db insight

Oracle Database 23ai puts AI in the driver's seat

Grounding gen AI with transaction data

Published May 2024

Executive Summary

Trigger

Oracle is hitting the next inflection point with the release of Oracle database 23ai. The highlights of 23ai, which will be Oracle's long-term support release, double down on the converged database theme by bringing AI, multimodel data types, caching, and perimeter security features into the heart of the platform. With databases embracing generative AI (gen AI), what unique advantage is Oracle bringing to the party?

Our Take

Amidst the constant pace of database platform updates in the era of the cloud, Oracle provides its customers definitive reference points with strategic releases with long-term support issued every four years. And that witching hour is now: Oracle has just released Database 23ai, the reference release for Oracle enterprise customers for the next 4 – 5 years. The change in suffix is a not-so-subtle hint as to where Oracle expects core demand for new use cases is going. We've been there before: Oracle added the suffix "i" with version 8 at the dawning of the commercial Internet back in 1998; "g" for grid computing with the 2003 v10 release; and "c" for scalable cloud deployment beginning with v12 in 2014. Now it's AI's turn.

The 'ai' suffix affirms that out of 300+ new features and enhancements, AI is the highlight. AI capabilities start with vector storage and indexing for supporting Retrieval-Augmented Generation (RAG). The new features make gen AI more accessible, through database development tooling and SQL extensions blending traditional query with vector similarity search. The guiding notion is that enterprise data grounds vector search while vectors in turn enrich enterprise data with stories and context.

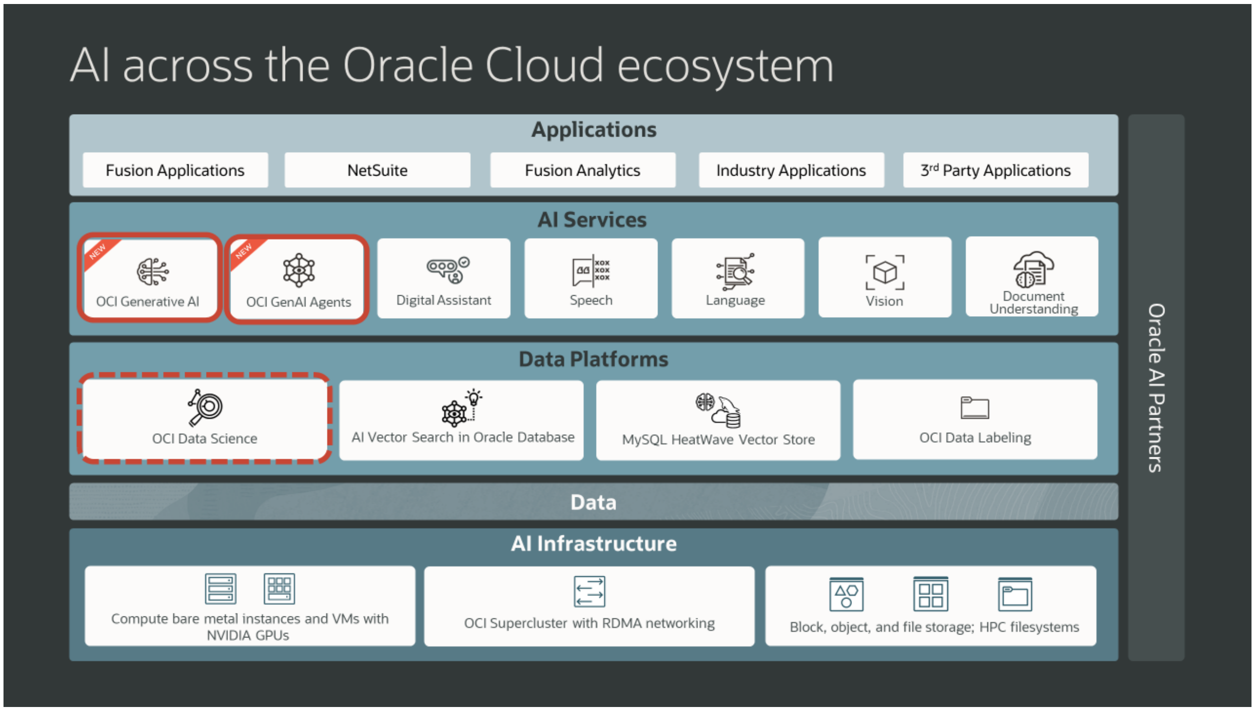
Clearly, AI has become a broad theme across the database landscape over the past year. Oracle's differentiator is under the hood where bringing in data of all forms and disparate functions directly in-database ensures results that are current and transactionally consistent, with security and access controls/permissions consistently enforced. An otherwise obscure detail, bringing in the network firewall in-database, closes a potentially serious back door by locking down in-database stored procedures that are otherwise left exposed by traditional perimeter firewalls. It's all about the details.

