

KNIME Benutzerhandbuch

KNIME AG, Zürich, Schweiz

Version 1.3 (letzte Aktualisierung am 2024-01-29)



Inhaltsverzeichnis

Einleitung	Einleitung
Interaktion mit KNIME Rando	Interaktion mit KNIME Rando
View Edge Clusters	View Edge Clusters
KNIME	KNIME
KNIME Edge Licensing	KNIME Edge Licensing
Lizenzstaaten	Lizenzstaaten
Überprüfung des Cluster	Überprüfung des Cluster
Registrierung von Ausführungs	Registrierung von Ausführungs
Löschen von Ausführungs	Löschen von Ausführungs
Verwaltung von Inference Deploy	Verwaltung von Inference Deploy
Erstellung einer Inference Dep	Erstellung einer Inference Dep
Inference Deployment	Inference Deployment
Überprüfung einer Inference D	Überprüfung einer Inference D
Aktualisierung einer Inference	Aktualisierung einer Inference
Löschung einer Inference Dep	Löschung einer Inference Dep
Protokollierung	Protokollierung
Abruf von Ausführungs	Abruf von Ausführungs

Einleitung

Diese Anleitung beschreibt die Anforderungen, Überlegungen und Schritte zur Interaktion mit einem KNIME Edge Cluster und Erstellung von Inference Deployments.



Die folgenden Beispiele gehen davon aus, dass die Edge-Workflows kopiert wurden über einen Edge-Bereich auf dem KNIME Business Hub.

KNIME Edge ist eine verteilte, Container-basierte Plattform, die den Verbrauch von Modellen bewegt direkt dort, wo Daten erzeugt werden. Auf Kubernetes gebaut, bietet KNIME Edge die Fähigkeit Inferenzorientierte Workflows als hochverfügbare und skalierbare Endpunkte bereitstellen. Das ermöglicht einen hohen Durchsatz und eine geringe Latenz sowie die Dezentralisierung der Ausführung durch die Bereitstellung in Rechenzentren, Fertigungsanlagen, mehrere Cloud-Anbieter und mehr.

Ein oder mehrere KNIME Edge-Cluster können per Fernsteuerung von KNIME Business verwaltet werden Hub. Mit KNIME Business Hub, ein Benutzer kann Workflows für alle angeschlossenen auswählen und bereitstellen KNIME Edge-Cluster. Sobald ein Workflow bereitgestellt wird, erstellt KNIME Edge lokal Verbrauchsmaterial Endpunkte während der Ausführung, Skalierung, Standzeit, Widerstandsfähigkeit und mehr, um Modell zu gewährleisten Anwendung kann nahtlos mit der Nachfrage skaliert werden.

Interaktion mit KNIME Rand der Steuerung Ebene

Das KNIME Edge Control Plan Workflows müssen an KNIME Business Hub eingesetzt werden, um mit KNIME Edge Cluster(s) interagieren. Die Steuerungsebenen-Workflows (und Datenapps) bieten Fähigkeiten für:

- Registrierung und Verwaltung von KNIME Ausführungsbildern.
- Erstellung, Aktualisierung und Löschung von Inference Deployments.
- Ansicht und Überwachung von Edge-Clustern und laufenden Inference Deployments

Für Anweisungen zur Bereitstellung der KNIME Edge-Steuerungsebenen-Workflows für KNIME Business Hub, siehe [Installation von KNIME Edge mit KNIME Business Hub und Kurl](#).

Edge Cluster anzeigen

Dieser Bedienebenen-Workflow bietet einen umfassenden Überblick über verfügbare Informationen für Edge Cluster, die sowohl aktive Cluster als auch Cluster umfassen, die in die Vergangenheit. Es zeigt nicht nur Details über die Cluster, wie ihre Namen, Typen und

Kapazitäten, aber auch Einblicke in die beigefügten Lizenzen und Status.



