

KNIME Edge User Guide

KNIME AG, Zurich, Switzerland
Version 1.3 (last updated on 2024-01-29)



Table of Contents

| | |
|--|----|
| Introduction..... | 1 |
| Interacting with KNIME Edge from the control plane..... | 1 |
| View Edge Clusters | 1 |
| KNIME Edge | 2 |
| KNIME Edge Licensing..... | 3 |
| License States | 3 |
| Inspecting the Cluster and License Status..... | 3 |
| Registering execution images..... | 4 |
| Deleting execution images | 5 |
| Managing Inference Deployments | 6 |
| Creating an Inference Deployment | 6 |
| Viewing Inference Deployments | 8 |
| Verifying an Inference Deployment | 8 |
| Updating an Inference Deployment..... | 9 |
| Deleting an Inference Deployment | 9 |
| Logging | 9 |
| Retrieving execution logs from an Inference Deployment | 10 |

Introduction

This guide outlines the requirements, considerations, and steps for interacting with a KNIME Edge cluster and creating Inference Deployments.



The examples below will assume that the Edge workflows have been copied over to an Edge space on KNIME Business Hub.

KNIME Edge is a distributed, container-based platform that moves consumption of models directly to where data is generated. Built on top of Kubernetes, KNIME Edge offers the ability to deploy inference-oriented workflows as highly available and scalable endpoints. This allows for high throughput and low latency while also decentralizing execution by deploying into datacenters, manufacturing facilities, multiple cloud providers, and more.

One or more KNIME Edge clusters can be remotely managed by leveraging KNIME Business Hub. Using KNIME Business Hub, a user can select and deploy workflows to any connected KNIME Edge clusters. Once a workflow is deployed, KNIME Edge creates locally consumable endpoints while managing execution, scaling, uptime, resiliency and more to ensure model application can scale seamlessly with demand.

Interacting with KNIME Edge from the control plane

The KNIME Edge control plane workflows must be deployed to KNIME Business Hub in order to interact with KNIME Edge cluster(s). The control plane workflows (and data apps) provide capabilities for:

- Registering and managing KNIME execution images.
- Creating, updating, and deleting Inference Deployments.
- Viewing and monitoring Edge clusters and running Inference Deployments

For instructions on deploying the KNIME Edge control plane workflows for KNIME Business Hub, see [Installing KNIME Edge with KNIME Business Hub and Kurl](#).

View Edge Clusters

This control plane workflow provides a comprehensive view of available information for Edge clusters, encompassing both currently active clusters and clusters that have been active in the past. It not only displays details about the clusters, such as their names, types, and

capacities, but also provides insights into the attached licenses and status.



Active clusters

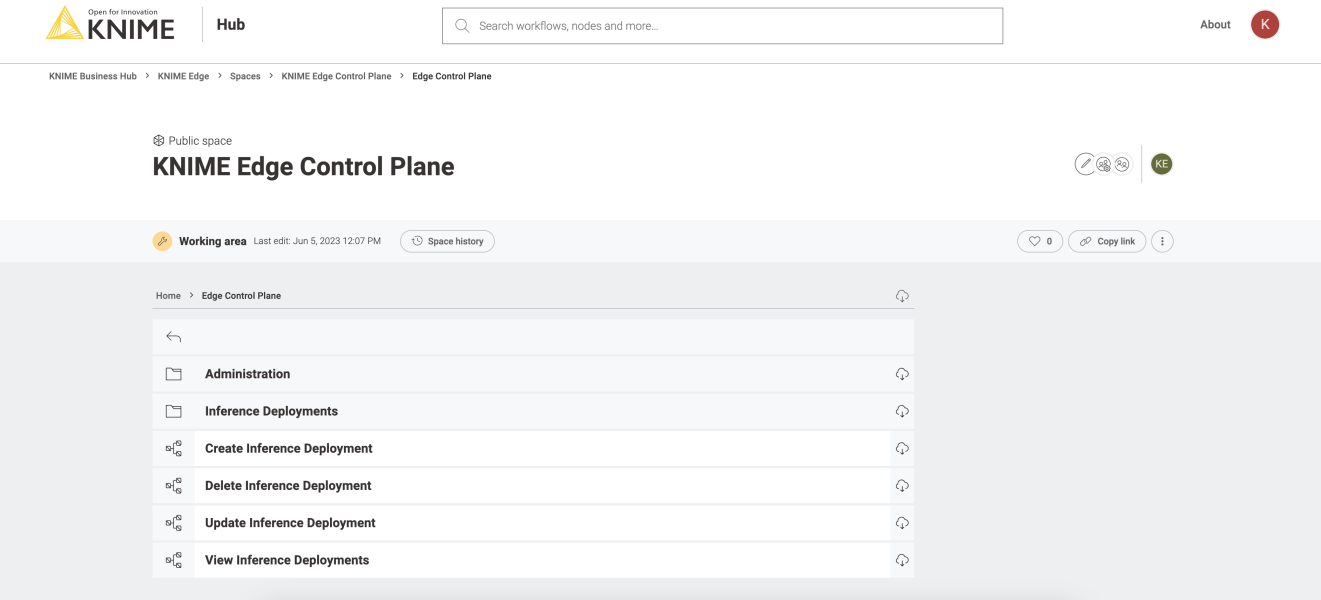
| | Cluster ID | Cluster Name | Description | Status | Location | License Expiration Date | Cluster Capacity | License Name | License Cores |
|---|------------|----------------|----------------------------|--------|-----------|--------------------------|------------------|-------------------|---------------|
| ■ | 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 Edge

