Research Brief:

ORACLE HEATWAVE GENAI

STEVE MCDOWELL, CHIEF ANALYST

JUNE 2024

ORACLE HEATWAVE GENAI

Oracle HeatWave GenAI is now generally available. The release features the industry's first in-database large language models (LLMs), automated vector store, scale-out vector processing, and natural language conversations informed by unstructured content.

The release enables enterprises to leverage generative AI with their data without needing AI expertise or moving data to a separate vector database. Available at no extra cost to HeatWave customers, these capabilities enhance data security and performance while reducing costs.

Key features include:

- **In-Database LLMs**: Simplify the development of generative AI applications, allowing for data search, content generation, and retrieval-augmented generation (RAG) with HeatWave Vector Store.
- **Automated Vector Store**: This enables easy use of generative AI with business documents without data transfer and AI expertise, automating the creation and embedding process.
- **Scale-Out Vector Processing**: This technology provides fast and accurate semantic search results using a new VECTOR data type and optimized distance function, enabling efficient parallel processing.
- **HeatWave Chat**: A Visual Code plug-in for MySQL Shell, offering a graphical interface for natural language or SQL queries and maintaining context for continuous, accurate conversation.

Let's take a deeper look at each.

IN-DATABASE LLMS

Oracle's HeatWave GenAI incorporates in-database LLMs, enabling several advanced capabilities designed to simplify and enhance the development and deployment of generative AI applications within enterprise environments.

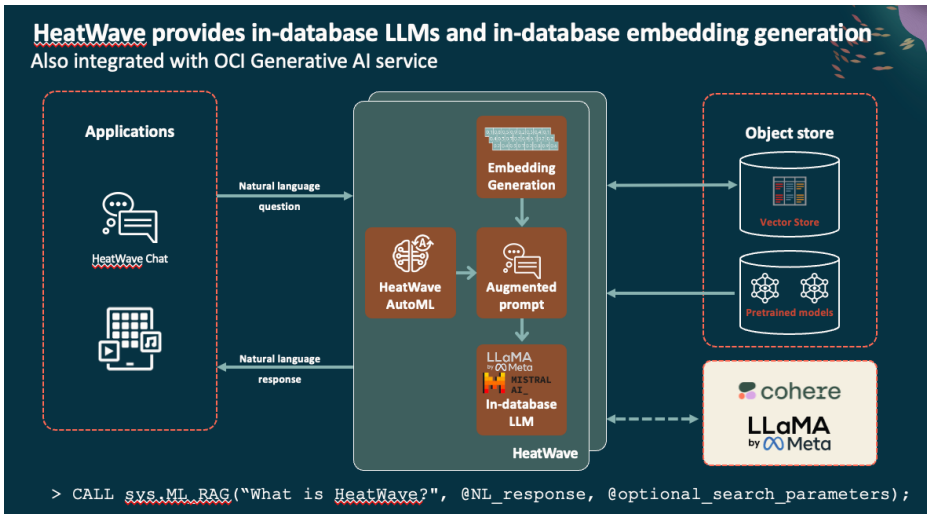

Figure 1. Oracle HeatWave In-Database LLMs (source: Oracle Corporation)

Here's what in-database large language models enable:

1. **Integrated LLMs**: These models are embedded directly within the database, eliminating the need for external LLM selection and integration, reducing complexity, and ensuring consistent availability across various cloud environments.

2. **Simplified Application Development**: With in-database LLMs, developers can leverage generative AI functionalities such as data search, content generation, and retrieval-augmented generation (RAG) without requiring extensive AI expertise, allowing faster and more efficient application development.

3. **Enhanced Data Security and Performance**: By keeping the data within the database, in-database LLMs minimize the need for data transfer, improving security and maintaining data integrity. This setup also takes advantage of HeatWave's scale and performance capabilities, negating the need for additional hardware like GPUs.

4. **Versatile Functionality**: The LLMs support various tasks, including generating and summarizing content, performing semantic searches, and enabling contextual natural language interactions. They can be combined with other built-in HeatWave features like AutoML to create more sophisticated applications.

5. **Contextual and Natural Language Processing**: Users can interact with enterprise data more intuitively, using natural language to query and retrieve information. The system maintains context through historical queries and document citations, enhancing the accuracy and relevance of responses.

6. **Seamless Integration with OCI Generative AI Service**: The in-database LLMs are integrated with Oracle Cloud Infrastructure's Generative AI service, which provides access to pre-trained models from leading LLM providers and further expands the range of generative AI capabilities available to users.

The new automated vector store in Oracle's HeatWave GenAI introduces several advanced features designed to streamline the integration and use of generative AI with enterprise data.