Active clusters

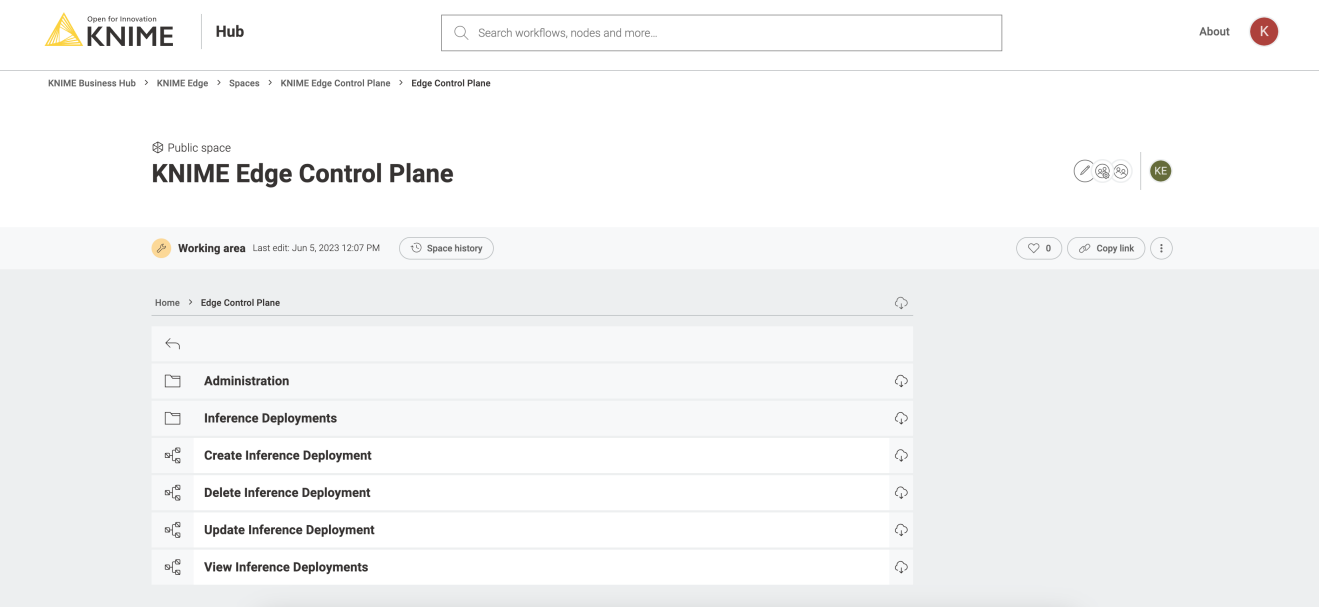
	Cluster ID	Cluster Name	Description	Status	Location	License Expiration Date	Cluster Capacity	License Name	License Cores
■	1	edge-cluster-1	some-cluster-description-1	Valid	us-east-1	0001-01-01T00:00:00.000Z	64	default-license-1	64
■	2	edge-cluster-2	some-cluster-description-2	Valid	us-east-1	0001-01-01T00:00:00.000Z	64	default-license-2	64

Inactive clusters

	Cluster ID	Cluster Name	Description	Status	Location	License Expiration Date	Cluster Capacity	License Name	License Cores
■	3	edge-cluster-3	some-cluster-description-2	Expired	us-east-1	0001-01-01T00:00:00.000Z	64	default-license-3	64

KNIME Rand

Nach der Einrichtung der KNIME Edge Workflows in einem neuen Raum auf dem KNIME Business Hub, dem Die Struktur dieses Raumes wird dem in der Abbildung unten entsprechen. Dieser Satz von Arbeitsabläufen und Verzeichnisse sind die "Kontrollebene" von KNIME Edge. Jeder Workflow ist eine Daten-App, die ermöglicht es Benutzern, mit KNIME Edge zu interagieren oder zu überwachen. KNIME Edge Installationsanleitung anzeigen für Details zum Hochladen der Steuerungsebenen-Workflows.






