AWS Machine Learning Blog

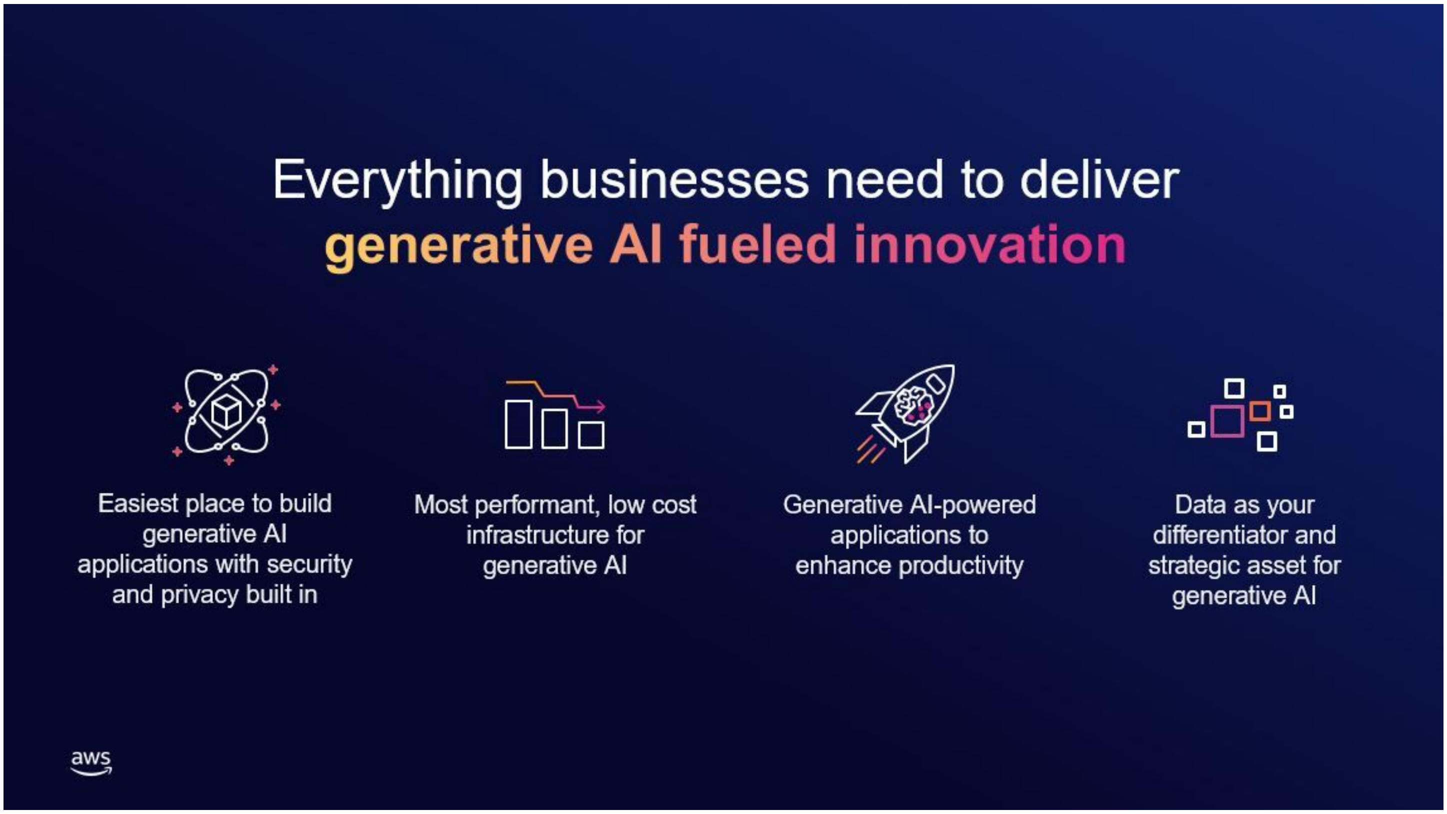
Announcing New Tools to Help Every Business Embrace Generative AI

by Swami Sivasubramanian | on 28 SEP 2023 | in Announcements, Artificial Intelligence, Generative AI | Permalink | 💬 Comments | ➦ Share

From startups to enterprises, organizations of all sizes are getting started with generative AI. They want to capitalize on generative AI and translate the momentum from betas, prototypes, and demos into real-world productivity gains and innovations. But what do organizations need to bring generative AI into the enterprise and make it real? When we talk to customers, they tell us they need security and privacy, scale and price-performance, and most importantly tech that is relevant to their business. We are excited to announce new capabilities and services today to allow organizations big and small to use generative AI in creative ways, building new applications and improving how they work. At AWS, we are hyper-focused on helping our customers in a few ways:

- Making it easy to build generative AI applications with security and privacy built in
- Focusing on the most performant, low cost infrastructure for generative AI so you can train your own models and run inference at scale
- Providing generative AI-powered applications for the enterprise to transform how work gets done
- Enabling data as your differentiator to customize foundation models (FMs) and make them an expert on your business, your data, and your company

To help a broad range of organizations build differentiated generative AI experiences, AWS has been working hand-in-hand with our customers, including BBVA, Thomson Reuters, United Airlines, Philips, and LexisNexis Legal & Professional. And with the new capabilities launched today, we look forward to enhanced productivity, improved customer engagement, and more personalized experiences that will transform how companies get work done.



Everything businesses need to deliver
generative AI fueled innovation

Easiest place to build generative AI applications with security and privacy built in

Most performant, low cost infrastructure for generative AI

Generative AI-powered applications to enhance productivity

Data as your differentiator and strategic asset for generative AI

Announcing the general availability of Amazon Bedrock, the easiest way to build generative AI applications with security and privacy built in

Customers are excited and optimistic about the value that generative AI can bring to the enterprise. They are diving deep into the technology to learn the steps they need to take to build a generative AI system in production. While recent advancements in generative AI have captured widespread attention, many businesses have not been able to take part in this transformation. Customers tell us they need a choice of models, security and privacy assurances, a data-first approach, cost-effective ways to run models, and capabilities like prompt engineering, retrieval augmented generation (RAG), agents, and more to create customized applications. That is why on April 13, 2023, we announced Amazon Bedrock, the easiest way to build and scale generative AI applications with foundation models. Amazon Bedrock is a fully managed service that offers a choice of high-performing foundation models from leading providers like 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. 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.

Since April, we have seen firsthand how startups like Coda, Hurone AI, and Nexxiot; large enterprises like adidas, GoDaddy, Clariant, and Broadridge; and partners like Accenture, BCG, Leidos, and Mission Cloud are already using Amazon Bedrock to securely build generative AI applications across industries. Independent software vendors (ISVs) like Salesforce are now securely integrating with Amazon Bedrock to enable their customers to power generative AI applications. Customers are applying generative AI to new use cases; for example, Lonely Planet, a premier travel media company, worked with our Generative AI Innovation Center to introduce a scalable AI platform that organizes book content in minutes to deliver cohesive, highly accurate travel recommendations, reducing itinerary generation costs by nearly 80%. And since then, we have continued to add new capabilities, like agents for Amazon Bedrock, as well as support for new models, like Cohere and the latest models from Anthropic, to offer our customers more choice and make it easier to create generative AI-based applications. Agents for Bedrock are a game changer, allowing LLMs to complete complex tasks based on your own data and APIs, privately, securely, with setup in minutes (no training or fine tuning required).

Today, we are excited to share new announcements that make it easier to bring generative AI to your organization:

- **General availability of Amazon Bedrock** to help even more customers build and scale generative AI applications
- **Expanded model choice with Llama 2** (coming in the next few weeks) and **Amazon Titan Embeddings** gives customers greater choice and flexibility to find the right model for each use case and power RAG for better results
- **Amazon Bedrock is a HIPAA eligible service and can be used in compliance with GDPR**, allowing even more customers to benefit from generative AI
- **Provisioned throughput** to ensure a consistent user experience even during peak traffic times

With the general availability of Amazon Bedrock, more customers will have access to Bedrock's comprehensive capabilities. Customers can easily experiment with a variety of top FMs, customize them privately with their data using techniques such as fine tuning and RAG, and create managed agents that execute complex business tasks—from booking travel and processing insurance claims to creating ad campaigns and managing inventory—all without writing any code. Since Amazon Bedrock is serverless, customers don't 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.

Second, model choice has been a cornerstone of what makes Amazon Bedrock a unique, differentiated service for our customers. This early in the adoption of generative AI, there is no single model that unlocks all the value of generative AI, and customers need the ability to work with a range of high-performing models. We are excited to announce the general availability of Amazon Titan Embeddings and coming in the next few weeks availability of Llama 2, Meta's next generation large language model (LLM) – joining existing model providers AI21 Labs, Anthropic, Cohere, Stability AI, and Amazon in further expanding choice and flexibility for customers. 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 can now build generative AI applications powered by Llama 2 13B and 70B parameter models, without the need to set up and manage any infrastructure.