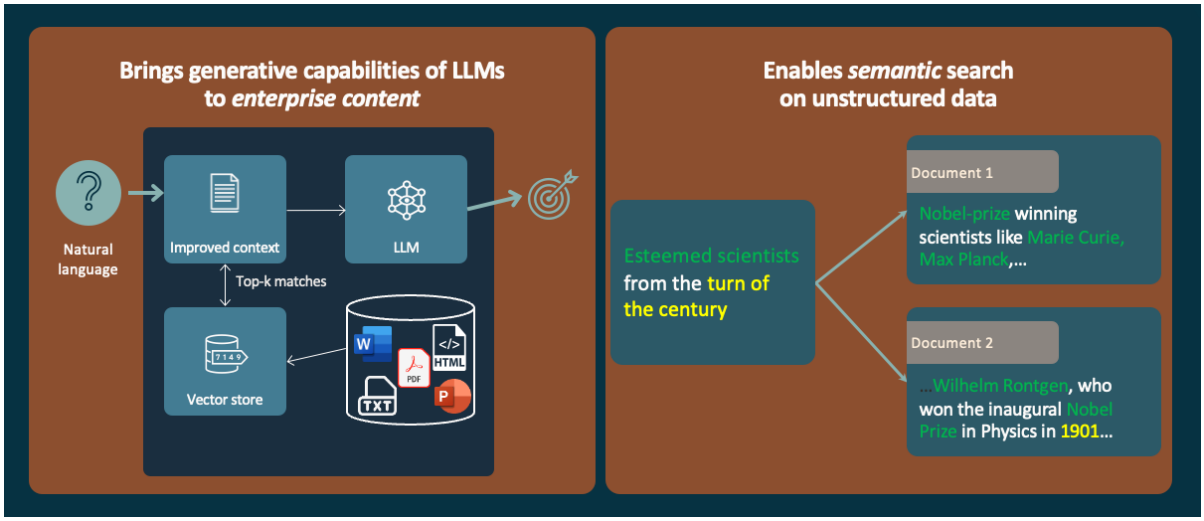

Figure 2. Oracle HeatWave Automated Vector Store (source: Oracle Corporation)

Here are the key aspects:

1. **In-Database Automation**: The vector store creation and management processes are fully automated within the database. This includes discovering documents in object storage, parsing them, generating vector embeddings, and inserting them into the vector store, all executed in a highly parallel and optimized manner.

2. **Ease of Use**: Users can create a vector store for enterprise unstructured content with a single SQL command using built-in embedding models. This significantly simplifies the process, reducing the need for specialized AI knowledge.

3. **Seamless Integration**: The automated vector store is built directly into the database, so there is no need to move data to a separate vector database. This integration ensures that data remains secure and consistent, eliminating the complexity and potential latency associated with data transfers.

4. **Efficiency and Performance**: The vector store utilizes HeatWave's extreme scale and performance capabilities. This allows for efficient vector embedding and retrieval operations without additional hardware, such as GPUs. The system's scale-out architecture and in-memory hybrid columnar representation enable rapid and accurate semantic searches.

5. **Enhanced Functionality**: The vector store supports retrieval-augmented generation (RAG), helping to solve the hallucination problem of large language models by allowing them to search proprietary data with appropriate context. This ensures more accurate and relevant responses from the AI models.

6. **Cost Savings**: By automating the vector store creation and embedding processes within the database, customers can avoid the additional costs typically associated with external AI infrastructure and specialized expertise. This leads to significant cost savings and reduced overall complexity of AI application development.

7. **Scalability**: The vector store's scale-out processing capabilities allow it to efficiently handle large volumes of data and queries. It can parallelize operations across up to 512 HeatWave nodes, ensuring users get rapid and reliable search results.

8. **Compatibility**: The vector store works seamlessly with various document formats, including PDF, PPT, Word, and HTML. This versatility makes it suitable for a wide range of enterprise applications and data types.

SCALE-OUT VECTOR PROCESSING

Oracle announced new scale-out vector processing capabilities in Oracle's HeatWave GenAI that deliver fast, efficient, and accurate semantic search and data processing within the database.

