Managing Security and Scalability in AWS SageMaker

Jorge Vásquez
SOFTWARE ENGINEER

@jorvasquez2301

Overview

Managing authentication and access control using IAM policies

Monitoring and troubleshooting endpoints using AWS CloudWatch

Configuring automatic scaling for endpoints

Managing Authentication and Access Control Using IAM Policies

AWS Authentication and Access Control Model

Every AWS resource is owned by an AWS account, and permissions to create or access a resource are governed by permissions policies.

AWS Authentication and Access Control Model

An account administrator can attach permissions policies to IAM identities

- Users
- Groups
- Roles

Permissions Policy Components

Who is getting the permissions (Principal)

For which resources

Specific actions you want to allow/deny

AWS SageMaker Resources

Notebook instances

Training/Tuning jobs

Models

Endpoint configurations

Endpoints

Demo

Globomantics has two teams that work with SageMaker on two different projects:

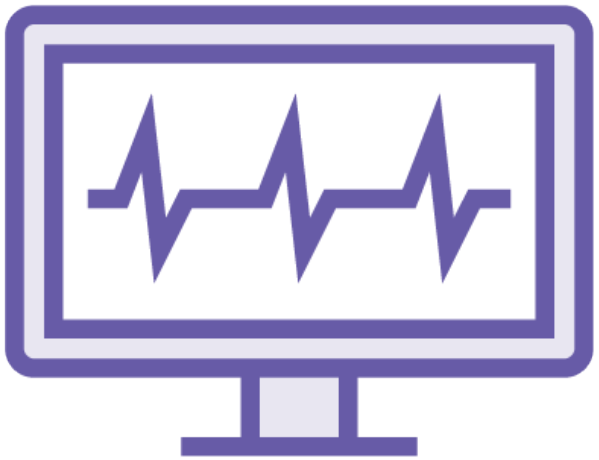
- Breast Cancer Detection
- Image Sentiment Analysis

You, as administrator, want each team not to have access to the notebook instances of the other team

Monitoring and Troubleshooting Deployed Models with AWS CloudWatch

Monitoring endpoints is an important part of maintaining the reliability, availability, and performance of AWS SageMaker

Monitoring AWS SageMaker Endpoints with CloudWatch



Once deployed, logs and metrics of SageMaker endpoints can be monitored with AWS CloudWatch

Endpoint Invocation Metrics

Invocation4XXErrors	Invocation5XXErrors	Invocations
InvocationsPerInstance	ModelLatency	OverheadLatency

Endpoint Hosting Instance Metrics

CPUUtilization

MemoryUtilization

GPUUtilization

GPUMemoryUtilization

DiskUtilization

Endpoints Logs in AWS CloudWatch

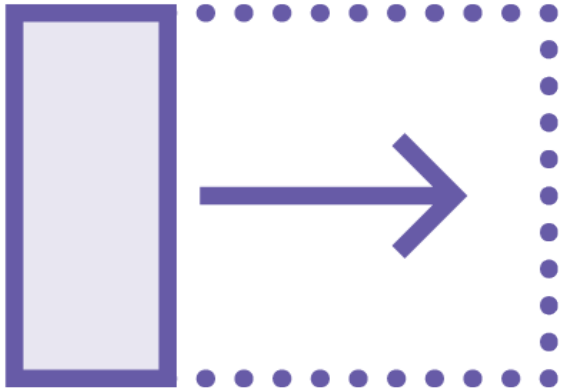
Log Group Name	Log Stream Name
/aws/sagemaker/Endpoints/[EndpointName]	[production-variant-name]/[instance-id]

Demo

Analyzing endpoint metrics and logs with AWS CloudWatch

Configuring Automatic Scaling for AWS SageMaker Endpoints

Automatic Scaling



Automatic scaling dynamically adjusts the number of instances provisioned for a production variant

Automatic Scaling

To use Automatic Scaling you define and apply a Scaling Policy

Scaling Policy Parameters

Target metric

Minimum and maximum capacity

Cool down period

Demo

Configuring automatic scaling for an AWS SageMaker endpoint using the AWS Console

Summary

- Controlling access to notebook instances
- Using AWS CloudWatch for analyzing SageMaker endpoints metrics and logs
- Configuring automatic scaling for adapting SageMaker endpoints to workloads