



INNOVATE

DATA AND AI/ML EDITION

22 February 2023

NLP Ops - Operationalize and automate your NLP pipeline with AWS

Hariharan Suresh

Senior Solutions Architect

Amazon Web Services

Natural language processing (NLP)

- Operational challenges
- Landscape on AWS
- Training and inference on AWS
- Ops and governance
- Key takeaways

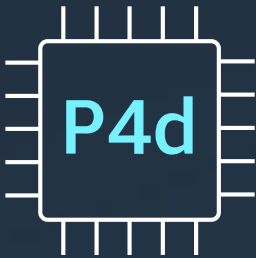
NLP Challenges - Performance vs Cost



Large datasets long training time
create bottlenecks



Training costs are an obstacle
to experimentation and innovation



Infrastructure

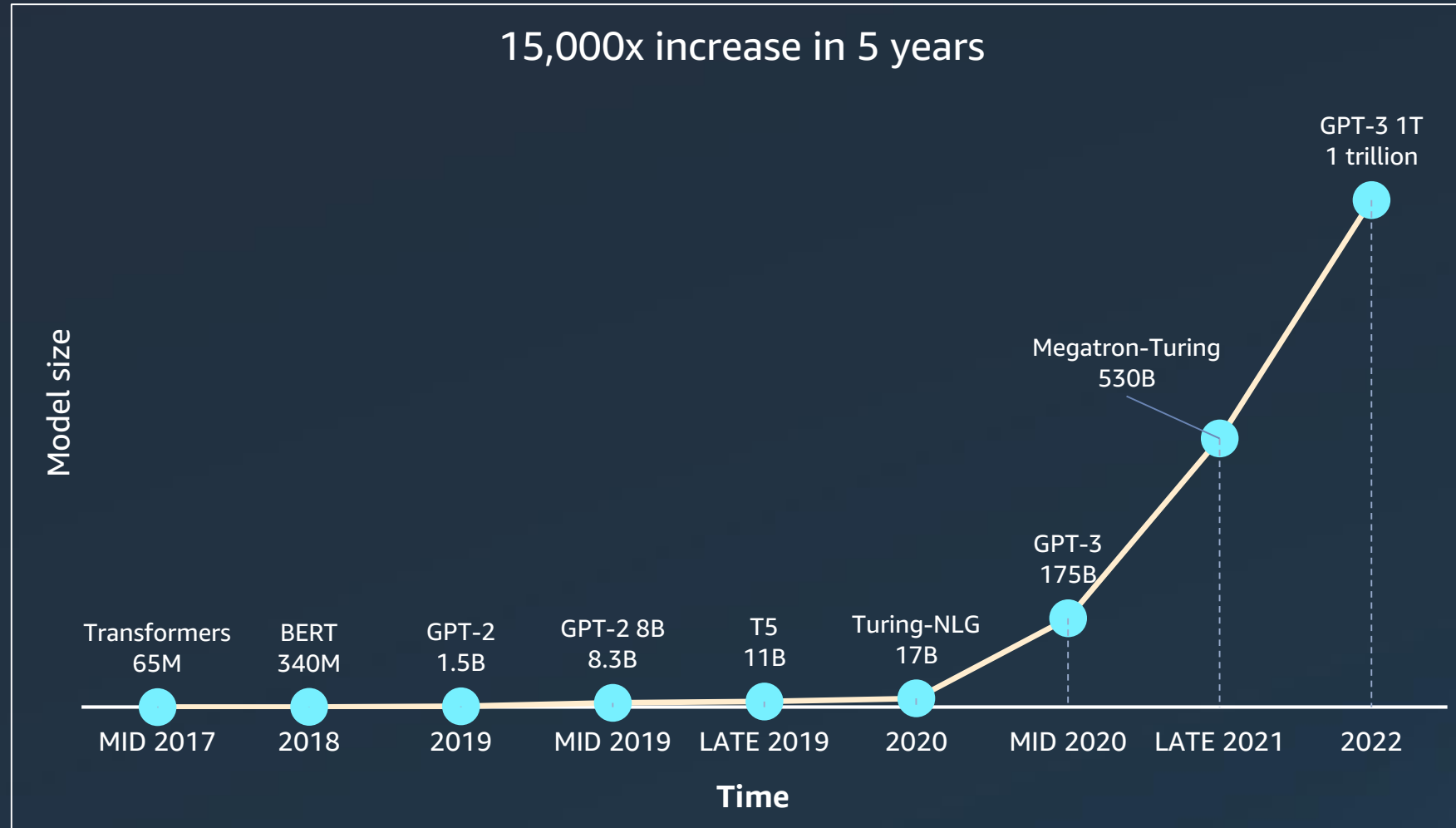


Distributed training



Skills and time

High training time - Model Size vs Productivity



Model	RoBERTa
Dataset	300+ GB
Cluster	64 p3dn.24xl
Training time	Several days

NLP Ops - Pre-processing tasks

Preliminary Tasks

Model Selection

Prepare Domain Data

Curate Features

Data Processing

NLP Ops - Training

Preliminary Tasks	Training	
Model Selection	Pre-Training	
Prepare Domain Data	Experimentation	
Curate Features	HPO/Task Tuning	
Data Processing	Incremental Training	
	Distributed Training	

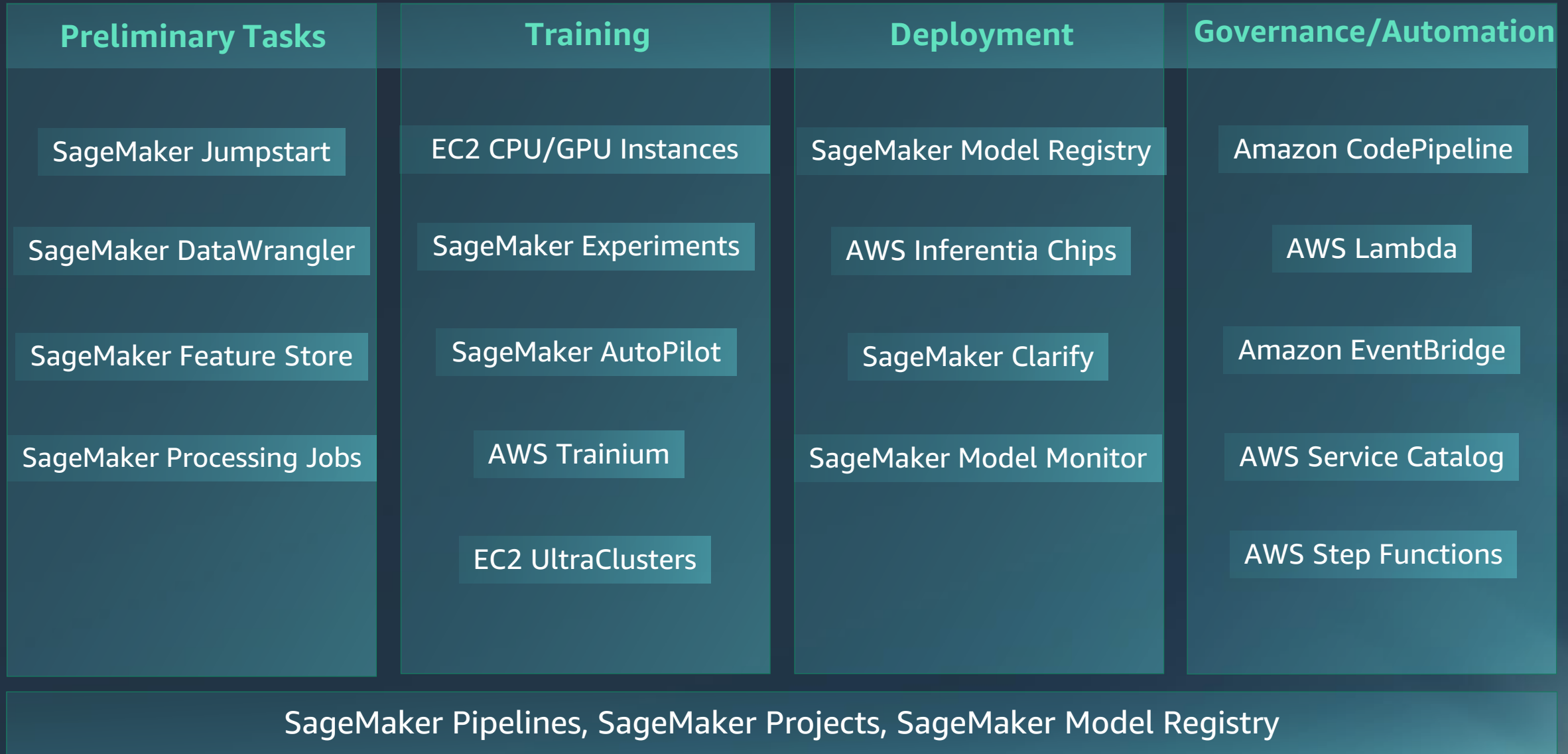
NLP Ops - Deployment

Preliminary Tasks	Training	Deployment	
Model Selection	Pre-Training	Model Versioning	
Prepare Domain Data	Experimentation	Inference Scaling	
Curate Features	HPO/Task Tuning	Model Deployment	
Data Processing	Incremental Training	Metrics/Monitoring	
	Distributed Training		