Wir beraten die Aufrechterhaltung der bestehenden Struktur, um unvorhergesehene Fehler im Zusammenhang mit dem Pfad zu verhindern Handhabung bei der Ausführung der Datenapps. Als allgemeine Leitlinie, die einzige Anforderung die Workflows "Create Inference Deployment" und die "Inference Deployments" Ordner auf der gleichen Ebene in der Verzeichnisstruktur. Die Daten-App ist konzipiert, um alle zu lesen Workflows in diesem Ordner. In der Tat sollte der Ordner "Inference Deployments" beide enthalten die Arbeitsabläufe, die eingesetzt wurden, und die, die bereit sind, eingesetzt zu werden. Das Ordner dient als zentraler Standort für die Verwaltung aller Workflows im Zusammenhang mit Inference Bereitstellungen.

KNIME Am Rande

Lizenzen

A KNIME Edge-Cluster kann in einem von drei verschiedenen Zuständen, angezeigt durch Farben:

-  **Grün:** Die Lizenz ist gültig und der Edge-Cluster funktioniert ohne Einschränkungen.
-  **Gelb:** Die Lizenz ist vor kurzem abgelaufen oder unterschrieben und arbeitet innerhalb einer Gnadenzeit. Der Edge-Cluster funktioniert immer noch ohne Einschränkungen, aber sein Zustand ist um rot zu werden.
-  **Rot:** Entweder wird keine Lizenz bereitgestellt oder die bereitgestellte Lizenz ist ungültig, abgelaufen oder unterbestimmt (und die Gnadenzeit ist vorbei). Bereits laufende Bereitstellungen halten die Arbeit, aber neue Einsatzbereiche werden abgelehnt.

Eine Lizenz wird als unterschrieben angesehen, wenn die Anzahl der bereitgestellten CPU-Kerne durch die Cluster überschreitet die Lizenzspezifikation. Beachten Sie, dass, wenn ein Knoten-Selektor konfiguriert ist, nur die CPU-Kerne der ausgewählten Knoten werden gezählt.

Überprüfung des Cluster- und Lizenzstatus

Der Lizenzzustand eines KNIME Edge-Clusters kann durch den Betrieb des Ansicht Rand
Cluster Arbeitsablauf. Darüber hinaus, bei der Erstellung neuer oder Aktualisierung bestehender Bereitstellungen, die der Zustand des Clusters wird mit einem farbigen Symbol (wie unten dargestellt) angezeigt. Ein Cluster mit einem rotes Icon wird neue Bereitstellungen ablehnen.

Select the edge locations for deployment:

Excludes

edge-cluster-3

>

>>

<

<<

Includes

edge-cluster-1

edge-cluster-2

Registrierung von Ausführungsbildern

Bevor eine Inference-Bereitstellung erstellt und einem KNIME Edge-Cluster zugeordnet werden kann, ist eine oder mehr Laufzeitausführungsbilder (d.h. Dockerbilder) müssen mit dem Cluster registriert werden. Überprüfen Sie die **knime-execution** Projekt zur [KNIME Art und Weise](#) für eine Liste von KNIME veröffentlichte Laufzeitbilder. Kontaktieren Sie Ihren Kundenbetreuer, wenn Sie Zugriff benötigen.

- KNIME Unternehmen Hub zieht nicht direkt oder nutzt diese Bilder, aber sie dienen als Katalog verfügbarer Laufzeitbilder, die für Workflows verwendet werden in KNIME Edge Clustern eingesetzt.
- Die von KNIME zur Verfügung gestellten Standardbilder können auf zusätzliche Erweiterungen und Fähigkeiten. Kontakt Support@knime.com wenn Sie brauchen Hilfe.

Um ein neues Ausführungsbild zu registrieren, führen Sie das **Hinzufügen von Execution Images**-App in der **Verwaltung** Verzeichnis der Stauerebenen-Workflows.

- Die **Docker Bild** Das Feld erwartet eine vollständige Registrierungs-URL für ein Bild, einschließlich der Bild-Tag (z. `registr.hub.knime.com/knime-execution/knime-inference-Agent:1.3.0-4.7.2`) Die **Warenbezeichnung**feld ist optional.



