Oracle Database@Azure redefines multi-cloud

Oracle CloudWorld 2023 trip report

Executive Summary

Trigger

The core themes from Oracle's recently concluded annual CloudWorld event centered on multicloud, multimodel database, low-code/no-code AppDev, and of course, Generative AI (GenAI). But the headline, the announcement that Oracle Cloud Infrastructure (OCI) services for Oracle Database (including Exadata Database Service and Autonomous Database) would now also be physically deployed in the Microsoft Azure Cloud, upstaged the data and GenAI announcements that came during CloudWorld. Oracle remains committed to the growth of Oracle Cloud regions (64 currently), but the Azure announcement adds new possibilities raising the question of where future Oracle Cloud Infrastructure (OCI) regions will turn up.

Our Take

The Oracle Database@Azure announcement proved a welcome respite from the "All-GenAI all-the-time" nonstop narrative that's been droning on since the spring, as documented in our recently published GenAI Trip Report. The Oracle-Microsoft Azure announcement represents a major turning point for Oracle Cloud that has major implications for Oracle Cloud's future, and adds a new twist in the discussion of what multi-cloud means.

Not that CloudWorld was short of GenAI announcements – Oracle placed its own stamp on GenAI and added nuance that we haven't heard so far, and they have put skin in the game with a minority investment in Cohere. One of the most popular enterprise use cases for GenAI is code generation, and where it came to SQL, Oracle became one of the first to add nuance: SQL generated from a large language model (LLM) is not likely to be perfect the first time around, and we'll still need humans in the loop. For vector indexing, Oracle provided far more detail than what we've heard from others so far with an offering that is *not* one-sized-fits-all. Not surprisingly, Oracle Database and MySQL HeatWave are adding vector storage and search support, although the plans for HeatWave were not fully disclosed. LLMs are also being used to autogenerate application blueprints for Oracle APEX no code/low code development tooling, which is getting newfound prominence as Oracle's new go-to programming language for upcoming Oracle SaaS functionality, with the re-write of Cerner's (now Oracle Health) applications being Exhibit One. (We'll do a deeper dive on the evolution of APEX in Oracle Health in an upcoming report.)

As for database, we saw a logical expansion of multimodel (or what Oracle terms, "converged") database support with new JSON Relational Duality and property graph views that will increase ease of working with non-relational data. As for MySQL HeatWave, the highlight was addition of a vector store for GenAI – a capability that is rapidly becoming a checkbox feature for all operational databases. There were also incremental enhancements to HeatWave building on existing capabilities (e.g., AutoML to Lakehouse, Lakehouse on AWS,

JSON acceleration, and autopilot indexing automation added to OLTP). As for Autonomous Database, it is entering a broader stage as Oracle's enterprise application SaaS services Fusion ERP and NetSuite are now ingraining it into their services.

Oracle Database goes native in Azure

As noted, announcement that Oracle database services will run inside the Azure cloud stole the show. Looking back, it's a logical outgrowth of an evolving relationship between two cloud providers who have embraced coopetition.

The back story was the unexpected 2019 announcement of the high-speed interconnect between OCI and Azure; both companies were moved by presence of a critical mass of joint customers using Microsoft 365 and Oracle database. It was a sea change for Oracle which, until then, was not pushing a multicloud strategy. To date, the interconnect has drawn 500 customers running across a dozen regions. The next step a couple years later was enabling Oracle database services, still running physically in OCI, but managed from Azure with the Azure console – making it *look like* an Azure native service and with identity federation auto-configured.

The new announcement closes the loop; now all Oracle Database services on OCI (Exadata platform, Autonomous Database, Exadata Database Service, and Base Database) will add the option for physically and *running inside* Azure data centers, and made available as Azure services through the Azure marketplace. There, joint customers can buy Oracle database services with Azure cloud credits. As in the Oracle Cloud, Oracle operates and manage OCI services directly within Microsoft's datacenters globally.

From a technology standpoint, the advance is both trivial and profound. From a geolocation perspective, the move is fairly inconsequential because the interconnects typically involved OCI instances situated either within the same campus or physical building. In some cases, OCI instances are moving just a few feet into the next room. But from an infrastructure perspective, the move is quite stark. It drops the equivalent of an Oracle Exadata Cloud Infrastructure rack literally inside the Azure data center. In Azure, Oracle databases will still operate on the same flat topology and RDMA architecture as they do inside an OCI region.

For Oracle, deploying OCI instances physically outside OCI data centers is nothing new. It has offered Cloud@Customer and Dedicated Region, which place Oracle-managed OCI racks inside customer data centers, for several years. For Microsoft, this was just a matter of dropping another bare metal environment into Azure.

