CLOUD NATIVE
COMPUTING FOUNDATION

Kubernetes Shouldn't Be Scary: Mastering Deployments and Scaling for Web Developers

Christopher Tineo
IDT | Deployments Engineer

CNCF Chapter in Santo Domingo

Cloud Native
Community Groups
Santo Domingo

Do not edit
How to change the design

CLOUD NATIVE
COMPUTING FOUNDATION

Join at slido.com
#1800077

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

CLOUD NATIVE
COMPUTING FOUNDATION

Audience Q&A

CLOUD NATIVE
COMPUTING FOUNDATION

Do you know Containers?

CLOUD NATIVE
COMPUTING FOUNDATION

Do you know Container Orchestration?

Agenda

- About me
- **Why Kubernetes Matters for any Developer**
- Key Kubernetes Concepts
- Live Demo: Deploying a Web App in Kubernetes
- **Scaling With KEDA (Kubernetes Event-Driven Autoscaling)**
- Best Practices & Tools
- Q&A and Closing Remarks

Cloud Native Community Groups
Santo Domingo

About Me

I'm a Community Organizer for the CNCF chapter in **Santo Domingo, Dominican Republic**.

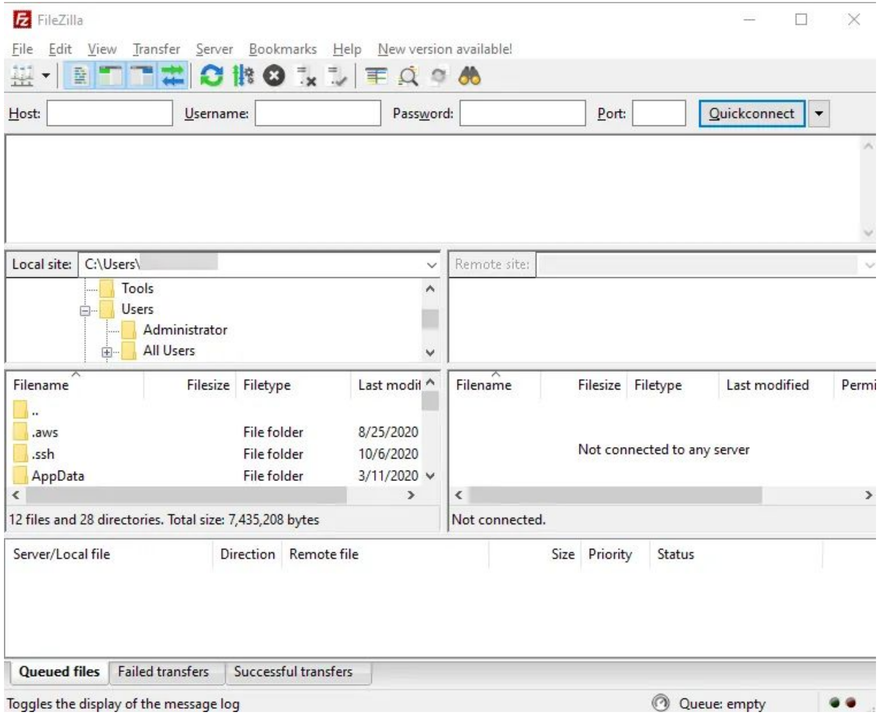
Enjoy giving talks, conferences and everything in the **open-source community.**

in **Christopher tineo**

Why Kubernetes Matters for any Developer

Cloud Native
Community Groups
Santo Domingo

Does your Deployments look like this?



Local site: C:\Users\[redacted]

- Tools
- Users
  - Administrator
  - All Users

Filename	Filesize	Filetype	Last modif
..			
.aws		File folder	8/25/2020
.ssh		File folder	10/6/2020
AppData		File folder	3/11/2020

12 files and 28 directories. Total size: 7,435,208 bytes

Filename	Filesize	Filetype	Last modified	Permi

Not connected to any server

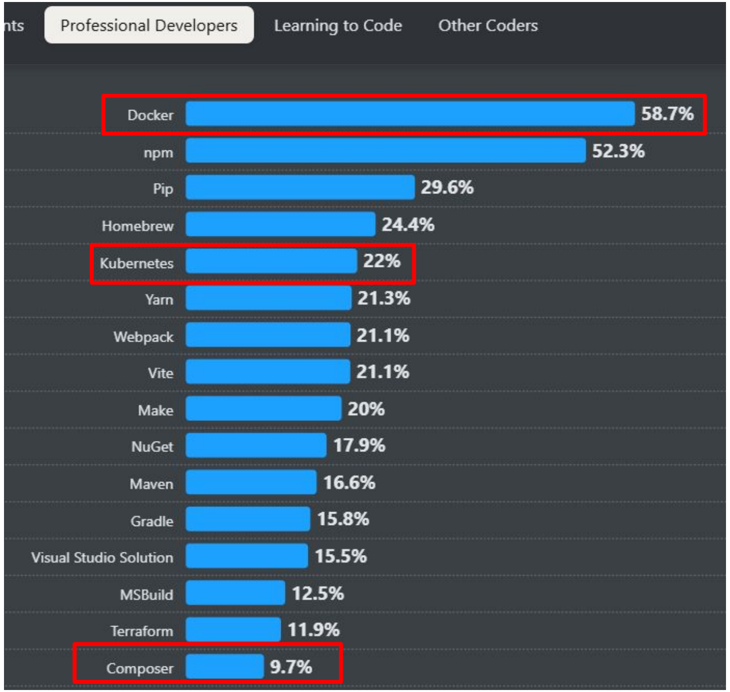
Not connected.

