AWS re:Invent

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

AIM328

Amazon SageMaker serverless inference (Preview)

Rishabh Ray Chaudhury
Sr. Product Manager, Amazon SageMaker

Ram Vegiraju
ML Solution Architect, Amazon SageMaker

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

Agenda

Why Serverless inference ?

Overview of Amazon SageMaker Serverless Inference

How to deploy on Serverless Inference?

Demo

Q&A

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

SageMaker inference options

NEW

Real-time inference

Low latency

Ultra high throughput

Multi-model endpoints

A/B testing

Batch transform

Process large datasets

Job-based system

Asynchronous inference

Near real-time

Large payloads (1 GB)

Long timeouts (15 mins)

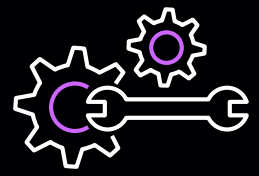
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Current customer challenges with ML inference



Workloads are intermittent

Some ML workloads have less predictable usage patterns and long periods of inactivity



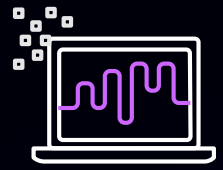
Challenging to provision capacity

Data scientists are challenged with selecting optimal instance types and managing autoscaling policies



End up over-provisioning capacity

Utilization is low and costs are high regardless of number of requests



Increases TCO

Spend lot of time in provisioning and managing servers

aws

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

PREVIEW

Introducing Amazon SageMaker Serverless Inference



First purpose built serverless ML inference in cloud



Fully managed



Pay only for what you use, billed in milliseconds

aws

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

What are the key benefits?

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

1. Fully managed offering



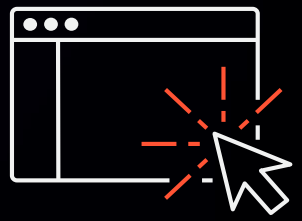
Managed infrastructure

Security

Monitoring

Logging

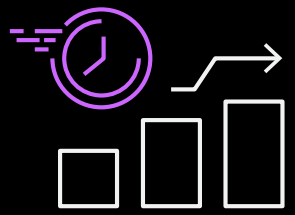
Built in availability and fault tolerance



Serverless

No need to select instance types or provision capacity

Choose memory options based on inference processing needs



Automatically scale out, in, and down to 0

No need to set scaling policies

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

2. Pay only for what you use

INFERENCE DURATION PRICING

Memory (MB)	Price per sec
1024	$0.000020
2048	$0.000040
3072	$0.000060
4096	$0.000080
5120	$0.000100
6144	$0.000120

Pay only for the duration to process each request and for data processing

No charge for idle periods

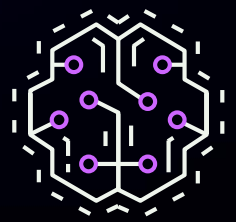
Billed in milli-seconds granularity

Free tier: 150k seconds free inference duration/month for the first 2 months

aws

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

3. Purpose built for ML



ML science

1P algorithms

Optimized ML frameworks

BYO



Seamlessly move from serverless to real-time

One-click update

No changes required to container

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

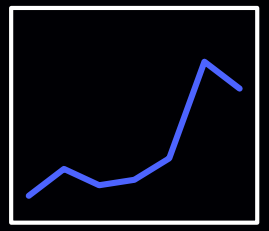
Example customer use case



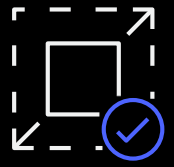
Sarah works as a DevOps Engineer for an online Food Delivery App



She uses ML to provide personalized restaurant recommendations to customers



However, the traffic pattern is highly unpredictable



1

Managing auto-scaling policies is a challenge

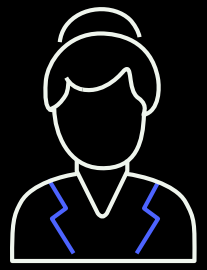


2

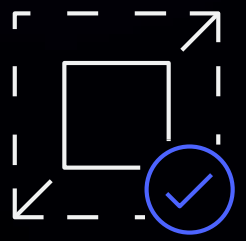
Inference costs go up as they end up over-provisioning instances

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

Example customer use case



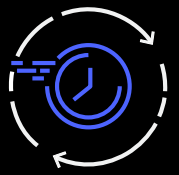
Sarah then decides to deploy on serverless



No need to manage auto-scaling policies, it scales automatically based on demand.

