



aws SUMMIT

INDIA | MAY 25, 2023

AIML001

Generative AI on AWS - Build and scale generative AI applications with foundation models

Praveen Jayakumar

Head of ML Solutions Architecture

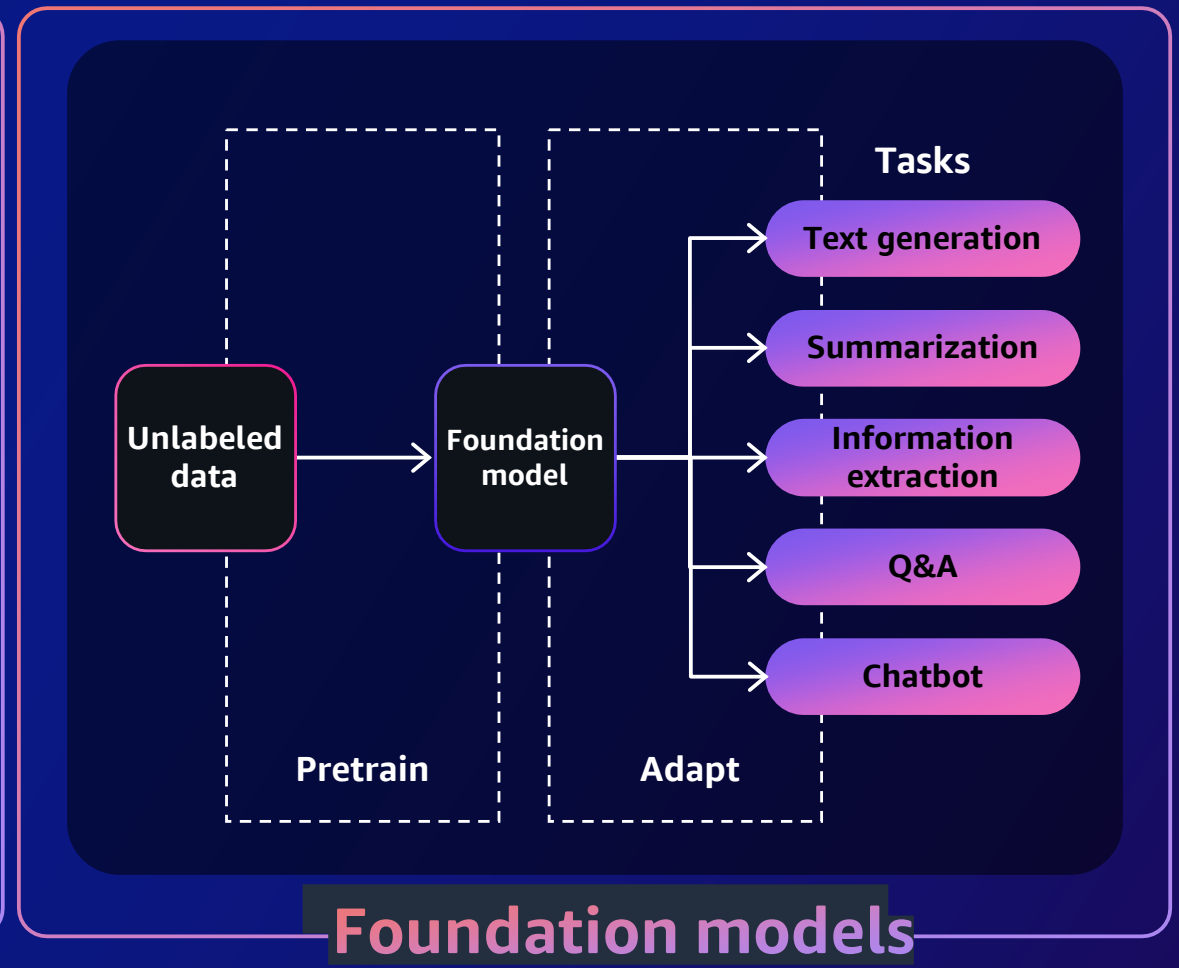
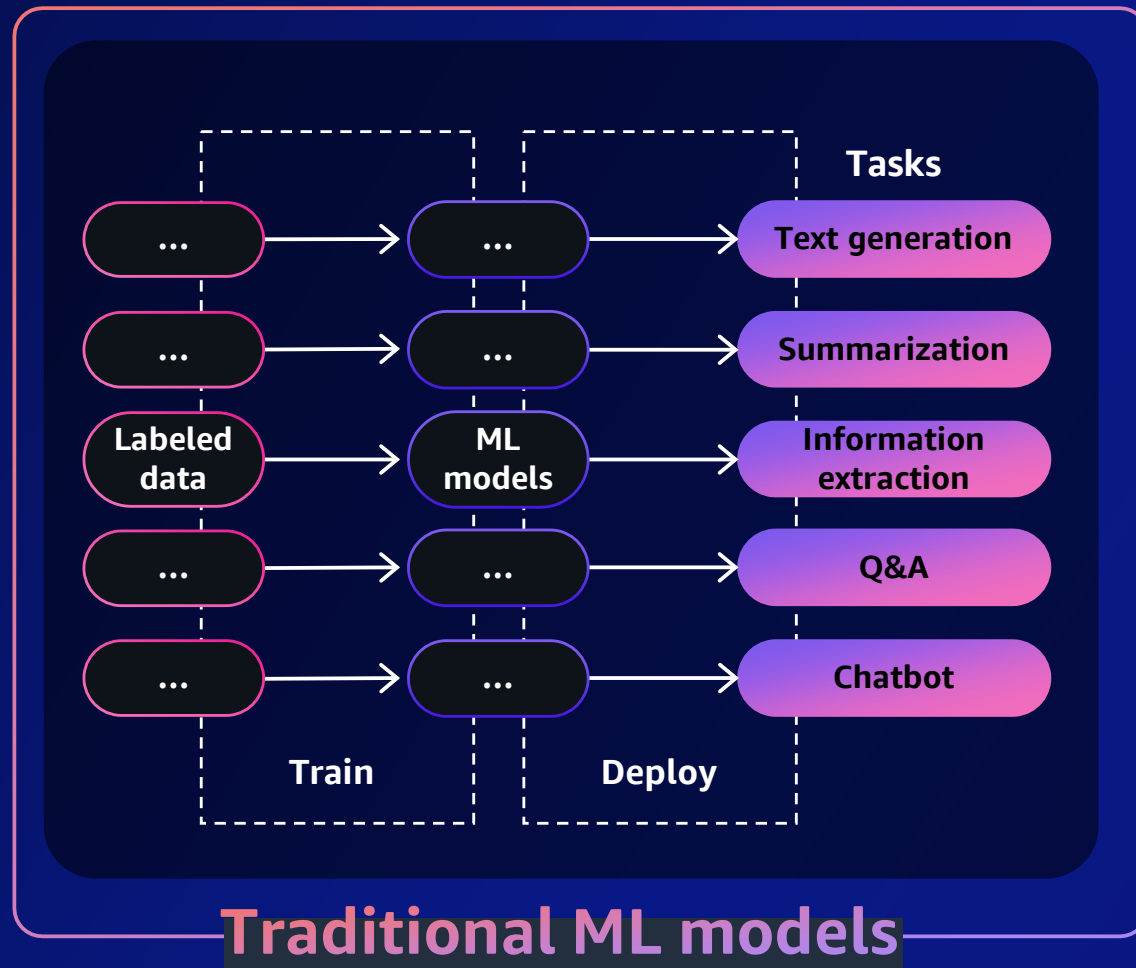
AWS India



