AWS re:Invent

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

AIM416

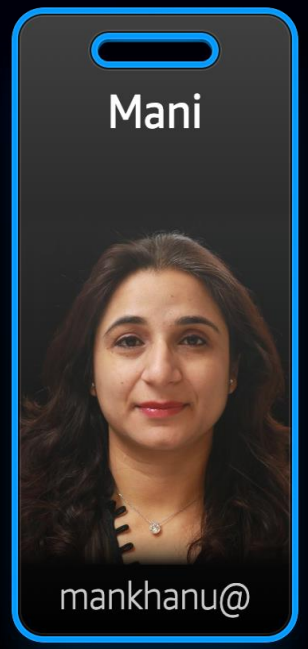
Using Hugging Face models on Amazon SageMaker

Mani Khanuja
AI/ML Specialist SA
Amazon Web Services

Pallavi Nargund
Sr. Solutions Architect
Amazon Web Services

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

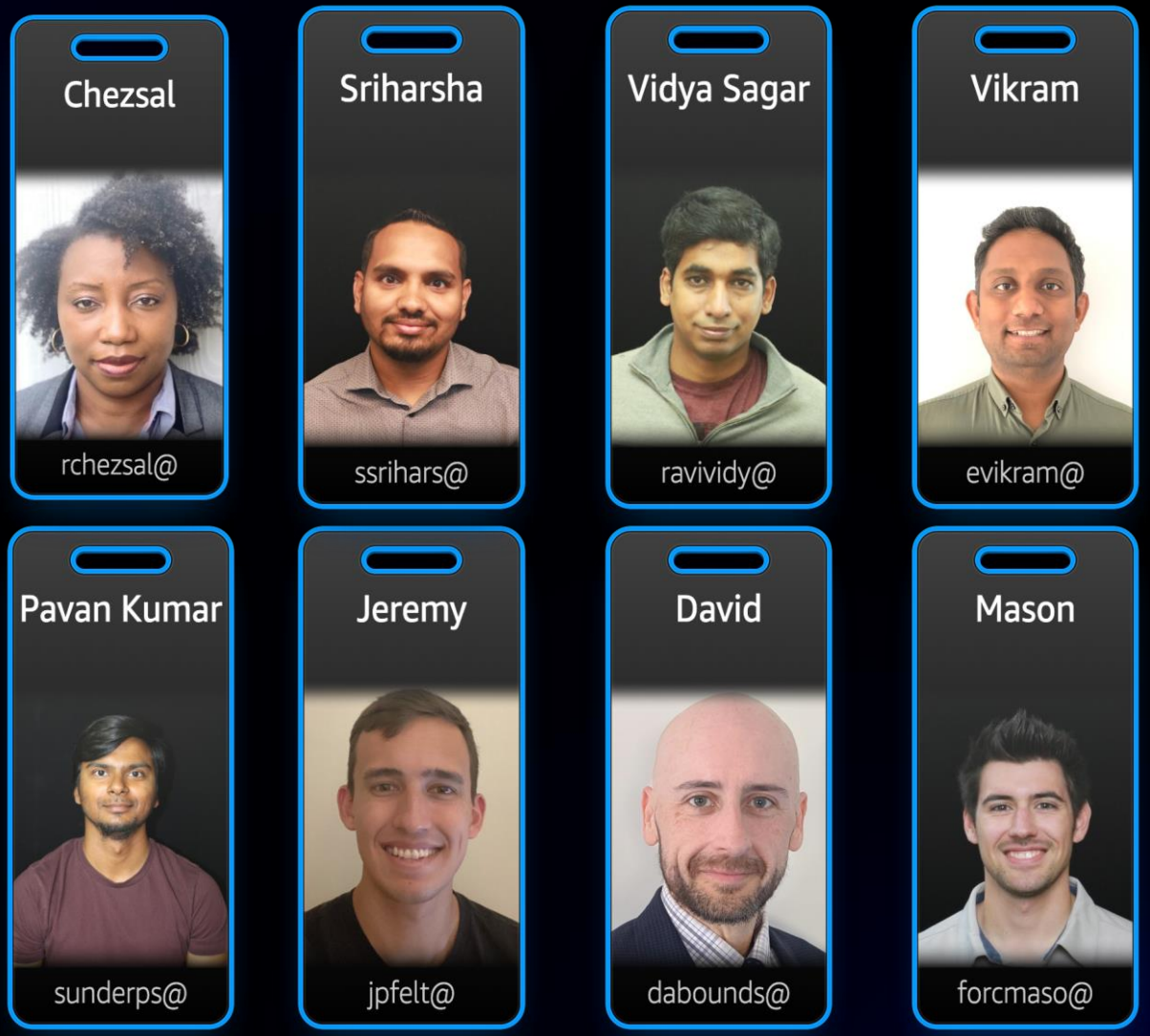
Session presenters





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

Workshop support



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

Agenda

- Overview of Natural Language Processing
- Overview of Hugging Face with Amazon SageMaker
- Hands-on labs