Add available KNIME runtime Docker images so that they can be assigned to Inference Deployments. Enter the full registry URL (incl. image tag) for the execution image and optionally a description.

Docker image:

registry.hub.knime.com/knime-execution/knime-inference-agent:1.3.0-4.7.2

Description:

Edge 1.3.0 - AP 4.7.2

Ein erfolgreich registriertes Bild wird ein grünes Statussymbol zurückgeben.

KNIME Edge

Result

	Image name	Image description	Error cause
	registry.hub.knime.com/knime-execution/knime-inference-agent:1.3.0-4.7.2	Edge 1.3.0 - AP 4.7.2	None

Result legend
 Image created
 Image not created

Relevante Fehler werden angezeigt, wenn Bilder aus irgendeinem Grund nicht registriert werden.

KNIME Edge

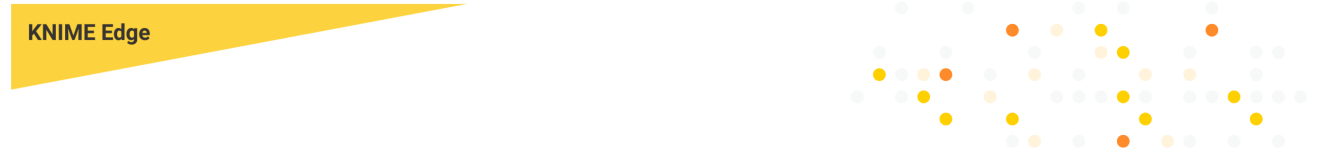
Result

	Image name	Image description	Error cause
	registry.hub.knime.com/knime-execution/knime-inference-agent:1.2.0-4.7.2	Edge 1.2.0 - AP 4.7.2	error : image name must be unique

Result legend
 Image created
 Image not created

Löschen von Ausführungsbildern

Die **Löschen von Execution Image** Daten App kann verwendet werden, um Bilder zu entfernen und zu reinigen, die nicht mehr erforderlich. Die Daten-App listet alle registrierten Ausführungsbilder auf und ermöglicht den Benutzern Wählen Sie die, die gelöscht werden sollten.



Select the images to delete

Only images that are not used by any Inference Deployment can be deleted.

<input type="checkbox"/>	Image name	⇅	Image description	⇅
<input type="checkbox"/>	registry.hub.knime.com/knime-execution/knime-inference-agent:1.2.0-4.7.2		Edge 1.2.0 - AP 4.7.2	

☐ Bilder, die von einem oder mehreren Inference Deployments referiert werden, können nicht gestrichen.



Deletion results

Show 5 entries

	Image name	⇅	Image description	⇅	Error Cause	⇅
<input checked="" type="checkbox"/>	registry.hub.knime.com/knime-execution/knime-inference-agent:1.2.0-4.7.2		Edge 1.2.0 - AP 4.7.2		"error" : "image is referenced by one or more deployment(s) and cannot be deleted"	

Previous1Next

Result legend

☒ Image deleted

☐ Image not deleted

The non-deletable images referenced by a deployment

Image name	⇅	Deployment name	⇅
registry.hub.knime.com/knime-execution/knime-inference-agent:1.2.0-4.7.2		some-deployment	

Verwalten von Inference Deployments

Schaffung einer Inferenz-Beschäftigung

Mit **Inference Deployment erstellen** Daten-App, können Sie eine neue Inference Deployment erstellen durch Auswahl eines Workflows zur Bereitstellung.

ANHANG Navigieren Sie auf dem KNIME Business Hub zum KNIME Edge-Steuerungsplan.

2. Im Raum navigieren Sie zum **Inference Deployment erstellen** daten app und ausführen.

3. Eine Daten-App zeigt eine Handvoll kritischer Parameter (siehe unten).

L 347 vom 20.12.2013, S. 1). Die restlichen Parameter können mit Standardwerten ausgefüllt oder nach Bedarf eingestellt werden.