Question: What is generative artificial intelligence (AI)?

- Creates new content and ideas, including conversations, stories, images, videos, and music
- Powered by large models that are pretrained on vast corpuses of data and commonly referred to as foundation models (FMs)

Why foundation models?



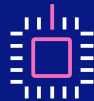
Why AWS for generative AI?



Flexibility



Secure
customization



The most cost-
effective
infrastructure



The easiest way
to build with FMs



Generative AI-
powered solutions

Why AWS for generative AI?



Flexibility



Secure customization



The most cost-effective infrastructure



The easiest way to build with FMs



Generative AI-powered solutions



Choose from a **wide selection of FMs built by AI21 Labs, Anthropic, Stability AI, and Amazon** to find the right model for your use case.

Why AWS for generative AI?



Flexibility



Secure customization



The most cost-effective infrastructure



The easiest way to build with FMs



Generative AI-powered solutions



Customize FMs for your business with just a few labeled examples. Since all data is encrypted and does not leave your Amazon Virtual Private Cloud (VPC), you can trust that your data will remain private and confidential.

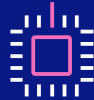
Why AWS for generative AI?



Flexibility



Secure
customization



The most cost-
effective
infrastructure



The easiest way
to build with FMs



Generative AI-
powered solutions



Get the **best price performance** for generative AI with infrastructure powered by **AWS-designed ML chips and NVIDIA GPUs**. Cost-effectively scale infrastructure to train and run FMs containing hundreds of billions of parameters.

Why AWS for generative AI?



Flexibility



Secure
customization



The most cost-
effective
infrastructure



The easiest way
to build with FMs



Generative AI-
powered solutions



Quickly integrate and deploy FMs into your applications and workloads running on AWS using familiar controls and integrations with the depth and breadth of AWS capabilities and services like Amazon SageMaker and Amazon S3.

Why AWS for generative AI?



Flexibility



Secure customization



The most cost-effective infrastructure



The easiest way to build with FMs

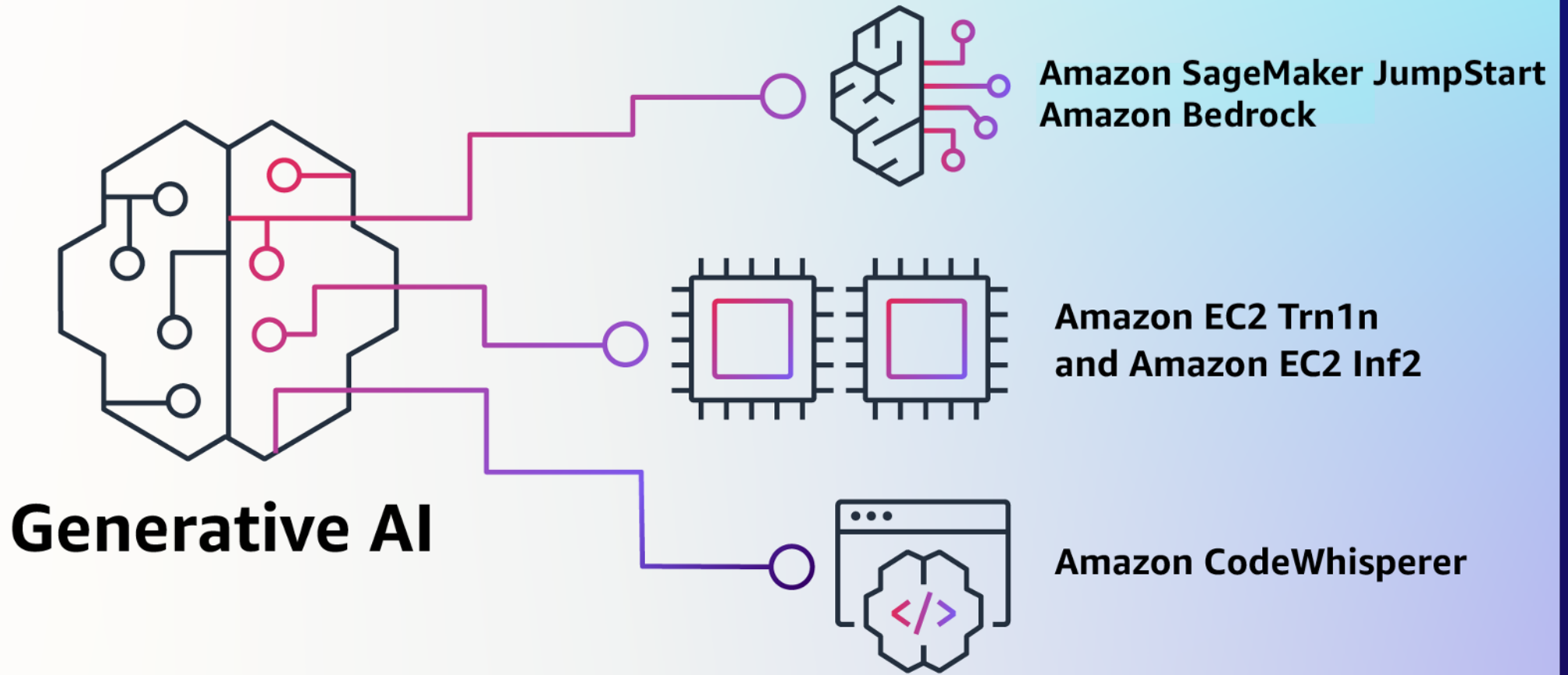


Generative AI-powered solutions



With generative AI built in, services such as **Amazon CodeWhisperer**, an AI coding companion, can help you improve productivity. In addition, you can **deploy common generative AI use cases like call summarization and question answering** using AWS sample solutions that combine AWS AI services with leading FMs.

