INTUIT turbotax creditkarma quickbooks mailchimp

Improving Service Availability

Scaling Ahead with Machine Learning for HPA Optimization

Estela Ramirez
Software Engineer, Intuit Kubernetes Service

Avni Sharma
Sr Product Manager, Intuit Kubernetes Service

Agenda

Background

Problem statement

Proposed solution

Demo

Learnings

QnA

Technology @ Intuit

Intuit is leading the way in building an AI-native development platform using cloud native open source technology. We're committed to building tools that scale and giving back to the open source community.

INTUIT turbotax creditkarma quickbooks mailchimp

~100M
customers

107B
consumer tax refunds per year

$2T+
invoices managed on our platform per year

18M
total US workers paid via QB payroll

INTUIT

KubeCon | CloudNativeCon
North America 2024

AI-native development platform

AI-powered app experiences

4M

Models running in production per day

AI-assisted development: coding, testing, debugging

8x

Developer velocity increase in past four years

AI-powered app centric runtime

60B

Machine learning predictions per day

Smart operations using AIOps

40M+

AIOps inferences/day

Problem statement

Problem statement

1

App + k8s can not scale up fast enough

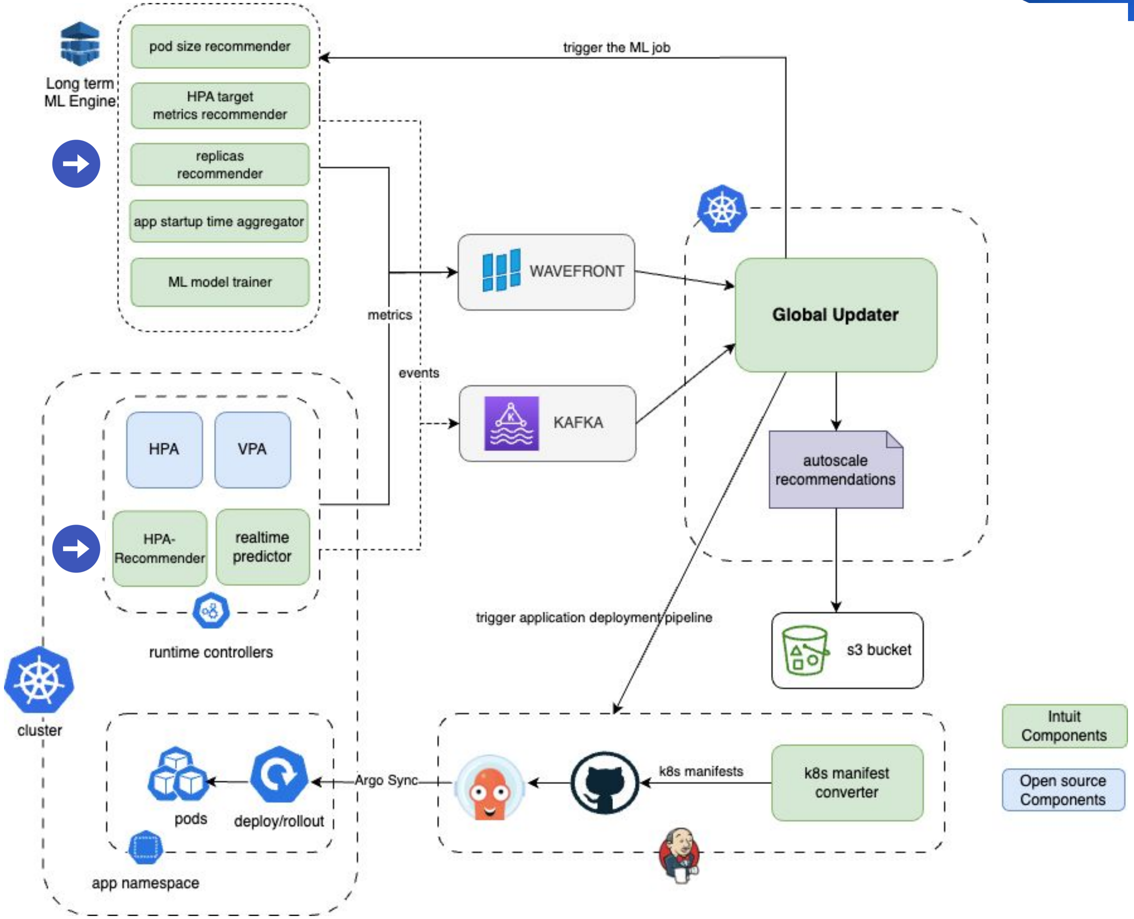
- High pod startup time
- Node scaling up take ~5m
- Image pulling
- 5xx errors, the app can't handle the surging traffic

2

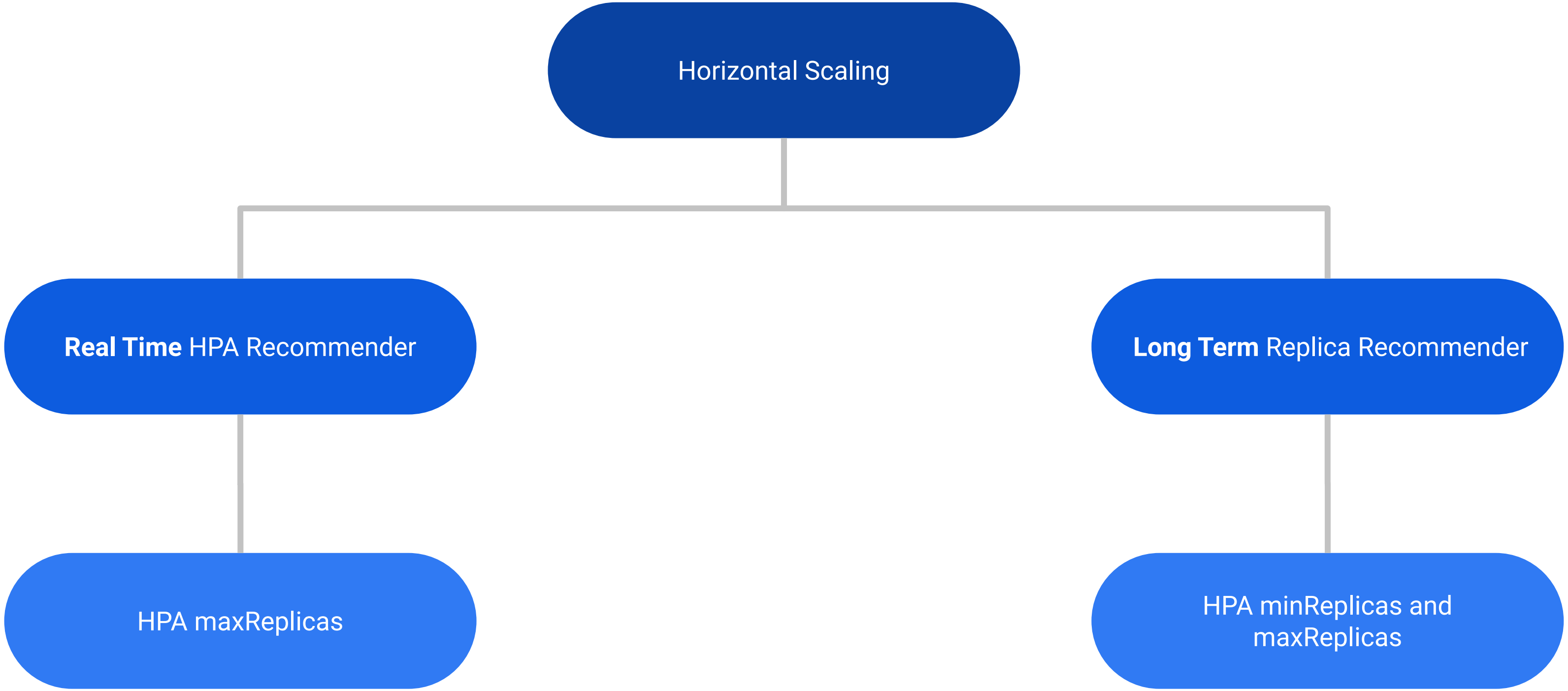
Developers have difficulty in right sizing the application and tune HPA settings

