
AWS
re:Invent

AIM313

Scaling MLOps on Kubernetes with Amazon SageMaker Operators

Alex Chung
Senior Product Manager
AWS

Srivathsan Canchi
Head of Engineering
Intuit ML Platform

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Machine Learning (ML) at Intuit

What is MLOps?

Open Source (OSS) ML and Amazon SageMaker

Integrating your own MLOps OSS platform with SageMaker

Demo

Summary



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



INTUIT powering prosperity

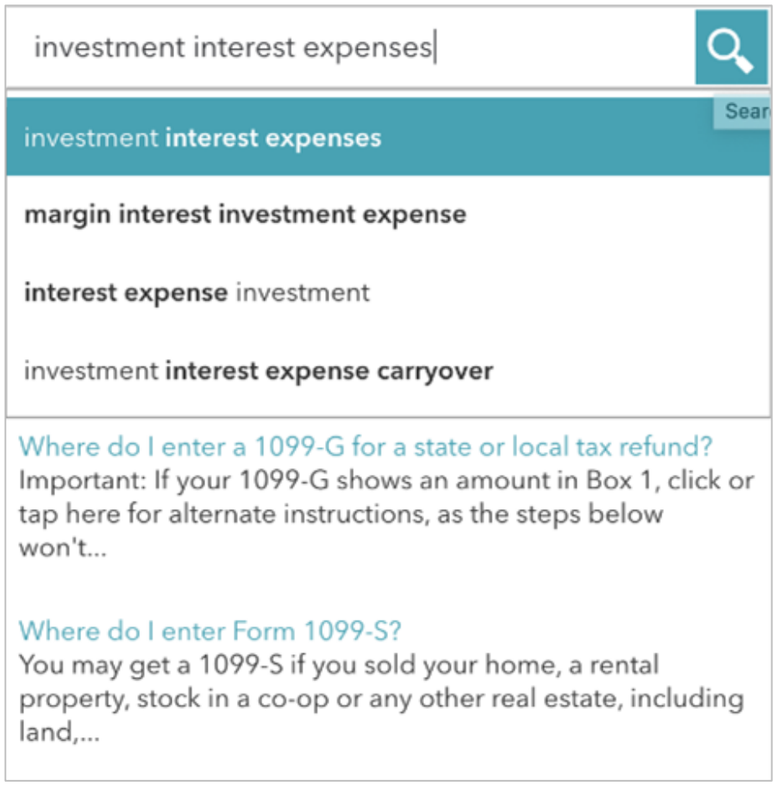
turbotax quickbooks proconnect mint

Who we serve:

Consumers
Small businesses
Self-employed

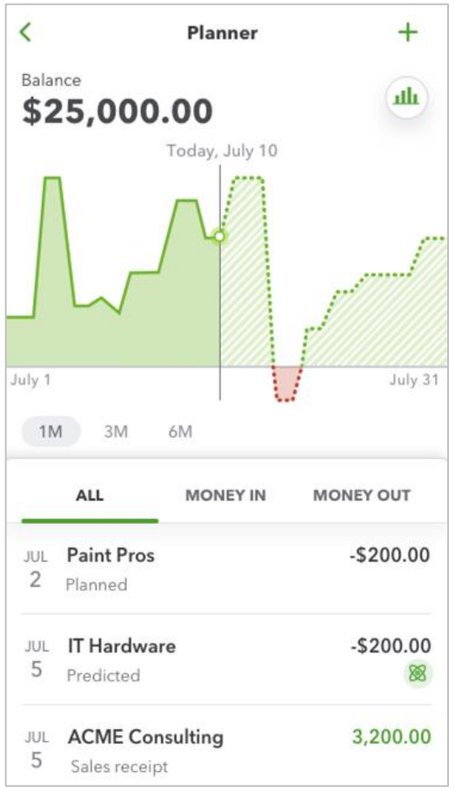
ML driven experiences

turbotax



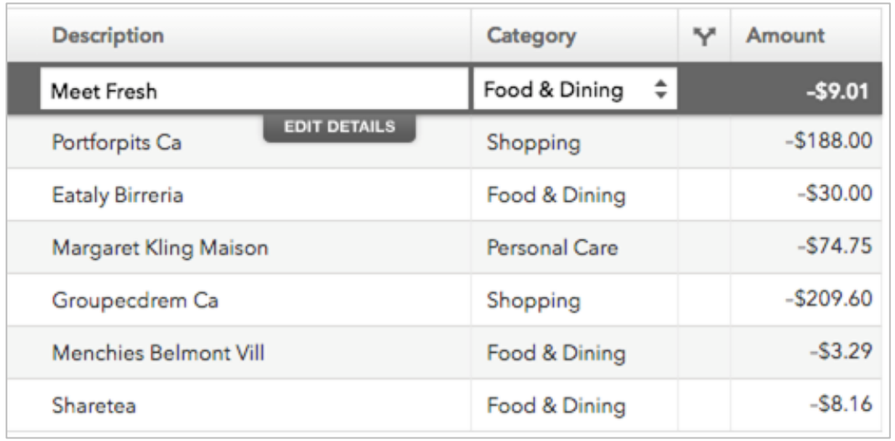
Self-help
in TurboTax

quickbooks



Forecasting cash flow
in QuickBooks

mint

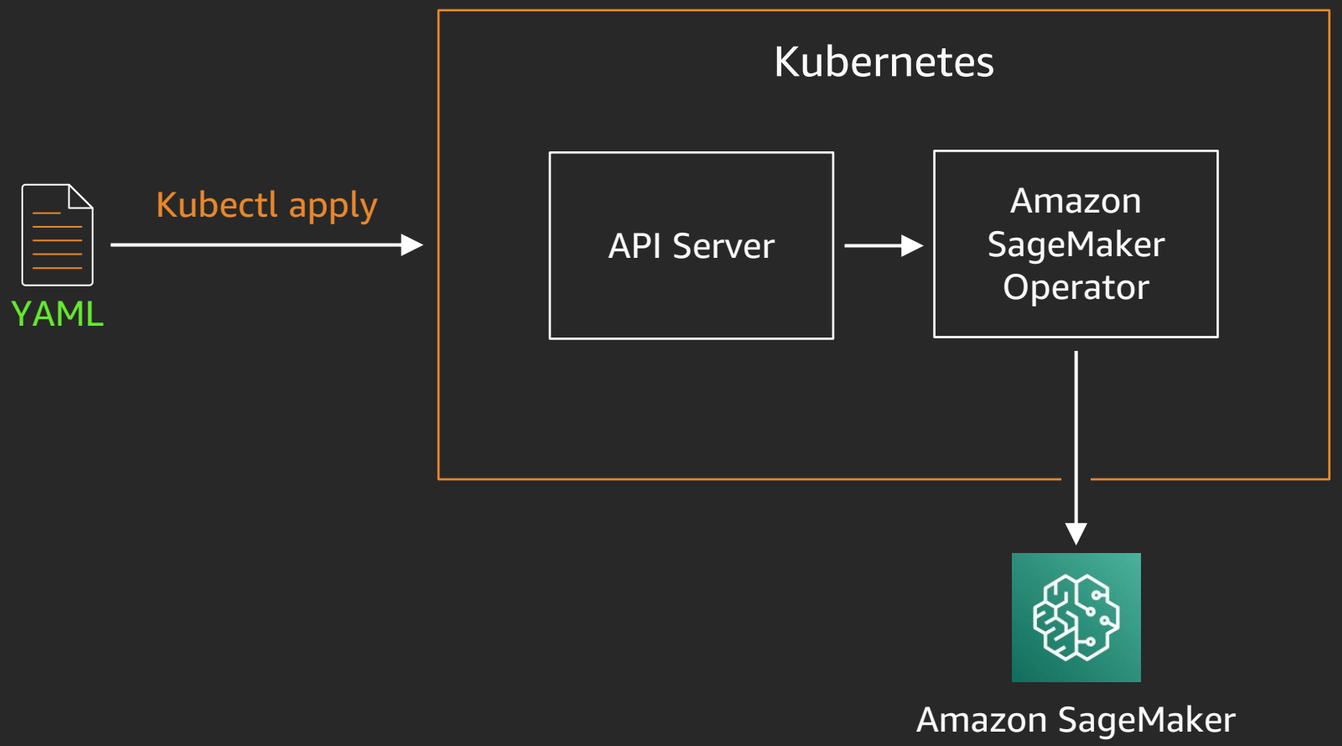


Automatic categorization
in Mint

Intuit is a leader on SageMaker and ML

- Customers of SageMaker since private Beta of SageMaker
- Beta tested nearly every feature of SageMaker
- Thought leaders and presenters at Kubecon, OpML, and other ML conferences
- Creators of open-source Argo for ML orchestration and data pipelines
- Moving to a Kubernetes native ecosystem

SageMaker Operators product



Key features

Amazon SageMaker Operators for training, tuning, inference

Natively interact with Amazon SageMaker jobs using Kubernetes tools (e.g., get pods, describe)

Stream and view logs from Amazon SageMaker in Kubernetes

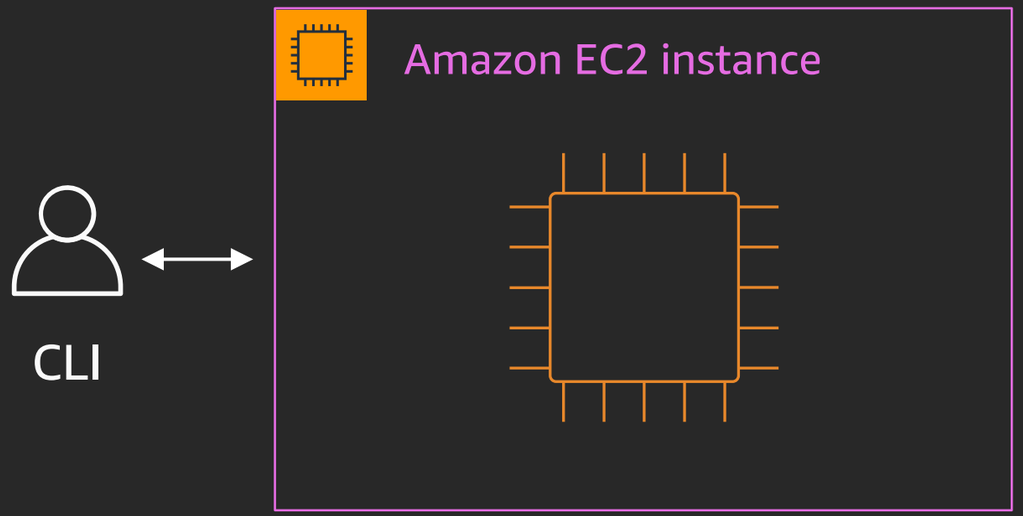
Helm Charts to assist with setup and spec creation

