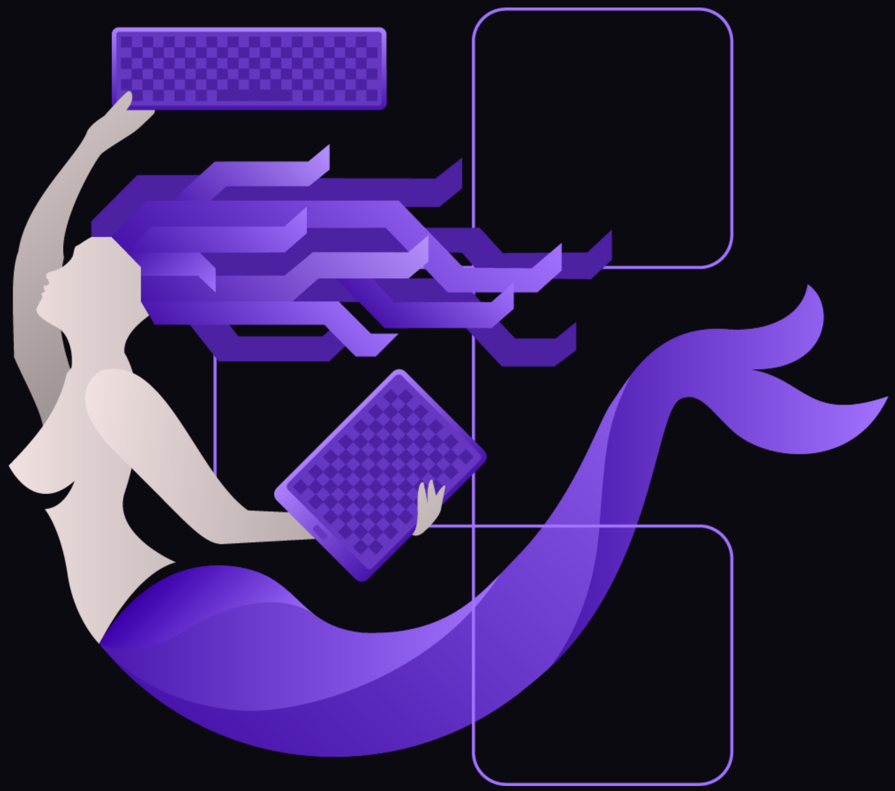
WARSAW
IT DAYS

Scaling to Success: Leveraging KEDA and Kubernetes for Optimal Azure DevOps Pipeline Performance



Wolfgang Ofner

Freelance Cloud Architect, ProgrammingWithWolfgang

MAIN ORGANIZER: AcademicPartners FUNDACJA   ORGANIZING COMMITTEE: dozens of organizations from the IT / data science sector (full list on the event website)

Wolfgang Ofner

Freelance Cloud Architect, Perth, Australia

Focus on Azure, Kubernetes, DevOps and .NET

https://programmingwithwolfgang.com

https://www.linkedin.com/in/wolfgangofner

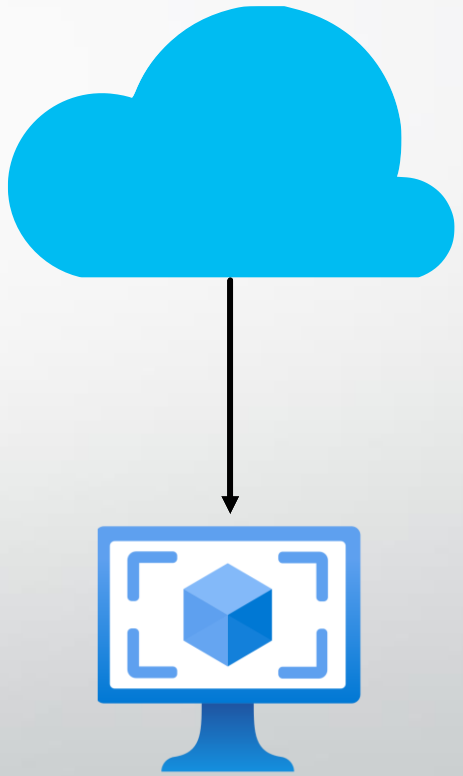
https://twitter.com/wolfgang_ofner

Agenda

- Architecture in SW projects
- Introduction to KEDA
- Scaling Azure DevOps Agents in Kubernetes
- KEDA Conclusion