Building with generative AI on AWS



Foundation models available through Amazon SageMaker JumpStart

Products / Machine Learning / Amazon SageMaker JumpStart

Getting started with Amazon SageMaker JumpStart

Amazon SageMaker JumpStart is a machine learning (ML) hub that can help you accelerate your ML journey. Explore how you can get started with built-in algorithms with pretrained models from model hubs, pretrained foundation models, and prebuilt solutions to solve common use cases. To get started, see documentation or example notebooks that you can quickly execute.

Reset Filters

Q foundation models X

Product Type

Sort By

Popularity

Text Tasks

- ☐ End-to-end Solution
- ☐ Text Classification
- ☐ Text Embedding
- ☐ Text Generation
- ☐ Text Summarization
- ☐ Named Entity Recognition
- ☐ Question Answering

FOUNDATION MODEL PREVIEW

Text Generation

Proprietary Models

Various Providers

Models from AI21 Labs, Cohere, and LightOn in preview. Sign-up for preview with JumpStart in us-east-1 or eu-west-1 SageMaker Console.

FOUNDATION MODEL FEATURED

Text to Image

stability.ai

Stable Diffusion 2

Stabilityai

Model ID: model-txt2img-stabilityai-stable-diffusion-v2. This is a text-to-image model from Stability AI and downloaded from HuggingFace. It takes a textual description as

FOUNDATION MODEL FEATURED

Text Generation

alexa

AlexaTM (20b)

Pytorch

Model ID: pytorch-textgeneration1-alexa20b. AlexaTM 20B is a multitask, multilingual, large-scale sequence-to-sequence (seq2seq) model, trained on a mixture of Common Crawl

FOUNDATION MODEL FEATURED

Text Generation

Bloom 1b7

Huggingface

Model ID: huggingface-textgeneration-bloom-1b7. This is a Text Generation model built upon a Transformer model from Hugging Face. It takes a text string as input and predicts next words in the sequence. This model has BigScience Responsible AI License v1.0. Please read the [terms] (<https://huggingface.co/spaces/b>



LightOn

AI for all, everywhere

AI21 labs



co:here



Proprietary Models in Gated Preview only



Demo



Amazon Bedrock

The easiest way to build and scale
generative AI applications with
foundation models (FMs)

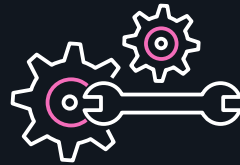
Amazon Bedrock key benefits



Accelerate development of generative AI applications using FMs through an API, without managing infrastructure



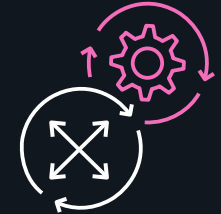
Choose FMs from AI21 Labs, Anthropic, Stability AI, and Amazon to find the right FM for your use case



Privately customize FMs using your organization's data



Enhance your data protection using comprehensive AWS security capabilities



Use AWS tools and capabilities that you are familiar with to deploy scalable, reliable, and secure generative AI applications

Bedrock supports a wide range of foundation models

FMs from Amazon



Titan Text



Titan
Embeddings

FMs from AI21 Labs, Anthropic, and Stability AI



Jurassic-2



Claude



Stable
Diffusion

Amazon Titan

INNOVATE RESPONSIBLY WITH HIGH-PERFORMING FOUNDATION MODELS (FMs) FROM AMAZON



Titan Text
focused on
NLP tasks



Titan Embeddings
for enterprise tasks
such as search and
personalization

Benefits

- Built with 20+ years of Amazon ML experience
- Automate language tasks such as summarization and text generation with Amazon Titan Text FM
- Enhance search accuracy and improve personalized recommendations with Amazon Titan Embeddings FM
- Support responsible use of AI by reducing inappropriate or harmful content

Foundation models from top AI startups

The logo for AI21 Labs, featuring the text "AI21" in black and "labs" in pink.