NLP Ops - Governance & Automation

Preliminary Tasks	Training	Deployment	Governance/Automation
Model Selection	Pre-Training	Model Versioning	DevOps
Prepare Domain Data	Experimentation	Inference Scaling	Developer Tooling
Curate Features	HPO/Task Tuning	Model Deployment	Trigger Re-training
Data Processing	Incremental Training	Metrics/Monitoring	Security & Governance
	Distributed Training		Workflow Orchestration

NLP Ops Lifecycle – Amazon SageMaker



Amazon SageMaker Training Compiler



Amazon SageMaker Training Compiler

The fast and easy way to train large NLP and deep learning models on GPUs



Accelerate deep learning model training

Speed up training by as much as 50%



Minimal code changes required

Enable in minutes without any changes to workflow



Lower training costs

Free to use on SageMaker / savings from shortened training jobs

Enable Amazon SageMaker Training Compiler

```
from sagemaker.pytorch import PyTorch
from sagemaker.pytorch import TrainingCompilerConfig

pytorch_estimator=PyTorch (
    entry_point='script.py',
    instance_count=1,
    instance_type='ml.p3.8xlarge',
    framework_version='1.12.0',
    py_version='py3',
    hyperparameters=hyperparameters,
    compiler_config=TrainingCompilerConfig(),
    distribution ={'pytorchxla' : { 'enabled': True }})

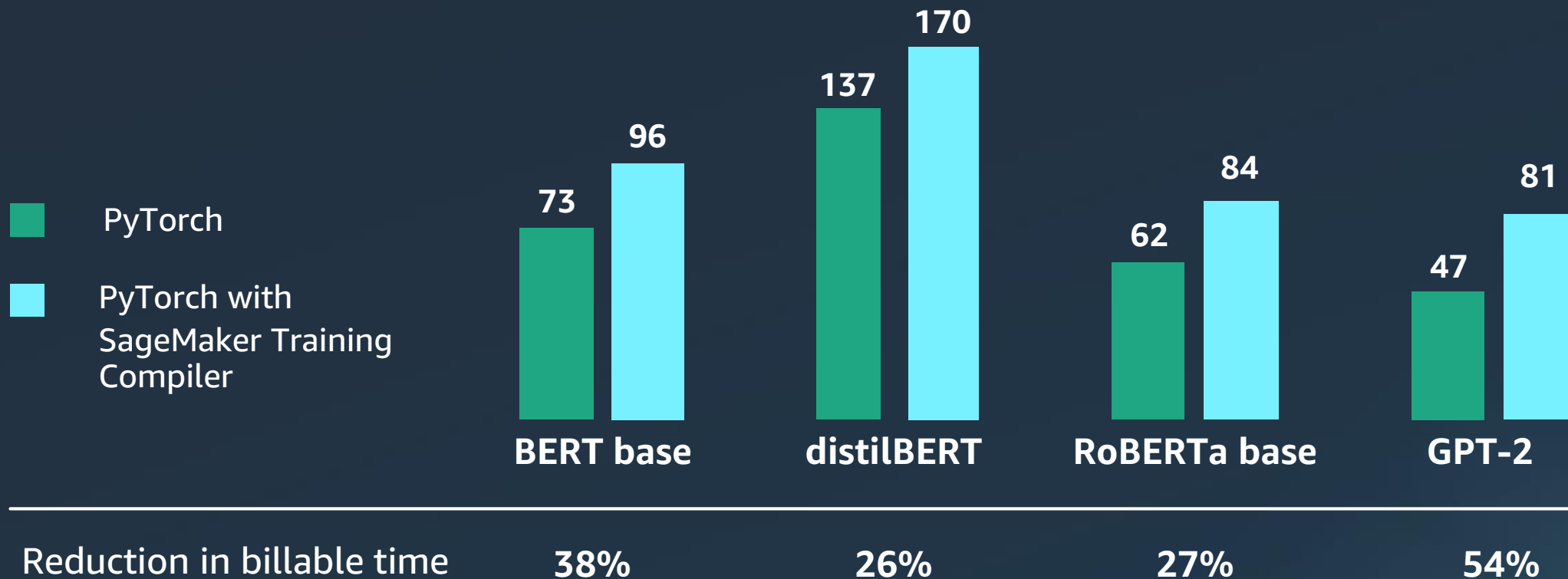
pytorch_estimator.fit()
```

What can you expect ?

- Add 2-line code to enable
- Supports single GPU and distributed training
- 100+ supported models
- Best performance on AWS for smaller clusters

Amazon SageMaker Training Compiler - decreases training time

Training sample throughput¹ (samples/second)

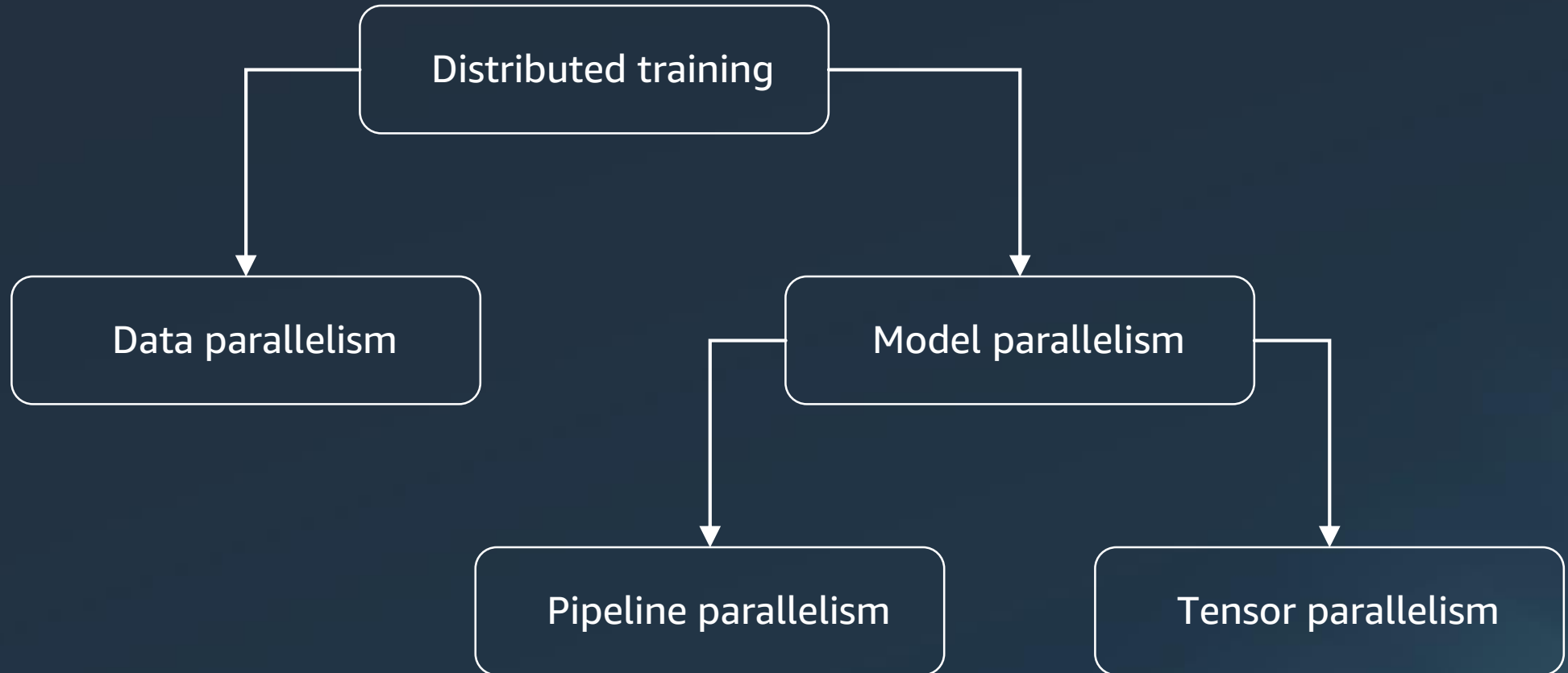


¹ Test parameters: ml.p3.2xlarge, PyTorch with Hugging Face Trainer API, 25 epochs, sequence length of 512
Baseline used the Hugging Face AWS Deep Learning Container from Amazon ECR