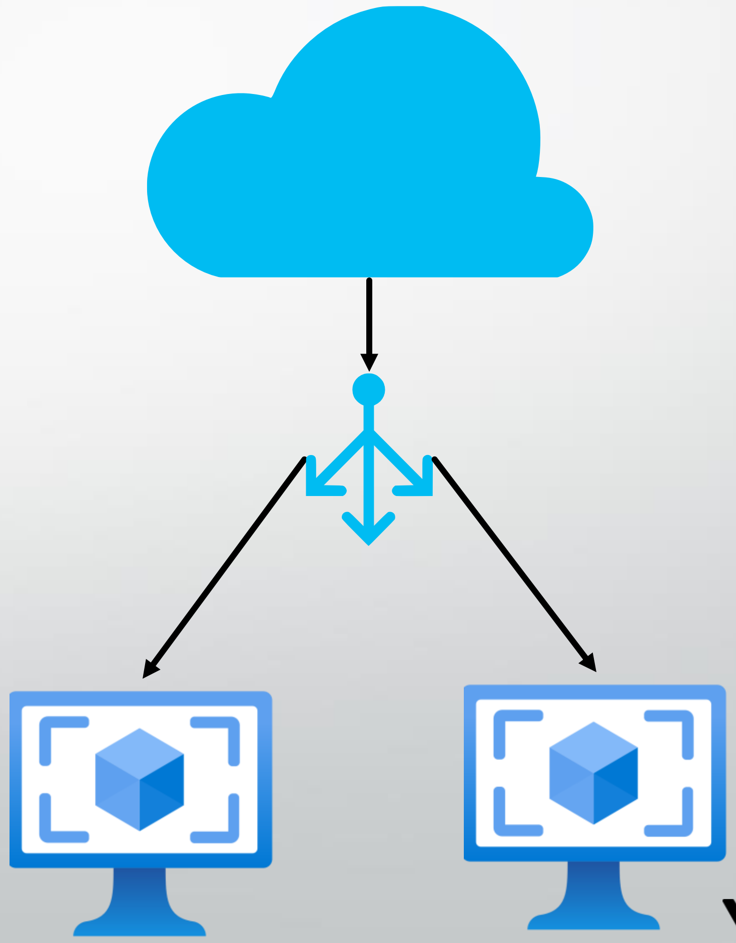
Simplified Architecture History

- Server – Client Architecture
- Only few clients
- No redundancy
- No high availability



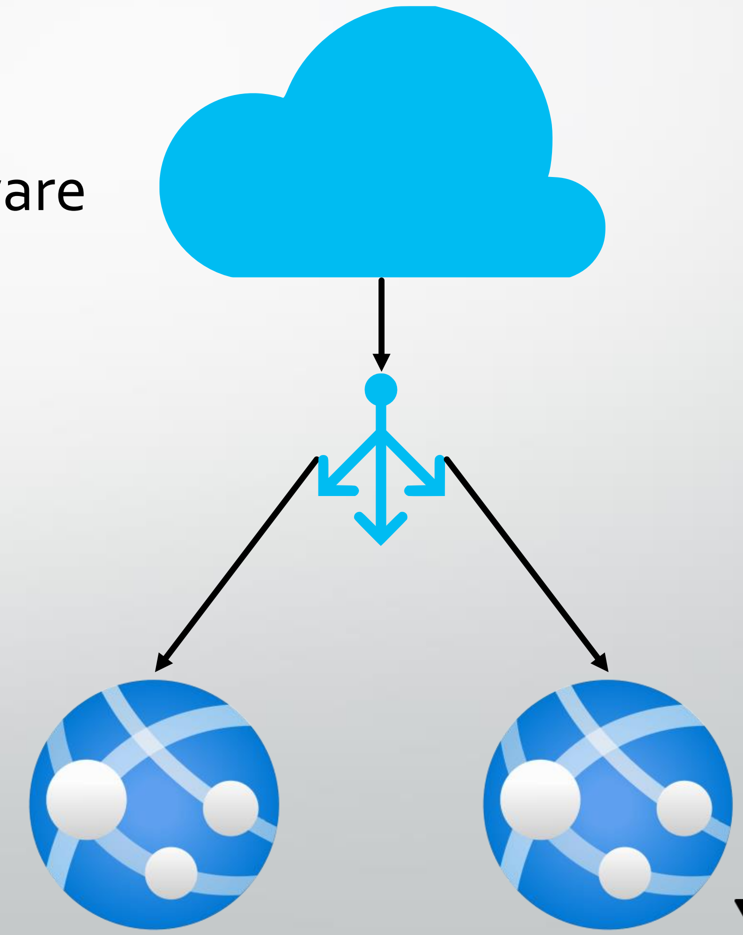
Simplified Architecture History

- Static load balancing
- New VMs need to be added by hand
- Expensive on-premises hardware