Jurassic-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch

The logo for Anthropic, featuring the word "ANTHROPIC" in black, all-caps, sans-serif font.

Claude

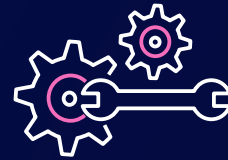
LLM for conversational and text processing tasks

The logo for Stability.ai, featuring the text "stability.ai" in black, lowercase, sans-serif font.

Stable Diffusion

Generation of unique, realistic, high-quality images, art, logos, and designs

Privately customize foundation models using your organization's data



Fine-tune

PURPOSE

Maximizing accuracy for specific tasks

DATA NEED

Small number of labeled examples

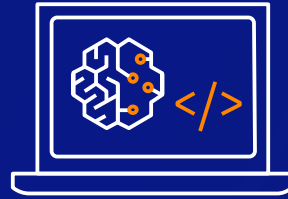
Amazon CodeWhisperer

Build applications faster and more securely
with an AI coding companion



CodeWhisperer

BUILD APPLICATIONS FASTER AND MORE SECURELY WITH YOUR AI CODING COMPANION



Code generation



AND

Go, Rust, PHP, Ruby, Kotlin, C, C++,
Shell scripting, SQL, and Scala



AND

CLion, GoLand, WebStorm, Rider,
PhpStorm, RubyMine, and DataGrip

Code generation

- Get multiple code suggestions in seconds based on natural language (English) description of coding task and surrounding code
- Code generation can range from single-line suggestions to full-function blocks
- Provides high-quality suggestions for popular AWS services
- Generated code matches developer style and patterns

```
J SQSIdentityFunction.java M src/main/java/com/amazonaws/services/sqs/SQSIdentityFunction.java
package com.amazonaws.services.sqs;
import java.util.Map.Entry;
import java.util.stream.Collectors;
import com.amazonaws.services.lambda.runtime.Context;
import com.amazonaws.services.lambda.runtime.RequestHandler;
import com.amazonaws.services.lambda.runtime.events.SQSEvent;
import com.amazonaws.services.lambda.runtime.events.SQSEvent.MessageAttribute;
import com.amazonaws.services.lambda.runtime.events.SQSEvent.SQSMessage;
import com.amazonaws.services.sqs.model.MessageAttributeValue;
import com.amazonaws.services.dynamodbv2.AmazonDynamoDBClientBuilder;
import com.amazonaws.regions.Regions;

/*Create a lambda function that stores the body of the SQS message
into a hash key of a DynamoDB table. */
public class SQSIdentityFunction implements RequestHandler<SQSEvent, String> {

    private static final String TABLE_NAME = "SQSMessage";
    private static final String HASH_KEY = "MessageId";

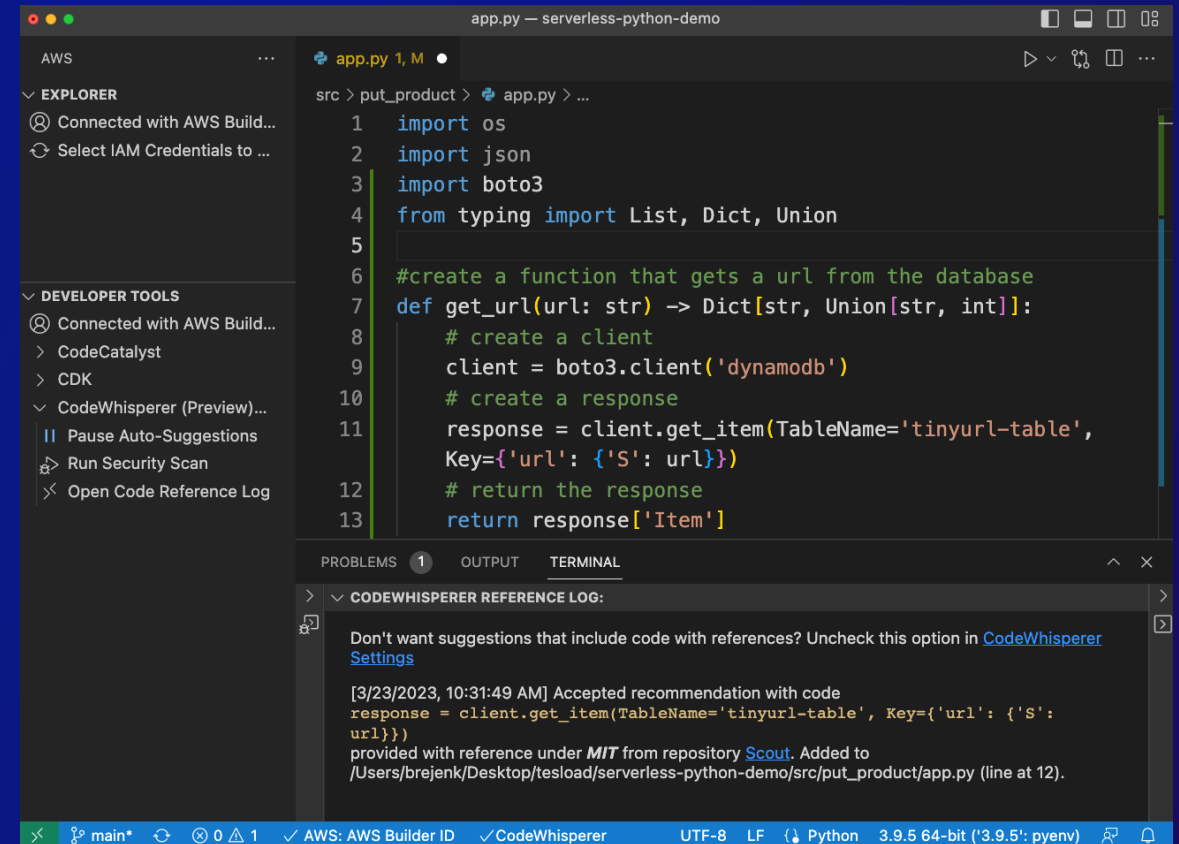
    private AmazonDynamoDBClientBuilder builder = AmazonDynamoDBClientBuilder.standard();
    private AmazonDynamoDB client = builder.withRegion(Regions.US_EAST_1).build();

    @Override
    public String handleRequest(SQSEvent event, Context context) {
        for (SQSMessage message : event.getRecords()) {
            String messageId = message.getMessageId();
            String messageBody = message.getBody();
            context.getLogger().log("MessageId: " + messageId + " MessageBody: " + messageBody);
            storeMessage(messageId, messageBody);
        }
        return "Success";
    }

    private void storeMessage(String messageId, String messageBody) {
        client.putItem(TABLE_NAME, HASH_KEY, messageId, messageBody);
    }
}
```

