



Distributed hosting on AWS with Amazon SageMaker

Generative AI Foundations on AWS

Emily Webber, Principal ML Specialist SA at AWS

Lesson 7 – **Level 400**

Today's activities

- Many ways to build your FM application
- Hosting options on AWS
- How to host a distributed model
- Optimizations
- Hands-on walk-through: large model serving container on AWS

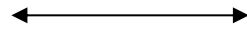


Reminder – everything we discuss today
is possible on AWS and SageMaker!

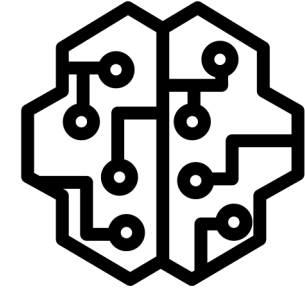
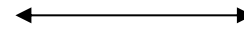
So you have a good model, now what?



Customers



Client application



Foundation model

Goal: Simplify customer experience

Goal: Streamline development lifecycle

Options for foundation model applications



Online



Offline



Queued



Embedded



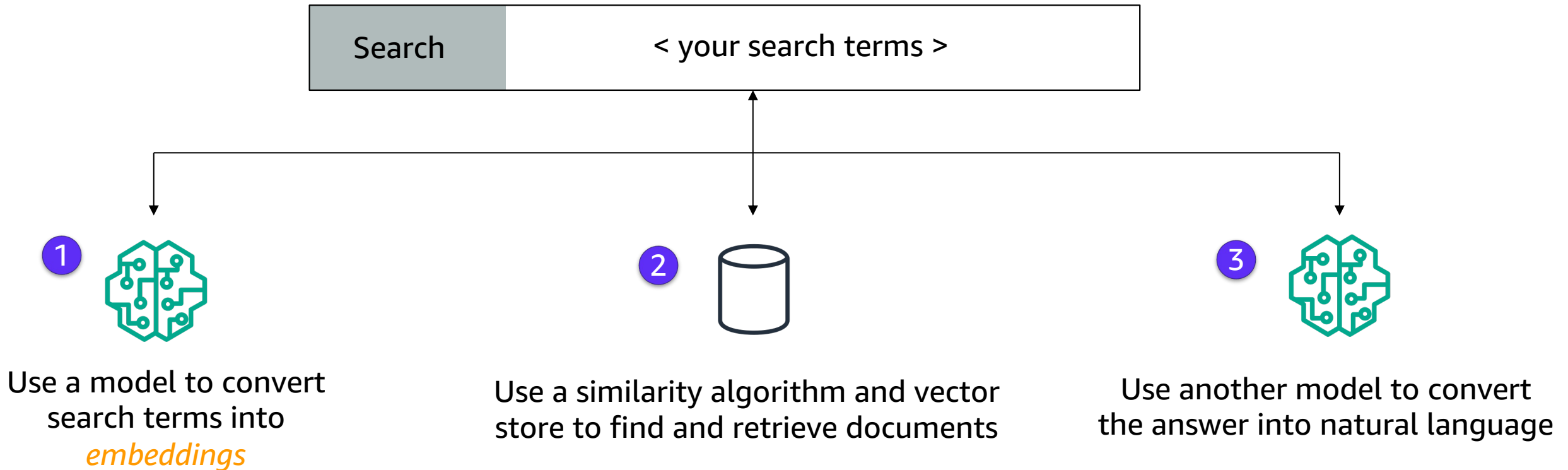
On-device



Serverless

Each option has **trade-offs** around latency, pricing, and model lifecycle

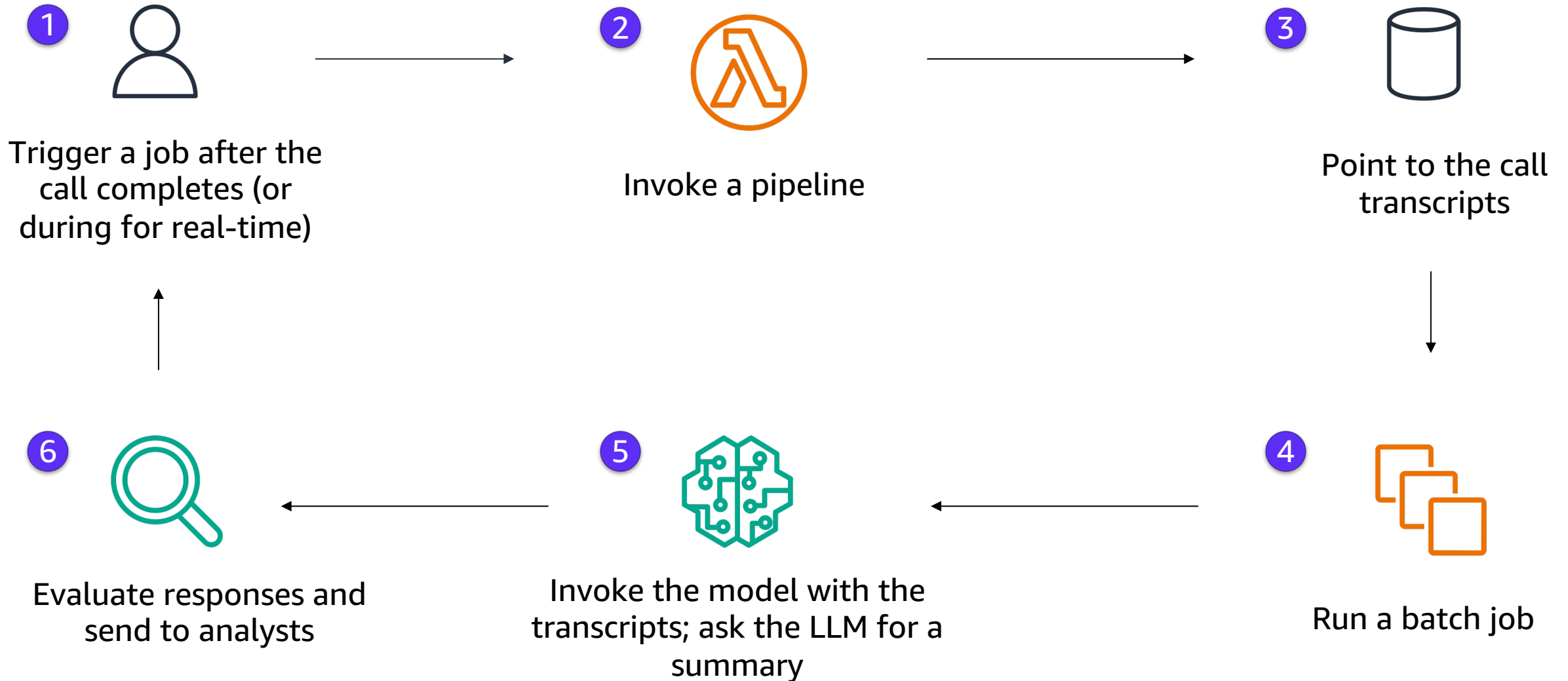
Online application example: search



Pro tip

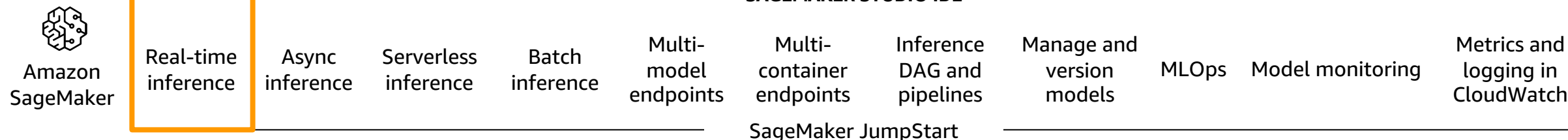
This is called *retrieval augmented generation*. It uses at least two models hosted *online* for immediate service to customers. Retrieving documents mitigates LLM hallucinations.