- Too conservative HPA minReplica and pod size
- Tune HPA
- High cost

How the new auto scaling system works



Horizontal scaling


Horizontal Scaling

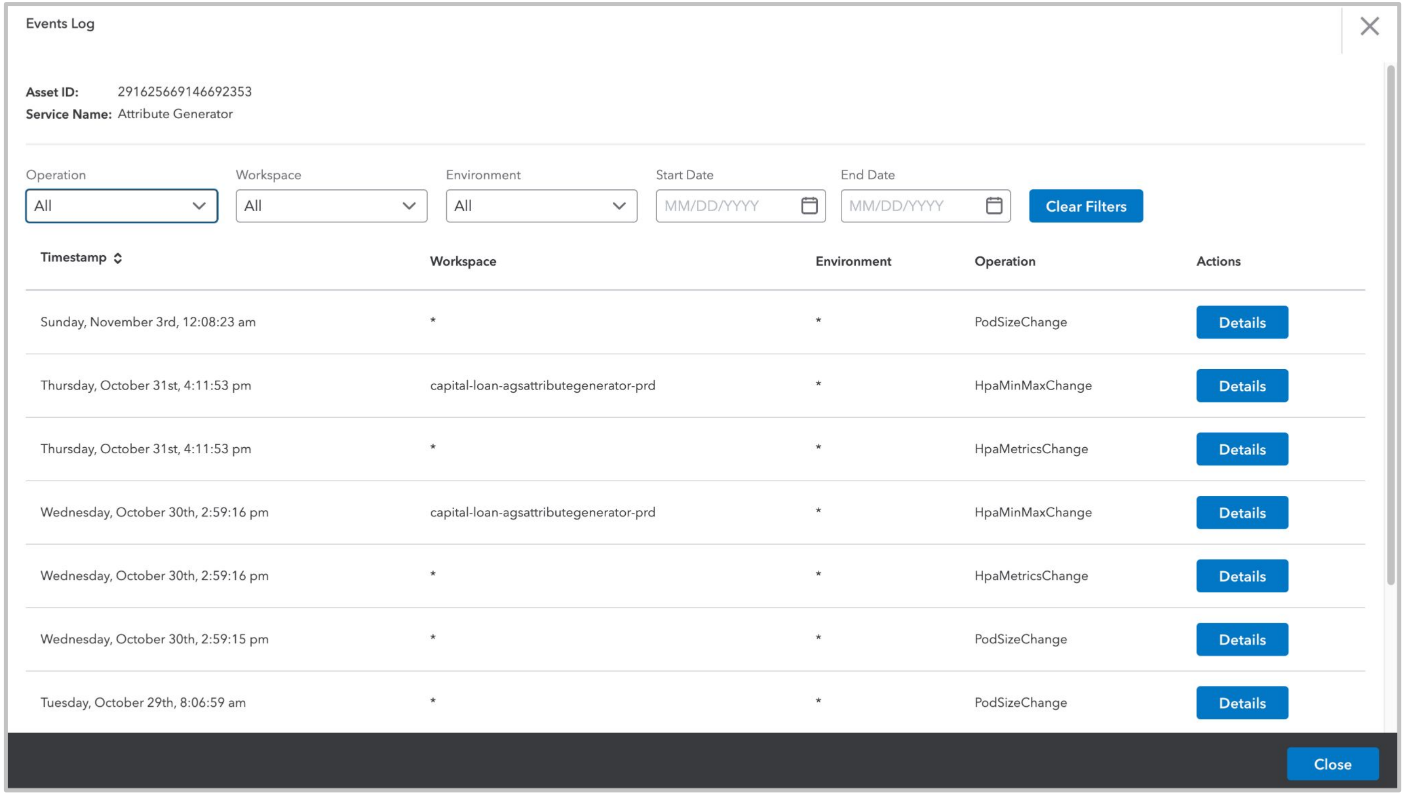
Real Time HPA Recommender

Long Term Replica Recommender

HPA maxReplicas

HPA minReplicas and maxReplicas

DeveloperPortal view

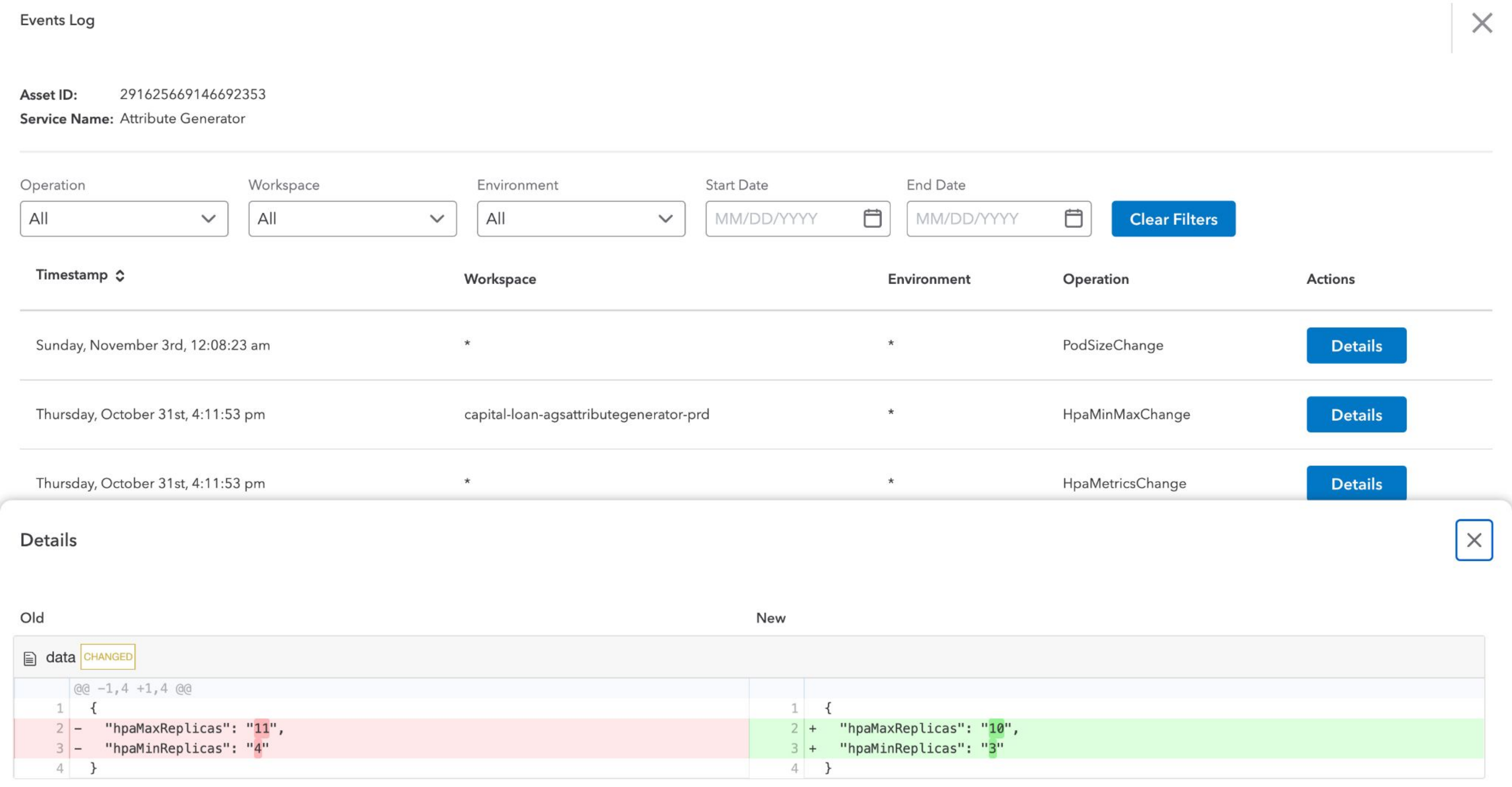


Events Log

Asset ID: 291625669146692353
Service Name: Attribute Generator

Operation: All
Workspace: All
Environment: All
Start Date: MM/DD/YYYY
End Date: MM/DD/YYYY

Timestamp	Workspace	Environment	Operation	Actions
Sunday, November 3rd, 12:08:23 am	*	*	PodSizeChange	Details
Thursday, October 31st, 4:11:53 pm	capital-loan-agsattributegenerator-prd	*	HpaMinMaxChange	Details
Thursday, October 31st, 4:11:53 pm	*	*	HpaMetricsChange	Details
Wednesday, October 30th, 2:59:16 pm	capital-loan-agsattributegenerator-prd	*	HpaMinMaxChange	Details