Reference tracking

- Trained on billions of lines of code
- Flags code similar to open-source training data
- Tracks accepted suggestions so that you can provide appropriate attribution
- Enterprise controls to more easily deactivate/filter code suggestions similar to open-source training data



The screenshot shows a code editor window titled 'app.py — serverless-python-demo'. The editor displays a Python function `get_url` that interacts with a DynamoDB database. A code suggestion is visible on line 12, showing a `response` variable assignment. Below the editor, the 'PROBLEMS' panel is open, showing a 'CODEWHISPERER REFERENCE LOG'. The log entry states: 'Don't want suggestions that include code with references? Uncheck this option in [CodeWhisperer Settings](#)'. It also records an accepted recommendation with code, the date and time, and the repository source (Scout) under the MIT license, including the file path and line number.

```
1 import os
2 import json
3 import boto3
4 from typing import List, Dict, Union
5
6 #create a function that gets a url from the database
7 def get_url(url: str) -> Dict[str, Union[str, int]]:
8     # create a client
9     client = boto3.client('dynamodb')
10    # create a response
11    response = client.get_item(TableName='tinyurl-table',
12                               Key={'url': {'S': url}})
13    # return the response
14    return response['Item']
```

PROBLEMS 1 OUTPUT TERMINAL

CODEWHISPERER REFERENCE LOG:

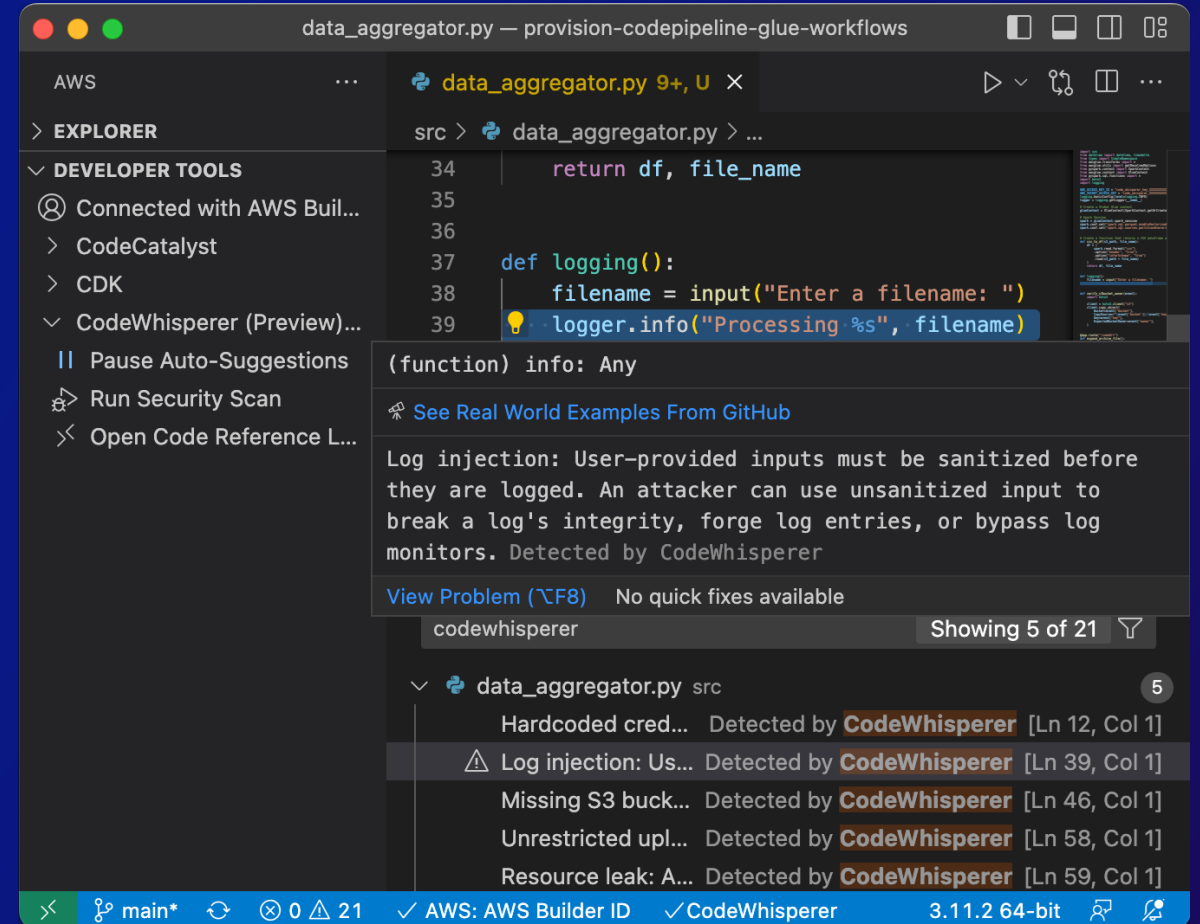
Don't want suggestions that include code with references? Uncheck this option in [CodeWhisperer Settings](#)

[3/23/2023, 10:31:49 AM] Accepted recommendation with code
`response = client.get_item(TableName='tinyurl-table', Key={'url': {'S': url}})`
provided with reference under **MIT** from repository [Scout](#). Added to
/Users/brejenk/Desktop/tesload/serverless-python-demo/src/put_product/app.py (line at 12).

main* 0 1 AWS: AWS Builder ID CodeWhisperer UTF-8 LF Python 3.9.5 64-bit ('3.9.5': pyenv)

Security scanning

- Scan generated and developer-written code to detect security vulnerabilities
- Receive vulnerability remediation suggestions
- Scan for hard-to-find security vulnerabilities
- Supports Python, Java, and JavaScript



Demo



Summary

- Amazon SageMaker Jumpstart – Foundation Model
- Amazon Bedrock
- Amazon CodeWhisperer

skillbuilder.aws 

Your time is now

Build in-demand cloud skills *your way*



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Praveen Jayakumar
Head of ML Solutions Architecture
AWS India



Please complete the
session survey