KubeCon | CloudNativeCon

North America 2024

KubeCon | CloudNativeCon

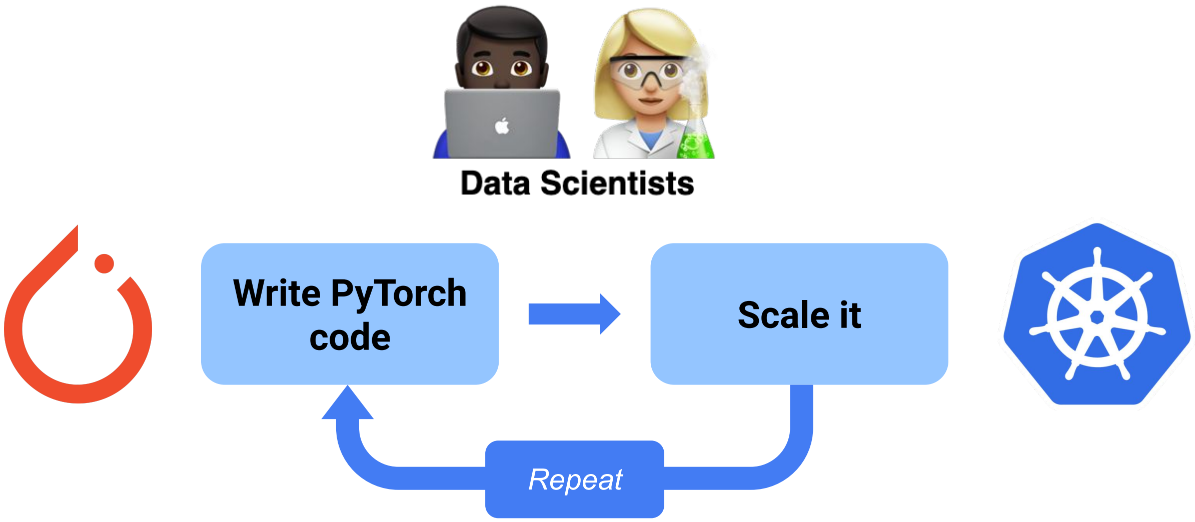
North America 2024

Democratizing AI Model Training on Kubernetes with Kubeflow TrainJob and JobSet

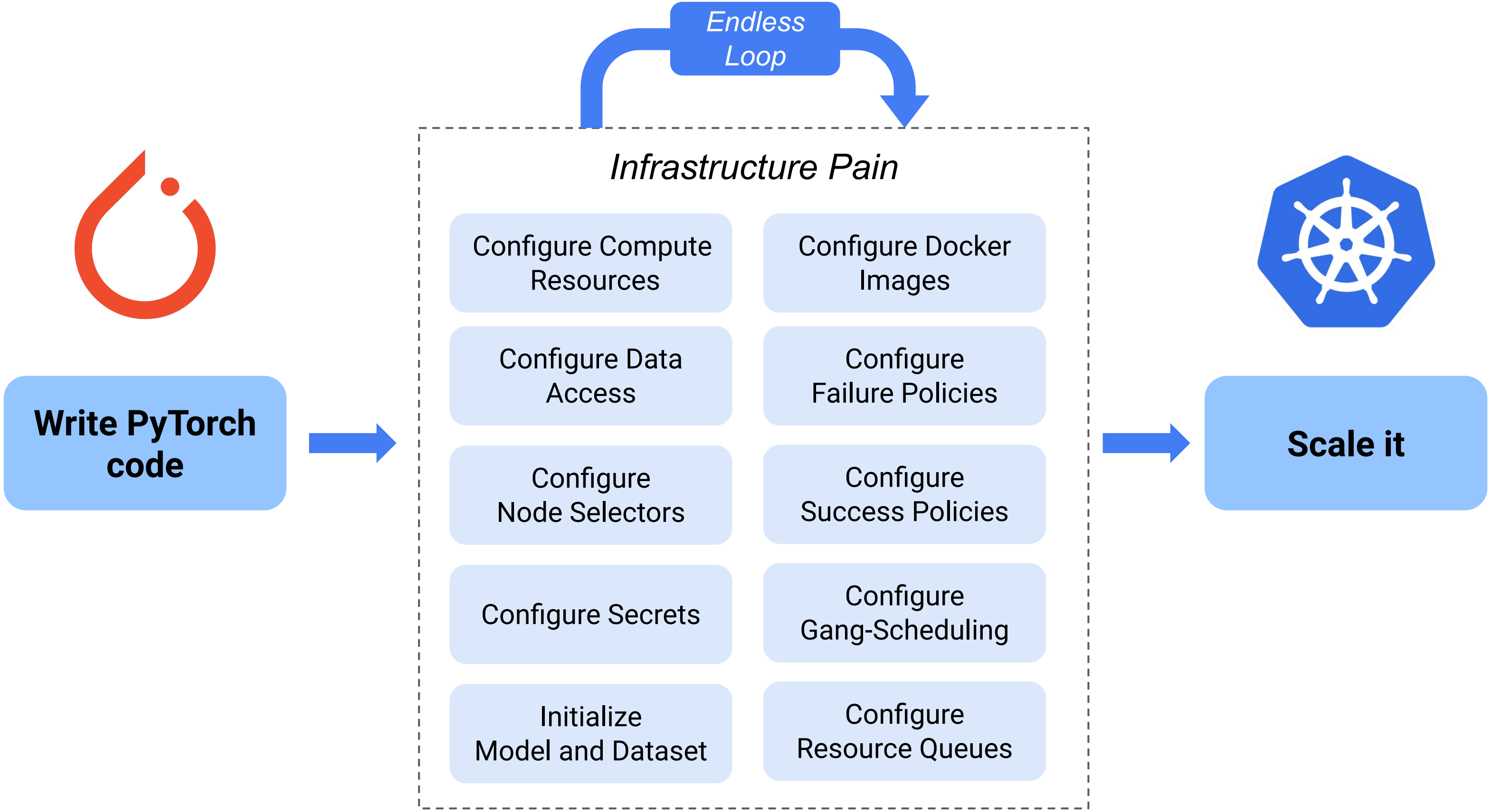
Andrey Velichkevich (@andreyvelich) - Apple

Yuki Iwai (@tenzen-y) - Cyber Agent, Inc.

