
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM335

Scalable Kubernetes ML applications with Kubeflow on AWS & Amazon SageMaker

Kanwaljit Khurmi

Sr. AI/ML Solutions Architect
AWS

Arindam Paul

Senior PMT - AWS ML Services
AWS

Victor Krylov

Principal Member of Technical Staff
athenahealth

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Why distributed deep learning and AWS services to support it

Discuss AWS Kubeflow distribution

Demonstrate Kubeflow pipeline, leveraging Amazon SageMaker for distributed training

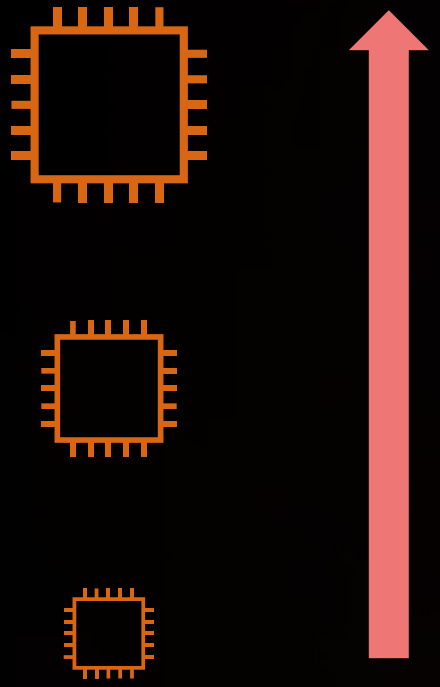
Discuss athenahealth's journey to create enterprise-ready ML platform using Kubeflow on AWS

Questions

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why scale? Scaling DL to get faster results

Scaling up



GPU instance size

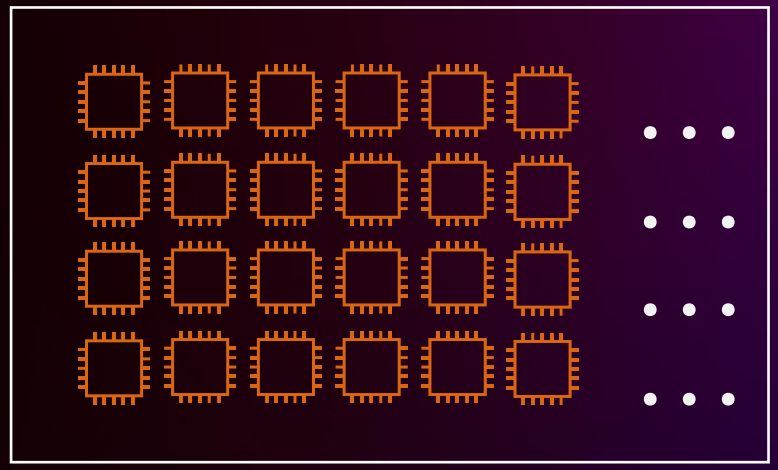
GPU memory

CPU cores

System memory

...

Scaling out





10s, 100s, and 1000s GPU-accelerated instances

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS services for scaling DL environments



Amazon SageMaker

SAGEMAKER CANVAS
No-code ML for business analysts

SAGEMAKER STUDIO LAB
Learn ML

SAGEMAKER GROUND TRUTH
Label data

SAGEMAKER STUDIO | RSTUDIO

Prepare data → Store features → Detect bias → Build with notebooks → Train models → Tune parameters → Deploy in production → Explain predictions → Manage & monitor

CI/CD

Manage edge devices

Management

Deployment, scheduling, scaling, and management of containerized applications



Amazon Elastic Container Service (Amazon ECS)



Amazon Elastic Kubernetes Service (Amazon EKS)



Compute

Where the containers run



Amazon Elastic Compute Cloud (Amazon EC2)

Image registry

Container image repository



Amazon Elastic Container Registry (Amazon ECR)

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kubernetes for machine learning