Amazon Titan FMs are a family of models created and pretrained by AWS on large datasets, making them powerful, general purpose capabilities built to support a variety of use cases. The first of these models generally available to customers, Amazon Titan Embeddings, is an LLM that converts text into numerical representations (known as embeddings) to power RAG use cases. FMs are well suited for 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. Data is the difference between a general generative AI application and one that truly knows your business and your customer. To augment FM responses with additional data, many organizations turn to RAG, a popular model-customization technique where an FM connects to a knowledge source that it can reference to augment its responses. To get started with RAG, customers first need access to an embedding model to convert their data into vectors that allow the FM to more easily understand the semantic meaning and relationships between data. Building an embeddings model requires massive amounts of data, resources, and ML expertise, putting RAG out of reach for many organizations. Amazon Titan Embeddings makes it easier for customers to get started 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. With new models and capabilities, it's easy to use your organization's data as a strategic asset to customize foundation models and build more differentiated experiences.

Third, because the data customers want to use for customization is such valuable IP, they need it to remain secure and private. With security and privacy built in since day one, Amazon Bedrock customers can trust that their data remains protected. None of the customer's data is used to train the original base FMs. All data is encrypted at rest and in transit. And you can expect the same AWS access controls that you have with any other AWS service. Today, we are excited to build on this foundation and introduce new security and governance capabilities – Amazon Bedrock is now a HIPAA eligible service and can be used in compliance with GDPR, allowing even more customers to benefit from generative AI. New governance capabilities include integration with Amazon CloudWatch to track usage metrics and build customized dashboards and integration with AWS CloudTrail to monitor API activity and troubleshoot issues. These new governance and security capabilities help organizations unlock the potential of generative AI, even in highly regulated industries, and ensure that data remains protected.

Finally, certain periods of the year, like the holidays, are critical for customers to make sure their users can get uninterrupted service from applications powered by generative AI. During these periods, customers want to ensure their service is available to all of its customers regardless of the demand. Amazon Bedrock now allows customers to reserve throughput (in terms of tokens processed per minute) to maintain a consistent user experience even during peak traffic times.

Together, the new capabilities and models we announced today for Amazon Bedrock will accelerate how quickly enterprises can build more personalized applications and enhance employee productivity. In concert with our ongoing investments in ML infrastructure, Amazon Bedrock is the best place for customers to build and scale generative AI applications.

To help customers get started quickly with these new features, we are adding a new generative AI training for

Amazon Bedrock to our collection of digital, on-demand training courses. Amazon Bedrock – Getting Started is a free, self-paced digital course that introduces learners to the service. This 60-minute course will introduce developers and technical audiences to Amazon Bedrock's benefits, features, use cases, and technical concepts.