What is MLOps

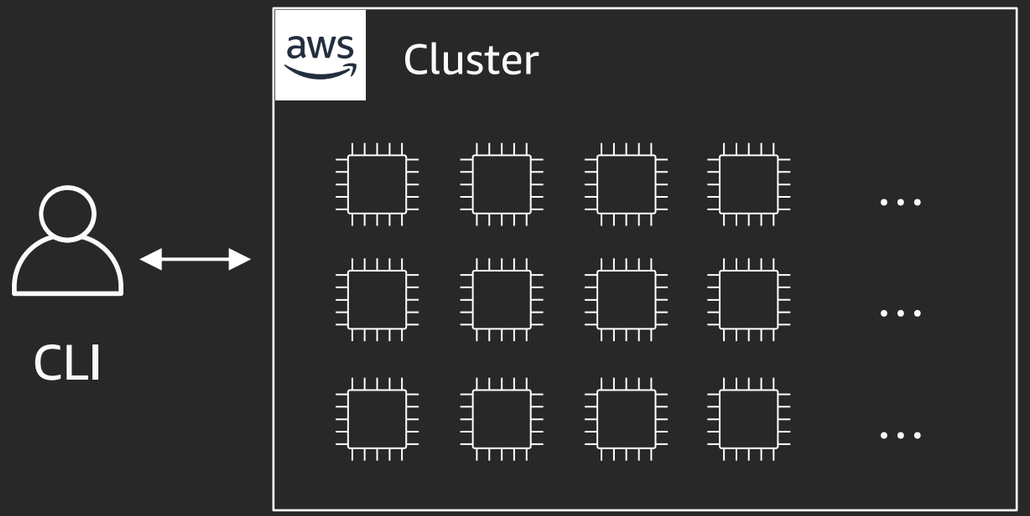
MLOps is the process of taking a model developed in a notebook or local environment and efficiently managing the deployment into production

Challenges

Scaling machine learning



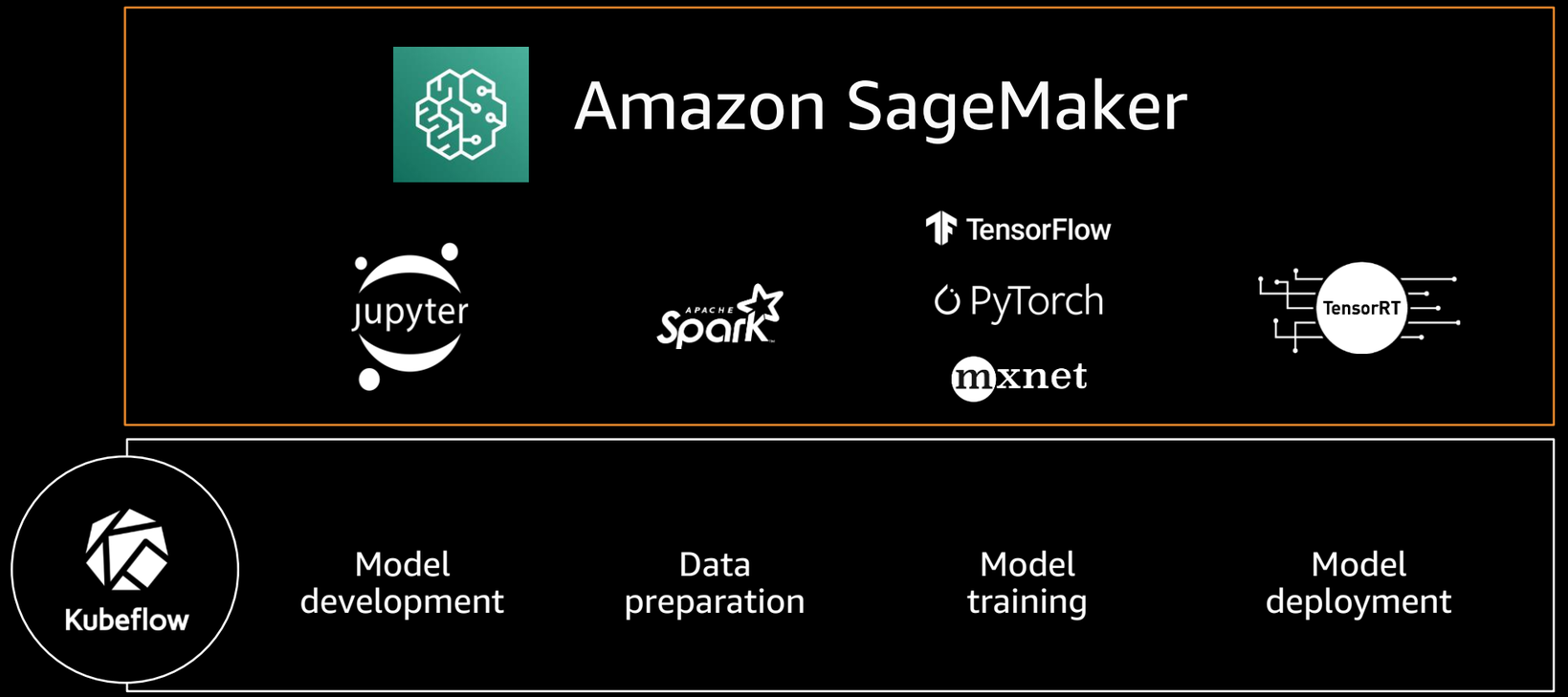
Compute (CPUs, GPUs) | Storage | Source control | ML frameworks



