Amazon Bedrock ENHANCEMENTS

STEVE MCDOWELL, CHIEF ANALYST
7/13/24

CONTEXT

Amazon Web Services (AWS) announced a series of significant enhancements to its Bedrock platform aimed at bolstering the capabilities and reliability of generative AI applications. These enhancements focus on improved data connectivity, advanced safety features, and robust governance mechanisms.

This research note provides an in-depth analysis of these enhancements and their strategic implications for enterprises leveraging AWS Bedrock.

BACKGROUND: WHAT IS AMAZON BEDROCK?

Amazon Bedrock is a fully managed service from Amazon Web Services (AWS) designed to help developers build and scale generative AI applications using foundational models (FMs).

Here's an overview of its key features and benefits:

Key Features

1. **Access to Foundational Models (FMs)**:
    - Provides access to pre-trained, state-of-the-art foundational models from leading AI model providers like Amazon, AI21 Labs, Anthropic, and Stability AI.
    - Offers a range of models for different tasks, including text generation, image generation, and more.
2. **Simplified Model Customization**:
    - Allows developers to customize foundational models with their data using simple API calls.

- Supports the fine-tuning of models to meet specific requirements and improve the relevance and accuracy of AI-generated outputs.

3. **Integrated Development Environment**:
   - Offers an intuitive development environment that integrates seamlessly with other AWS services.
   - Provides tools for building, deploying, and scaling generative AI applications without the need for deep expertise in machine learning.
4. **Scalability and Performance**:
   - Ensures high performance and scalability, enabling developers to handle large-scale generative AI workloads.
   - Leverages the robust infrastructure of AWS to provide reliable and efficient AI model serving.
5. **Secure and Compliant**:
   - Ensures that data is handled securely, with support for encryption and compliance with industry standards and regulations.
   - Provides mechanisms for data privacy and protection, crucial for enterprise applications.
6. **Flexible Pricing**:
   - Offers a flexible pricing model, allowing developers to pay only for the resources they use.
   - Supports cost-effective development and deployment of AI applications.

ENHANCED: AGENTS FOR BEDROCK

Amazon Bedrock introduces two new fully managed capabilities in preview for its generative AI applications: memory retention across interactions and code interpretation support. These enhancements allow agents to perform multistep tasks across various systems and data sources more effectively.

Key Features

1. **Memory Retention Across Multiple Interactions**:

- **Persistent Memory**: Agents can retain conversation summaries with users, allowing for a smoother, more adaptive experience in complex tasks like booking flights or processing insurance claims.
- **Context Awareness**: Facilitates personalized and efficient interactions by remembering user preferences and ongoing interactions.
- **Secure Storage**: Each user's conversation history is stored under a unique memory identifier (ID), ensuring privacy and separation between users.

2. **Code Interpretation Support**:
   - **Dynamic Code Generation**: Agents can generate and run code snippets within a secure, sandboxed environment, addressing complex use cases such as data analysis, visualization, text processing, equation solving, and optimization problems.
   - **File Processing**: Supports diverse data types and formats, including CSV, XLS, YAML, JSON, DOC, HTML, MD, TXT, and PDF.
   - **Enhanced Data Interpretation**: Allows agents to generate charts, making data analysis more accessible.

UPDATED: BEDROCK GUARDRAILS

Guardrails for Amazon Bedrock allows safeguards based on application requirements and responsible AI policies. These guardrails help prevent undesirable content, block prompt attacks (such as prompt injection and jailbreaks), and remove sensitive information for privacy.

Guardrails can be configured for different scenarios and applied across foundation models (FMs) on Amazon Bedrock as well as custom and third-party FMs. Additionally, Guardrails integrates with Agents and Knowledge Bases for Amazon Bedrock.

KEY FEATURES AND CAPABILITIES

1. **Customizable Safeguards**:
   - **Denied Topics, Content Filters, Sensitive Information Filters, and Word Filters**: These initial safeguards were introduced at general availability in April 2024.

- **Blocks Harmful Content**: Up to 85% more harmful content is blocked than native FM protections.
- **Filters Hallucinated Responses**: Over 75% of hallucinated responses are filtered for RAG and summarization workloads.

2. **Integration and Centralized Governance**:
   - **ApplyGuardrail API**: Evaluates input prompts and model responses for all FMs, enabling centralized governance across generative AI applications.
   - **Contextual Grounding Checks**: Detects hallucinations in model responses based on reference sources and user queries, ensuring responses are grounded in enterprise data and relevant to user queries.

NEW CAPABILITIES

1. **Contextual Grounding Checks**:
   - **Grounding**: Ensures responses are factually correct based on reference sources. Responses below the grounding threshold are blocked.
   - **Relevance**: Ensures responses are relevant to the user's query. Responses below the relevance threshold are blocked.

2. **Use Cases**:
   - Enhances response quality in applications like RAG, summarization, and information extraction.
   - Improves the trustworthiness of AI applications by filtering inaccurate responses not grounded in enterprise data.

UPDATED: KNOWLEDGE BASES FOR AMAZON BEDROCK

Knowledge Bases for Amazon Bedrock enable foundation models (FMs) and agents to retrieve contextual information from your company's private data sources for Retrieval-Augmented Generation (RAG), enhancing the relevance, accuracy, and customization of responses.

Amazon has updated the feature to supplement its existing Amazon S3 support and also to support connections to web domains, Confluence,

Salesforce, and SharePoint (in preview). This allows RAG applications to access public data from web domains (e.g., company social media feeds) and existing company data from Confluence, Salesforce, and SharePoint

ANALYSIS

The new capabilities of Amazon Bedrock, specifically memory retention and code interpretation, represent a substantial enhancement in the platform's ability to handle complex, multistep tasks and provide personalized, adaptive experiences.

These features broaden the applicability of generative AI and enhance its value proposition for businesses across various industries. By focusing on improving user experience, operational efficiency, and data security, Amazon Bedrock is poised to drive significant advancements in the adoption and effectiveness of AI-driven solutions.

The enhancements to AWS Bedrock offer enterprises comprehensive tools to enhance AI applications' safety, relevance, and efficiency. These developments not only bolster AI governance and compliance but also expand the functional capabilities of AI, enabling more sophisticated and reliable solutions. As AI plays a critical role in business innovation, AWS Bedrock's enhancements provide a robust foundation for enterprises to build upon, ensuring secure, effective, and cutting-edge AI deployment.

.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.