Train AI Models - Expectations



Train AI Models - Reality



Write PyTorch code

Endless Loop

Infrastructure Pain

- Configure Compute Resources
- Configure Docker Images
- Configure Data Access
- Configure Failure Policies
- Configure Node Selectors
- Configure Success Policies
- Configure Secrets
- Configure Gang-Scheduling
- Initialize Model and Dataset
- Configure Resource Queues

Scale it

Challenges for Model Training on Kubernetes

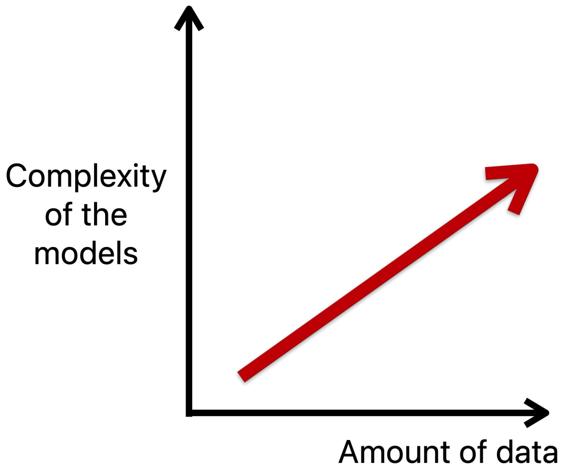
Models are becoming more complex

Large datasets need to be distributed across training nodes

Efficent management of compute resources is essential

Diversity of ML frameworks is increasing

New distributed technologies need to be adopted



What Users Want

“I just want to scale my PyTorch code”