Scaling | Managing infrastructure | Managing pipelines

kubernetes

SageMaker + Kubeflow for machine learning



TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

OSS operators and managed compared

Pipeline Step	OSS Offering	Managed Service Offering
Feature Processing Feature Loading	Spark Operator	SageMaker Processing
Hyperparameter Optimization	Katib Controller	SageMaker HyperParameter Tuning
Model Training	PyTorch / TensorFlow Operator	SageMaker Training
Model Hosting	KFServing	SageMaker Hosting Deployment
Gang Scheduling	Training + Volcano Scheduler	SageMaker Batch Inference

SageMaker Hosting vs Kubeflow Serving

	SageMaker	KFServing
Traffic Routing	Easy	Med
Cost Transparency	✓	
AWS-Standard Security Permissions	✓	
Endpoint Tracking	✓	

```yaml
apiVersion: "serving.kubeflow.org/v1alpha2"
kind: "InferenceService"
metadata:
  name: "sklearn-iris"
spec:
  default:
    predictor:
      sklearn:
        storageUri: "gs://kfserving-samples/models/sklearn/iris"
```

```yaml
apiVersion: sagemaker.aws.amazon.com/v1
kind: HostingDeployment
metadata:
  name: hosting-deployment
spec:
  region: us-east-2
  productionVariants:
    - variantName: AllTraffic
      modelName: xgboost-model
      initialInstanceCount: 1
      instanceType: ml.r5.large
      initialVariantWeight: 1
  models:
    - name: xgboost-model
      executionRoleArn: arn:aws:iam::123456789012:role/service-role/AmazonSageMaker-ExecutionRole
      containers:
        - containerHostname: xgboost
          modelDataUrl: s3://my-bucket/xgboost/model.tar.gz
          image: 825641698319.dkr.ecr.us-east-2.amazonaws.com/xgboost:latest
```

Auto scaling inference endpoints

	SageMaker	KFServing
Granular Options	✓	
Spec Transparency	✓	
AWS-Standard Conventions	✓	

```yaml
apiVersion: "serving.kubeflow.org/v1alpha2"
kind: "InferenceService"
metadata:
  name: "flowers-sample"
  annotations:
    autoscaling.knative.dev/target: "10"
spec:
  default:
    predictor:
      tensorflow:
        storageUri: "gs://kfserving-samples/models/tensorflow/flowers"
```

```yaml
apiVersion: sagemaker.aws.amazon.com/v1
kind: HostingAutoscalingPolicy
metadata:
  name: hap-custom
spec:
    resourceId:
      - endpointName: Endpoint-20200410010239-FZFY
        variantName: variant-name-1
    region: us-east-2
    policyName: SageMakerEndpointInvocationScalingPolicy
    policyType: TargetTrackingScaling
    minCapacity: 1
    maxCapacity: 2
    suspendedState:
        dynamicScalingInSuspended: true
    targetTrackingScalingPolicyConfiguration:
        targetValue: 60.0
        scaleInCooldown: 700
        scaleOutCooldown: 300
        customizedMetricSpecification:
          metricName: CPUUtilization
          namespace: /aws/sagemaker/Endpoints
          dimensions:
            - name: EndpointName
              value: Endpoint-20200410010239-FZFY
            - name: VariantName
              value: variant-name-1
          statistic: Average
          unit: Percent
```

SageMaker for better open source

MODEL MONITOR AND MODEL DRIFT

- Baseline training data and measure new inference endpoints against existing endpoints
- Kick off retraining through alarming in Amazon CloudWatch and AWS Lambda
- Track lineage of all models created using SageMaker Experiment



Amazon SageMaker

Fully-managed infrastructure	Ground Truth labeling	Automatic model turning	Built-in optimized algorithms
Managed Spot Training	Scalable inference endpoints	Model monitoring	Easy-to-do scalability

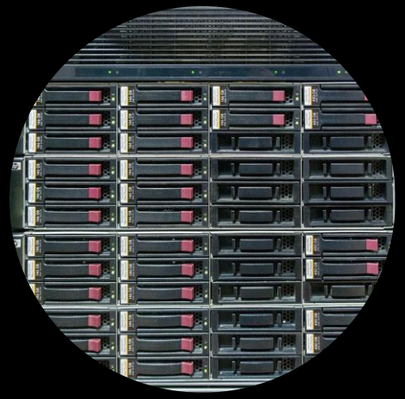
Amazon SageMaker
accessible from Kubernetes

