AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

AIM408

Achieve high performance and cost-effective model deployment

Michael Hsieh (he/him)
Sr. AI/ML Specialist Solutions Architect
AWS

Pravallika Etoori (she/her)
Machine Learning Engineer
Goldman Sachs

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Model deployment options in Amazon SageMaker

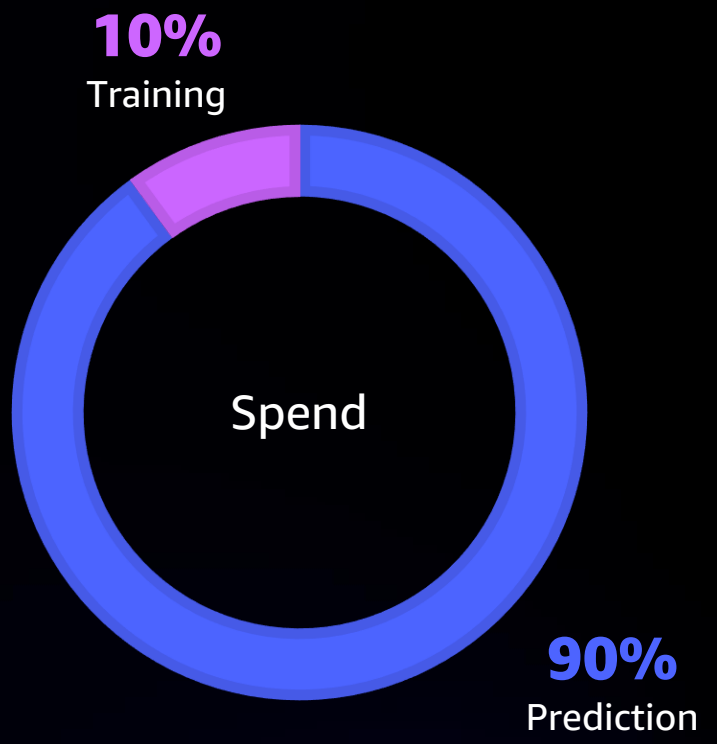
How to optimize model deployment

Goldman Sachs's SageMaker journey

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why optimize model deployment

Predictions drive complexity and cost in production



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How to choose your inference architecture

CONSIDERATIONS



SLA	Cost	Compute	Data	Models
Latency	Budget	CPU, GPU, AI accelerator	Collection frequency	Multiple models
Throughput	Monitoring	Infrastructure management	Payload size	Size, runtime, frameworks
Variable workload	Control	Efficient utilization	Data drift	Updates, AB testing

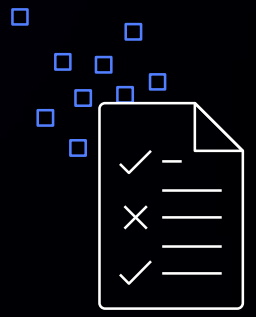
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

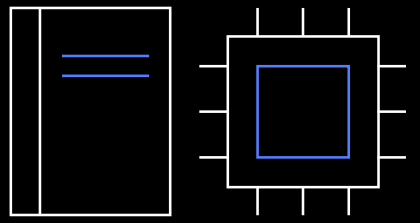
Deploying models in SageMaker



Easy deployment of ML models



Online and offline scoring



Fully managed infrastructure

```
predictor = estimator.deploy(
      initial_instance_count = 1,
      instance_type = 'ml.c5.4xlarge')

prediction = predictor.predict(x_test)
```

```
transformer = estimator.transformer(
      instance_count = 1,
      instance_type = 'ml.m5.xlarge’)

transformer.transform(
      test_data_s3,
      content_type = 'text/csv')
```

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker model deployment options

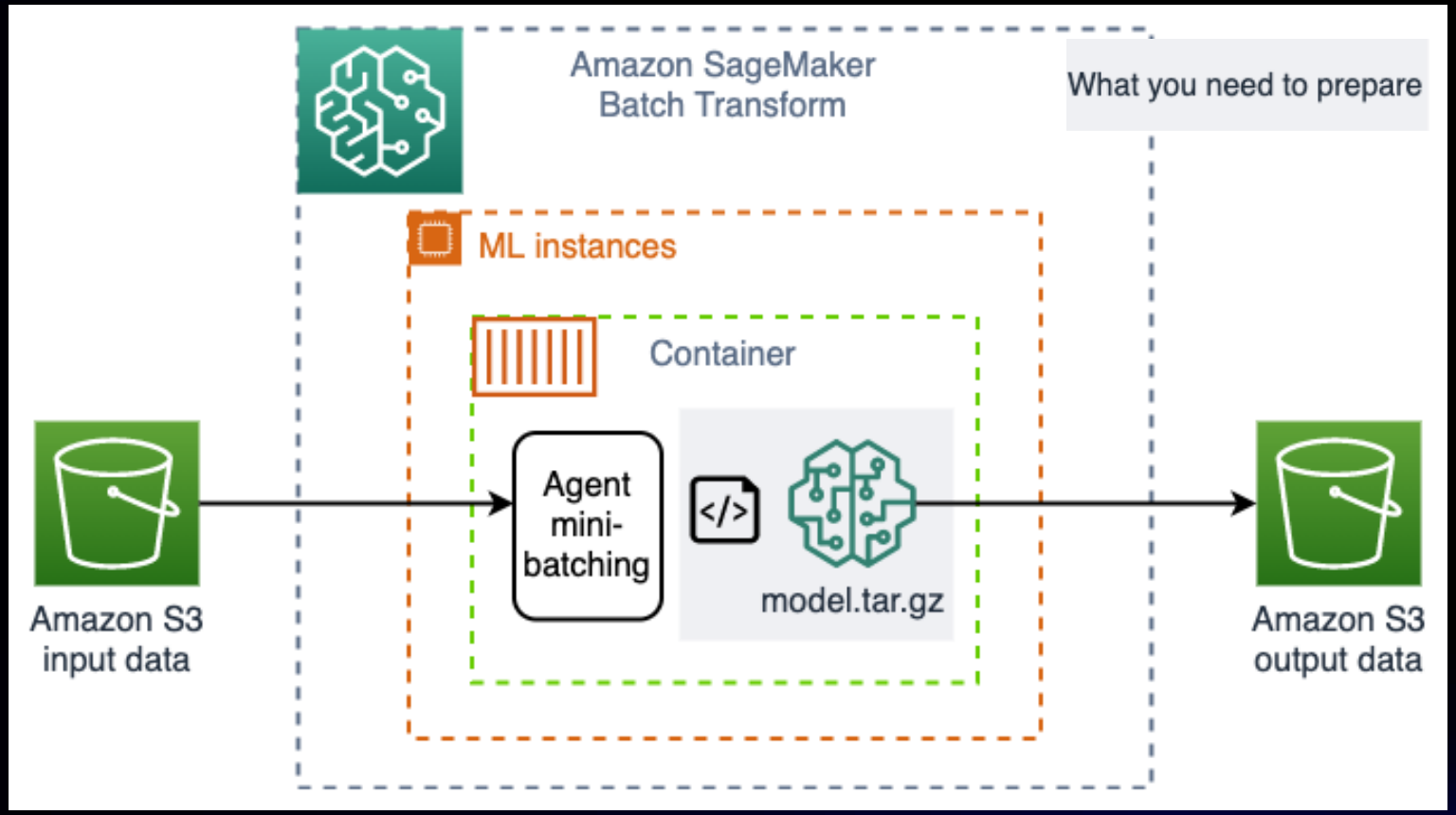
SageMaker Batch Transform



Fully managed mini-batching for large data

Pay only for what you use

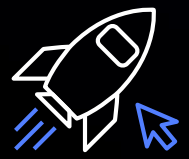
Suitable for periodic arrival of large data



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker model deployment options

