Autoscaling trifft Analytics

Posit-Infrastrukturen im Kubernetes Cluster

Experten

Stefan Eikenbusch

Solution Architect

Josef Petermann

Solution Architect

eoda

daten ▸ wissen ▸ nutzen

Online Session

Autoscaling trifft Analytics

- Die Posit Produkte in Kubernetes
- Skalierbare Posit-Infrastruktur in Kubernetes
- "How to run" - Best Practices aus echten Projekten
- Herausforderungen und Vorteile

FRAGEN & ANTWORTEN

Wir freuen uns auf Ihre Fragen im Chat.

eoda
daten ▸ wissen ▸ nutzen

„Unternehmen verfügen über große Datenmengen.

Wieso nutzen sie das darin enthaltene Wissen nicht besser?“

–

Oliver Bracht | 2010

Mitgründer & CEO

eoda

daten ▸ wissen ▸ nutzen

INFRASTRUKTUR

LÖSUNGEN

STRATEGIE

DATA SCIENCE & KI

Mit uns von der Vision bis zum Produktiveinsatz

eoda
daten ▸ wissen ▸ nutzen

> 50 Mitarbeitende, > 400 Kunden – eine Mission: Aus Daten Mehrwerte generieren

eoda
daten ▸ wissen ▸ nutzen

Online Session

Autoscaling trifft Analytics

Die Posit Produkte in Kubernetes

Skalierbare Posit-Infrastruktur in Kubernetes

"How to run" - Best Practices aus echten Projekten

Herausforderungen und Vorteile

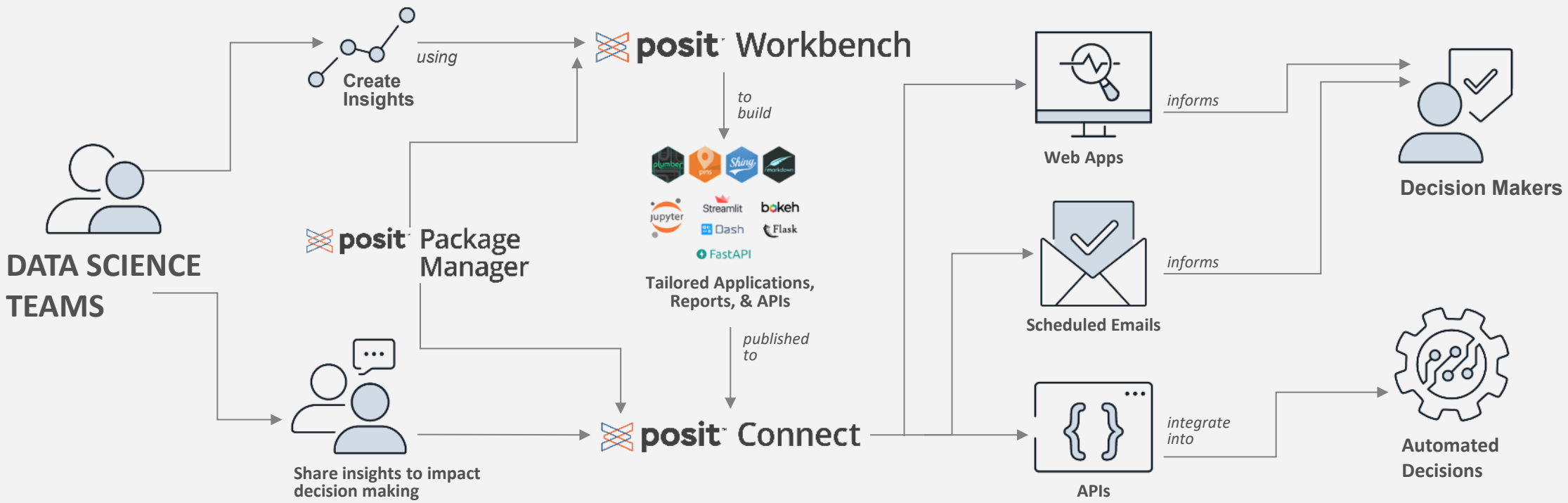
FRAGEN & ANTWORTEN

Wir freuen uns auf Ihre Fragen im Chat.

eoda

daten ▸ wissen ▸ nutzen

Der Posit Team Stack



Bildquelle: posit.co

Posit-Infrastruktur in Kubernetes

Alle Posit Produkte lassen sich auch in Kubernetes installieren

Installation via Helm-Chart (externer Link) aktuell der Standard

Ein sog. Posit Team Operator (externer Link) befindet sich aktuell von Posit in der Entwicklung und soll die Installation und Verwaltung in Zukunft vereinfachen