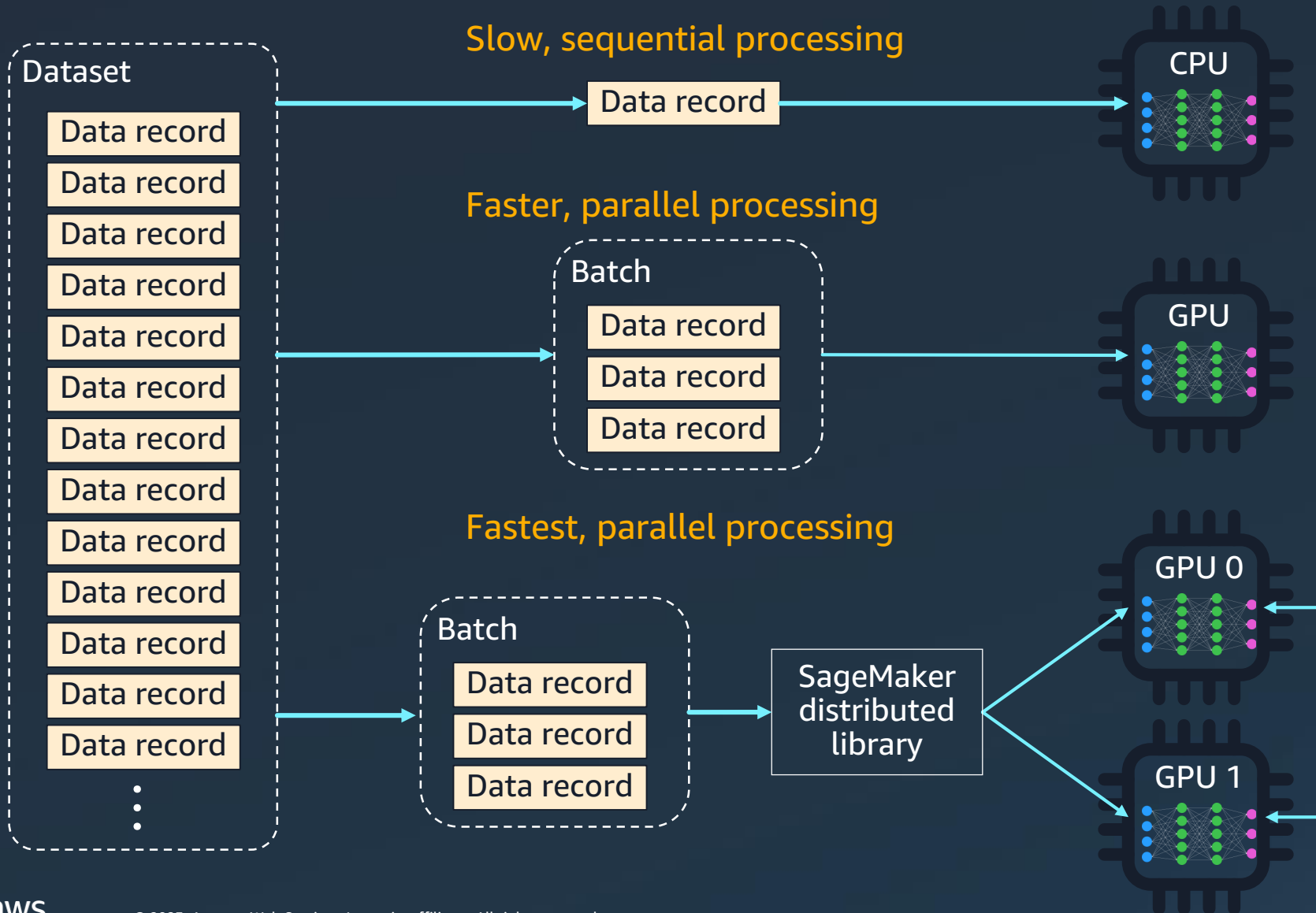
Amazon SageMaker - Distributed Training



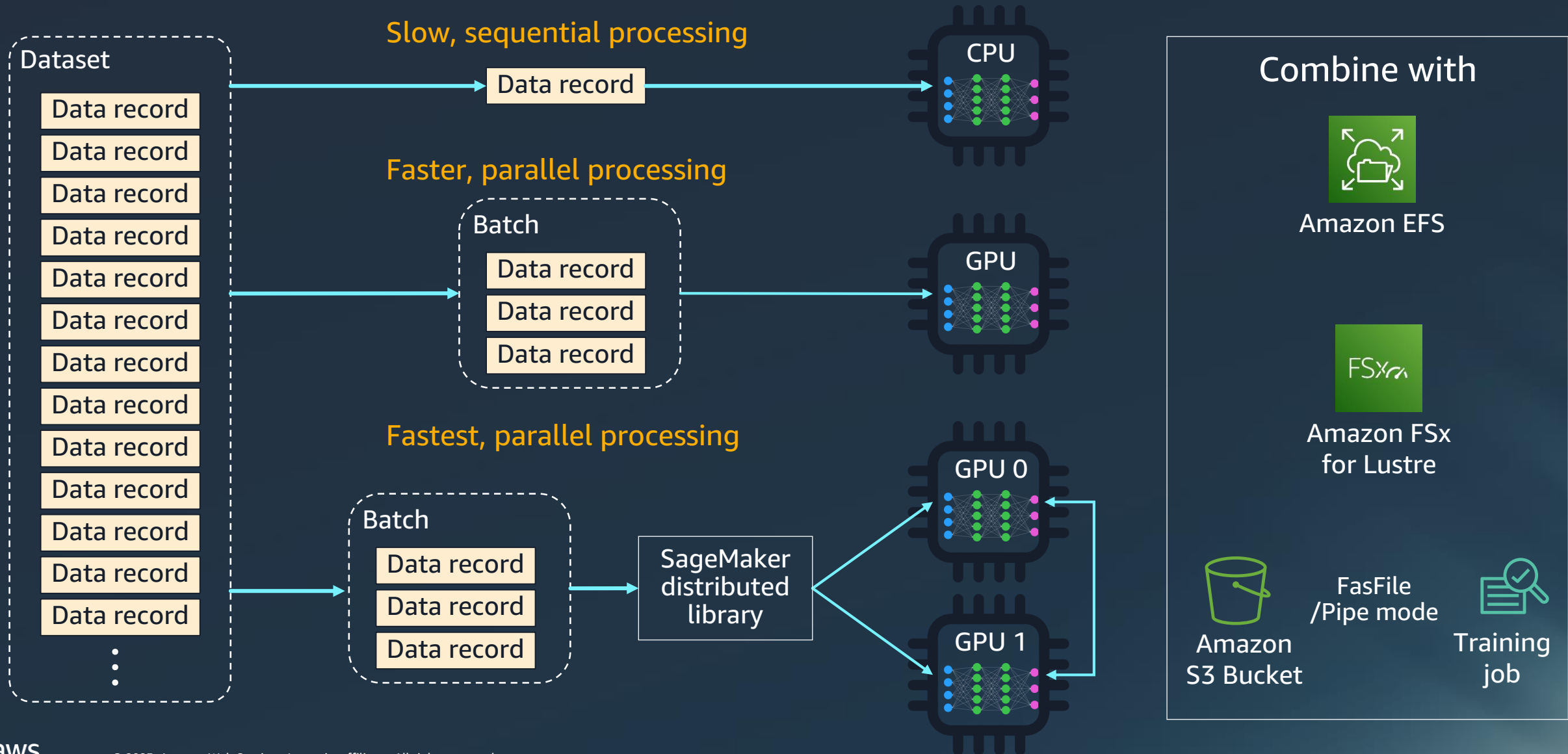
Distributed Training



Data Parallelism



Data Parallelism – Storage Acceleration



Data Parallelism – Takeaways



Fully managed training on replicas on workers



Supports popular APIs like Horovod and DistributedDataParallel



Automatically synchronizes workers across multiple GPUs

- Max Model size = Single GPU Memory
- Larger models -> OOM errors
- Model replicated across all GPUs
- Wasteful when model is huge