Parameter	Warenbezeichnung
Name der Bereitstellung	Der hier bereitgestellte Wert wird zur API Endpunkt für den Inferenzeinsatz.
Beschreibung der Bereitstellung	Freundlicher Text, um den Zweck zu beschreiben intent der Inferenz-Bereitstellung.
Wählen Sie einen Workflow zur Bereitstellung	Ein visueller Dateinavigator zum Finden und Auswählen der Ziel-Workflow zur Bereitstellung.
Wählen Sie eine Version des Workflows aus	Die Workflow-Version zum Einsatz als Inference-Bereitstellung.
Wählen Sie eine Basisbildinstanz	Das Ausführungsbild, das die Inference-Bereitstellung.
Wählen Sie die Randplätze für die Bereitstellung	Das KNIME Rand-Cluster, die Inferenz-Bereitstellung wird weiterverbreitet.

KNIME Edge

Deployment name (duplicates not allowed)

some-deployment

Deployment description

This is my deployment.

Select a workflow to deploy. A new version of the selected workflow will be created and used. The deployment description will be set as comment of the version.

Calculate Length

Sentiment_Predictor

Select a version of the workflow:

<Create and deploy new version>

Get versions

Select a base image instance:

registry.hub.knime.com/knime-execution/knime-inference-agent:1.2.0-4.7.2

Select the edge locations for deployment:

Excludes

No entries in this list

Includes

edge

Logging level

WARN

Minimum size

1

Desired size

1

Maximum size

2

Target CPU utilization threshold

75

Request memory allocation

1024

Limit memory allocation (MB)

2048

Request CPU allocation

1000

Limit of CPU usage

2000

☒ Executor initialization

☒ Enable workflow polling

Polling frequency (empty or zero values mean no polling)

0h0m0s0ms

Validate inputs

Validation Message

All inputs are correct

Inferenzentwicklungen anzeigen

ANHANG Navigieren Sie auf dem KNIME Business Hub zum KNIME Edge-Steuerungsplan.

2. Im Raum navigieren
- Inference Deployments anzeigen
- Workflow und ausführen.
3. Eine Ansicht wird angezeigt, mit der Sie einen Edge-Cluster auswählen und alle Inferenz anzeigen können dem ausgewählten Cluster zugewiesene Bereitstellungen.

Überprüfung eines Inferenzeinsatzes

Der folgende Befehl, der von einem Edge-Cluster-Standort ausgeführt werden soll, gibt alle Bereitstellungen zurück die aktiv sind:

```
% curl -sL http://localhost:8081 | jq
{
  "InferenceDeployments": [
    {
      "endpoint": "http://localhost:8081/"
    }
  ]
}
```

Der folgende Befehl zeigt, wie man mit dem eingesetzten Workflow interagiert. Die Anfrage JSON ist spezifisch für die oben genannten REST API für Sentiment Analyse Arbeitsablauf und müsste für andere Workflows angepasst werden:

```
% curl -s -X POST -H "Content-Type: application/json" --data '{"content":["happy happy
Freude", "sad bad mad mad"]}' http://localhost:8081/ |jq
(
  {
    "Prädiktion (Dokumentenklasse)": 0.6459559392130747,
    "Prädiktion (Sentiment)": "Positive"
  },
  {
    "Prädiktion (Dokumentenklasse)": 0.5321978494130959,
    "Prädiktion (Sentiment)": "Positive"
  }
]
```

Aktualisierung einer Inferenz-Beschäftigung

ANHANG Navigieren Sie auf dem KNIME Business Hub zum KNIME Edge-Steuerungsplan.

2. Im Raum navigieren Sie zum **Update Inference Deployment** Daten app und ausführen.
3. Eine Ansicht wird angezeigt, mit der Sie die Parameter einer vorhandenen Inferenz ändern können
Bereitstellung. Der mit der Inferenz-Bereitstellung verbundene Workflow kann nicht
geändert, jedoch kann eine andere Version des Workflows gewählt werden.