Posit stellt vorgefertigte Container Images bereit (externer Link)

eoda
daten ▸ wissen ▸ nutzen

Kubernetes

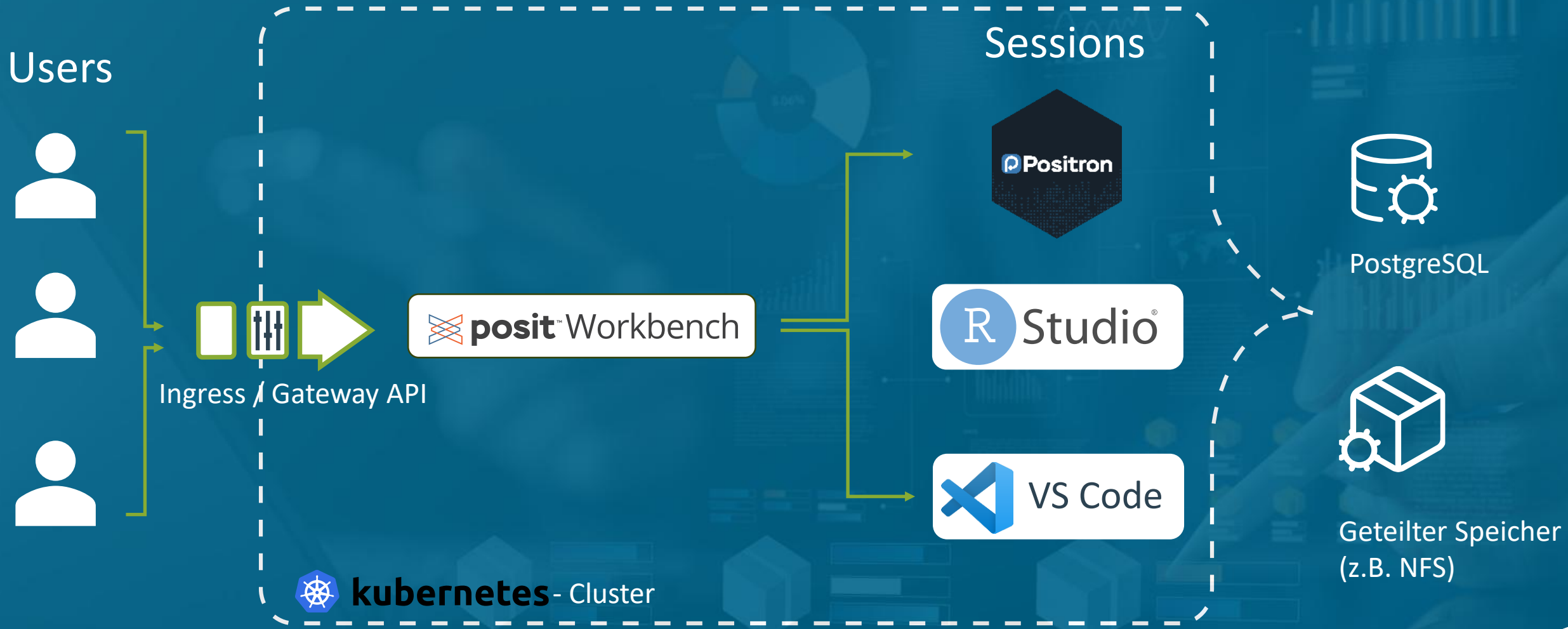
- **Container-Orchestrierungsplattform** für die *Bereitstellung*, *Skalierung* und *Verwaltung* von containerisierten Anwendungen
- Open-Source-Projekt, ursprünglich von Google entwickelt
- Zu den Kernfunktionen von Kubernetes gehören:
  - Automatische Skalierung – passt die Anzahl der Container basierend auf der Ressourcennutzung an
  - "Self-Healing" - fehlerhafte Container werden automatisch ersetzt
  - Load Balancing / Hochverfügbarkeit - Datenverkehr wird auf dem gesamten Cluster verteilt
  - Rolling Updates – Anwendungen werden (je nach Art) ohne Ausfallzeit aktualisiert

