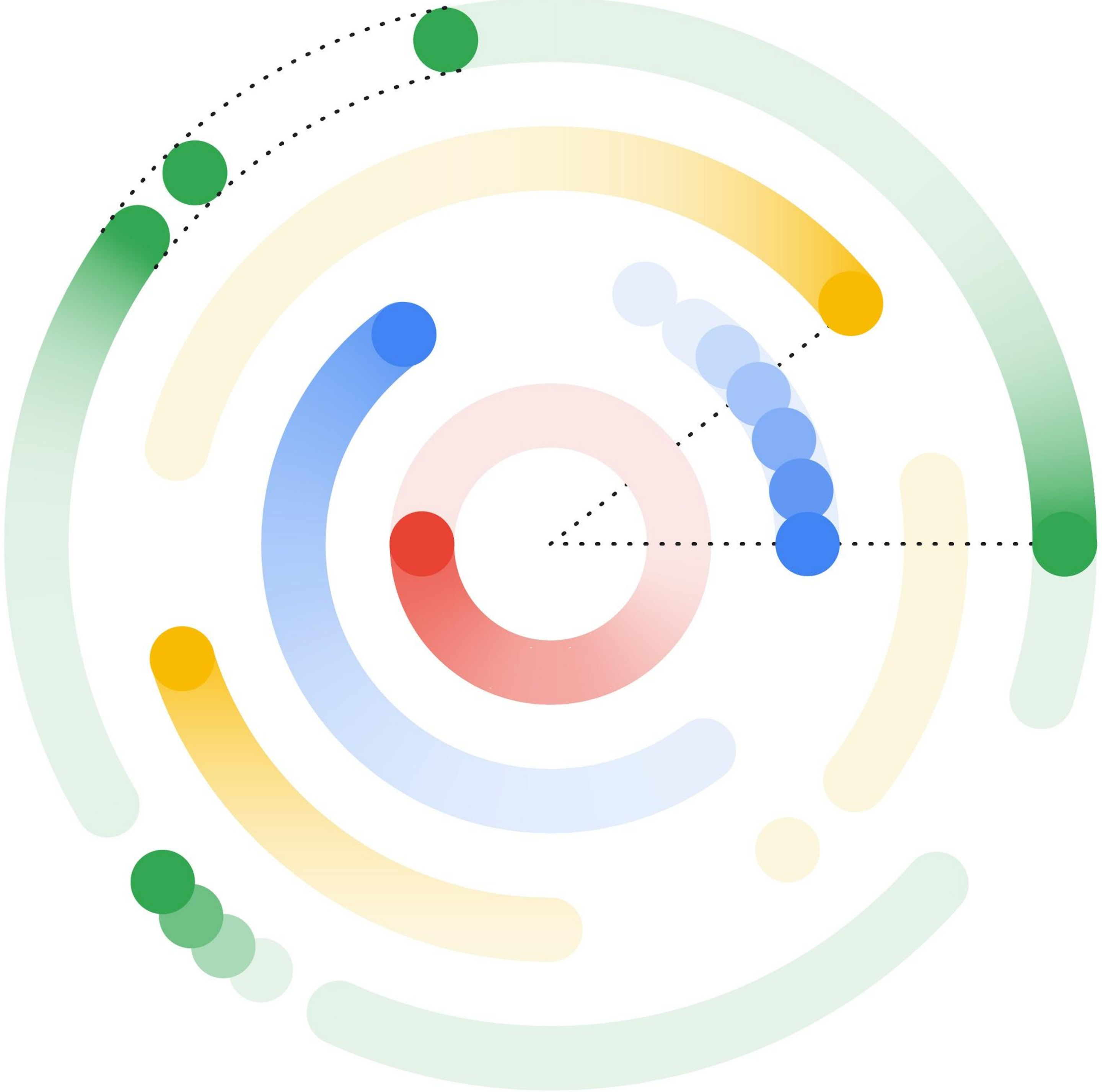
Google Cloud

Scalable ML Deployment Using PyTorch and Kubeflow Pipelines

Google Cloud Applied ML Summit

Solving for the future.