 PyTorch **DDP**



Distributed Data Platform

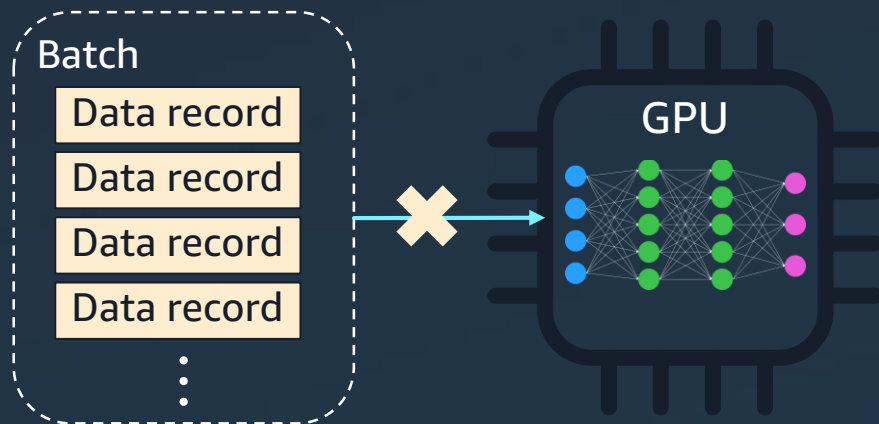
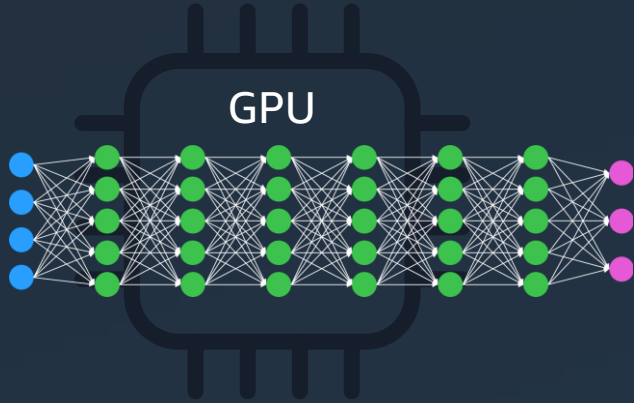
 TensorFlow



DeepSpeed

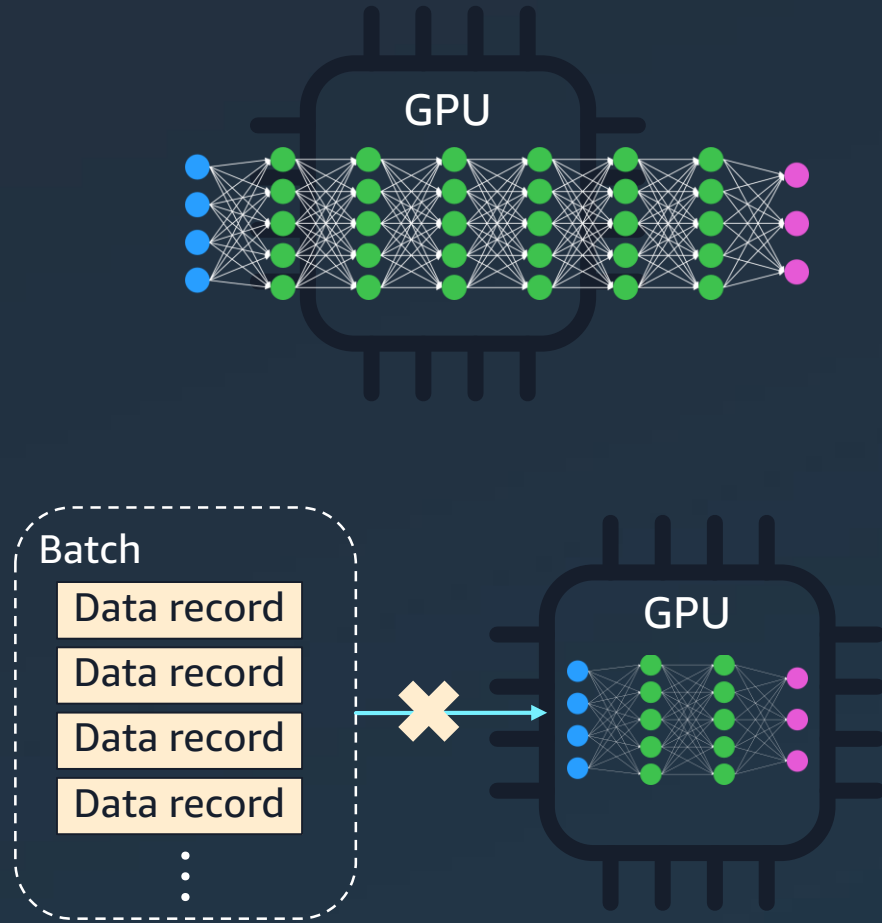
Data Parallelism – Drawbacks

Out-of-memory error



Data Parallelism – Mitigation Options

Out-of-memory error



Try first

Reduce network size



Reduce batch size



Reduce data size



Model Parallelism - Takeaways



Analyzes variables and graph structure



Automatically partitions the model and sends subgraphs to devices



Managed distributed training on pipelined microbatches

- Single model replica partitioned across GPUs
- Combine memory of all GPUs
- No model replication -> saves additional memory
- Servers communicate during forward and backward pass

 PyTorch FSDP

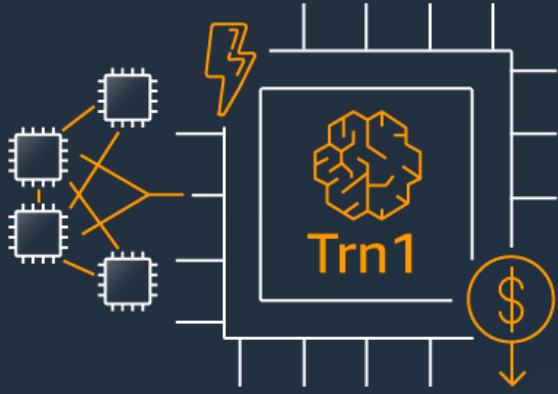


Distributed Data Platform



DeepSpeed

Amazon SageMaker Training & Inference - Enhancers

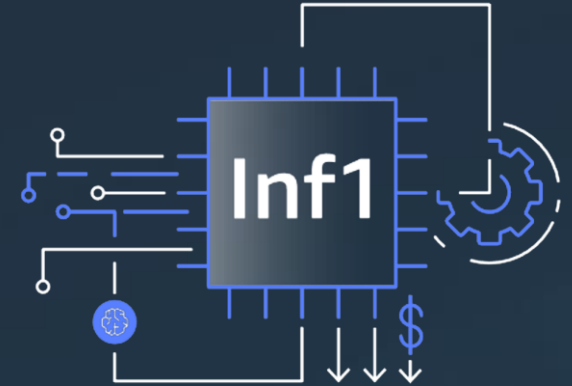


- Purpose built for high-performance ML training
- 60% higher accelerator memory with 800 Gbps network bandwidth
- Integrated with major frameworks



 PyTorch **mxnet**  TensorFlow

- SDK - Compiler + Runtime + Profiling tools
- Enables graph & loop optimizations
- Scheduling and allocation operations
- Integrated with major frameworks




- Purpose built for inference
- Inf2 - 4x higher throughput
- Inf2 support scale-out distributed inference