Server/Local file	Direction	Remote file	Size	Priority	Status

Toggles the display of the message log

Queue: empty

Stack Developer Survey 2024



Essential tools for a Dev in 2025

Container Engine (docker/podman)

Container orchestrator

84%

Companies were using or Evaluating Kubernetes as of 2023

*Based on CNCF Annual Report 2023**

Cloud Native
Community Groups
Santo Domingo

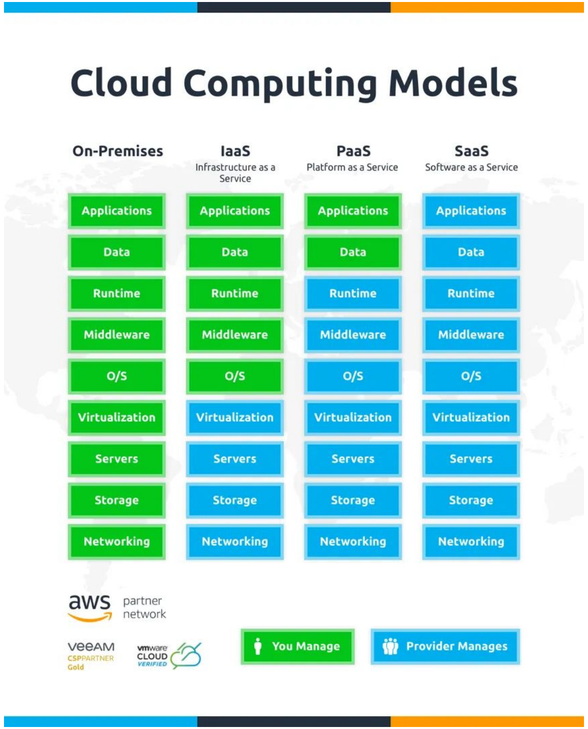
What Cloud Providers Say

1. **Trade fixed expense for variable expense**
2. Benefit from massive economies of scale
3. Stop guessing capacity
4. **Increase speed and agility (HA & Resilience)**
5. **Stop spending money running and maintaining data centers (Spot Instances)**
6. **Go global in minutes (Multi region)**

From AWS Website: **Six advantages of cloud computing**

What Cloud Providers Don’t Say

1. You will need to deal with **Vendor Lock-In.**
2. Your code needs to **adapt** to your **provider services and platform**.
3. You must decide how much control you’re **willing to give up** when selecting between (IaaS/PaaS/SaaS).



Cloud Computing Models

On-Premises	IaaS Infrastructure as a Service	PaaS Platform as a Service	SaaS Software as a Service
🟩 Applications	🟩 Applications	🟩 Applications	🟦 Applications
🟩 Data	🟩 Data	🟩 Data	🟦 Data
🟩 Runtime	🟩 Runtime	🟦 Runtime	🟦 Runtime
🟩 Middleware	🟩 Middleware	🟦 Middleware	🟦 Middleware
🟩 O/S	🟩 O/S	🟦 O/S	🟦 O/S
🟩 Virtualization	🟦 Virtualization	🟦 Virtualization	🟦 Virtualization
🟩 Servers	🟦 Servers	🟦 Servers	🟦 Servers
🟩 Storage	🟦 Storage	🟦 Storage	🟦 Storage
🟩 Networking	🟦 Networking	🟦 Networking	🟦 Networking

🟩 You Manage 🟦 Provider Manages

Build once, Deploy everywhere

This should be solved, right?

Cloud Native Community Groups
Santo Domingo

CNCF Cloud Native Definition v1.1

Cloud native practices empower organizations to develop, build, and deploy workloads in **computing environments (public, private, hybrid) …**

CLOUD NATIVE
COMPUTING FOUNDATION

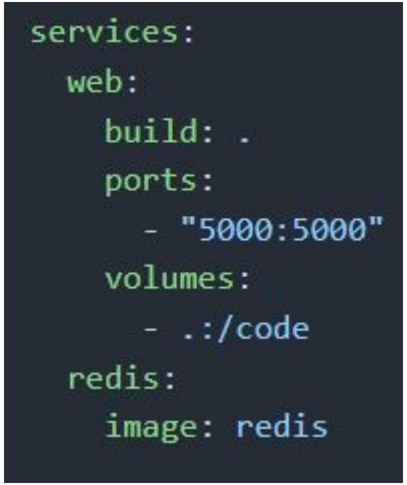
Cloud Native
Community Groups
Santo Domingo

Key Kubernetes Concepts

Cloud Native
Community Groups
Santo Domingo

What you’re probably familiar with

Docker compose

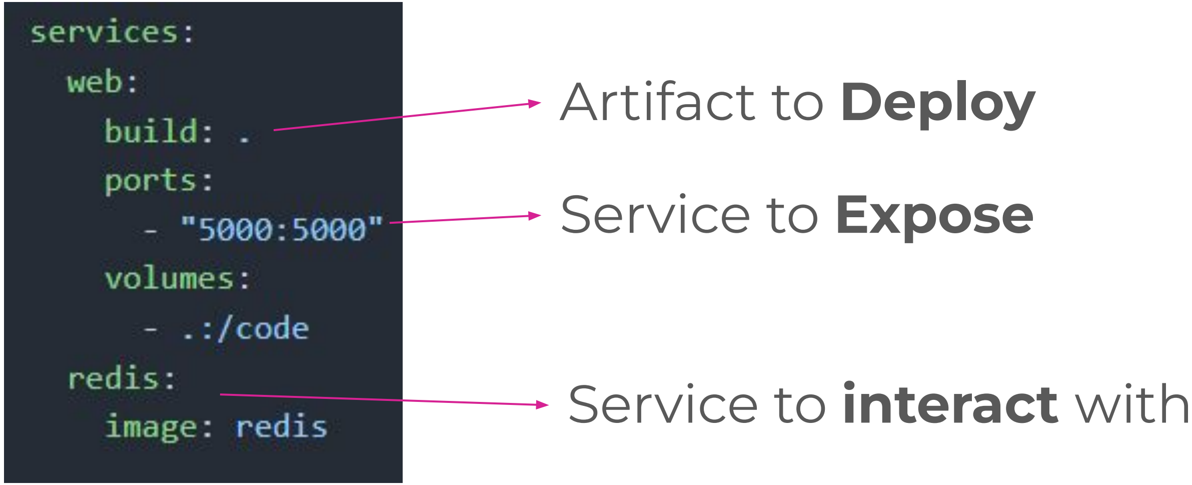


```
services:
  web:
    build: .
    ports:
      - "5000:5000"
    volumes:
      - .:/code
  redis:
    image: redis
```

Cloud Native
Community Groups
Santo Domingo

What you’re probably familiar with

Docker compose



```
services:
  web:
    build: .
    ports:
      - "5000:5000"
    volumes:
      - .:/code
  redis:
    image: redis
```

Artifact to **Deploy**

Service to **Expose**

Service to **interact** with

Service

Is a set of pods **(artifacts)** that are **exposed** within the cluster network.