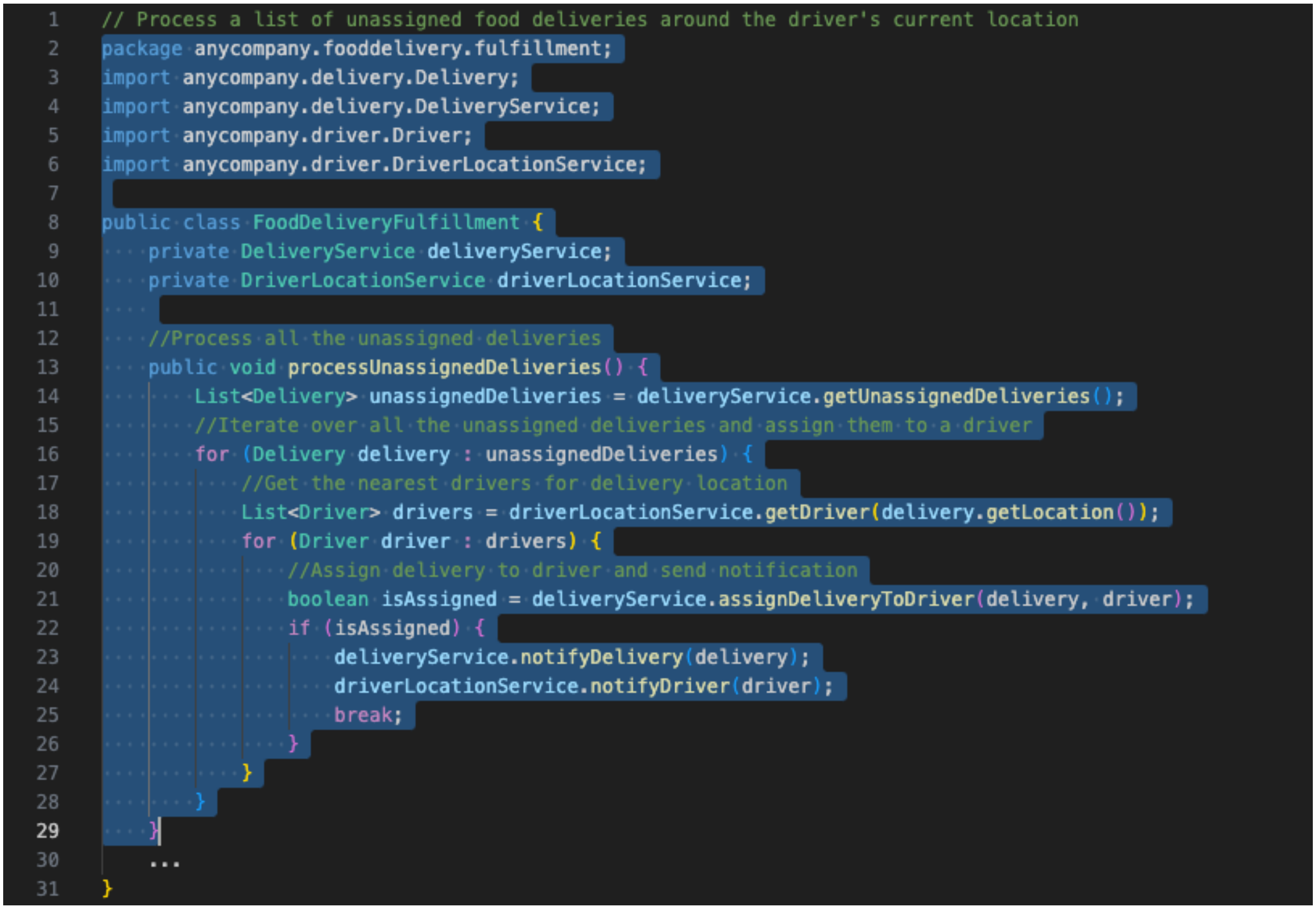
Announcing Amazon CodeWhisperer customization capability to generate more relevant code recommendations informed by your organization's code base

At AWS, we are building powerful new applications that transform how our customers get work done with generative AI. In April 2023, we announced the general availability of Amazon CodeWhisperer, an AI coding companion that helps developers build software applications faster by providing code suggestions across 15 languages, based on natural language comments and code in a developer's integrated developer environment (IDE). CodeWhisperer has been trained on billions of lines of publicly available code to help developers be more productive across a wide range of tasks. We have specially trained CodeWhisperer on high-quality Amazon code, including AWS APIs and best practices, to help developers be even faster and more accurate generating code that interacts with AWS services like Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), and AWS Lambda. Customers from Accenture to Persistent to Bundesliga have been using CodeWhisperer to help make their developers more productive.

Many customers also want CodeWhisperer to include their own internal APIs, libraries, best practices, and architectural patterns in its suggestions, so they can speed up development even more. Today, AI coding companions are not able to include these APIs in their code suggestions because they are typically trained on publicly available code, and so aren't aware of a company's internal code. For example, to build a feature for an ecommerce website that lists items in a shopping cart, developers have to find and understand existing internal code, such as the API that provides the description of items, so they can display the description in the shopping cart. Without a coding companion capable of suggesting the correct, internal code for them, developers have to spend hours digging through their internal code base and documentation to complete their work. Even after developers are able to find the right resources, they have to spend more time reviewing the code to make sure it follows their company's best practices.

Today, we are excited to announce a new **Amazon CodeWhisperer customization capability**, which enables CodeWhisperer to generate even better suggestions than before, because it can now include your internal APIs, libraries, best practices, and architectural patterns. This capability uses the latest model and context customization techniques and will be available in preview soon as part of a new CodeWhisperer Enterprise Tier. With this capability, you can securely connect your private repositories to CodeWhisperer, and with a few clicks, customize CodeWhisperer to generate real-time recommendations that include your internal code base. For example, with a CodeWhisperer customization, a developer working in a food delivery company can ask CodeWhisperer to provide recommendations that include specific code related to the company's internal services, such as "Process a list of unassigned food deliveries around the driver's current location." Previously, CodeWhisperer would not know the correct internal APIs for "unassigned food deliveries" or "driver's current location" because this isn't publicly available information. Now, once customized on the company's internal code base, CodeWhisperer understands the intent, determines which internal and public APIs are best suited to the task, and generates code recommendations for the developer. The CodeWhisperer customization capability can save developers hours spent searching and modifying sparsely documented code, and helps onboard developers who are new to the company faster.