AWS Neuron SDK Samples

```
!pip config set global.extra-index-url https://pip.repos.neuron.amazonaws.com
# Install Neuron PyTorch
!pip install torch-neuron neuron-cc[tensorflow]
```

PyTorch Installation

Neuron Model Compilation




```
import torch
import torch.neuron

# [...]
# model and sample input are created before
model_neuron = torch.neuron.trace(model, sample_inputs)


model_neuron.save('model_neuron.pt')

# load model to Neuron Cores on Inf1 instance
loaded_model = torch.jit.load('model_neuron.pt')
```

Neuron Dynamic Batching & DataParallel API



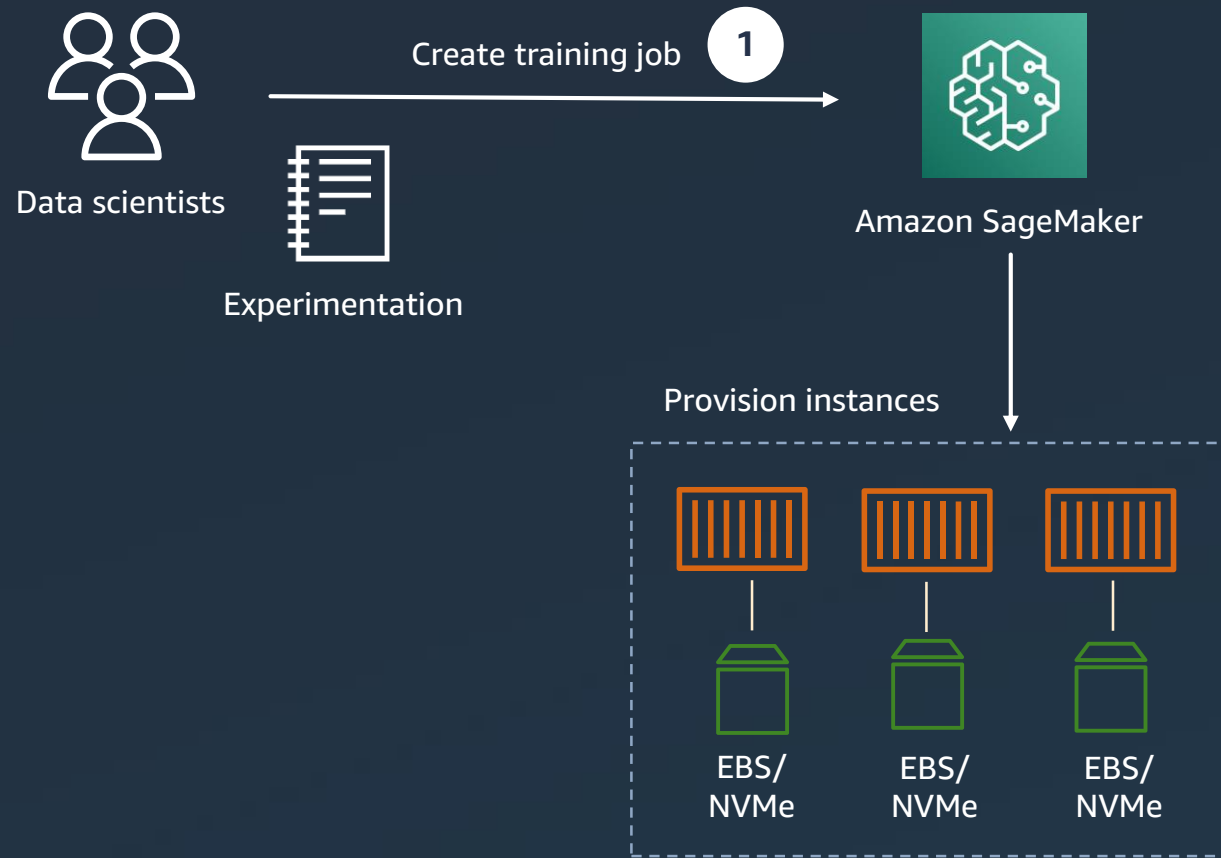
```
model_neuron = torch.neuron.trace(model,
                                   neuron_inputs,
                                   dynamic_batch_size=True)
```



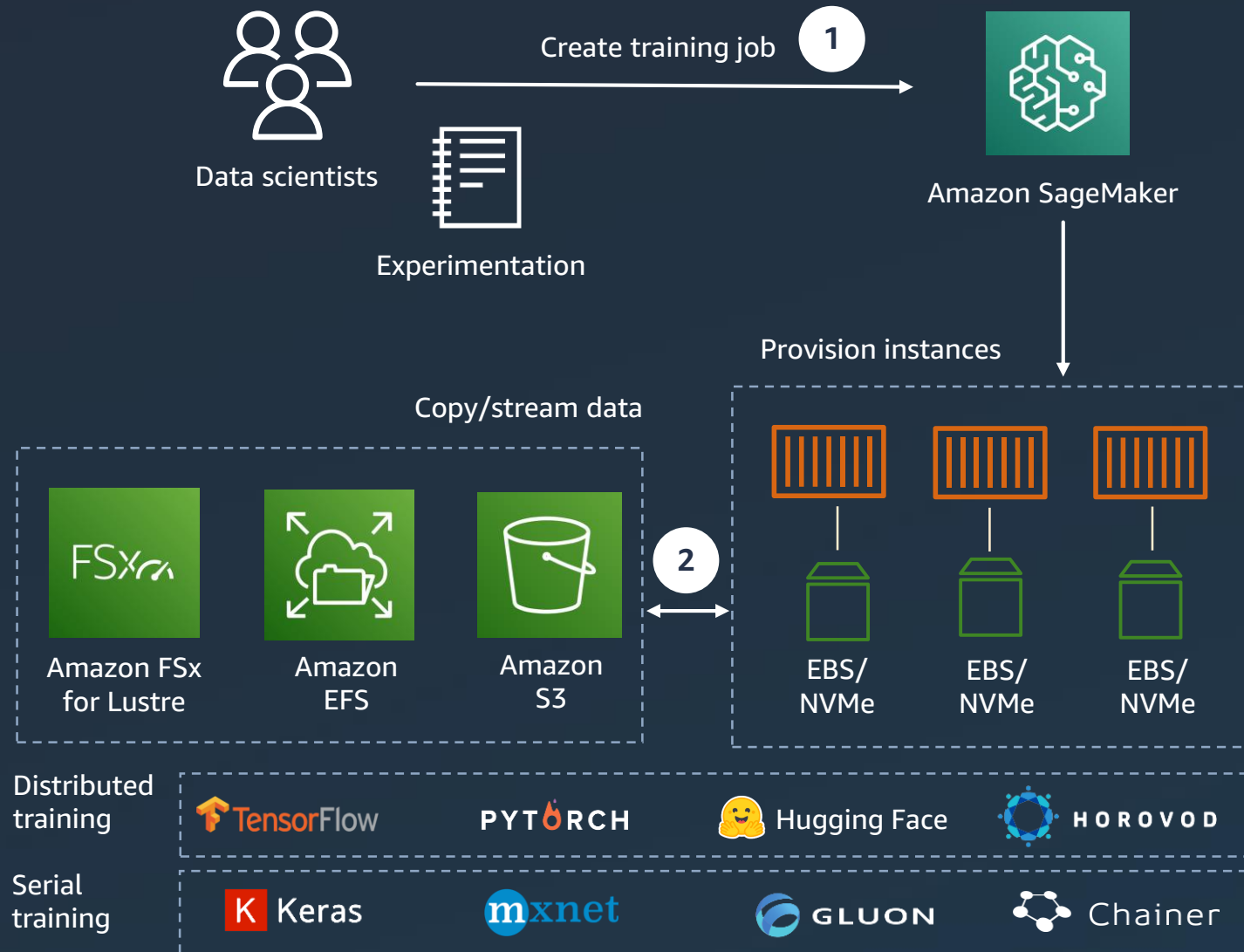
```
model_neuron = torch.neuron.trace(model, image)
# Create the DataParallel module
model_parallel = torch.neuron.DataParallel(model_neuron)
```