SageMaker Real-time Inference

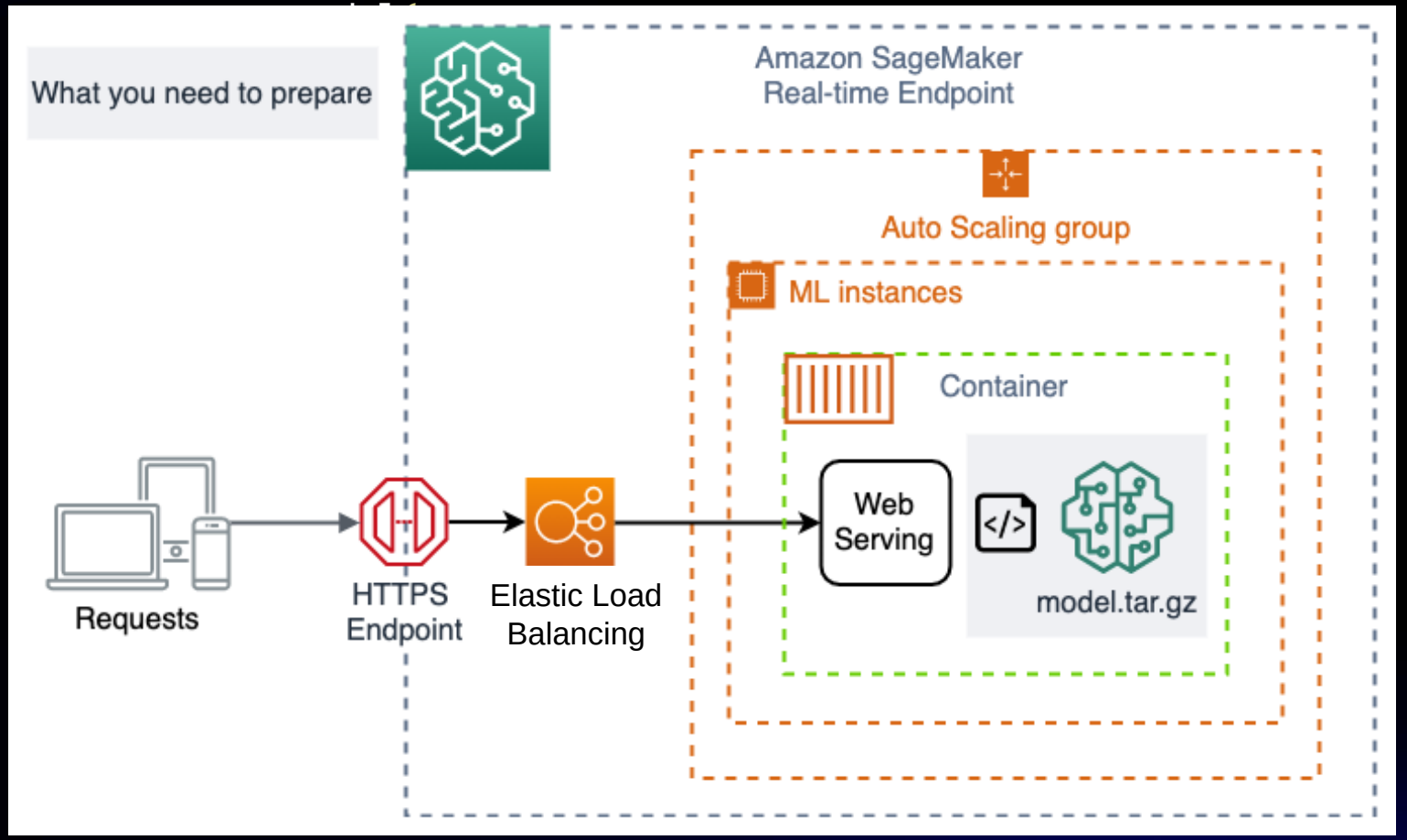


Create a long-running microservice

Instant response for payload up to 6MB

Accessible from an external application

Autoscaling

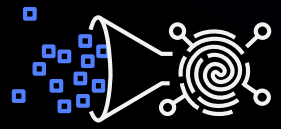


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

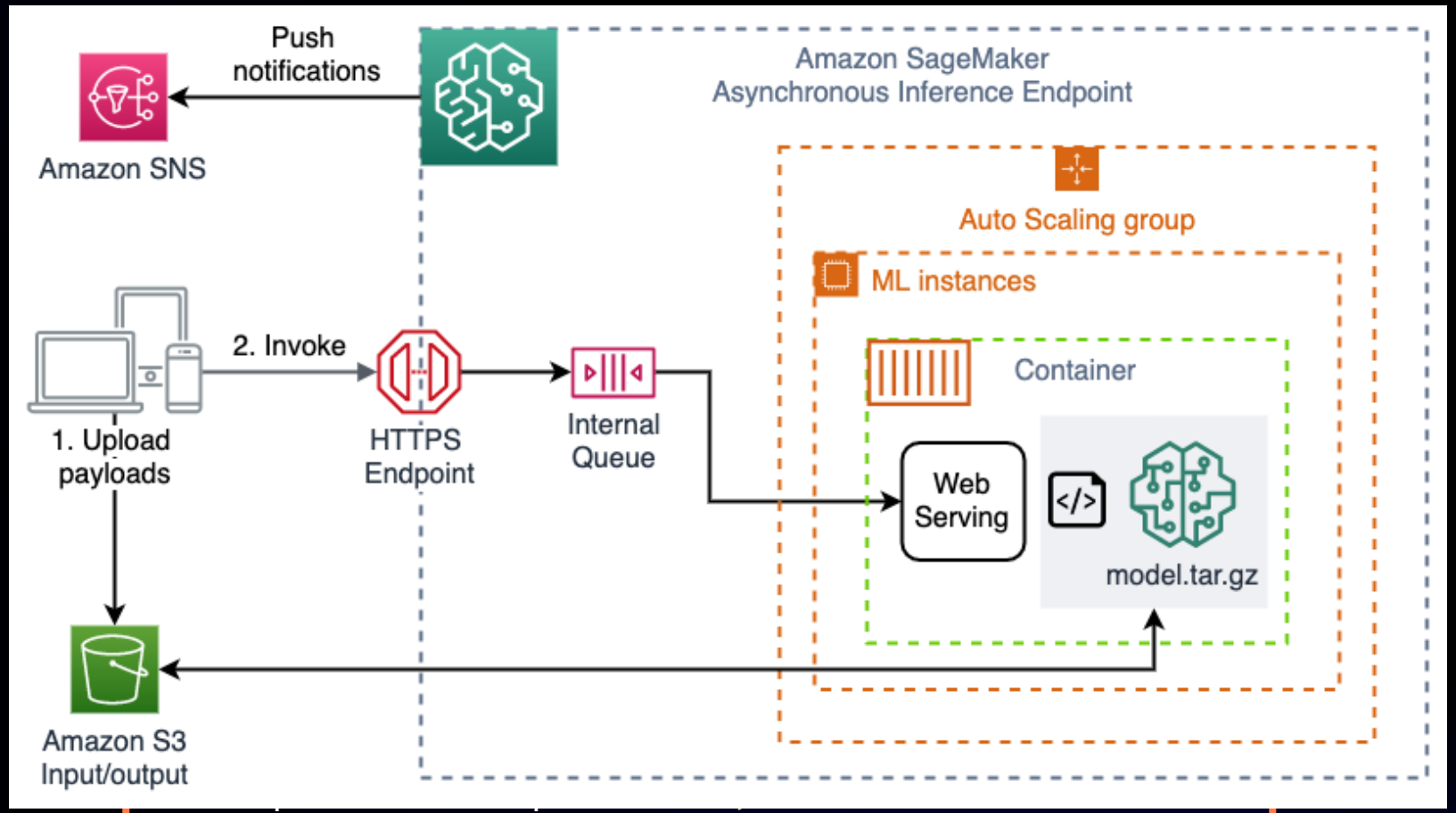
SageMaker model deployment options

NEW

SageMaker Asynchronous Inference



- Ideal for large payload up to 1GB
- Longer processing timeout up to 15 min
- Autoscaling (down to 0 instance)
- Suitable for CV/NLP use cases

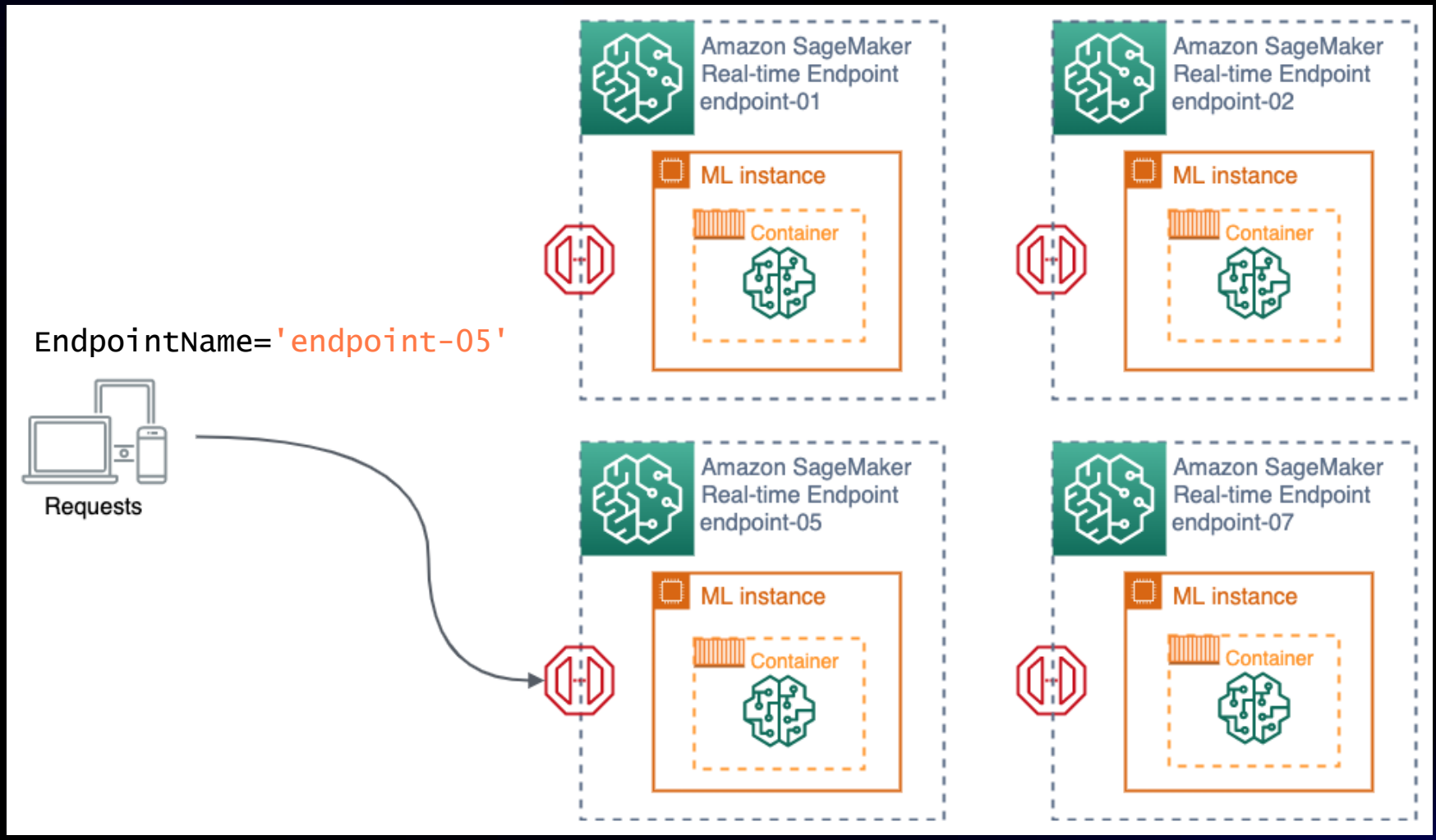



```
InputLocation='s3://my-bucket/inference/video.mpeg')
```