In the following example, after creating a private customization, AnyCompany (a food delivery company) developers get CodeWhisperer code recommendations that include their internal APIs and libraries.



```java
// Process a list of unassigned food deliveries around the driver's current location
package anycompany.fooddelivery.fulfillment;
import anycompany.delivery.Delivery;
import anycompany.delivery.DeliveryService;
import anycompany.driver.Driver;
import anycompany.driver.DriverLocationService;

public class FoodDeliveryFulfillment {
    private DeliveryService deliveryService;
    private DriverLocationService driverLocationService;
    
    //Process all the unassigned deliveries
    public void processUnassignedDeliveries() {
        List<Delivery> unassignedDeliveries = deliveryService.getUnassignedDeliveries();
        //Iterate over all the unassigned deliveries and assign them to a driver
        for (Delivery delivery : unassignedDeliveries) {
            //Get the nearest drivers for delivery location
            List<Driver> drivers = driverLocationService.getDriver(delivery.getLocation());
            for (Driver driver : drivers) {
                //Assign delivery to driver and send notification
                boolean isAssigned = deliveryService.assignDeliveryToDriver(delivery, driver);
                if (isAssigned) {
                    deliveryService.notifyDelivery(delivery);
                    driverLocationService.notifyDriver(driver);
                    break;
                }
            }
        }
    }
    ...
}
```

We conducted a recent study with Persistent, a global services and solutions company delivering digital engineering and enterprise modernization services to customers, to measure the productivity benefits of the CodeWhisperer customization capability. Persistent 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.

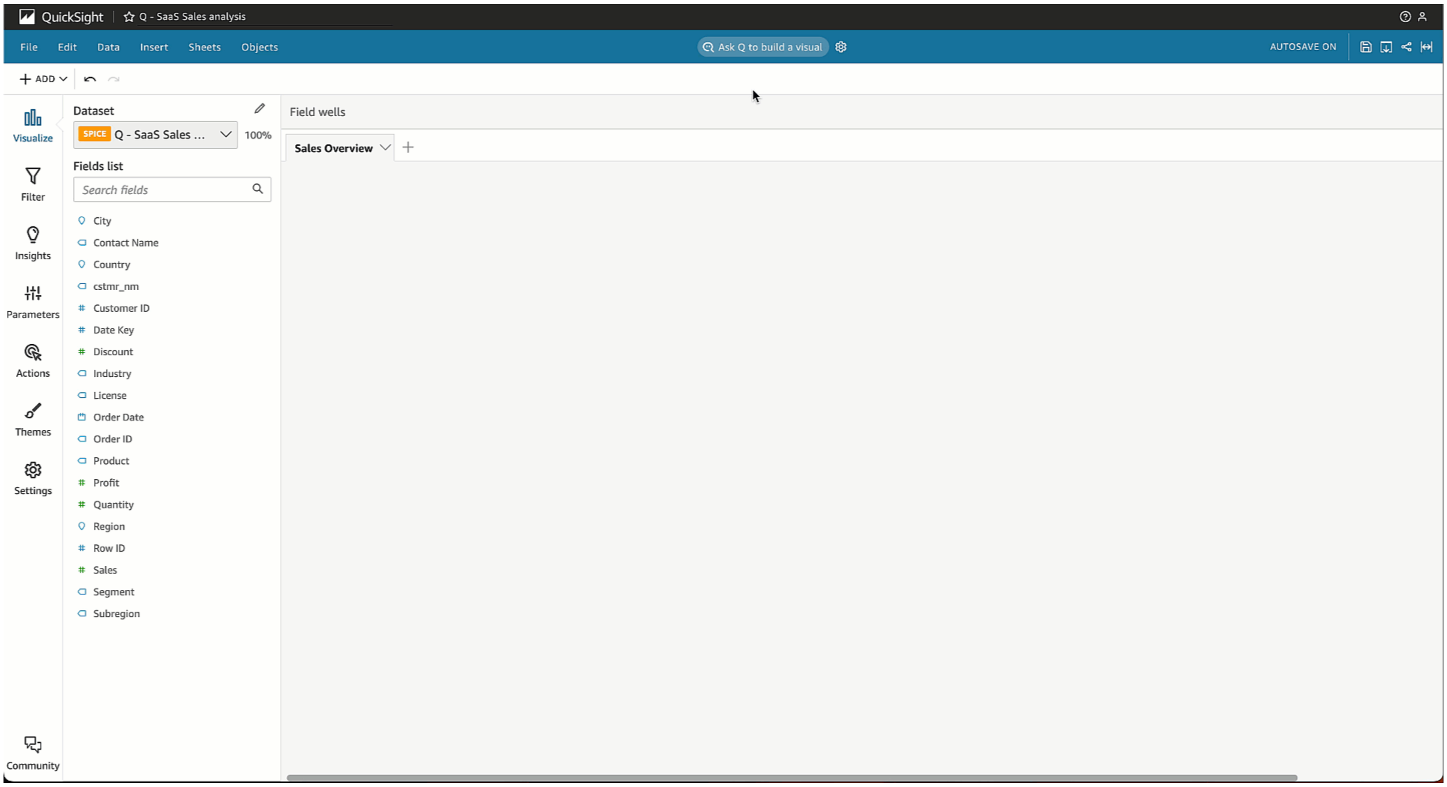
We designed this customization capability with privacy and security at the forefront. Administrators can easily manage access to a private customization from the AWS Management Console, so that only specific developers have access. Administrators can also ensure that only repositories that meet their standards are eligible for use in a CodeWhisperer customization. Using high-quality repositories helps CodeWhisperer make suggestions that promote security and code quality best practices. Each customization is completely isolated from other customers and none of the customizations built with this new capability will be used to train the FM underlying

