HealthManagement
Promoting Management and Leadership

AWS Announces Powerful New Offerings to Accelerate Generative AI Innovation

Now generally available, Amazon Bedrock is a fully managed service that offers a choice of high-performing foundation models from leading AI companies, along with a broad set of capabilities to build generative AI applications, simplifying development while maintaining privacy and security

With the addition of Amazon Titan Embeddings and Meta's Llama 2 models, Amazon Bedrock gives customers even greater choice and flexibility to find the right models for each use case

New Amazon CodeWhisperer capability will deliver customized, generative AI-powered code suggestions that leverage an organization's own internal codebase, increasing developer productivity

Generative BI dashboard-authoring capabilities in Amazon QuickSight make it faster and easier for business analysts to explore data and create compelling visuals simply by describing what they want in natural language

Companies such as adidas, BMW Group, GoDaddy, Merck, NatWest Group, Persistent, the PGA TOUR, Takenaka Corporation, and Traeger Grills among customers applying generative AI innovations from AWS to transform their products and services

Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company announced five generative artificial intelligence (AI) innovations, so organizations of all sizes can build new generative AI applications, enhance employee productivity, and transform their businesses. Today's announcement includes the general availability of Amazon Bedrock, a fully managed service that makes foundation models (FMs) from leading AI companies available through a single application programming interface (API). To give customers an even greater choice of FMs, AWS also announced that Amazon Titan Embeddings model is generally available and that Llama 2 will be available as a new model on Amazon Bedrock–making it the first fully managed service to offer Meta's Llama 2 via an API. For organizations that want to maximize the value their developers derive from generative AI, AWS is also announcing a new capability (available soon in preview) for Amazon CodeWhisperer, AWS's AI-powered coding companion, that securely customizes CodeWhisperer's code suggestions based on an organization's own internal codebase. To increase the productivity of business analysts, AWS is releasing a preview of Generative Business Intelligence (BI) authoring capabilities for Amazon QuickSight, a unified BI service built for the cloud, so customers can create compelling visuals, format charts, perform calculations, and more–all by simply describing what they want in natural language. From Amazon Bedrock and Amazon Titan Embeddings to CodeWhisperer and QuickSight, these innovations add to the capabilities AWS provides customers at all layers of the generative AI stack—for organizations of all sizes and with enterprise-grade security and privacy, selection of best-in-class models, and powerful model customization capabilities. To get started with generative AI on AWS, visit aws.amazon.com/generative-ai/.

"Over the last year, the proliferation of data, access to scalable compute, and advancements in machine learning (ML) have led to a surge of interest in generative AI, sparking new ideas that could transform entire industries and reimagine how work gets done," said Swami Sivasubramanian, vice president of Data and AI at AWS. "With enterprise-grade security and privacy, a choice of leading FMs, a data-first approach, and our high-performance, cost-effective infrastructure, organizations trust AWS to power their businesses with generative AI solutions at every layer of the stack. Today's announcement is a major milestone that puts generative AI at the fingertips of every business, from startups to enterprises, and every employee, from developers to data analysts. With powerful, new innovations, AWS is bringing greater security, choice, and performance to customers, while also helping them to tightly align their data strategy across their organization, so they can make the most of the transformative potential of generative AI."

Organizations of all sizes and across industries want to get started with generative AI to transform their operations, reimagine how they solve

tough problems, and create new user experiences. While recent advancements in generative AI have captured widespread attention, many businesses have not been able to take part in this transformation. These organizations want to get started with generative AI, but they are concerned about the security and privacy of these tools. They also want the ability to choose from a wide variety of FMs, so they can test different models to determine which works best for their unique use case. Customers also want to make the most of the data they already have by privately customizing models to create differentiated experiences for their end users. Finally, they need tools that help them bring these new innovations to market quickly and the infrastructure to deploy their generative AI applications on a global scale. That is why customers such as adidas, Alida, Asurion, BMW Group, Clariant, Genesys, Glide Publishing Platform, GoDaddy, Intuit, LexisNexis Legal & Professional, Lonely Planet, Merck, NatWest Group, Perplexity AI, Persistent, Quext, RareJob Technologies, Rocket Mortgage, SnapLogic, Takenaka Corporation, Traeger Grills, the PGA TOUR, United Airlines, Verint, Verisk, WPS, and more have turned to AWS for generative AI.