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost concern

HOSTING INDIVIDUAL ENDPOINTS

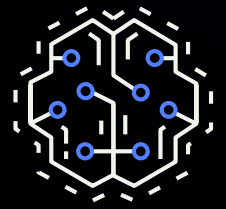


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

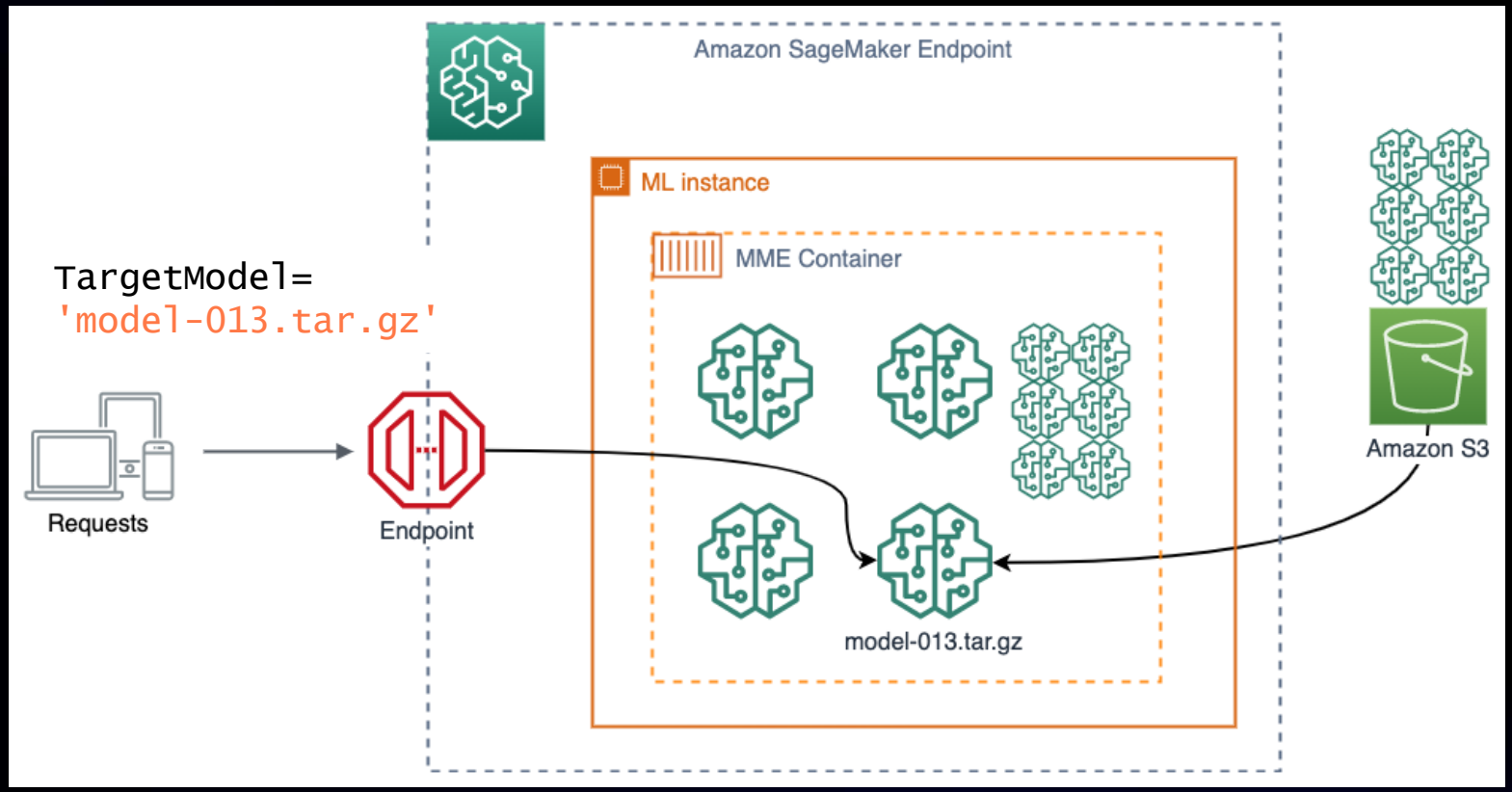
SageMaker model deployment options

COST-SAVING OPPORTUNITY

SageMaker Multi-Model Endpoint



- Host multiple models in one container
- Direct invocation to target model
- Improves resource utilization
- Dynamic loading model from Amazon S3

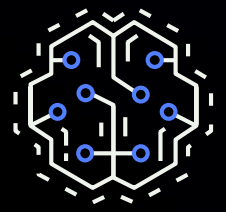


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker model deployment options

COST-SAVING OPPORTUNITY

SageMaker Multi-Model Endpoint



Host multiple models in one container

Direct invocation to target model

Improves resource utilization

Dynamic loading model from Amazon S3

```
container = {
  'Image': mme-supported-image,
  'ModelDataUrl': 's3://my-bucket/folder-of-tar-gz',
  'Mode': 'MultiModel'}

sm.create_model(
   Containers = [container], ...)

sm.create_endpoint_config();sm.create_endpoint()

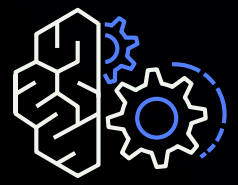
smrt.invoke_endpoint(
   EndpointName = endpoint_name,
   TargetModel = 'model-007.tar.gz',
   Body = body, ...)
```

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker model deployment options

