Oracle Rounding out Generative AI support

Next stop? Fusion Applications

Published February 2024

Executive Summary

Trigger

Oracle is following up its opening shots with Gen AI with the general release of OCI Generative AI service, plus unveiling of new services for building more complex Gen AI workloads with Retrieval Augmented Generation (RAG). As icing on the cake, Oracle is adding new low code/no code actions to the Oracle Data Science service that could make generative applications event-driven, and is expanding the infusion of Fusion Cloud Applications with Gen AI capabilities to complete its full-stack approach. How do the new Gen AI introductions and features stack up to what's being offered by the hyperscalers, and what are the implications for Oracle's home base of database and enterprise applications?

Our Take

With the current announcements, Oracle has filled some of the blanks in its cloud AI offerings with Gen AI. Oracle has a broad strategy addressing four tiers of the stack, from applications to data, prepackaged AI services, and infrastructure, where it leverages high-performance, OCI topology for offering Nvidia GPUs as "superclusters." For databases and applications in its sweet spot, Oracle has already begun adding Gen AI capabilities such as vector data support in Oracle Database 23c and MySQL HeatWave, along with the beginnings of generative support at the SaaS tier with Oracle Fusion Cloud CX and HCM.

Oracle's Generative AI model serving service (now GA) is a good start, but as yet its portfolios of foundation models (FMs) are dwarfed by AWS and Google Cloud. But watch this space for more news soon. Oracle's Generative AI Agents, just announced for beta, steal a move on rivals by making Gen AI semantic query services reactive and able to take actions on behalf of the users. Although initially confined to automating RAG processes with unstructured data, Oracle will likely fill the gap with structured data when it later goes GA. Oracle hasn't neglected data scientists on this round of announcements, adding no code/low code assistance for the pedestrian tasks of model deployment and scaling, which will make them more productive.

Ultimately, Oracle's biggest differentiator is at the data and application tiers, and this is where the current spate of announcements leave off. But not for long.