Native Kubernetes

Hybrid infrastructures	Reuse existing infrastructure	Portability	Composability



Intuit's Machine Learning Platform



400+

ML Models

25+

Trainings / day



8000+

Features

15B

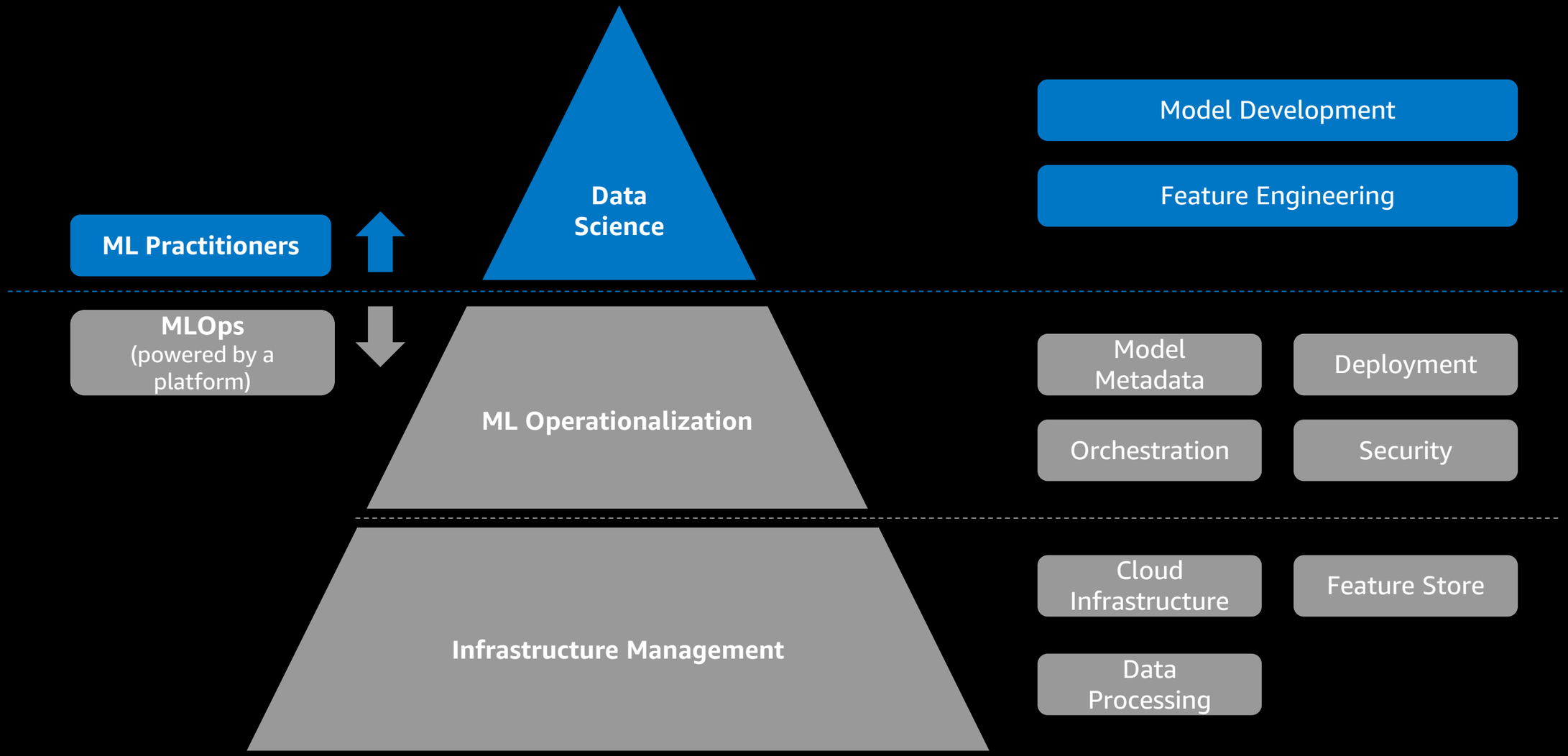
Feature updates / day



ML Tech

SageMaker
Kubernetes
Argo Workflows
Spark
Beam

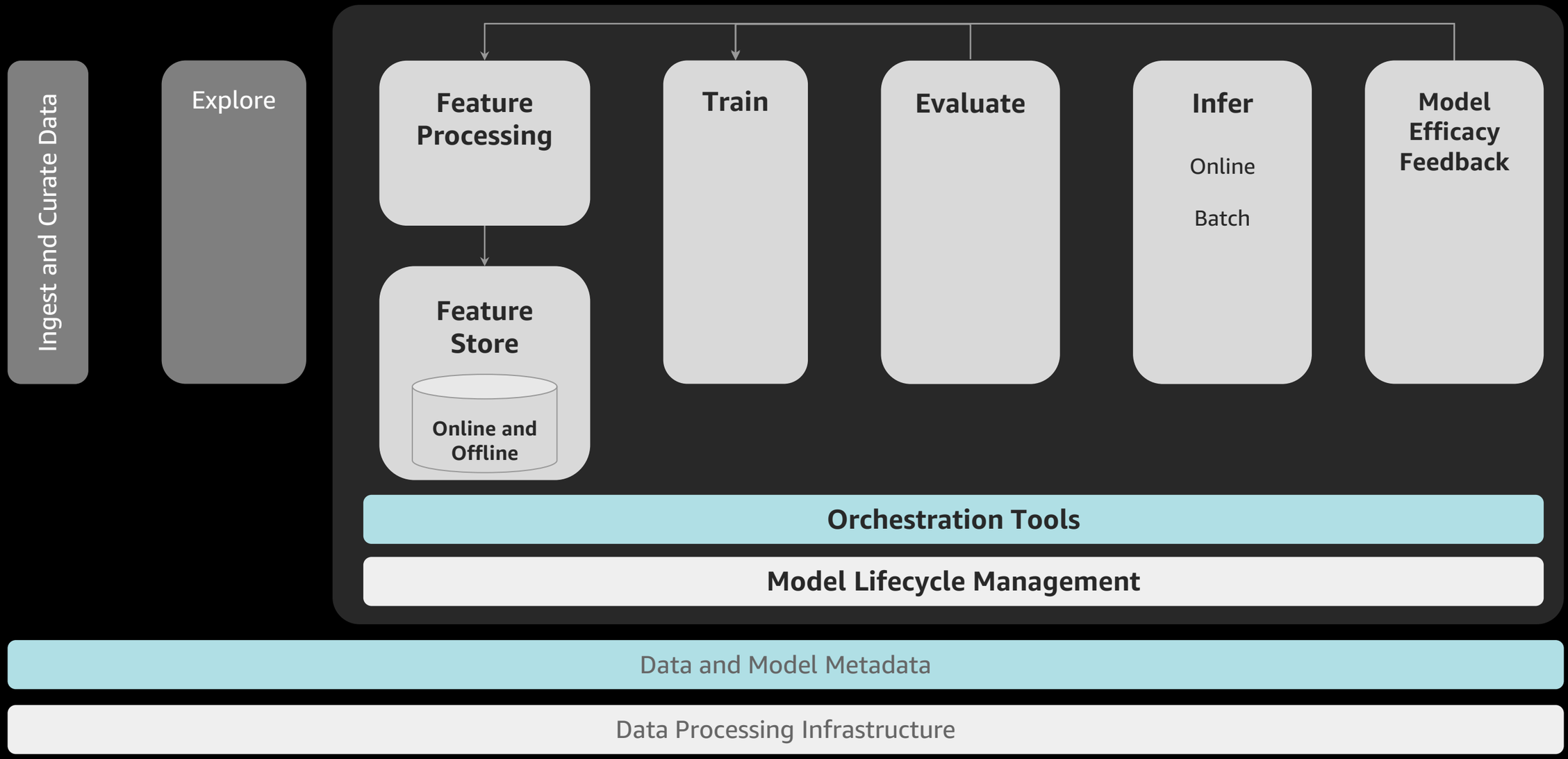
How can ML practitioners focus on their craft?