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

Prerequisites

Familiarity with

- AWS Management Console
- Amazon SageMaker
- JupyterLab
- Natural language processing (NLP)

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

Natural Language Processing (NLP)

NLP serves many human-friendly functions and its usage can have huge impact on the business of a company



Language translation

NLP has made machine translation so fast and accurate that translation becomes accessible to many applications



Spell check and grammar check

This insures higher data quality from the very source of data, reducing complex and tedious data cleansing operations



Smart assistants

Offering human-friendly interactions using voice to a broad category of devices



Natural query and search

NLP allows powerful query and search capabilities using natural language questions and human-oriented interfaces



Text analytics

Extracting topics or sentiments from unstructured text to serve analytics and data-driven business decisions

aws

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

What are the Hugging Face libraries?



Open source
Datasets, tokenizers and transformers



Popular
53,000+ GitHub stars, 1,000,000+ downloads per month



Intuitive
NLP-specific Python frontends based on PyTorch or TensorFlow



State-of-the-art
Transformer-based models are state-of-the-art, enable transfer-learning and scale



Comprehensive
Model hub with 15,000+ model architectures, 240+ languages

aws

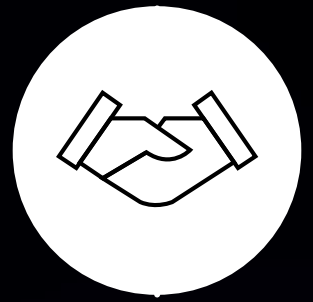
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

A strong collaboration to make NLP easy and accessible for all

Hugging Face



Hugging Face is the most popular open source company providing state-of-the-art NLP technology



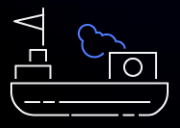
AWS



Amazon SageMaker offers high performance resources to train and use NLP models

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

Introducing a new Hugging Face experience in Amazon SageMaker



Deep learning containers (DLCs) developed with Hugging Face for both training and inference for the PyTorch and TensorFlow frameworks



A Hugging Face estimator in the SageMaker SDK to launch NLP scripts on scalable, cost-effective SageMaker training jobs without worrying about Docker



An example gallery to find readily usable high-quality samples of Hugging Face scripts on Amazon SageMaker



Maintained and supported by AWS

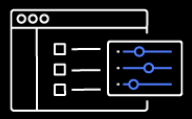
aws

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

What are the benefits of running Hugging Face in Amazon SageMaker?



Cost effective – SageMaker optimizes scale, performance, and effectiveness to reduce costs; Amazon EC2 Spot Instances further save costs for non-critical workloads



ML ops ready – Includes automated metadata persistence and search in the SageMaker metastore, log extraction to Amazon CloudWatch, monitoring with SageMaker Debugger and Profiler and experiment management



Scalable – Able to run on clusters of GPUs, with efficient data-parallel and model-parallel distribution provided by SageMaker; ability to launch several concurrent jobs at the same time with the async mode of the API