Wednesday, October 30th, 2:59:16 pm	*	*	HpaMetricsChange	Details
Wednesday, October 30th, 2:59:15 pm	*	*	PodSizeChange	Details
Tuesday, October 29th, 8:06:59 am	*	*	PodSizeChange	Details



Events Log

Asset ID: 291625669146692353
Service Name: Attribute Generator

Operation: All
Workspace: All
Environment: All
Start Date: MM/DD/YYYY
End Date: MM/DD/YYYY

Timestamp	Workspace	Environment	Operation	Actions
Sunday, November 3rd, 12:08:23 am	*	*	PodSizeChange	Details
Thursday, October 31st, 4:11:53 pm	capital-loan-agsattributegenerator-prd	*	HpaMinMaxChange	Details
Thursday, October 31st, 4:11:53 pm	*	*	HpaMetricsChange	Details

Details

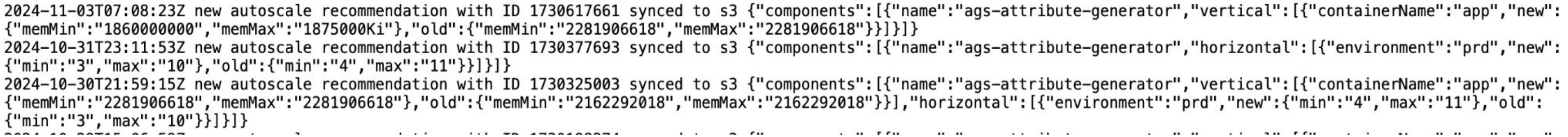
Old

```
data CHANGED
@@ -1,4 +1,4 @@
 {
-   "hpaMaxReplicas": "11",
-   "hpaMinReplicas": "4"
 }
```

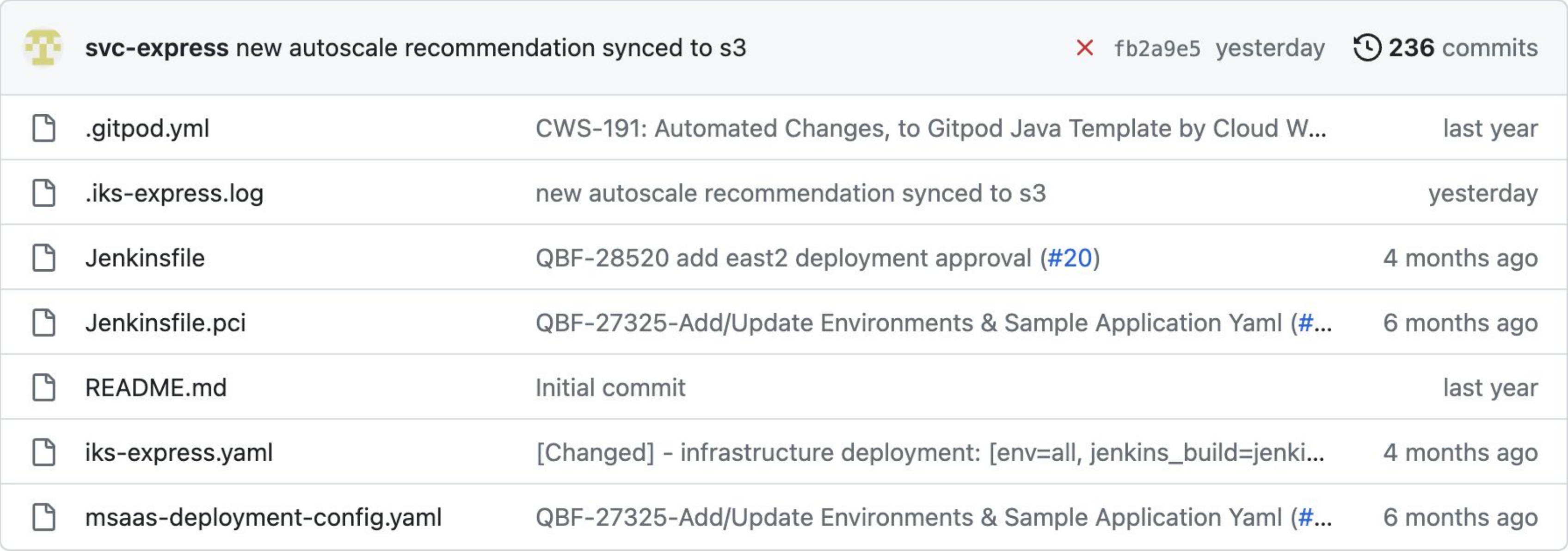
New

```
 {
+   "hpaMaxReplicas": "10",
+   "hpaMinReplicas": "3"
 }
```