Amazon Bedrock is now generally available to help more customers build and scale generative AI applications

Amazon Bedrock is a fully managed service that offers a choice of high-performing FMs from leading AI companies including AI21 Labs, Anthropic, Cohere, Meta, Stability AI, and Amazon, along with a broad set of capabilities that customers need to build generative AI applications, simplifying development while maintaining privacy and security. The flexibility of FMs makes them applicable to a wide range of use cases, powering everything from search to content creation to drug discovery. However, a few things stand in the way of most businesses looking to adopt generative AI. First, they need a straightforward way to find and access high-performing FMs that give outstanding results and are best-suited to their purposes. Second, customers want application integration to be seamless, without managing huge clusters of infrastructure or incurring large costs. Finally, customers want easy ways to use the base FM and build differentiated apps with their data. Since the data customers want for customization is incredibly valuable IP, it must stay completely protected, secure, and private during that process, and customers want control over how the data is shared and used.

With Amazon Bedrock's comprehensive capabilities, customers can easily experiment with a variety of top FMs and customize them privately with their proprietary data. Additionally, Amazon Bedrock offers differentiated capabilities like creating managed agents that execute complex business tasks—from booking travel and processing insurance claims to creating ad campaigns and managing inventory—without writing any code. Since Amazon Bedrock is serverless, customers do not have to manage any infrastructure, and they can securely integrate and deploy generative AI capabilities into their applications using the AWS services they are already familiar with. Built with security and privacy in mind, Amazon Bedrock makes it easy for customers to protect sensitive data. Customers can use AWS PrivateLink to establish a private, secure connection between Amazon Bedrock and their virtual private cloud (VPC) without exposing any traffic to the public internet. And, for customers in highly regulated industries, Amazon Bedrock is a HIPAA eligible service and can be used in compliance with GDPR, allowing even more customers to benefit from generative AI.

Amazon Bedrock continues to expand its model selection with Amazon Titan Embeddings and Llama 2 to help every customer find the right model for their use case

No single model is optimized for every use case, and to unlock the value of generative AI, customers need access to a variety of models to discover what works best based on their needs. That is why Amazon Bedrock makes it easy for customers to find and test a selection of leading FMs, including models from AI21 Labs, Anthropic, Cohere, Meta, Stability AI, and Amazon, through a single API. Additionally, as part of a recently announced strategic collaboration, all future FMs from Anthropic will be available within Amazon Bedrock with early access to unique features for model customization and fine-tuning capabilities. With today's announcement, Amazon Bedrock continues to broaden its selection of FMs with access to new models:

- **Amazon Titan Embeddings now generally available:** Amazon Titan FMs are a family of models created and pretrained by AWS on large datasets, making them powerful, general purpose capabilities to support a variety of use cases. The first of these models generally available to customers, Amazon Titan Embeddings is a large language model (LLM) that converts text into numerical representations called embeddings to power search, personalization, and retrieval-augmented generation (RAG) use cases. FMs are well-suited to a wide variety of tasks, but they can only respond to questions based on learnings from the training data and contextual information in a prompt, limiting their effectiveness when responses require timely knowledge or proprietary data. To augment FM responses with additional data, many organizations turn to RAG, a popular model-customization technique where the FM connects to a knowledge source that it can reference to augment its responses. To get started with RAG, customers must first access an embedding model to convert their data into embeddings that allow the FM to more easily understand the semantic meaning and relationships among data. Building an embeddings model requires massive amounts of data and resources as well as deep ML expertise, making it impractical for many customers to build themselves and putting RAG out of reach for many organizations. Amazon Titan Embeddings makes it easier for customers to start with RAG to extend the power of any FM using their proprietary data. Amazon Titan Embeddings supports more than 25 languages and a context length of up to 8,192 tokens, making it well-suited to work with single words, phrases, or entire documents based on the customer's use case. The model returns output vectors of 1,536 dimensions, giving it a high degree of accuracy, while also optimizing for low-latency, cost-effective results.
- **Llama 2 coming in the next few weeks:** Amazon Bedrock is the first fully managed generative AI service to offer Llama 2, Meta's next-generation LLM, through a managed API. Llama 2 models come with significant improvements over the original Llama models, including being trained on 40% more data and having a longer context length of 4,000 tokens to work with larger documents. Optimized to provide a fast response on AWS infrastructure, the Llama 2 models available via Amazon Bedrock are ideal for dialogue use cases. Customers will be able to build generative AI applications powered by the 13B and 70B parameter Llama 2 models, without the need to setup and manage any infrastructure.

New Amazon CodeWhisperer capability will allow customers to securely customize CodeWhisperer suggestions using their private codebase to unlock new levels of developer productivity

Trained on billions of lines of Amazon and publicly available code, Amazon CodeWhisperer is an AI-powered coding companion that improves developer productivity. While developers frequently use CodeWhisperer for day-to-day work, they sometimes need to incorporate their organization's internal, private codebase (e.g., internal APIs, libraries, packages, and classes) into an application, none of which are included in CodeWhisperer's training data. However, internal code can be difficult to work with because documentation may be limited, and there are no public resources or forums where developers can ask for help. For example, to write a function for an ecommerce website that removes an item from a shopping cart, a developer must first understand the existing APIs, classes, and other internal code used to interact with the application. Previously, a developer might have spent hours examining previously written internal code to find information they needed and understand how it works. Even after finding the right resources, they must inspect the code closely to ensure it adheres to company coding best practices and does not repeat any flaws or vulnerabilities present in the reference code.

Amazon CodeWhisperer's new customization capability will unlock the full potential of generative AI-powered coding by securely leveraging a customer's internal codebase and resources to provide recommendations that are customized to their unique requirements. Developers save time through improved relevancy of code suggestions across a range of tasks. To start, an administrator connects to their private code repository from a source, such as GitLab or Amazon S3, and schedules a job to create their own customization. When creating a customization, CodeWhisperer leverages a variety of model- and context-customization techniques to learn from the customer's repository and enhance its real-time code suggestions, so developers spend less time searching for the right answer to undifferentiated problems and more time focused on creating new and differentiated experiences. Administrators can then centrally manage all customizations from the AWS Console, allowing them to view evaluation metrics, estimate how each customization will perform, and selectively deploy them to specific developers across the company to restrict access to sensitive code. By selectively choosing only the highest-quality repositories, administrators can ensure the customizations that CodeWhisperer provides omit deprecated code and meet the organization's quality and security standards. Built with enterprise-grade security and privacy in mind, the capability keeps customizations completely private, and the underlying FM powering CodeWhisperer does not use the customizations for training, protecting customers' valuable intellectual property. This customization capability will be available soon to customers in preview, as part of a new CodeWhisperer Enterprise Tier. CodeWhisperer customizations are also secure by default, and AWS does not store or log any customer content when handling requests from a developer's IDE that use the Amazon CodeWhisperer Professional Tier or Enterprise Tier.

New Generative BI authoring capabilities in Amazon QuickSight help business analysts easily create and customize visuals using natural-language commands

Amazon QuickSight is a unified BI service built for the cloud that offers interactive dashboards, paginated reports, and embedded analytics, plus natural-language querying capabilities using QuickSight Q, so every user in the organization can access insights they need in the format they prefer. Business analysts often spend hours with BI tools exploring disparate data sources, adding calculations, and creating and refining visualizations before providing them in dashboards to business stakeholders. To create a single chart, an analyst must first find the correct data source, identify the data fields, set up filters, and make necessary customizations to ensure the visual is compelling. If the visual requires a new calculation (e.g., year-to-date sales), the analyst must identify the necessary reference data and then create, verify, and add the visual to the report. Organizations would benefit from reducing the time that business analysts spend manually creating and adjusting charts and calculations so that they can devote more time to higher-value tasks.