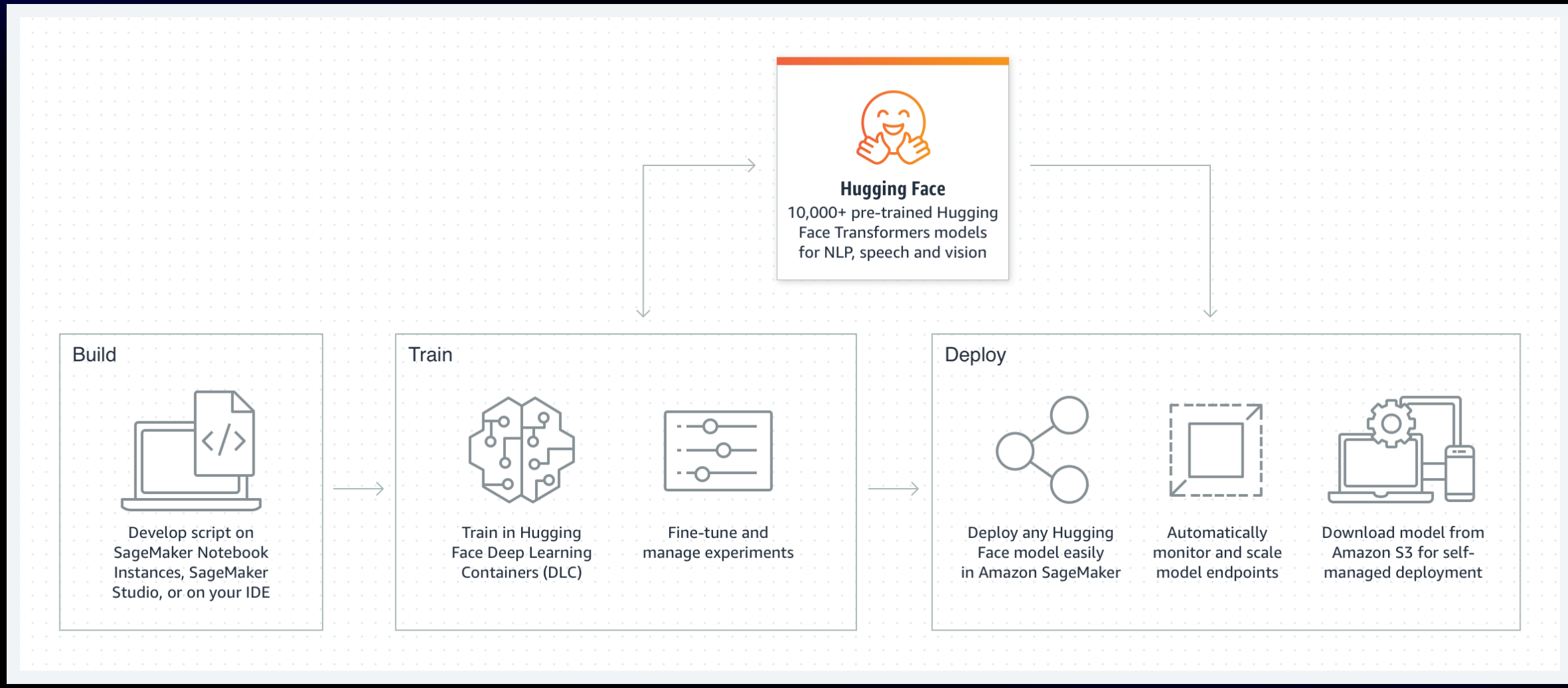


Secure – High bar on security, with available mechanisms including encryption at rest and in transit, VPC connectivity and fine-grained IAM permissions

aws

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

Integrated workflow with Amazon SageMaker



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

Training on Amazon SageMaker

Hugging Face estimator

```
# metric definition to extract the results
metric_definitions=[
     {"Name": "train_runtime", "Regex": "train_runtime.*=\D*(.*?)$"},
     {'Name': 'train_samples_per_second', 'Regex': "train_samples_per_second.*=\D*(.*?)$"},
     {'Name': 'epoch', 'Regex': "epoch.*=\D*(.*?)$"},
     {'Name': 'f1', 'Regex': "f1.*=\D*(.*?)$"},
     {'Name': 'exact_match', 'Regex': "exact_match.*=\D*(.*?)$"}]

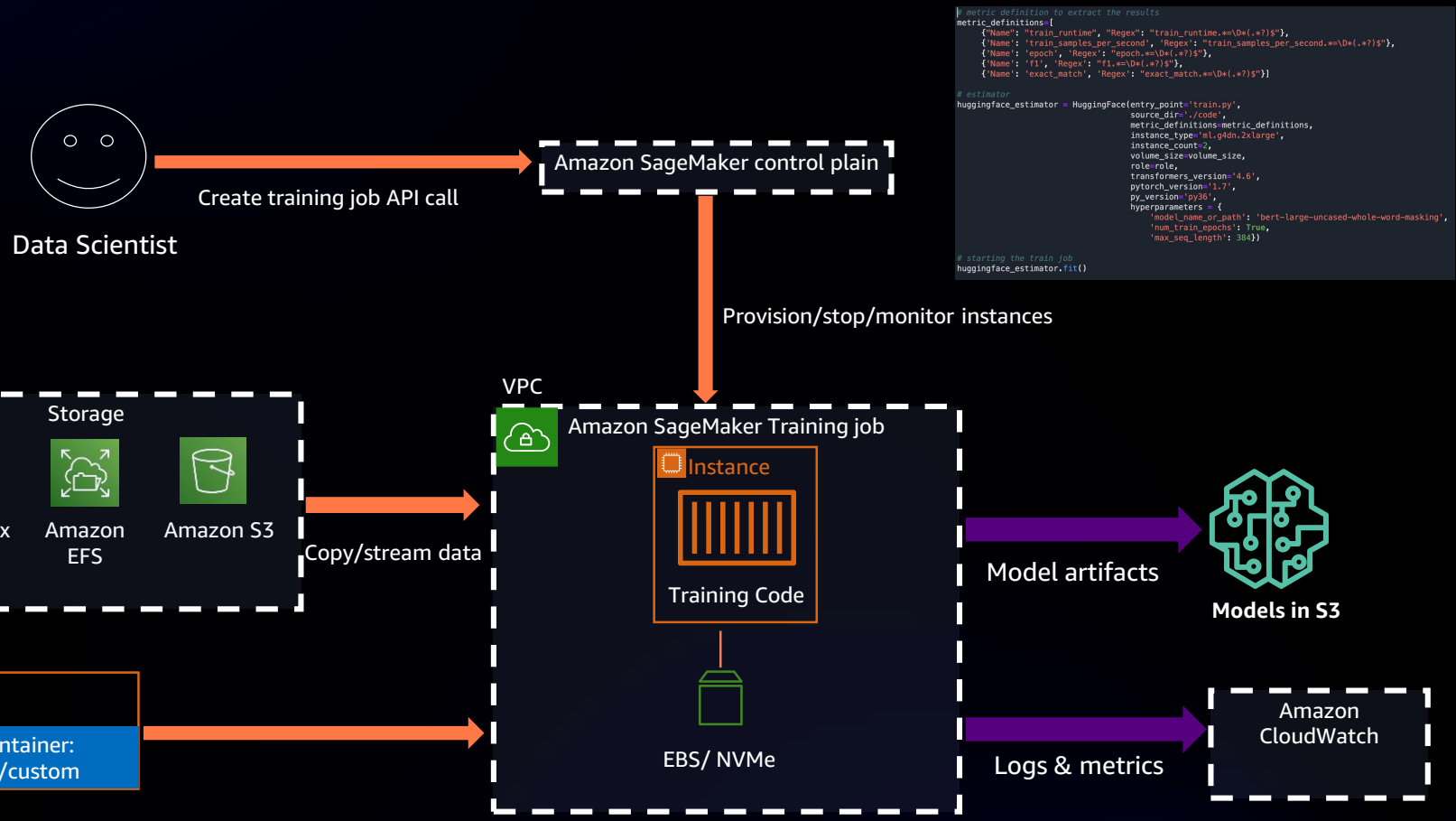
# estimator
huggingface_estimator = HuggingFace(entry_point='train.py',
                                    source_dir='./code',
                                    metric_definitions=metric_definitions,
                                    instance_type='ml.g4dn.2xlarge',
                                    instance_count=2,
                                    volume_size=volume_size,
                                    role=role,
                                    transformers_version='4.6',
                                    pytorch_version='1.7',
                                    py_version='py36',
                                    hyperparameters = {
                                        'model_name_or_path': 'bert-large-uncased-whole-word-masking',
                                        'num_train_epochs': True,
                                        'max_seq_length': 384})

# starting the train job
huggingface_estimator.fit()
```