Recommendation



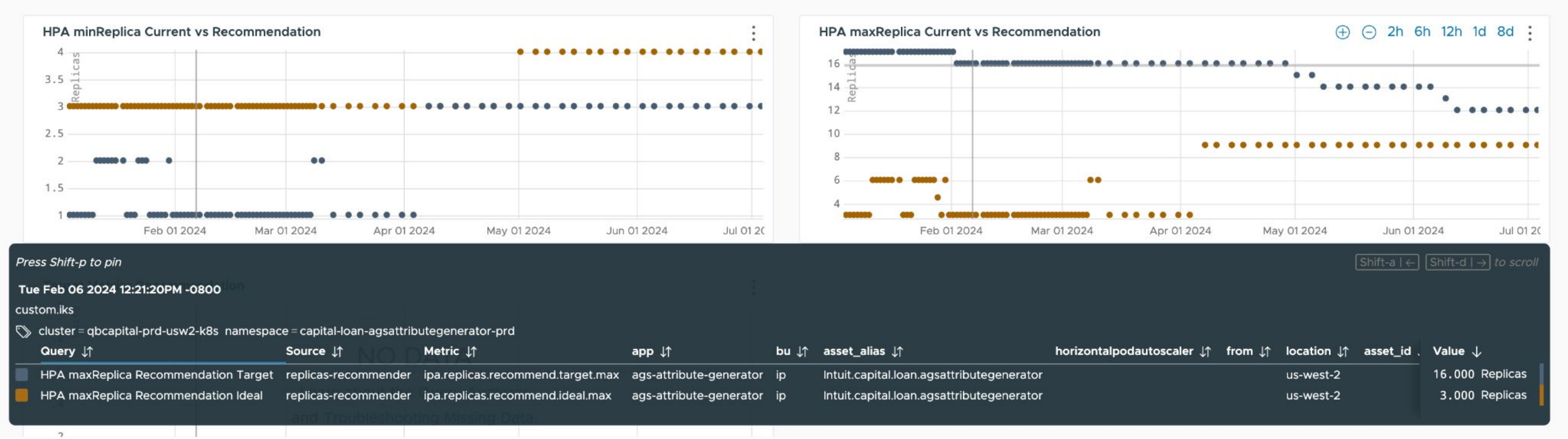
```
2024-11-03T07:08:23Z new autoscale recommendation with ID 1730617661 synced to s3 {"components":[{"name":"ags-attribute-generator","vertical":[{"containerName":"app","new":
{"memMin":"1860000000","memMax":"1875000Ki"},"old":{"memMin":"2281906618","memMax":"2281906618"}}]}]}
2024-10-31T23:11:53Z new autoscale recommendation with ID 1730377693 synced to s3 {"components":[{"name":"ags-attribute-generator","horizontal":[{"environment":"prd","new":
{"min":"3","max":"10"},"old":{"min":"4","max":"11"}}]}]}
2024-10-30T21:59:15Z new autoscale recommendation with ID 1730325003 synced to s3 {"components":[{"name":"ags-attribute-generator","vertical":[{"containerName":"app","new":
{"memMin":"2281906618","memMax":"2281906618"},"old":{"memMin":"2162292018","memMax":"2162292018"}}],"horizontal":[{"environment":"prd","new":{"min":"4","max":"11"},"old":
{"min":"3","max":"10"}}]}]}
```



svc-express new autoscale recommendation synced to s3 ✕ fb2a9e5 yesterday **236** commits

.gitpod.yml	CWS-191: Automated Changes, to Gitpod Java Template by Cloud W...	last year
.iks-express.log	new autoscale recommendation synced to s3	yesterday
Jenkinsfile	QBF-28520 add east2 deployment approval (#20)	4 months ago
Jenkinsfile.pci	QBF-27325-Add/Update Environments & Sample Application Yaml (#...	6 months ago
README.md	Initial commit	last year
iks-express.yaml	[Changed] - infrastructure deployment: [env=all, jenkins_build=jenki...	4 months ago
msaas-deployment-config.yaml	QBF-27325-Add/Update Environments & Sample Application Yaml (#...	6 months ago