Kubernetes Komponenten

Zu den Hauptkomponenten von Kubernetes gehören

- Cluster – Gesamtheit der Kubernetes Ressourcen (control plane + worker nodes)
- Nodes – Linux Server auf denen die Container ausgeführt werden
- Pod – kleinste Einheit in Kubernetes, die aus mehreren Containern bestehen kann
- Deployments – definiert den gewünschten Zustand einer containerisierten Anwendung
- Services – definiert den Zugriff auf die Pods
- Ingress / Gateway API – Zugriff von außen

Posit Workbench in Kubernetes



Online Session

Autoscaling trifft Analytics

Die Posit Produkte in Kubernetes

Skalierbare Posit-Infrastruktur in Kubernetes

"How to run" - Best Practices aus echten Projekten

Herausforderungen und Vorteile

FRAGEN & ANTWORTEN

Wir freuen uns auf Ihre Fragen im Chat.

eoda
daten ▸ wissen ▸ nutzen

Skalierbare Workloads in Kubernetes

- Ermöglichen von großen Workloads zB. GPU Server
- Kosten im Griff behalten
- Pod Scaling vs. Node Scaling
- Scale to Zero: Teure Instanzen sollen nicht permanent laufen

Autoscaling Node Groups

Aktive Nutzer

"System" Node Group

- K8s Controlplane
- Workbench Server
- Cluster Autoscaler

Session Node Groups

- Standard Pool (always warm)
  - Standard Session
  - Standard Session
- High Memory Pool (min-size=0)
  - HiMem Session
- GPU Pool (min-size=0)
  - GPU Session

Cloud Infrastructure (VMs)

- Standard VM
- Standard VM
- High Memory VM
- GPU VM

eoda
daten ▸ wissen ▸ nutzen

Online Session

Autoscaling trifft Analytics

Die Posit Produkte in Kubernetes

Skalierbare Posit-Infrastruktur in Kubernetes

"How to run" - Best Practices aus echten Projekten

Herausforderungen und Vorteile

FRAGEN & ANTWORTEN

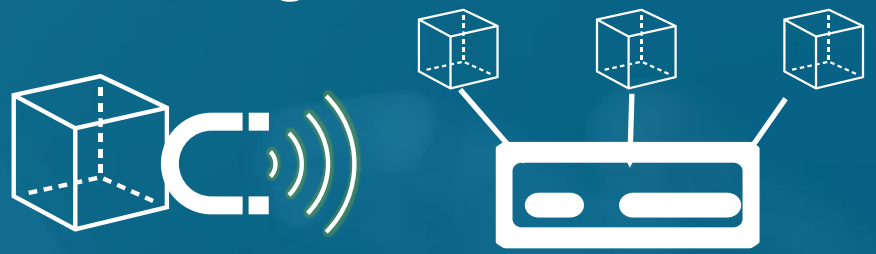
Wir freuen uns auf Ihre Fragen im Chat.

eoda

daten ▸ wissen ▸ nutzen

Scheduling – Warum Labels allein nicht reichen

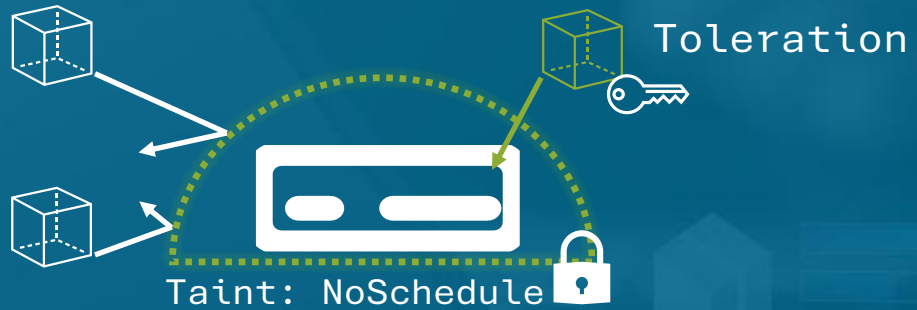
/ The Magnet



Labels ziehen an, blockieren aber nicht.

Ein Pod fordert einen spezifischen Node (`nodeSelector`), aber ohne Schutz können auch Standard-Pods auf dem teuren 128 GB- Node landen und Scale-Downs verhindern.