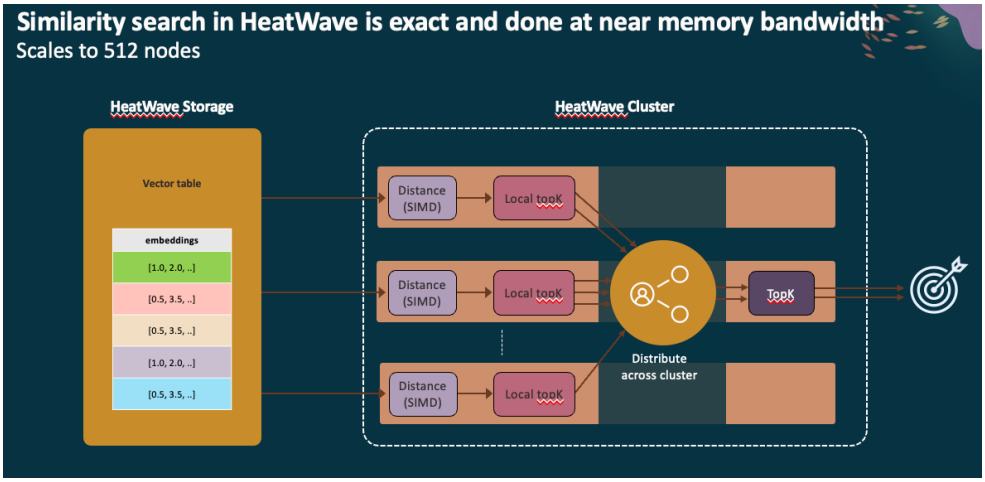

Figure 3. Oracle HeatWave Scale-Out Vector Processing (source: Oracle Corporation)

Here are the key features and benefits:

1. **Native VECTOR Data Type**: HeatWave introduces a new, native VECTOR data type optimized explicitly for vector processing, allowing for efficient storage and manipulation of vector embeddings directly within the database.

2. **Optimized Distance Function**: The system includes an optimized implementation of the distance function, which is crucial for performing accurate semantic queries. This function handles vector calculations efficiently, ensuring rapid processing times.

3. **In-Memory Hybrid Columnar Representation**: HeatWave uses an in-memory hybrid columnar representation for vector data. This leverages the speed of in-memory processing while maintaining the efficiency of columnar storage, enabling high-speed data access and manipulation.

4. **Scale-Out Architecture**: HeatWave's scale-out architecture allows vector processing tasks to be distributed across multiple nodes, up to 512 HeatWave nodes. Its parallel processing capability ensures that large-scale vector operations can be handled efficiently, significantly speeding up query response times.

5. **Near-Memory Bandwidth Execution**: The vector processing capabilities execute at near-memory bandwidth speeds. Data processing tasks are performed almost as quickly as the system's memory can handle, minimizing latency and maximizing throughput.

6. **High Parallelization**: HeatWave ensures that even complex and large-scale queries are completed quickly by parallelizing vector processing tasks across multiple nodes. This high degree of parallelization is key to maintaining performance as data volumes and query complexities increase.

7. **Semantic Search Integration**: The scale-out vector processing capabilities enable advanced semantic search functions. Users can perform semantic queries using standard SQL commands, which are processed efficiently thanks to the optimized vector handling.

8. **Combination with Other SQL Operators**: Users can combine semantic search with other SQL operators to perform complex queries. For example, they can join multiple tables with different documents and perform similarity searches across all documents, enhancing the richness and utility of search results.

9. **Benchmark Performance**: HeatWave GenAI's vector processing capabilities have been benchmarked to significantly outperform other systems. For instance, Oracle claims that it's 30 times faster than Snowflake and 15 times faster than Databricks while also being more cost-effective.

Oracle provided benchmarks showing HeatWave GenAI outperforms competitors like Snowflake, Databricks, and Google BigQuery in speed and cost efficiency. If these hold in real-world applications, Oracle HeatWave will be positioned as a formidable offering in the AI and data analytics market.

	HeatWave (1MySQL + 1 HeatWave node)	Snowflake on AWS (X-small)	Databricks on AWS (25 units+2xsmall)	Google BigQuery (100 slots)
Cost/hour	$1.52	$2	$9.6	$4
Total Time	16 sec	466 sec (30x)	238 sec (15x)	288 sec (18x)
Price-perf	$0.007	$0.259 (39x)	$0.637 (96x)	$0.32 (48x)

Figure 4. Vector Processing Performance Comparison (source: Oracle Corporation)

The scale-out vector processing capabilities in HeatWave GenAI provide a robust and high-performance solution for handling vector-based data operations. This enables enterprises to perform rapid and accurate semantic searches and other vector-related tasks directly within their database environment.

HeatWave Chat is a powerful new feature in Oracle's HeatWave GenAI that enhances user interaction with the database through natural language processing and intuitive graphical interfaces.