Recommendation over time



Press Shift-p to pin

Shift-a | ← Shift-d | → *to scroll*

Tue Feb 06 2024 12:21:20PM -0800

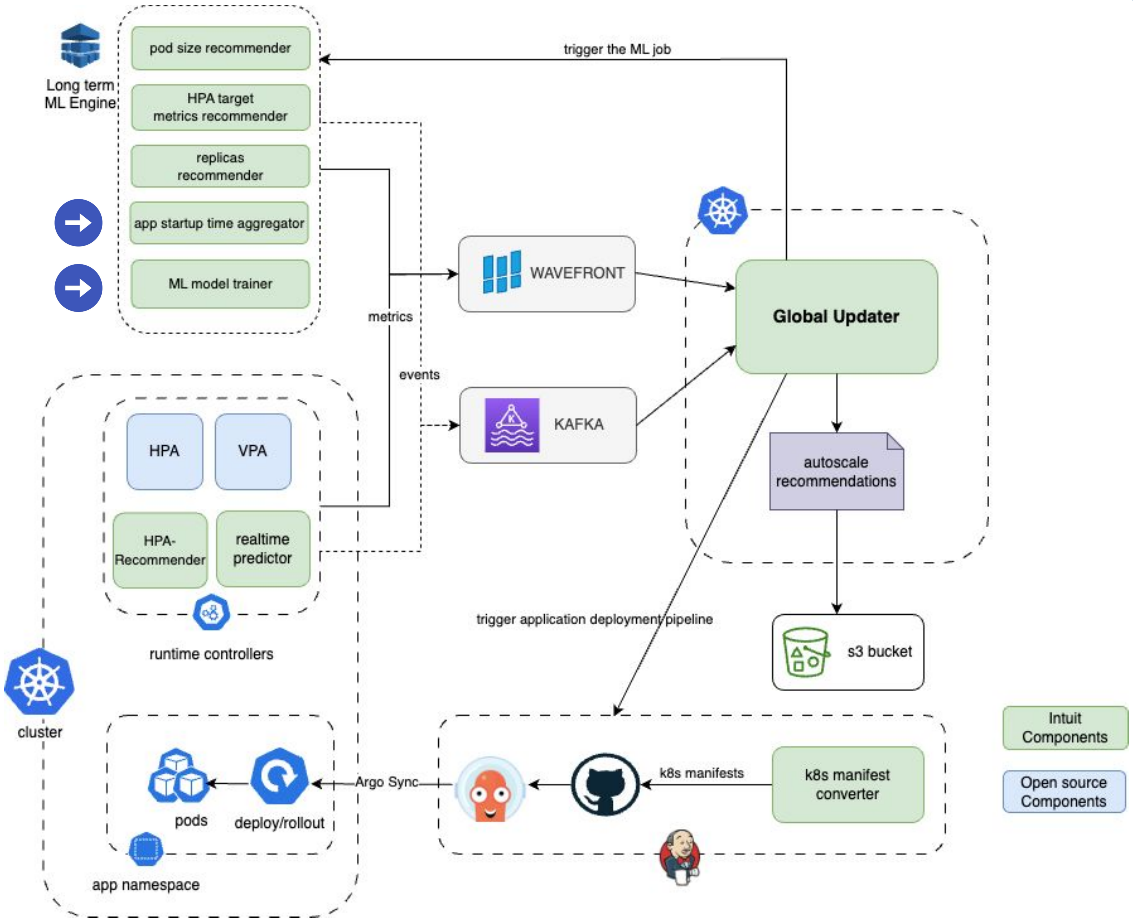
custom.iks

cluster = qbcapital-prd-usw2-k8s namespace = capital-loan-agsattributegenerator-prd

	Query ⇅	Source ⇅	Metric ⇅	app ⇅	bu ⇅	asset_alias ⇅	horizontalpodautoscaler ⇅	from ⇅	location ⇅	asset_id	Value ↓
■	HPA maxReplica Recommendation Target	replicas-recommender	ipa.replicas.recommend.target.max	ags-attribute-generator	ip	Intuit.capital.loan.agsattributegenerator			us-west-2		16.000 Replicas
■	HPA maxReplica Recommendation Ideal	replicas-recommender	ipa.replicas.recommend.ideal.max	ags-attribute-generator	ip	Intuit.capital.loan.agsattributegenerator			us-west-2		3.000 Replicas

Conservative, with a bias towards availability and correctness

How the new auto scaling system works

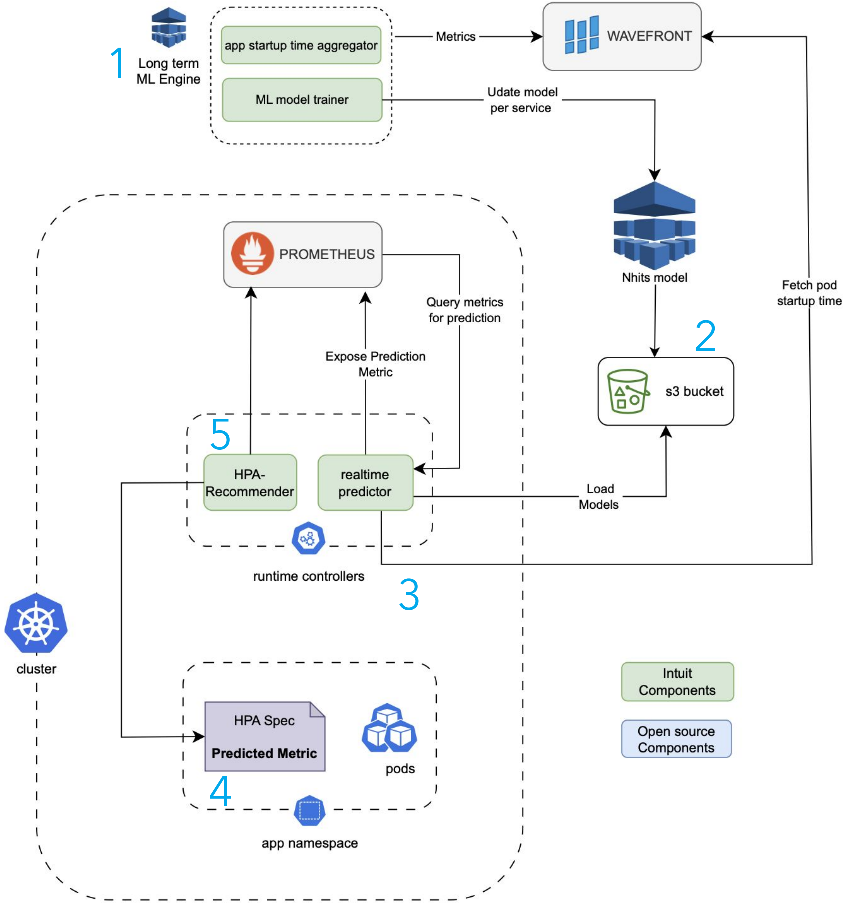


Proposed solution

Time series forecasting for real-time cpu usage prediction

- Integrate predicted metrics with Prometheus
- Use predicted metrics as custom metric in HPA
- Proactively increases desired replicas x minutes ahead of time, based on pod start up time