/ The Forcefield



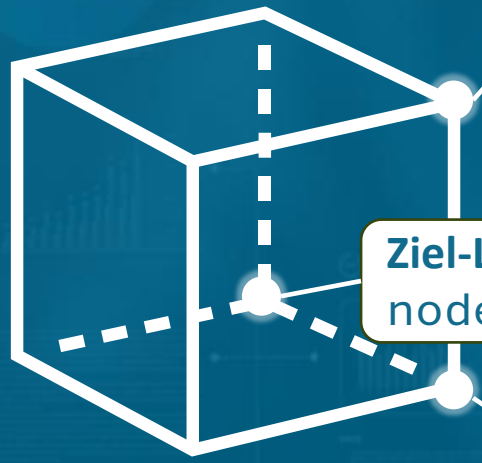
Taints stoßen ab, Tolerations gewähren Zutritt.

Ein Taint (NoSchedule) weist alle Pods ab. Nur Pods mit einer expliziten Toleration (Schlüssel) dürfen auf dem dedizierten Node ausgeführt werten

Workload-Integration: Posit Workbench Profile

3. Template-Engine (`job.tpl`):
Übersetzt `Constraints` in Kubernetes-Scheduling-Regeln (benötigt `use-templating=1`)

1. UI-Definition (`resources.conf`):
Definiert, was die User im Launcher sehen (z.B. GPU Session).



4. Generierter Pod

Ressourcen:
`limits.nvidia.com/gpu: 1`

Ziel-Label:
`nodeSelector:workload=gpu`

Schlüssel:
`Toleration` für den GPU-Taint

2. Logik & Berechtigungen (`profiles.conf`):
Steuert Zugriffe und definiert `placementConstraints`.

End-To-End-System-Workflow

Eine GPU-Session

Step 1: Request
User wählen GPU-Profil in der Posit-Workbench

Step 2: Pending
Job generiert einen Pod. Da `min=0` gibt es keine GPU-Nodes .
Pod bleibt auf „`pending`“.

Step 3: Autoscaler Match
Cluster Autoscaler liest die Labels des leeren Pools, findet einen Match und löst API-Call für `0->1` Skalierung aus

Step 4: Boot & Taint
Node wird gestartet. Bootstrap-Daemon appliziert den GPU-Taint

Step 5: Execution
Pool toleriert den Taint und wird gestartet. Die Session läuft nun.

Step 6: Scale-Down
User beendet Session Pod terminiert. Nach ca. 5 Minuten Leerlauf (`--scale-down-unneeded time`) skaliert der Autoscaler den Pool zurück auf 0

eoda
daten ▸ wissen ▸ nutzen

Online Session

Autoscaling trifft Analytics

Die Posit Produkte in Kubernetes

Skalierbare Posit-Infrastruktur in Kubernetes

"How to run" - Best Practices aus echten Projekten

Herausforderungen und Vorteile

FRAGEN & ANTWORTEN

Wir freuen uns auf Ihre Fragen im Chat.