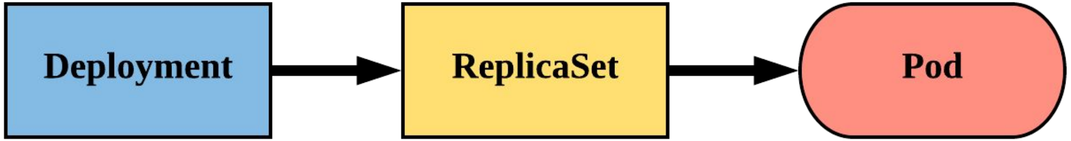
- Have an unique static IP
- Have their own dns record.

```
<service-name>.<namespace>.svc.cluster.local
frontend.default.svc.cluster.local
```

Cloud Native
Community Groups
Santo Domingo

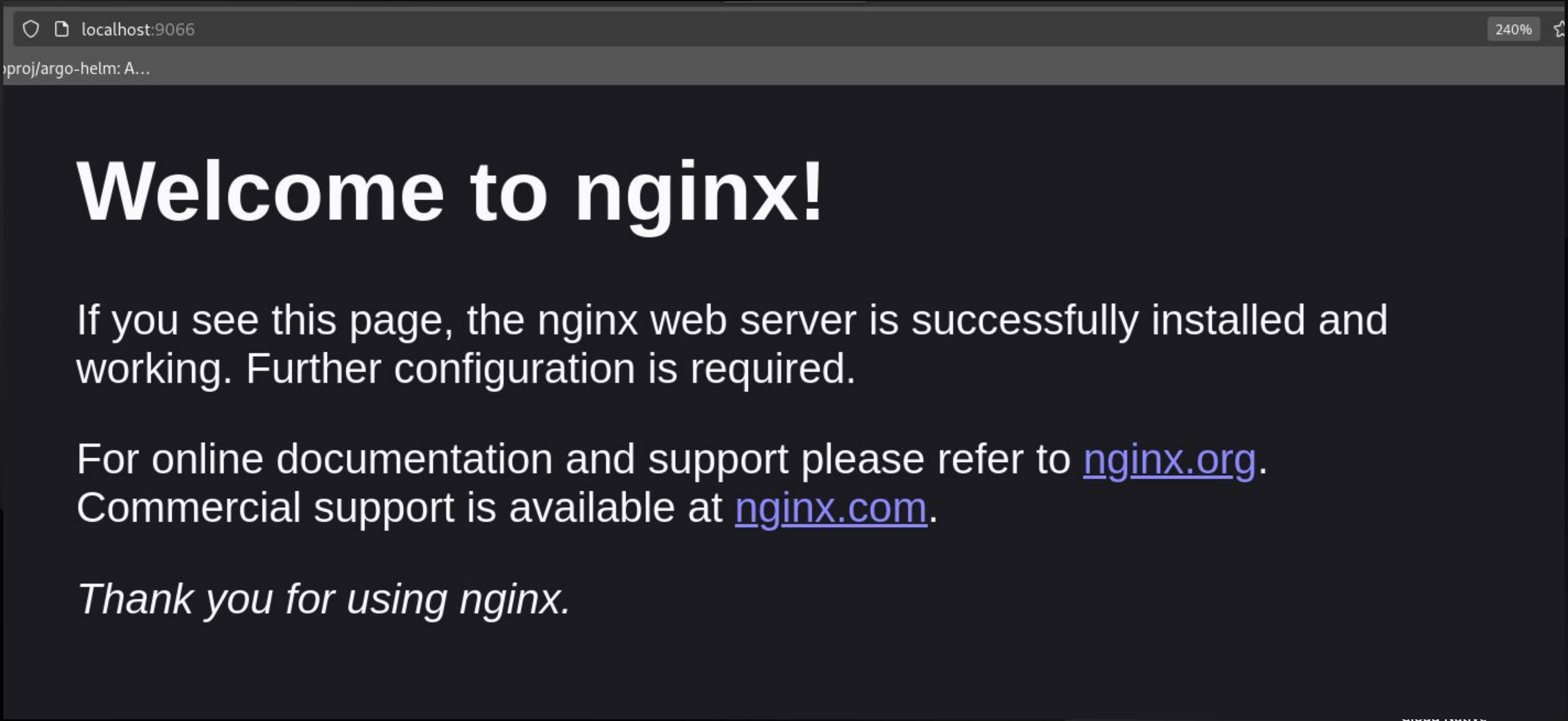
Deployments

Is a resource whose job is to **guarantee** that their desired amount of replicas **(artifacts)** are **up and running** correctly.