No need to over-provision instances

1



Fully managed service, so saves her time and helps her prioritize other projects

2



Inference costs go down as they only pay for the usage and not for idle time

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

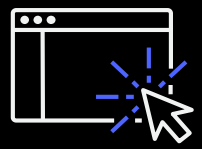
How to deploy on SageMaker Serverless Inference?

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

Simple one-click deployment

1 
ECR image location for inference code

2 
S3 location for model artifacts

3 
Choose a memory size

Serverless Inference Endpoint



Automatically spins up and manages compute resources

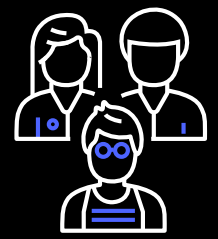
Automatic scaling based on demand

Managed logging and monitoring

Sends inference requests

Inference results

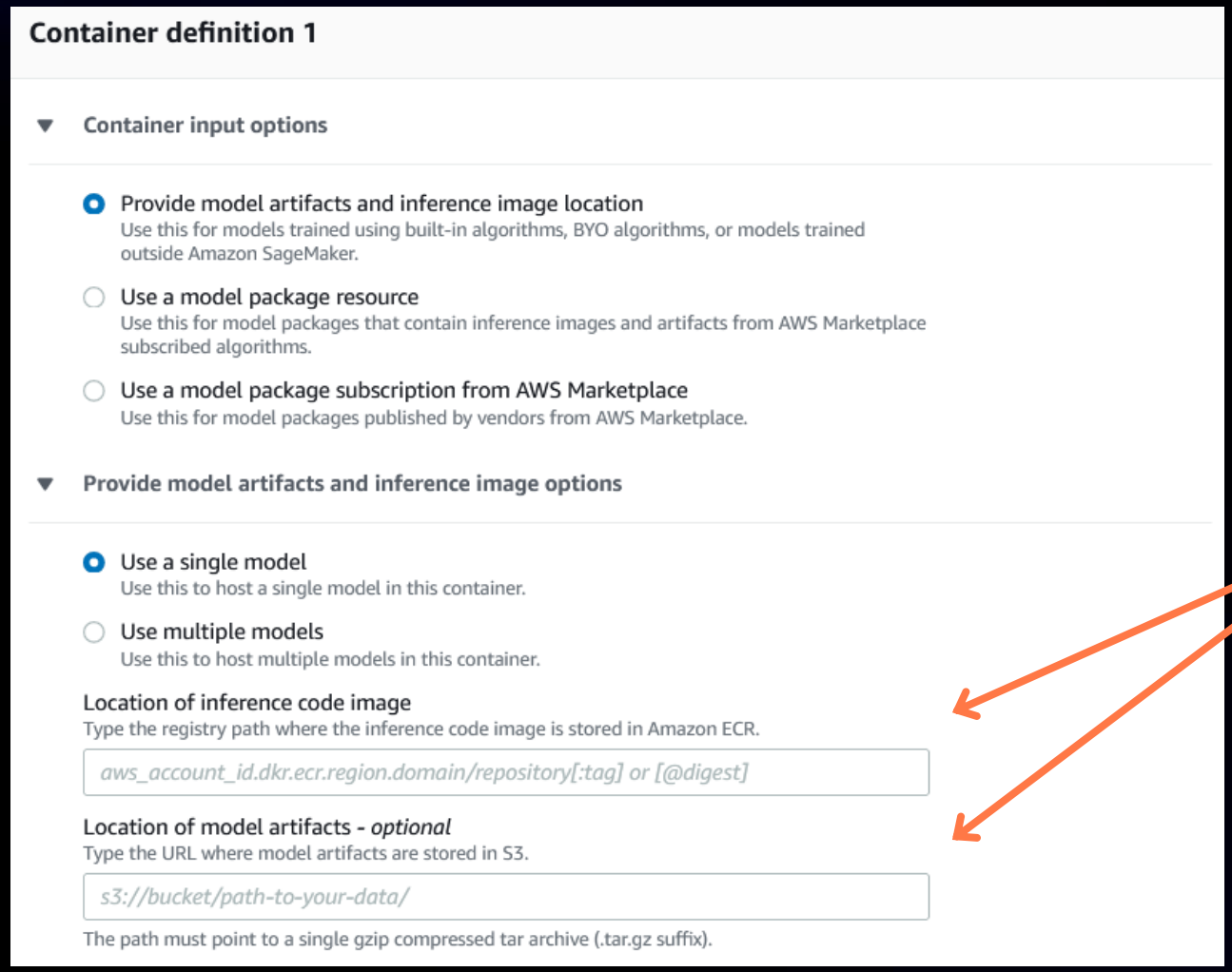
Trigger from client application or other AWS services