Eine Inferenzentschädigung löschen

ANHANG Navigieren Sie auf dem KNIME Business Hub zum KNIME Edge-Steuerungsplan.

2. Im Raum navigieren Sie zum **Inference Deployment löschen** Workflow und ausführen.
3. Eine Ansicht wird angezeigt, mit der Sie die gewünschte Inferenz-Bereitstellung auswählen können
löschen.

Protokoll

Inference Deployments erstellen Protokolle während der Ausführung, die hilfreich sein können, um Fehler zu beheben
Probleme. Jedes Mal, wenn ein Workflow ausgeführt wird, werden Ereignisse wie in KNIME Analytics protokolliert
Plattform. Bei der Erstellung oder Aktualisierung eines Inference Deployment (siehe oben), der Protokollierungsebene
kann entweder DEBUG, INFO, WARN oder ERROR angegeben werden. Von ERROR bis DEBUG die Protokolle
werden ausführlicher und verbose. Standardmäßig wird der Protokollierungspegel auf WARN eingestellt. Wenn Sie
Probleme mit einer Inferenz-Deployment zu begegnen und detailliertere Informationen zu benötigen, Sie
kann die Inference Deployment (siehe oben) aktualisieren und eine mehr Verbose-Logging-Ebene festlegen.

Abruf von Ausführungsprotokollen aus einer Inference Deployment

Um die Protokolle für Inference Deployments inspizieren zu können, benötigen Sie das Terminal.

Der folgende Befehl zeigt, wie man Logs vom Inference Deployment erhält; in dieses Beispiel entspricht das Protokoll dem zuvor ausgeführten Scoring-Job gegen

Vorbemerkung [REST API für Sentiment Analyse](#) Arbeitsablauf:

```
> kubectl [-n -
01. September 2021 3:31:02 PM org.apache.cxf.bus.osgi.CXFExtensionBundleListener addExtensions
INFO: Hinzufügen der Erweiterungen von Bundle org.apache.cxf.cxf-rt-frontend-jaxrs (388)
[org.apache.cxf.jaxrs.JAXRSBindingFactory]
01. September 2021 3:31:02 PM org.apache.cxf.bus.osgi.CXFExtensionBundleListener addExtensions
INFO: Hinzufügen der Erweiterungen von Bundle org.apache.cxf.cxf-rt-transport-http (391)
[org.apache.cxf.transport.http.HTTPTransportFactory,
org.apache.cxf.transport.http.HTTPWSDLExtensionLoader,
org.apache.cxf.transport.http.policy.HTTPClientAssertionBuilder,
org.apache.cxf.transport.http.policy.HTTPServerAssertionBuilder,
org.apache.cxf.transport.http.policy.NoOpPolicyInterceptorProvider]
01. September 2021 3:31:02 PM org.apache.cxf.bus.osgi.CXFExtensionBundleListener addExtensions
INFO: Hinzufügen der Erweiterungen von Bundle org.apache.cxf.cxf-rt-transport-http-hc (392)
[org.apache.cxf.transport.http.HTTPConduitFactory,
org.apache.cxf.transport.ConduitInitiator]
Initialisierung von Scoring Agent...
WARN KNIME-Worker-1-Document Vector Applier 5:303 Node Die Strukturen beider
aktive Eingabedatentabellen sind nicht kompatibel.
WARN KNIME-Worker-2-Kategorie zur Klasse 5:275 Node Die Strukturen beider aktiven
Eingabedatentabellen sind nicht kompatibel.
WARN KNIME-Worker-0-Gradient Boosted Trees Predictor (deprecated) 5:371 Node
Die Strukturen beider aktiven Eingabedatentabellen sind nicht kompatibel.
WARN KNIME-Worker-1-Regelmaschine 5:352 Node Die Strukturen beider aktiven Eingänge
Datentabellen sind nicht kompatibel.
```

KNIME AG
Talacker 50
8001 Zürich, Schweiz
www.knime.com
Info@knime.com