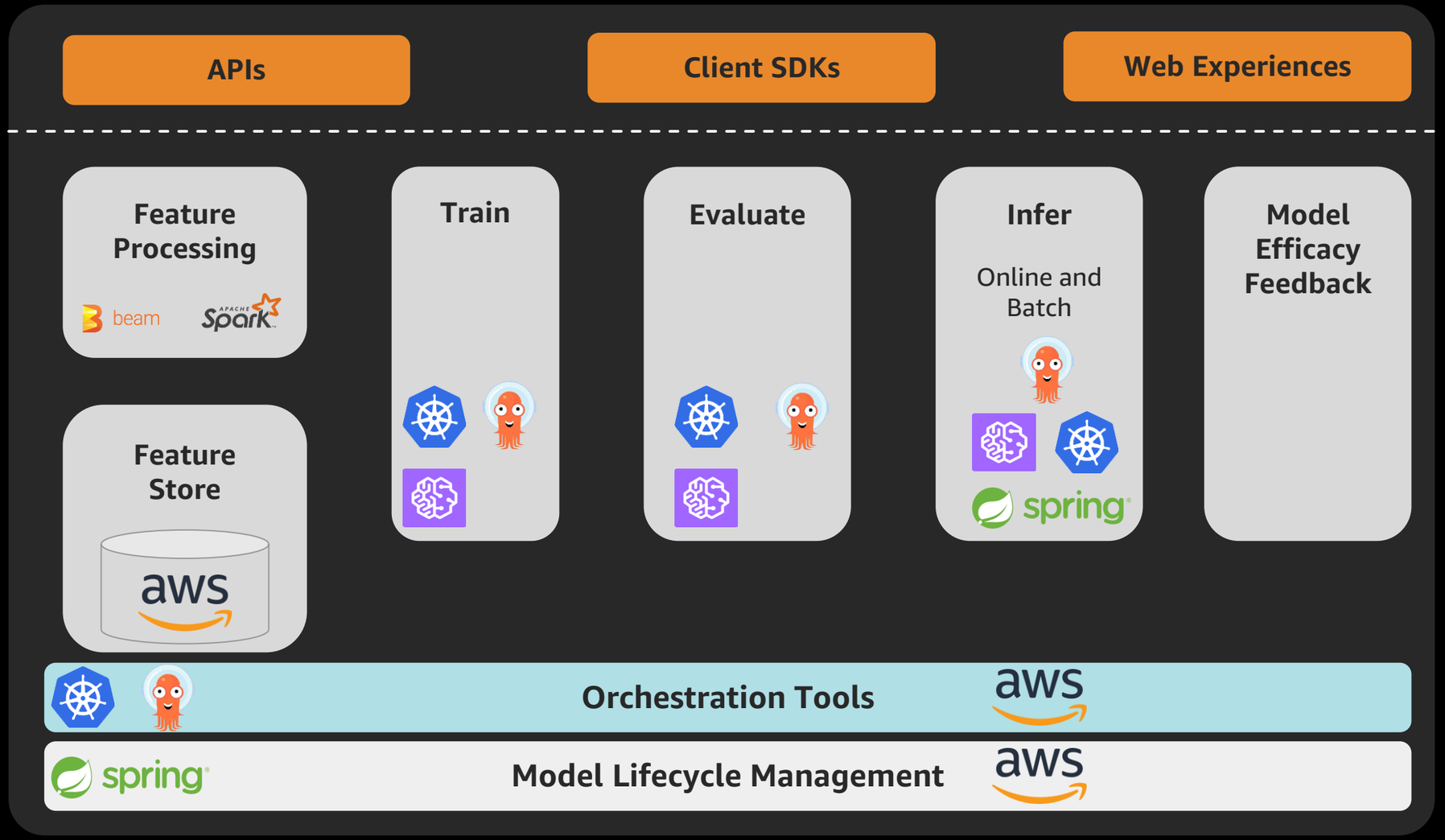
Enterprise challenges without MLOps

- Cloud infrastructure management
- Enterprise security and compliance
- Observability
- Interfaces: *Web*, *API*, *SDK,* and *CLI*
- Burden is on the data scientists to maintain and manage these touchpoints
- How can we simplify MLOps?

Simplifying MLOps: Intuit's ML Platform