COST-SAVING OPPORTUNITY

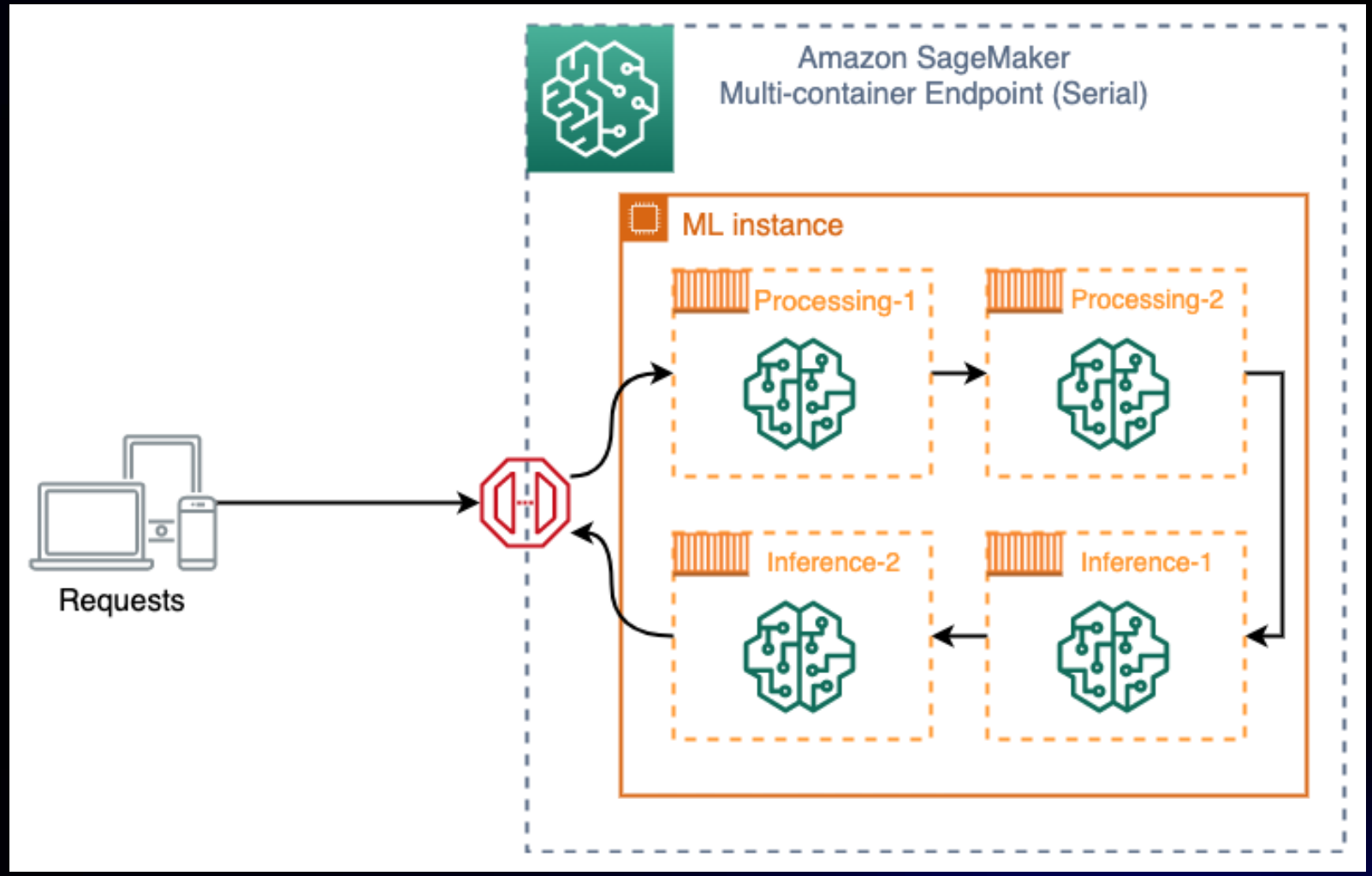
SageMaker multi-container endpoint



Host up to 15 distinct containers

Direct or serial invocation

No cold start vs. Multi-Model Endpoint

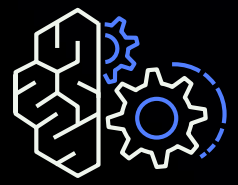


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker model deployment options

COST-SAVING OPPORTUNITY

SageMaker multi-container endpoint



Host up to 15 distinct containers

Direct or serial invocation

No cold start vs. Multi-Model Endpoint

```
container1 = {
   'Image': container,
   'ContainerHostname': 'firstContainer'}; ...

sm.create_model(
   InferenceExecutionConfig = {'Mode': 'Direct'},
   Containers = [container1, container2, ...], ...)

sm.create_endpoint_config();sm.create_endpoint()

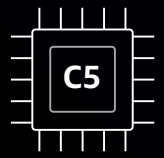
smrt.invoke_endpoint(
   EndpointName = endpoint_name,
   TargetContainerHostname = 'firstContainer',
   Body = body, ...)
```

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker ML instance options

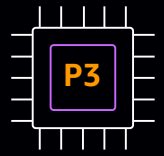
BALANCING BETWEEN COST AND PERFORMANCE

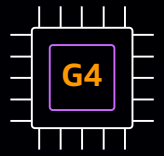
CPU INSTANCES



Low throughput, low cost, most flexible

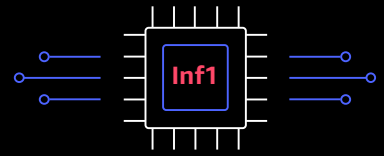
GPU INSTANCES





High throughput, and low-latency access to CUDA

CUSTOM CHIP



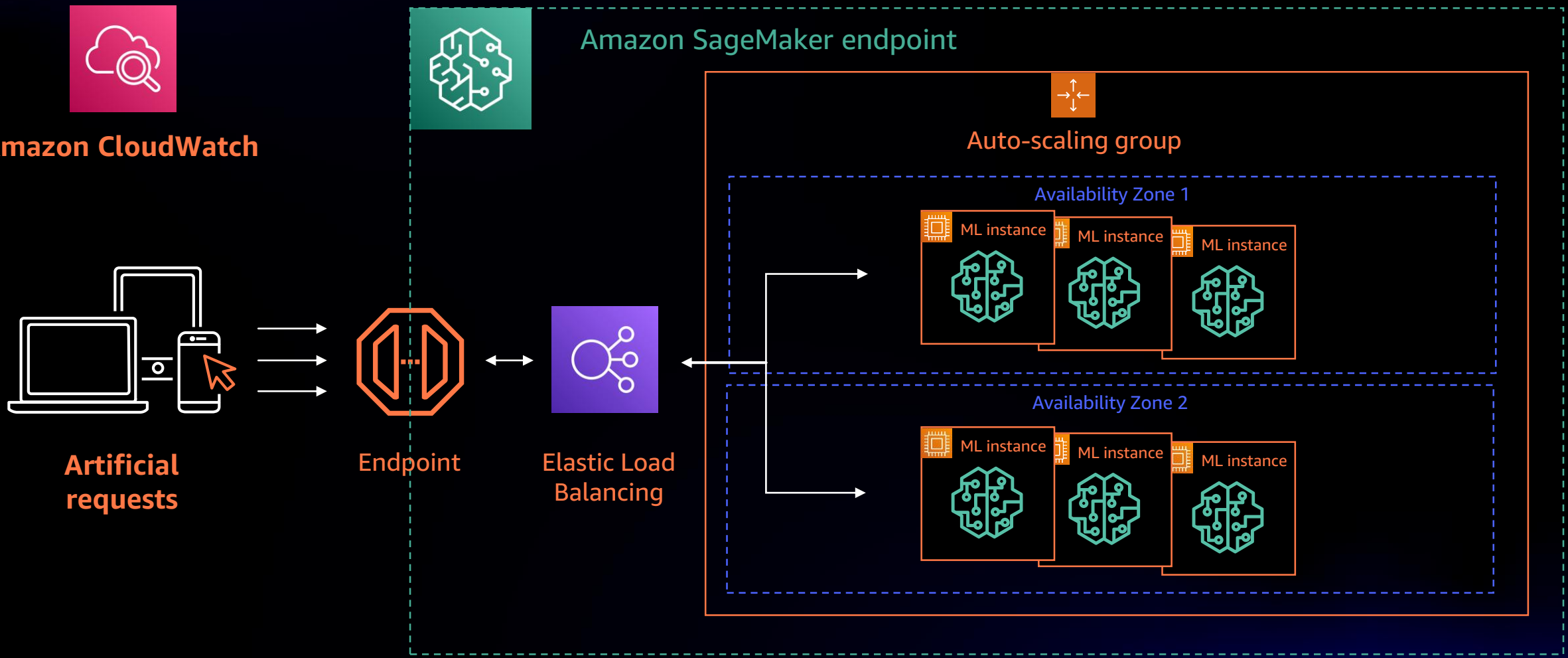
Inf1: High throughput, high performance, and lowest cost in the cloud

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Load testing

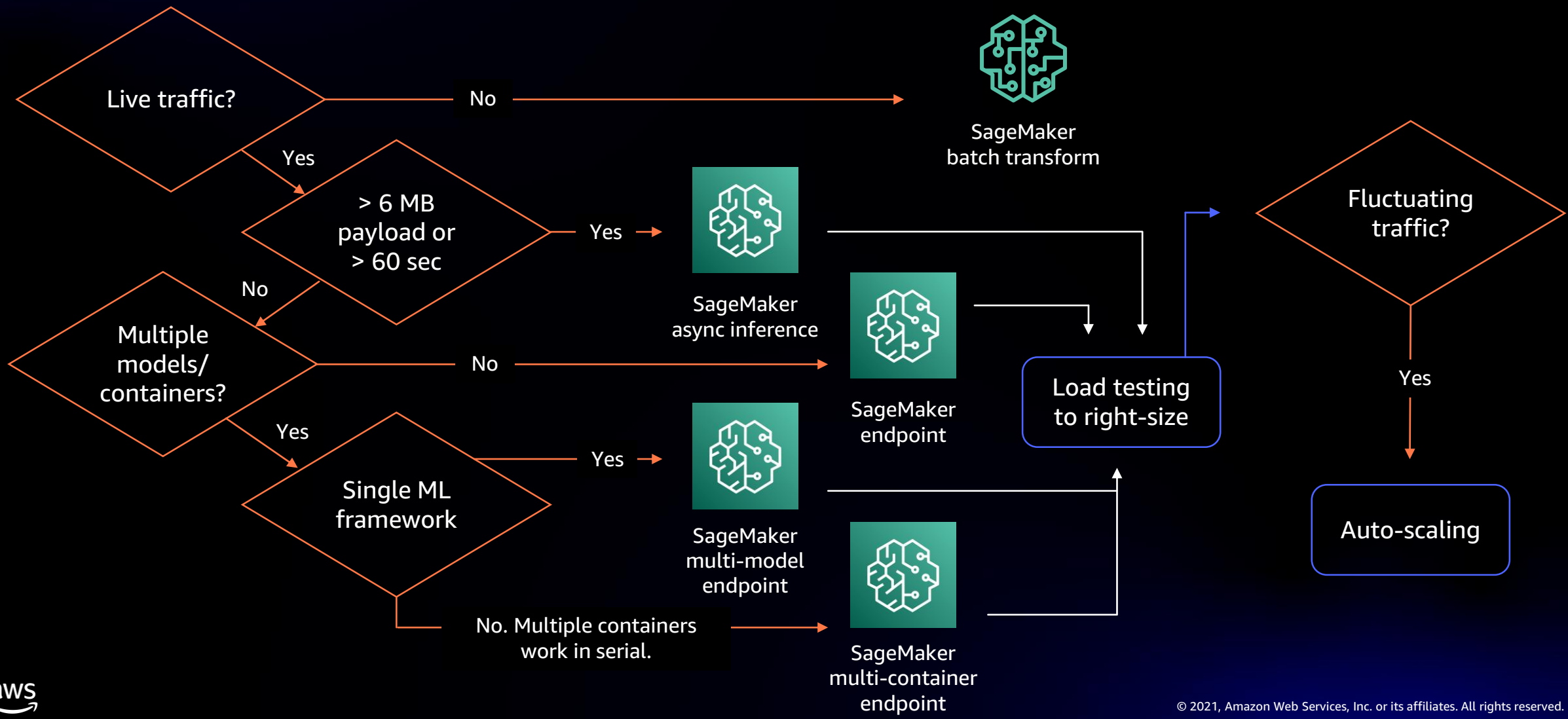
KNOW YOUR ENDPOINTS



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How to choose your inference architecture