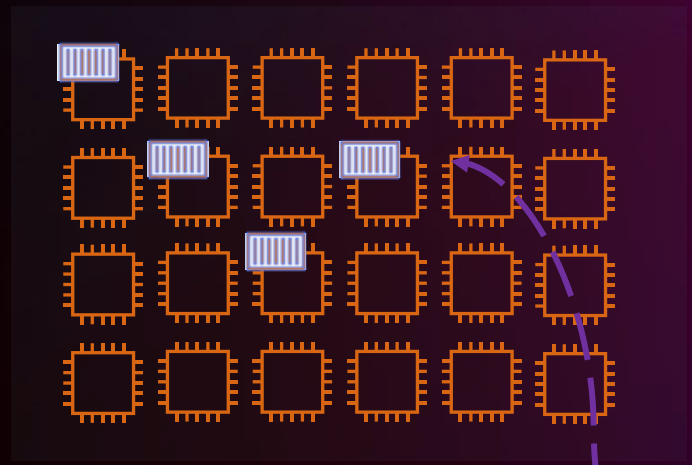
Desired state
- Train on 2 GPUs
- Host model on 10 CPUs



```
apiVersion: batch/v1
kind: Job
metadata:
  name: eks-training-job
spec:
  template:
    spec:
      restartPolicy: Never
      containers:
       - name: keras-example
         image: .dkr.ecr.us-west-2.amazonaws.com/ek
```

Orchestrator

Amazon EKS



Amazon ECR

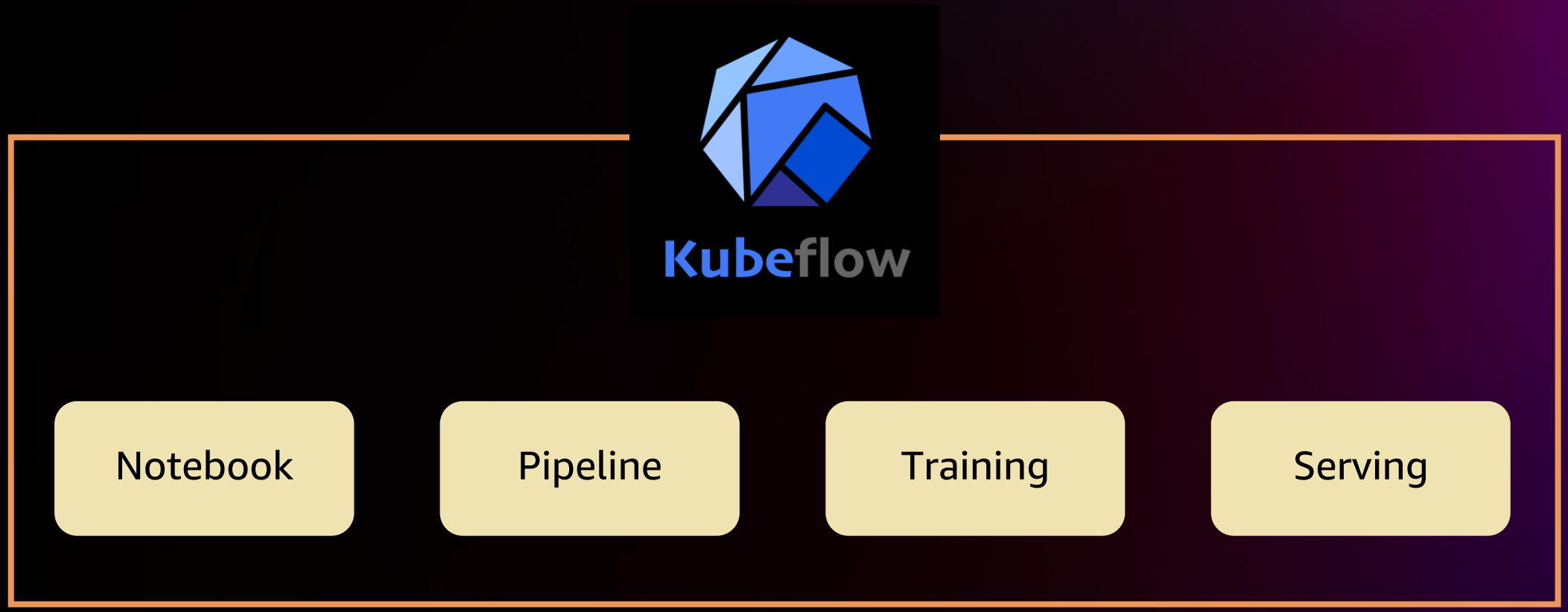
Kubernetes for ML

- ✓ Not data scientist friendly
- ✓ No ML experiment and workflow management
- ✓ Operational overheads in managing worker nodes

