RHOAI Admin Config

Agenda

- Users and groups
- Custom notebook images
- Custom serving runtimes
- Accelerators and GPUs
- Managing RHOAI resources

Authentication

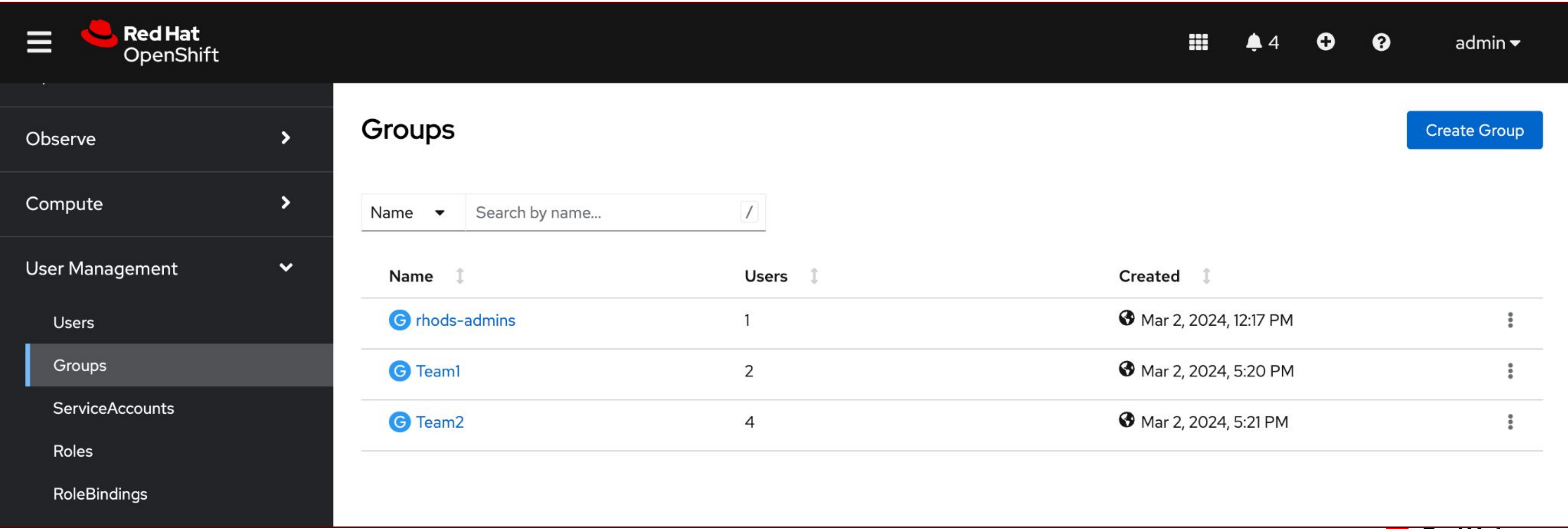
OpenShift AI - Authentication

- OpenShift AI uses the same Authentication mechanisms used by the underlying OpenShift Cluster
- See this page for more details on identity providers
  - htpasswd
  - Keystone
  - LDAP
  - Basic authentication
  - Request header
  - GitHub or GitHub Enterprise
  - GitLab
  - Google
  - OpenID Connect

Users and groups

OpenShift AI - Group Membership

- Group membership is managed at the OpenShift Level:

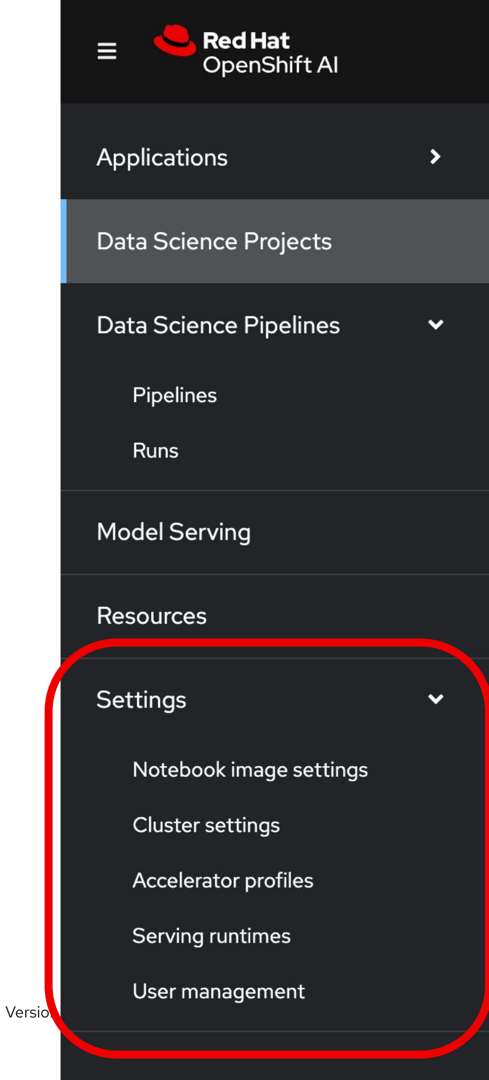


Groups

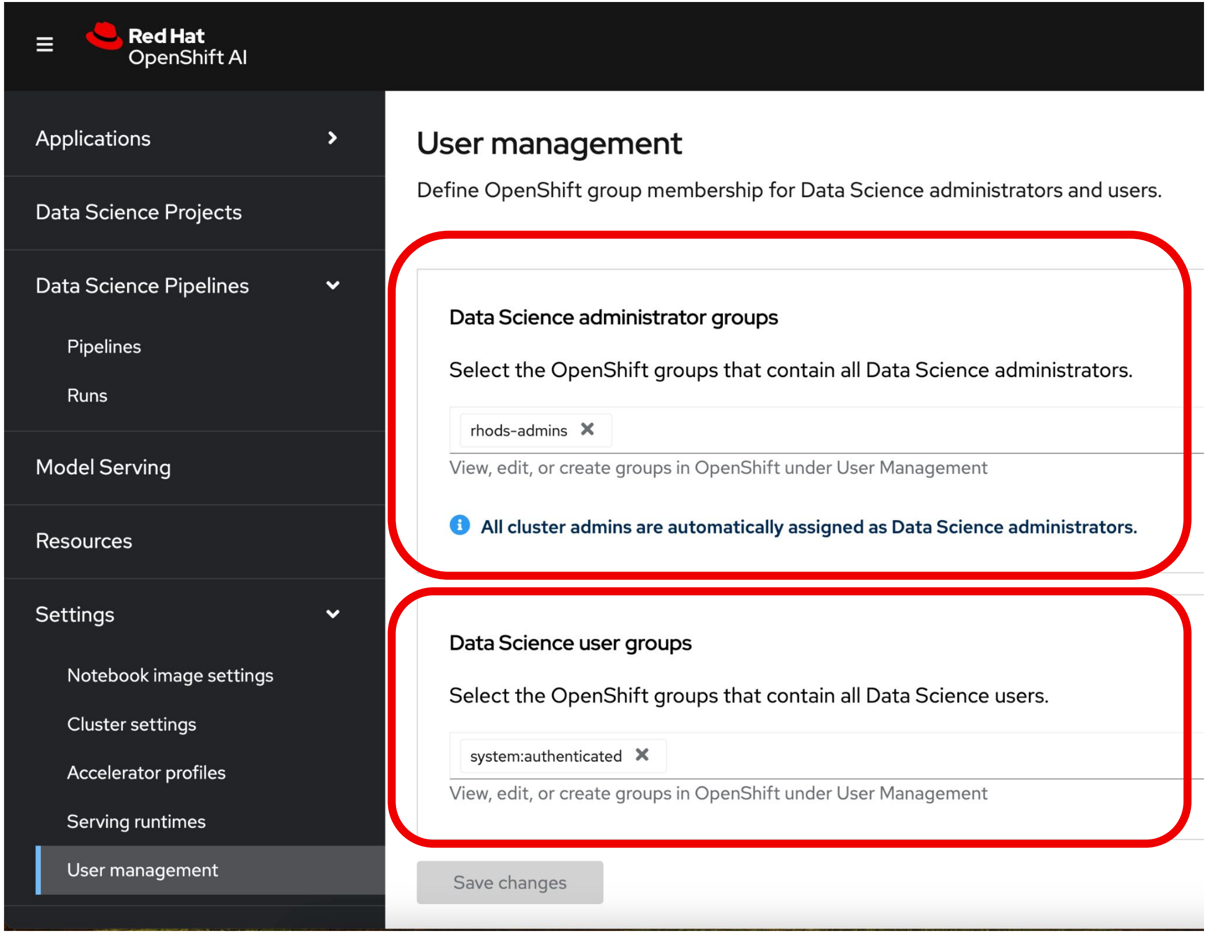
Name	Users	Created
rhods-admins	1	Mar 2, 2024, 12:17 PM
Team1	2	Mar 2, 2024, 5:20 PM
Team2	4	Mar 2, 2024, 5:21 PM

OpenShift AI - Users and Admins

- Only RHOAI admins can see the settings panel:



- User Management:
  - Choose the OpenShift Groups that should be RHOAI admins and RHOAI users
- "system:authenticated" means "any user declared in the identity provider". i.e. All users



User management

Define OpenShift group membership for Data Science administrators and users.

Data Science administrator groups

Select the OpenShift groups that contain all Data Science administrators.

rhods-admins ✕

View, edit, or create groups in OpenShift under User Management

All cluster admins are automatically assigned as Data Science administrators.