CodeWhisperer, protecting customers' valuable intellectual property.

Announcing the preview of Generative BI authoring capabilities in Amazon QuickSight to help business analysts easily create and customize visuals using natural-language commands

AWS has been on a mission to democratize access to insights for all users in the organization. Amazon QuickSight, our unified business intelligence (BI) service built for the cloud, allows insights to be shared across all users in the organization. With QuickSight, we've been using generative models to power Amazon QuickSight Q, which enable any user to ask questions of their data using natural language, without having to write SQL queries or learn a BI tool, since 2020. In July 2023, we announced that we are furthering the early innovation in QuickSight Q with the new LLM capabilities to provide Generative BI capabilities in QuickSight. Current QuickSight customers like BMW Group and Traeger Grills are looking forward to further increasing productivity of their analysts using the Generative BI authoring experience.

Today, we are excited to make these LLM capabilities available in preview with **Generative BI dashboard authoring capabilities** for business analysts. The new Generative BI authoring capabilities extend the natural-language querying of QuickSight Q beyond answering well-structured questions (such as “what are the top 10 products sold in California?”) to help analysts quickly create customizable visuals from question fragments (such as “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. QuickSight Q also offers related questions to help analysts clarify ambiguous cases when multiple data fields match their query. When the analyst has the initial visualization, they can 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.



Creating visuals using Generative BI capabilities in Amazon QuickSight

Generative AI tools and capabilities for every business

Today's announcements open generative AI up to any customer. With enterprise-grade security and privacy, choice of leading FMs, a data-first approach, and a highly performant, cost-effective infrastructure, organizations trust AWS to power their innovations with generative AI solutions at every layer of the stack. We have seen exciting innovation from Bridgewater Associates to Omnicom to Asurion to Rocket Mortgage, and with these new announcements, we look forward to new use cases and applications of the technology to boost productivity. This is just the beginning—across the technology stack, we are innovating with new services and capabilities built for your organization to help tackle some of your largest challenges and change how we work.

Resources

To learn more, check out the following resources:

- Explore generative AI on AWS
- Learn about Amazon Bedrock, the easiest way to build and scale generative AI applications with FMs
- Learn more about Llama2 on Amazon Bedrock
- Learn about Amazon Titan, high-performing FMs from Amazon to innovate responsibly
- Learn how you can use the Amazon CodeWhisperer customization capability
- Learn more about Generative BI features for QuickSight

- Discover generative AI solutions from AWS Partners in AWS Marketplace

About the author

Swami Sivasubramanian is Vice President of Data and Machine Learning at AWS. In this role, Swami oversees all AWS Database, Analytics, and AI & Machine Learning services. His team’s mission is to help organizations put their data to work with a complete, end-to-end data solution to store, access, analyze, and visualize, and predict.

Comments