There are some industry precedents for Oracle's move, but none of them exactly size up to actually plopping physical infrastructure directly inside a foreign hyperscaler environment. For instance, AWS has small availability zones, including Wavelength which is placed inside

5G telco switching centers, and Local Zones, which run in colocation facilities near a core group of clients, such as in Los Angeles where a local zone is available for Hollywood entertainment industry clients. Customers can also rent an AWS Outposts rack that sits inside their data center, just as with Oracle Cloud@Customer (in both cases, the hyperscaler manages it). Then there is Google Cloud Omni for BigQuery, which Google supports running inside AWS and Azure, but it is a software-defined cloud running as an Anthos Kubernetes cluster. By the way, don't confuse BigQuery Omni with the similarly-branded AlloyDB Omni, which is *not* a Google-managed service, but is instead a sandbox for developers to run on their laptop (that opens a Pandora's Box, as AWS offers a developer DynamoDB sandbox).

For joint Oracle Database/Azure customers, the differences will manifest as performance and convenience: the latencies will shrink from milliseconds to microseconds. That will improve the performance for use cases such as Power BI running against Oracle Database to the point of being instantaneous. And as noted above, being able to buy Oracle database services with Azure cloud credits will simplify life by dispensing with the need for customers to bring their own existing Oracle database licenses.

The obvious synergies are between Oracle Database and Microsoft GenAI services, where Oracle customers can leverage Copilot and similar services coming out of the Microsoft-OpenAI partnership. But Microsoft is not giving up its database and Oracle is hardly outsourcing GenAI services. Instead, the biggest ramification will impact OCI itself. Oracle indicates that it would like to depend on kindness from more strangers with similar cohabitation deals, and we wouldn't be surprised to see Oracle ink another such deal sometime in the following year.

But what does this mean for the future growth of OCI? For now, the interconnect is not going away, as the Azure deal doesn't cover Oracle SaaS or IaaS services. From Microsoft's standpoint, database is where they perceive to get the biggest bang for the Oracle buck, but never say never. As we noted a few months back, Oracle's goal is to saturate the world with OCI regions (which are the equivalent of availability zones with the other hyperscalers); our take is that a large chunk of the growth will be inside host hyperscalers.

Internalizing GenAI

It seems like the data, analytics, and cloud worlds have entered the era of "All-GenAI all-the-time." And so there was little surprise that GenAI took prominence at CloudWorld. Oracle has taken a sizable minority investment in Cohere and is using it to build out its conversational (natural language) APIs both for making analytic queries and generating SQL code. With the latest announcement of Amazon taking a strategic stake in Anthropic, which will make AWS its prime cloud partner, all of the hyperscalers have now put skin in the game with LLM model first movers, with Oracle joining the wave.

Using LLMs for query and code generation

The viral popularity of ChatGPT has shown that the best understood, and for the moment, most popular use case for GenAI is for English (and soon other spoken languages) to become the default API or SDK.

So it's not surprising that Oracle is hopping on the bandwagon with support for conversational query and SQL code generation, courtesy of foundation models provided by Cohere. Oracle APEX will be using Cohere to generate blueprints. For the Oracle Database 23c JSON Relational Duality feature, it will also be creating a utility that uses the LLM to generate relational schemas from JSON document collections. And, like other hyperscalers and others in the data and analytics space, Oracle won't restrict support to models developed using its own preferred LLM; customers can use the FM of choice with Oracle database development.

What's different is how Oracle is positioning SQL code generation for Oracle Database; it is far more sober in its expectations for how optimized the resulting code will be coming off an LLM model (we heard assessments that Oracle would score the results at seven on a scale of ten). Oracle insists that human SQL coders must stay in the loop. We haven't heard such caution, at least out loud, from other providers introducing SQL code generation.

Oracle explains that, compared to LLMs that harvest the general Internet, there isn't any single repository of SQL code that has sufficient critical mass at least yet. Given that Oracle databases tend to be enterprise-scale deployments, often with mixed workloads and complex table structures, generating the best SQL will test the limits of language models. It will certainly be more demanding compared to generating SQL for departmental databases that often use other platforms. Yet on the APEX side, Oracle is more sanguine, with LLMs adding another layer of automation to the process for generating database application blueprints based on database metadata. But, in this case, it is working with a well-defined corpus, which is the customer's existing table structures.