The new Generative BI authoring capabilities extend the natural-language querying of QuickSight Q beyond answering well-structured questions (e.g., "What are the top 10 products sold in California?") to help analysts quickly create customizable visuals from question fragments (e.g., "top 10 products"), clarify the intent of a query by asking follow-up questions, refine visualizations, and complete complex calculations. Business analysts simply describe the desired outcome, and QuickSight generates compelling visuals that can be easily added to a dashboard or report with a single click. For example, an analyst can ask QuickSight Q to create a visualization for the "monthly trend for sneaker sales in 2022 and 2023," and the service automatically selects the appropriate data and plots the information using the chart format (e.g., line chart or bar chart) that makes the most sense based on the request. QuickSight Q will also offer related questions to help analysts clarify ambiguous cases when multiple data fields match their query (e.g., should the chart include the total dollar value of sneaker sales or the number of units sold). Once the analyst has the initial visualization, they can also add complex calculations, change chart types, and refine visuals using natural-language prompts. The new Generative BI authoring capabilities in QuickSight Q make it fast and easy for business analysts to create compelling visuals and reduce the time to deliver the insights needed to inform data-driven decisions at scale.

Customers across industries are leveraging generative AI services from AWS to create new applications, accelerate developer productivity, and help analysts derive insights faster

adidas is one of the largest sports brands in the world. "We were excited to be part of the Amazon Bedrock preview and get our hands on the service," said Daniel Eichten, vice president of Enterprise Architecture at adidas. "Amazon Bedrock quickly became a highly valued addition to our generative AI toolkit, allowing us to focus on the core aspects of our LLM projects, while letting it handle the heavy lifting of managing infrastructure. Using Amazon Bedrock, we have developed a generative AI solution that gives the community of adidas engineers the ability to find information and answers from our knowledge base through a single conversational interface, covering everything from getting started to highly technical questions."

GoDaddy is a leading domain registrar, commerce, and web hosting company serving more than 20 million customers. "At GoDaddy, we aim to help everyday entrepreneurs succeed by giving them the tools for establishing their business, creating a website and brand, marketing to their customers, and managing their work," said Travis Muhlestein, chief data and analytics officer at GoDaddy. "Today, one of the biggest challenges facing entrepreneurs and microbusinesses is the lack of funding, time, and resources. We heard from customers that they want to accelerate content creation for end-user engagement, thereby enabling them to expand their business. We are using Amazon Bedrock to build a generative

AI service that will help customers easily set up their businesses online, and to more efficiently connect them to relevant suppliers, consumers, resources, and funding opportunities."

Merck is a research-intensive biopharmaceutical company that has been discovering and developing innovative medicines and vaccines to save and improve lives for more than 130 years. "All across the pharma value chain, there are manual, time-intensive processes that detract from more impactful work, as well as data that is not being effectively harnessed to improve employee, customer, and patient experiences," said Suman Giri, executive director of Data Science for Merck. "With Amazon Bedrock, we have quickly built generative AI capabilities to make things like knowledge-mining and market research more efficient. In our U.S. patient-level analytics workflows, we can use those capabilities to provide patient insights, improve lives, and grow commercial reach, while closing gaps in data-sharing and building our data governance ecosystem for responsible generative AI."

NatWest Group is a leading bank in the U.K., serving more than 19 million people and supporting communities, families, and businesses. "The world has changed over the past 12 months with the expansion of generative AI technology," said Zachery Anderson, chief analytics and data officer at NatWest Group. "This technology has raised the bar in terms of the types of services, products, and support that our customers expect in meeting their financial goals. Amazon Bedrock allows us to leverage the latest generative AI models in a secure and scalable platform, which our teams of data scientists, engineers, and technologists are using to experiment and build new services. With these tools, we will be able to combat the next generation of threats from financial crime, as well as allow customers and colleagues access to the information they need, in the format they want, when they need it."

