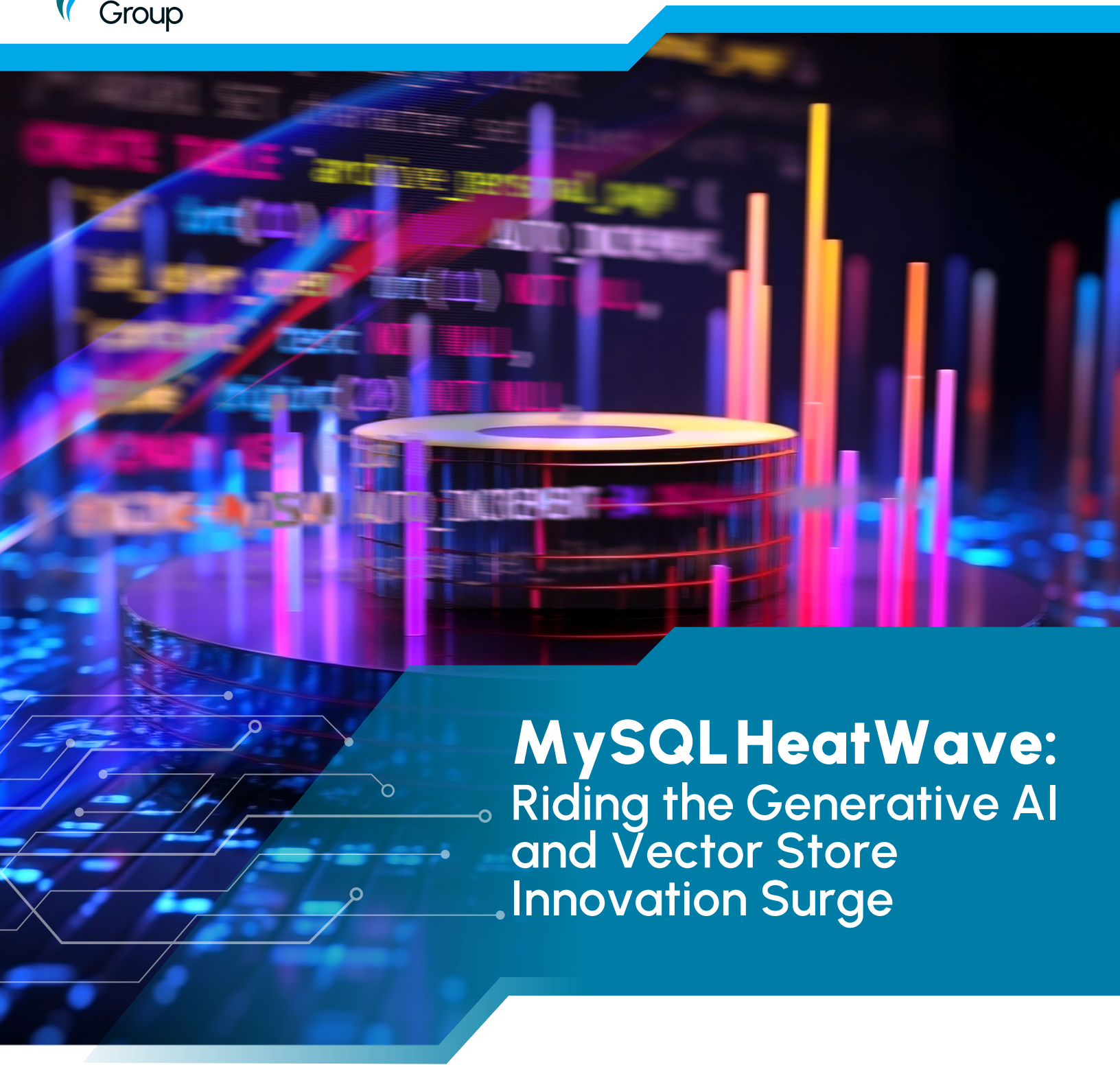
The Futurum Group



MySQL HeatWave: Riding the Generative AI and Vector Store Innovation Surge

AUTHOR **Ron Westfall**
Research Director | The Futurum Group

IN PARTNERSHIP WITH

ORACLE

NOVEMBER 2023

Oracle Propels MySQL HeatWave Innovation Momentum With New Generative AI and Vector Store Features.