Vector storage and retrieval

The importance of Retrieval-Augmented Generation (RAG) for making large language foundation models updated and relevant to the domain has placed the spotlight on vector storage and indexing. While specialized vector databases have emerged, we believe that for most use cases, they will be *features* of existing databases. This is useful for scenarios where the answer to a question could involve summarization of findings from text or image that is supplemented with an analytic visualization of tabular data (e.g., pick a favorite picture of a house, then look through a database tabular data of what houses are on the market, then sort through associated pictures, which are represented by vectors, to find closest matches).

Oracle is the latest to add vector storage, and is implementing extensions to SQL (which it plans to submit to the SQL standards community) for vector similarity search queries for both

Database 23c and MySQL HeatWave. The extensions are necessary because similarity searches are different from SQL queries; they look for approximate matches, rather than the exact data that is customary for SQL. In preview, Oracle Database 23c is adding vector search and indexing, differentiating by offering a choice of two different schemes:

- *Neighbor Graph Vector Indexes* for similarity searches of index compact enough to fit into memory. There are parallels with the Facebook-created Faiss open source library that is designed to fit in memory, and is one of the approaches supported by Microsoft Azure Cognitive search.
- *Neighbor Partition Vector Indexes* for similarity searches that must span a much wider corpus of data. As the indexes are partitioned, they are optimized for scale-out, parallel search. This takes advantage of several capabilities: Oracle RAC for scale-out and workload isolation, and Exadata's ability to offload processing to its smart storage tier.

By comparison, most of Oracle's rivals rely on more generic search index approaches for supporting vector similarity searches. The indexing scheme for MySQL HeatWave in private preview has not yet been disclosed.

Oracle Database 23c gets more Duality

Oracle Database 23c, the new long-term release, is entering general release initially on Oracle Base Database Service in OCI, and we expect that it will percolate to on-premises, Exadata, and Autonomous Database respectively within the next 6 - 12 months. Here are several highlights out of the 300+ new features for 23c.

The top story is expanding on the JSON Relational Duality capability unveiled a year ago. Duality harnesses GraphQL to represent relational data as JSON documents. The new enhancement adds views for JSON document formats relying on the underlying GraphQL to construct the views. Compared to traditional materialized views, which physically instantiate the view, Oracle's views are virtual, generated on the fly. JSON Relational Duality views extend a capability that Oracle has long had on the relational side, both for transaction data (e.g., custom table joins) and analytics (multidimensional views).

Database 23c also adds support for SQL/PGQ syntax for querying property graphs. We expected this, as Oracle was one of the key participants (alongside Neo4J and others) in developing this emerging standard for property graph query.

Another key enhancement for 23c is a new real-time capability for automated SQL query plan management. The going notion is that the new plans that optimizers automatically choose might not work as well as the previous plan; the real-time enhancement will allow the older proven plan to get reinstated on the fly. This enhancement augments the Oracle Database

Adaptive Query Optimization set of capabilities introduced a decade ago that enable the optimizer to make run-time adjustments to execution plans and discover additional information that can lead to better statistics. This new rollback approach will come in handy when existing statistics are not sufficient to generate an optimal plan.

Oracle is not the first to get there; about five years ago, Adaptive Query Processing and Automatic Query Plan Correction capabilities that activated while queries were running debuted in Microsoft SQL Server 2017.

MySQL extends the Lakehouse

Last year, Oracle announced major extensions to MySQL HeatWave with AWS and lakehouse support. Not surprisingly, this year, the major new feature is vector storage along with a private preview of vector indexing support, as noted earlier.

Otherwise, the bulk of announcements build on existing features. For instance, the lakehouse now also runs on AWS, supporting existing HeatWave AutoML capabilities for training, inference, and explanations on data in object storage in addition to data in the MySQL database. HeatWave AutoML was enhanced with support for text processing across anomaly detection, forecasting, classification, regression, and recommender system tasks, as well as an enhanced recommender system to generate personalized recommendations. JSON performance also gets a needed upgrade, enabling real-time analytics to be performed by HeatWave on JSON documents.

Other enhancements include:

- A new feature for Autopilot that automatically determines what secondary indexes should be created or deleted based on workload patterns;
- Adaptive Query Execution that optimizes query execution based on actual data distribution at run time;
- Autocompression that determines the optimal compression algorithm for each column, which Oracle claims could cut memory usage costs by up to 25%;
- Implementing a new open source MySQL 8.0.33 feature. For automating data unloads via stored procedures; and
- Automatic data unloading via a stored procedure (this actually implements a capability introduced in open source MySQL 8.0.33).