aws

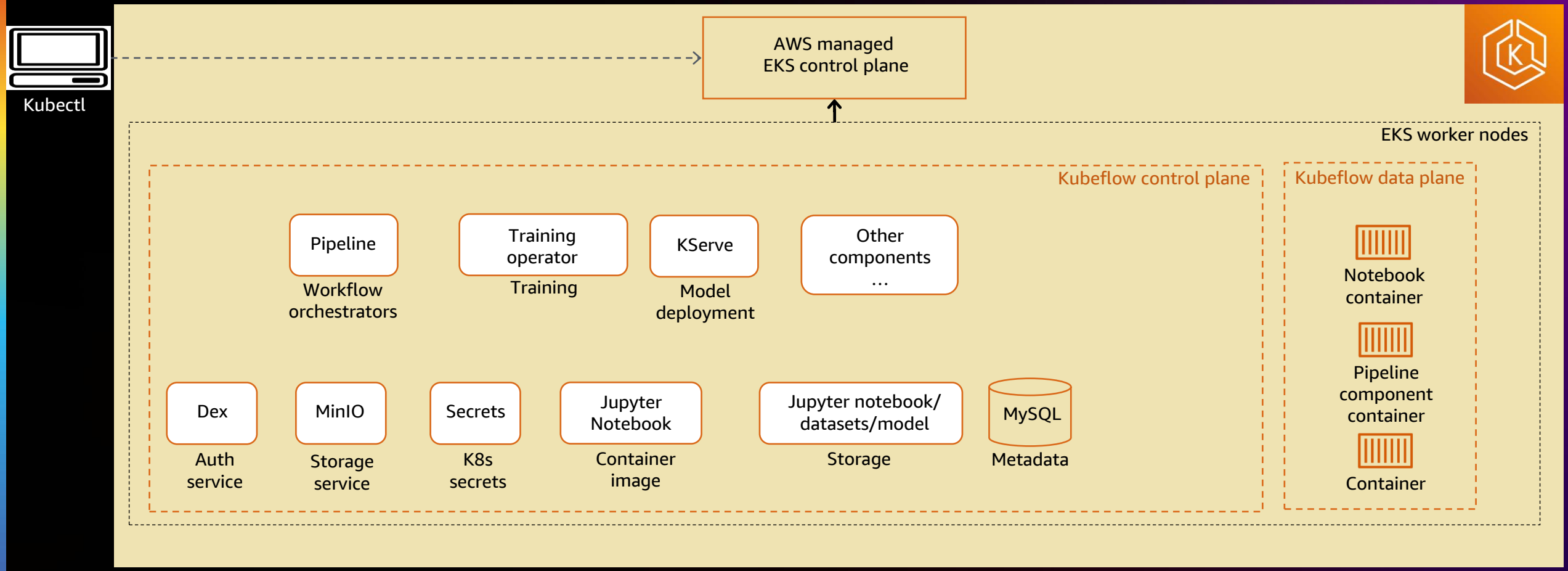
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introducing Kubeflow