MySQL HeatWave Increases Ease of Use for Customers Through Generative AI, Vector Store, AutoML, and Lakehouse Enhancements Plus JSON Acceleration and JavaScript Support.

Introduction: How to Best Meet Today's Database Generative AI Challenges

Oracle has unveiled substantial enhancements to **MySQL HeatWave**, the company's fully managed MySQL database service, powered by the HeatWave in-memory query accelerator. The new enhancements encompass support for vector store, generative AI, new HeatWave Lakehouse capabilities, new in-database machine learning (ML) features, MySQL Autopilot enhancements, support for JavaScript, acceleration of JSON queries, and support for new analytical operators. Oracle's new enhancements fully address the brave new uncertainties organizations are facing in adopting generative AI and vector store capabilities, especially in easing the adoption of emerging generative AI technology and ensuring optimization of business outcomes throughout their AI/ML journeys.

In direct parallel, organizations are prioritizing the best path for unifying their customer data, analytics, and AI capabilities, especially across lakehouse environments. In addition, they are most keen on avoiding the difficulties of moving their data to separate services - whether for transaction processing, analytics, ML, AI, or lakehouse — and gaining the thorough efficiency and productivity benefits of integration that MySQL HeatWave offers.

For context, MySQL HeatWave is a fully managed cloud database service, powered by the HeatWave in-memory query accelerator. We find that it is the only cloud service that combines transactions, real-time analytics across data warehouses and data lakes, and ML in one MySQL database – all without the complexity, latency, risks, and cost of extract, transform, load (ETL) duplication.

With the enhancements, customers will be able to use the vector store (currently in private preview) to combine the power of large language models (LLMs) with their proprietary data and obtain answers that are more accurate than using models that have been trained on public data only. Through generative AI and vector store capabilities, customers will interact with MySQL HeatWave in natural language and efficiently search documents across various file formats in HeatWave Lakehouse. This ushers major breakthroughs for MySQL HeatWave in securely advancing customers' organization-wide AI journeys.

Executive Summary

- Oracle has unveiled substantial enhancements to MySQL HeatWave, including support for vector store, generative AI, new HeatWave Lakehouse capabilities, new in-database ML features, MySQL Autopilot enhancements, support for JavaScript, acceleration of JSON queries, and support for new analytical operators.
- The new MySQL HeatWave generative AI and vector store enhancements are targeted directly at meeting the challenges organizations face in making sure that their unfolding AI journeys are proceeding on the right track.
- MySQL HeatWave has also demonstrated unmatched lakehouse performance and price-performance because of its scale-out architecture that leverages AMD EPYC™ processors.
- The HeatWave AutoML support for HeatWave Lakehouse delivers swift training, inference, and explanation benefits on data in object store to customers.
- The HeatWave AutoML differentiation is further enhanced by new Bayesian personalized ranking enhancements to its recommender system.
- Other enhancements that further strengthen the overall value of MySQL HeatWave now include that HeatWave is accelerating JSON, enabling faster query processing and real-time analytics.

Oracle MySQL HeatWave Generative AI and Vector Store Breakthroughs

The new MySQL HeatWave generative AI and vector store enhancements are targeted directly at meeting the challenges organizations face in making sure that their unfolding AI journeys are proceeding on the right track. Key to locking down a sound foundation is ensuring that users can query and retrieve information in natural language. This functionality includes the built-in efficient searching of documents in various formats in HeatWave Lakehouse.

As a result, we see customer use cases for generative AI expanding across critical capabilities and key verticals such as digital marketing, gaming, healthcare, and fintech. Such use case capabilities include:

- **Zero-Shot Classification**. Organizations can easily classify customer queries to the relevant departments, perform sentiment analysis for customer reviews and content, and streamline product categorization.

- **Natural Language to SQL.** With automated SQL generation from natural language queries, customers can access information in both the data lake and the database/data warehouse.

- **Summarization.** Log summarization accelerates root cause analysis with document summarization features that speed up research and development missions. Summarization of customer interactions is an additional key benefit.

- **Retrieval.** Customers can conduct question-answering on public and proprietary data using the vector store plus access intelligent context-aware applications.

HeatWave Lakehouse Capabilities Gain Tremendous Boost

We view the enhancements augmenting HeatWave Lakehouse capabilities, now generally available on **Oracle Cloud Infrastructure** (OCI) and in limited availability on **Amazon Web Services (AWS)**, as delivering a tremendous portfolio development boost to the overall Oracle data lakehouse proposition. For clarity, we define a data lakehouse as a modern, open architecture that stores, understands, and analyzes all data.