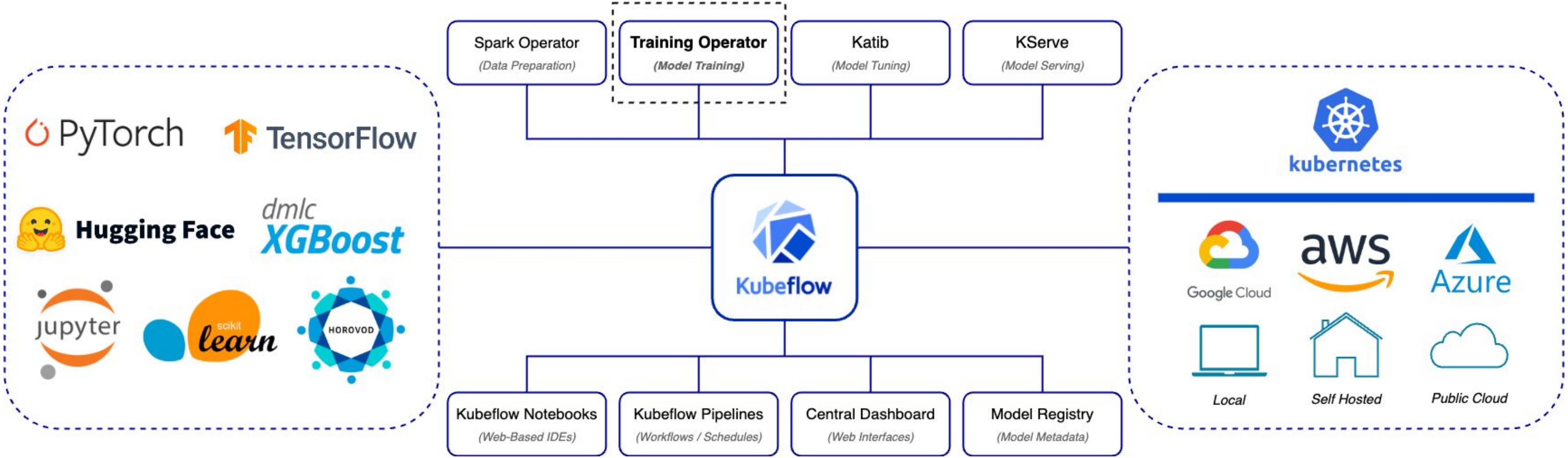
Data Scientists

Simplicity

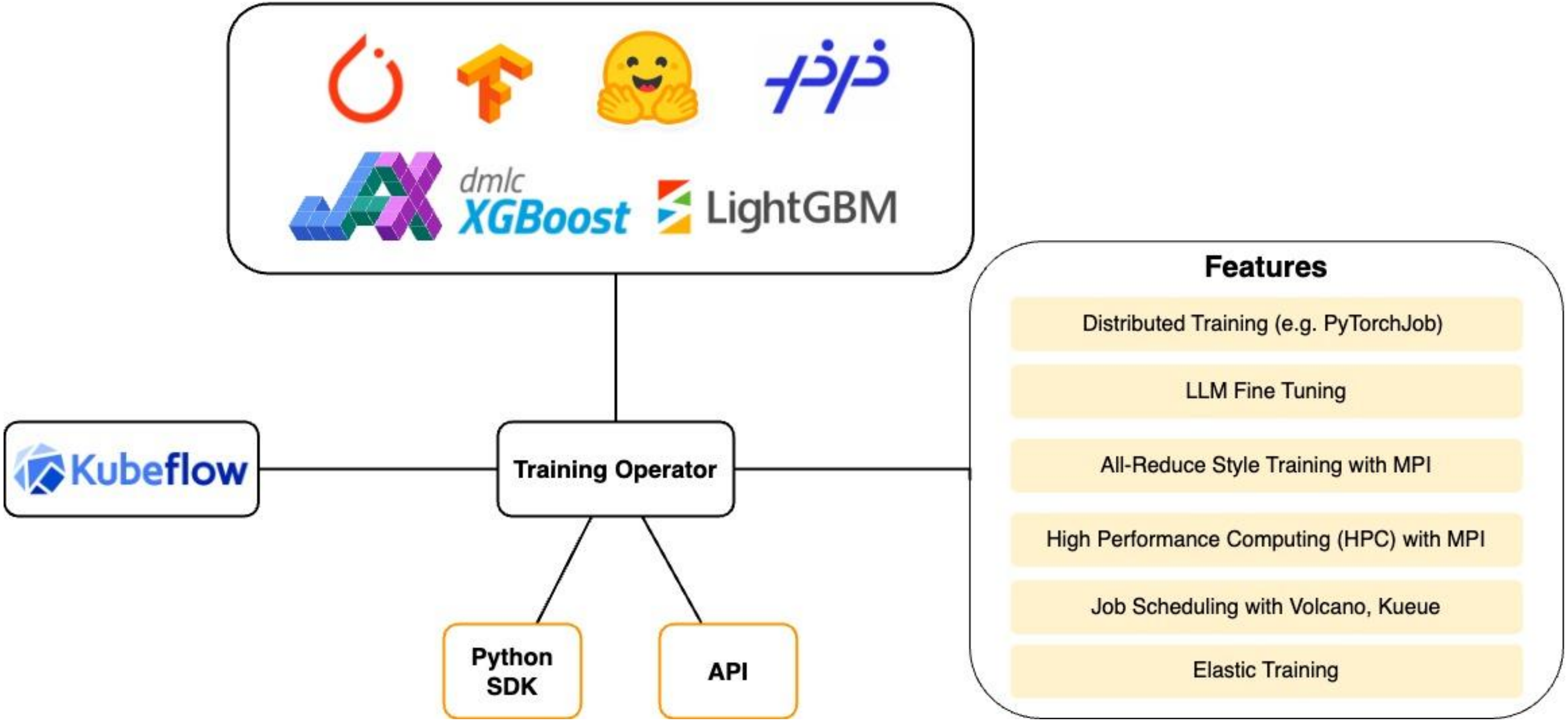
Flexibility

Scalability & Cost Saving

What is Kubeflow?



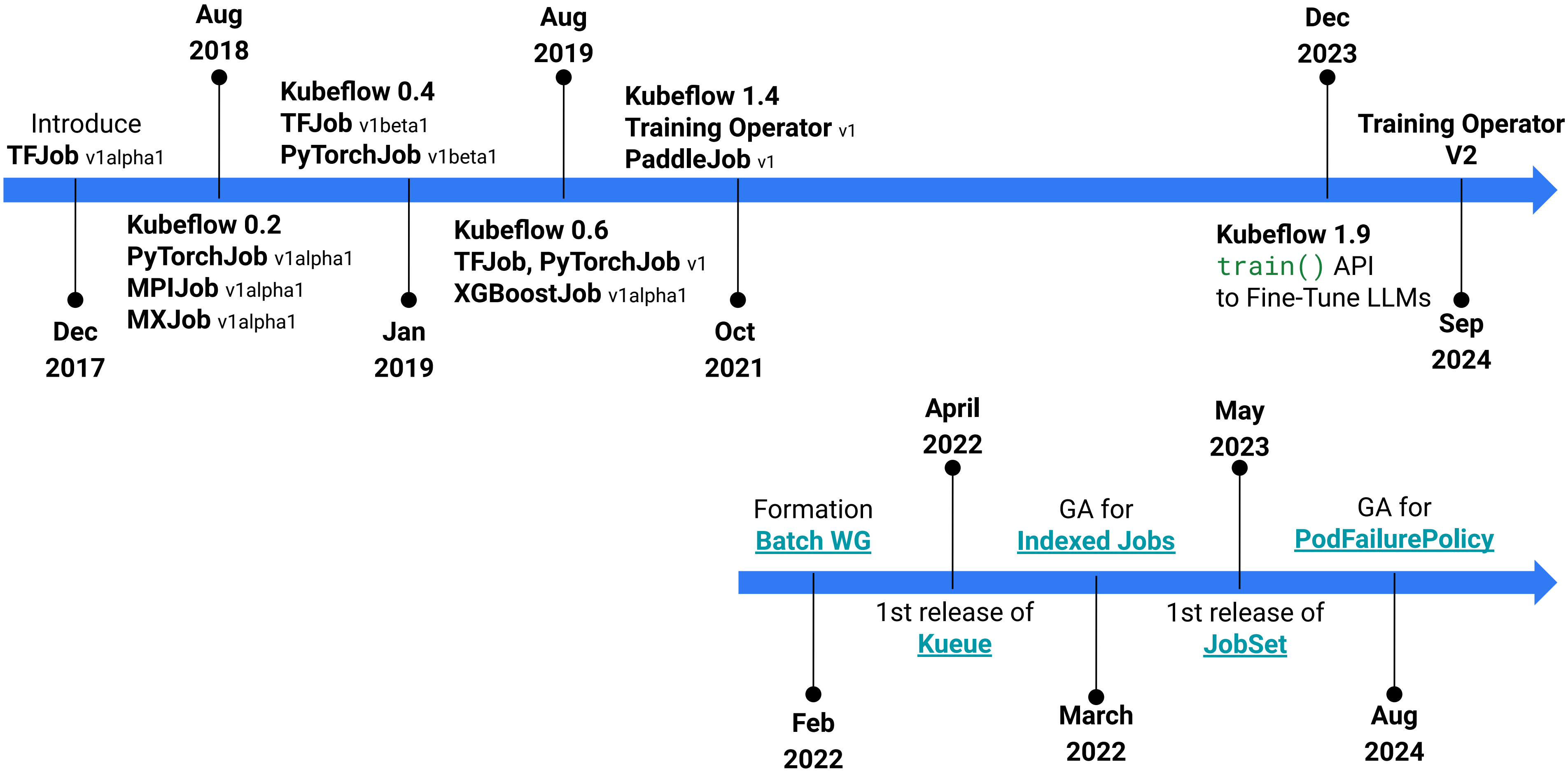
What is Training Operator?



Features

- Distributed Training (e.g. PyTorchJob)
- LLM Fine Tuning
- All-Reduce Style Training with MPI
- High Performance Computing (HPC) with MPI
- Job Scheduling with Volcano, Kueue
- Elastic Training

History of Training Operator and Batch WG



Dec 2017
Introduce
TFJob v1alpha1

Aug 2018
Kubeflow 0.2
PyTorchJob v1alpha1
MPIJob v1alpha1
MXJob v1alpha1

Jan 2019
Kubeflow 0.4
TFJob v1beta1
PyTorchJob v1beta1

Aug 2019
Kubeflow 0.6
TFJob, PyTorchJob v1
XGBoostJob v1alpha1

Oct 2021
Kubeflow 1.4
Training Operator v1
PaddleJob v1

Dec 2023
Kubeflow 1.9
`train()` API
to Fine-Tune LLMs

Sep 2024
Training Operator V2

Feb 2022
Formation
[Batch WG]

April 2022
1st release of
[Kueue]

March 2022
GA for
[Indexed Jobs]

May 2023
1st release of
[JobSet]

Aug 2024
GA for
[PodFailurePolicy]