06/10/21



Geeta Chauhan

Head of AI Partner
Engineering
Facebook AI

Table of Contents

PyTorch Community Growth 01

Challenges for Scalable ML 02

Kubeflow Pipelines + PyTorch 03

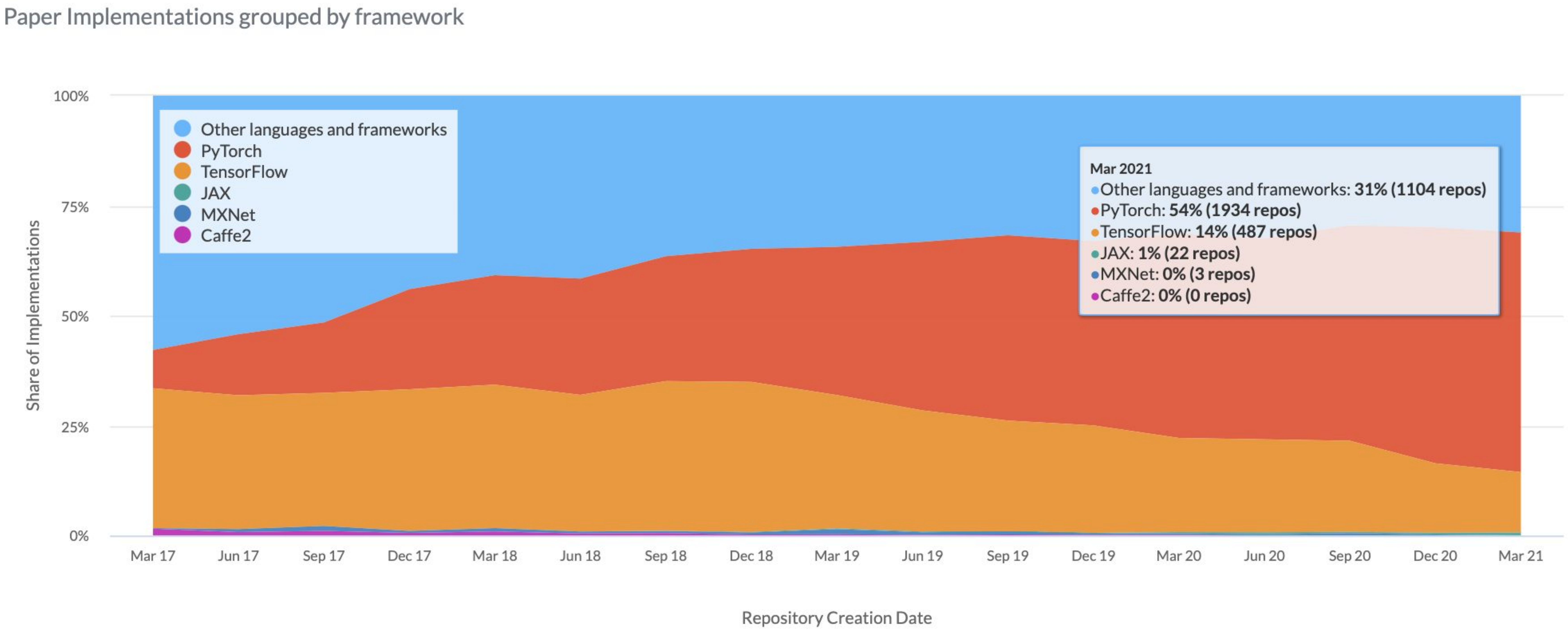
NLP Workflow + Demo 04

Serverless Scalable Pipelines 05

Computer Vision Workflow + Demo 06

Reference 07

PyTorch Community Growth