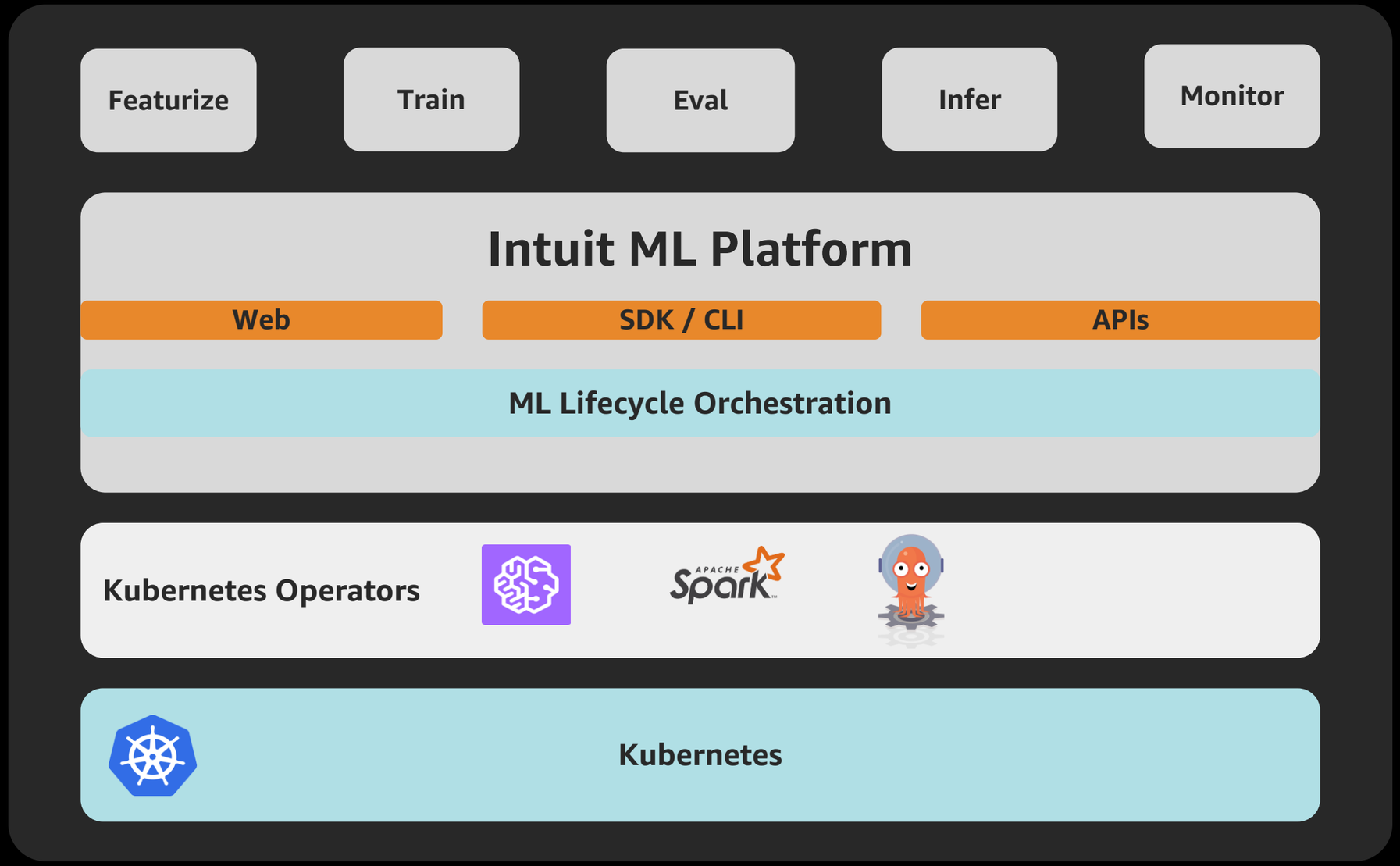
Lot of technologies at play



Can the platform be made better?

- Intuit has a heavy investment in Kubernetes
  - Intuit's Kubernetes Service
  - Argo Workflows / CD
- Kubernetes Operators are everywhere
  - Argo
  - SageMaker
  - Spark
- Can this ecosystem be leveraged for offering a unique interface for data scientists?