Autoscaling ahead of time



- ML Model Training
- Exposing Prometheus Metric
- Use Custom metric in HPA

Models under consideration

We evaluated 4 separate time series forecasting models

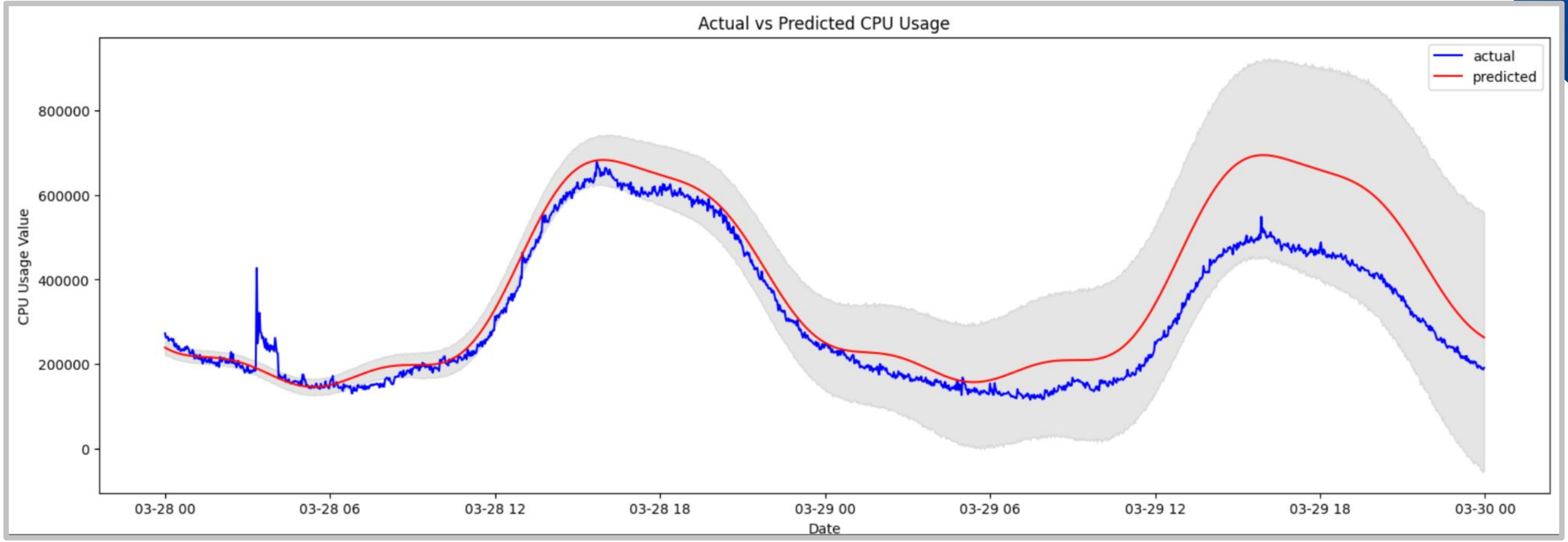
Prophet model

TimesFM forecasting model

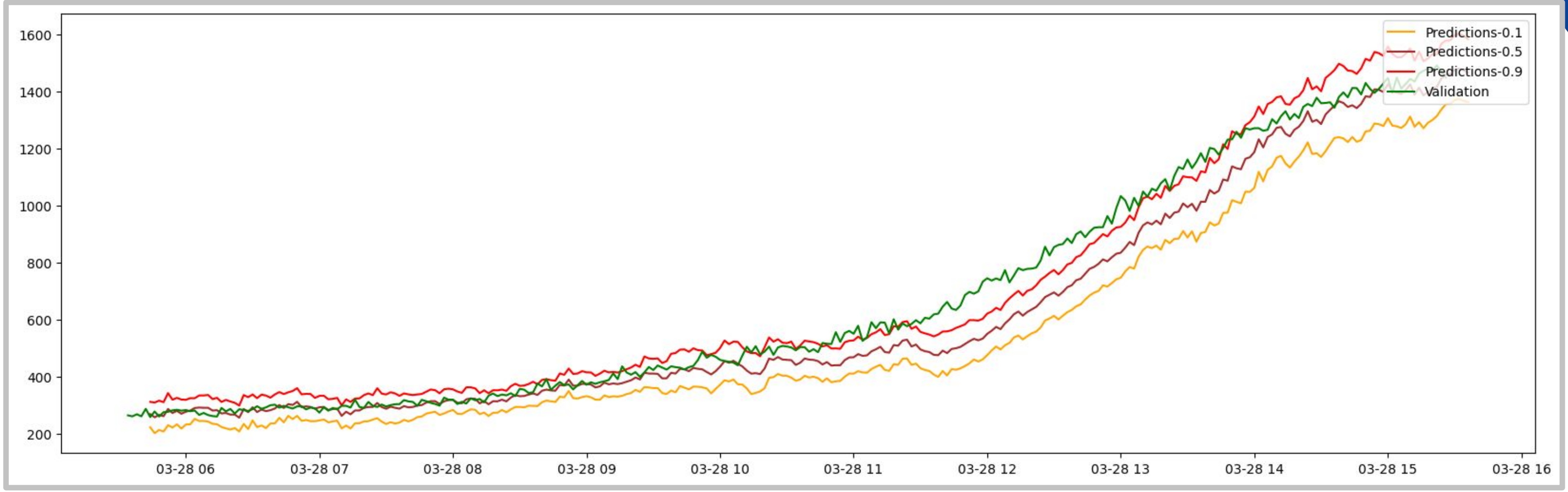
In-**House GRU** model

Nhits forecasting model

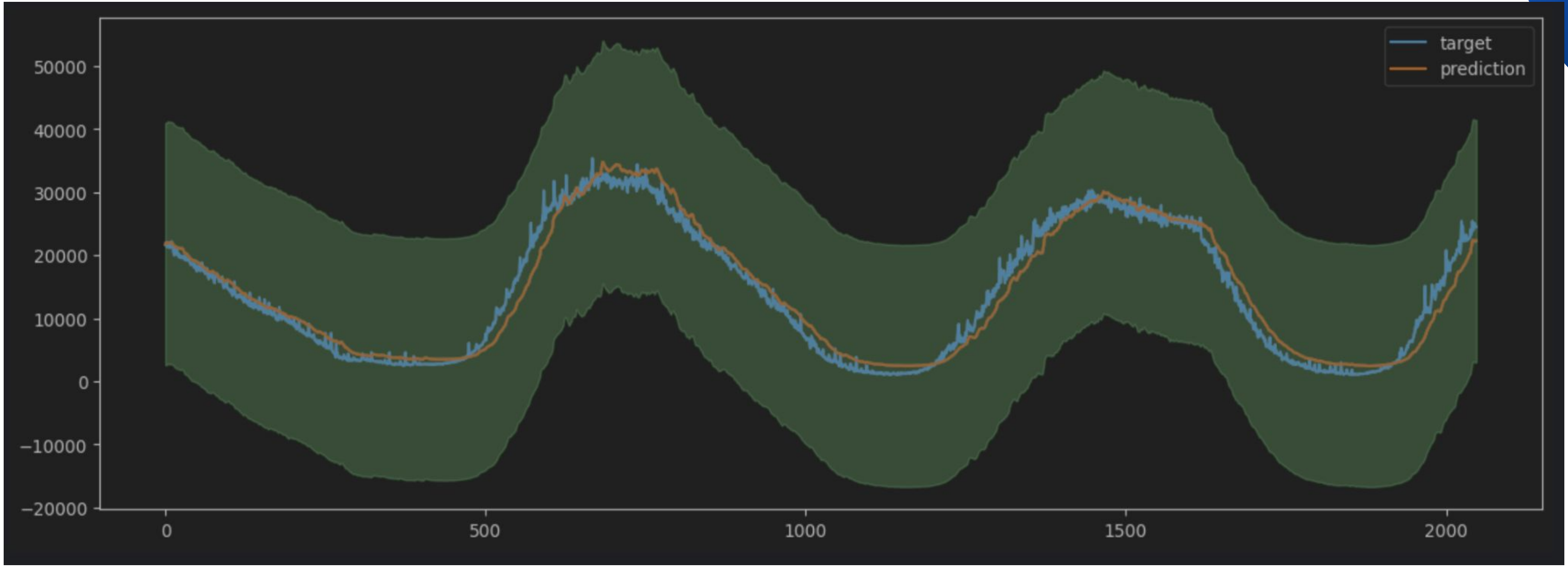
Prophet model



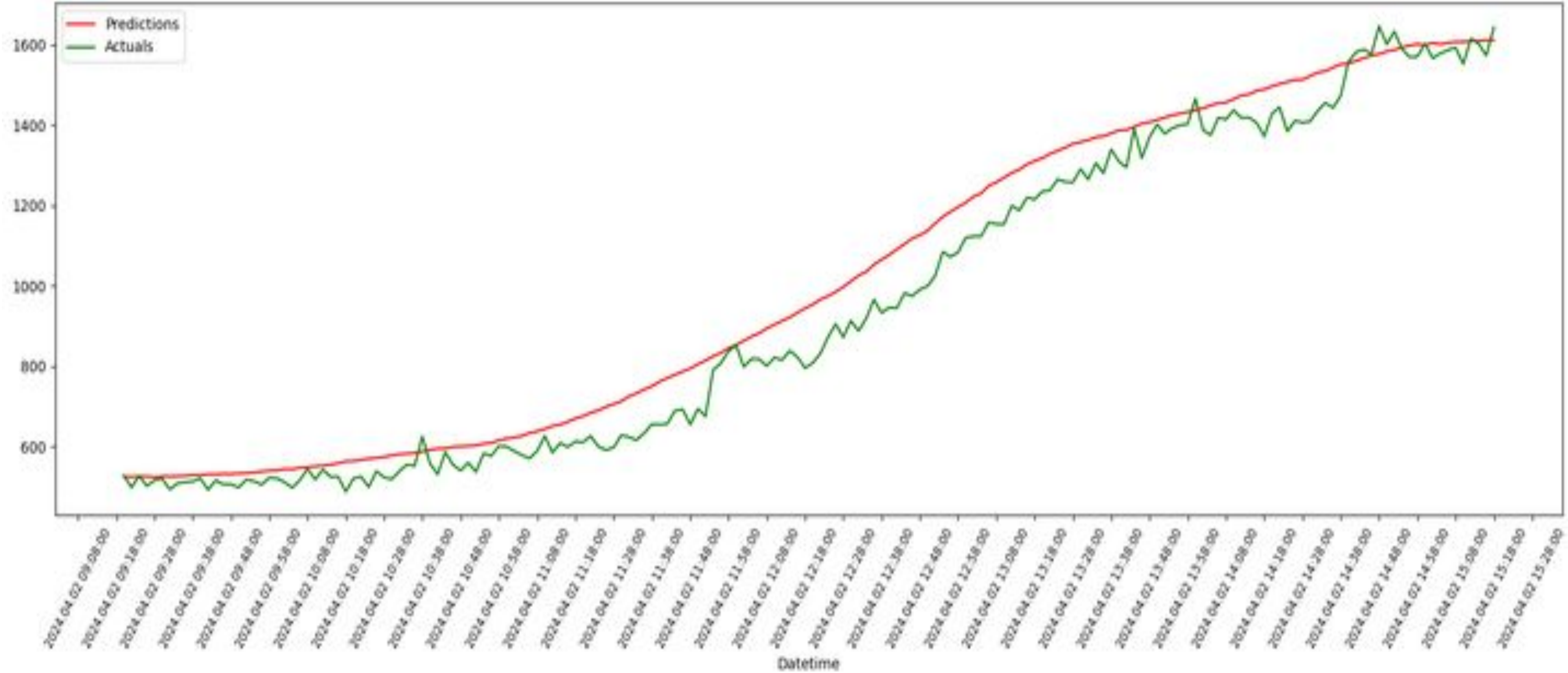
Google's TimesFM model



In house RNN forecasting model

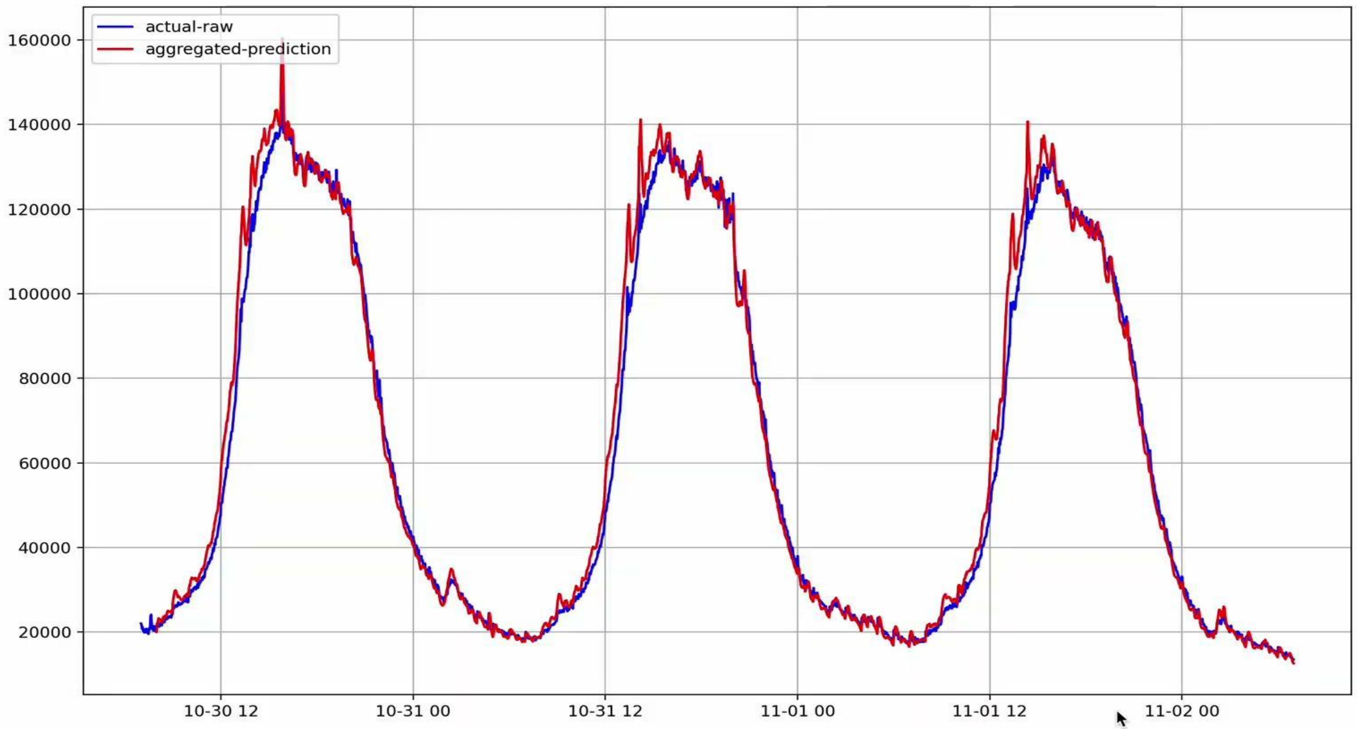


Nhits model

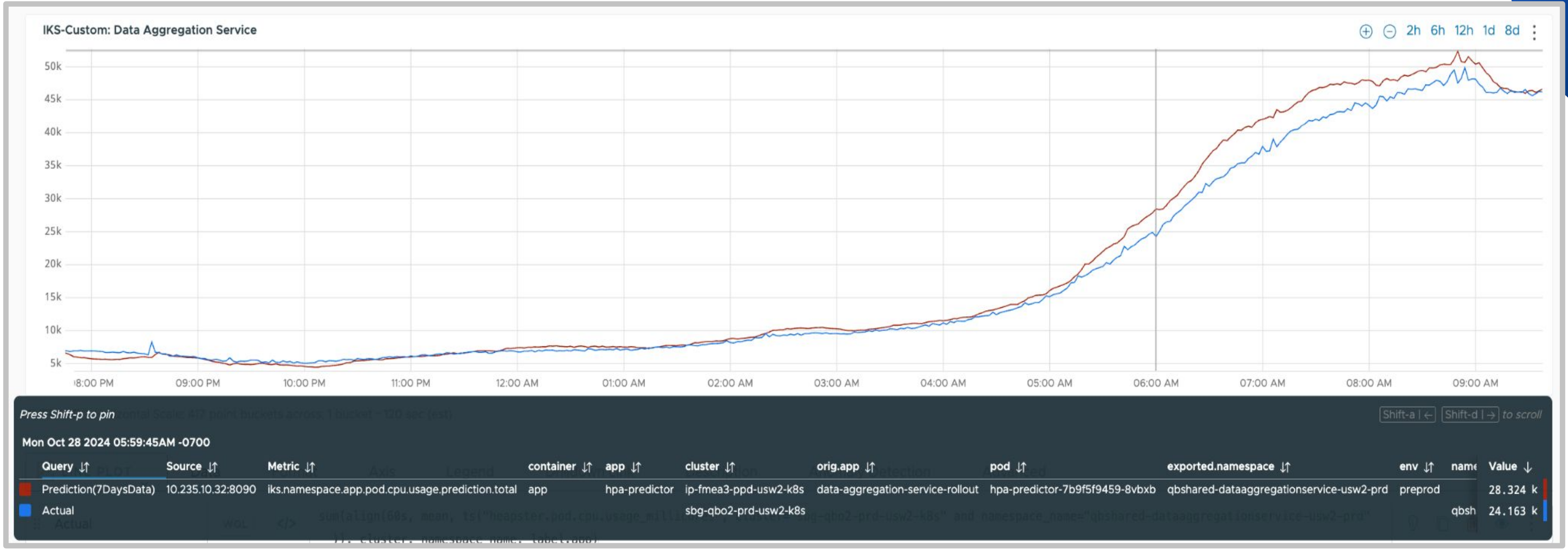

Predictions
Actuals
1600
1400
1200
1000
800
600
2024 04 02 09:08:00
2024 04 02 09:18:00
2024 04 02 09:28:00
2024 04 02 09:38:00
2024 04 02 09:48:00
2024 04 02 09:58:00
2024 04 02 10:08:00
2024 04 02 10:18:00
2024 04 02 10:28:00
2024 04 02 10:38:00
2024 04 02 10:48:00
2024 04 02 10:58:00
2024 04 02 11:08:00
2024 04 02 11:18:00
2024 04 02 11:28:00
2024 04 02 11:38:00
2024 04 02 11:48:00
2024 04 02 11:58:00
2024 04 02 12:08:00
2024 04 02 12:18:00
2024 04 02 12:28:00
2024 04 02 12:38:00
2024 04 02 12:48:00
2024 04 02 12:58:00
2024 04 02 13:08:00
2024 04 02 13:18:00
2024 04 02 13:28:00
2024 04 02 13:38:00
2024 04 02 13:48:00
2024 04 02 13:58:00
2024 04 02 14:08:00
2024 04 02 14:18:00
2024 04 02 14:28:00
2024 04 02 14:38:00
2024 04 02 14:48:00
2024 04 02 14:58:00
2024 04 02 15:08:00
2024 04 02 15:18:00
2024 04 02 15:28:00
Datetime

Demo

1. Train Nhits model



2. Expose metric