After setting the KNIME Edge workflows up in a new space on the KNIME Business Hub, the structure of that space will match the one in the screenshot below. This set of workflows and directories is named the “control plane” of KNIME Edge. Each workflow is a data app that allows users to interact with or monitor KNIME Edge. See the KNIME Edge Installation Guide for details on uploading the control plane workflows.






We advise maintaining the existing structure to prevent any unforeseen errors related to path handling during the execution of the data apps. As a general guideline, the only requirement is to have the "Create Inference Deployment" workflows and the "Inference Deployments" folder at the same level in the directory structure. The data app is designed to read all workflows within that folder. In fact, the "Inference Deployments" folder should contain both the workflows that have been deployed and the ones that are ready to be deployed. This folder serves as a central location for managing all the workflows related to Inference deployments.

KNIME Edge Licensing

License States



A KNIME Edge cluster can be in one of three different states, indicated by colors:

-  **Green:** The license is valid and the Edge cluster works without restrictions.
-  **Yellow:** The license has recently expired or become underspecified and is working within a grace period. The Edge cluster still works without restrictions but its state is about to turn red.
-  **Red:** Either no license is provided or the provided license is invalid, expired or underspecified (and the grace period is over). Already running deployments keep working but new deployments are rejected.

A license is considered to be underspecified if the number of provisioned CPU cores by the cluster exceeds the license specification. Note that if a node selector is configured, only the CPU cores of the selected nodes will be counted.

Inspecting the Cluster and License Status

The license state of a KNIME Edge cluster can be inspected by running the `View Edge Clusters` workflow. Additionally, when creating new or updating existing deployments, the state of the cluster will be indicated with a colored icon (as depicted below). A cluster with a red icon will reject new deployments.

Select the edge locations for deployment:**Excludes** edge-cluster-3**Includes** edge-cluster-1
 edge-cluster-2

Registering execution images

Before an inference deployment can be created and assigned to a KNIME Edge cluster, one or more runtime execution images (i.e. docker images) must be registered with the cluster. Check the **knime-execution** project on the [KNIME Artifact Registry](#) for a list of KNIME published runtime images. Contact your customer care representative if you need access.



KNIME Business Hub does not directly pull or use these images, but they serve as a catalog of available runtime images that will be used for workflows deployed to KNIME Edge clusters.



The standard images provided by KNIME can be extended to include additional extensions and capabilities. Contact support@knime.com if you need assistance.

To register a new execution image, run the **Add Execution Image** data app found in the **Administration** directory of the control plane workflows.



The **Docker image** field expects a full registry URL for an image, including the image tag (e.g. `registry.hub.knime.com/knime-execution/knime-inference-agent:1.3.0-4.7.2`). The **Description** field is optional.

KNIME Edge

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

A successfully registered image will return a green status icon.

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

Relevant errors will display if images fail to be registered for any reason.

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

Deleting execution images

The **Delete Execution Image** data app can be used to remove and clean up images that are not required anymore. The data app lists all registered execution images and allows users to select the ones that should be deleted.



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 |



Images that are referenced by one or more Inference Deployments cannot be deleted.



Deletion results

Show 5 entries

| | 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 is referenced by one or more deployment(s) and cannot be deleted" |

Previous 1 Next

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 |

Managing Inference Deployments


Creating an Inference Deployment

With the **Create Inference Deployment** data app, you can create a new Inference Deployment by selecting a workflow to deploy.