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

Training on Amazon SageMaker

A REFRESHER




```
# metric definition to extract the results
metric_definitions=[
     {"Name": "train_runtime", "Regex": "train_runtime.*=\D*(.*?)$"},
     {'Name': 'train_samples_per_second', 'Regex': "train_samples_per_second.*=\D*(.*?)$"},
     {'Name': 'epoch', 'Regex': "epoch.*=\D*(.*?)$"},
     {'Name': 'f1', 'Regex': "f1.*=\D*(.*?)$"},
     {'Name': 'exact_match', 'Regex': "exact_match.*=\D*(.*?)$"}]

# estimator
huggingface_estimator = HuggingFace(entry_point='train.py',
                                    source_dir='./code',
                                    metric_definitions=metric_definitions,
                                    instance_type='ml.g4dn.2xlarge',
                                    instance_count=2,
                                    volume_size=volume_size,
                                    role=role,
                                    transformers_version='4.6',
                                    pytorch_version='1.7',
                                    py_version='py36',
                                    hyperparameters = {
                                        'model_name_or_path': 'bert-large-uncased-whole-word-masking',
                                        'num_train_epochs': True,
                                        'max_seq_length': 384})

# starting the train job
huggingface_estimator.fit()
```

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

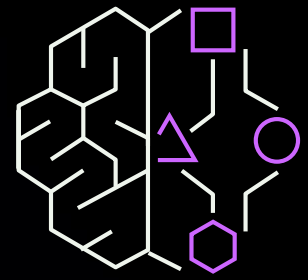
Amazon SageMaker distributed training libraries