Announcing

Kubeflow Training V2

Kubeflow Training V2 Goals

Simple to use and scale

Python is the main user interface

Enables quick fine-tuning of LLMs

Provides robust support for the ML ecosystem

Streamline dataset and pre-trained model initialization

Consolidates efforts between Kubernetes and Kubeflow communities

Demo

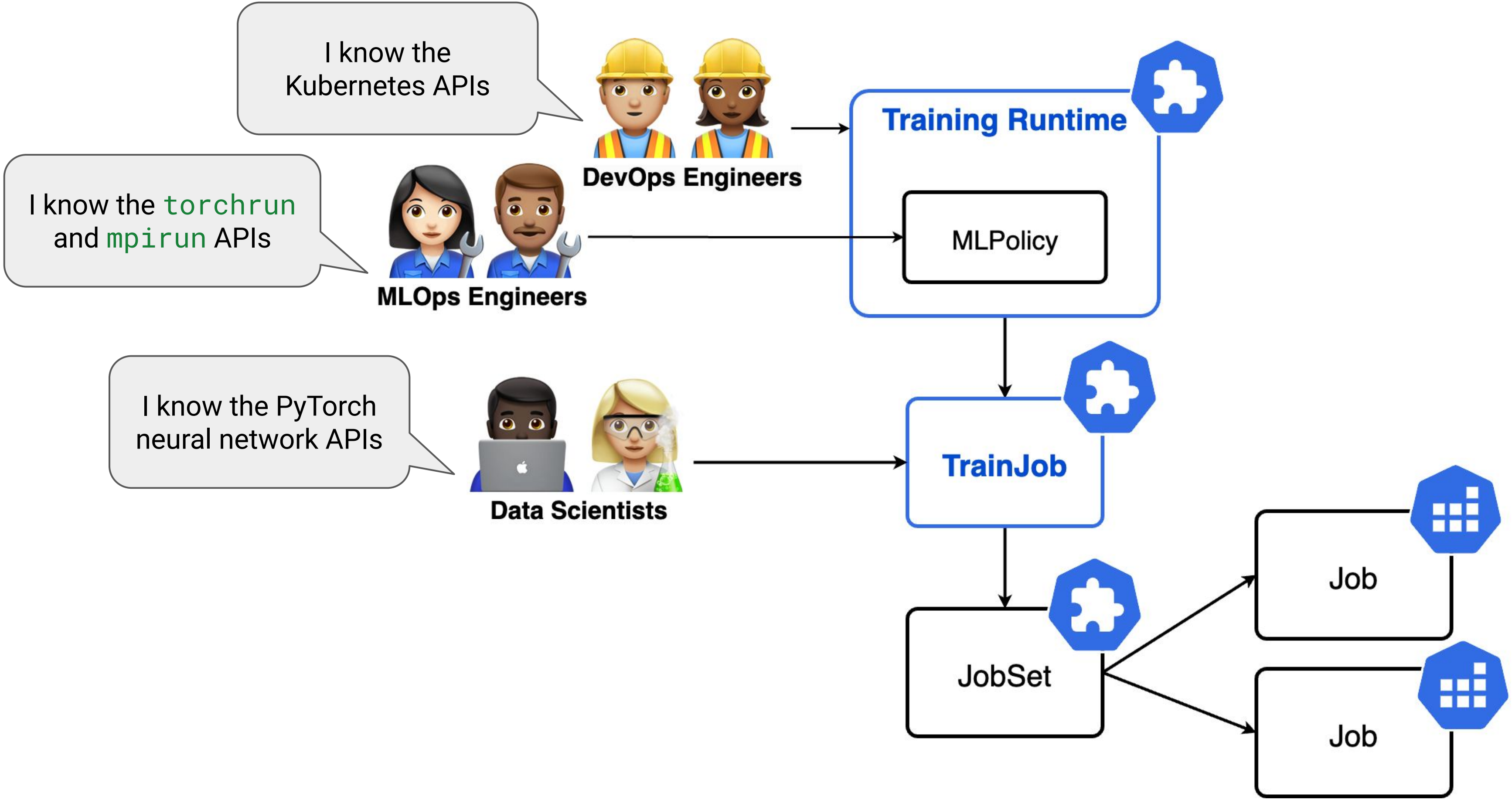
Torch Distributed &
LLM Fine-Tuning Demo