MySQL HeatWave is optimized for AMD EPYC processors on OCI, scaling to 512 nodes across the large number of cores available on EPYC CPUs. The MySQL HeatWave code takes advantage of various system resources on EPYC processors, including compute, large L3 cache, DRAM bandwidth, input/output (I/O), and cache lookup to achieve the best performance and price-performance in the industry for data warehouse and lakehouse workloads.

In a **500 TB TPC-H benchmark**, the query performance of MySQL HeatWave Lakehouse powered by AMD EPYC processors on OCI is:

- **9x** faster than Amazon Redshift, delivering **8x** better price performance
- **17x** faster than Databricks, delivering **18x** better price performance
- **17x** faster than Snowflake, delivering **22x** better price performance
- **36x** faster than Google BigQuery, delivering **30x** better price performance

Now, with its support for generative AI, users can interact with MySQL HeatWave in natural language. Both user queries and the response from the system are generated in natural language using an LLM. LLMs are trained on public data, and for organizations planning to use LLM capabilities to answer user questions, the results can prove erroneous due to the hallucination problems of LLMs, and/or because of their lack of enterprise knowledge. We find that the introduction of a vector store in MySQL HeatWave is indispensable in alleviating such difficulties.

The vector store uses a language encoder to create vector embeddings from documents in HeatWave Lakehouse that can be stored in an array of formats such as ppt, pdf, and text. It also considers the question asked by the user to create vector embeddings and does a similarity search in an n-dimensional space. The output of the vector store is the additional context gleaned from internal documents, which is included along with the users' question in the prompt as inputs to the LLM. The LLM uses this information to generate a response that will then include proprietary information from the documents in MySQL HeatWave Lakehouse. This feature allows enterprises to receive more relevant and precise answers to their queries.

Moreover, MySQL HeatWave Lakehouse capabilities are now further strengthened by support for the Apache Avro format. Oracle is fulfilling fast-growing customer demand for Avro file format support, as it is the leading serialization format for record data and data pipelines. Avro is a row-oriented format that uses compact, binary data types and has embedded schema metadata, which fuels its expanding popularity.

Of key importance, HeatWave Lakehouse supports different compression algorithms that provide uncompressed, snappy, and deflate codecs for Avro files. The high query and load performance of HeatWave Lakehouse is identical across all supported file formats.

MySQL Autopilot also supports the Avro file format, enabling customers to automatically infer schema from source Avro files and present it as an easy-to-use 'create table' SQL statement. Moreover, MySQL Autopilot automatically estimates the optimal size of the HeatWave cluster for a given workload, and estimates the time required to load Avro and other supported files such as Parquet and CSV.

HeatWave AutoML Turbocharges HeatWave Lakehouse Training, Inference, and Explanations

Integral to the HeatWave Lakehouse vision is making sure that customers can query half a petabyte of data in object storage and leverage all the benefits of HeatWave even when their data is stored outside a MySQL database. With HeatWave AutoML, developers and data analysts can build, train, deploy, and explain ML models in MySQL HeatWave without moving data to a separate ML service.

The HeatWave AutoML support for HeatWave Lakehouse delivers swift training, inference, and explanation benefits on data in the object store to customers. Such benefits include enabling models to be kept up to date frequently, in-database simplicity, fully automated training, explainable outputs, and no extra cost.

Now text processing with HeatWave AutoML is available, enabling users to perform ML tasks on text columns. We discern that support for TFIDF and BERT further crystallizes the relevance of words in documents and a better understanding of the context of words in a sentence. As a result, these models generate embedding, which provides critical input to AutoML, meeting the expanding demand for ML-driven text column capabilities. HeatWave AutoML provides sharp competitive advantages, as demonstrated by the ability to perform 25x faster machine learning model training than Amazon Redshift ML.

We identify additional key advantages, including the fact that HeatWave AutoML is built into the database, whereas Redshift ML is not, and Snowflake's Snowpark ML is only in preview mode for the same capability. This approach directly aligns with the simplicity of combining five AWS services into one with MySQL HeatWave, contrasting sharply with the Redshift ML requirement for more complex integration across multiple services, and Snowpark ML, which compels data scientists to use a third-party ML library.