Distributed Training - Architecture

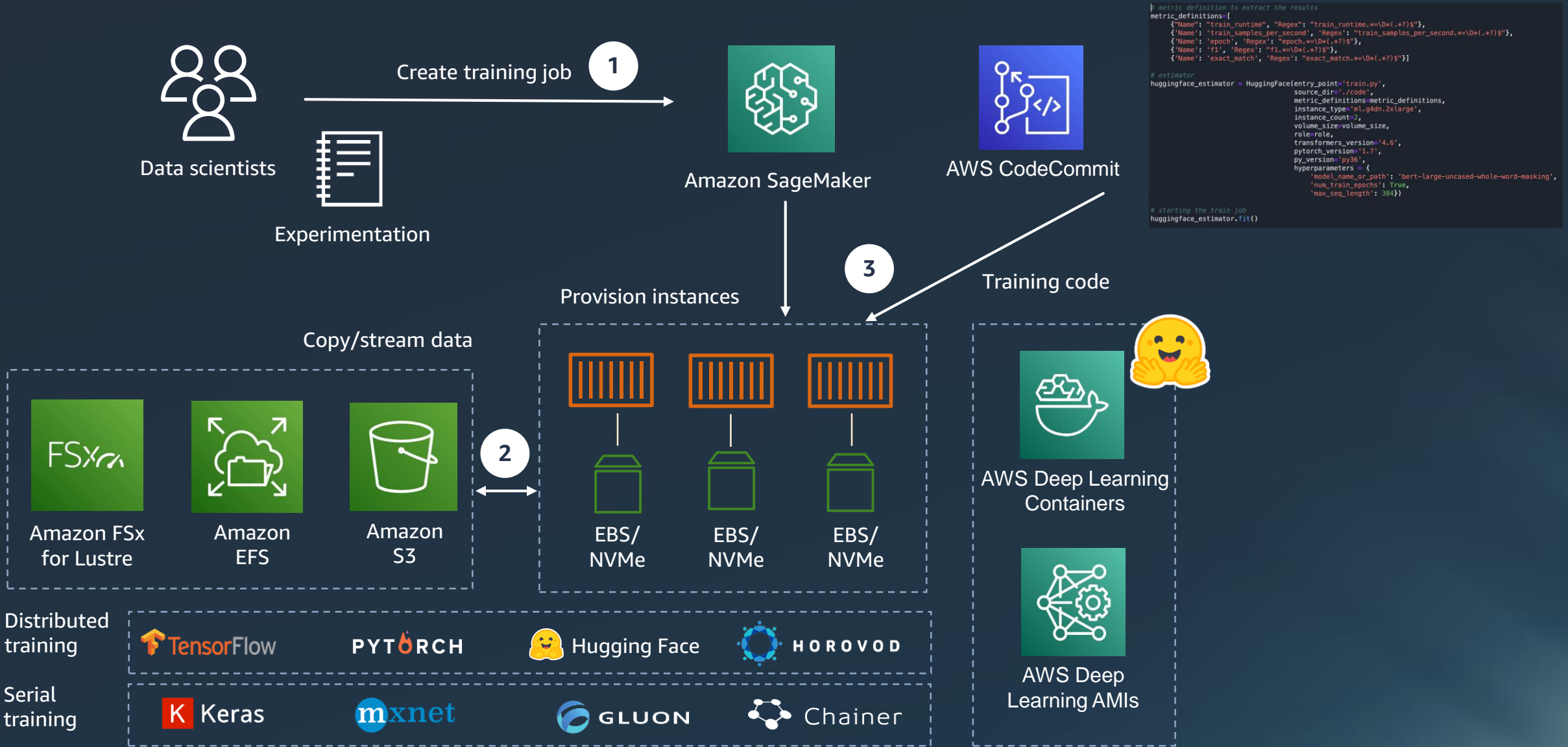
Amazon SageMaker - NLP Training



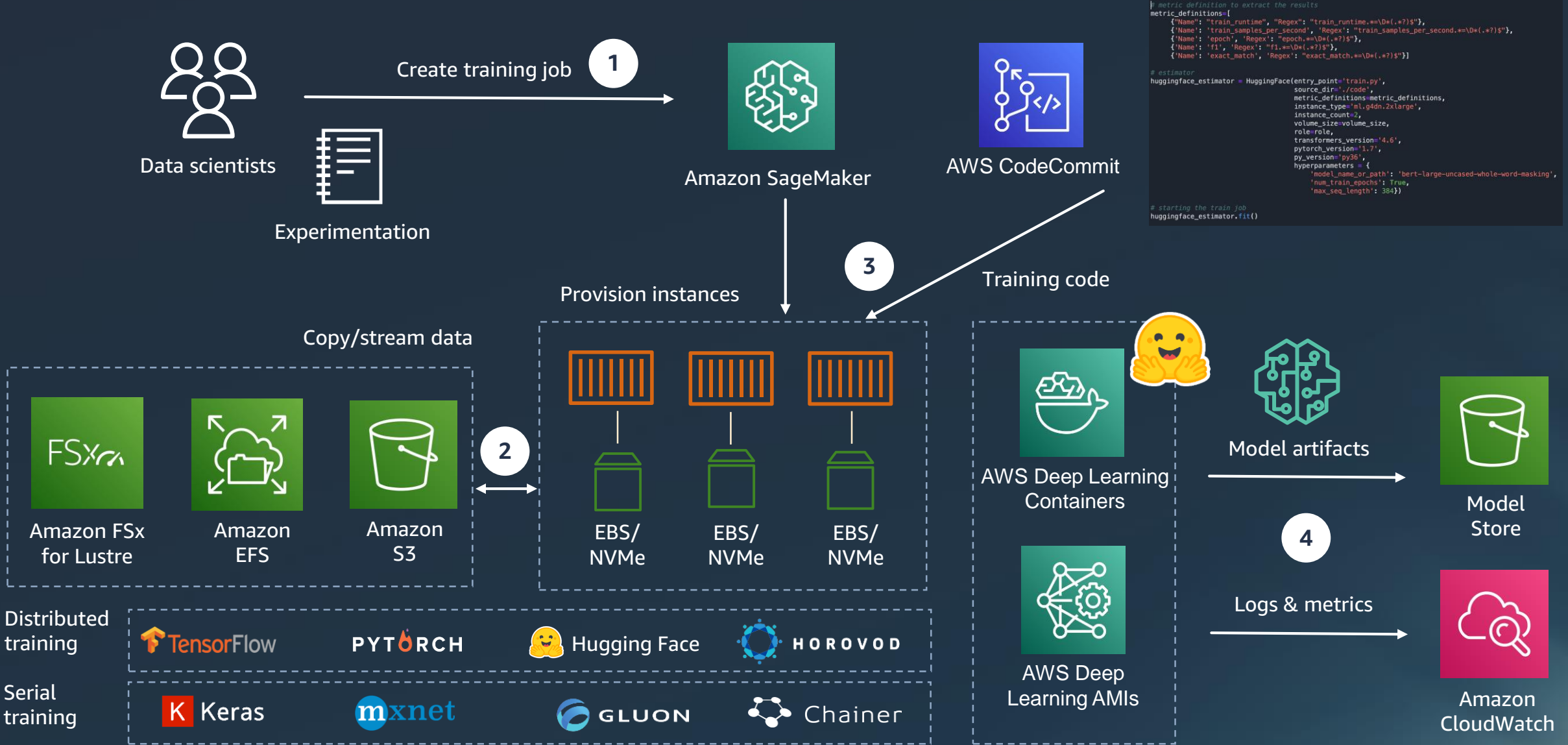
Amazon SageMaker - NLP Training



Amazon SageMaker - NLP Training



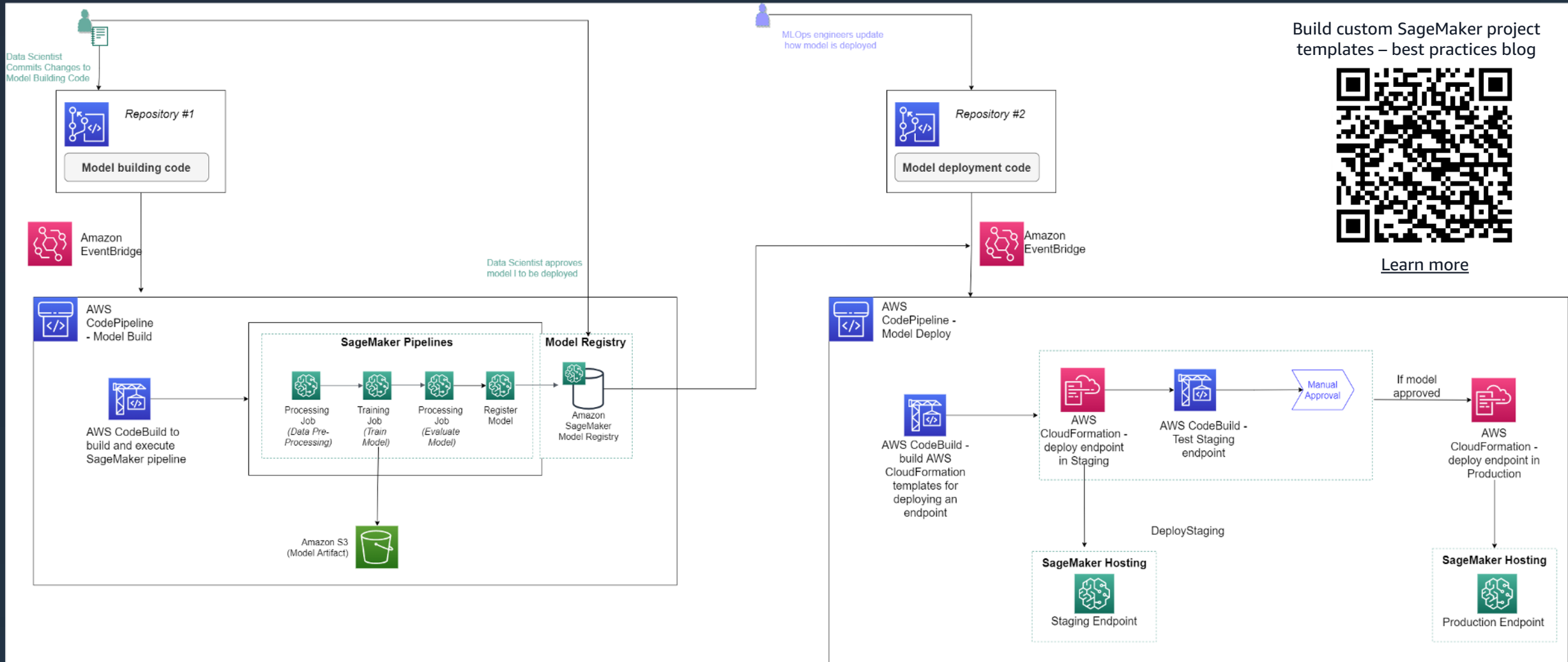
Amazon SageMaker - NLP Training



Demo – Amazon SageMaker NLP model selection & experimentation with transfer learning