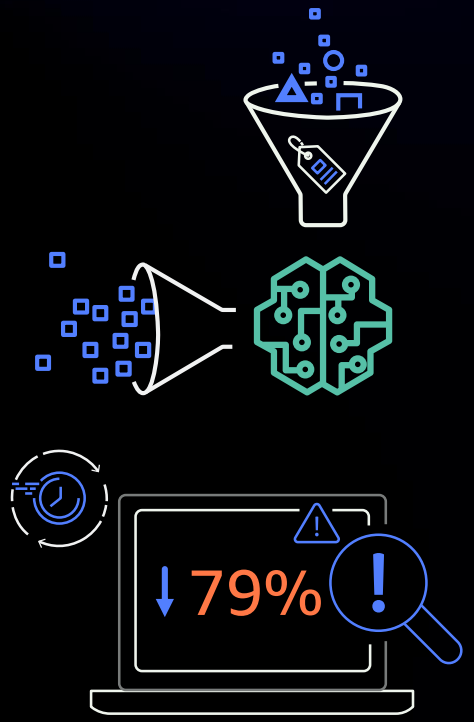
A DECISION TREE



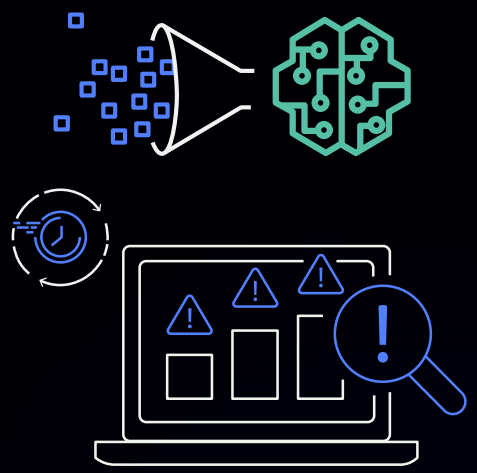
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Model Monitor

OPTIMIZING MODEL ACCURACY



Model quality drift



Data drift



SageMaker Clarify

Feature importance drift & data bias

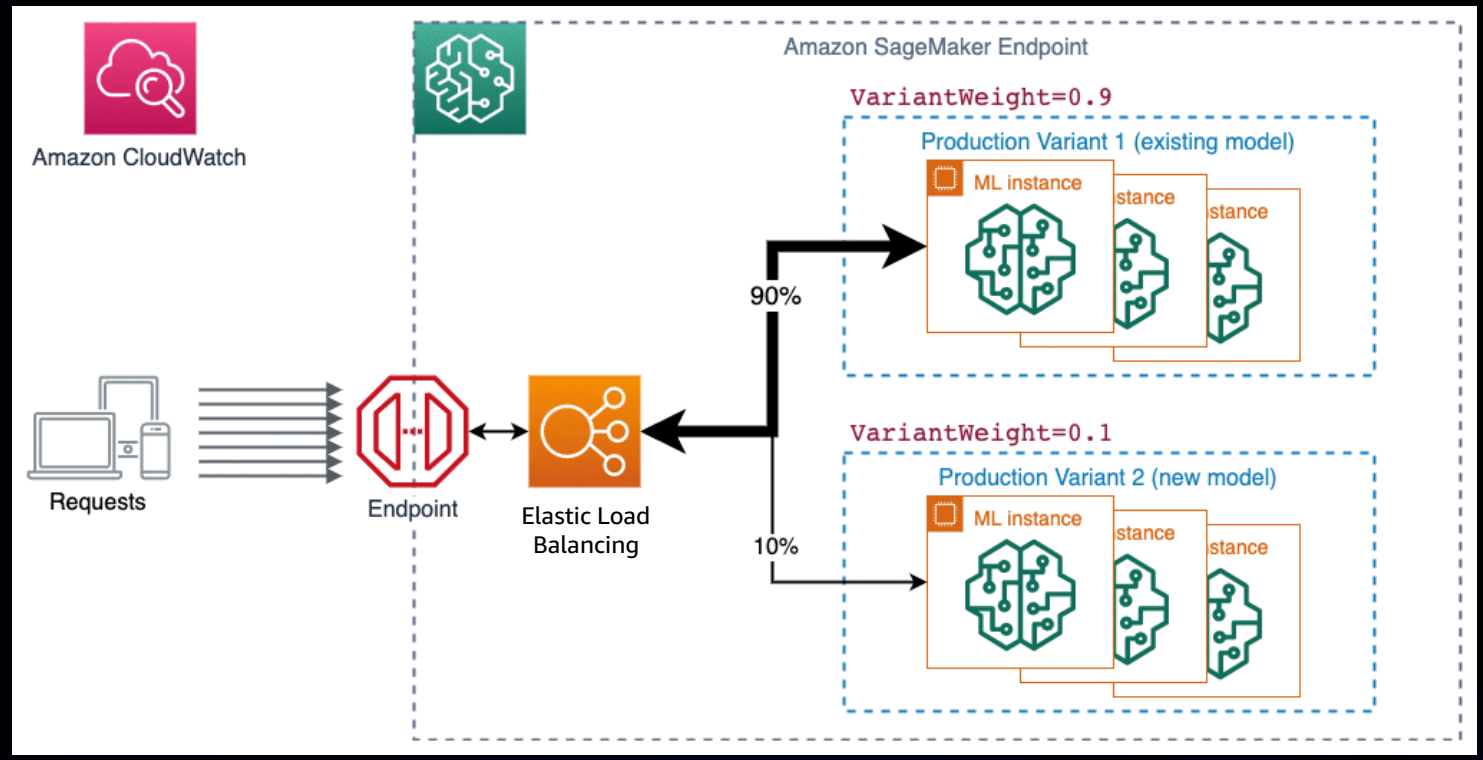
aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

A/B testing

USING PRODUCTION VARIANTS

```
sm.update_endpoint_weights_and_capacities(
  EndpointName=endpoint_name,
  DesiredWeightsAndCapacities=[
   {
     "DesiredWeight": 0.1,
     "VariantName": ”new-model”
   },
   {
     "DesiredWeight": 0.9,
     "VariantName": ”existing-model”
   }
  ]
)
```



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



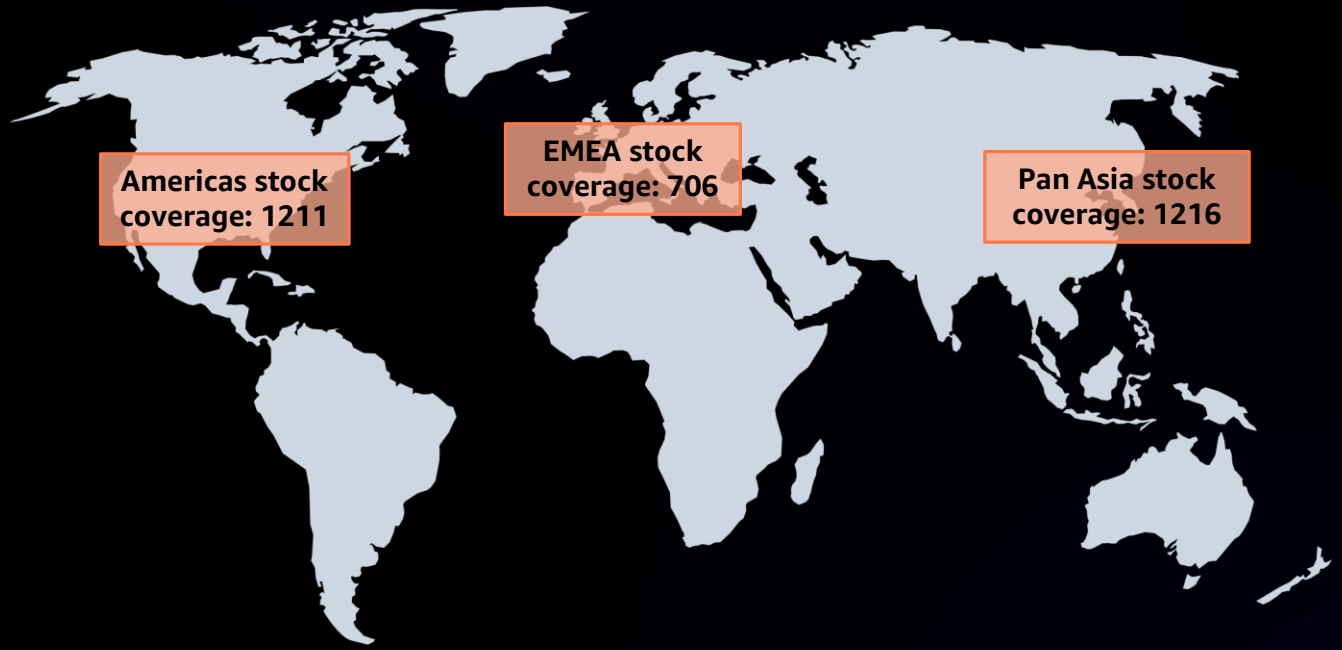
Goldman Sachs's SageMaker journey

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Global Investment Research (GIR)