We're still awaiting MySQL HeatWave Lakehouse to follow the lead of Autonomous Data Warehouse for supporting Apache Iceberg, the open source table format that is becoming the de facto industry standard. Watch this space.

Let's not forget Autonomous Database

Oracle Autonomous Database is the first, and so far only self-driving database on the market. Over its 5+ years on the market, it has found a niche mostly with midsized enterprises or departments of larger customers; significantly, some of the midmarket base is new to Oracle. But among larger enterprises, there is an entrenched belief that deployments are too exceptional to entrust to optimizations perceived to be generic, not to mention perceptions that improvements to the performance of one or two database instances might not be worthwhile.

Oracle is addressing the cost/benefit of Autonomous Database in several ways. First, it is phasing in *elastic resource pools*, a new scheme that allows customers to bundle multiple databases to a single virtual instance. Instead of dedicating compute resources to individual databases, they can be pooled together. There are parallels with an Oracle scheme to support multitenancy (bundle multiple databases under a common management scheme) decade ago with Oracle Database 12c. This will lower the barrier to cost (pools do not have to be dedicated) and convenience (resources for each autonomous instance will not have to be configured separately).

Secondly, Oracle is taking out an impediment to scaling Autonomous Data Warehouse (ADW) by slashing the cost of block storage to have parity with cloud object storage. This has interesting implications for Oracle's recent ADW lakehouse support, which added support of Iceberg open tables to extend the platform's reach. It will draw ADW Lakehouse customers to a question of tradeoff: opt for placing data in Exadata storage for higher performance vs. the advantages of keeping it open to other analytics platforms in cloud storage running Iceberg.

Additionally, the Autonomous Database will move to a wider stage as it becomes the default database for Oracle's SaaS services, including NetSuite NXN and Fusion Applications (Fusion EPM, Warehouse Management, and CX Unity). While the shift should be invisible to SaaS customers, it should lower operational costs. We'd like to see Oracle seize the moment and slash SaaS costs accordingly.

Takeaways

Oracle's newfound ability to run its database services *inside* the Azure cloud was the biggest takeaway from CloudWorld, even though the actual announcement came the previous week. Although, at first glance, it appears as the logical next step to a series of developments that have brought the two cloud providers into closer collaboration, the significance is proof that Oracle's unique OCI cloud architecture is portable. As we noted previously, OCI is a classic second generation leapfrog design. In combination, its flat topology and RDMA features yield extremely high performance thanks to elimination of several layers of overhead from the operating system and network hops. Bringing this to Azure proves that Oracle database cloud

services can operate with the same SLAs and get exposed to a wider audience of customers. In this case, that would be joint customers seeking to take advantage of Microsoft 365 and related BI, Copilot and other GenAI services stemming from its OpenAI partnership.

In practice, OCI regions tend to be much more compact than the availability zones of the major hyperscalers. That is largely attributable to Oracle's later start in the market and the fact that workloads tend to cluster around the database and enterprise applications for which Oracle is known; Oracle does not have the same variety of services of AWS. Oracle has planned to grow its regions like inkblots scattered across the globe. And, as we noted, Oracle is hardly alone as AWS is likewise building out miniature presences adjacent to concentrations of key customers, such as near Hollywood for entertainment industry clients.

For now, the Azure arrangement is with Oracle database services only. Omitting MySQL HeatWave, SaaS, and IaaS services, we don't expect that the need for OCI interconnects with Azure, and potentially other hyperscalers, will go away anytime soon. But we believe that going forward, the primary trend for OCI growth will involve colocations, and Oracle has publicly stated that it wants to cohabitate with other hyperscalers. That prompts the obvious question: Who's next?

Author

Tony Baer, Principal, dbInsight

tony@dbinsight.io

LinkedIn https://www.linkedin.com/in/onstrategies/

About dbInsight

dbInsight LLC® provides an independent view on the database and analytics technology ecosystem. dbInsight publishes independent research, and from our research, distills insights to help data and analytics technology providers understand their competitive positioning and sharpen their message.

Tony Baer, the founder and principal of dbInsight, is a recognized industry expert on data-driven transformation. *Onalytica* named him as a Top Cloud Influencer for 2022 for the fourth straight year. *Analytics Insight* named him one of the 2019 Top 100 Artificial Intelligence and Big Data Influencers. His combined expertise in both legacy database technologies and emerging cloud and analytics technologies shapes how technology providers go to market in an industry undergoing significant transformation. A founding member of The Data Gang, Baer is a frequent guest on *theCUBE* and other video and podcast channels.

dbInsight® is a registered trademark of dbInsight LLC.