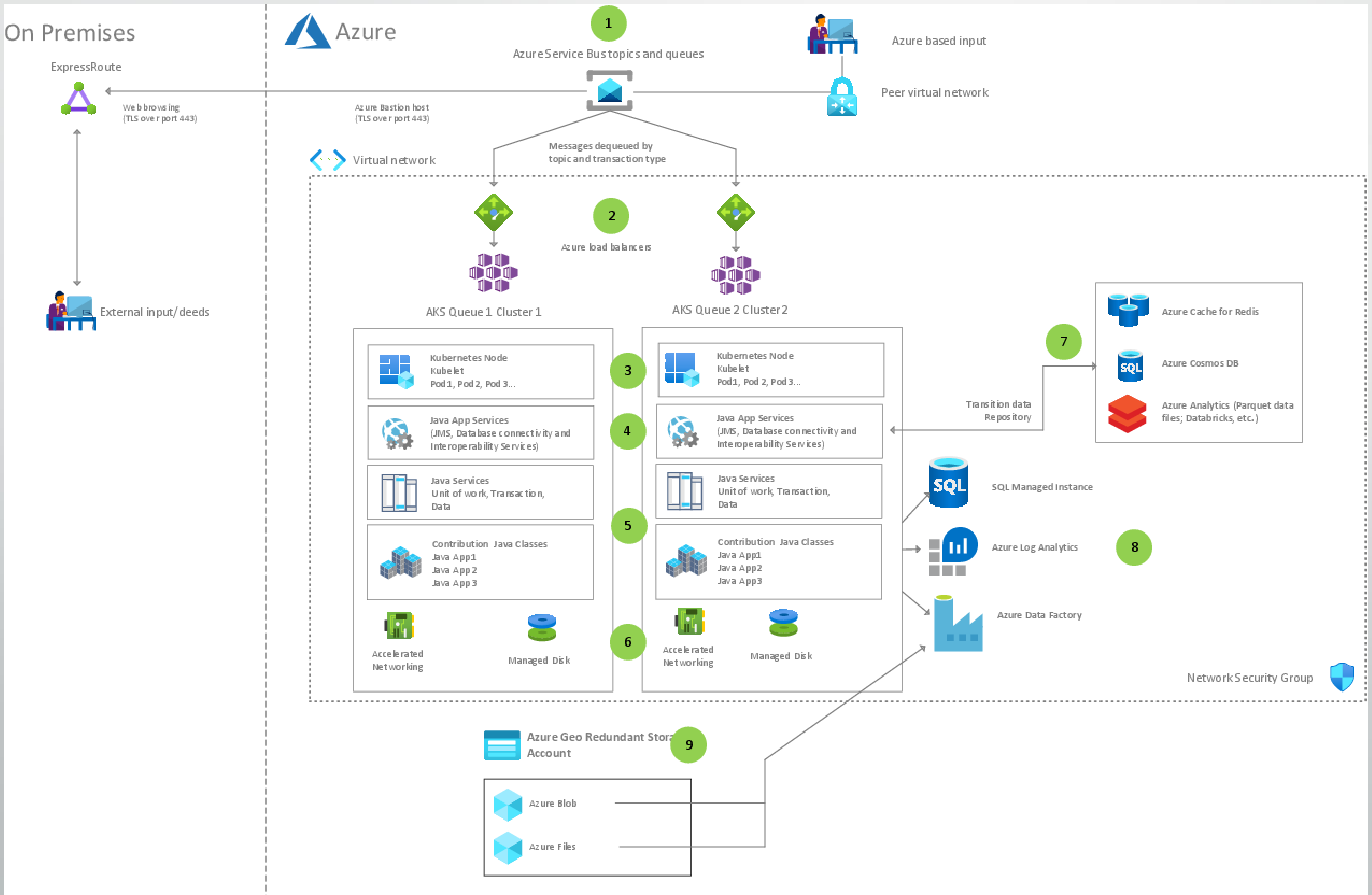
Simplified Architecture History

- Automatically adding additional hardware
- Pay only what you need
- Mostly CPU or RAM based scaling



WARSAW IT DAYS

Modern Architecture



Kubernetes

- Horizontal Pod Autoscaler (HPA)
  - Scaling according to CPU and/or RAM
- Architectures get more and more complex
- Dependencies on external components
- Applications have to react to events
  - Database
  - Service Bus
  - Streams

Horizontal Pod Autoscaler

- Scales Deployments or StatefulSets
- Adds or removes pods
- Scaling based on CPU or RAM usage
- Scaling on custom metrics
  - Query custom metrics from Kubernetes API
  - Prometheus
  - requests per second