Overview

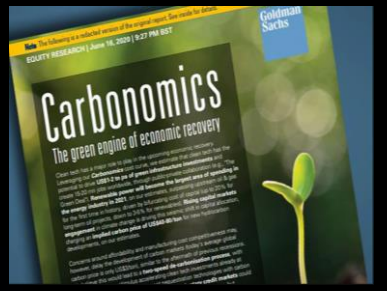
GIR provides original, fundamental insights and analysis for clients in the equity, fixed income, currency, and commodities markets



Coverage

GIR consists of **~1,000 people** in **24 offices** around the world, covering **~3,000 securities** over **45 economies** and every major market

Thematic research













© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Reaching our clients

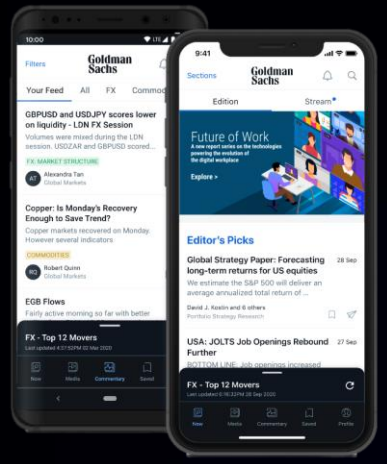
GIR publishes **~350** reports, blogs, and financial models daily

Over **1 million** clients

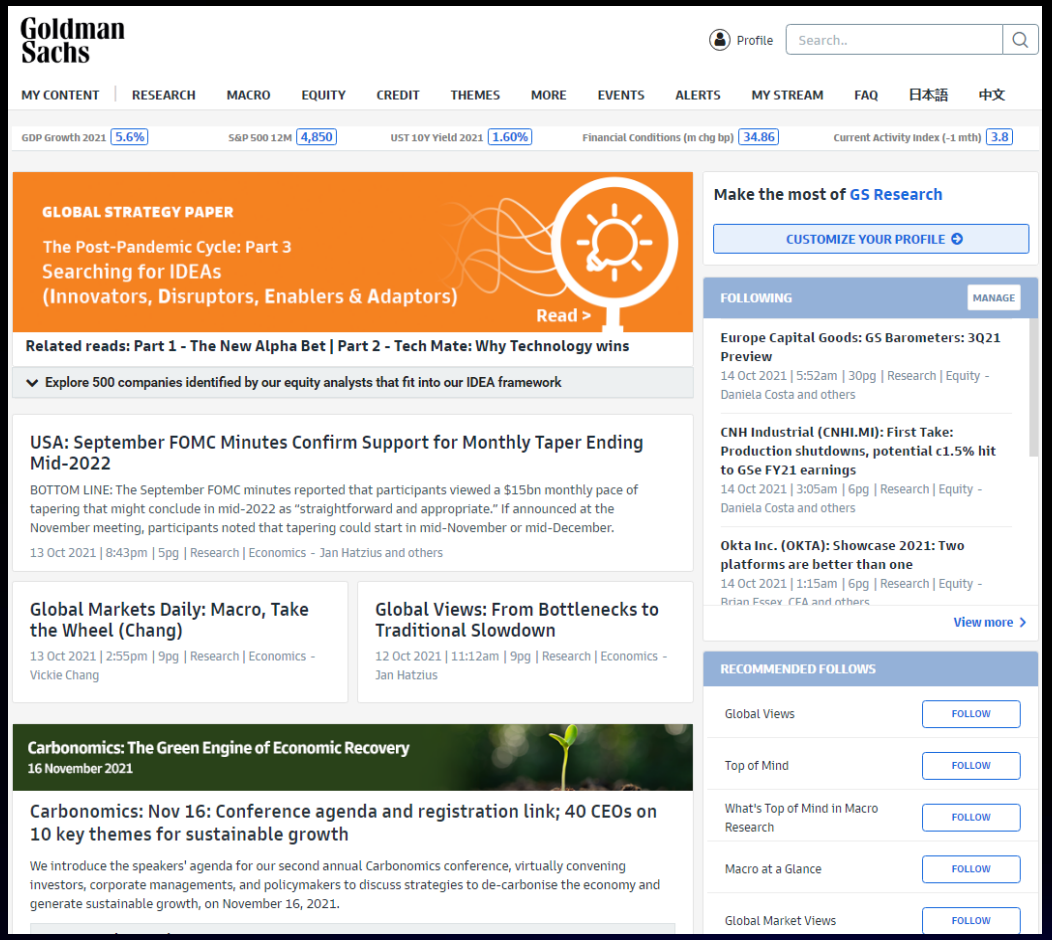
Recommend most relevant content to our clients