Cloud Native Community Groups
Santo Domingo

Live Demo: Deploying a Web App in Kubernetes

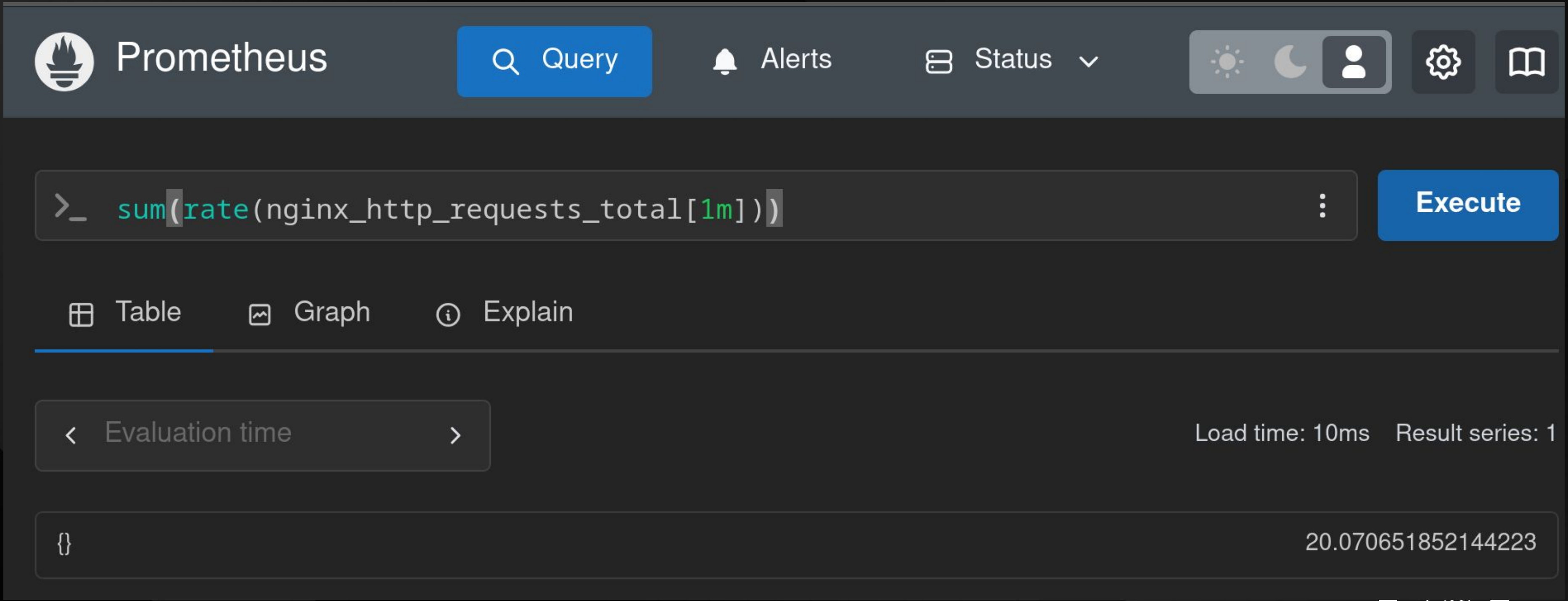


Welcome to nginx!

If you see this page, the nginx web server is successfully installed and working. Further configuration is required.

For online documentation and support please refer to nginx.org.
Commercial support is available at nginx.com.

Thank you for using nginx.

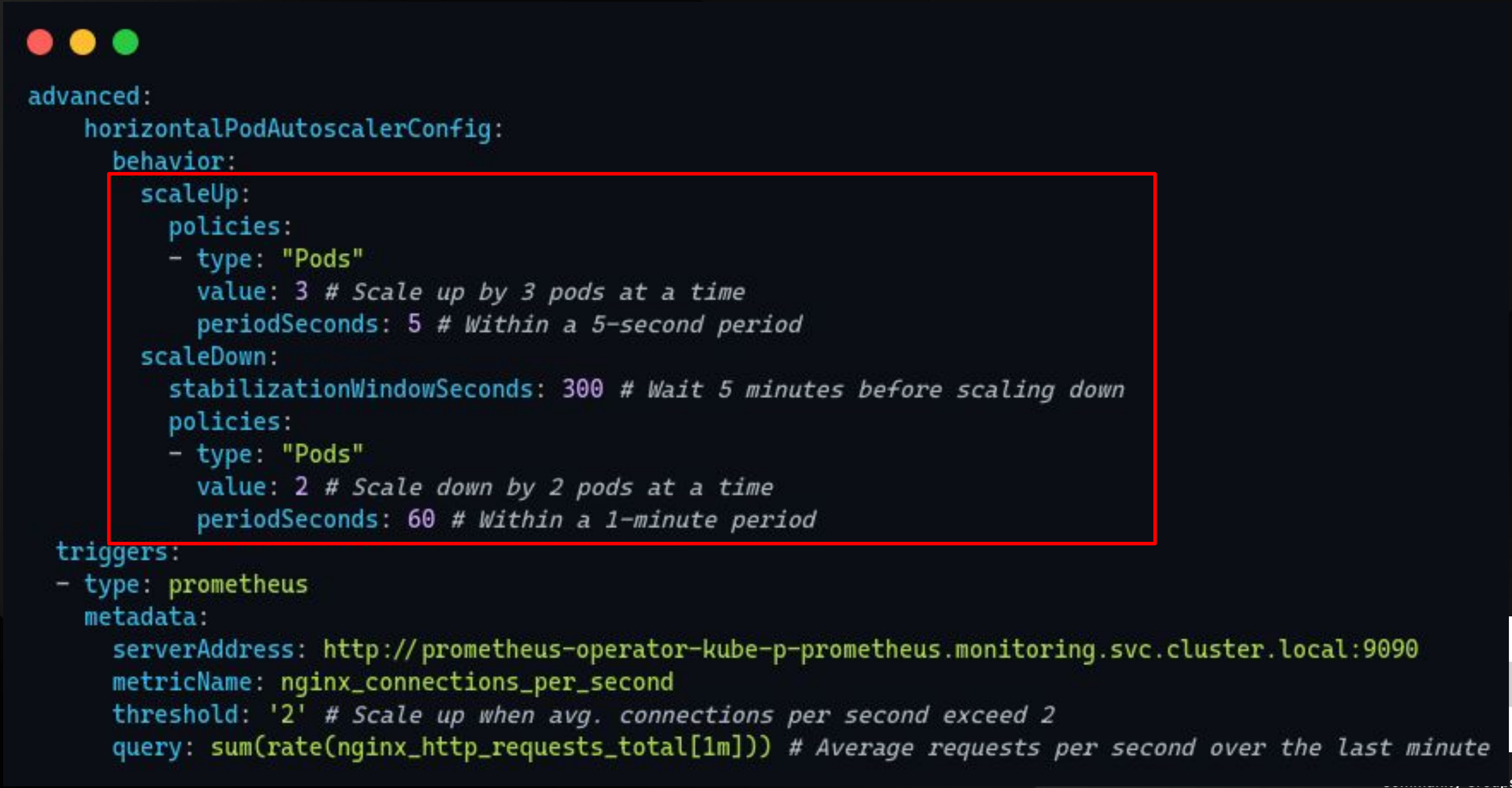


```
sum(rate(nginx_http_requests_total[1m]))
```

