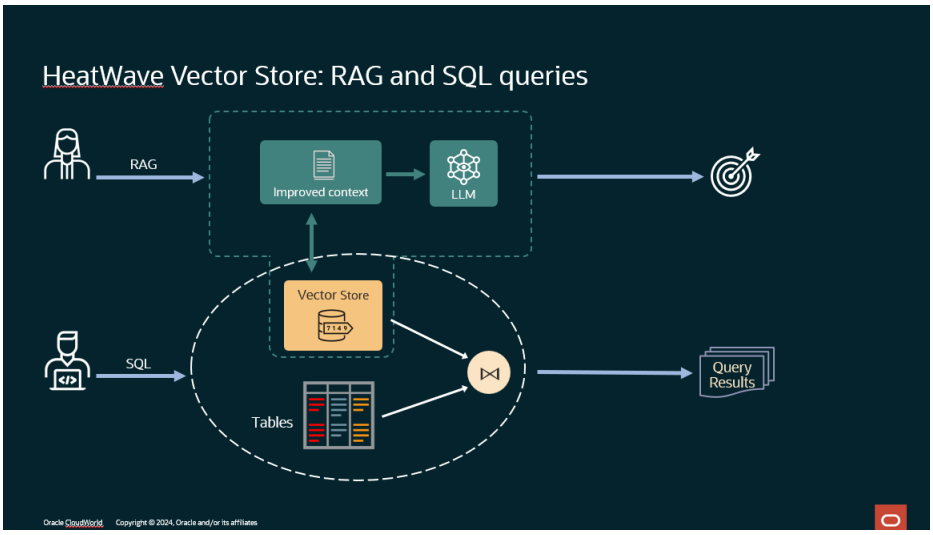
Text in images

Many organizations receive information via printed media such as invoices, contracts, and other documents. They often store these documents by scanning them into a digital image format. With OCR support, HeatWave GenAI can now convert the text information in these scanned documents into vector embeddings, enabling organizations to query the documents to perform similarity searches and retrieval augmented generation (RAG).

Using HeatWave Vector Store

Once the vector store is created, it can be used by applications to perform similarity search using native SQL operators and to provide proprietary enterprise context for RAG use cases.

HeatWave Vector Store stores does not only store the vector embeddings but also the raw text and document metadata. Using familiar and native SQL operators, applications can filter embeddings on these fields, find the most similar embeddings to input queries, and combine these results with other MySQL tables using JOINs, Aggregates, etc. HeatWave Vector Store thus seamlessly fits into the SQL world while introducing novel enterprise data search capabilities.



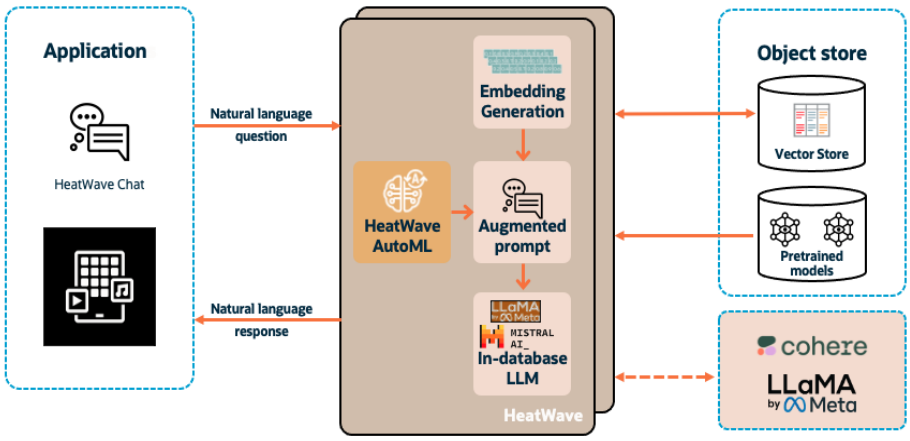
Providing the in-database LLMs with context from the vector store (RAG), allows users to ask questions about their proprietary data, and to receive more accurate and contextually relevant responses in natural language.