The database as AI linchpin

As gen AI has dominated technology headlines over the past year, Oracle has been putting the pieces in place for bringing gen AI into the database in conjunction with accompanying OCI cloud infrastructure and services; Azure OpenAI services integration; and an equity investment investment in Cohere.

Oracle's database AI support is part of a broader palette of AI services from OCI public cloud that covers all the tiers from the application/SaaS level to platform-based managed services through infrastructure level (see Figure 1).

Figure 1. OCI AI portfolio



AI across the Oracle Cloud ecosystem

Applications

Fusion Applications | NetSuite | Fusion Analytics | Industry Applications | 3rd Party Applications

AI Services

OCI Generative AI (NEW) | OCI GenAI Agents (NEW) | Digital Assistant | Speech | Language | Vision | Document Understanding

Data Platforms

OCI Data Science | AI Vector Search in Oracle Database | MySQL HeatWave Vector Store | OCI Data Labeling

Data

AI Infrastructure

Compute bare metal instances and VMs with NVIDIA GPUs | OCI Supercluster with RDMA networking | Block, object, and file storage; HPC filesystems

Oracle AI Partners

Source: Oracle

While the OCI cloud provides standalone AI services such as Document Understanding and Data Science development environments/workbenches, the database is the linchpin. And here, Oracle's strength is its leading presence with transaction systems with large enterprises, giving it the opportunity to flip the script about gen AI. While the common impression is that gen AI augments query with natural language, Oracle's ability to converge SQL query and vector search backs the narrative that *transaction data augments conversational query*.