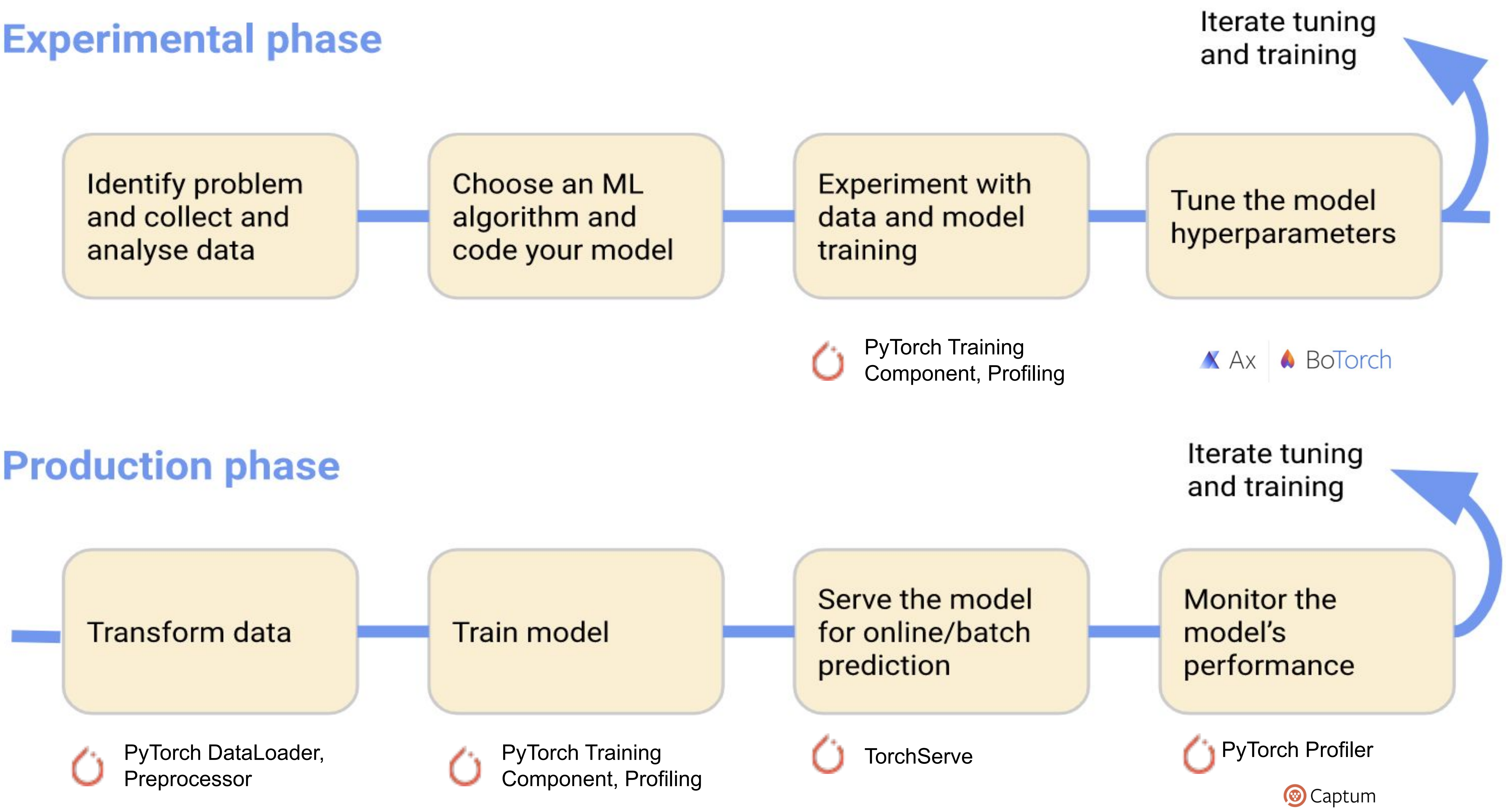


Source: https://paperswithcode.com/trends

Challenges for Scalable ML

- Building and Managing pipelines is hard
- Continuous iterative process, optimize for metric
- Over time data changes, model drift
- Experiment tracking is difficult
- Model artifacts get lost
- Diverse deployment environments
- What happens when you hit scale?