"HeatWave is taking a big step in making generative AI and Retrieval-Augmented Generation (RAG) more accessible by pushing all the complexity of creating vector embeddings under the hood. Developers simply point to the source files sitting in cloud object storage, and HeatWave then handles the heavy lift."

Tony Baer
Founder and CEO
dbinsight

In-database LLMs

HeatWave GenAI offers an integrated environment for Generative AI, where every aspect of the Generative AI pipeline is integrated within the database.



Key features include:

In-database Large Language Models

HeatWave GenAI supports LLMs hosted and executed within HeatWave, in addition to external LLMs. These LLMs are smaller quantized versions of popular public LLMs such as Mistral and Llama3 and run within HeatWave. In-HeatWave LLMs help increase security as inference happens in-database. They provide HeatWave customers an alternative to GPU dependent LLMs at no additional cost. The smaller size of enterprise data compared to the large corpus of public data, the precise context gleaned from HeatWave Vector Store, and well-engineered prompts make these LLMs nearly as accurate as larger external LLMs for a variety of enterprise use cases like RAG.

External Generative AI Services

HeatWave GenAI is integrated with OCI Generative AI Service and Amazon Bedrock which uses large foundation models, such as the Cohere command and LLaMA by Meta, to generate high-quality responses. This provides the flexibility for customers to invoke LLMs of their choice.

In-database embedding generation of input questions

HeatWave Vector Store provides valuable context for LLMs for RAG use cases. With HeatWave GenAI, user questions are transparently and automatically encoded with the same embedding model as HeatWave Vector Store (a necessity for RAG), without additional user input or invoking a separate service.

Multi-lingual support

As mentioned above, HeatWave GenAI supports documents in more than two dozen of languages for vector store creation, similarity search and natural language interaction. For RAG use cases, users can ask questions about their documents in HeatWave Vector Store in the same language.

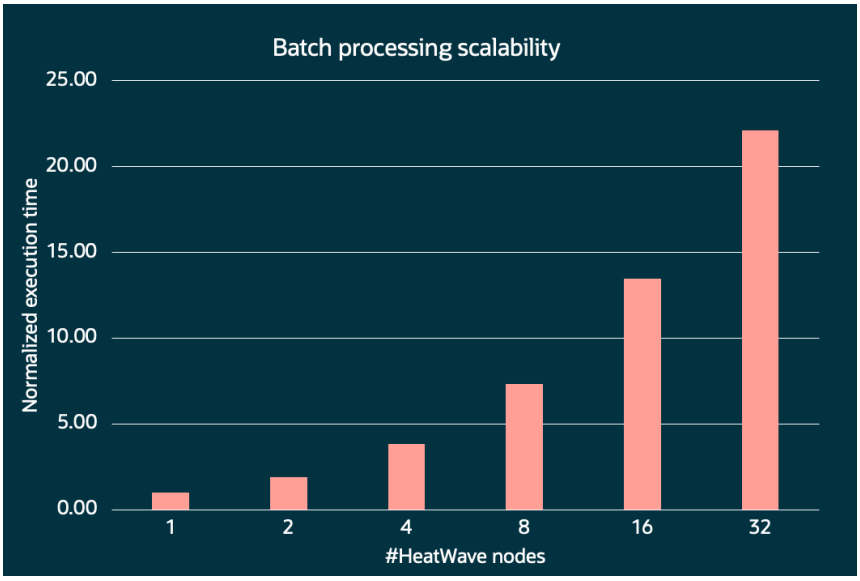
Batch processing of GenAI LLM inference requests

HeatWave GenAI allows users to provide LLM generation, summarization or embedding requests as a batch in the form of a table. Each row in the table corresponds to a single request, which are then processed concurrently across different nodes in the HeatWave cluster. The output of these requests can then be placed in a column in the same table or a new table. This

> "HeatWave in-database LLMs, in-database vector store, scale-out in-memory vector processing, and HeatWave Chat are very differentiated capabilities from Oracle that democratize generative AI and make it very simple, secure, and inexpensive to use."
>
> **Eric Aguilar**
> Founder
> Aiwifi

improves LLM inference performance and throughput while keeping the inference latency of each of the request unchanged.