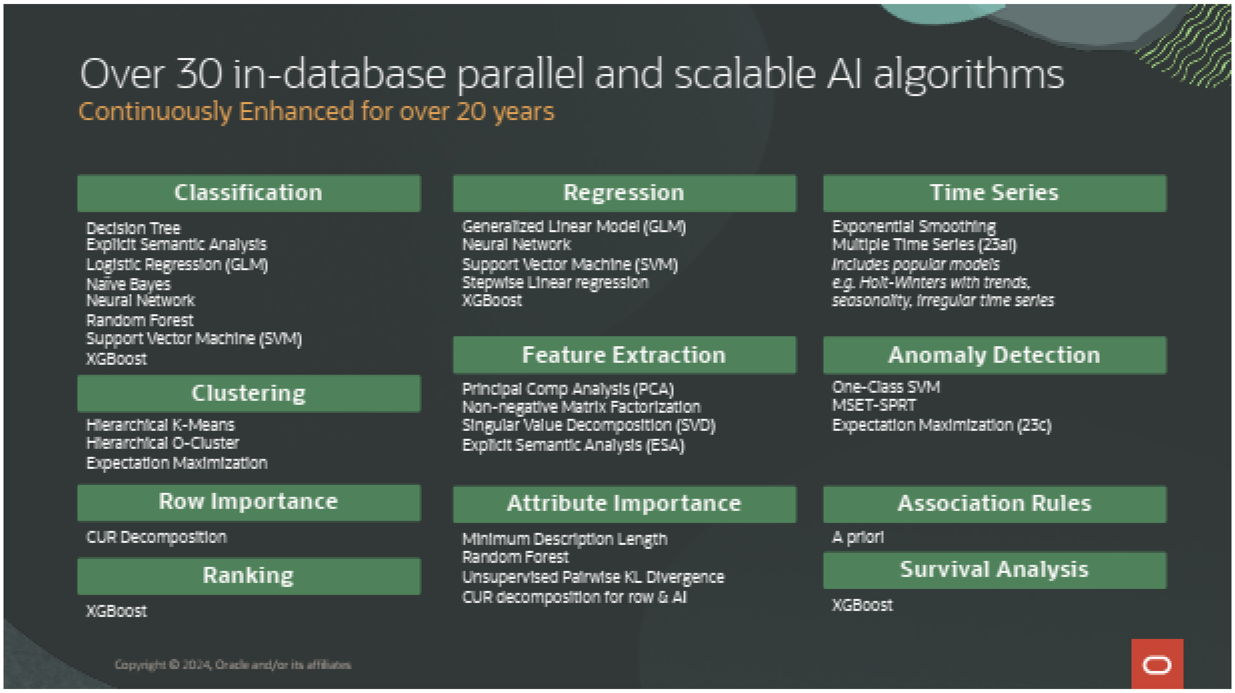
The pillars of Oracle's AI support in Database 23ai encompass:

- Vector storage and search for supporting retrieval-augmented generation (RAG);
- Making gen AI accessible to SQL developers;

- Grounding up gen AI answers with enterprise data (and vice versa); and simplifying paths to adoption for developers with enhanced tooling.

This strategy is complemented by Oracle's existing support of in-database machine learning, and for Exadata customers, leveraging Exadata for running vector similarity searches in parallel across its scale-out storage servers (see Figure 2).

Figure 2. OCI in-database ML model support



Over 30 in-database parallel and scalable AI algorithms

Continuously Enhanced for over 20 years

Classification

Decision Tree
Explicit Semantic Analysis
Logistic Regression (GLM)
Naive Bayes
Neural Network
Random Forest
Support Vector Machine (SVM)
XGBoost

Clustering

Hierarchical K-Means
Hierarchical O-Cluster
Expectation Maximization

Row Importance

CUR Decomposition

Ranking

XGBoost

Regression

Generalized Linear Model (GLM)
Neural Network
Support Vector Machine (SVM)
Stepwise Linear regression
XGBoost

Feature Extraction

Principal Comp Analysis (PCA)
Non-negative Matrix Factorization
Singular Value Decomposition (SVD)
Explicit Semantic Analysis (ESA)

Attribute Importance

Minimum Description Length
Random Forest
Unsupervised Pairwise KL Divergence
CUR decomposition for row & AI

Time Series

Exponential Smoothing
Multiple Time Series (23ai)
Includes popular models
e.g. Holt-Winters with trends,
seasonality, irregular time series

Anomaly Detection

One-Class SVM
MSET-SPRT
Expectation Maximization (23c)

Association Rules

A priori

Survival Analysis

XGBoost

Copyright © 2024, Oracle and/or its affiliates

Source: Oracle

Vector support

Vector storage and indexing have grown commonplace over the past year with operational databases, and more recently with data warehouses. While some specialized vector databases (e.g., Milvus, Pinecone) have emerged, the more common pattern is that vector support is becoming a *feature* of off-the-shelf databases. The driver is supporting RAG – the notion of running gen AI against the enterprise's own data as a query rather than constantly refining a large language model with enterprise data.

As we noted last fall, Oracle Database 23ai added 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.

In our 2024 Data and AI Outlook, we turned the spotlight on vector *indexes* as a dark horse hotbed for innovation this year because of the fact that not all vector similarity searches have the same requirements, and therefore, one style or size vector index won't fit all.