1. Navigate to the KNIME Edge control plane workflows on KNIME Business Hub.
2. In the space, navigate to the **Create Inference Deployment** data app and execute it.
3. A data app will display a handful of critical parameters (see below).
4. The remaining parameters can be filled out using default values or adjusted as needed.

| Parameter | Description |
|--|--|
| Deployment name | The value provided here becomes the API endpoint for the inference deployment. |
| Deployment description | Friendly text to describe the purpose & intent of the inference deployment. |
| Select a workflow to deploy | A visual file navigator to find and select the target workflow to deploy. |
| Select a version of the workflow | The workflow version to deploy as an inference deployment. |
| Select a base image instance | The execution image which will host the inference deployment. |
| Select the edge locations for deployment | This KNIME Edge cluster(s) which the inference deployment will be propagated to. |

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

Viewing Inference Deployments

1. Navigate to the KNIME Edge control plane workflows on KNIME Business Hub.
2. In the space, navigate to **View Inference Deployments** workflow and execute it.
3. A view will display that allows you to select an Edge cluster and view all inference deployments assigned to the selected cluster.

Verifying an Inference Deployment

The following command, to be run from an edge cluster location, returns any deployments that are active:

```
% curl -sL http://localhost:8081 | jq
{
  "inferenceDeployments": [
    {
      "endpoint": "http://localhost:8081/<edge_deployment_name>"
    }
  ]
}
```

The following command demonstrates how to interact with the deployed workflow. The request JSON is specific to the aforementioned **REST API for Sentiment Analysis** workflow and would need to be adapted for other workflows:

```
% curl -s -X POST -H "Content-Type: application/json" --data '{"content":["happy happy joy joy","sad bad mad mad"]}' http://localhost:8081/<edge_deployment_name> |jq
[
  {
    "Prediction (Document class) (Confidence)": 0.6459559392130747,
    "Prediction (Sentiment)": "Positive"
  },
  {
    "Prediction (Document class) (Confidence)": 0.5321978494130959,
    "Prediction (Sentiment)": "Positive"
  }
]
```

Updating an Inference Deployment

1. Navigate to the KNIME Edge control plane workflows on KNIME Business Hub.
2. In the space, navigate to the **Update Inference Deployment** data app and execute it.
3. A view will display that allows you to change the parameters of an existing inference deployment. The workflow associated with the inference deployment cannot be changed, but a different version of the workflow can be selected.

Deleting an Inference Deployment

1. Navigate to the KNIME Edge control plane workflows on KNIME Business Hub.
2. In the space, navigate to the **Delete Inference Deployment** workflow and execute it.
3. A view will display that allows you to select the inference deployment you want to delete.

Logging

Inference Deployments create logs during execution that can be helpful to troubleshoot problems. Each time a workflow is executed, events are logged similar as in KNIME Analytics Platform. When creating or updating an Inference Deployment (see above), the logging level can be specified to be either DEBUG, INFO, WARN or ERROR. From ERROR to DEBUG the logs become more detailed and verbose. By default, the logging level is set to WARN. If you encounter problems with an Inference Deployment and need more detailed information, you can update the Inference Deployment (see above) and set a more verbose logging level.

Retrieving execution logs from an Inference Deployment

To inspect the logs for Inference Deployments you will need to use the terminal.

The following command demonstrates how to get logs from the Inference Deployment; in this example the log corresponds to the previously run scoring job run against the aforementioned **REST API for Sentiment Analysis** workflow:

```
> kubectl [-n <namespace>] logs <edge_deployment_name>-<podID>
Sep 01, 2021 3:31:02 PM org.apache.cxf.bus.osgi.CXFExtensionBundleListener addExtensions
INFO: Adding the extensions from bundle org.apache.cxf.cxf-rt-frontend-jaxrs (388)
[org.apache.cxf.jaxrs.JAXRSBindingFactory]
Sep 01, 2021 3:31:02 PM org.apache.cxf.bus.osgi.CXFExtensionBundleListener addExtensions
INFO: Adding the extensions from 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]
Sep 01, 2021 3:31:02 PM org.apache.cxf.bus.osgi.CXFExtensionBundleListener addExtensions
INFO: Adding the extensions from bundle org.apache.cxf.cxf-rt-transport-http-hc (392)
[org.apache.cxf.transport.http.HTTPConduitFactory,
org.apache.cxf.transport.ConduitInitiator]
Initializing Scoring Agent...
WARN      KNIME-Worker-1-Document Vector Applier 5:303 Node      The structures of both
active input data tables are not compatible.
WARN      KNIME-Worker-2-Category To Class 5:275 Node      The structures of both active
input data tables are not compatible.
WARN      KNIME-Worker-0-Gradient Boosted Trees Predictor (deprecated) 5:371 Node
The structures of both active input data tables are not compatible.
WARN      KNIME-Worker-1-Rule Engine 5:352 Node      The structures of both active input
data tables are not compatible.
```

KNIME AG
Talacker 50
8001 Zurich, Switzerland
www.knime.com
info@knime.com