Kubernetes native MLOps



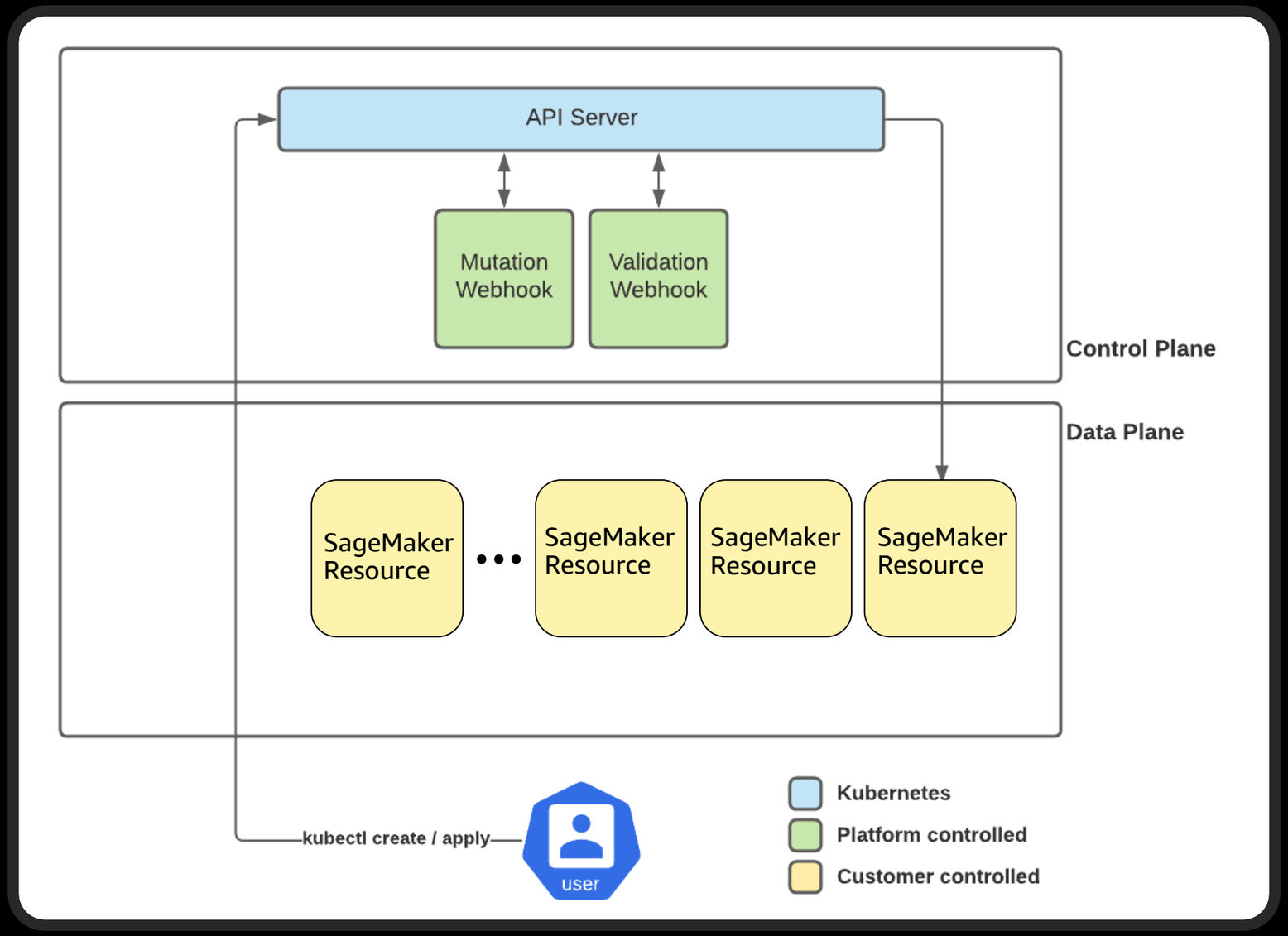
Kubernetes native MLOps

DECLARATIVELY EXPRESSING ML WORKFLOWS



Kubernetes native MLOps

LEVERAGING KUBERNETES PROGRAMMING MODEL



Kubernetes native MLOps

KUBERNETES MANIFESTS FOR INTERFACING WITH OPERATORS

```yaml
apiVersion: sagemaker.aws.amazon.com/v1
kind: HostingDeployment
metadata:
  name: creditcard-approval
spec:
  region: us-west-2
  productionVariants:
  - variantName: AllTraffic
    modelName: creditcard-approval
    initialInstanceCount: 1
    initialVariantWeight: 1
    instanceType: ml.t2.medium
  models:
  - name: creditcard-approval
    executionRoleArn: arn:aws:iam::238251263086:role/sagemaker-default
    containers:
    - containerHostname: creditcard-approval
      modelDataUrl: s3://
      image:
```

Re-imagined data scientist experience

1. Every operation in the model lifecycle is represented in code (gitOps)
2. It is a cloud-native pattern
   - Use tools such as Argo CD to control ML pipelines
   - Kubectl for local and pre-prod testing
3. One place for everything
   - Building a model dashboard
   - Control actions, status, logs, and metrics all in one place
4. Easier maintainability
   - New releases of the underlying technologies get incorporated easily

intuit.

Credits

Amazon SageMaker

Jason Gelman
Kartik Kalamadi
Kevin Mould
Meghna Baijal
Nicholas Thomson
Suraj Kota
Roberto Bruno Martins

Intuit

Nagaraj Janardhana
David Farr
Kevin Marquardsen
Ian Sebanja
Rachit Chauhan
Tobias Wenzel
Deepen Mehta

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.