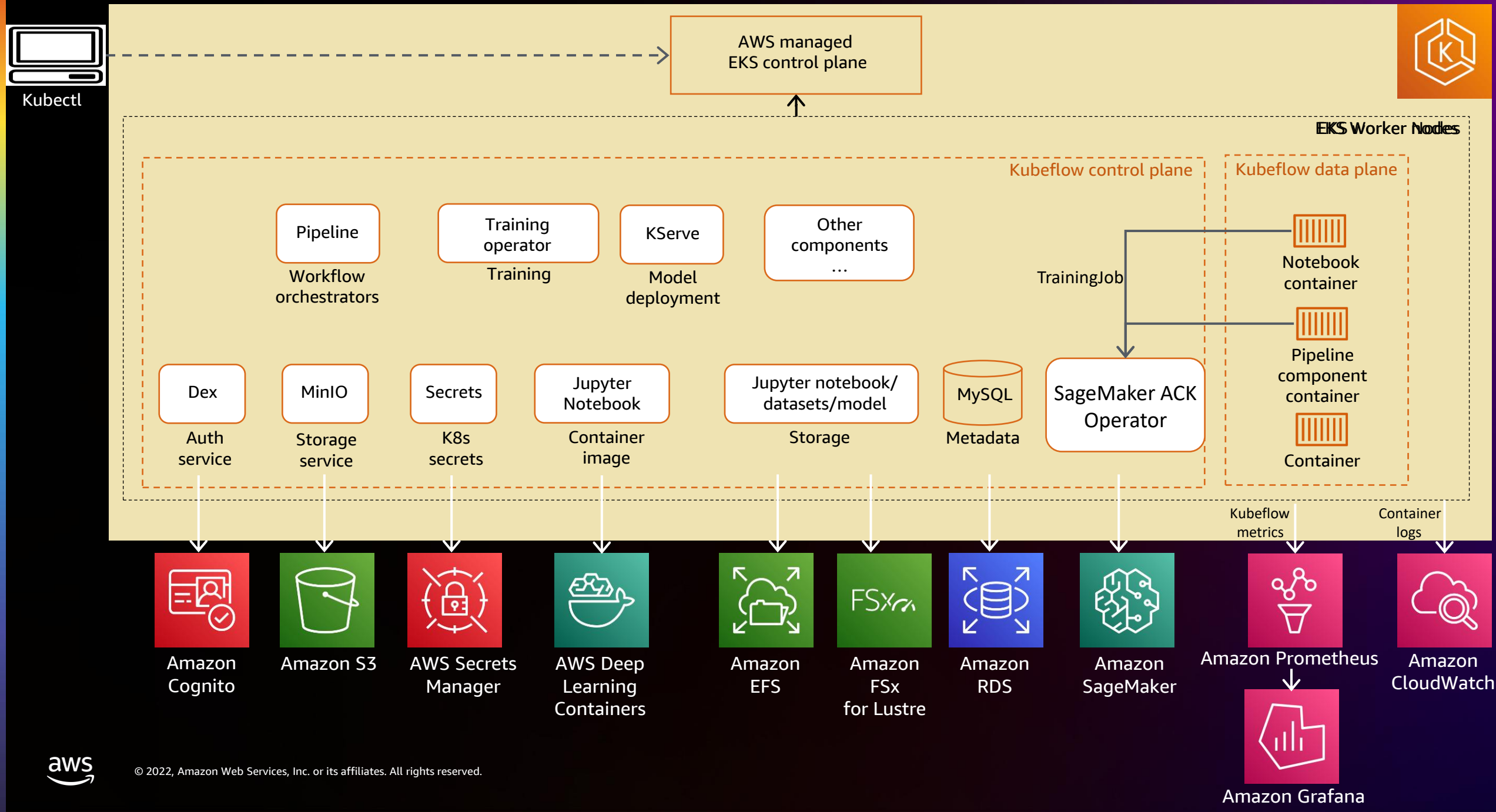


Kubeflow is the machine learning (ML) toolkit for Kubernetes

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



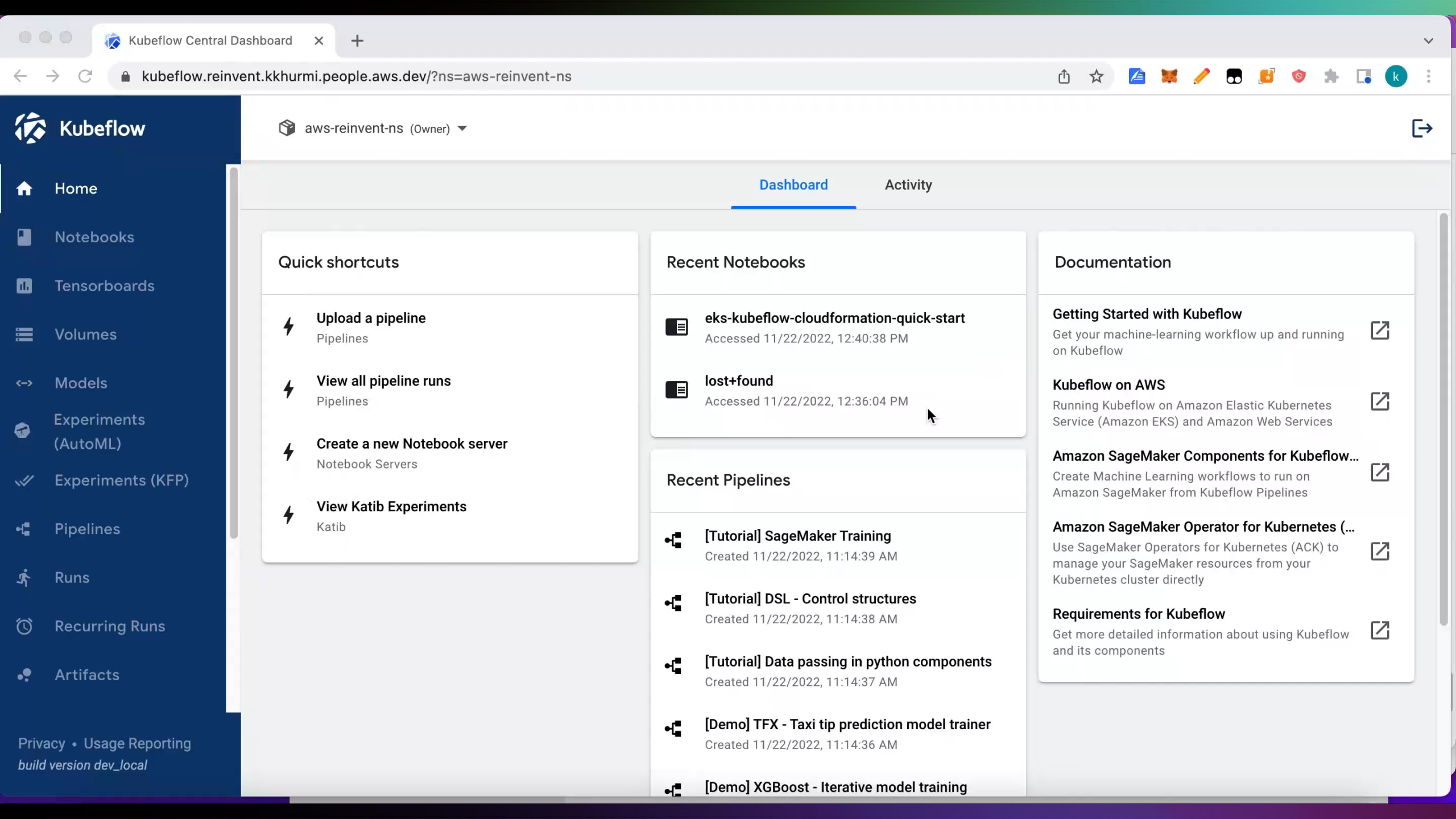
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo: Create a pipeline for running distributed training on Amazon SageMaker