The below chart shows the normalized execution time taken for 1000 RAG queries on the HeatWave cluster, achieving a speedup of 22X for 32 HeatWave nodes in comparison to a single HeatWave node.



Synergy with other built-in HeatWave capabilities

Customers can easily combine HeatWave GenAI capabilities with other built-in HeatWave capabilities, such as machine learning. This helps customers develop innovative applications while improving accuracy and performance as well as reducing costs.

HeatWave GenAI uses a familiar SQL interface and is easy to use for content generation, summarization, and RAG.

Content generation: User can query the LLM in HeatWave in natural language.

```
MySQL > SELECT sys.ML_GENERATE("What is Heatwave ?",
JSON_OBJECT("context", "MySQL has a service called Heatwave")));
```

Content summarization: User can receive a summary of a document loaded into MySQL.

```
MySQL > SET @document =
LOAD_FILE("/user/documents/prod_manual.txt");
MySQL > SELECT sys.ML_GENERATE(@document, JSON_OBJECT("task",
"summarization")));
```

Retrieval Augmented Generation: Users can search through enterprise documents using HeatWave Vector Store.

```
MySQL> CALL sys.ML_RAG("What is Heatwave ?", @output,  NULL);
MySQL> SELECT JSON_PRETTY(@output) \G
*********************************** 1. row
***************************************
JSON_PRETTY(@output): {
```

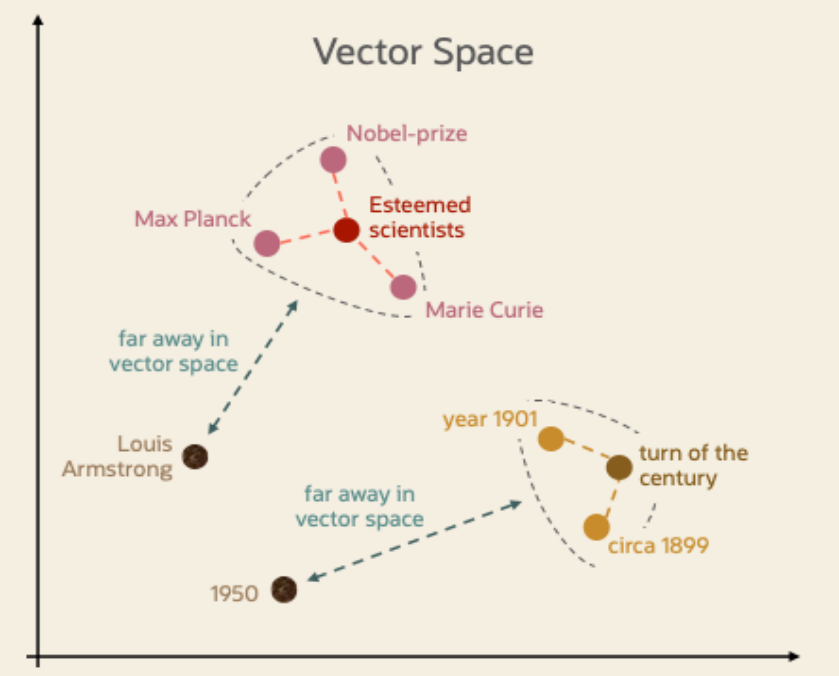
"HeatWave has been instrumental in our strategy to leverage on GenAI and Machine Learning capabilities. AutoML for predictive analytics on data, in-database LLMs, in-database Vector Store, and RAG within HeatWave, have been a cornerstone, to easily secure the adoption of Generative Artificial Intelligence with our enterprise data, making the things simple and faster than other solutions, combining all the capabilities into a single data platform service. We look forward to improving Toks customer experience, powered by HeatWave GenAI, Lakehouse and AutoML."