Kubeflow Pipelines + PyTorch



Experimental phase

Identify problem and collect and analyse data

Choose an ML algorithm and code your model

Experiment with data and model training

PyTorch Training Component, Profiling

Tune the model hyperparameters

Ax | BoTorch

Iterate tuning and training

Production phase

Transform data

PyTorch DataLoader, Preprocessor

Train model

PyTorch Training Component, Profiling

Serve the model for online/batch prediction

TorchServe

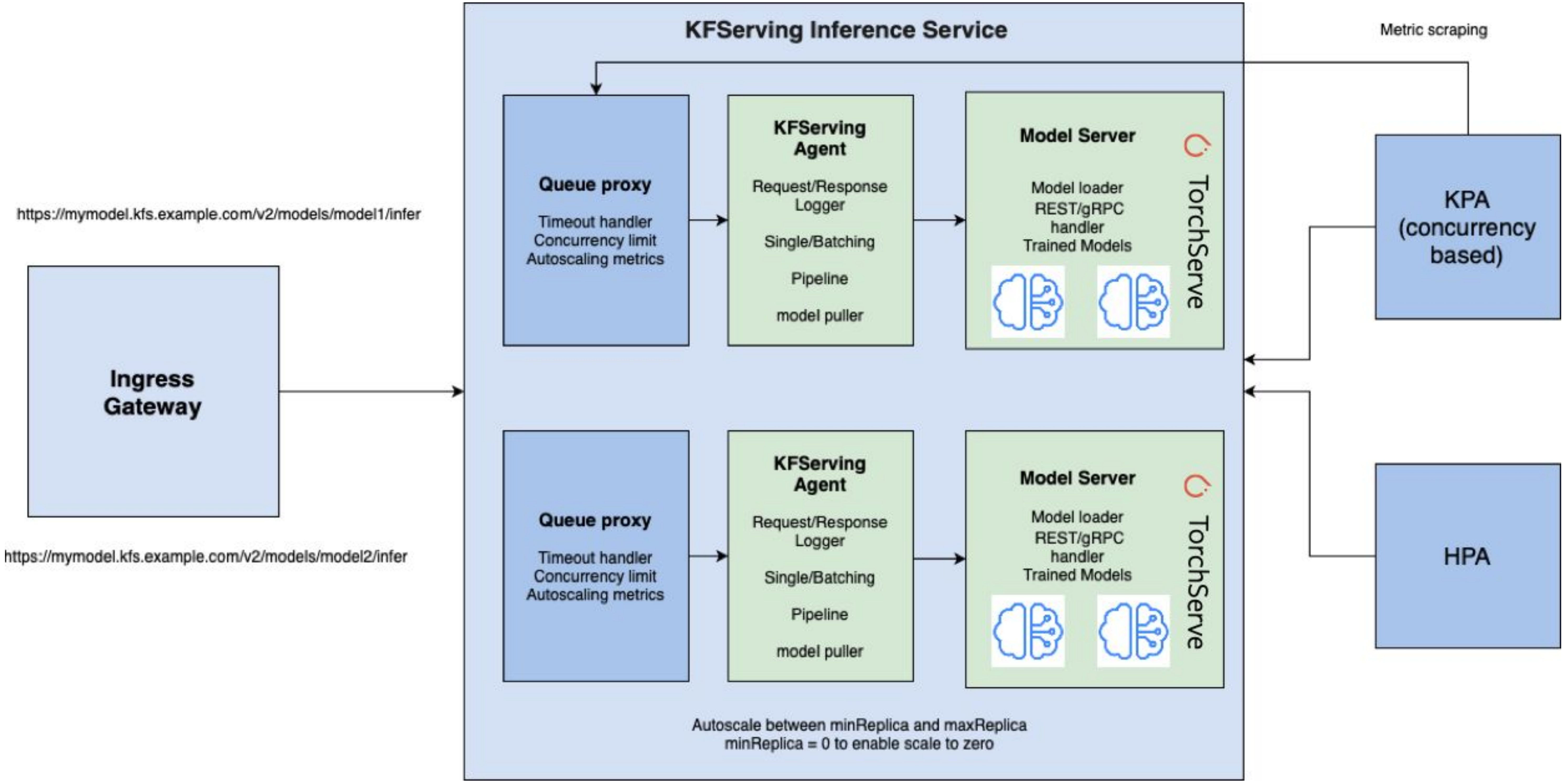
Monitor the model's performance

PyTorch Profiler

Captum

Iterate tuning and training

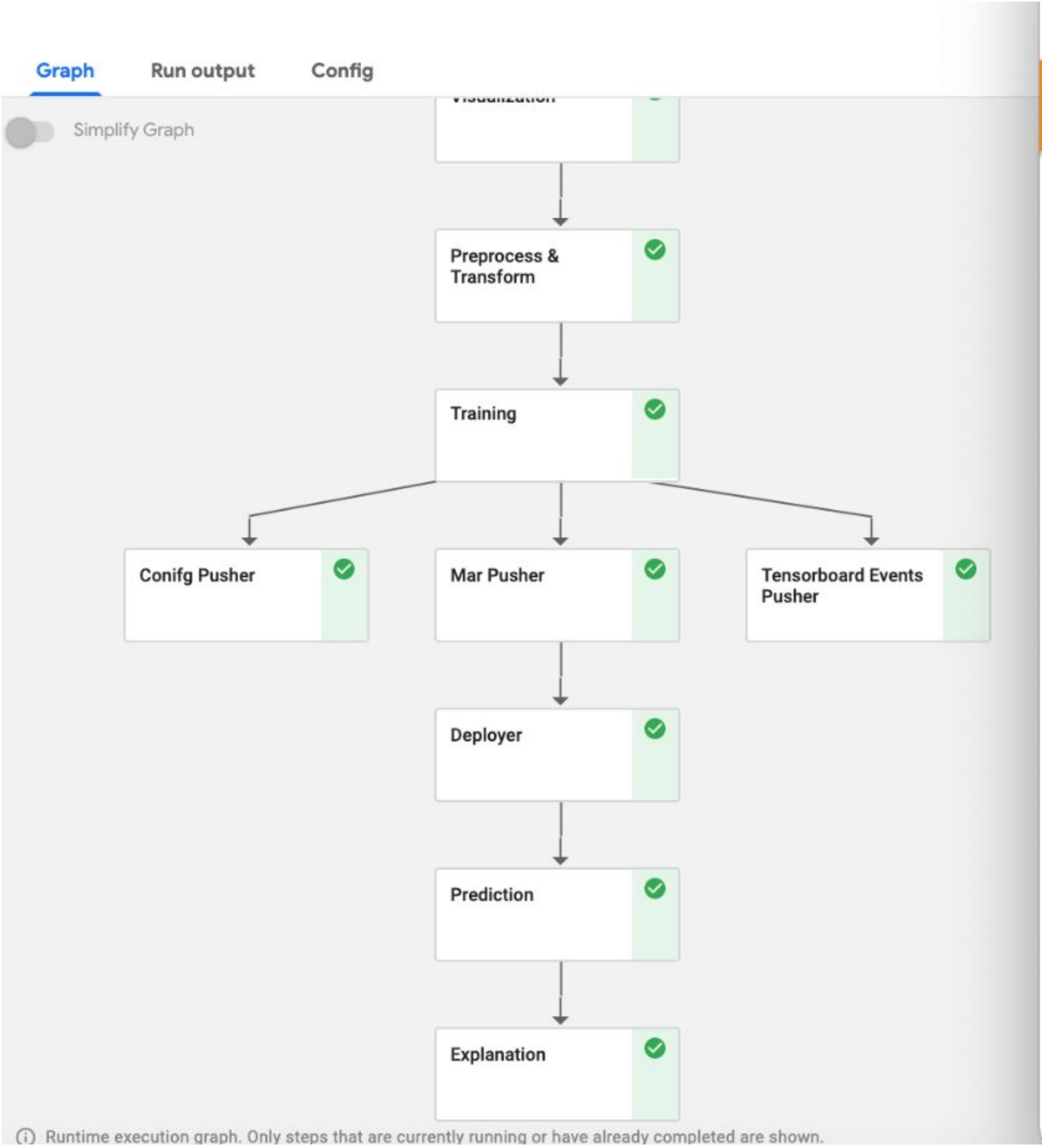
KFServing for Model Serving with TorchServe