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

Challenges with large datasets and models



Large datasets take a long time to train, creating a bottleneck

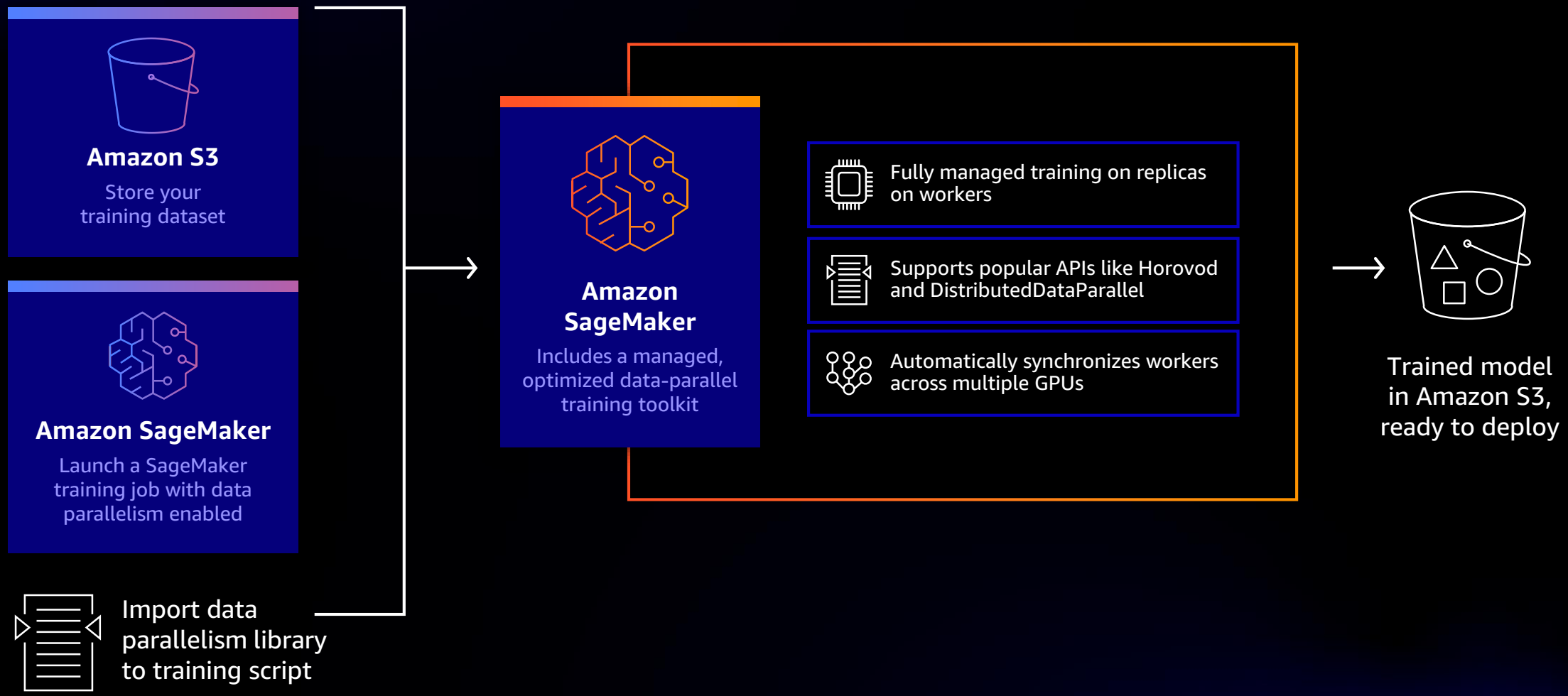


Large models or models with many parameters can exceed the memory capacity of a single GPU

aws

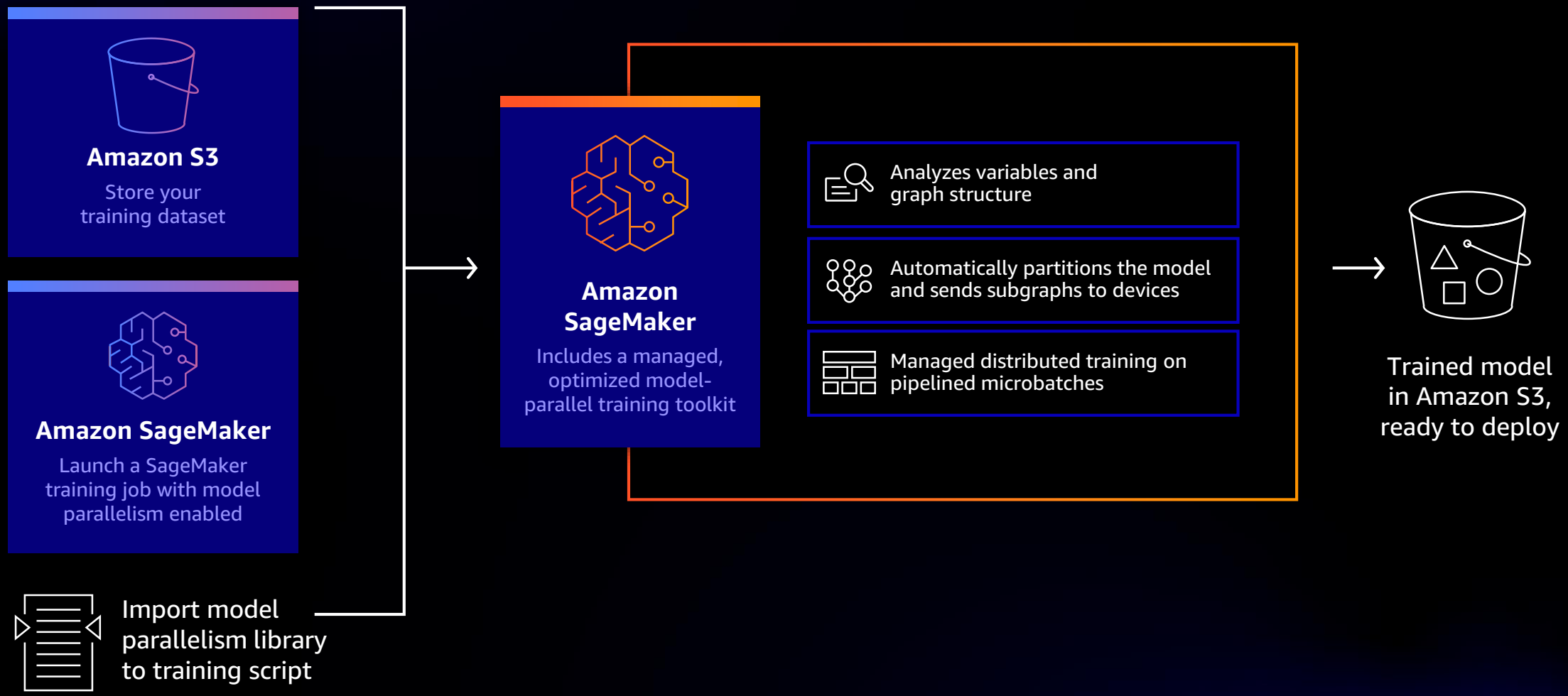
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How it works: Data parallelism library