Horizontal Pod Autoscaler Configuration

```yaml
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: customerapi
  namespace: customerapi-test
spec:
  maxReplicas: 10
  minReplicas: 1
  averageCpuUtilization: 50
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: customerapi
```

```yaml
behavior:
  scaleDown:
    policies:
    - type: Pods
      value: 4
      periodSeconds: 60
    - type: Percent
      value: 10
      periodSeconds: 60
    selectPolicy: Min
```

```yaml
scaleUp:
  policies:
  - type: Pods
    value: 5
    periodSeconds: 60
  - type: Percent
    value: 12
    periodSeconds: 60
  selectPolicy: Max
```

Limitation of the HPA

- Black Friday
- Thousands of orders are stored in a queue
- Scaling using CPU or RAM is not sufficient
- No option for scaling in this scenario

KEDA – Kubernetes Event-driven Autoscaling

- Kubernetes Event-driven Autoscaling
- Open source
- CNCF Project
- Maintained by
  - Docplanner Tech
  - Microsoft
  - Red Hat

KEDA

- ~61 built-in Scaler
  - Apache Kafka
  - Azure Blob Storage
  - Azure Monitor
  - Azure Service Bus
  - Elastic Search
  - MongoDB
  - Prometheus
  - Redis Streams

KEDA Use Cases

- Scale according to external events
- Scale to Zero
  - Bring serverless to your datacenter
  - Recreate Azure Functions architecture
  - Better resource usage

WARSAW IT DAYS

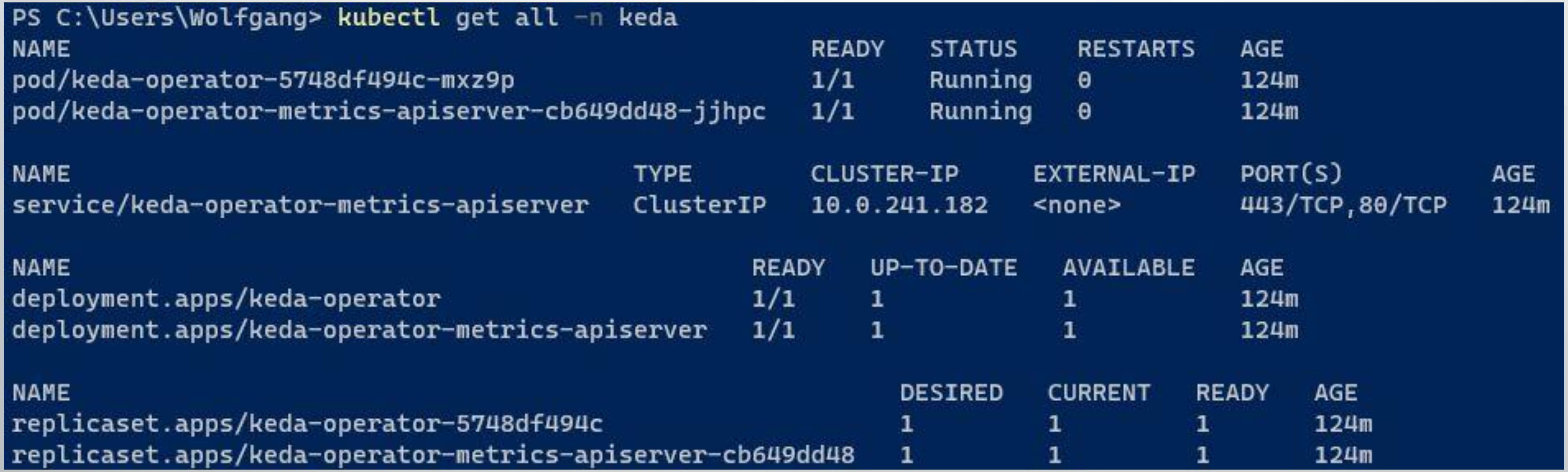
KEDA Installation

- Installation via Helm charts
- Namespace: keda

KEDA Installation

```
kubectl create namespace keda

helm repo add kedacore https://kedacore.github.io/charts
helm repo update
helm install keda kedacore/keda --namespace keda
```