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Kubeflow Central Dashboard
kubeflow.reinvent.kkhurmi.people.aws.dev/?ns=aws-reinvent-ns
Kubeflow
aws-reinvent-ns (Owner)
Home
Notebooks
Tensorboards
Volumes
Models
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Privacy • Usage Reporting
build version dev_local
Dashboard
Activity
Quick shortcuts
Upload a pipeline
Pipelines
View all pipeline runs
Pipelines
Create a new Notebook server
Notebook Servers
View Katib Experiments
Katib
Recent Notebooks
eks-kubeflow-cloudformation-quick-start
Accessed 11/22/2022, 12:40:38 PM
lost+found
Accessed 11/22/2022, 12:36:04 PM
Recent Pipelines
[Tutorial] SageMaker Training
Created 11/22/2022, 11:14:39 AM
[Tutorial] DSL - Control structures
Created 11/22/2022, 11:14:38 AM
[Tutorial] Data passing in python components
Created 11/22/2022, 11:14:37 AM
[Demo] TFX - Taxi tip prediction model trainer
Created 11/22/2022, 11:14:36 AM
[Demo] XGBoost - Iterative model training
Documentation
Getting Started with Kubeflow
Get your machine-learning workflow up and running on Kubeflow
Kubeflow on AWS
Running Kubeflow on Amazon Elastic Kubernetes Service (Amazon EKS) and Amazon Web Services
Amazon SageMaker Components for Kubeflow...
Create Machine Learning workflows to run on Amazon SageMaker from Kubeflow Pipelines
Amazon SageMaker Operator for Kubernetes (...
Use SageMaker Operators for Kubernetes (ACK) to manage your SageMaker resources from your Kubernetes cluster directly
Requirements for Kubeflow
Get more detailed information about using Kubeflow and its components