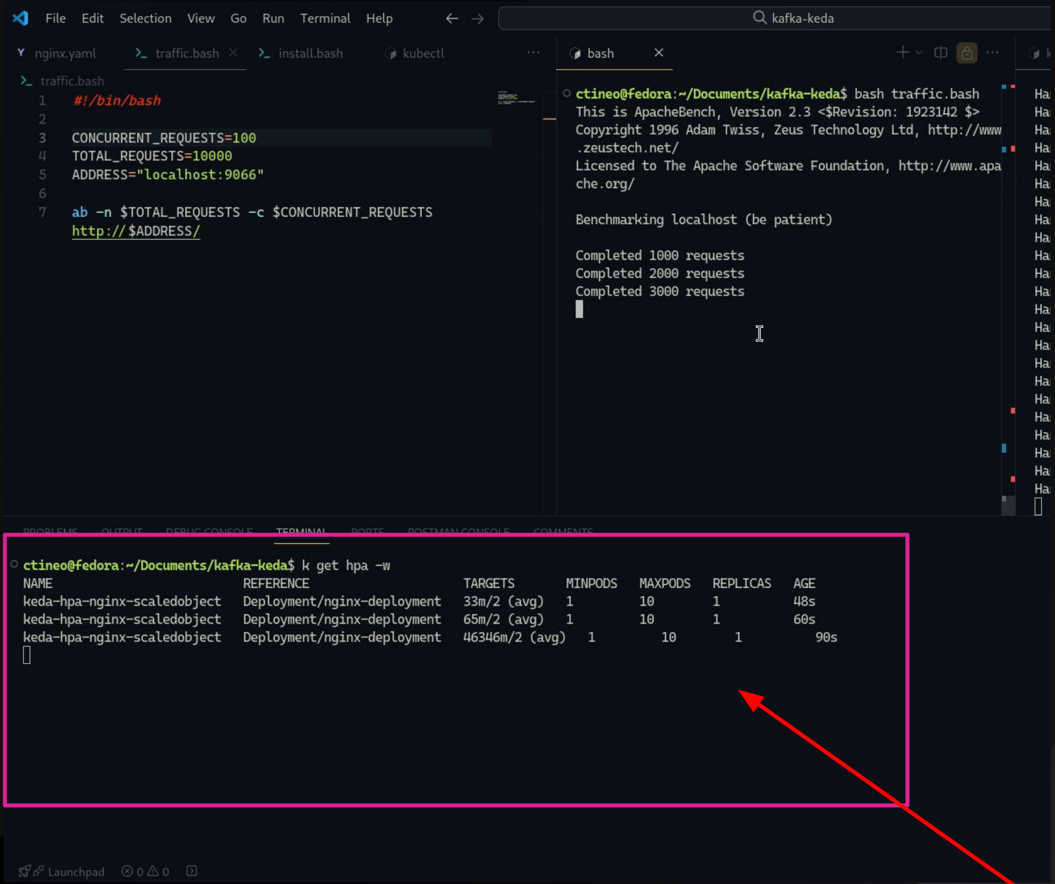
Load time: 10ms Result series: 1

{}	20.070651852144223

Cloud Native
Community Groups
Santo Domingo



```yaml
advanced:
    horizontalPodAutoscalerConfig:
      behavior:
        scaleUp:
          policies:
          - type: "Pods"
            value: 3 # Scale up by 3 pods at a time
            periodSeconds: 5 # Within a 5-second period
        scaleDown:
          stabilizationWindowSeconds: 300 # Wait 5 minutes before scaling down
          policies:
          - type: "Pods"
            value: 2 # Scale down by 2 pods at a time
            periodSeconds: 60 # Within a 1-minute period
  triggers:
  - type: prometheus
    metadata:
      serverAddress: http://prometheus-operator-kube-p-prometheus.monitoring.svc.cluster.local:9090
      metricName: nginx_connections_per_second
      threshold: '2' # Scale up when avg. connections per second exceed 2
      query: sum(rate(nginx_http_requests_total[1m])) # Average requests per second over the last minute
```



traffic.bash

```bash
#!/bin/bash

CONCURRENT_REQUESTS=100
TOTAL_REQUESTS=10000
ADDRESS="localhost:9066"

ab -n $TOTAL_REQUESTS -c $CONCURRENT_REQUESTS
http://$ADDRESS/
```

```
ctineo@fedora:~/Documents/kafka-keda$ bash traffic.bash
This is ApacheBench, Version 2.3 <$Revision: 1923142 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www
.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apa
che.org/

Benchmarking localhost (be patient)

Completed 1000 requests
Completed 2000 requests
Completed 3000 requests
```

```
ctineo@fedora:~/Documents/kafka-keda$ k get hpa -w
NAME                          REFERENCE                       TARGETS        MINPODS   MAXPODS   REPLICAS   AGE
keda-hpa-nginx-scaledobject   Deployment/nginx-deployment   33m/2 (avg)    1         10        1          48s
keda-hpa-nginx-scaledobject   Deployment/nginx-deployment   65m/2 (avg)    1         10        1          60s
keda-hpa-nginx-scaledobject   Deployment/nginx-deployment   46346m/2 (avg)   1         10        1          90s
```

Cloud Native
Community Groups
Santo Domingo

But my app is running Okey

Why even bother?