Offline application example: call center summarizations



Amazon SageMaker model deployment stack

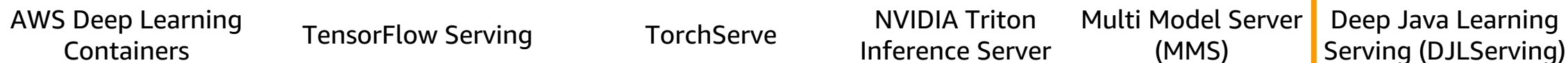
Amazon SageMaker



FRAMEWORKS



MODEL SERVERS



ML COMPUTE INSTANCES & ACCELERATORS

CPUs GPUs Inferentia & Trainium Graviton (ARM)

DEEP LEARNING COMPILERS AND RUNTIMES

SageMaker Neo NVIDIA TensorRT/cuDNN Intel oneDNN ARM Compute Library

What you need to host a model on SageMaker



Model artifact stored
in a bucket



Inference image
hosted in a registry



Managed ML
instances

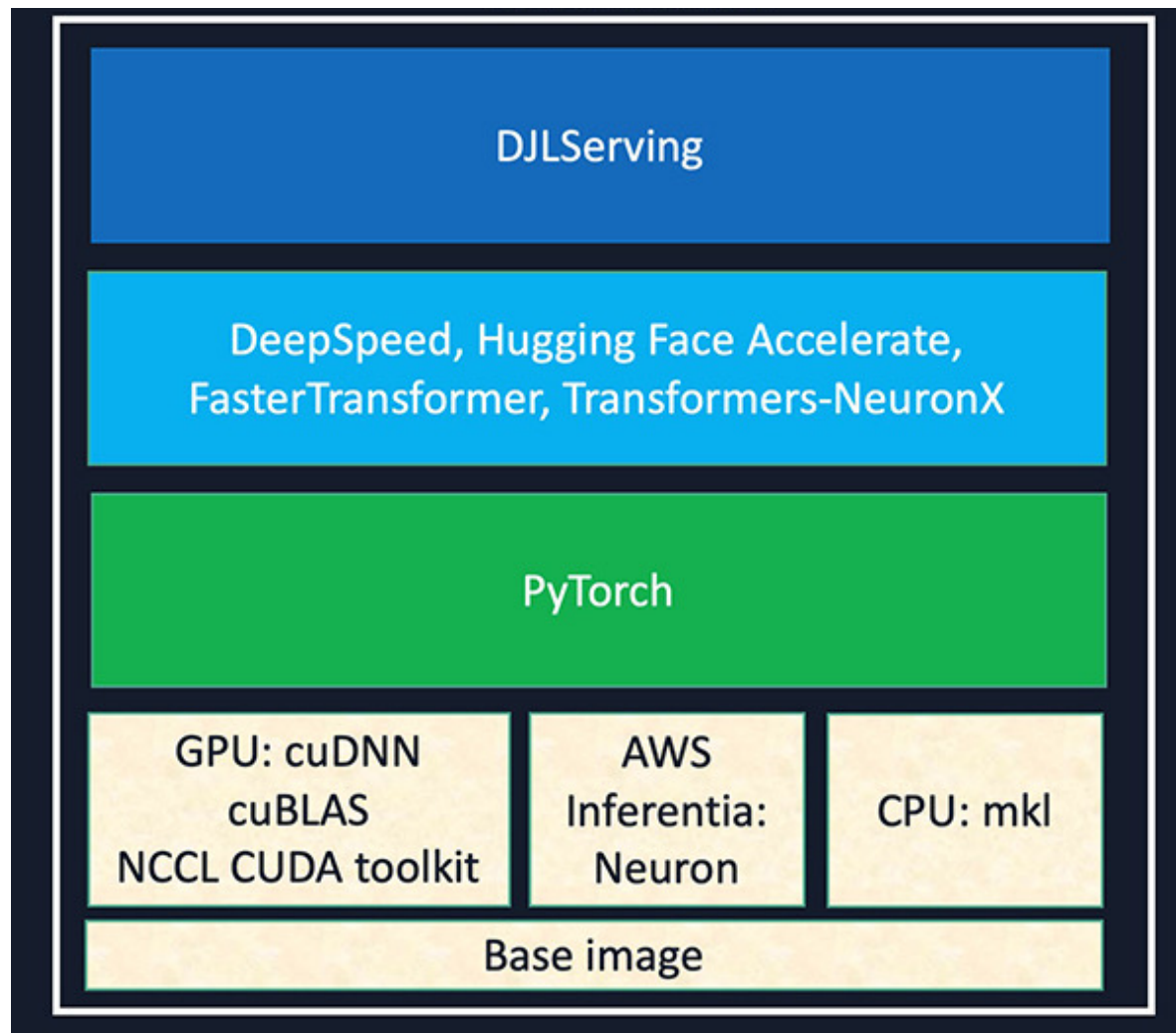
Pro tip

Point to the latest *AWS Deep Learning Container* to easily run your inference script.
Use the one built for your framework, and pick *large model hosting*.



