Deploying Machine Learning Models Using AWS SageMaker

Jorge Vásquez
SOFTWARE ENGINEER

@jorvasquez2301

Overview

Deploying and testing models in AWS SageMaker Hosting Services

- Built-in Image Classification algorithm
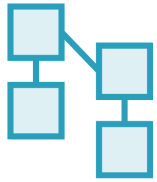- Custom Tensorflow Algorithm
- Custom MXNet Algorithm

Integrating with AWS API Gateway and AWS Lambda

Deploying and Testing Machine Learning Models in AWS SageMaker Hosting Services

Deploying to AWS SageMaker

Create a **Model**

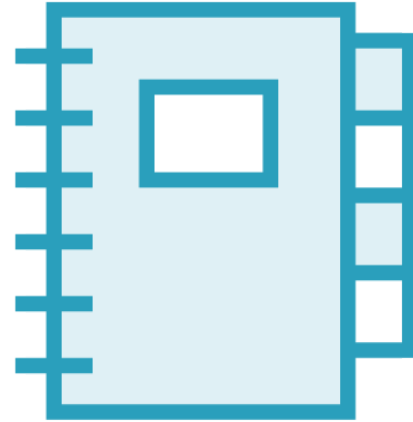Create an **Endpoint Configuration for an HTTPS Endpoint**

Create an **HTTPS Endpoint**

If you specify that you want two or more instances for deploying a model, AWS SageMaker launches them in multiple Availability Zones

Testing Deployed Models in AWS SageMaker

You can send sample requests and get inferences directly from Jupyter

Demo

Deploying and testing the trained model based on the built-in Image Classification algorithm, using the low-level AWS SDK for Python

Demo

Deploying and testing the trained model based on the built-in Image Classification algorithm, using the high-level SageMaker Python Library

Demo

Deploying and testing the trained model based on a custom Tensorflow algorithm, using the high-level SageMaker Python Library

Demo

Deploying and testing the trained model based on a custom MXNet algorithm, using the high-level SageMaker Python Library

Integrating Endpoints with AWS API Gateway and AWS Lambda

AWS API Gateway

Create

Publish

Maintain

Monitor

Secure

AWS API Gateway

With a few clicks, you can create REST APIs that act as a "front door" to external applications

AWS Lambda

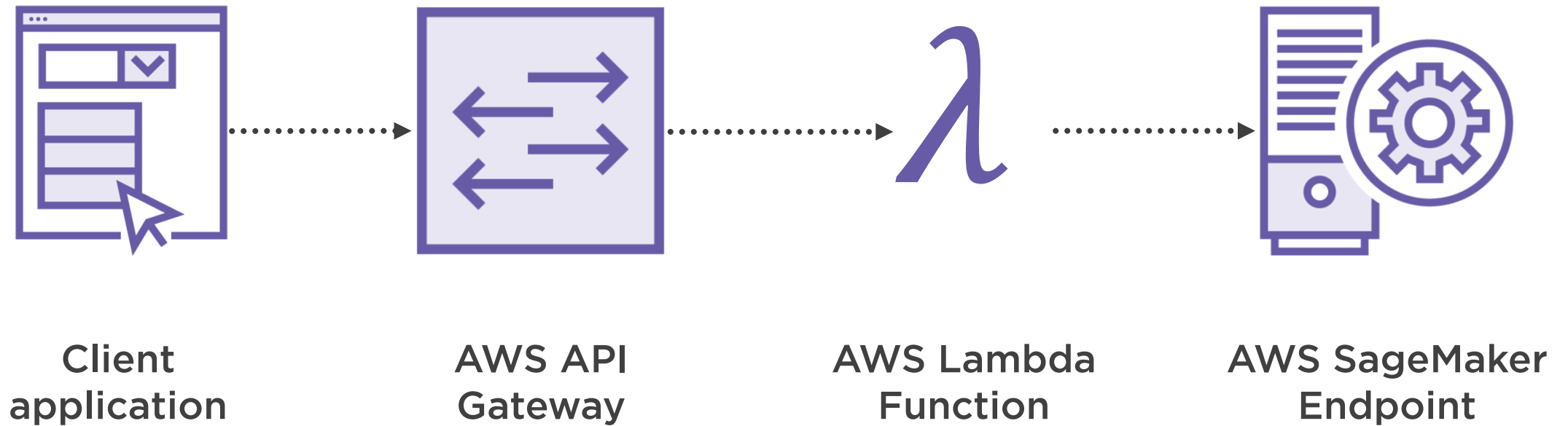**Run code without provisioning or managing servers**

Pay only for the compute time you consume

Automatically triggered from other AWS services

Integrating AWS SageMaker with AWS API Gateway and AWS Lambda

Client application

AWS API Gateway

AWS Lambda Function

AWS SageMaker Endpoint

Demo

Integrating an AWS SageMaker Endpoint with AWS API Gateway and AWS Lambda:

- Creating a Lambda function

- Creating an API Gateway

- Testing the API Gateway with Postman

Summary

Creating endpoints for built-in or custom algorithms is a straightforward process

Integrating AWS SageMaker with AWS API Gateway and AWS Lambda makes it easier to create scalable REST APIs