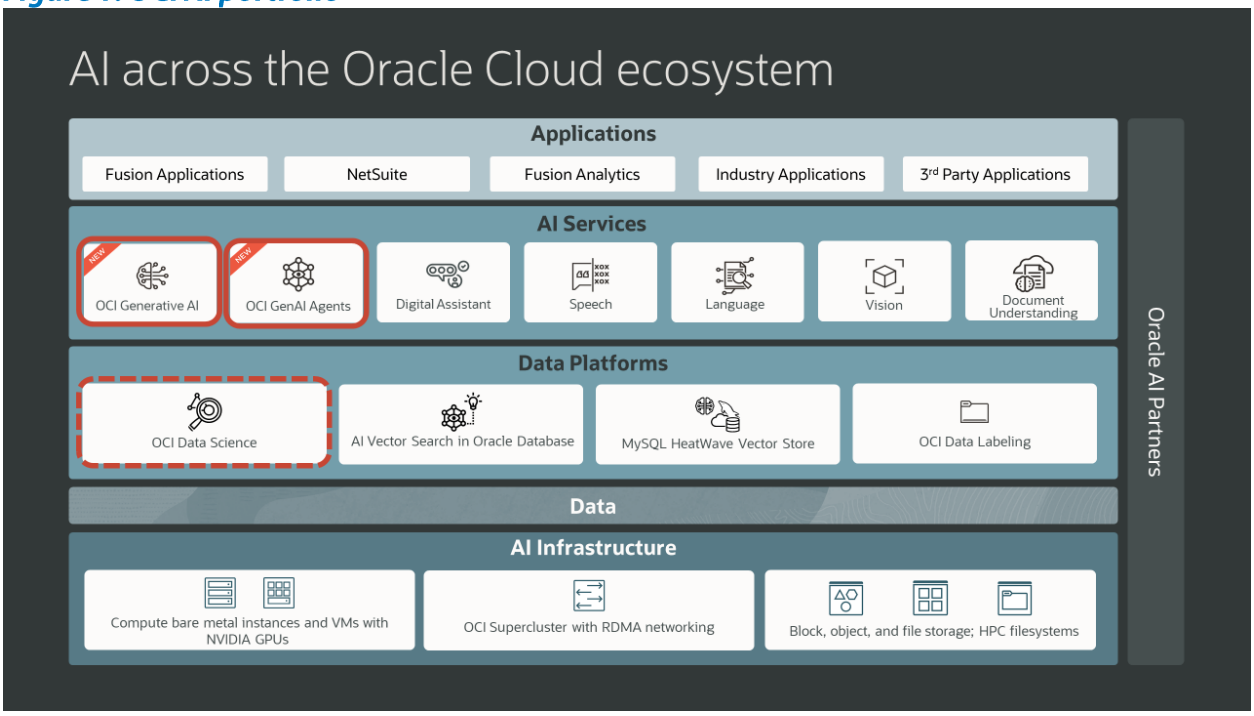
The onrush to Gen AI

Since Chat-GPT was unleashed to the public just over a year ago, the big question facing data, analytics, and AI solution providers wasn't if, but when, they would start rolling out production services for the enterprise. We saw hotspots emerging with language models, propelling English (and other spoken languages) as the world's most popular new API;

copilots cropping up as smart digital assistants; databases adding vector storage and indexing support for Retrieval Augmented Generation (RAG); and infrastructure becoming cool again, courtesy of rare and precious GPUs.

That was the backdrop last fall when Oracle made its initial foray with the preview of OCI Generative AI Service. As we noted at the time, given that the data, analytics, and cloud worlds entered the era of "All-GenAI all-the-time," we weren't surprised that GenAI took prominence at CloudWorld. The newly-announced Gen AI service supplements a critical mass of machine learning (ML) and deep learning (DL) prepackaged services that are not necessarily unique to OCI, but table stakes for all cloud data and analytics providers.

Figure 1. OCI AI portfolio



Source: Oracle

Oracle's AI starting points

Figure 1 shows the portfolio of AI-related services available from OCI, with the current announcements highlighted in solid or dashed red lines. It covers all the tiers from the application/SaaS level to platform-based managed services through infrastructure level. And so, like hyperscaler rivals, Oracle has the basic checkboxes at different tiers filled in. Of course, the maturity and breadth of AI-related services varies. For instance, the OCI portfolio carries

the ML and DL-based services that one would expect from a hyperscaler. But Gen AI is a fast emerging segment, with Oracle and hyperscalers still in the early stages of rolling out their services. In general, we expect CY 2024 will be primarily an investment rather than adoption year for Gen AI. Cloud and solution providers (Oracle included) are in build and extend mode with first-generation offerings, and a new long tail of third parties will introduce purpose-built foundation models with more modest (and sustainable from cost and environmental perspectives) compared to early general-purpose LLMs. Aside from conversational interfaces and copilots, we expect most enterprise implementations will be at proof of concept stage as organizations identify use cases and *the right mix* of models and services to support them. And in 2025, we expect that hyperscalers (Oracle included) will introduce marketplaces for the emerging long tail of foundation models. It has taken the first steps with Nvidia AI Enterprise and Nvidia DGX Cloud listing in the Oracle marketplace.

As noted below, Oracle's selection of foundation models is currently dwarfed by rivals such as AWS and Google Cloud, but we view this as a moving target. And as we'll note later, Oracle's advantages are, not surprisingly, on its home turf of databases and enterprise application SaaS services. Hold those thoughts.

Application tier

Oracle liberally infuses ML predictive analytics into its enterprise applications SaaS stack – as have rivals such as SAP, Microsoft, Salesforce, and Infor. Within the Fusion Cloud applications, there are predictive analytics built into the analytics services for core ERP, HCM (human capital management), supply chain, and CX (customer experience). They come in the form of prebuilt analytics (e.g., collection risk prediction for ERP accounts receivable or diversity analysis for HCM), and custom options providing data scientists access to a library of 30+ ML algorithms for running against Fusion application data in-database. By infusing Gen AI into those applications the goal is making them more interactive and easier to use.

Prepackaged AI services

At the next level down there are AI-specific services on OCI that parallel what is offered by other hyperscalers. Among them are prepackaged deep learning services for translating speech to text, image and text recognition, digital assistants (bots), and document text extraction. And then there is Oracle Data Science as a PaaS offering for data scientists seeking to run the algorithms of their choice.