AWS Deep Learning
Containers

SageMaker Large Model Inference Container



- Use pre-built images to distribute your model over multiple GPUs
- Implements *tensor and pipeline parallelism*
- Data parallel is not needed, because there's no backward pass
- Can also run with inferentia
- Faster model loading with *s5cmd*
- Integrates with top open-source frameworks

Large Model Inference Containers

Framework	Job Type	Accelerator	Python Version Options	Example URL
DJLServing 0.22.1 with FasterTransformer 5.3.0, HuggingFace Transformers 4.27.3, and HuggingFace Accelerate 0.17.1	inference	GPU	3.9 (py39)	763104351884.dkr.ecr.us-west-2.amazonaws.com/djl-inference:0.22.1-fastertransformer5.3.0-cu118
DJLServing 0.22.1 with DeepSpeed 0.8.3, HuggingFace Transformers 4.27.4, Diffusers 0.14.0 and HuggingFace Accelerate 0.18.0	inference	GPU	3.9 (py39)	763104351884.dkr.ecr.us-west-2.amazonaws.com/djl-inference:0.22.1-deepspeed0.8.3-cu118
DJLServing 0.22.1 with Neuron SDK 2.10.0, TransformersNeuronX 0.3.0 and HuggingFace Transformers 4.28.1	inference	Neuron	3.8 (py38)	763104351884.dkr.ecr.us-west-2.amazonaws.com/djl-inference:0.22.1-neuronx-sdk2.10.0
DJLServing 0.21.0 with FasterTransformer 5.3.0, HuggingFace Transformers 4.25.1, and HuggingFace Accelerate 0.15.0	inference	GPU	3.9 (py39)	763104351884.dkr.ecr.us-west-2.amazonaws.com/djl-inference:0.21.0-fastertransformer5.3.0-cu117
DJLServing 0.21.0 with DeepSpeed 0.8.3, HuggingFace Transformers 4.26.0, and HuggingFace Accelerate 0.16.0	inference	GPU	3.9 (py39)	763104351884.dkr.ecr.us-west-2.amazonaws.com/djl-inference:0.21.0-deepspeed0.8.3-cu117
DJLServing 0.20.0 with DeepSpeed 0.7.5, HuggingFace Transformers 4.23.1, and HuggingFace Accelerate 0.13.2	inference	GPU	3.8 (py38)	763104351884.dkr.ecr.us-west-2.amazonaws.com/djl-inference:0.20.0-deepspeed0.7.5-cu116

Extend a prebuilt deep learning container

SageMaker PyTorch image

FROM 763104351884.dkr.ecr.us-east-1.amazonaws.com/pytorch-training:1.5.1-cpu-py36-ubuntu16.04

ENV PATH="/opt/ml/code:\${PATH}"

this environment variable is used by the SageMaker PyTorch container to determine our user code directory.