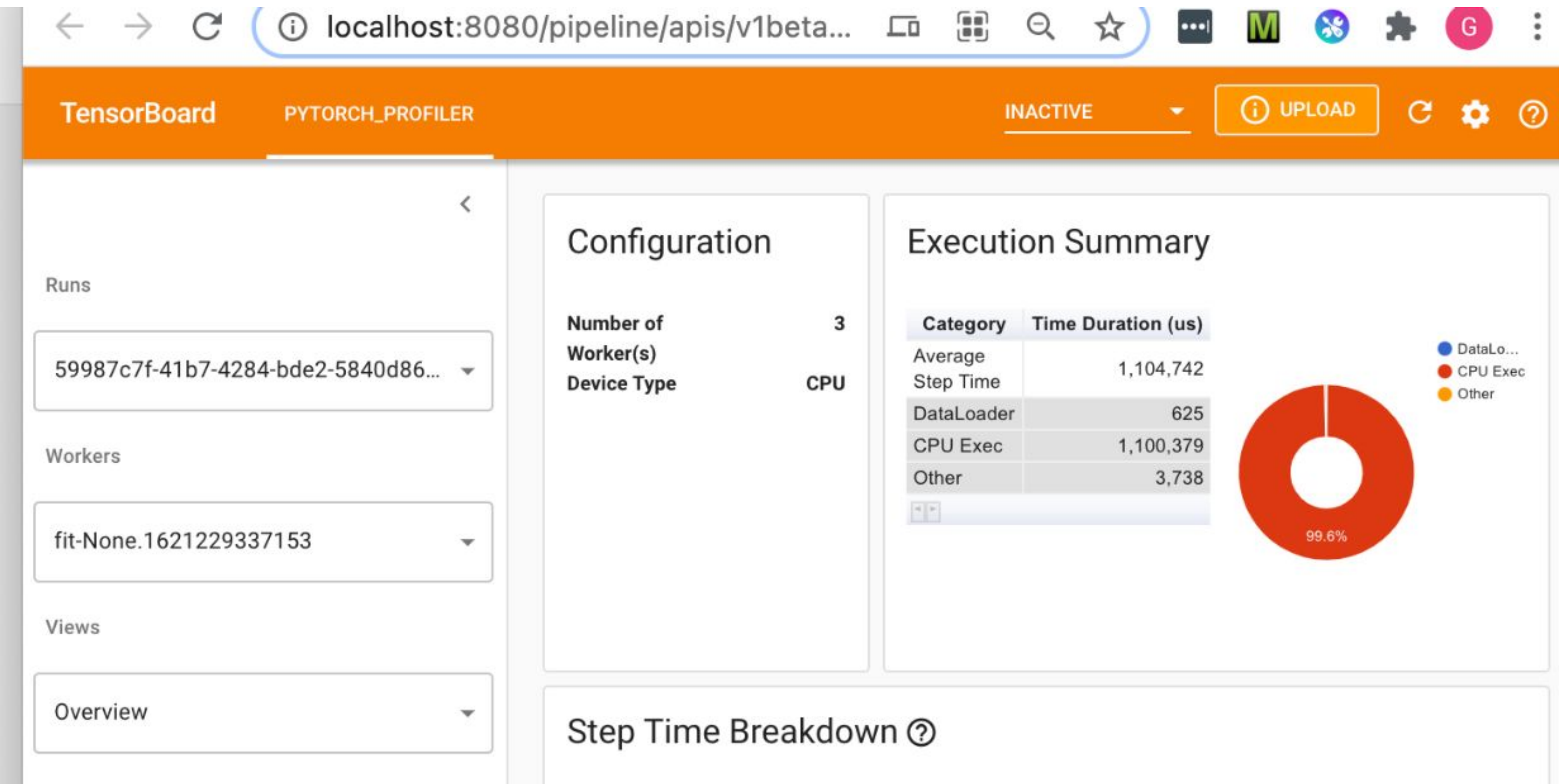


Kubeflow Pipelines + PyTorch

- Dataloading and preprocessing
- Model Training
- Hyper Parameter Optimization using Ax/Botorch
- Model deployment & Serving using TorchServe + KFServing w/ canary rollouts, autoscaling, prometheus monitoring
- Visualizations using Tensorboard PyTorch Profiler
- Model Interpretability using Captum
- Artifact Lineage Tracking
- Open Source KFP: on-prem or any cloud
- GCP Vertex AI Pipelines - Serverless

NLP BERT workflow - Open source KFP





Runs: 59987c7f-41b7-4284-bde2-5840d86...

