NAND RESEARCH

RESEARCH NOTE

INSIDE ORACLE'S NEW GENERATIVE AI FEATURE FOR OCI

STEVE MCDOWELL, CHIEF ANALYST
JANUARY 23, 2024

CONTEXT

Oracle announced the general availability of its OCI Generative AI Service and several substantial enhancements to its data science and cloud offerings. Let's take a look at what Oracle announced.

ANNOUNCEMENT OVERVIEW

Oracle announced a broad range of capabilities centered on generative AI. Key highlights of the announcement include:

- **General Availability of OCI Generative AI Service**: Oracle has launched the OCI Generative AI service, a managed platform designed to integrate large language models (LLMs) into a broad spectrum of enterprise use cases.
- **Support for Multiple Models**: The service now supports models like Meta's Llama 2, Cohere's Command 52/6B models and multilingual embeddings feature for over 100 languages.
- **Enhanced Functionality**: Enhancements include LangChain integration, improved endpoint management, content moderation, and an upgraded GPU cluster management experience with multi-endpoint support.
- **Flexible Model Fine-Tuning**: OCI Generative AI now offers flexible fine-tuning for models like Cohere's Command 52/6B, allowing enterprises to tailor AI applications to their specific business contexts.
- **Diverse Enterprise Use Cases**: The service is positioned to address various enterprise needs, including customer operations automation, marketing personalization, virtual sales assistance, contract acceleration, and strategy and finance monitoring.
- **Beta Release of OCI Generative AI Agents**: This new service translates user queries into tasks for Generative AI components, enhancing enterprise data access through natural language. The initial release includes a retrieval augmented generation (RAG) agent.

- **OCI Data Science AI Quick Actions Feature**: This no-code feature within the OCI Data Science service allows easy access to a range of open-source LLMs, including those from Meta and Mistral AI.
- **Integration in Oracle Databases and Cloud Apps**: Oracle is embedded generative AI capabilities into Oracle Fusion Cloud Applications, Oracle NetSuite, and other industry applications, focusing on tasks like summarization and assisted authoring.
- **Upcoming Enhancements**: Oracle plans further enhancements in AI, including improvements to Oracle Digital Assistant, OCI Language, Document Translation Experience, OCI Vision, OCI Speech, OCI Document Understanding, and OCI Data Science.

Oracle's announcements revolve around making generative AI more accessible and integrated within its broad range of enterprise products and services, signaling a significant commitment to embedding AI at every layer of its technology offerings.

OCI GENERATIVE AI SERVICE

Oracle has made its OCI Generative AI service generally available across its Oracle Cloud Infrastructure. The new offering incorporates enhanced functionality designed to streamline and optimize the use of LLMs in enterprise applications.

Key enhancements within the service include:

- **Support for Multiple Models**: The service extends its capabilities by supporting prominent AI models like Meta's Llama 2 and Cohere's Command 52/6B models, catering to diverse enterprise AI needs.
- **Multilingual Embeddings**: A significant feature is the multilingual embedding capability, which supports over 100 languages, thus broadening the service's applicability across different geographies and linguistic contexts.
- **LangChain Integration**: The service integrates LangChain, which enhances its ability to work with LLMs, making the processing and handling of natural language data more efficient and effective.
- **Improved Endpoint Management**: The service includes enhanced endpoint management, which is crucial for maintaining the performance and scalability of AI applications in a distributed computing environment.
- **Content Moderation**: Content moderation is a key functionality added to the service, essential for ensuring that the outputs generated by AI models align with organizational standards and ethical guidelines.
- **Upgraded GPU Cluster Management**: The service features an improved GPU cluster management experience, including support for hosting multiple endpoints and scaling clusters to handle more model requests. This enhancement is critical for enterprises that require robust computing power for their AI applications.
- **Endpoint Analytics**: The inclusion of endpoint analytics provides valuable insights into the performance and usage of the AI models, enabling better decision-making and resource optimization.

- **Flexible Fine-Tuning**: The service offers flexible fine-tuning capabilities for models like Cohere's Command 52/6B, allowing businesses to customize AI models to fit their specific requirements and contexts.

OCI GENERATIVE AI AGENTS

The new OCI Generative AI Agents, which Oracle announced is now in beta, are a significant development in enterprise AI applications.

These agents are designed to act as intermediaries that translate user queries into actionable tasks for Generative AI components. This allows for a more efficient and intuitive interaction between users and AI systems, enhancing the overall utility of the AI models in business settings.