eoda

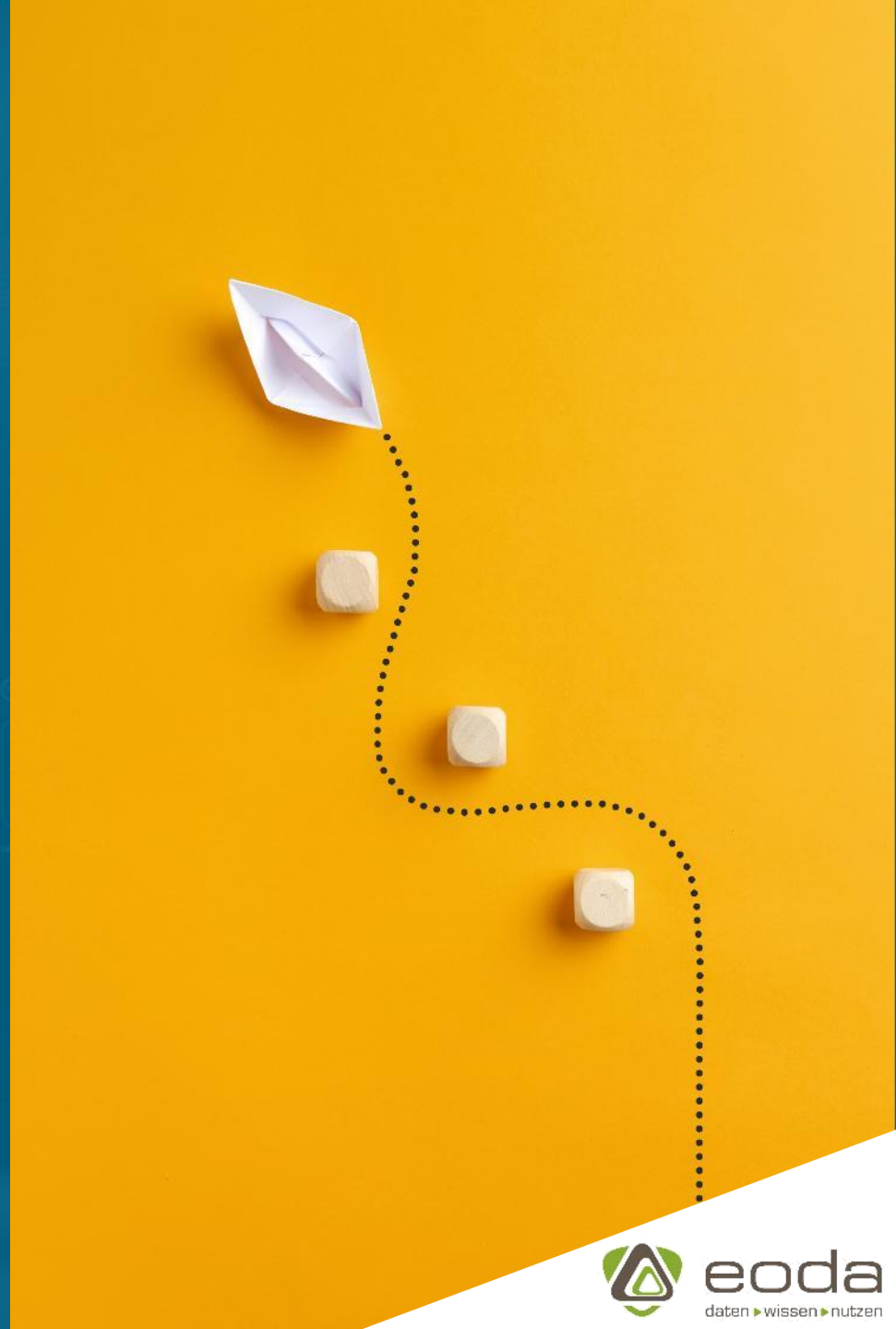
daten ▸ wissen ▸ nutzen

Vorteile

- Einfache horizontale Skalierung der Infrastruktur
- Ressourcen können bei Bedarf dem Cluster hinzugefügt werden
- "Scale to zero" - nicht benötigte Ressourcen müssen nicht dauerhaft vorgehalten werden
- Dadurch potentielle Kostenreduzierung bei der Bereitstellung einer Posit-Infrastruktur
- Bei der Nutzung von Kubernetes bereits eingebaute Hochverfügbarkeit

Herausforderungen

- Installation und Konfiguration nicht trivial
- Je nach Art der Kubernetes-Installation (händisch vs. verwaltet durch z.B. IONOS oder AWS) sind unterschiedliche Funktionen vorhanden
- Wenn umgebungsspezfische Einstellungen benötigt werden, müssen die von Posit bereitgestellten Container Images angepasst werden
  - Vorinstallierte R- oder Python-Pakete
  - CA-Zertifikate
- Ein CI/CD-Prozess und eine Container Registry werden idealerweise benötigt



Online Session

Autoscaling trifft Analytics

- Die Posit Produkte in Kubernetes
- Skalierbare Posit-Infrastruktur in Kubernetes
- "How to run" - Best Practices aus echten Projekten
- Herausforderungen und Vorteile

FRAGEN & ANTWORTEN

Wir freuen uns auf Ihre Fragen im Chat.

eoda
daten ▸ wissen ▸ nutzen

eoda begleitet Sie von der Idee bis zum erfolgreichen Betrieb

Beratung & Entwicklung	Posit Services	Assessment & Konzeption	Implemen-tierung	Betrieb & Support

eoda
daten ▸ wissen ▸ nutzen

Ausblick: Unsere nächste Online Session

Intuitiv, klar, barrierefrei: Wie Sie Daten und Analysen erfolgreich kommunizieren
Best Practices für Dashboards & Visualisierungen

16. Juni 2026 | 11:00-11:45 Uhr

Jetzt anmelden

Rückblick: Einblicke in zurückliegende Online Sessions erhalten Sie hier.

eoda
daten ▸ wissen ▸ nutzen

Ihre Ansprechpartner für Dateninfrastrukturen

Josef Petermann
Solution Architect

Stefan Eikenbusch
Solution Architect

+49 561 87948-370
infrastructure@eoda.de

eoda GmbH | Universitätsplatz 12 | 34127 Kassel

We are social

eoda
daten ▸ wissen ▸ nutzen