Mon Oct 28 2024 05:59:45AM -0700

	Query ⇅	Source ⇅	Metric ⇅	container ⇅	app ⇅	cluster ⇅	orig.app ⇅	pod ⇅	exported.namespace ⇅	env ⇅	name	Value ↓
🟥	Prediction(7DaysData)	10.235.10.32:8090	iks.namespace.app.pod.cpu.usage.prediction.total	app	hpa-predictor	ip-fmea3-ppd-usw2-k8s	data-aggregation-service-rollout	hpa-predictor-7b9f5f9459-8vbxb	qbshared-dataaggregationservice-usw2-prd	preprod		28.324 k
🟦	Actual					sbg-qbo2-prd-usw2-k8s					qbsh	24.163 k

3. Use predicted metric in HPA

```
spec:
  maxReplicas: 20
   metrics:
   - object:
       describedObject:
         apiVersion: apps/v1
         kind: Deployment
         name: test-app
      metric:
         name: avg_cpu_utilization_metric
       target:
         type: Value
         value: "50"
       type: Object
   - object:
       describedObject:
         apiVersion: apps/v1
         kind: Deployment
         name: test-app
       metric:
        name: avg_cpu_utilization_predictive_metric
       target:
         type: Value
         value: "50"
     type: Object
     minReplicas: 9
```

Learnings

Learnings

1. Performance test: small ramp up / ramp down time
    a. Prediction was having delay recognizing pattern in first 10 minutes
    b. Solution: HPA uses max of both actual and predicted metrics
2. Difficult to predict for spiky cpu usage
    a. Creates a spiky prediction metric
    b. Solution: Smooth the training data, prediction metric
3. Not enough data
    a. 14 days or less of data

Future enhancements

Apps with seasonal traffic pattern

- Tax peak