Goal One: Simplicity and Flexibility

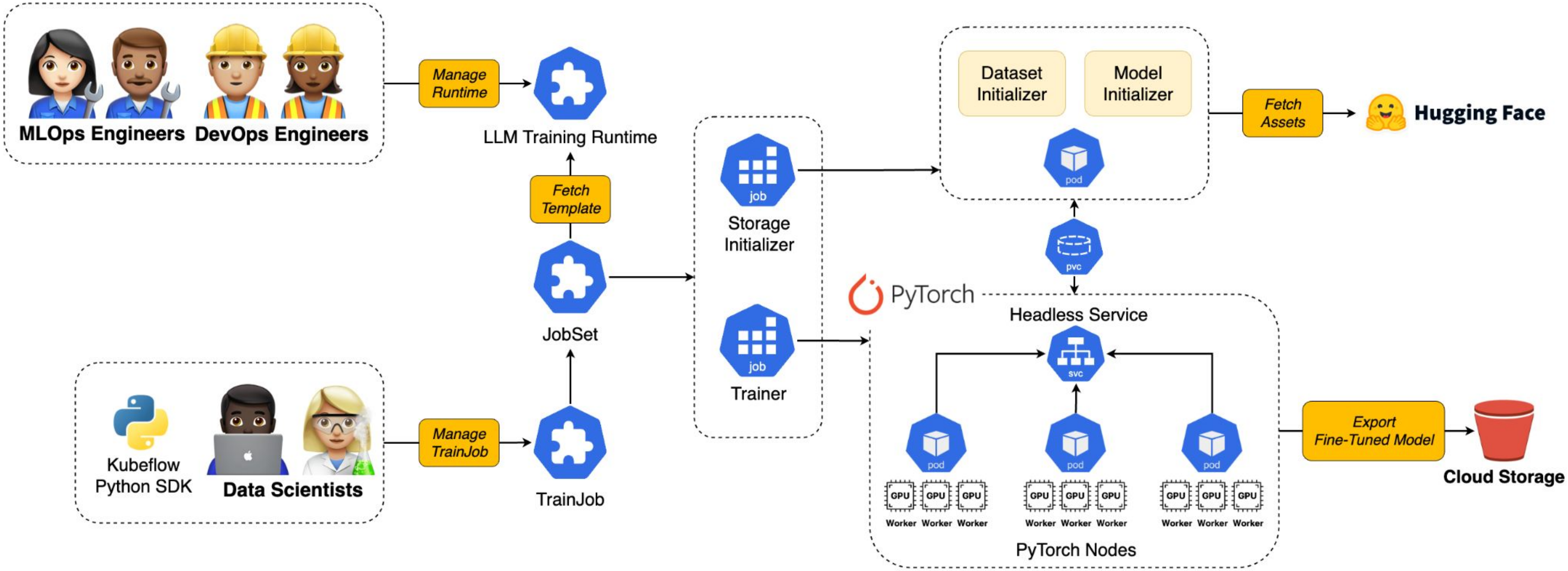
Simplicity

Flexibility

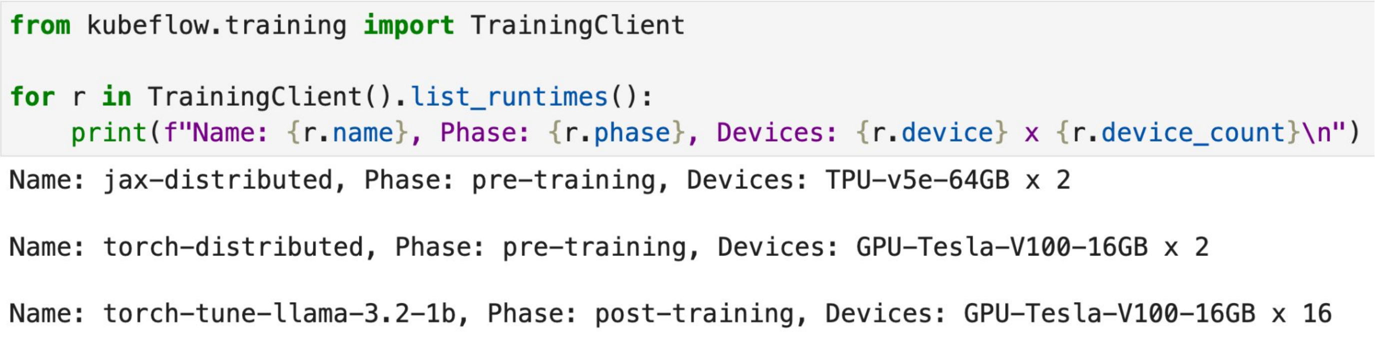
Simplicity: CRDs for Different Personas


I know the Kubernetes APIs
DevOps Engineers
I know the torchrun and mpirun APIs
MLOps Engineers
I know the PyTorch neural network APIs
Data Scientists
Training Runtime
MLPolicy
TrainJob
JobSet
Job
Job

LLM Runtime Lifecycle for Fine-Tuning



Simplicity: Python is the Main Interface



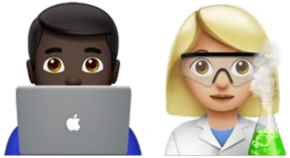
```python
from kubeflow.training import TrainingClient

for r in TrainingClient().list_runtimes():
    print(f"Name: {r.name}, Phase: {r.phase}, Devices: {r.device} x {r.device_count}\n")
```

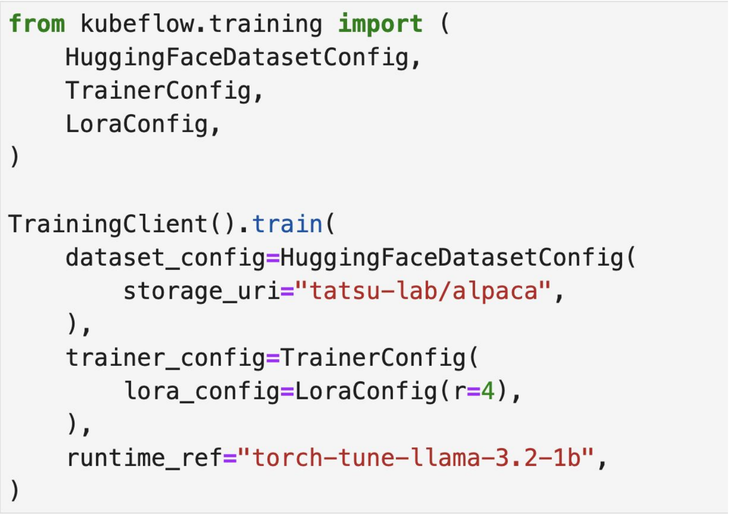
```
Name: jax-distributed, Phase: pre-training, Devices: TPU-v5e-64GB x 2

Name: torch-distributed, Phase: pre-training, Devices: GPU-Tesla-V100-16GB x 2

Name: torch-tune-llama-3.2-1b, Phase: post-training, Devices: GPU-Tesla-V100-16GB x 16
```

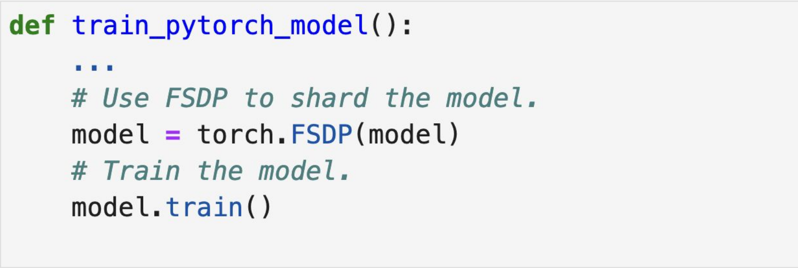


Data Scientists

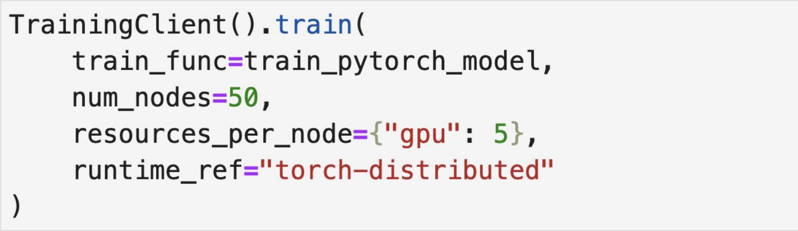


```python
from kubeflow.training import (
    HuggingFaceDatasetConfig,
    TrainerConfig,
    LoraConfig,
)

TrainingClient().train(
    dataset_config=HuggingFaceDatasetConfig(
        storage_uri="tatsu-lab/alpaca",
    ),
    trainer_config=TrainerConfig(
        lora_config=LoraConfig(r=4),
    ),
    runtime_ref="torch-tune-llama-3.2-1b",
)
```



```python
def train_pytorch_model():
    ...
    # Use FSDP to shard the model.
    model = torch.FSDP(model)
    # Train the model.
    model.train()
```



```python
TrainingClient().train(
    train_func=train_pytorch_model,
    num_nodes=50,
    resources_per_node={"gpu": 5},
    runtime_ref="torch-distributed"
)
```