Key features of the OCI Generative AI Agents include:

- **Task Translation**: The agents are capable of interpreting user queries and converting them into specific tasks that the underlying Generative AI components can execute. This feature simplifies the process of leveraging AI for complex operations.
- **Integration with Generative AI Components**: These agents seamlessly interact with various generative AI components, such as search tools, document corpora, and LLMs. This integration enables them to perform multiple functions based on user requests.
- **Retrieval Augmented Generation (RAG) Agent**: The initial release of the OCI Generative AI Agents includes an RAG agent. This agent combines the general knowledge of LLMs with internal data using Oracle Cloud Infrastructure OpenSearch. It provides contextually relevant answers by accessing and understanding enterprise data.
- **Natural Language Processing**: Users can interact with these agents using natural language, making the AI more accessible and user-friendly, especially for those without specialized skills.
- **Application in Diverse Enterprise Settings**: The agents are designed to handle various use cases like summarizing HR policies or guiding users through business transactions, making them versatile tools in enterprise environments.
- **Beta Release**: The announcement includes the beta release of these agents, indicating ongoing development and refinement based on user feedback and application needs.

The OCI Generative AI Agents service is a significant addition to Oracle's AI offerings. It provides enterprises with advanced tools to automate AI interactions, streamline processes, and extract more value from their AI investments.

DATA SCIENCE AI QUICK ACTIONS

The recent OCI announcements include significant new features in data science, particularly enhancements in AI and machine learning.

Here's what's new in OCI Data Science service:

- **OCI Data Science AI Quick Actions**: This is a notable no-code feature within the OCI Data Science service. It provides easy access to many open-source LLMs, including those developed by Meta, Mistral AI, and others. This feature is designed to democratize access to advanced AI models, making them more accessible to a broader range of users.
- **User-Friendly Access to Pre-Tested Models**: OCI Data Science AI Quick Actions offers a curated list of pre-tested models, including LLMs. This enables users to select and fine-tune models that best suit their specific needs, streamlining the model selection and deployment process.
- **Comprehensive Ecosystem with Integrated Workflows**: The service provides an integrated ecosystem with user-friendly workflows, telemetry, visualizations, and simplified execution processes. This ecosystem aims to make the user experience more intuitive and efficient, especially for those needing more extensive coding expertise.
- **No-Code Interface for Model Deployment**: AI Quick Actions offers a no-code interface that simplifies the execution of tasks, such as fine-tuning, with guided steps to facilitate user interaction with the models.
- **Support for a Range of Open Source LLMs**: The feature provides access to various open-source LLMs, giving users multiple options based on their unique requirements.
- **Flexibility and Customization**: Users can search, filter, and select models that meet their needs. The service also provides options for executing fine-tuning tasks in a few clicks, making it easier to adapt the models for specific use cases.

The enhancements in OCI Data Science are part of Oracle's broader strategy to make AI and machine learning more accessible and applicable across different business scenarios. By providing a no-code platform with access to various pre-tested, open-source models, Oracle is positioning itself to cater to the evolving needs of businesses looking to leverage AI for data analysis, prediction, and other advanced applications.

ANALYSIS

Enterprises need capabilities and use cases that impact business outcomes from generative AI, such as models fine-tuned on an organization's data to deliver unique, organization-specific outputs. This requires a reliable partner to handle various needs like infrastructure configuration and model integration with business applications, a role Oracle has fulfilled for decades across industries.

While every major cloud provider offers some AI-as-a-service capability, none provides an experience as broad and deep as Oracle's. Microsoft comes closest with its various "co-pilot" and Azure AI service offerings, but those are building blocks still need to be assembled.

Oracle has integrated generative AI into every layer of its stack, offering an end-to-end experience built on high-performing AI infrastructure that integrates generative AI capabilities into its extensive range of databases and cloud applications, including recent updates to Oracle Database 23c with AI Vector Search and MySQL HeatWave with Vector Store.

Oracle's integration extends to its cloud applications, such as Oracle Fusion Cloud Applications, Oracle NetSuite, and industry-specific applications like Oracle Health. These applications now come embedded with generative AI capabilities, focusing on functionalities like summarization and assisted authoring.

Oracle's approach ensures that generative AI features are not just add-ons but deeply integrated into its technological offerings' fabric. This gives users a more intuitive experience, making AI more accessible and customizable according to specific business needs. It's a compelling enterprise story.

The OCI-focused AI announcements are simply the next step in Oracle's continuing strategic moves towards enhancing the utility and efficiency of AI-enabled business processes across various Oracle offerings. By embedding AI capabilities directly into its database and cloud applications, Oracle simplifies the AI integration process and enables businesses to leverage AI more effectively in their core operations.

For enterprises interested in a carefully thought-out approach to generative AI—with a service designed to provide a wide range of options, from dedicated AI clusters to fine-tuning models and even on-premises cloud deployments—the OCI Generative AI Service sets the stage for AI solutions that put the customer's needs first.