Plus, HeatWave AutoML natively supports time-series forecasting, which we find is a considerable advantage over Redshift ML's complete lack of built-in forecasting capabilities, and Snowpark obliging users to rely on scikit-learn. Notably, HeatWave AutoML supports native anomaly detection while Redshift ML's anomaly detection is only in development using more complex clustering techniques. Snowpark ML requires data scientists to turn to the scikit-learn library to perform anomaly detection.

HeatWave AutoML vs. Snowpark ML: Summation of Competitive Advantages

We believe that the competitive advantages of HeatWave AutoML over Snowpark ML are important. In summary, Snowpark ML liabilities include the following:

- To gain systemwide ML benefits, Snowpark ML requires data scientist skills to access scikit-learn, adding cost and mplexity.
- The Snowpark ML model training steps are manual. There are no built-in automation benefits.
- The process to set up parameters to start model training takes at least 45 minutes, leading to diminished efficiency and productivity.

New Bayesian Personalized Ranking Gems in HeatWave AutoML's Recommender System

From our perspective, the HeatWave AutoML differentiation is further enhanced by new Bayesian personalized ranking enhancements to its recommender system.

The HeatWave AutoML recommender system is a built-in recommendation engine that uses ML algorithms to match relevant products or content to the target audience. It now supports a broader array of feedback, including implicit feedback—such as past purchases and browsing history—and explicit feedback, such as ratings and likes to produce more accurate personalized recommendations.

By accessing tailored recommendations, MySQL HeatWave customers can further refine their organization's personalization strategies to optimize CX, drive engagement, and boost revenue. As a result, HeatWave AutoML delivers more precise predictions. Some examples include highlighting items for sale or content that a user will like, pinpointing users who will like a specific item, and the ratings an item will receive. It can enable organizations to achieve greater ad spend or marketing ROI by identifying similar users given a specific user profile, and similar items given a specific item.

The new Bayesian personalized ranking capabilities further differentiate HeatWave AutoML, as can be seen in the following comparison with other systems:

	RedShift ML	Azure AutoML	BigQuery ML	HeatWave AutoML
Rec sys support	✗	✗	✓	✓
Explicit feedback	✗	✗	✓	✓
Implicit Feedback	✗	✗	✓	✓
Recommend items to users?	✗	✗	✓	✓
Recommend users for an item?	✗	✗	✗	✓
Recommend similar users?	✗	✗	✗	✓
Recommend similar items?	✗	✗	✗	✓
Content based	✗	✗	✗	✗

HeatWave AutoML especially shines in comparison to Redshift ML, Azure ML, and Google BigQuery ML across the recommend users for an item, recommend similar users, and recommend similar items categories. Only HeatWave AutoML supports these capabilities, further strengthening overall competitive differentiation.

MySQL Autopilot Strengths and New Enhancements

MySQL Autopilot provides workload-aware, ML-powered automation of various aspects of the application lifecycle, including provisioning, data loading, query execution, and failure handling. Additionally, it provides capabilities for online transaction processing (OLTP) workloads. Within the interactive console (available on AWS), users can access the MySQL Autopilot Auto Shape Prediction advisor, which continuously monitors the OLTP workload to recommend the most appropriate compute shape for best price-performance at any given time.

The visual representation within the console makes it easy for database users to upsize or downsize their database shape. Recommendations are supported by the visual analyses of historic performance trends, including buffer pool hit rate and throughput. We find these features are vital for enabling customers to attain the best price-performance outcomes.

New MySQL Autopilot capabilities include:

- **Adaptive Query Execution in HeatWave.** Dynamically adjusts data structures and system resources after query execution has started, and independently optimizes query execution for each node based on the actual data distribution at run time. This functionality directly results in improved ad hoc query performance by up to 25%, and skew handling.

- **Auto Load and Unload.** Automatically loads tables into HeatWave according to user workloads, and automatically unloads tables that are rarely or never queried to decrease memory costs. Accordingly, developers are liberated from manually loading and unloading tables, improving productivity.

- **Auto Column Compression.** Introduces multiple compression algorithms as well as HeatWave selection of optimal compression algorithms for each column considering compression ratio and performance. This feature improves load and query performance due to faster compression and decompression. Additionally, the higher compression ratio reduces memory usage up to 25%. Users also gain the major benefit of column compression optimization according to workload characteristics.