David Leo
IT Director
Toks

```
  "text": " HeatWave is a unified MySQL cloud database service
that provides transactions, real-time analytics across data
warehouses and data lakes, and machine learning capabilities. It
is designed to simplify ETL batch jobs ...",
  "citations": [
    {"segment": "long-running ETL batch jobs to refresh the
data. MySQL HeatWave provides a unified MySQL cloud database
service ...",
     "distance": 0.36409956216812134,
     "document_name":
"https://objectstorage...com/n/user_namespace/b/user_bucket/o/he
atwave_doc.pdf"},
    {"segment": "A MySQL HeatWave instance is a cluster composed
of a MySQL instance and multiple HeatWave nodes. When HeatWave
is enabled, ...",
     "distance": 0.36607372760772705,
     "document_name":
"https://objectstorage...com/n/user_namespace/b/user_bucket/o/he
atwave_doc.pdf"},
  ],
  "vector_store": [
    "`user_documents`.`vector_pdf`"
  ]
}
```


Scale-out Vector Processing

Traditional keyword-based search without context can fall short of delivering relevant results. For example, if you are looking for “esteemed scientists” you might not receive results about “Nobel prize winning physicists”, even though they are clearly relevant.



Advanced search such as similarity/semantic search bridges the gap by using machine learning to understand the meaning and context behind the

“The integration of Generative AI in Oracle MySQL HeatWave is a major leap forward for us at SOCOBOX. By bringing in-database LLMs, automated vector processing, AutoML, and Lakehouse into our workflows, we can now deliver powerful AI-driven insights and applications without the overhead of external tools. This comprehensive approach not only simplifies our operations but also ensures real-time, cost-effective solutions that resonate with the demands of our customers.”

Hans Ospina
Founder
SOCOBOX SAS

data. Using machine learning models, data is converted into numerical representations called vector embeddings (vectors), capturing the context of the data and relationships to other data.

Vectors are points in a multi-dimensional vector space. Each point represents the vector embedding of the corresponding data. Data similar in semantic meaning are closer in vector space. Using the previous example, if a user searches for “esteemed scientists”, similarity search will return relevant results such as “Nobel-prize”, “Marie Curie”, and “Max Planck”. This similarity in meaning is derived from the context of the data and is represented by small distances between the vector embeddings of these data.

Vector data type

Storing and maintaining a vector store in an optimized storage format both in long-term storage and in-memory is crucial for supporting scalable semantic search queries on millions of vectors. With HeatWave GenAI, VECTOR is a natively supported data type.

```
mysql> CREATE TABLE wikipedia (page_data TEXT,
       page_embedding VECTOR(384));
```

Vector Distance Functions

Vector distance functions measure the similarity between vectors by calculating the mathematical distance between two multi-dimensional vectors. A smaller *distance* equates to *higher* similarity.

The embedding model used to encode the vectors dictates the distance function most suited for comparing those vectors. HeatWave natively supports the commonly used distance functions such as COSINE, DOT, L2/EUCLIDEAN, L1/MANHATTAN, L1^2/MANHATTAN_SQUARED, L2^2/EUCLIDEAN_SQUARED, HAMMING.

```
mysql> SELECT id, title FROM dbpedia
       WHERE text NOT LIKE '%Italian pronunciation:%'
       ORDER BY DISTANCE(embedding, @query_embedding,
                         'COSINE') LIMIT 10;
```

Predictable and exact answers

One of the advantages of HeatWave’s hybrid columnar engine is that it does not depend on indexes to achieve high performance for query processing. The same applies for the vector store and similarity search. HeatWave performs an exact search predictably finding the most similar vectors for each query.