Workers: fit-None.1621229337153

Views: Overview

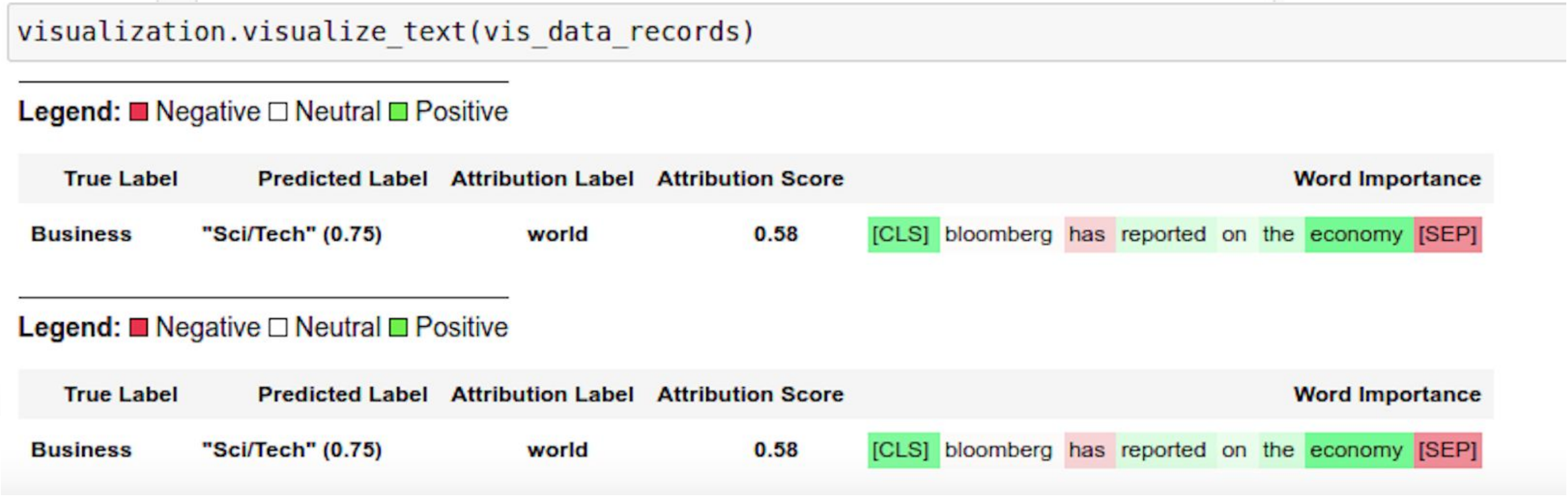
Configuration

Number of Worker(s)	3
Device Type	CPU

Execution Summary

Category	Time Duration (us)
Average Step Time	1,104,742
DataLoader	625
CPU Exec	1,100,379
Other	3,738

Step Time Breakdown



```
visualization.visualize_text(vis_data_records)
```

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Business	"Sci/Tech" (0.75)	world	0.58	[CLS] bloomberg has reported on the economy [SEP]