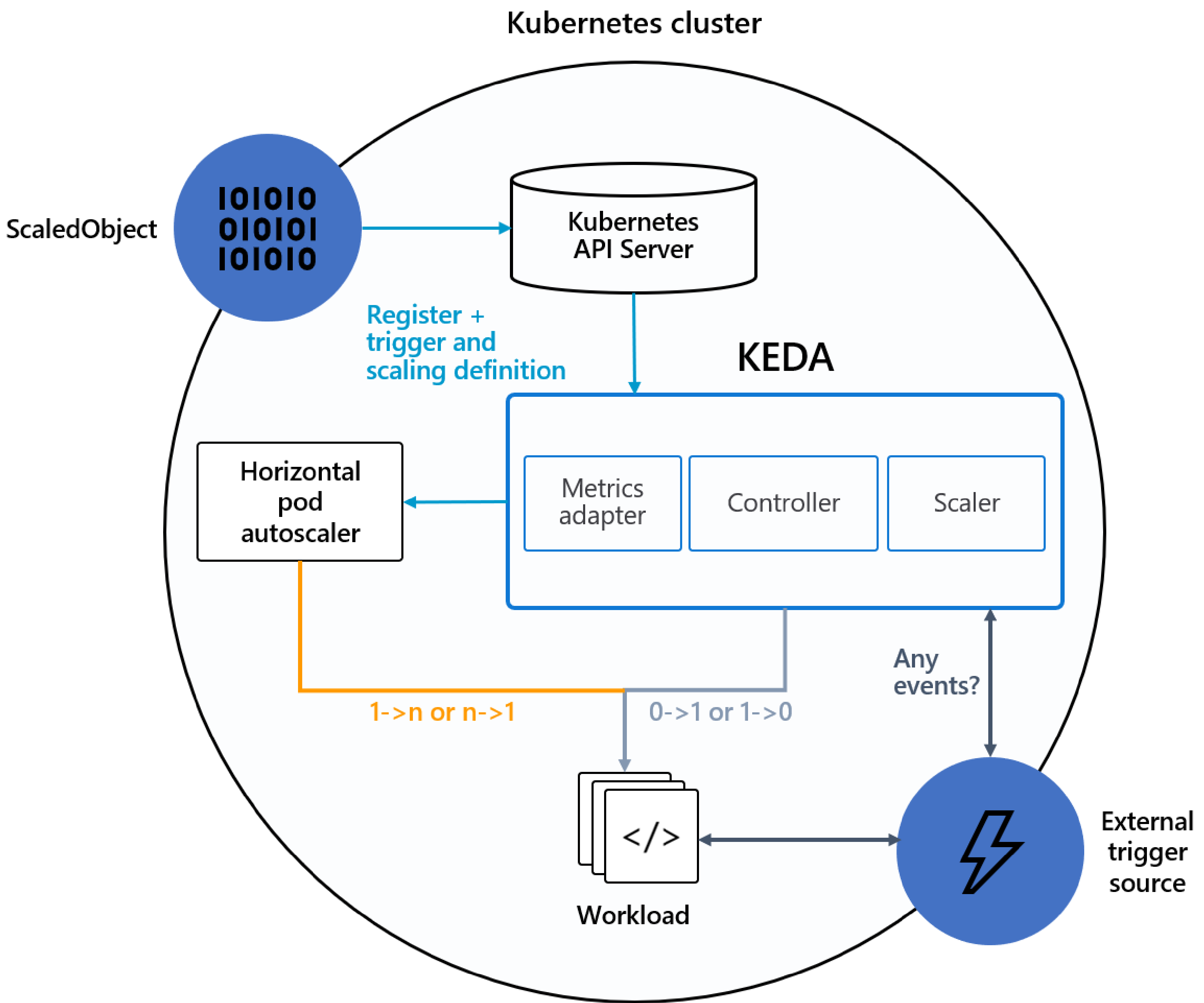
KEDA Resources



```
PS C:\Users\Wolfgang> kubectl get all -n keda
NAME                                              READY   STATUS    RESTARTS   AGE
pod/keda-operator-5748df494c-mxz9p                1/1     Running   0          124m
pod/keda-operator-metrics-apiserver-cb649dd48-jjhpc   1/1     Running   0          124m

NAME                                      TYPE        CLUSTER-IP     EXTERNAL-IP   PORT(S)          AGE
service/keda-operator-metrics-apiserver   ClusterIP   10.0.241.182   <none>        443/TCP,80/TCP   124m

NAME                                              READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/keda-operator                     1/1     1            1           124m
deployment.apps/keda-operator-metrics-apiserver   1/1     1            1           124m

NAME                                                        DESIRED   CURRENT   READY   AGE
replicaset.apps/keda-operator-5748df494c                    1         1         1       124m
replicaset.apps/keda-operator-metrics-apiserver-cb649dd48   1         1         1       124m
```

KEDA Architecture

- 2 components for KEDA
  - Agent or Operator
  - Metrics Server
- Uses HPA for scaling
- Seamless integration into existing architecture