The PGA TOUR is the world's premier membership organization for touring professional golfers. "Creating unique and engaging fan experiences is a top priority for the PGA TOUR," said Scott Gutterman, senior vice president of Digital Operations at the PGA TOUR. "Together with AWS, we have been transforming the way golf content is created, distributed, and experienced. Now using Amazon Bedrock, we will break new ground as we reimagine how golf fans connect with and follow our sport. Leveraging generative AI will enable us to create new touchpoints for our fans, and create an AI platform to evaluate players' game performance and make recommendations for adjustments on different holes or courses. AWS allows us to unlock more value from our data while providing a secure environment to protect our intellectual property."

Founded in 1610, Takenaka Corporation is one of Japan's leading construction companies and collaborated with AWS to develop the Building 4.0 Digital Platform, which uses data and analytics to drive efficiencies and new value creation across its business operations. "To raise productivity and accelerate business developments, architecture, engineering, and construction firms need to focus on digitizing their entire operations, including physical sites," said Dr. Keizo Iwashita, executive officer, Digital Division at Takenaka Corporation. "Generative AI is poised to deliver significant efficiency gains and is one of the key focuses of Takenaka Corporation's digital transformation efforts. We plan to use Amazon Bedrock and Amazon Kendra to build an application that enables employees to easily find information from vast amounts of construction industry laws and regulations, internal guidelines, and best practices to make smarter, faster business decisions, and improve work-life balance."

Persistent is a global services and solutions company delivering digital engineering and enterprise modernization services to customers. "At Persistent, we are equipping our 16,000-plus engineering organization with Amazon CodeWhisperer to build and deliver industry applications faster and more securely in a responsible way," said Pandurang Kamat, chief technology officer at Persistent. "While we have already seen CodeWhisperer accelerate productivity across a range of tasks, our developers also need to work with internal code that is not included in CodeWhisperer's suggestions, which limits the productivity gains when working with recommended code. Several teams have started to leverage CodeWhisperer's new customization capability to help maximize the benefits from generative AI-powered code suggestions, and we are already seeing great results. In a recent study conducted in collaboration with AWS, we found that developers using the customization capability were able to complete their coding tasks up to 28% faster, on average, than developers using standard CodeWhisperer, and in certain instances. We're excited to expand access to this new capability across more teams so our developers can increase our productivity even further."

The BMW Group is a global manufacturer of premium automobiles and motorcycles. "Here at BMW, our regional specialists are focused on optimizing inventory throughout our supply chain," said Christoph Albrecht, data engineering and analytics expert at BMW Group. "They get regular requests from stakeholders like our board members or supply chain specialists to create new dashboard views for them to analyze the latest trends. QuickSight's new Q-powered authoring experience is a huge time saver to create calculations without stopping for reference, build visuals fast, and then refine the visual presentation for a precise experience, all with natural language. The regional specialists can impress our business users with a quick turnaround and they can make critical decisions more quickly."

Traeger Grills is a leading provider of smokers, grills, and barbeque products. "Our business is constantly evolving and developing new data needs, which led us to create and update dashboards and reports," said Corey Savory-Venzke, vice president of Customer Experience at Traeger Grills. "QuickSight enables our operations teams to deliver data to users across a variety of use cases, from distribution-center forecasts to reporting Amazon Connect call-center metrics. QuickSight Q has shown us the power of natural-language experiences to accelerate data work by helping our business users get insights instantly. We are excited to see the additional Generative BI capabilities for authors because they can raise our speed to respond to these changing business needs to a new level. Natural-language experiences like these are fundamentally changing the way people work."

Source: Amazon Web Services

Published on : Thu, 28 Sep 2023