- **MySQL Autopilot Indexing.** Uses ML to recommend secondary indexes for OLTP workloads, considering both query and data manipulation language (DML) performance, and recommends the CREATE/DROP functions of indexes. Generation of data definition languages (DDLs) is also included for index creation/drop. Customers can see the overall workload throughput advances, storage impact, and per query latency improvement without creating the indexes and without compute or storage overhead. This capability is currently in limited availability. Once generally available, it will help customers eliminate the time-consuming tasks of creating and maintaining optimal indexes for their OLTP workloads—while helping them to ensure the best performance and the lowest storage cost.

Additional MySQL HeatWave New Enhancements = More Benefits

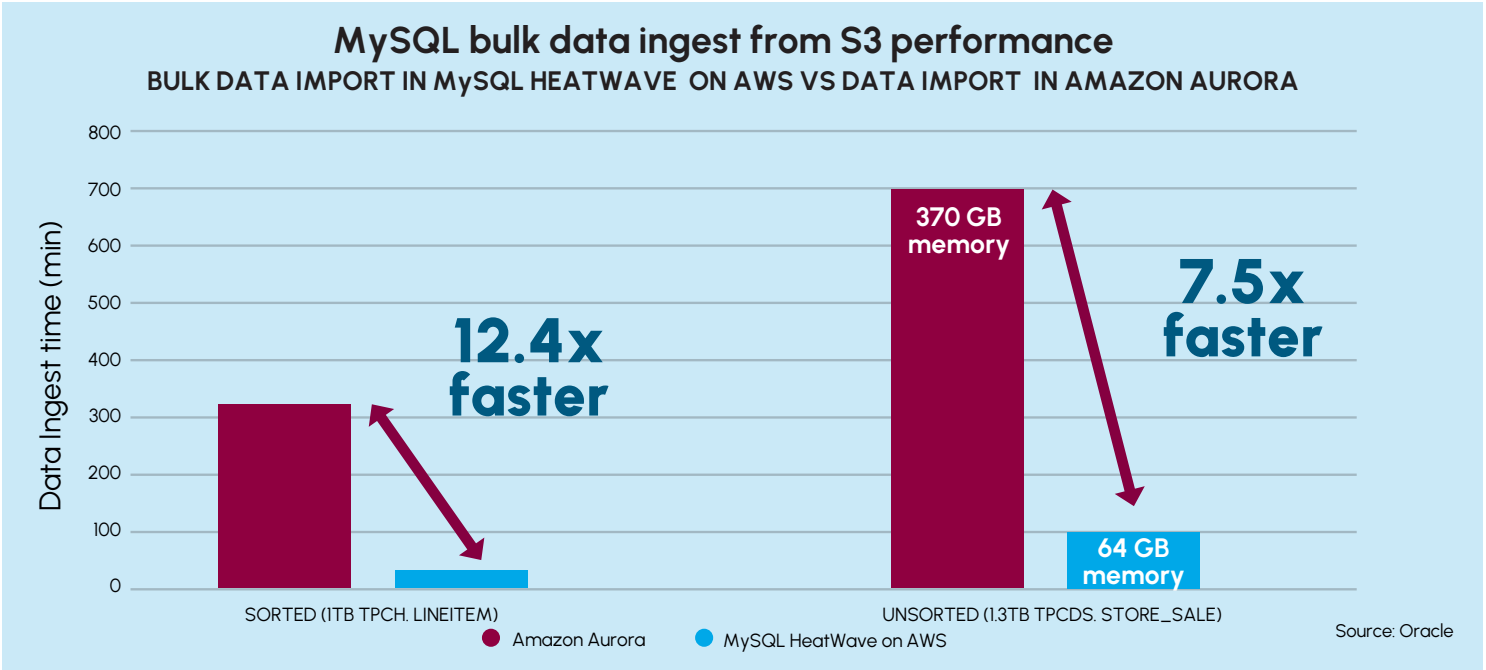
In addition to the compelling set of new generative AI/vector store, AutoML, and Autopilot capabilities now offered in MySQL HeatWave, we would be remiss not to highlight the other enhancements that robustly strengthen the overall value of MySQL HeatWave. Now, HeatWave is accelerating JSON, enabling faster query processing and real-time analytics on JSON documents. In MySQL Database, it is labor-intensive to optimize JSON; you must extract JSON data into a virtual column and then build secondary indexes in the virtual column. However, with this new MySQL enhancement, no indexes are needed, propelling faster query processing times.