Kubernetes cluster
ScaledObject
101010
010101
101010
Kubernetes API Server
Register + trigger and scaling definition
KEDA
Metrics adapter
Controller
Scaler
Horizontal pod autoscaler
Any events?
1->n or n->1
0->1 or 1->0
</>
Workload
External trigger source

KEDA Architecture

- 2 components for KEDA
  - Agent
  - Metrics Server
- Uses HPA for scaling
- Seamless integration into existing architecture
- 2 custom K8s resources for scaler
  - ScaledObject
  - TriggerAuthentication

WARSAW IT DAYS

ScaledObject

```yaml
apiVersion: keda.sh/v1alpha1
kind: ScaledObject
metadata:
  name: kedademoapi-scaler
```

```yaml
spec:
  scaleTargetRef:
    name: kedademoapi
  minReplicaCount: 0
  maxReplicaCount: 10
  pollingInterval: 30
  cooldownPeriod: 30
```

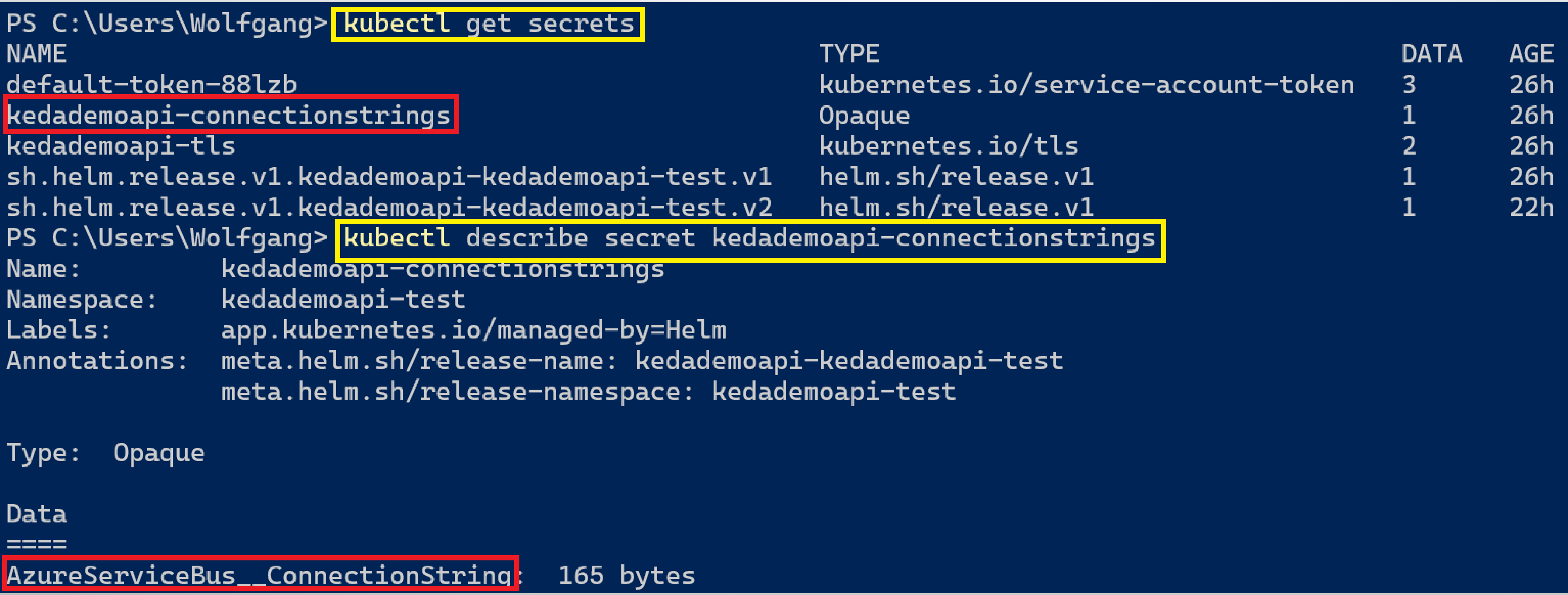
```yaml
triggers:
  - type: azure-servicebus
    metadata:
      queueName: KedaDemo
      queueLength: '5'
    authenticationRef:
      name: trigger-authentication-kedademoapi
```

TriggerAuthentication

```yaml
apiVersion: keda.sh/v1alpha1
kind: TriggerAuthentication
metadata:
  name: trigger-authentication-kedademoapi
spec:
  secretTargetRef:
  - parameter: connection
    name: kedademoapi-connectionstrings
    key: AzureServiceBus__ConnectionString
```