Cloud Native
Community Groups
Santo Domingo

Risks of not scaling

Under Provisioning during Traffic Spikes

Overprovisioning in Low Traffic Periods

Inability to Handle Unpredictable Workloads

Operational Complexity of Manual Scaling

Cloud Native Community Groups
Santo Domingo

Kubernetes Autoscaling

Autoscaling options for Kubernetes

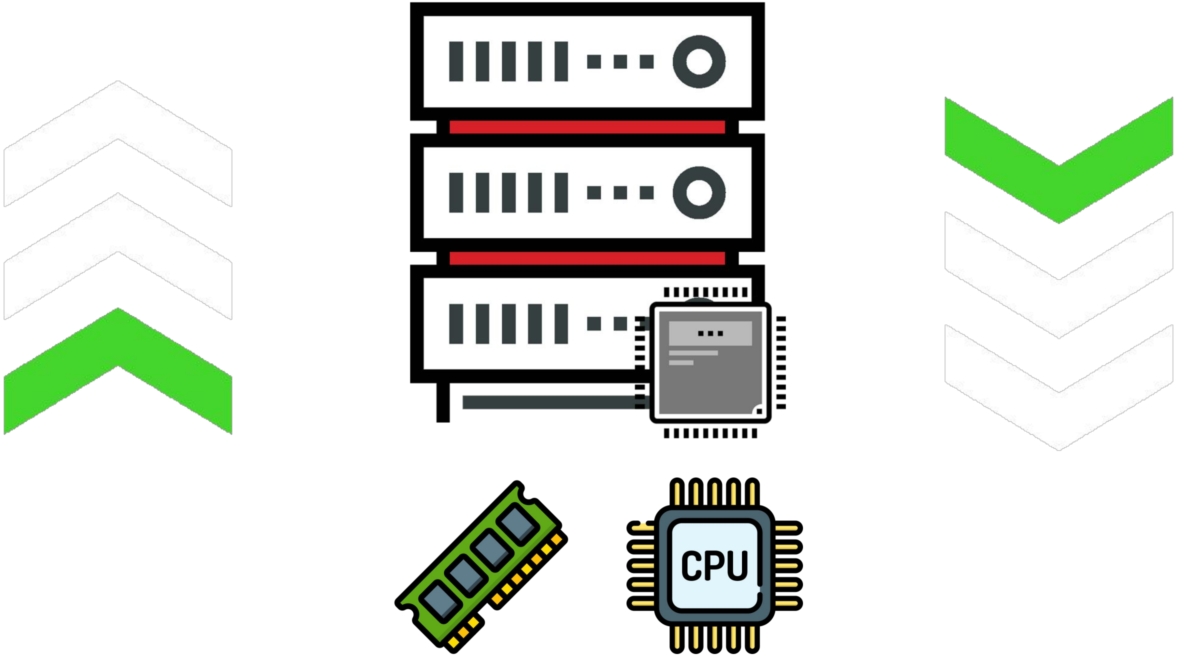
Cluster Autoscaler ✓
Adjusts the **size of a Kubernetes Cluster** based on resource demands and **optimizing cost**.

Horizontal and Vertical Pod Autoscaler ✓
Adjust the resources allocated to pods or spread the load across a **fleet of pods**.

Event Driven Autoscaler ✓
Adjust your workloads based on events.
Customer Orders, Processing time, Users connected

Vertical Scaling

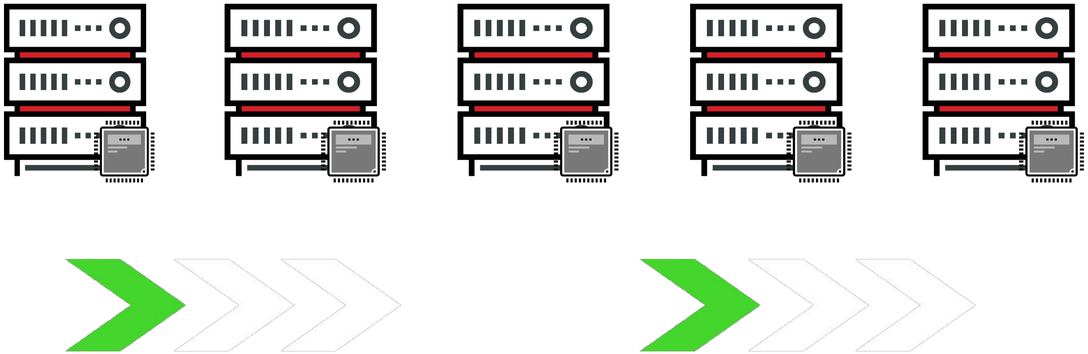
Scaling Up



Cloud Native
Community Groups
Santo Domingo

Horizontal Scaling

Scaling Out



Cloud Native
Community Groups
Santo Domingo

Autoscaling

Using Cloud Native Practices

The ability of a system to **scale automatically**, typically, in terms of computing resources. With an auto scaling system, **resources are automatically added** when needed and can scale **to meet fluctuating user demands.**

Reactive

Scale according to workload

Great option when latency is not a consideration

Scheduled

Schedule auto scaling of resources

Can plan ahead to avoid latency disruption

Predicted

Scaling with AI/ Machine Learning

Intelligent Autoscaling

Benefits of Event-Driven Autoscaling

Scaling based on what your business matters

1. Amount of orders in queue
2. Amount pending transactions
3. Users connected simultaneously
4. Average response time of your services

And the best of it, you could define yours.

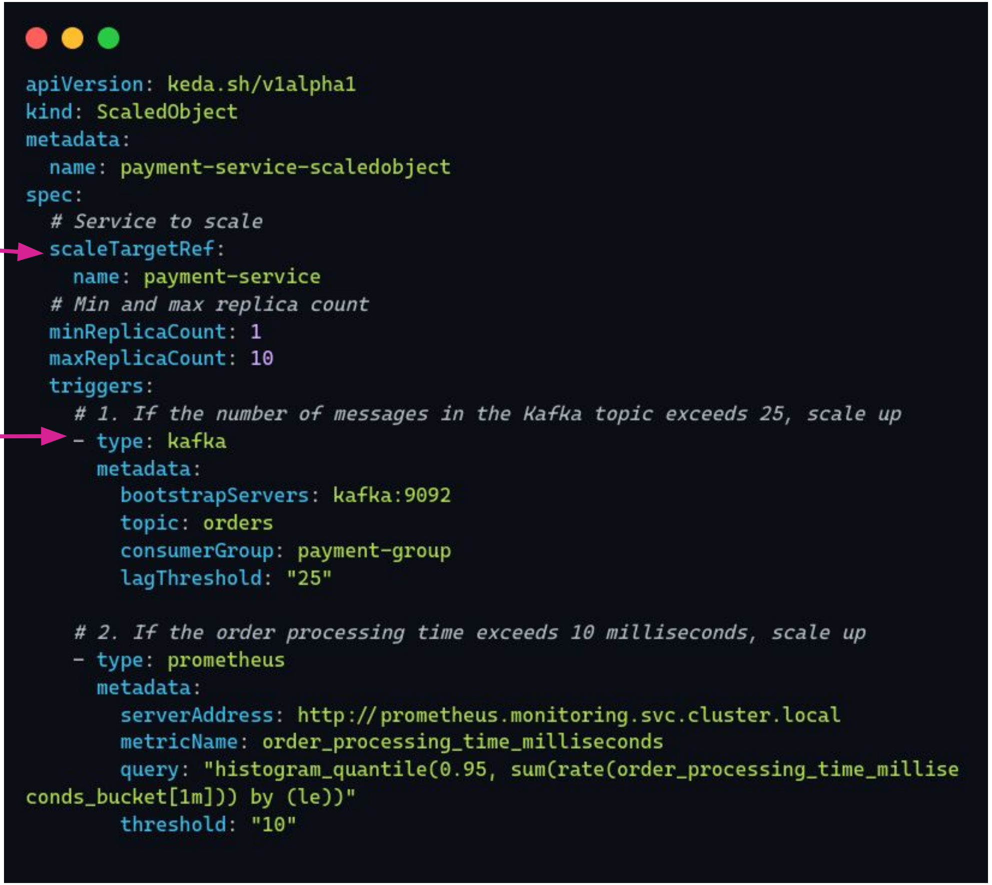
Kubernetes Event-driven Autoscaling

With KEDA, you can drive the scaling of any container in Kubernetes based on events.