For instance, "quick and dirty" similarity searches are likely sufficient for use cases such as content generation, whereas, conversely, clinical research identifying promising molecular candidates for new drug discovery will require more exhaustive searches, with less premium on speed of response. While most databases have debuted with a single, generic indexing scheme, Oracle is one of the first outside of specialized vector stores (e.g., Milvus, Pinecone) to embrace different use cases; Google's AlloyDB with its SCaNN index provides another example.

Nonetheless, we term indexing a dark horse disruptor because it is not likely going to be topping the list on how customers differentiate database provider RAG support. It will become more important to customers once they gain more experience with gen AI and learn that one size doesn't fit all.

Lowering the barriers to AI Development

There are several angles to empowering developers. It can come by making business data accessible through natural language; enabling them to tell stories through natural language; blending querying of structured and unstructured data; and empowering developers with digital code assistants. In Oracle database 23ai, key development-related features include:

- Natural language query via support from Cohere (that is the default option), or the language model of the customer's choice. Oracle is enabling customers to swap in different language models through the open source ONNX (Open Neural Net exchange), a protocol that allows running a vector embedding model remotely on GPUs or directly in-database eliminating the need to move data.

- Select AI, a capability that allows SQL users to trigger vector searches through SQL SELECT statements. While this is a well-established capability for invoking ML models, Oracle is the first to support this capability for gen AI.
- LangChain support, providing an open source framework for developers to build language models into their applications; LangChain support is becoming increasingly common across analytics and AI development platforms.
- Extending JSON Duality for enabling an LLM to generate relational schemas from JSON document collections.

The brass ring is bringing the world of relational and generative query together, because for most business use cases, a query of unstructured data will likely require some reference to the structured data of record. Oracle characterizes this as "augmenting [natural language] prompts with private database content." This is the high level differentiating theme we referred to earlier.

By comparison, most operational databases and (more recently) data warehouses have added vector data storage and retrieval. But most have treated vector similarity search as separate operations from traditional database queries, regarding application logic to synthesize them together. For instance, Microsoft Fabric and Cosmos DB for PostgreSQL require separate Python programs for vector search. In turn, Google BigQuery has added a Studio offering for Python developers to author queries running against Gemini or other models with the data staying in-database.

Oracle brings the two together by enabling developers to add vector searches to familiar SELECT statements. There is ample precedent for this approach, as most major cloud data warehousing platforms have extended their SQL capabilities for triggering ML models; for vector search, we expect Oracle rivals to follow suit. It also supports Python developers in a couple ways, including a Python AI (Oracle Machine Learning for Python, a.k.a., OML4Py) for running statistical analysis and ML algorithms, or execution of Python user-define functions (UDFs) in-database.

Supporting Distributed AI

Oracle's design philosophy is to move all functions to the extent possible inside the database so they can all be managed within the same high availability, and scalability. security and management umbrella. This has been true ever since the days of Oracle Forms, its original database development language where logic ran inside the database. The guiding notion is working with data in place, avoiding movement wherever possible. And, to the extent possible, internalizing applications logic, AI models, and vector embedding means that apps and models are always working with the latest state of the data.

But reality often intrudes, and it is not always possible to have all the data reside in one happy database. Many Oracle customers operate fleets of multiple instances. But if they haven't consolidated their instances as pluggable databases in Exadata, their Oracle 23ai upgrades with their vector stores will reside in separate silos. A workaround is to replicate data, with Oracle's GoldenGate providing the ability to populate a central Oracle 23ai instance where a language model can generate vector embeddings. The latest release of Oracle GoldenGate 23ai, adds the ability to replicate vector embeddings from Oracle and/or non-Oracle databases as well, closing the loop. When the autonomous database gets updated to 23ai, there will be another option for its globally distributed edition that will include lighter weight RAFT-based replication. As most enterprises are just starting to get up to speed with gen AI and are likely starting with a single instance, Oracle is clearly ahead of the curve with its GoldenGate support for replicating vector embeddings.