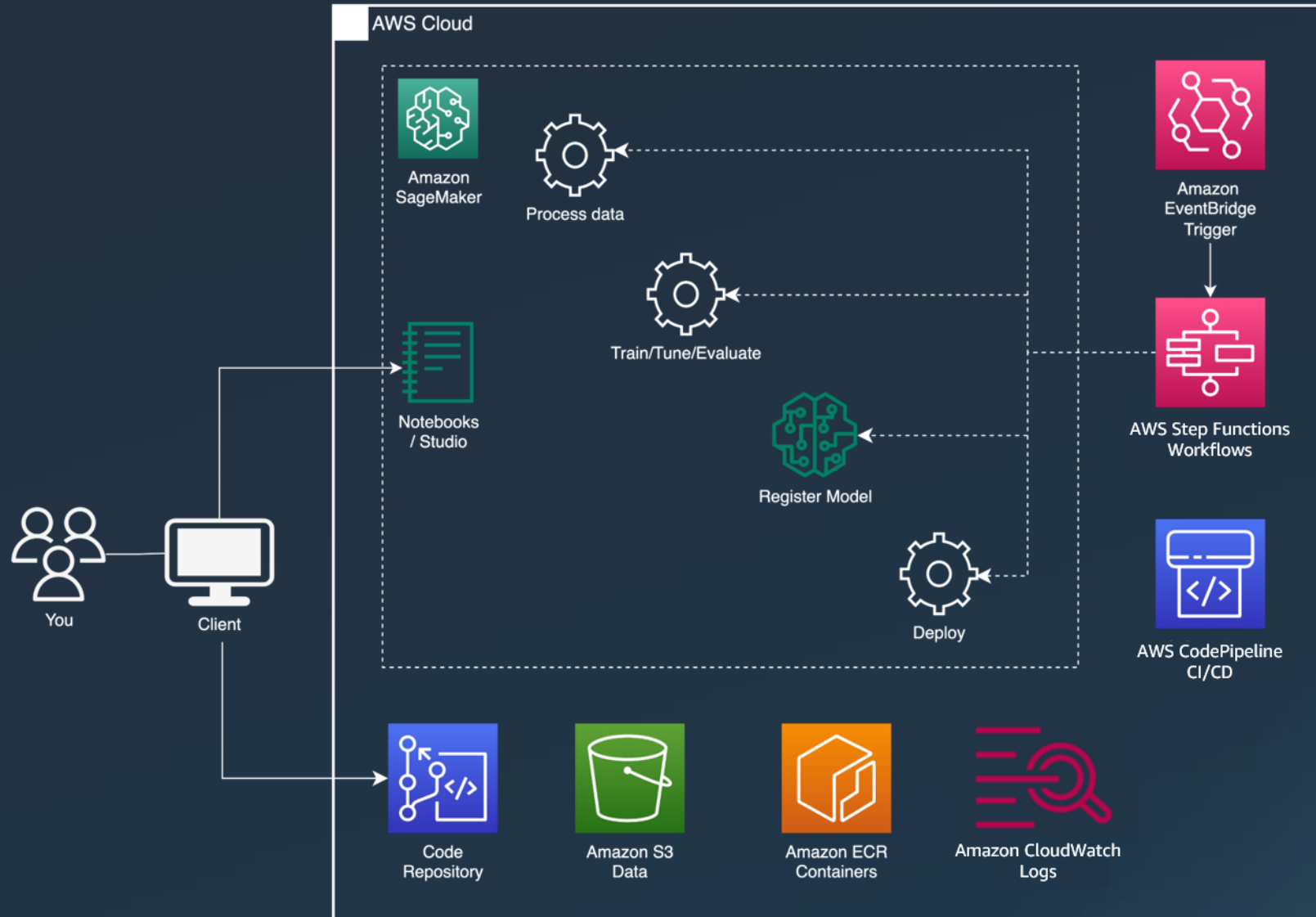
NLP Ops and Governance

NLP Ops – End to End ML lifecycle

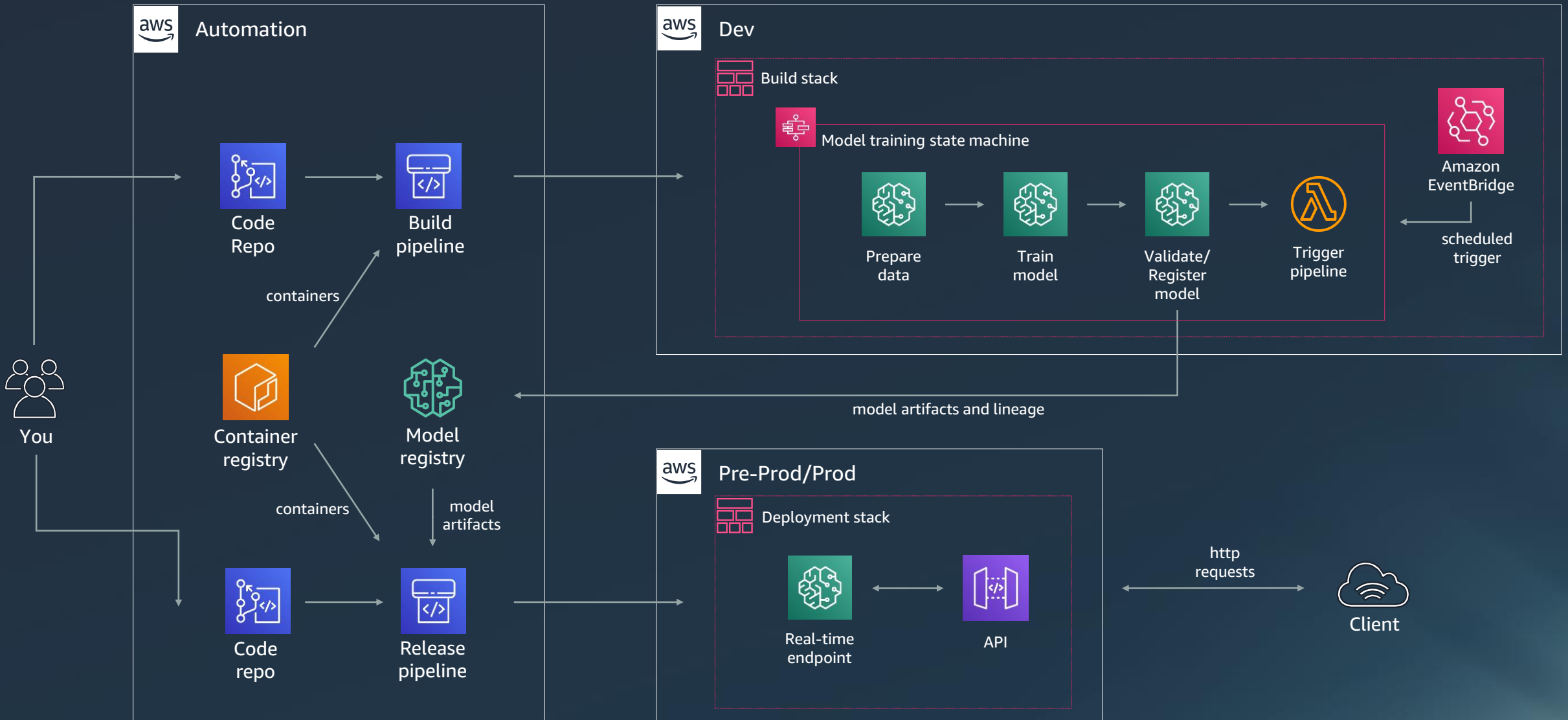


Demo – Amazon SageMaker Pipeline

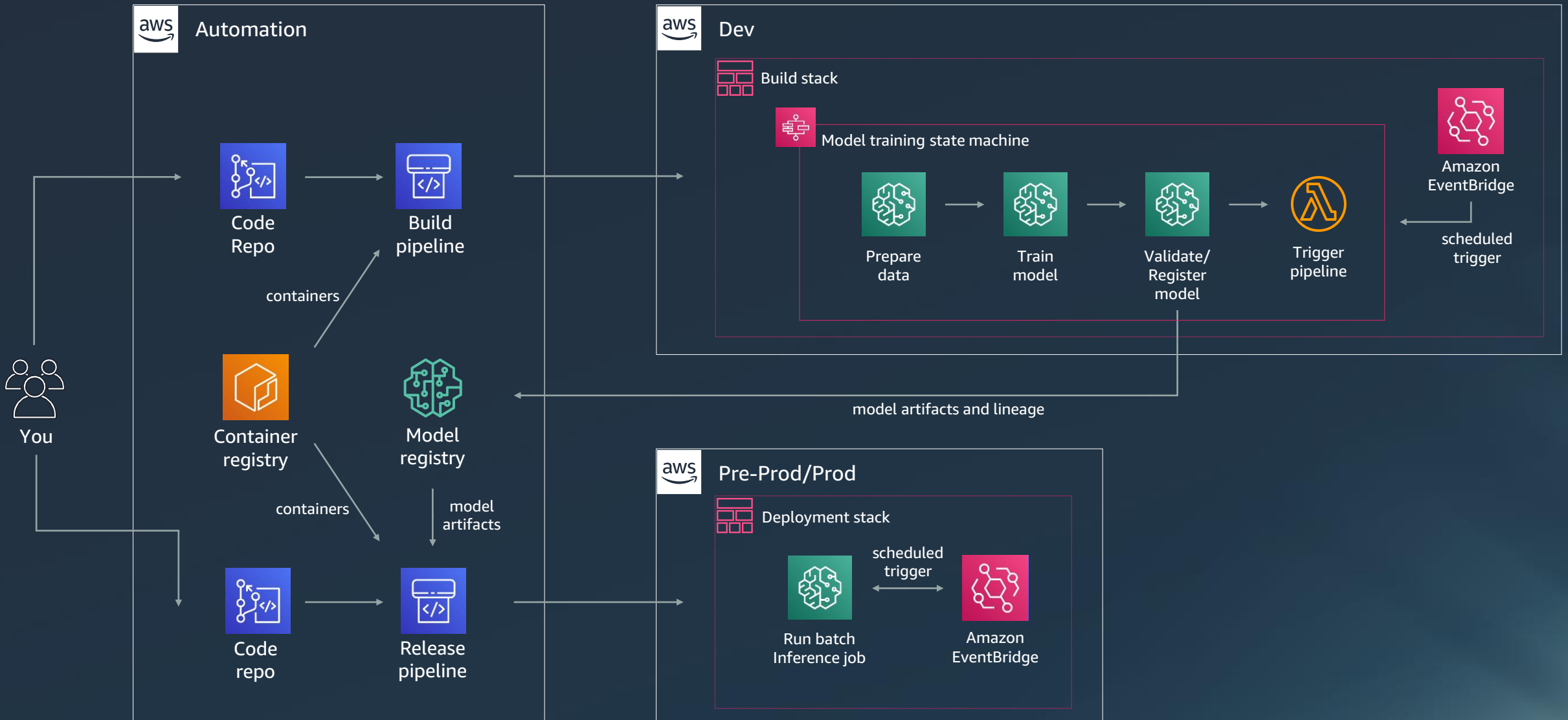
NLP – Execute workflows and track ML operations



Real-time inference deployment pattern



Batch inference deployment pattern



NLP Operations Lifecycle – Amazon SageMaker Recap

Preliminary Tasks	Training	Deployment	Governance/Automation
SageMaker Jumpstart	EC2 CPU/GPU Instances	SageMaker Model Registry	Amazon CodePipeline
SageMaker DataWrangler	SageMaker Experiments	AWS Inferentia Chips	AWS Lambda
SageMaker Feature Store	SageMaker AutoPilot	SageMaker Clarify	Amazon EventBridge
SageMaker Processing Jobs	AWS Trainium	SageMaker Model Monitor	AWS Service Catalog
	EC2 UltraClusters		AWS Step Functions
SageMaker Pipelines, SageMaker Projects, SageMaker Model Registry			

Get started

Amazon SageMaker
Training Compiler



[Link](#)

Amazon SageMaker
Training Compiler Samples



[Link](#)

Distributed Training
Options



[Link](#)

Amazon SageMaker
Pipelines Custom



[Link](#)

Amazon SageMaker
Project Templates



[Link](#)

Amazon Trainium



[Link](#)

Amazon SageMaker
Custom Project Templates



[Link](#)

AWS Neuron



[Link](#)

AWS Inferentia



[Link](#)



Visit the Data & AI/ML resource hub

Dive deeper into these resources, get inspired and learn how you can use AI and machine learning to accelerate your business outcomes.