ScaledObject

Target Service

Events (1...n)



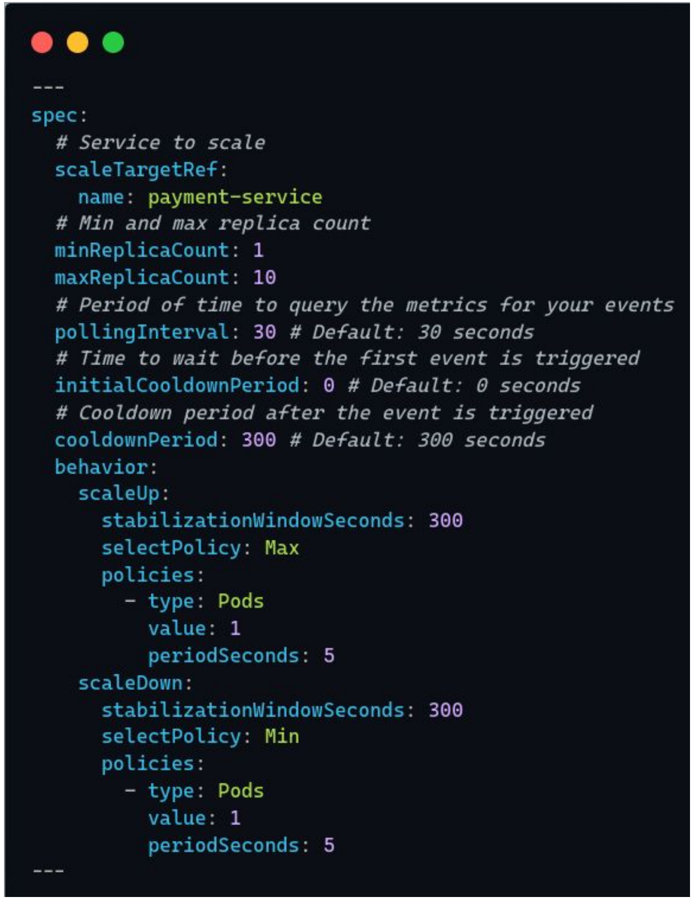
```yaml
apiVersion: keda.sh/v1alpha1
kind: ScaledObject
metadata:
  name: payment-service-scaledobject
spec:
  # Service to scale
  scaleTargetRef:
    name: payment-service
  # Min and max replica count
  minReplicaCount: 1
  maxReplicaCount: 10
  triggers:
    # 1. If the number of messages in the Kafka topic exceeds 25, scale up
    - type: kafka
      metadata:
        bootstrapServers: kafka:9092
        topic: orders
        consumerGroup: payment-group
        lagThreshold: "25"

    # 2. If the order processing time exceeds 10 milliseconds, scale up
    - type: prometheus
      metadata:
        serverAddress: http://prometheus.monitoring.svc.cluster.local
        metricName: order_processing_time_milliseconds
        query: "histogram_quantile(0.95, sum(rate(order_processing_time_millise
conds_bucket[1m])) by (le))"
        threshold: "10"
```

ScaledObject

How should I scale?

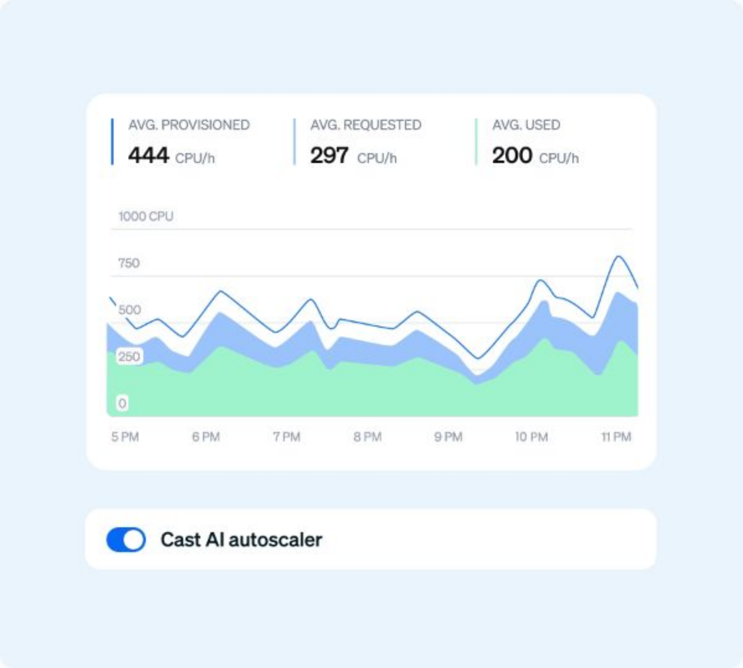
As the typical IT guy, “it depends”

My recommendation is: “Scale up aggressively and scale down conservatively”



```yaml
---
spec:
  # Service to scale
  scaleTargetRef:
    name: payment-service
  # Min and max replica count
  minReplicaCount: 1
  maxReplicaCount: 10
  # Period of time to query the metrics for your events
  pollingInterval: 30 # Default: 30 seconds
  # Time to wait before the first event is triggered
  initialCooldownPeriod: 0 # Default: 0 seconds
  # Cooldown period after the event is triggered
  cooldownPeriod: 300 # Default: 300 seconds
  behavior:
    scaleUp:
      stabilizationWindowSeconds: 300
      selectPolicy: Max
      policies:
        - type: Pods
          value: 1
          periodSeconds: 5
    scaleDown:
      stabilizationWindowSeconds: 300
      selectPolicy: Min
      policies:
        - type: Pods
          value: 1
          periodSeconds: 5
---
```