athenahealth machine learning platform

Background

ML development in Kubeflow

Demonstrate CI/CD in development cycle

Nonintrusive Kubeflow cluster upgrade

Monitoring

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Background

- athenahealth

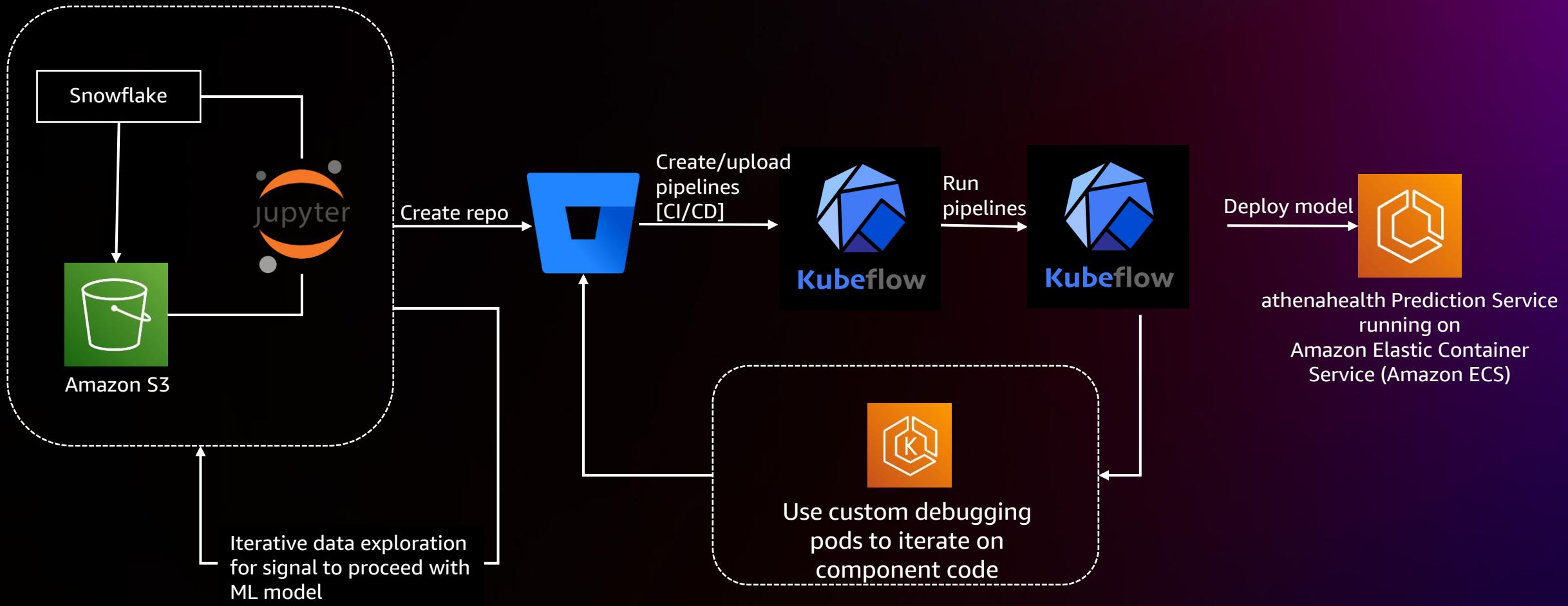
  We partner with healthcare organizations across the care continuum to drive clinical and financial results. Our expert teams build modern technology on an open, connected ecosystem, yielding insights that make a difference for our customers and their patients

- Our move to Kubeflow on AWS
  - Transition the existing array of training tools for growth
  - Kubeflow on AWS
    - Active roadmap and open-source community
    - Works with existing tooling
    - Robust support and integration with other AWS services

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

athenahealth development cycle on Kubeflow



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ATHN-100-re-invent [Data Scien | Kubeflow Central Dashboard | Kubeflow Central Dashboard

prod-kf.ml.ai.athena.io/_/pipeline/?ns=misty#/runs/details/27190eb8-0c4a-42ac-ad55-c415667df285

Kubeflow

misty

Home
Notebooks
Tensorboards
Volumes
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Executions
Grafana
Kibana

Privacy • Usage Reporting
build version dev_local

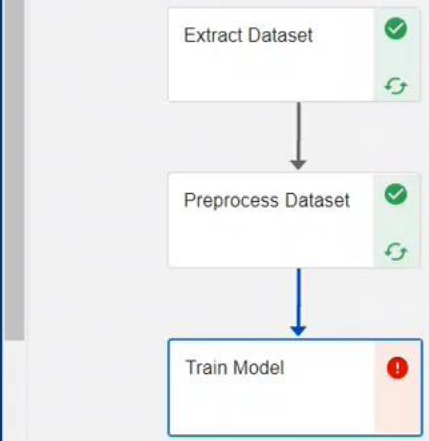
Experiments › svc-athn-ai-ds-mltp.misty.athn-100

Retry | Clone run | Terminate | Archive

end_to_end_pipeline.c303fd3.2022-11-22_19-54-28

Graph | Run output | Config

Simplify Graph



Runtime execution graph. Only steps that are currently running

end-to-end-of-misty-pipeline-pjvcs-3965362795

· This step is in Failed state with this message: Error (exit code 1)

Input/Output | Visualizations | Details | Volumes | Logs | Pod | Events | ML Metadata

