Research Note

BY RICHARD WINTER

ADVANCES IN MACHINE LEARNING WITH MYSQL *HEATWAVE*

Open Source Cloud Database Service Now Provides In-Database Machine Learning, Elastic Scaling and Lower Entry Price

ORACLE HAS BEEN ADDING MAJOR NEW CAPABILITIES to *MySQL HeatWave,* its open source cloud database service, at a striking pace. In fifteen months, we have seen an integrated, in-memory query acceleration engine for both OLAP and OLTP; Autopilot for automated management; and now in-database machine learning, elastic scaling, and a lower price entry point. In addition, Oracle has released query processing and machine learning benchmarks that show superior performance and price/performance over other popular cloud database offerings. Customers with an interest in an open source cloud database with advanced integrated functionality really ought to take a serious look at *MySQL HeatWave*.

Released in Dec. 2020, *MySQL HeatWave* is the first MySQL implementation with a built-in, highly parallel query accelerator, making it a viable option for data warehouses on Oracle Cloud Infrastructure (OCI), as well for the transaction processing workloads it has long handled.

In 10 TB TPC-DS benchmark results recently released by Oracle, *MySQL HeatWave* demonstrates a large performance and price/performance advantage against *Amazon Redshift*, *Amazon Aurora* and *Snowflake* on *AWS* (test results and code are available here).

IN DECEMBER 2020, Oracle extended MySQL in remarkably significant ways, giving it a highly parallel architecture with the ability to process data warehouse queries on a database also handling near real-time OLTP transactions at the same time. In Aug. 2021 Oracle introduced a feature called *MySQL Autopilot* that learns from the behavior of its workload to enhance various MySQL operations like data load queries. These changes made MySQL suitable for many applications it could not have handled well in the past, including mixed workloads and analytic applications. These changes positioned the open source cloud database service, now known as *MySQL HeatWave*, with significant advantages over the alternatives.

Now, less than a year later, Oracle has once again released major enhancements to *MySQL HeatWave*, addressing some key issues for customers in today's market: machine learning, elastic scaling, and a lower price entry point. In addition, Oracle has released new proofs of the performance and cost advantages of *MySQL HeatWave*.

WinterCorp
www.wintercorp.com

SPONSORED RESEARCH PROGRAM

Methodology

PURPOSE AND METHODOLOGY FOR THIS REPORT

This *WinterCorp Research Note* covers the recent announcement of *Oracle MySQL HeatWave* and its implications for database customers. In developing this report, WinterCorp drew on its own independent research and experience, interviewed employees, attended Oracle events and analyzed Oracle documentation and literature. Oracle was provided an opportunity to comment on the paper with respect to facts, in its capacity as the sponsor of this research. WinterCorp has final editorial control over the content of this publication and is solely responsible for any opinions expressed.

Machine Learning

With most databases, customers who want to apply machine learning on the data in the database must extract the data from the database and employ special machine learning tools and techniques in a separate environment. Often scalability is an issue. Further, the data extraction is not a one-time process: data is often extracted initially for experimentation and training models and then later, on a larger scale, for production execution of the models. In production use, data may be extracted in periodic batch runs or it may be re-trieved constantly to support online processes. The result is a time-consuming and error-prone development environment and a complex and often fragile operational architecture with security gaps. The copying of data out of, and back into, the database can itself lead to security and privacy problems.

It is much more desirable to do the machine learning in place, within the database, with a scale-out architecture such as that provided by *MySQL HeatWave*. In this approach, data scalability is built-in. Further, development, experimentation, training and production are all simpler. End-to-end development and deployment times are reduced.

The newest version of *MySQL HeatWave* offers such in-database machine learning with the added benefit of a high degree of automation. The user defines the prediction to be made and data to be used in making it. The system selects the prediction algorithm and generates the model, tunes it and seeks to select the optimal features and hyper parameter values. In a demo, Oracle shows how this machine learning capability can automatically find the factors in the U.S. Census database that best predict who will earn an annual income of more than $50,000. The system not only generates the prediction but will explain it, on request. Explainability has become important not only for regulatory requirements — why a credit card was approved or rejected, for example — but also to provide a higher degree of trust in machine learning for its users in all applications.

This level of automation is intended to enable the use of machine learning by business analysts without requiring that they have special data science skills. It also has the potential to enable customers to deploy machine learning models throughout their business processes at a much higher rate than was previously practical.

Online Elastic Scaling

Online Elastic scaling lets the customer increase or decrease the number of nodes in the *MySQL HeatWave* cluster while the system is running. After the customer specifies the new number of nodes, the system makes all the necessary changes automatically. While it does so, queries and updates continue to run as usual. There is a short pause, which Oracle says will be in the tens of microseconds, for updating the in-memory partition maps.

The usual approach to changing the number of nodes in a scale-out cluster involves a "shuffle" — a parallel redistribution of data across the nodes. *MySQL HeatWave* avoids the more costly and disruptive shuffle of data between nodes by instead loading the new nodes in parallel from object storage. This operation employs scale-out parallelism; thus, loading ten nodes takes approximately the same elapsed time as reloading, say, two nodes.