Cast AI

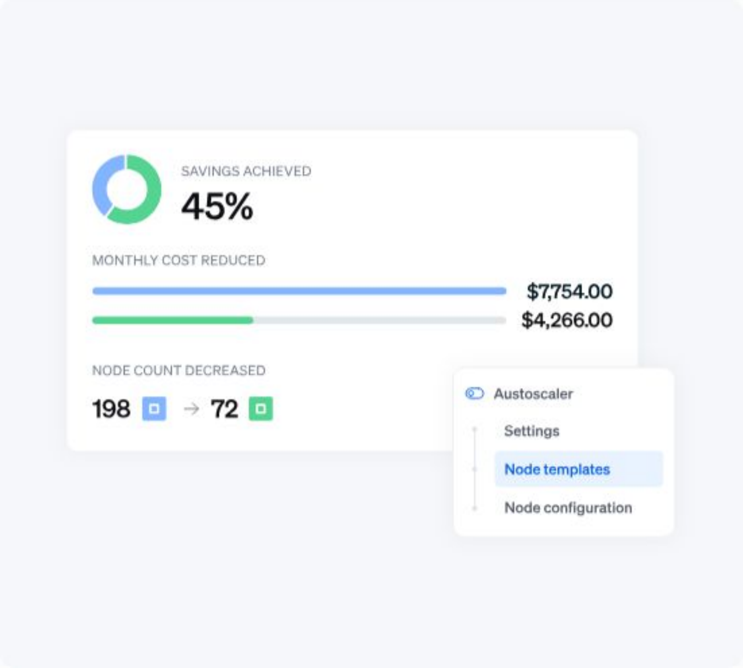


AVG. PROVISIONED 444 CPU/h

AVG. REQUESTED 297 CPU/h

AVG. USED 200 CPU/h

Cast AI autoscaler



SAVINGS ACHIEVED 45%

MONTHLY COST REDUCED

$7,754.00

$4,266.00

NODE COUNT DECREASED

198 → 72

Other tools to consider

Karpenter

kubecost

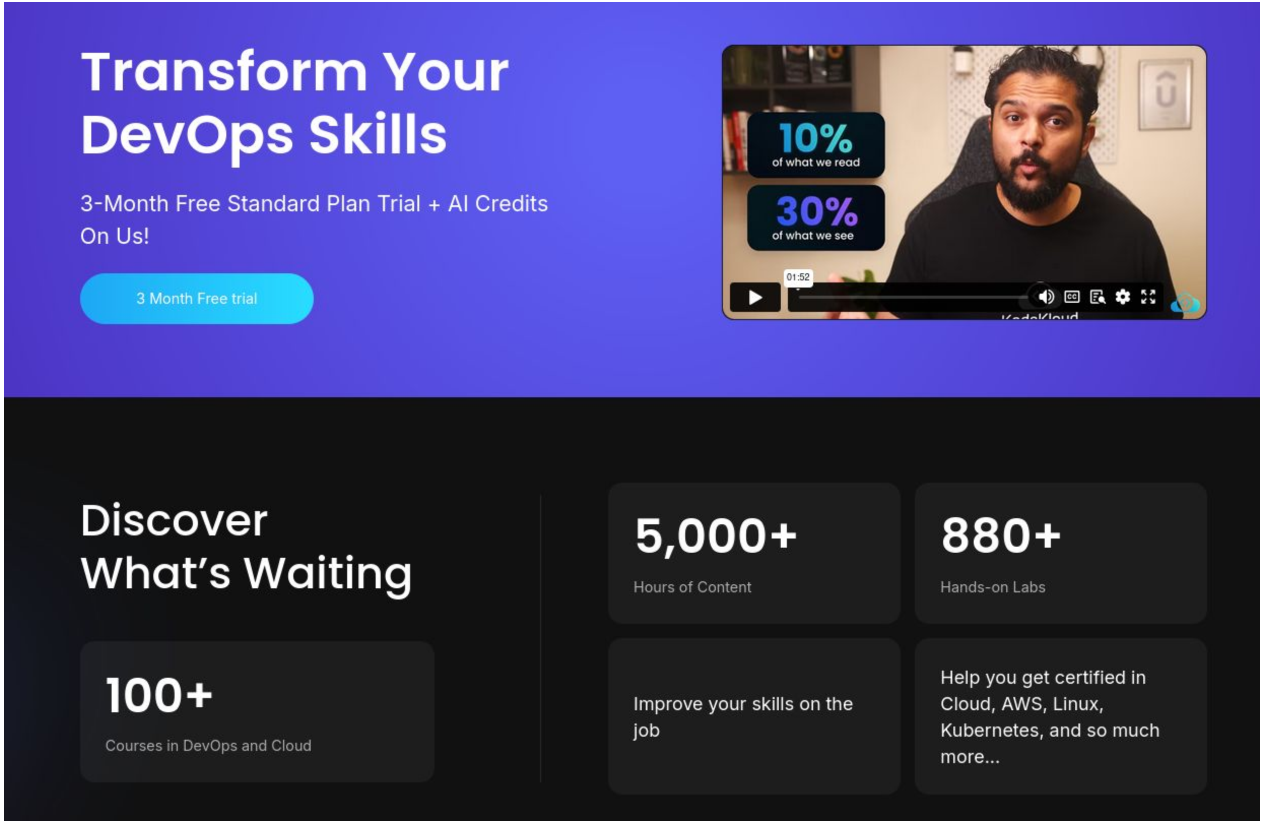
cast.ai

You want to learn more?

Cloud Native
Community Groups
Santo Domingo

3 months of KodeKloud Free

KodeKloud



Transform Your DevOps Skills

3-Month Free Standard Plan Trial + AI Credits On Us!

3 Month Free trial

Discover What's Waiting

100+
Courses in DevOps and Cloud

5,000+
Hours of Content

880+
Hands-on Labs

Improve your skills on the job

Help you get certified in Cloud, AWS, Linux, Kubernetes, and so much more...

Christopher Tineo

CLOUD NATIVE
COMPUTING FOUNDATION

Audience Q&A