End user

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

Console experience



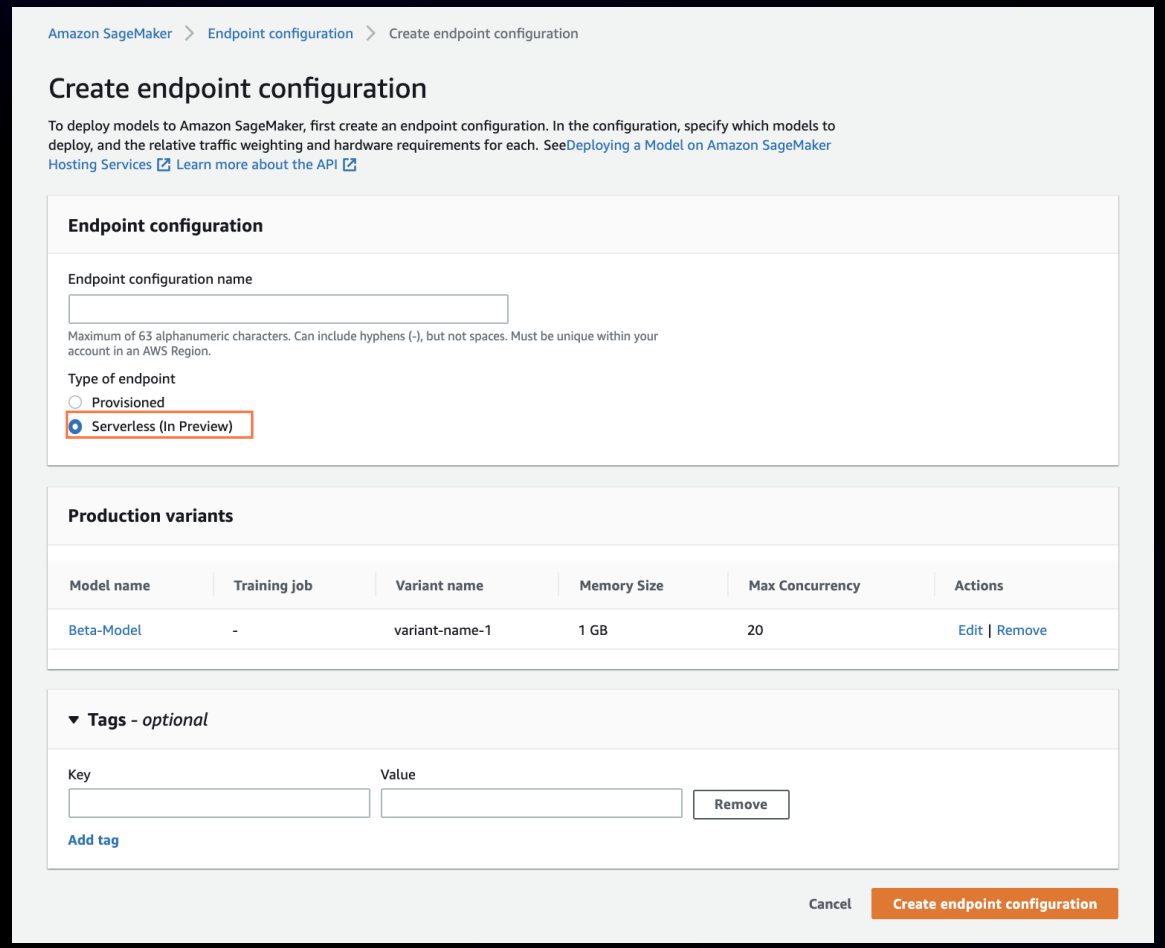
First create a model by providing ECR image location and S3 location for model artifacts