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

How it works: Model parallelism library

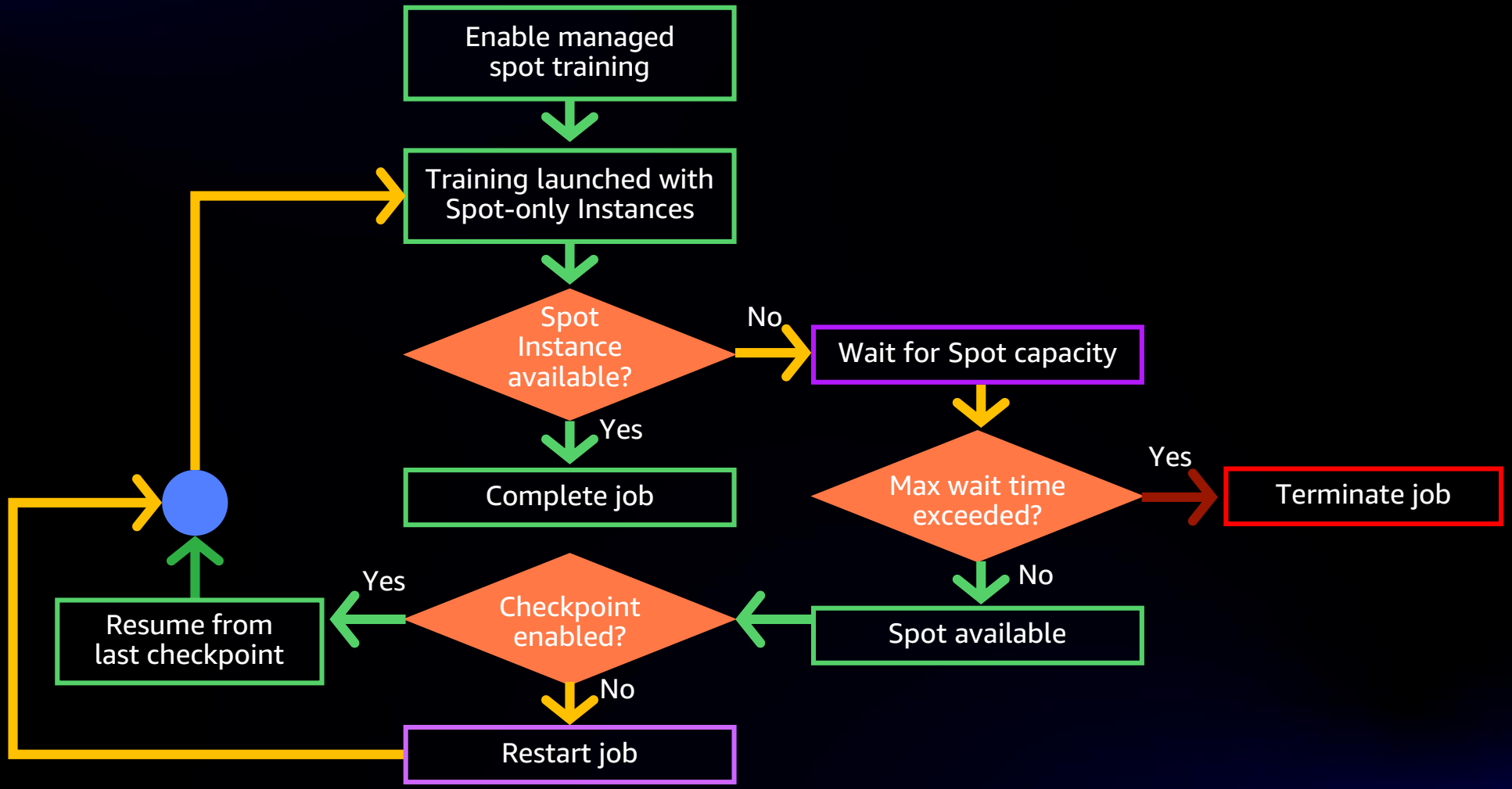


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Managed spot training – Save up to 90% on model training costs