About WinterCorp

WinterCorp is an independent consulting firm, expert in the strategy, architecture and scalability of the modern analytic data ecosystem.

Since our founding in 1992, we have architected and engineered solutions to some of the toughest and most demanding analytic data challenges, worldwide.

We help customers define their data-related business interests; develop their data strategies and architectures; select their data platforms; and, engineer their solutions to optimize business value.

Our customers get business results with analytics in which their return is often ten or more times their investment.

When needed, we create and conduct benchmarks, proofs-of-concept, pilot programs and system engineering studies that help our clients manage profound technical risks, control costs and reach business goals.

With our in-depth knowledge and experience, we deliver unmatched insight into the issues that impede scalability and into the technologies and practices that enable business success.

WinterCorp
www.wintercorp.com
TYNGSBORO, MA
617-695-1800

This elastic scaling lets the customer increase capacity to handle a peak workload or to accelerate the performance of large queries, which would take too long on a small cluster. At the same time, the customer incurs the expense of running a large cluster only when it is really needed. A distinctive feature is that the number of nodes in use can be any number between 2 and 64. A demo available on YouTube here shows the number of nodes being increased from 5 nodes to 29 nodes and then later decreased to 11 nodes.

Some other systems do not offer this flexibility; for example, with *Snowflake* the number of CPUs need to be a power of 2. A *Snowflake* customer with 16 CPUs who wanted more capacity, say 17 CPUs, would need to go to 32 CPUs doubling the per minute cost and getting unneeded capacity. By contrast, a *MySQL HeatWave* customer could go from 16 cpus to 17 cpus — or any other number up to 64 — to provide just the right amount of needed capacity. So, the *MySQL HeatWave* customer can meet the temporary demand while saving on cost. By analogy, think of the situation where you want a sandwich for lunch, but your only option is to buy a foot-long sub. You pay too much, and you have no way of using the large amount you have been forced to buy.

Performance and Price/Performance

Oracle has published TPC-DS benchmark data at 10TB scale showing that *MySQL HeatWave* delivers much higher performance and much better price performance than comparable cloud database services offered by other leading cloud database vendors, such as *Amazon Redshift*, *Google BigQuery*, and *Snowflake*. When viewed through the lens of modern data warehouse requirements, TPC-DS is more realistic than the older TPC-H test, in that it features a wider variety of queries; skewed data; and other characteristics.

While WinterCorp has not independently validated these benchmarks, they do seem to indicate very large advantages in both performance and cost of operation. Even the best, independently audited, standard benchmarks are only suggestive of what customers might experience with their own data and queries, so customers should always interpret standard benchmarks with caution.

While TPC-DS represents a step forward, it is not representative of all the requirements that a high-end data warehouse must satisfy in practice. Still, it shows *MySQL HeatWave* in a favorable light performing some of the types of data warehouse queries that many customers run every day. And, to Oracle's credit, customers can validate the benchmarks on their own time, as all the code, scripts and configurations are publicly available on *GitHub*, for both *HeatWave* ML and TPC-DS. Lastly, no competitor has yet to publicly challenge any of the *MySQL HeatWave* benchmarks that have been published over the last 15 months.

With the fully transparent publication of these tests, Oracle has quantified its claims in three separate announcements for performance and price performance and provided publicly available evidence to support them. It is worth noting that the advantages Oracle shows in these tests are rather large. Customers considering *MySQL HeatWave* should, however, perform their own evaluations and tests before relying on any particular level of performance or cost savings for their applications and workloads.

Recommendation

CUSTOMERS WITH AN INTEREST in open-source database cloud solutions, including data warehouse solutions, ought to take a close look at *MySQL HeatWave*. This is MySQL in a way we have never seen it before, with an integrated high-performance query accelerator; machine learning based automation of the cloud service; in database machine learning support; online elastic scaling; and, a lower price entry point. Here customers gain the benefits of open source with vendor backed support integrated into Oracle Cloud Infrastructure with its high security, high availability, and high performance database operations. On the basis of the *MySQL HeatWave* architecture alone, no other MySQL option currently available comes close.

I don't recommend *MySQL HeatWave* for enterprise-scale or mission-critical data warehouse requirements for which one would currently use Oracle Database. However, *MySQL HeatWave* is a clearly superior solution for open-source database users looking for cost-effective, high performance for their application-driven data warehouses with built-in machine learning and elastic scaling, while avoiding the headaches of managing separate databases for transaction processing and analytics.

This announcement puts *MySQL HeatWave* much further ahead of comparable offerings on *AWS*. The closest alternatives would be the combination of *AWS Aurora* and *Redshift* or the combination of *AWS Aurora* and *Snowflake*. Both of these combined offerings would also require a host of chargeable *AWS* ETL tools and services to provide the capabilities built into *MySQL HeatWave*. In the case of *MySQL HeatWave*, one database is clearly better than two. ●