Legend: 🟥 Negative □ Neutral 🟩 Positive

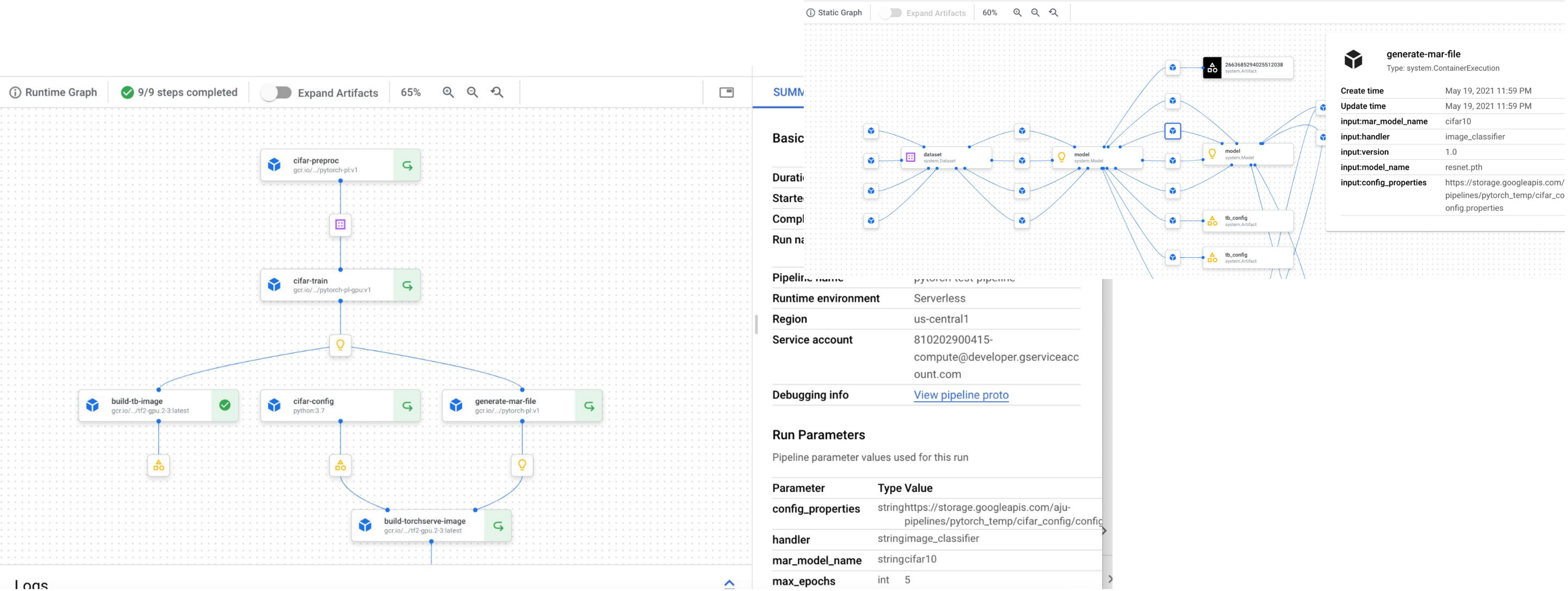
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Business	"Sci/Tech" (0.75)	world	0.58	[CLS] bloomberg has reported on the economy [SEP]

Demo

Scalable Serverless Vertex AI Pipelines

- Automated, Scalable, Serverless, no need to manage cluster
- Cost Effective - Pay only for what you use
- Build using familiar open source KFP SDK
- Metadata and Lineage for all Artifacts, metrics across and execution for the workflow
- Pipeline run analysis, monitoring and debugging
- Security using Cloud IAM, VPC-SC and CMEK

Computer Vision Workflow - Vertex AI Pipeline



Basic

Durati

Starte

Compl

Run na

Pipeline name	pytorch test pipeline
Runtime environment	Serverless
Region	us-central1
Service account	810202900415-compute@developer.gserviceaccount.com
Debugging info	View pipeline proto

Run Parameters

Pipeline parameter values used for this run

Parameter	Type Value
config_properties	stringhttps://storage.googleapis.com/aju-pipelines/pytorch_temp/cifar_config/config
handler	stringimage_classifier
mar_model_name	stringcifar10
max_epochs	int 5

generate-mar-file

Type: system.ContainerExecution

Create time	May 19, 2021 11:59 PM
Update time	May 19, 2021 11:59 PM
input:mar_model_name	cifar10
input:handler	image_classifier
input:version	1.0
input:model_name	resnet.pth
input:config_properties	https://storage.googleapis.com/pipelines/pytorch_temp/cifar_config.properties

Demo

References

- KFP Examples Github: http://bit.ly/pt-kfp-demos
- Vertex AI Pipeline - Colab Notebook: http://bit.ly/PT-Vertex-AI
- Vertex AI Pipelines: https://cloud.google.com/vertex-ai/docs/pipelines
- Captum.ai: https://captum.ai/tutorials/CIFAR_TorchVision_Captum_Insights
- PyTorch Tensorboard Profiler: https://pytorch.org/blog/introducing-pytorch-profiler-the-new-and-improved-performance-tool/
- Operationalize, Scale & Infuse Trust in AI using KFServing: https://blog.kubeflow.org/release/official/2021/03/08/kfserving-0.5.html

Google Cloud Summit

Q&A

Contact:

Email: gchauhan@fb.com

Linkedin:
https://www.linkedin.com/in/geetachauhan/

Google Cloud

Google Cloud Summit

Thank you for joining