Data tier

Oracle's databases have added vector storage and indexing for supporting Retrieval Augmented Generation (RAG), a practice for keeping Gen AI models current and relevant. RAG enables enterprises to exploit their own data, either supplementing or replacing the

training data sets for existing foundation models (FMs). Oracle's vector support largely echoes that of many operational databases that treat vector data as just another data type; it is consistent with Oracle's positioning of its flagship database as a "converged" database supporting relational and nonrelational data types. Oracle is one of the few database providers that offers a choice of vector indexes (for performance or scale), giving it a headstart in an area where we expect significant innovation in 2024. Oracle also offers a service that streamlines labelling of training data.

Infrastructure

This tier has grown in importance given the heavy compute requirements of Gen AI transformer models, and the fact that GPUs have become rare and precious commodities. Hardware has become as cool as Nvidia CEO Jensen Huang's trademarked black leather jacket. Unlike AWS or Google Cloud, Oracle is not developing its own silicon, but it differentiates in how it is optimizing Nvidia GPUs with OCI's unique, high-performance backplanes. It is deploying Nvidia H100s in "supercluster' configurations ganging up to 4000+ instances totaling 32,000+ H100 GPUs. Key to this, of course, is OCI's differentiated implementation of RoCE (RDMA over Converged Ethernet), a backplane which bypasses the OS providing direct memory access that delivers super-fast, low latency compute. Size evidently matters; while Microsoft Azure also has its own Nvidia instances, it is relying on OCI superclusters to support Bing conversational searches.

Rounding out the Gen AI portfolio

OCI Generative AI Service

As Oracle's first entry into Gen AI model serving, the OCI Generative AI service was announced for preview last fall at Oracle CloudWorld; it is now entering general release. It provides a selection of pre-trained foundation models and the ability to fine tune them (e.g., compare outputs of test data sets) on dedicated AI clusters.

The Gen AI service is Oracle's answer to Amazon Bedrock, Google Model Garden, and Azure Open AI in that these services provide API-based access to a curated portfolio of supported Gen AI transformer models as a managed SaaS service. Currently the service supports two families of models, including Meta's open source LLaMA 2, and several models from Cohere, a company in which Oracle has invested along with Salesforce Ventures, Nvidia, and others. The service will have multilingual support for over 100 speaking languages. Customers have a choice of two modes of deployment through a serverless pay-as-you-go option or through dedicated, provisioned AI clusters for predictable price/performance that is primarily intended for stable, inference workloads. This is akin to how AWS delivers Amazon Bedrock, a comparable service.

Published February 2024

Specifically, the portfolio includes three models from Cohere that include:

- **Command**, a general-purpose LLM, which is available in two configurations: Light, which reduces the number of computed parameters to 6 billion for higher throughput and lower cost; and XL, which has 52 billion parameters and where the key requirement is high precision and accuracy;
- **Summarize**, for document summarization; and
- **Embed**, a specialized model that converts language to vector embeddings for English and other spoken languages. Oracle offers a full-size and a smaller model for both English and multilingual versions. The smaller model is less performant though it is faster and designed for lower cost.

The other model currently supported by OCI Generative AI service is Meta's LLaMA 2, which is offered with its latest 70-billion parameter configuration. There are some subtle distinctions in how Oracle packages and delivers these models. Both are tuned for OCI superclusters that can be used in their vanilla pretrained form or supplemented with enterprise data via RAG.

Atop these models, Oracle has refined some of the capabilities of the Generative AI service on entering GA, with the highlights being the ability to run "stacked" models (concurrently serving multiple fine-tuned custom models from the same set of GPUs) and a simplified interface for scaling and managing clusters. The stacked modeling feature is especially significant given that we expect, going forward, that more specialized models will become available, so you don't have to always run a general-purpose 70-billion parameter model for applying some highly specialized domain context. In the long run, we expect that enterprises will gravitate to relying on models with more compact compute (and carbon) footprints for specialized or domain-specific tasks. On the roadmap, the service will add integration with LangChain, an open source framework for developing custom applications with language models.