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

Managed spot training workflow



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

Train on Amazon SageMaker

HANDS-ON LAB – DURATION: 45 MINUTES

- Prepare and upload train and test datasets to Amazon S3
- Prepare a fine-tuning script to be used with Amazon SageMaker training jobs
- Launch a training job and store the trained model in Amazon S3

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

Let’s get started . . .

Works on both Amazon SageMaker Studio notebook and Amazon SageMaker notebook instances

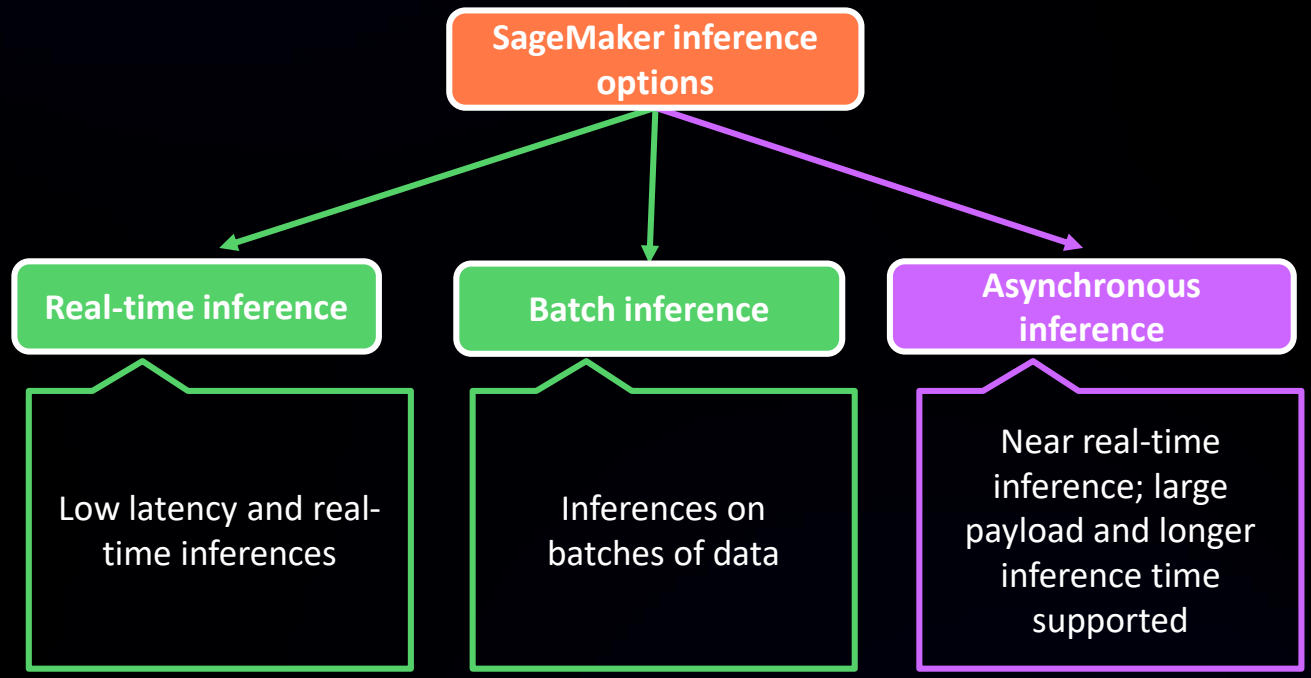
https://bit.ly/AIM416-2021

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

Managed deployment

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

Amazon SageMaker inference options

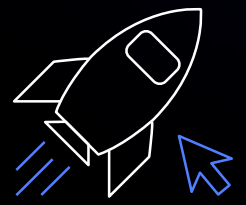


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deployment flexibility

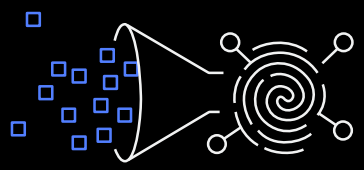
REAL-TIME OR BATCH INFERENCE

SageMaker hosting



- Create a long-running microservice
- Effectively instant response for small data batches
- Accessible from an external application

SageMaker batch transform



- Bulk data transformation
- Pay only for what you use
- Periodic arrival of big chunks of data