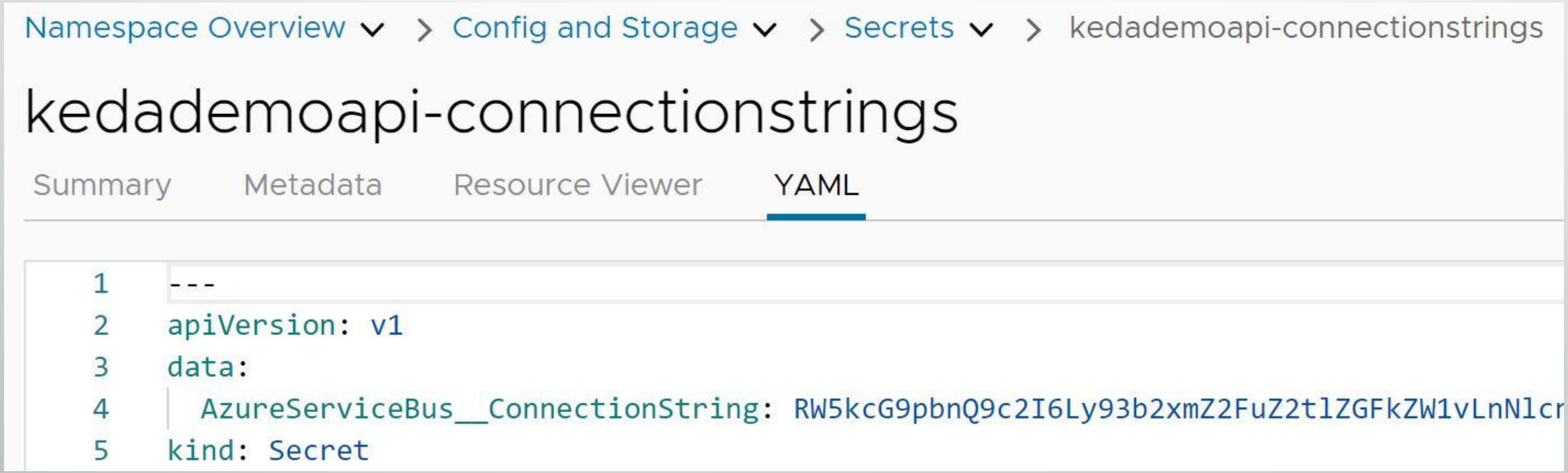
WARSAW IT DAYS

Kubernetes Secret



```
PS C:\Users\Wolfgang> kubectl get secrets
NAME                                               TYPE                                  DATA   AGE
default-token-88lzb                                kubernetes.io/service-account-token   3      26h
kedademoapi-connectionstrings                      Opaque                                1      26h
kedademoapi-tls                                    kubernetes.io/tls                     2      26h
sh.helm.release.v1.kedademoapi-kedademoapi-test.v1   helm.sh/release.v1                  1      26h
sh.helm.release.v1.kedademoapi-kedademoapi-test.v2   helm.sh/release.v1                  1      22h
PS C:\Users\Wolfgang> kubectl describe secret kedademoapi-connectionstrings
Name:         kedademoapi-connectionstrings
Namespace:    kedademoapi-test
Labels:       app.kubernetes.io/managed-by=Helm
Annotations:  meta.helm.sh/release-name: kedademoapi-kedademoapi-test
              meta.helm.sh/release-namespace: kedademoapi-test

Type:  Opaque

Data
====
AzureServiceBus__ConnectionString:  165 bytes
```

Kubernetes Secret



Namespace Overview > Config and Storage > Secrets > kedademoapi-connectionstrings

kedademoapi-connectionstrings

Summary
Metadata
Resource Viewer
YAML

```yaml
---
apiVersion: v1
data:
  AzureServiceBus__ConnectionString: RW5kcG9pbnQ9c2I6Ly93b2xmZ2FuZ2tlZGFkZW1vLnNlcr
kind: Secret
```

Azure DevOps Agent with KEDA

Demo: Scaling ADO Agent with KEDA

- Azure DevOps preparation
- Build Docker image
- Test locally
- Deploy to Kubernetes
- Apply KEDA scaling

Azure DevOps Limitations

- ADO Pipelines support scale to zero but need at least one agent registered
- ADO Pipelines can not queue a job with an empty agent pool
- Licensing limits parallel jobs

KEDA ADO Scaling Limitations

- Cancelling a pipeline does not stop running pods
- KEDA does not remove completed pods
- Azure DevOps does not remove offline agents from the agent pool

KEDA in Production

- Microsoft uses KEDA for Azure Services
  - Azure Container Apps
  - Azure App Services with Azure Arc