GS Now App



GS Research Portal



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recommendation engines

Report recommendations

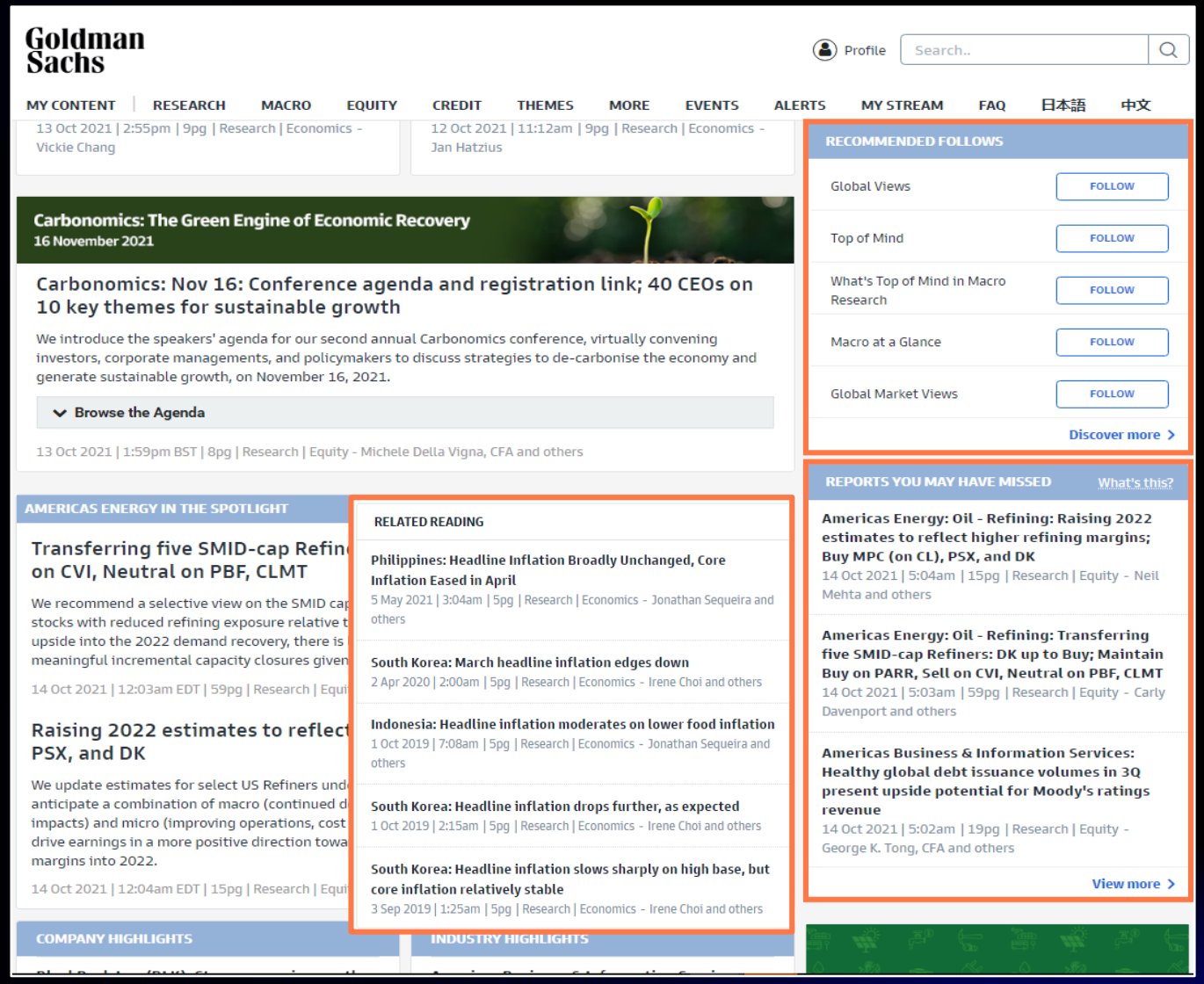
XGBoost

Follow recommendations

Collaborative filtering

Related reading

Universal sentence encoder



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our requirements



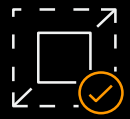
Cloud-native



Fast model predictions



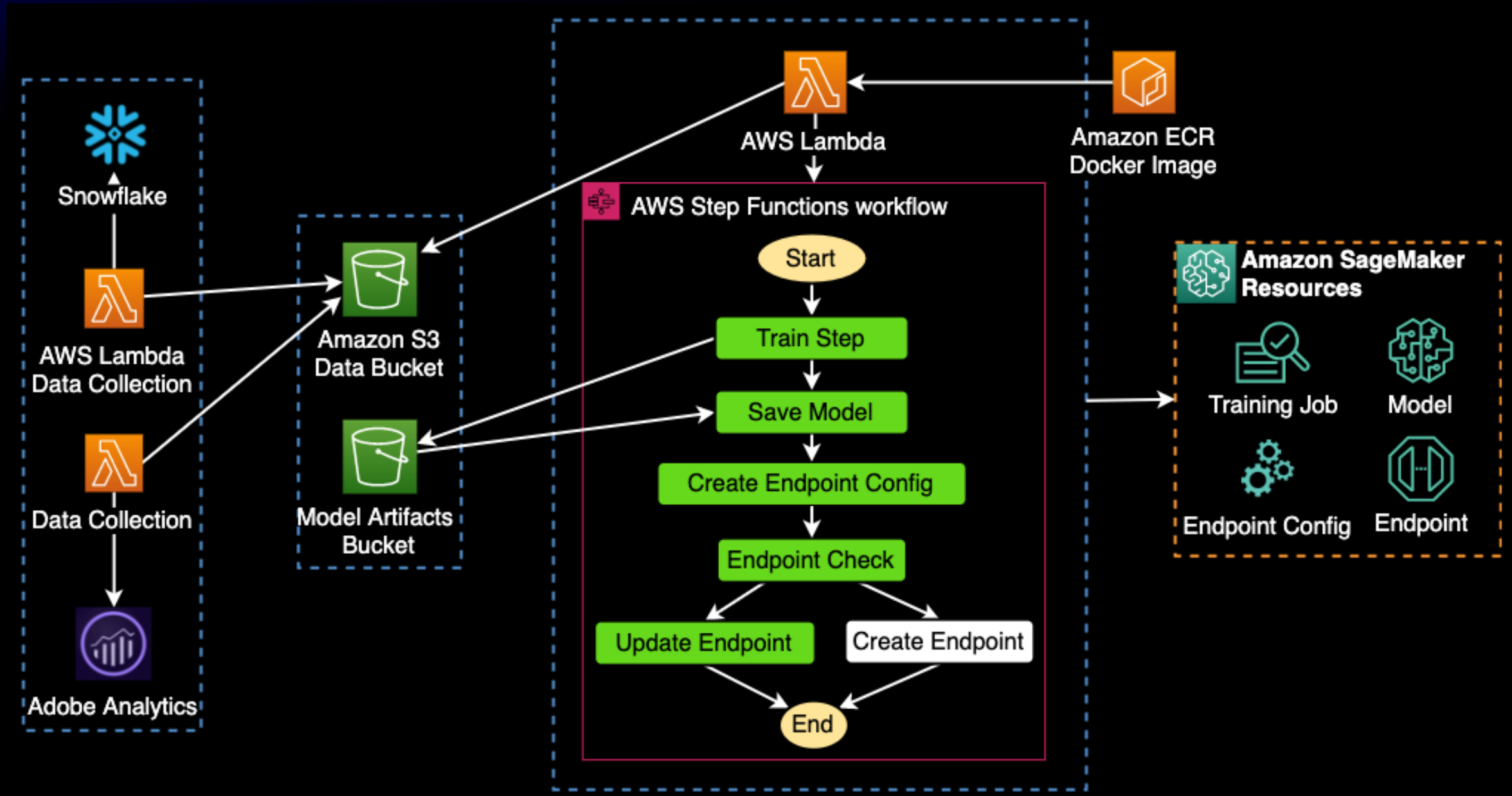
Quick to deploy models



Highly available and scalable

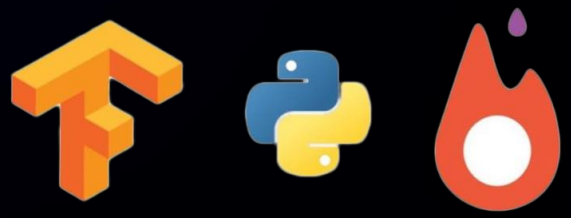
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

System Architecture



Deployment

- Custom container



- Ease of choosing instance type

```
ENDPOINT_INSTANCE_TYPE: 'ml.c5.xlarge'
ENDPOINT_INSTANCE_COUNT: 3
```

- Resiliency: Multiple instances

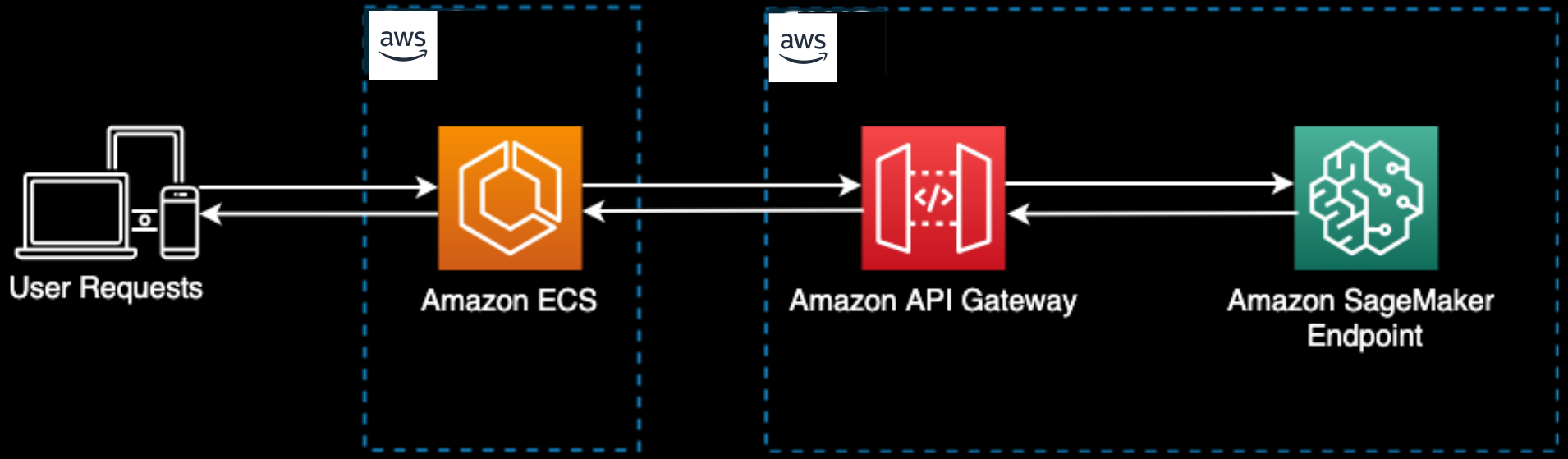
```
endpoint_config_step = EndpointConfigStep(
  "Create Endpoint Config",
  endpoint_config_name=execution_input['ModelName'],
  model_name=execution_input['ModelName'],
  initial_instance_count=ENDPOINT_INSTANCE_TYPE,
  instance_type=ENDPOINT_INSTANCE_COUNT)
```