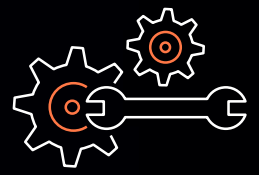
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Problem – Emerging use cases for near real-time inference



Inference payloads are large

Customers want to process large payloads (100s of MBs or GBs) such as high-def images and video frames



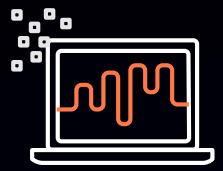
Model complexity is increasing

New deep learning models are complex and take long (in orders of minutes to process)



Latency-tolerant workloads

Real-time inference payloads have 5 MB payload limit and 60-second timeout; data needs to be processed as it arrives so cannot process offline as a batch; need near real-time inference



Need to control costs

Workloads don’t need millisecond latency so can tolerate cold starts; deep learning workloads need to control costs

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

Deploy trained model on Amazon SageMaker

HANDS-ON LAB OPTIONS: 30 MINUTES

Choose one of the following labs:

- Real-time inference – Load a pre-trained model from Amazon S3 and deploy on Amazon SageMaker endpoint
- Batch inference – Load a pre-trained model from Amazon S3 and deploy on Amazon SageMaker

Bonus lab

- Asynchronous inference – Load a pre-trained model from Amazon S3 and deploy on Amazon SageMaker endpoint

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

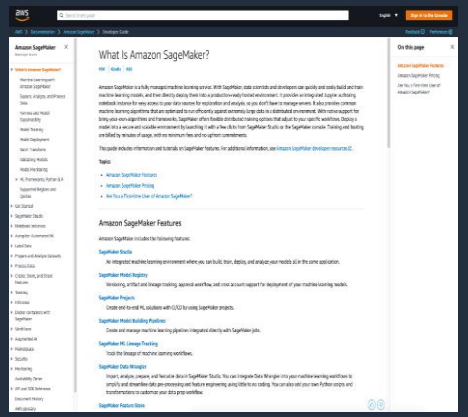
Where to get help?

Sample notebooks



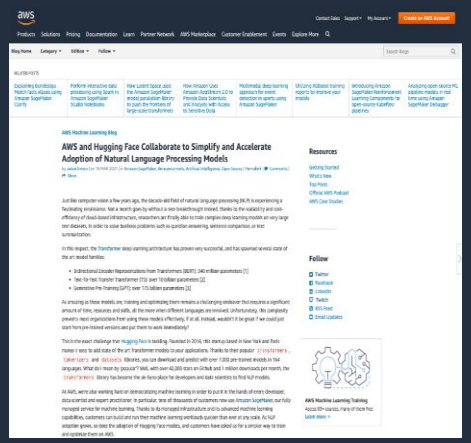
Access GitHub (https://github.com/huggingface/notebooks/tree/master/sagemaker)

Documentation



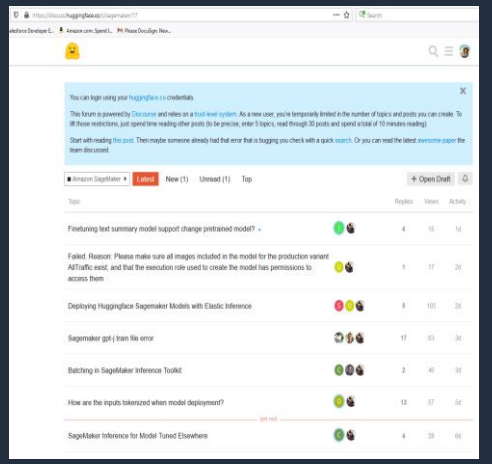
AWS Documentation (https://docs.aws.amazon.com/sagemaker/latest/dg/hugging-face.html)

Launch blog



Read the blogs
Training (https://aws.amazon.com/blogs/machine-learning/aws-and-hugging-face-collaborate-to-simplify-and-accelerate-adoption-of-natural-language-processing-models/)
Inference (https://aws.amazon.com/blogs/machine-learning/announcing-managed-inference-for-hugging-face-models-in-amazon-sagemaker/)

Hugging Face forum



https://discuss.huggingface.co/c/sagemaker/

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

Additional SageMaker sessions

- AIM402 – Develop your ML project with Amazon SageMaker
- AIM415 – Real-time recommendations using Amazon SageMaker Feature Store
- AIM328 – Automatically scale Amazon SageMaker endpoints for inference
- AIM405 – Rightsizing SageMaker compute instances for ML model training & inference
- AIM320 – Implementing MLOps practices with Amazon SageMaker, featuring Vanguard

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

Join us for a Happy Hour on Wednesday

Hosted by SageMaker Distributed Training Team

6–8 PM on Dec 1st

At Venetian Hotel Royal Brittania Pub

RSVP now https://reinventawsdthappyhour.splashthat.com/



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

Thank you!

Mani Khanuja
mankhanu@amazon.com

Pallavi Nargund
nargundp@amazon.com

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