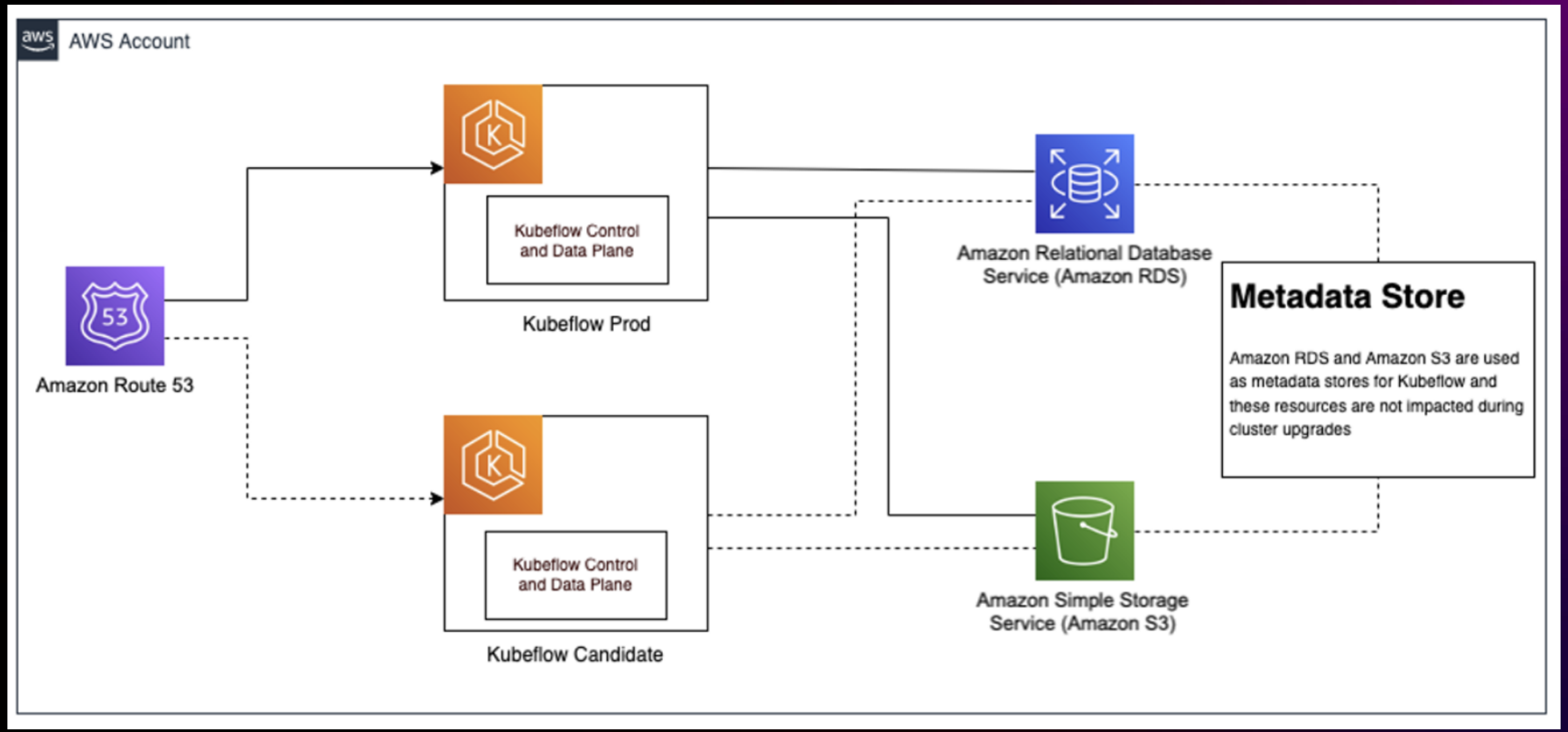
```
1  2022-11-22 20:03:10.409714: E tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit: UNKNOWN ERROR (34)
2  Kick off training.
3  Loading Dataset.
4  2022-11-22 20:03:10.409759: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running
5  2022-11-22 20:03:10.410055: I tensorflow/core/platform/cpu_feature_guard.cc:193] This TensorFlow binary is optimized with oneAPI D
6  To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
7  Traceback (most recent call last):
8    File "/opt/.venv/bin/run_train", line 6, in <module>
9      sys.exit(main())
10   File "/opt/repo/builds/train/src/train.py", line 75, in main
11     train(
12   File "/opt/repo/builds/train/src/train.py", line 40, in train
13     tensorboard = tf.keras.callbacks.TensorBoard(
14   File "/opt/venvs/misty-train-rmORz8gh-py3.8/lib/python3.8/site-packages/keras/callbacks.py", line 2225, in __init__
15     self._init_profile_batch(profile_batch)
16   File "/opt/venvs/misty-train-rmORz8gh-py3.8/lib/python3.8/site-packages/keras/callbacks.py", line 2415, in _init_profile_batch
17     profile_batch = tf.nest.map_structure(int, profile_batch)
18   File "/opt/venvs/misty-train-rmORz8gh-py3.8/lib/python3.8/site-packages/tensorflow/python/util/nest.py", line 916, in map_struct
19     structure[0], [func(*x) for x in entries],
20   File "/opt/venvs/misty-train-rmORz8gh-py3.8/lib/python3.8/site-packages/tensorflow/python/util/nest.py", line 916, in <listcomp>
21     structure[0], [func(*x) for x in entries],
22 ValueError: invalid literal for int() with base 10: 'mistake'
23
```