So how does this stack up vs. Oracle's rivals? In general, most competing model serving services currently have broader model selections:

- AWS Bedrock is a managed service that provides access to Jurassic-2 (from AI21 Labs); Claude (from Anthropic); Command and Embed (from Cohere); LLaMa 2 (from Meta); Stable Diffusion (from Stability AI); and Amazon's own Titan. Bedrock is more than just a model server library API; it also has capabilities for testing and tuning models, comparing them, and configuring prompts. A fun fact is that Bedrock has recently introduced an agents feature that helps automate tasks such as prompt engineering and secure access to data. This is a feature that Oracle is separately previewing for a different purpose (semantic search), as noted below.

- Google Cloud Model Garden offers a choice of 130+ foundation models from Google, third parties, and open source. Google's Duet strategy aims to offer copilots across all Google Cloud services, a goal that will take several years to roll out. For open source models in the portfolio, functions such as model tuning, deployment, and collaboration are handled through Google's Vertex AI developer service.
- IBM Granite is providing access to foundation models that include language models and document entity extraction, but also target other types of data and use cases such as geospatial; molecular chemistry (often used for drug discovery); IT events (for addressing IT operations); and code generation (a fairly common use case).
- Microsoft is best known for its exclusive partnership with OpenAI. But it is starting to crack the door open with other models including LLaMa 2 and bespoke models designed for Microsoft 365, not to mention its ambitious plans for copilots addressing front office apps and coding.

Upcoming forks in the road for model serving

Model serving services, such as OCI Generative AI, will face several decision points in 2024.

The first is whether they will simply focus on serving and leave lifecycle management to their AI developer services that are geared toward data scientists with coding skills. Or will they extend their role to handling some lifecycle management functions, as Amazon Bedrock has already started to add? For now, the primary tweaks that customers can do is fine tune them when running inferencing on dedicated AI clusters. Our vote is for the latter as this will help democratize usage of Gen AI.

Secondly, there will be the need to broaden model selection. We expect that there will be a Cambrian explosion of third party foundation models because we expect demand for more compact models that are more energy (and cost)-efficient, and are tailored for specific business domains. Get used to the term "Small Language Model." And also get used to the fact that there will be use cases for Gen AI transformer models beyond language (e.g., for images, geospatial, molecular structure, etc.). Model serving services such as OCI Generative AI will need to add new tiers or marketplaces supporting this emerging long tail of foundation models. Additionally, they should make provision for customers to bring their own models to take advantage of the managed automated deployment environment that model serving services provide.

OCI Generative AI Agents

This new service, launched as beta, is designed to outfit the RAG process with a natural/conversational language front end, and then automate and orchestrate the steps to

trigger a semantic search, retrieve documents, generate the embeddings, re-rank the results through an LLM before returning a conversational response with citations to the end user. They are based on a framework outlined in the ReAct paper published by researchers from Google and Princeton University for enabling LLMs to not only provide conversational responses, but prompt follow-up reasoning and actions.

The beta starts with support for retrieving data from text documents indexed by OpenSearch. That covers text, which of course could extend to documents, call transcripts, emails, and other forms of messaging. This is exactly the kind of use case that has brought Gen AI, and services like Chat GPT, to prominence over the past year; it picked up where traditional database and BI queries left off, not to mention keyword-based search.

With the beta release, Oracle is also disclosing its near term roadmap, and the next pieces are logical follow-ons. For starters, extending semantic query to its flagship databases through the vector stores already supported by Oracle and MySQL HeatWave databases, are next on the agenda (both in preview now). This goes to Oracle's sweet spots, extending semantic query to the structured data assets that Oracle is known for. And that picks up from an early gap in Gen AI query that left off with structured data. The brass ring will be bridging the two worlds, where a conversational query could generate a textual response embellished with tables or visualizations from structured data. Besides Oracle, we expect BI providers to support this as well, with unifying semantic query across structured and unstructured data to become table stakes this year.

Oracle is also promising to enhance Gen AI agents that automate semantic query with another next logical step: having responses that include agents that take follow-up actions, such as calling APIs that trigger new enforcement steps by manufacturing quality control systems when factory defects exceed certain thresholds; or providing follow-up responses for call center agents.