Simplicity: TrainJob API

```yaml
apiVersion: kubeflow.org/v2alpha1
kind: TrainJob
metadata:
  name: k29520669946
spec:
  trainer:
    env:
    - name: LORA_CONFIG
      value: '{"r": 4, "lora_alpha": 16}'
  datasetConfig:
    storageUri: hf://tatsu-lab/alpaca
  modelConfig:
    output:
      storageUri: oci://registry/my-llm
  runtimeRef:
    apiGroup: kubeflow.org
    kind: ClusterTrainingRuntime
    name: torch-tune-llama-3.2-1b
```

Trainer Config

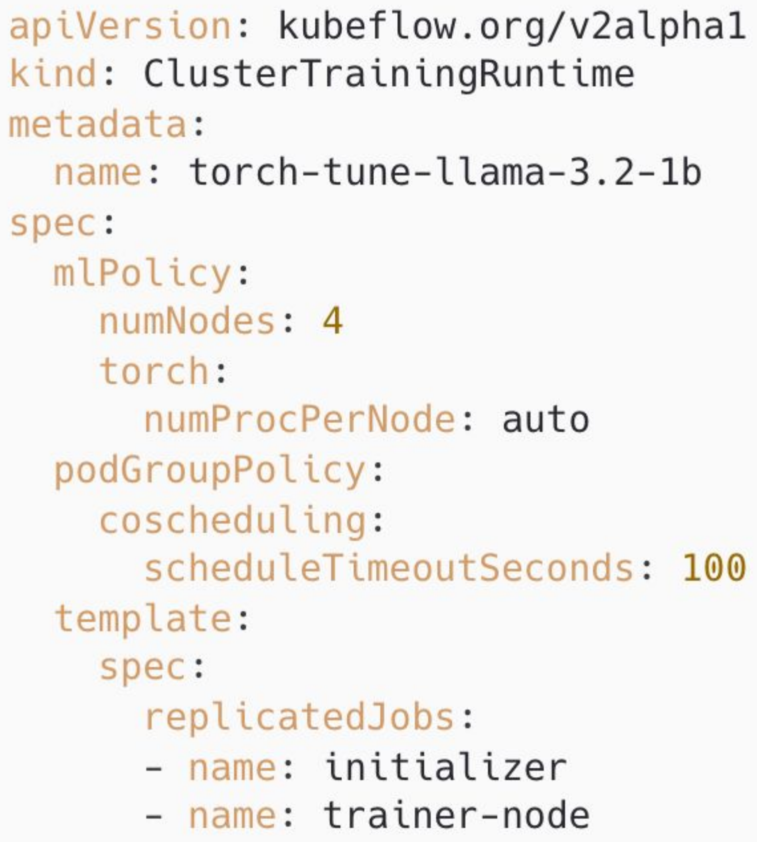
Dataset Config

Model Config

Runtime reference

Data Scientists

Flexibility: TrainingRuntime API



```yaml
apiVersion: kubeflow.org/v2alpha1
kind: ClusterTrainingRuntime
metadata:
  name: torch-tune-llama-3.2-1b
spec:
  mlPolicy:
    numNodes: 4
    torch:
      numProcPerNode: auto
  podGroupPolicy:
    coscheduling:
      scheduleTimeoutSeconds: 100
  template:
    spec:
      replicatedJobs:
      - name: initializer
      - name: trainer-node
```

ML configuration
(e.g. MPI, Torch)