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference

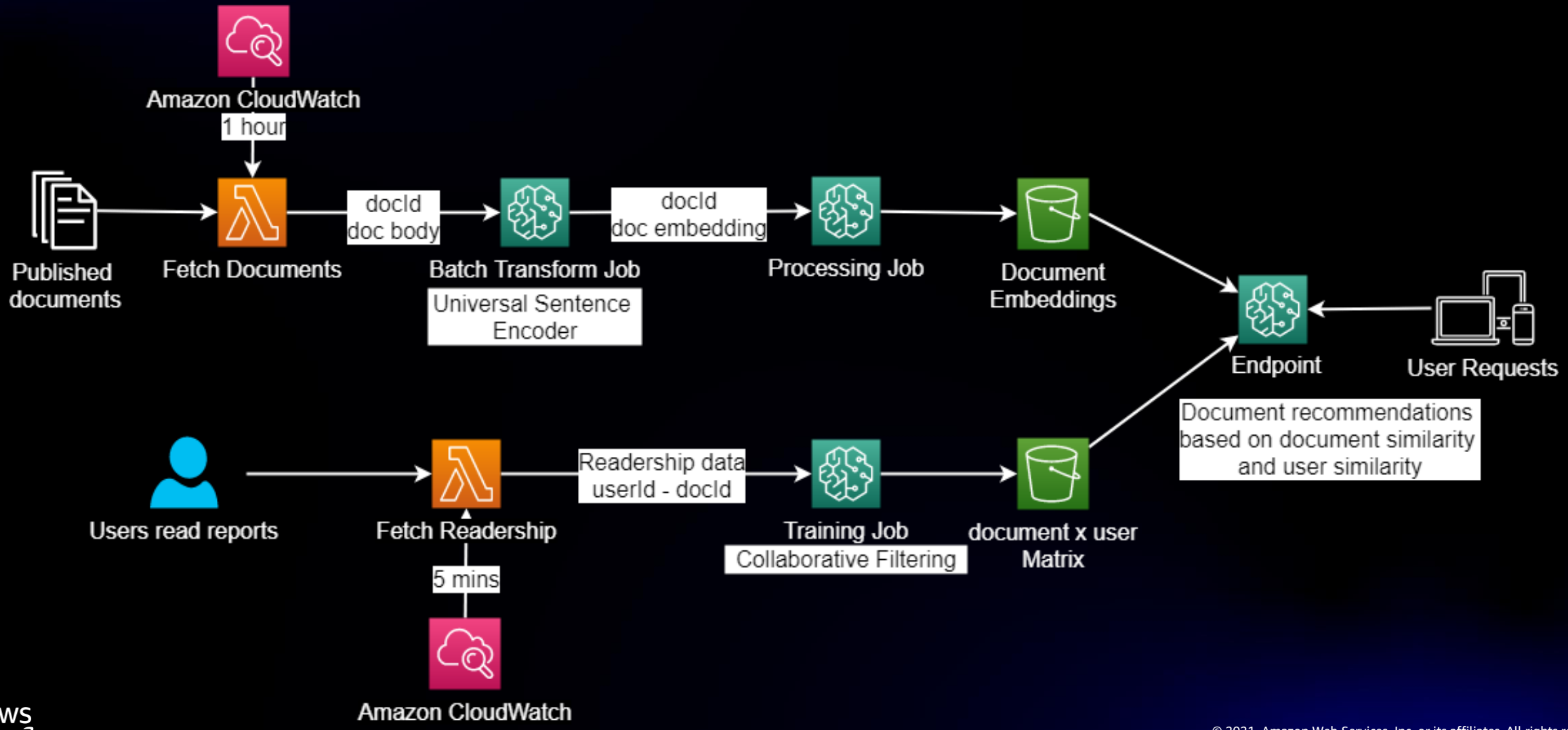
REAL-TIME



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference

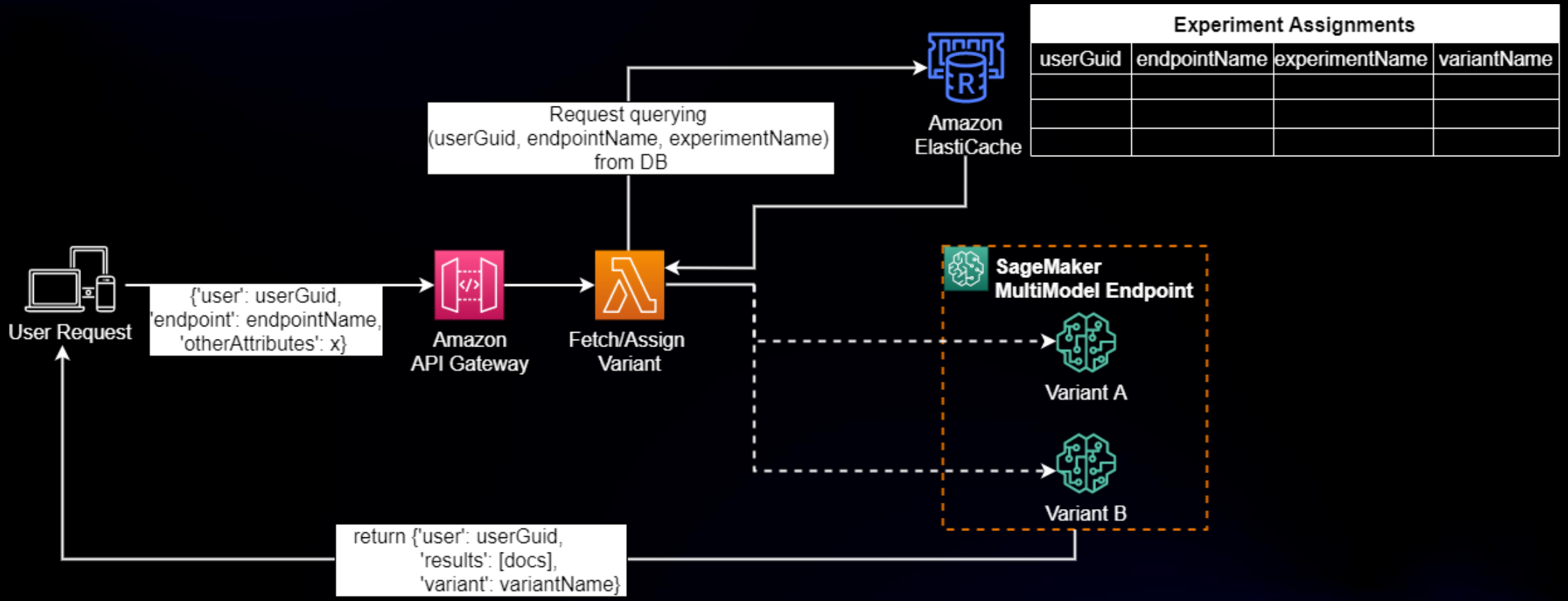
BATCH TRANSFORM



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inference

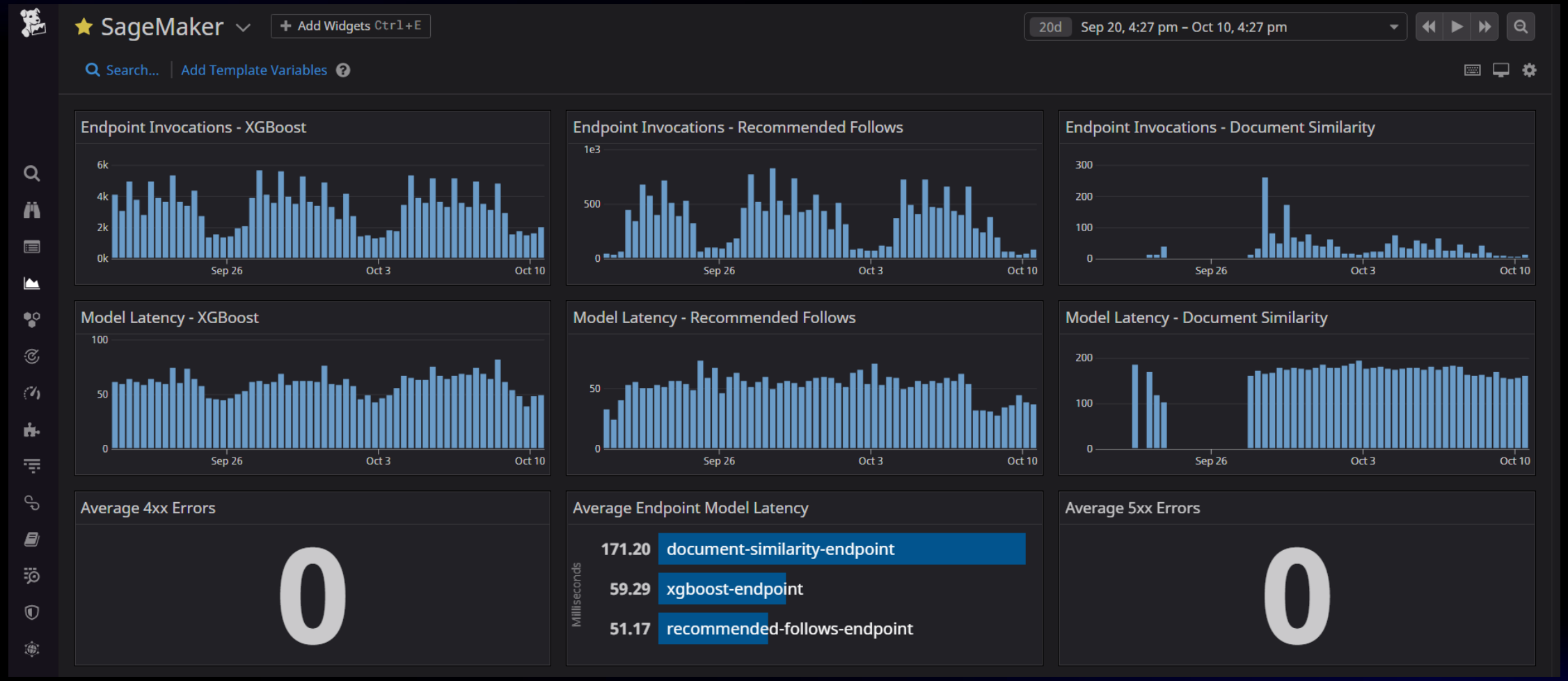
MULTI-MODEL ENDPOINTS



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Monitoring

DATADOG – SAGEMAKER INTEGRATION



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

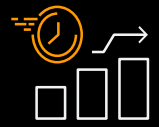
What we achieved?



Low-latency predictions: About 150ms



Highly available endpoints: No downtime



Ease of productionizing ML models: 60% more efficient

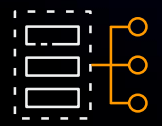


Cost-effective: 1/3 compared to on-premise

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Highlights



Cloud-native end-to-end architecture



MLOps best practices



Support from AWS team



4 models in production in ~1 year

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What next?

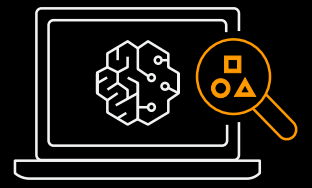
FUTURE WORK



SageMaker Model Monitor



SageMaker Feature Store



SageMaker Clarify

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Michael Hsieh

GitHub/LinkedIn
@michaelhsieh42

Pravallika Etoori

LinkedIn: https://uk.linkedin.com/in/etoori

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.