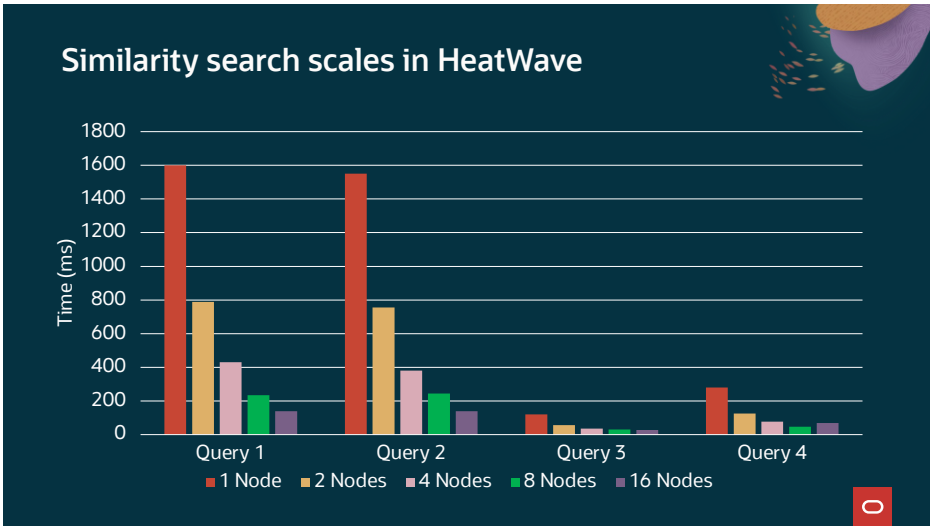
Scale-out Performance

Vectors are stored natively in HeatWave’s hybrid-columnar format, enabling highly performant and scalable similarity search in HeatWave, executed at near-memory bandwidth. This is the result of deep processor specific optimizations (SIMDization of operators), and efficient distributed and parallel algorithms – translating into faster results and lower cost for users.

> “HeatWave GenAI makes it extremely simple to take advantage of generative AI. With in-database LLMs, in-database vector store, and HeatWave Chat, HeatWave GenAI is a very comprehensive offering that greatly simplifies the development of a new class of apps. The ability to have a human-like conversation with follow-up questions, without going through complex manual operations, is very valuable.”
>
> **Fabricio Rucci**
> IT Director
> Natura

As demonstrated by a third-party benchmark using a variety of similarity search queries on tables ranging from 1.6GB to 300GB in size, HeatWave GenAI is 30X faster than Snowflake at 25% lower cost, 15X faster than Databricks at 85% lower cost, and 18X faster than Google BigQuery at 60% lower cost.

Similarity search scales with the size of the HeatWave cluster, providing great flexibility to customers.



JavaScript support

As generative AI and LLMs primarily handle textual and JSON data, JavaScript is a natural choice for manipulating this data. We've added native support for the vector data type in JavaScript and the ability to invoke HeatWave GenAI capabilities from a JavaScript program. Developers can seamlessly use HeatWave GenAI APIs in their JavaScript functions, which are executed directly in the database. Developers can pre-process the prompt or post-process the generated response using JavaScript's powerful string and JSON processing capabilities.

The following example shows how one can write a sentiment analysis application in just a few lines of code. Let's assume we have a table containing customer reviews for a product in HeatWave MySQL. Using the HeatWave GenAI support in JavaScript, we can easily identify the sentiment of each review using a simple prompt and add the generated sentiment back to the table.

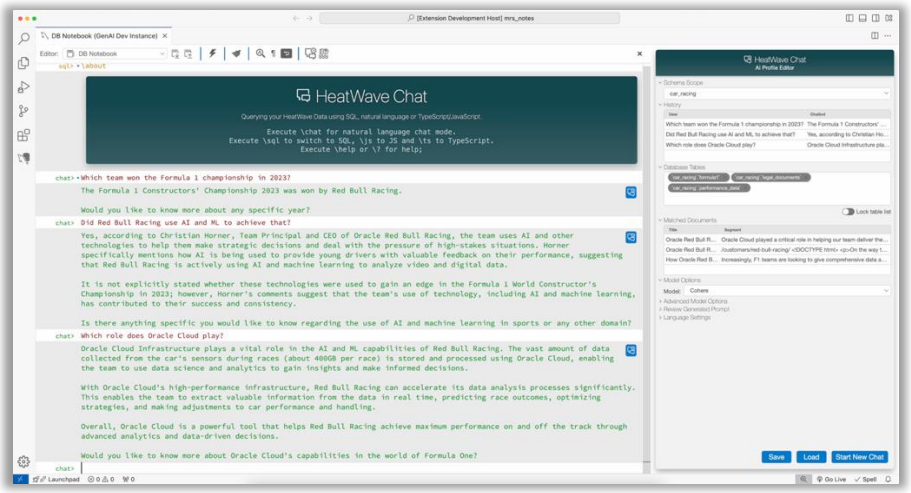
```
CREATE PROCEDURE SENTIMENT_ANALYSIS(
  IN review TEXT,
  IN review_id INT
) LANGUAGE JAVASCRIPT AS $$
  let prompt = `Classify the review into NEGATIVE or POSITIVE
\n${review}. Please provide a single word to describe the sentiment:
“POSITIVE” or “NEGATIVE”. \nSentiment:`;

  let sentiment = ml.generate(prompt);
  let processed_sentiment = sentiment.search(“POSITIVE”) ?
“POSITIVE” : “NEGATIVE”;
  let sql = session.prepare(`UPDATE reviews SET sentiment = ? WHERE
id = ?`);
  sql.bind(processed_sentiment, review_id).execute();
$$;
```