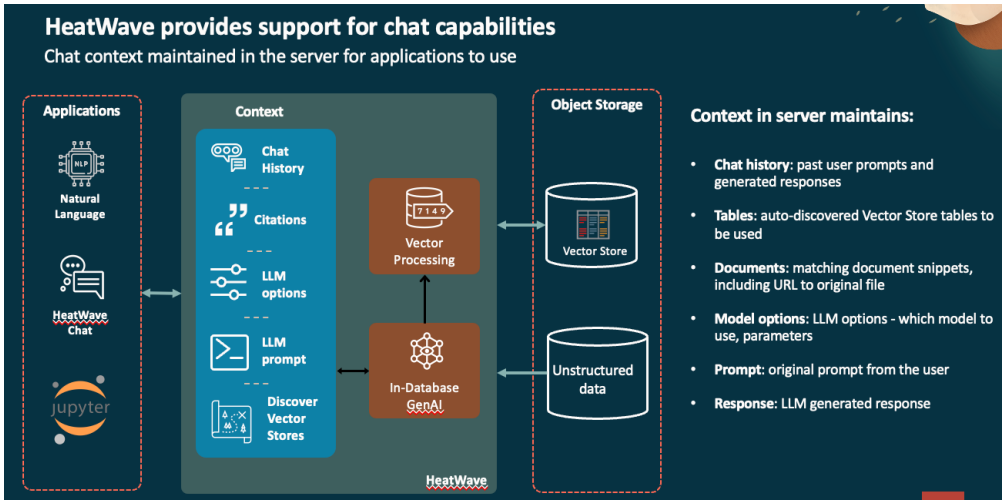

Figure 5. Oracle HeatWave Chat (source: Oracle Corporation)

Here are the critical aspects of HeatWave Chat:

1. **Visual Code Plug-In for MySQL Shell**: HeatWave Chat is integrated as a Visual Code plug-in within MySQL Shell, providing a graphical interface that allows developers to interact with the database quickly and efficiently.

2. **Natural Language and SQL Queries**: Developers can ask questions and perform queries in natural language or SQL, making it easier for users who may not be familiar with SQL syntax to extract information and interact with the database.

3. **Contextual Conversations**: HeatWave Chat maintains the context of the conversation, including the history of questions asked, citations of the source documents, and the prompts given to the LLM, enabling more accurate and relevant responses over the course of a conversation.

4. **Lakehouse Navigator Integration**: The integrated Lakehouse Navigator enables users to select files from object storage and create vector stores directly from within the graphical interface, simplifying the process of managing and utilizing unstructured data.

5. **Semantic Search Capabilities**: Users can perform semantic searches across the entire database or restrict searches to specific folders, allowing for more precise and context-aware data querying.

6. **Source Verification**: HeatWave Chat includes features that allow users to verify the sources of answers generated by the LLM. This transparency helps ensure the accuracy and reliability of the information provided by the AI.

7. **Context Maintenance for Applications**: The context maintained by HeatWave Chat is available to any application using HeatWave. This allows for seamless integration and continuity of data and query context across different applications and use cases.

8. **Efficiency and Productivity**: HeatWave Chat helps developers and users work more efficiently and productively by enabling natural language interactions and simplifying complex queries. This reduces the time and effort required to retrieve and analyze data.

HeatWave Chat provides a user-friendly interface for interacting with Oracle's HeatWave GenAI. It leverages natural language processing and contextual awareness to make data querying and management more intuitive and efficient. This feature enhances accessibility, improves user experience, and supports more accurate and relevant data interactions.

ANALYSIS

Oracle's HeatWave GenAI is a leap forward in integrating generative AI capabilities within enterprise data environments. Taken together, the new capabilities ensure that enterprises can leverage AI to gain insights and make data-driven decisions with unprecedented speed and accuracy.

While its use of in-database large language models, which streamline AI deployment by eliminating the need for external LLMs and additional infrastructure, is the standout feature, the other capabilities are equally compelling.

Oracle’s implementation of automated in-database vector store and scale-out vector processing in Heatwave GenAI allows users to perform high-speed, accurate semantic searches and contextual natural language interactions directly within the database; this could set a new standard for data processing performance and usability.

HeatWave Chat further enhances this offering by providing an intuitive graphical interface for natural language and SQL queries. This makes advanced AI accessible to a broader range of users, including those without specialized technical skills. Integrating with the Lakehouse Navigator and the maintenance of query context improves the user experience and the reliability of AI-generated responses.

In summary, HeatWave GenAI is a game-changer for enterprise AI, delivering high performance, ease of use, and cost savings. Its integrated, automated approach democratizes access to generative AI, enabling organizations to build richer applications and achieve significant productivity gains. Oracle's innovative strides with HeatWave GenAI will likely set new benchmarks and drive broader AI adoption across industries.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.