ENV SAGEMAKER_SUBMIT_DIRECTORY /opt/ml/code

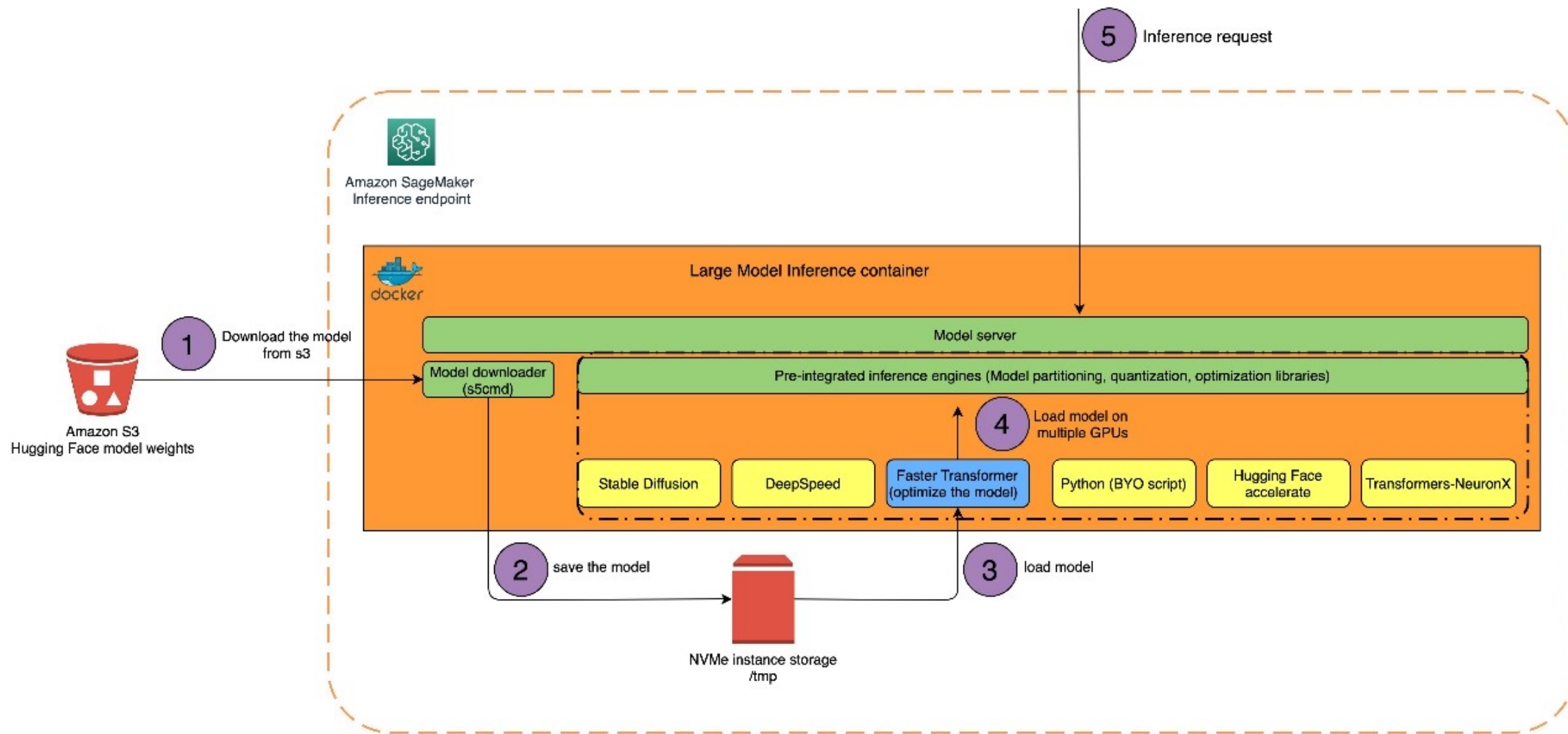
/opt/ml and all subdirectories are utilized by SageMaker, use the /code subdirectory to store your user code.