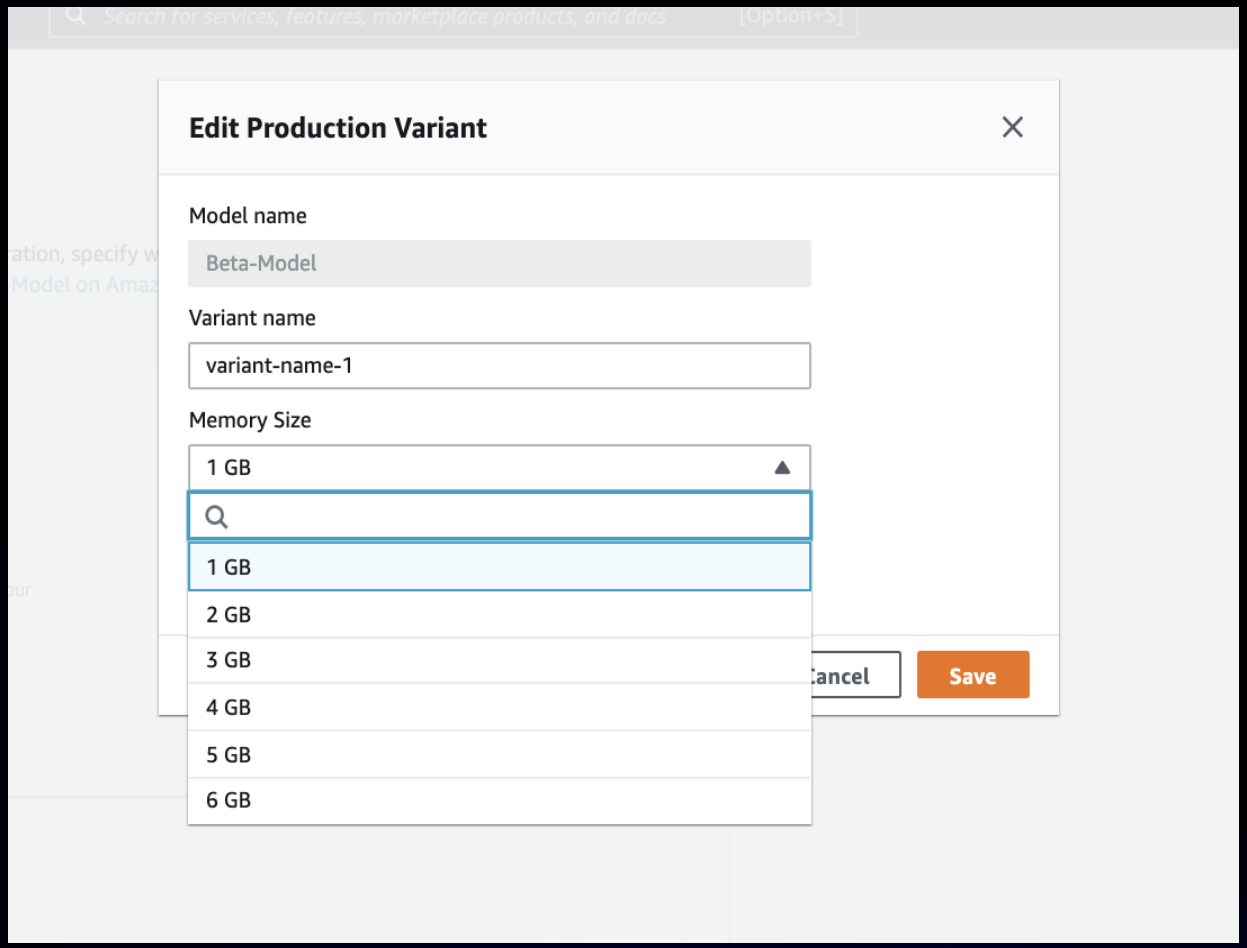
No changes to current experience

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

Console experience

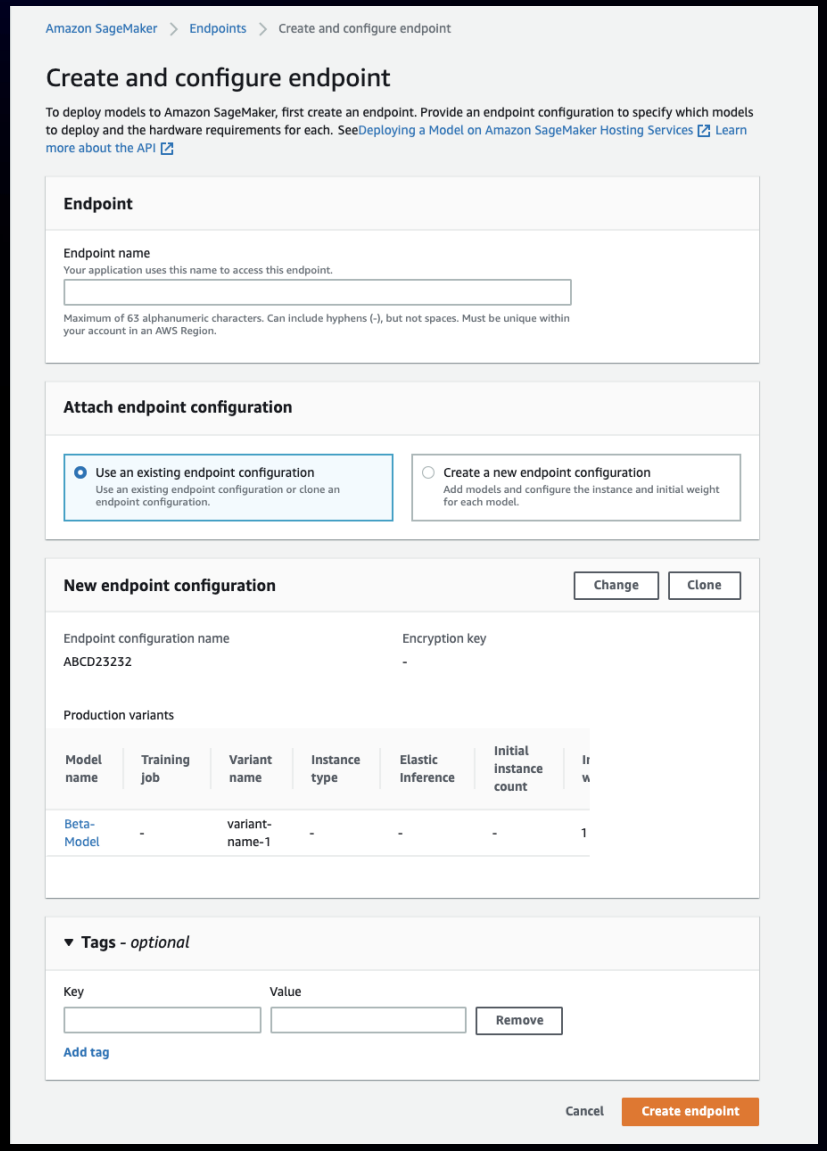
THEN CREATE ENDPOINT CONFIGURATION BY SELECTING SERVERLESS OPTION





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

Console experience



Last step is to create an endpoint using the configuration you chose earlier

No changes to current experience

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

How to get started?

Regions in preview

U.S. East (Northern Virginia), U.S. East (Ohio), U.S. West (Oregon), E.U. (Ireland), Asia Pacific (Tokyo), and Asia Pacific (Sydney)

Documentation—https://docs.aws.amazon.com/sagemaker/latest/dg/serverless-endpoints.html

FAQs—https://aws.amazon.com/sagemaker/faqs/

Example Notebook —https://github.com/aws/amazon-sagemaker-examples/tree/master/serverless-inference/Serverless-Inference-Walkthrough.ipynb

aws

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

Thank you!

Rishabh Ray Chaudhury
rishrayc@amazon.com

Ram Vegiraju
rvegira@amazon.com

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