HeatWave Chat

HeatWave GenAI enables applications to interact with data using natural language. HeatWave Chat exposes an SQL API and a graphical interface integrated with the existing Visual Studio Code plugin for MySQL Shell.

HeatWave preserves the history of the chat as well as other options exposed to the user. Thus, the SQL API works out-of-the-box from any MySQL client or application, and customers can directly interact with their data in HeatWave Vector Store using in-HeatWave LLMs or external ones.

Despite its simplicity, the HeatWave Chat API is extremely powerful and can be used by developers to build customized chat applications by specifying custom settings, prompt, chat history length, number of citations to be used, and many more advanced options. By default, HeatWave Chat will search for an answer to users' queries across all ingested documents by automatically discovering available vector stores and will return the answer along with relevant citations. However, a user can easily limit the scope of search to specific document collections – either certain vector stores or documents to include in the search.



Ingesting documents into HeatWave Vector Store is further simplified through this integrated drag and drop User Interface. Countless optimizations helping to improve efficiency and to lower cost are built-in, and parameters are automatically tuned.

Application Examples

HeatWave GenAI can help you solve challenging real-world problems such as:

- Fraud detection
- Technical support improvement

Detection of Fraudulent Bank Transactions

Banks spend significant time and effort to detect and prevent fraudulent transactions, including money theft or money laundering. Such fraud can result in significant monetary loss as well as reputational damage to the bank.

HeatWave provides all the components necessary to build a fraud detection solution in-database by combining HeatWave AutoML's unsupervised anomaly detection coupled with HeatWave GenAI. Each banking transaction goes through the Anomaly Detection (AD) model first. The AD model predicts whether a transaction is normal or anomalous and provides the probability of each transaction being anomalous. The application can set a threshold for anomaly probability and filter transactions with highest probabilities for further analysis and summarization by an LLM model. As a result, each potentially fraudulent transaction is accompanied by a natural language description of the transaction details, a diagnosis of why the transaction is suspicious, and any other relevant information that a bank operator may need to perform further investigation. This saves tremendous amount of time for bank operators and help them prevent fraud.

AskME – HeatWave Technical Support Application

Technical support is an important component of any business to keep customer satisfaction high. However, it requires technical personnel to spend time answering customer questions. In the absence of technical support,

customers must search through large amounts of documentation, which is often composed of dense technical details.

HeatWave GenAI enables semantic search on proprietary documentation in HeatWave Vector Store. “AskME” is an example application that enables users to ask questions, and answers are generated based on the technical documentation, manuals, and bug records in real-time. The application can also help resolve and find the root cause of issues based on known bugs and limitations. It provides a natural language interface and indicates references to documents from which answers are derived.

See a demo of the AskME application here.

Summary

Generative AI is reshaping our world but can be complex to implement. HeatWave GenAI, with industry’s first automated and in-database vector store and in-database LLMs, plus scale out vector processing and the ability to have contextual conversations in natural language, enables you to take advantage of generative AI without AI expertise, data movement, or additional cost. You can also easily combine HeatWave GenAI with other built-in HeatWave capabilities, and benefit from the enhanced data security of an in-database solution.

HeatWave GenAI is a powerful new tool to help you develop new, innovative applications. Try it out!

Connect with us

Call +**1.800.ORACLE1** or visit **oracle.com**. Outside North America, find your local office at: **oracle.com/contact**.

blogs.oracle.com facebook.com/oracle twitter.com/oracle