- KEDA 1.0.0 → 17. Nov 2019
- Currently 2.10
- Over 6k GitHub stars

Resources

- Demo Application
  - https://github.com/WolfgangOfner/MicroserviceDemo/tree/master/KedaDemoApi
  - https://github.com/WolfgangOfner/Ado-Agent-Keda
- KEDA
  - https://keda.sh
- KEDA GitHub
  - https://github.com/kedacore/keda
- KEDA Architecture Screenshot
  - https://keda.sh/docs/2.6/concepts/#architecture

WARSAW IT DAYS

Contact

Level Up your Kubernetes Scaling with KEDA

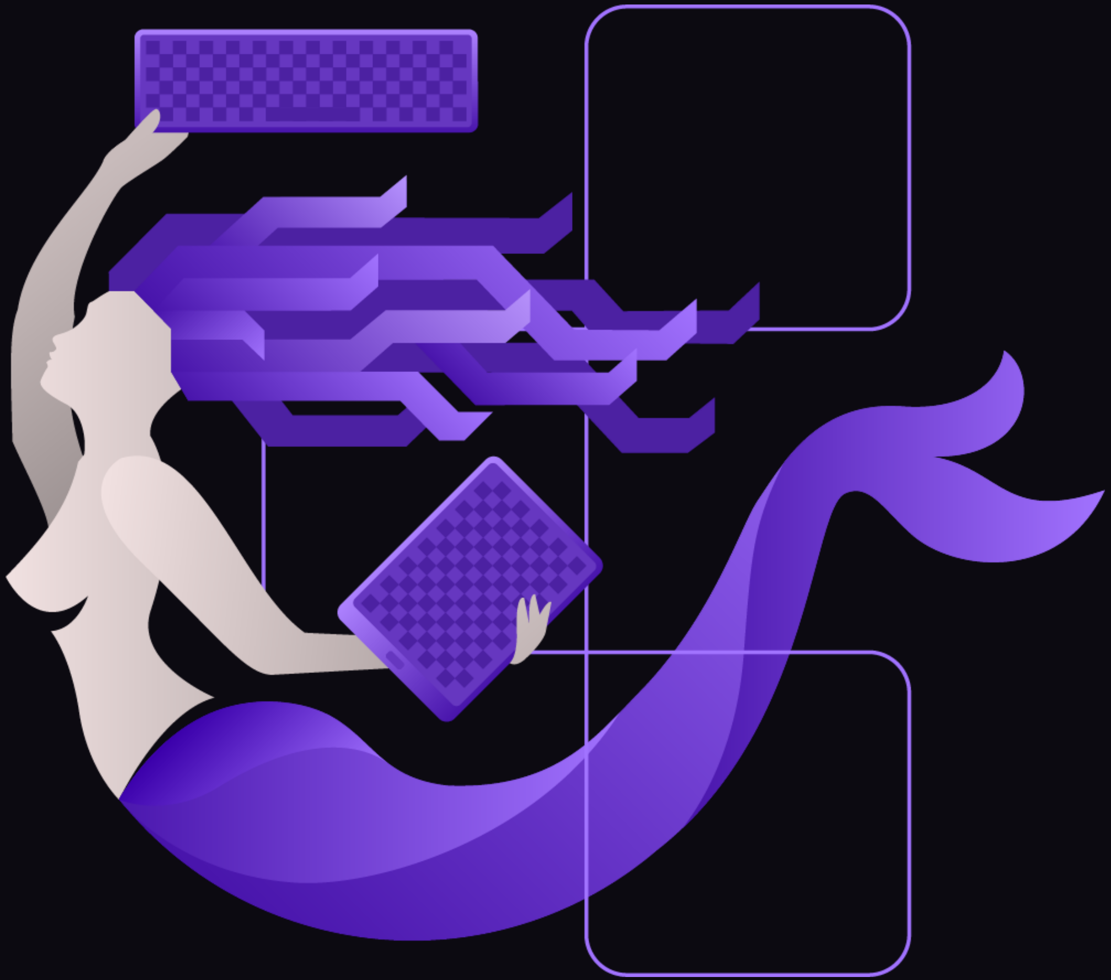
Wolfgang Ofner

https://programmingwithwolfgang.com

https://www.linkedin.com/in/wolfgangofner

https://twitter.com/wolfgang_ofner

WARSAW IT DAYS



Thank you for watching!

Remember to rate the presentation and leave your questions in the section below.

www.WarszawskieDniInformatyki.pl

31 March - 1 April 2023

Politechnika Warszawska + online

MAIN ORGANIZER: AcademicPartners FUNDACJA

ORGANIZING COMMITTEE: dozens of organizations from the IT / data science sector (full list on the event website)