Gang scheduling config

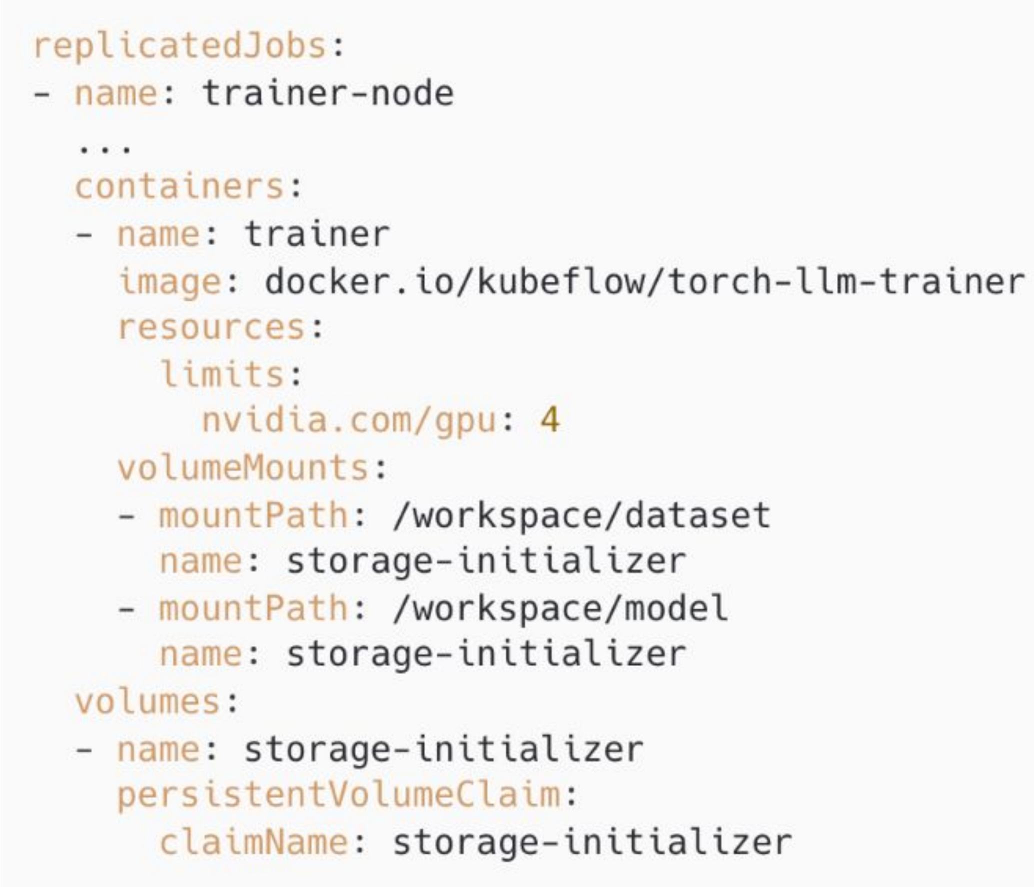
JobSet template

DevOps Engineers

MLOps Engineers

Flexibility: TrainingRuntime API

DevOps Engineers



```yaml
replicatedJobs:
- name: trainer-node
  ...
  containers:
  - name: trainer
    image: docker.io/kubeflow/torch-llm-trainer
    resources:
      limits:
        nvidia.com/gpu: 4
    volumeMounts:
    - mountPath: /workspace/dataset
      name: storage-initializer
    - mountPath: /workspace/model
      name: storage-initializer
  volumes:
  - name: storage-initializer
    persistentVolumeClaim:
      claimName: storage-initializer
```