COPY cifar10.py /opt/ml/code/cifar10.py

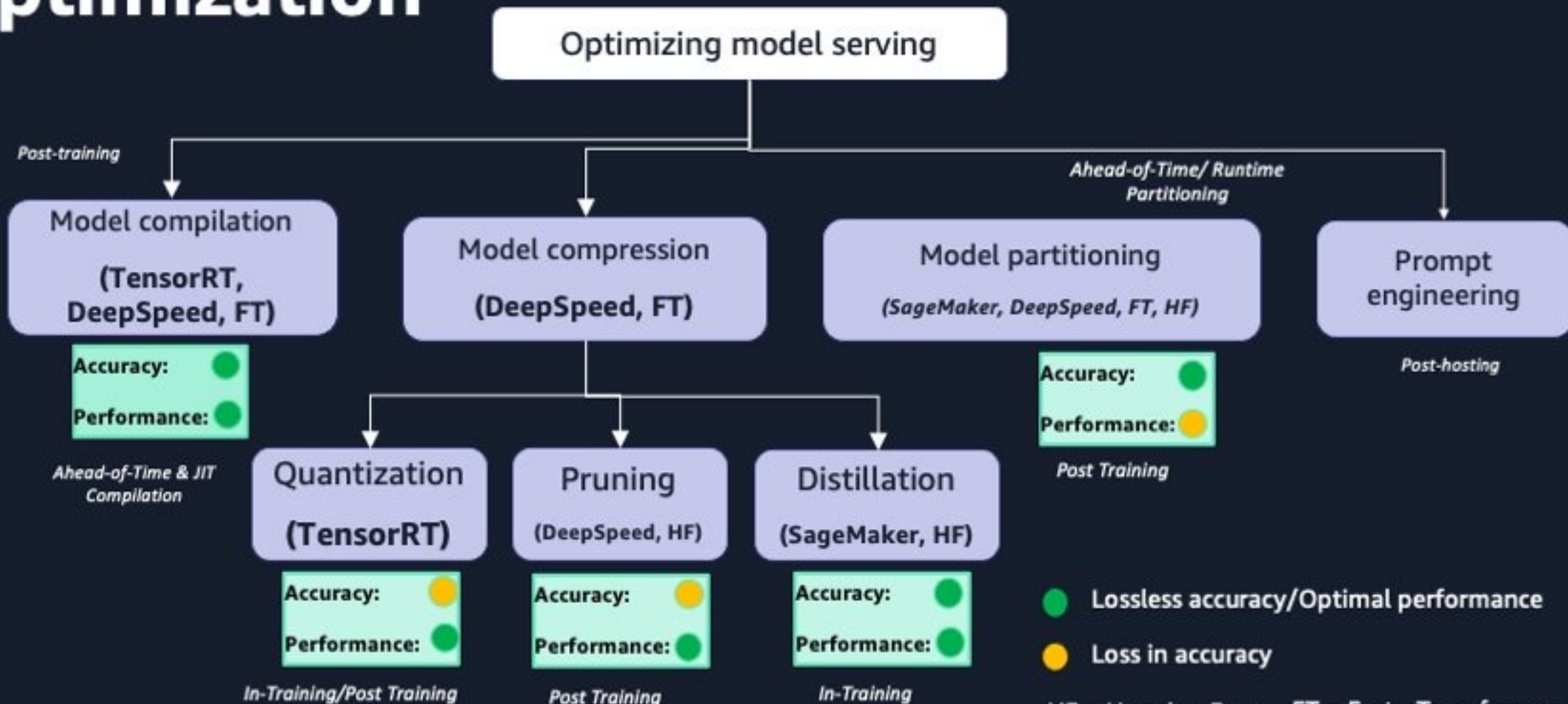
Defines cifar10.py as script entrypoint

ENV SAGEMAKER_PROGRAM cifar10.py

Distributed model hosting on SageMaker



Large generative model inference optimization



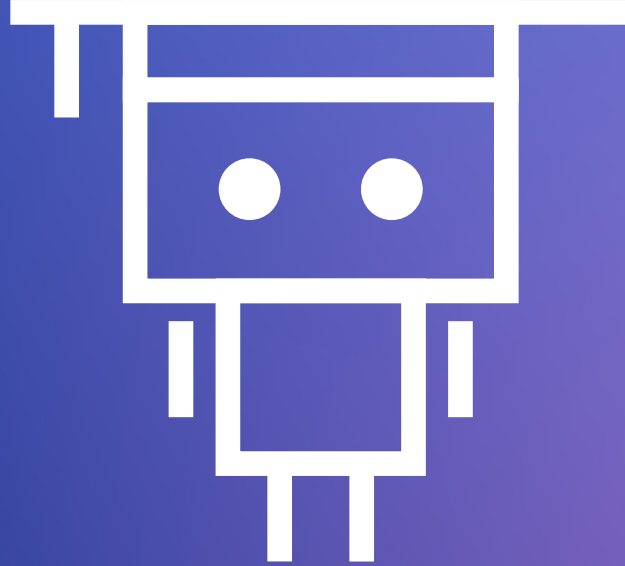


<https://bit.ly/sm-nb-7-hosting>

Hands-on demo



[amazon-sagemaker-examples](#) / [inference](#) / [nlp](#) / [realtime](#) / [llm](#) / [bloom_176b](#) / [djl_deepspeed_deploy.ipynb](#)



Thank you!

Type: Corrections, feedback, or other questions?
Contact us at <https://support.awsamazon.com/#/contacts/aws-academy>.
All trademarks are the property of their owners.