Victor Krylov

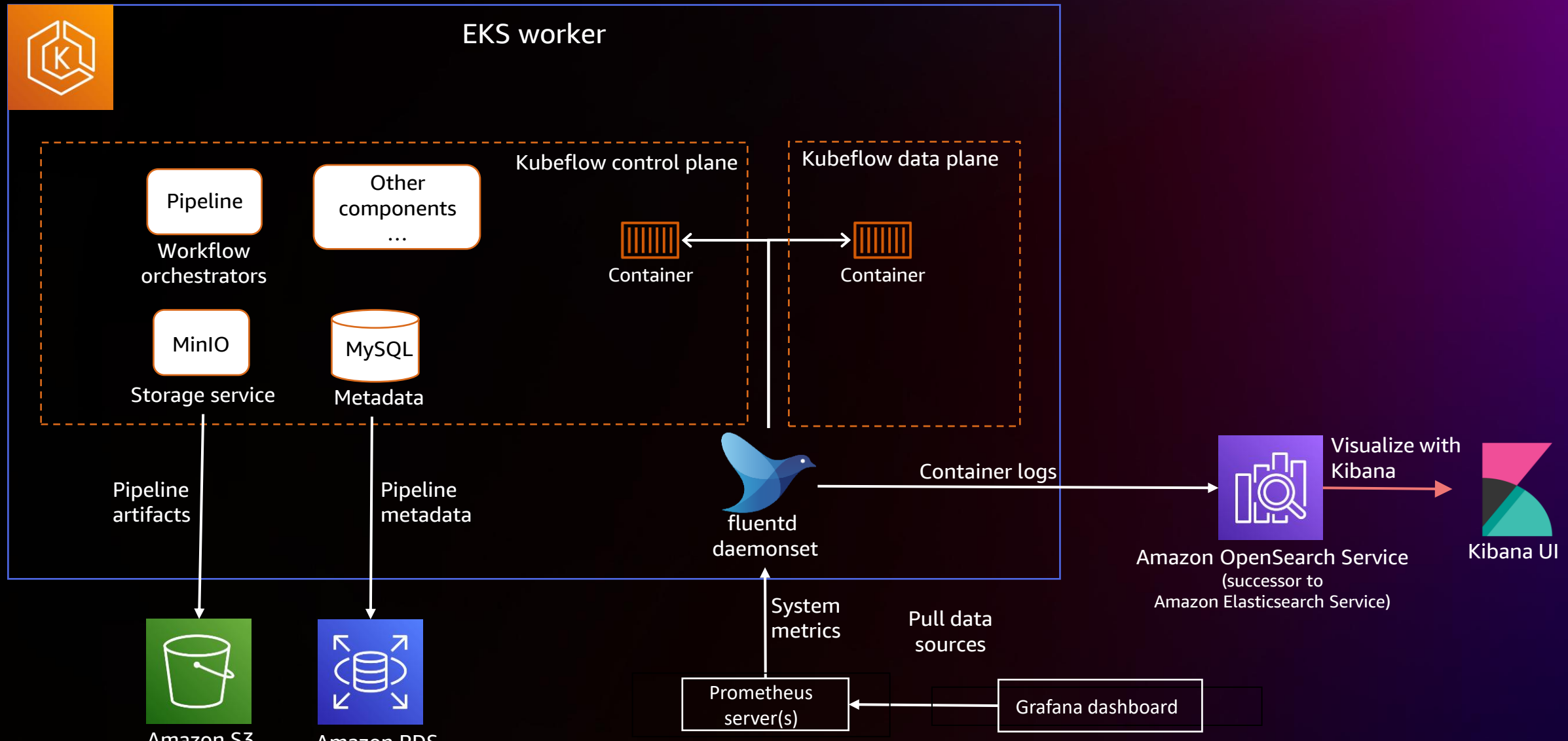
9:23 AM
11/29/2022

Safe Kubeflow cluster upgrade



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kubeflow logging



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Arindam Paul
aripauly@amazon.com

Kanwaljit Khurmi
kkhurmi@amazon.com

Victor Krylov
vkrylov@athenahealth.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.