Trainer

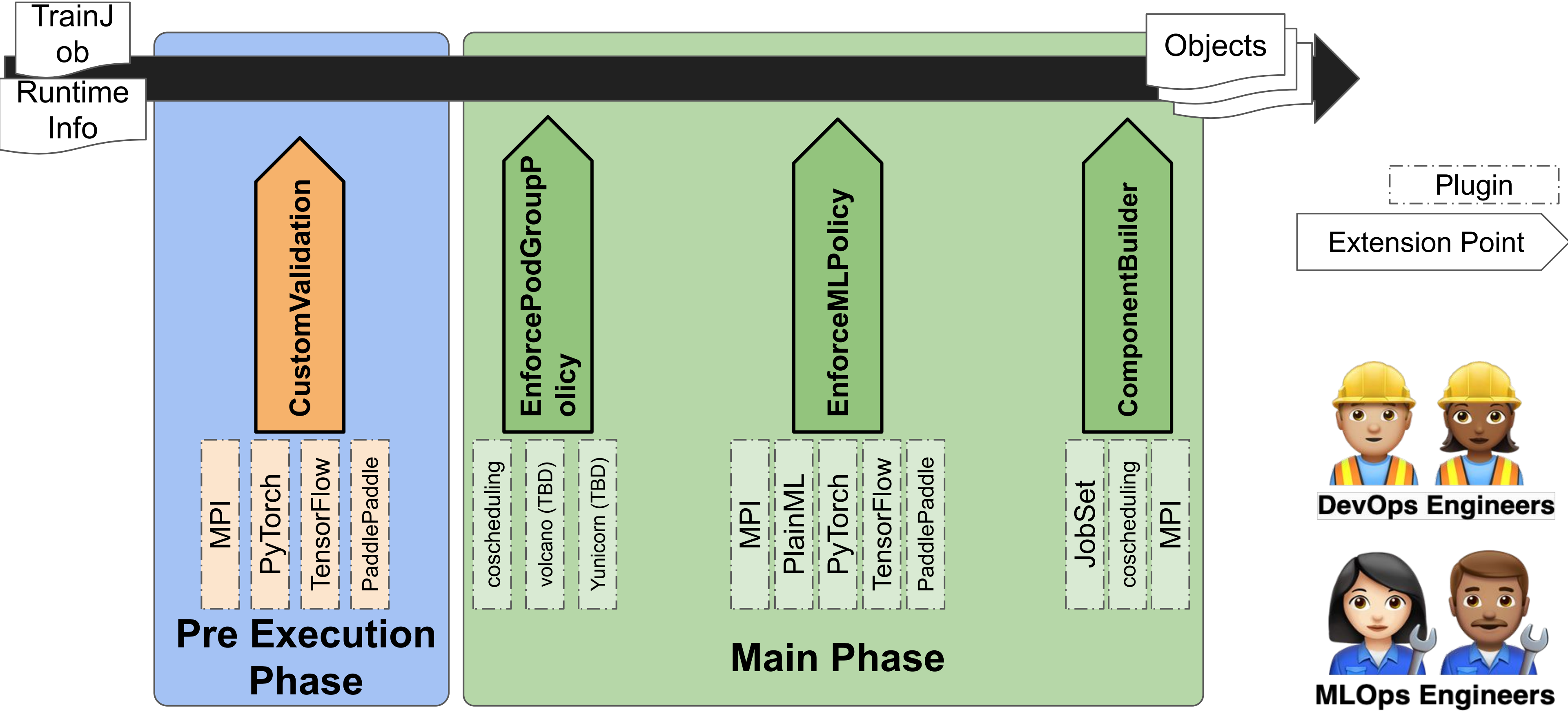
Storage Volume

Robust Support for ML Ecosystem

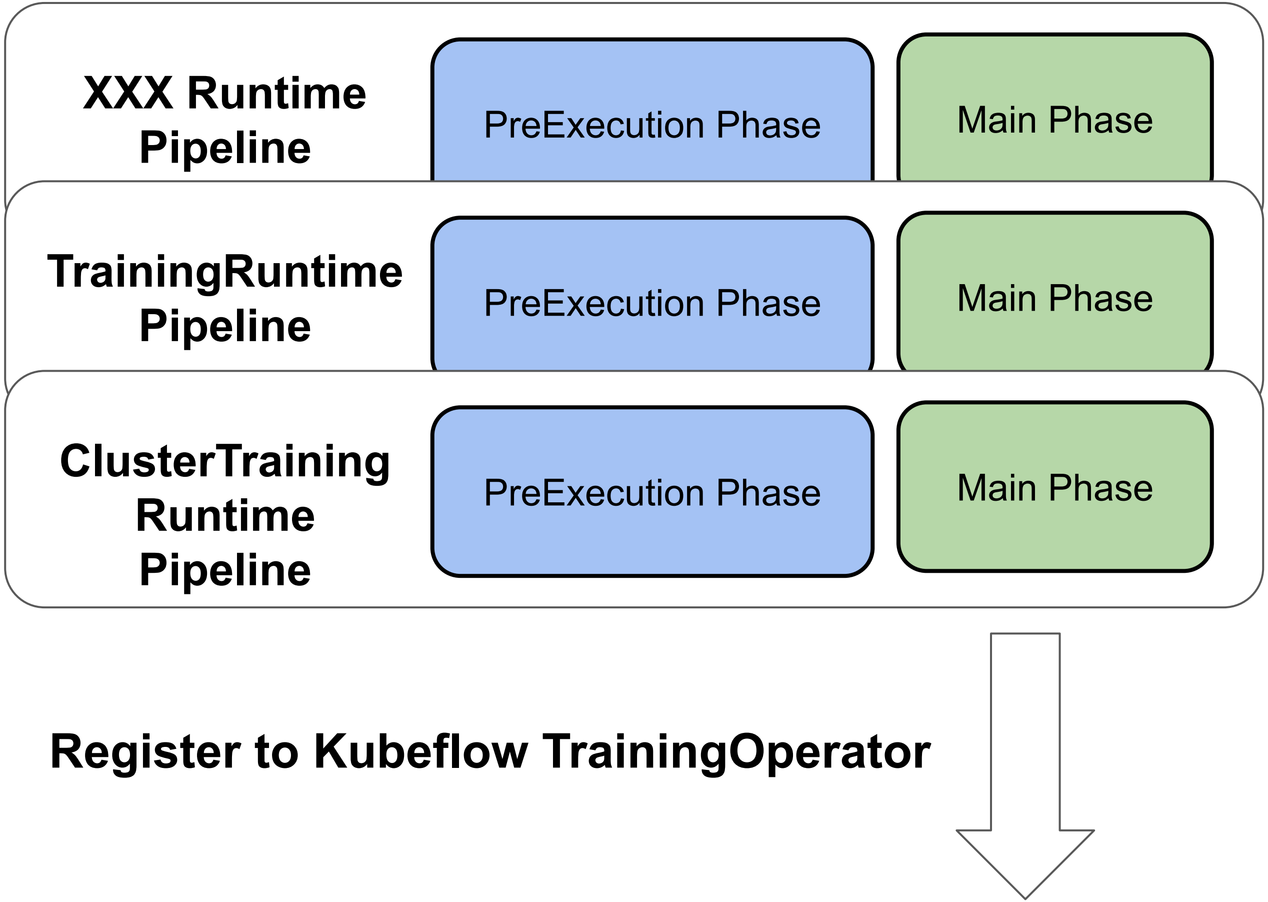
Frameworks	KF V2 Plugin	Phase	KF V1 CRD
PyTorch	Torch	Supported	PyTorchJob
TensorFlow	TensorFlow	In Progress	TFJob
Hugging Face	Torch & HF	In Progress	PyTorchJob
deepspeed	MPI	In Progress	MPIJob
MLX	MPI	Design	**No Support**
JAX	PlainML	Design	JAXJob
PaddlePaddle	PlainML	Design	PaddleJob
DMLC XGBoost	PlainML	Design	XGBoostJob

Open MPI

Flexibility: Kubeflow Job Pipeline Framework



Flexibility: Kubeflow Job Pipeline Framework



XXX Runtime Pipeline

PreExecution Phase

Main Phase

TrainingRuntime Pipeline

PreExecution Phase

Main Phase

ClusterTraining Runtime Pipeline

PreExecution Phase

Main Phase

Register to Kubeflow TrainingOperator

DevOps Engineers

MLOps Engineers

Goal Two: Scalability & Cost

Scalability & Cost Saving

TrainJob is Scalable and Cost Efficient

TrainJob optimizes GPU cost by delegating I/O tasks to CPU nodes

TrainJob is using Kubernetes native workloads: JobSet + Job

TrainJob operates Pods concurrently, but Kubeflow Training V1 creates Pods sequentially

Kubeflow Training V2 Summary

- ✓ Simple to use and scale
- ✓ Python is the main user interface
- ✓ Enables quick fine-tuning of LLMs
- ✓ Provides robust support for the ML ecosystem.
- ✓ Streamline dataset and pre-trained model initialization
- ✓ Consolidates efforts between Kubernetes and Kubeflow communities.

Future Work

Implement more runtimes for LLMs fine-tuning

Support for MPI V2 and other ML frameworks

KEP-2170: Kubeflow Training V2

Improve Kubernetes for AI training workloads

Serial Job Execution: sigs.k8s.io/jobset#680
Elastic JobSet: sigs.k8s.io/jobset#463
Stateful Index Jobs for volume management: sigs.k8s.io/jobset#572
Multi Cluster Job dispatching with Kueue
Support Quota management and Job Queueing with Kueue

Get Involved

Kubeflow AutoML and Training WG

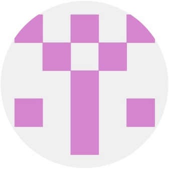
- Join the CNCF Slack
  - #kubeflow-training
- Participate in the Kubeflow Training V2
- AutoML and Training WG bi-weekly meetings:
  - Wednesdays 2pm UTC
  - Wednesdays 5pm UTC

Kubernetes Batch WG

- Join the Kubernetes Slack
  - #wg-batch
- Participate in the the WG Batch
- Batch WG bi-weekly meetings:
  - Thursdays 3pm CET
  - Thursdays 3pm PT

Thanks to our Contributors!

ahg-g


akshaychitneni

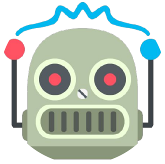
andreyvelich

Bobbins228

champon1020

danielvegamyhre

deepanker13


droctothorpe

franciscojavierarceo

johnugeorge

helenxie-bit

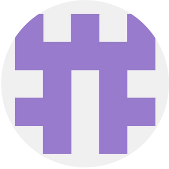
kannon92

lowang-bh

mimowo

mszadkow

rpemsel


saileshd1402

sandipanpanda

seanlaii

Syulin7

tariq-hasan

tenzen-y

terrytangyuan

YosiElias

varshaprasad96

vsoch

Thank you!

Andrey Velichkevich

- Email: andrey.velichkevich@gmail.com
- GitHub: andreyvelich
- Slack
- LinkedIn
- BlueSky and Twitter

Yuki Iwai

- Email: yuki.iwai.tz@gmail.com
- GitHub: tenzen-y
- Slack
- LinkedIn

Please scan the QR Code above to leave feedback on this session