We see the introduction of new analytics operators in MySQL HeatWave (CUBE, Hyper Log Log, Qualify, and Table sample) as very important to facilitate the migration of non-MySQL workloads. This introduction aligns with the support of SET operations by MySQL and HeatWave, which we find delivers faster performance compared with other analytics services. Moreover, we identify the new partitioning support in MySQL HeatWave as essential to improving scalability.

For the new JavaScript support, JavaScript stored programs ensure seamless MySQL-to-JavaScript type conversion for I/O arguments. JavaScript stored programs can be used anywhere SQL stored functions can be used, e.g., SELECT, WHERE, and ORDER BY in parallel with support for DML, DDL, and Views. Existing XDev application programming interfaces (APIs) can be used to execute SQL inside JavaScript including support for prepared statements, transactions, and session states. As a result, JavaScript stored programs are now first-class objects in MySQL HeatWave, simplifying the execution of complex operations. With the inclusion of the JavaScript for MySQL HeatWave feature, developers can now express complex programming logic directly inside the MySQL server. This feature allows developers to push data-intensive parts of their applications close to their data, reducing data movement cost. This new capability is currently in limited availability.

For OLTP, new bulk ingest capabilities in MySQL HeatWave provide parallel sort and merge and parallel build of index sub-trees as well as sequential writes of sorted data into disk. This eliminates random disk I/O, enabling pipelining of internal stages that overlap compute with disk I/O.

From our view, MySQL HeatWave bulk ingest performance further amplifies the competitive advantages of its database service proposition. Data can be queried sooner, and the system resources used for loading data are freed up much faster, lowering costs for customers. Strikingly, MySQL HeatWave is tenfold faster than Amazon Aurora while using less memory as validated in the following comparison chart:



Conclusion and Recommendations

In summary, we believe that the new generative AI and vector store enhancements in MySQL HeatWave deliver remarkable improvements and enhance its competitive advantages. So do the new lakehouse, AutoML, Autopilot, analytics, JSON acceleration, JavaScript, and bulk ingest capabilities. Now MySQL HeatWave is even better positioned to meet data, analytics, and AI priorities in a single cloud database service. Vector store and generative AI bring the vigor of LLMs to customers, providing them with an intuitive way to interact and the ability to combine external and internal data to get the accurate answers that they need to boost their business outcomes.

Overall, the innovation train keeps accelerating faster for the MySQL HeatWave offering through the raft of new enhancements and innovations. A new vector store and support for LLMs of choice enable users to interact with MySQL HeatWave in natural language. Plus, customers can use HeatWave AutoML to perform ML operations on data loaded directly from the Lakehouse; JavaScript stored programs execute in the MySQL database improving performance, and with JSON acceleration in HeatWave, queries can run up to orders of magnitude faster.

For its competitors such as Redshift, Aurora, Snowflake, Databricks, and Big Query, the heat is on as we see that customers have numerous more reasons to make the switch to HeatWave.

The key takeaways on the new MySQL HeatWave enhancements for IT decision-makers include:

- **Vector Store Makes Generative AI Valuable to Business.** Organizations can augment the power of LLMs with their proprietary data, getting more valuable answers and alleviating LLM difficulties such as hallucinations.

- **MySQL HeatWave Lakehouse Capabilities Strengthen.** HeatWave Lakehouse capabilities are now further strengthened by support for the Apache Avro format. Customers can significantly reduce complexity and costs by replacing five AWS services with one – running transaction processing, real-time analytics across data warehouses and data lakes, and ML in MySQL HeatWave.

- **HeatWave AutoML Delivers Major Training, Inference, and Explanation Gains.** The HeatWave AutoML for HeatWave Lakehouse delivers immediate training, inference, and explanation benefits on data in the object store. Such benefits include enabling models to be re-trained frequently, in-database simplicity alleviating the need to move data to a separate ML service, fully automated training, and explainable outputs all at no extra cost.

Important Information About this Report

CONTRIBUTORS

Ron Westfall
Research Director | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ORACLE®

ABOUT ORACLE

Oracle is a leading global technology company specializing in cloud computing, database management, and enterprise software solutions. With a rich history of innovation, Oracle empowers businesses of all sizes to streamline their operations, optimize data management, and harness the power of cutting-edge technologies, ensuring they remain at the forefront of the digital age.Learn more here.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION

The Futurum Group LLC | futurumgroup.com | (833) 722-5337