But there is one key requirement for making vector replication from multiple source work: embeddings must be generated by the same language model and version, and trained on the same corpus of data. Otherwise the embeddings from remote databases with different models may look like gibberish when combined. The management overhead of ensuring that all the language models remain in sync won’t be trivial, which is why we believe this will be at best an interim approach. Going to such trouble makes sense if the customer is planning this as one of the first steps toward database consolidation (which of course often comes with cloud transformation or the outcome of corporate restructuring).

But there is another approach to distributed AI that we believe will be more sustainable: the geographical, shared-nothing sharding that Oracle just introduced with the Globally Distributed Autonomous Database that we analyzed in a previous report, and will soon be updated to 23ai. *Here, the database and the model are a single logical instance*. Furthermore, the globally distributed autonomous adds support for RAFT (a lightweight consensus protocol) for active-active replication scenarios within regions for high availability. While introduced to support rising requirements for data sovereignty, we believe there will be a knock-on effect for globalizing gen and/or classical AI. We can imagine Oracle extending this distributed AI capability for these highly localized global deployments.

Takeaways

Oracle’s key differentiator with Database 23ai is extending the full lifecycle for gen AI applications into the database. We view it as a logical extension of Oracle’s “converged” database architecture, in this case venturing to the AI and the application tier. The guiding notion is simplifying application and AI development by working with data *in place* while managing it all under a common security and management umbrella and guarantees transactional valid results.

The converged architecture is best known for bringing all types of structured data into the database; in 23ai it adds vector embeddings representing unstructured data to the mix. That in turn enables Database 23ai's SELECT AI feature. SELECT AI enables database developers to use SQL to combine queries of structured data (or JSON data via Duality Views) with similarity searches of vectors embeds under a common SQL select statement.

But as noted above, we also view the converged theme to extending beyond data to other tiers, such as app tier and perimeter protection. In some respects, that's hardly news for Oracle, which has had its own portfolio of tools for in-database app development dating back to Oracle Forms, followed by Java, in-database processing of JavaScript stored procedures (via the multi-lingual engine), and more recently the low code/no code APEX. The guiding notion is (1) not moving data; (2) ensuring apps get transactional valid data; and (3) placing all supporting functions under the Oracle management umbrella.

In 23ai, there's some new features converging the app tier. True Cache adds an in-database caching tier, flattening and simplifying the architecture and ensuring transactional consistency between the cache and persistent storage. Although commonly associated with Memcached or Redis, the need for separate caches actually dates back to the dot com days with the need for appservers because, at the time, databases were overwhelmed by Internet scale. With True Cache, Oracle comes full circle from its WebLogic days. Another new feature, Property Graph views, provides the ability to work with property graphs through the emerging GQL property graph SQL extensions that are now part of the ANSI standard.

Now that Oracle has interwoven SQL and vector searches, and with ONNX, opened up the database to in-situ processing with external language models, we'd like to see them venture beyond language models. Images, sound, geospatial, molecular, and other firms of data that will also become fair game for gen AI. For instance, Google's BigQuery has ventured beyond the text (language) API of its Gemini model to images as well. Oracle says that the door is open for that as we saw a demonstration of SELECT AI for a home buyer app where they can search for listings that look like their submitted photo that are within their price range. The demo combined similarity searches from a photo with lookups of pricing and inventory from structured data to display candidate listings. This is just the tip of the iceberg.

Related reports

Further background on Oracle's emerging gen AI strategy is available in the following reports:

- Oracle's Autonomous Database can now be globally distributed (March 2024)
- Oracle rounding out Generative AI Support (February 2024)
- Oracle Database @ Azure redefines multi-cloud (October 2023)

Author

Tony Baer, Principal, dbInsight

tony@dbinsight.io

Linked In https://www.linkedin.com/in/dbinsight/

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.

Published May 2024