- Super Bowl event

Use other metrics

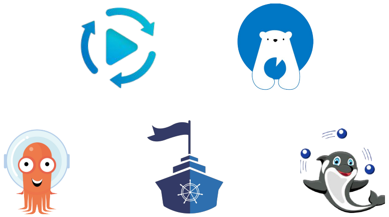
- Custom metrics
- Jvm metrics
- Tps

Multidimensional prediction

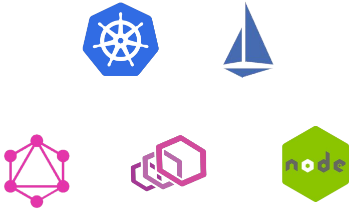
- Prediction metric not based on single metric
- Prevent from influencing real cpu usage data

We believe in open source and open collaboration

Recipient of the **End User Award** in 2019 & 2022



Created, open-sourced, used, and maintained **by Intuit**



End user of cloud native and mobile open source tech

bit.ly/intuit-oss

Stay in the loop

FOLLOW

Intuit Open Source

Don’t miss on exciting OSS events, activities & news

Scan or visit
bit.ly/intuit-oss

Visit our Booth

Get some exciting OSS swag - while supplies last

Check out our past presentations

Platform Eng Day
https://www.youtube.com/watch?v=z6ltgXM4RxE

Autoscaling
https://www.youtube.com/watch?v=h2zmITPG3GM

Debuggability
https://www.youtube.com/watch?v=bPa1PjY-Hg4

Proposed solution

Proposed solution

Time series forecasting for real-time cpu usage prediction

- Integrate predicted metrics with Prometheus
- Use predicted metrics as custom metric in HPA
- Proactively increases desired replicas x minutes ahead of time, based on pod start up time

Cost saving

- Recommend an optimal minReplica value

Benefits

Improve service availability

Proactively increases desired replicas ~x minutes ahead of time

Promising result for weekly/daily traffic pattern app

The ML Model is able to predict ahead of time

Cost saving

Reduce number of minReplicas