- 6 steps to machine learning success e-book
- 7 leading machine learning use cases e-book
- Machine learning at scale e-book
- Achieving transformative business results with machine learning e-book
- Tackling our world's hardest problems with machine learning e-book
- Accelerating machine learning innovation through security e-book
- ... and more!



<https://bitly.co/FqdC>

Visit resource hub



AWS Training and Certification

Access the AI & ML learning plan courses built by AWS experts on AWS Skill Builder

- Get started with digital self-paced, on-demand training and ramp-up guides to help you grow your technical skills
- Learn how to apply machine learning, artificial intelligence, and deep learning to unlock new insights and value in your role
- Take the steps today, towards validating your expertise with an AWS Certified Machine Learning – Specialty Certification



<https://bit.ly/3FnxDH7>

Learn your way [explore.skillbuilder.aws](https://skillbuilder.aws) »



Thank you for attending AWS Innovate – Data & AI/ML Edition

We hope you found it interesting! A kind reminder to **complete the survey**.
Let us know what you thought of today's event and how we can improve the event experience for you in the future.



aws-apj-marketing@amazon.com



twitter.com/AWSCloud



facebook.com/AmazonWebServices



youtube.com/user/AmazonWebServices



slideshare.net/AmazonWebServices



twitch.tv/aws

Thank you!