OCI Data Science AI Quick Actions

This is about making the data scientist more productive through automation of routine tasks that they must perform when deploying and running LLMs. AI Quick Actions for OCI Data Science, also introduced as beta, will provide no-code access to LLMs (e.g., LLaMa 2 and Mistral 7B) directly from Jupyter notebooks for deployment, fine-tuning, integrating, and scaling LLM models being sent to production.

Specifically, it will support automating model deployment using text generation inference for Hugging Face, vLLM, and Nvidia Triton, along with distributed training with PyTorch, Hugging Face Accelerate or DeepSpeed, along with tasks such as mounting object storage and file systems as a service for checkpointing tasks.

Takeaways

There is little question that Gen AI has become a fast moving target. As another case where consumer technology spurred enterprises to "think different," we would have been surprised if Oracle did not step up to the plate and unleash new Gen AI services. As we noted last year, Gen AI is fast making conversational interfaces the world's most popular API.

We're not surprised that Oracle's Gen AI service moved fast from beta to GA; the fine print with this – and all Gen AI services – is that delivering an MVP approach (minimum viable product) has become the expectation, and that customers want to see where their technology providers are going. That also explains Oracle's openness to talking near-term product roadmaps, something that is normally kept under wraps.

The current spate of announcements addresses the checkboxes and formalizes Oracle's four tiered strategy encompassing:

- **Applications**, which along with databases, are Oracle's forte. As noted below, this is currently Oracle's least mature area with respect to Gen AI, but Oracle is promising Gen AI enhancements in dozens of processes coming with the upcoming Fusion Cloud Applications 24A release. Stay tuned.
- **Data, with** in-database ML and vector/RAG support, already established in Oracle Database 23c and MySQL HeatWave.
- **Prebuilt AI services**, where Oracle's Gen AI introduction are supplementing an existing portfolio of packaged ML and DL services addressing speech, text, vision, and document entity extraction.
- **Infrastructure,** where Oracle features highly scaled "superclusters" powered by Nvidia GPUs.

As we noted above, Oracle's Generative AI services are works in progress. As others currently offer broader selection of foundation models, we expect that Oracle will significantly expand this portfolio in coming months. We would also like to see Oracle introduce capabilities for core model lifecycle management tasks for an audience that would rather work with models through APIs rather than notebooks. For instance, Amazon Bedrock goes beyond model access to automate model evaluation through use of curated data sets and predefined metrics; it also offers a debugging-like capability to trace how models orchestrate processing ("Chain of Thought reasoning").

Oracle is not alone with this challenge. Aside from Bedrock, most of these model via AI services are still rudimentary and often require working with separate services to complete basic tasks. For instance tuning models chosen from Google Cloud's Model Garden requires

Figure 2. Fusion Cloud Applications 24A Gen AI roadmap

Source: Oracle

opening up Vertex AI. There is significant opportunity to add automation and low-code/no-code capabilities for common model lifecycle management functions.

Oracle's big differentiators are on its home turf of databases and enterprise applications. Yes, Oracle has in-database ML, but that has become a checkbox feature across most cloud analytic platforms. But more importantly, Oracle's distinction is how it brings databases and enterprise applications together.

For instance, Figure 2 shows Oracle's ambitious Gen AI plans for Fusion Applications 24A; it is quite an expansive list of processes that will get conversational interfaces; note that SAP has a similarly expansive roadmap, and of course, with its latest generation of S/4 HANA applications, also optimizes on its own database. Oracle's differentiator in the enterprise apps space is that it owns its cloud infrastructure (SAP counters that it won't lock you into any hyperscaler). For Oracle, owning from top to bottom of the stack has a long history, from the engineered systems of Exadata to the design of OCI infrastructure that runs in the public cloud or on-premises via OCI Dedicated region. It could play a key advantage with Gen AI, where compute costs (and carbon footprint) could benefit heavily from optimization and a very fast compute backplane.

Author

Tony Baer, Principal, dbInsight

tony@dbinsight.io

LinkedIn 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.

Published February 2024