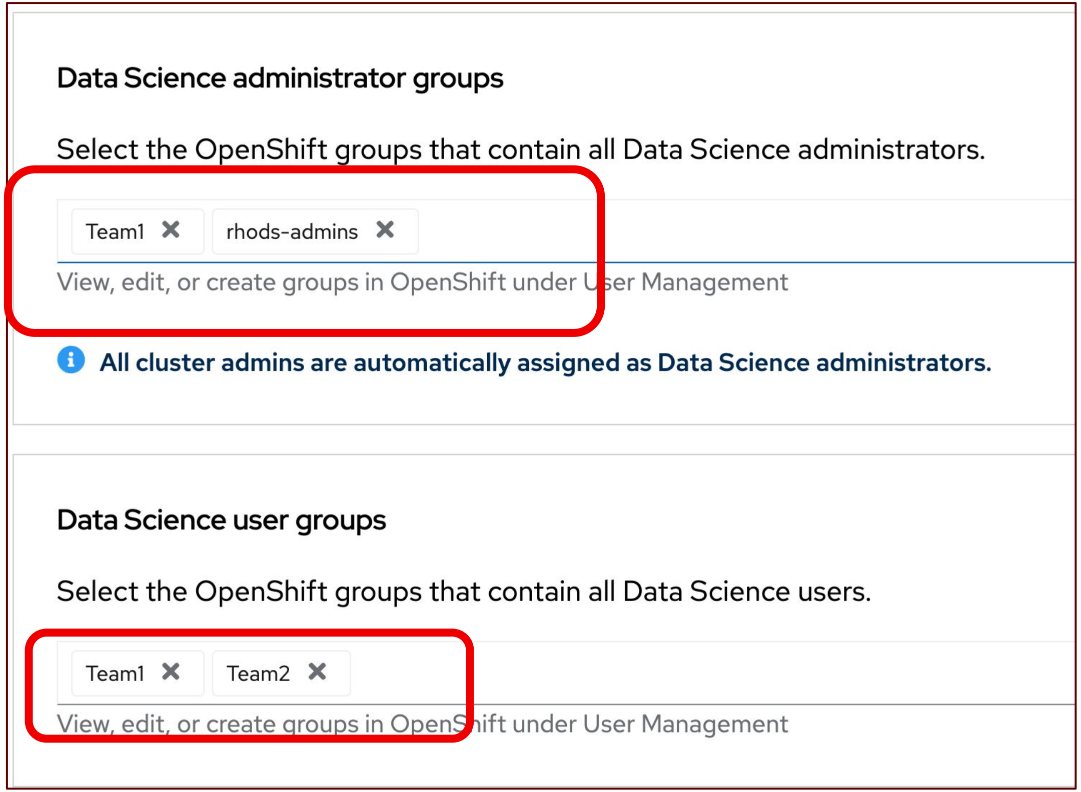
Data Science user groups

Select the OpenShift groups that contain all Data Science users.

system:authenticated ✕

View, edit, or create groups in OpenShift under User Management

- Example configuration



Data Science administrator groups

Select the OpenShift groups that contain all Data Science administrators.

Team1 ✕ rhods-admins ✕

View, edit, or create groups in OpenShift under User Management

All cluster admins are automatically assigned as Data Science administrators.

Data Science user groups

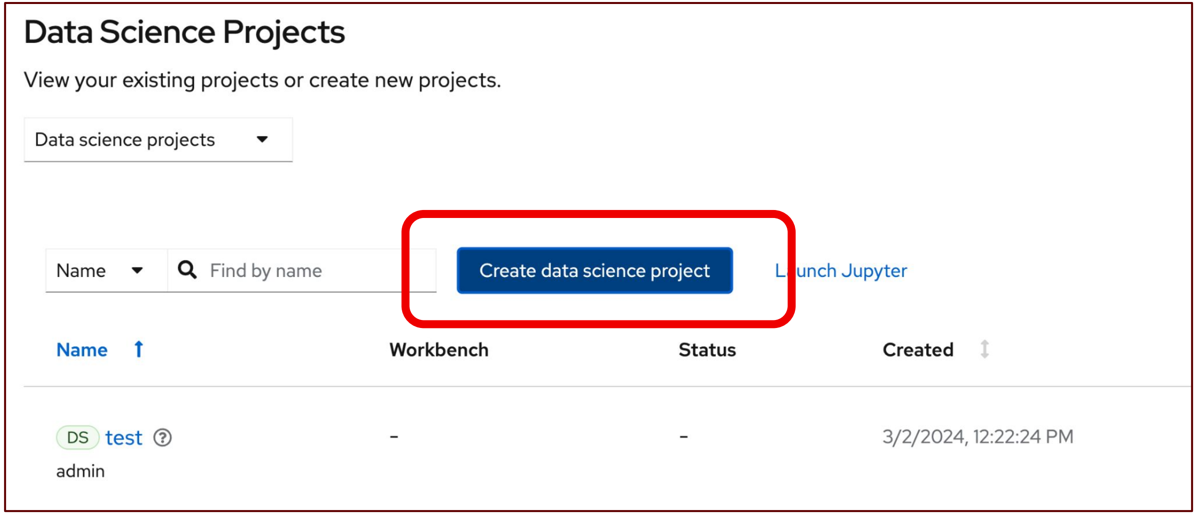
Select the OpenShift groups that contain all Data Science users.

Team1 ✕ Team2 ✕

View, edit, or create groups in OpenShift under User Management

Creating Projects

- OpenShift users can only create projects if they have the Project Self-Provisioner role
- The OpenShift admin will have configured which users (if any) have this role.
- To disable/enable, consult this documentation.



Data Science Projects

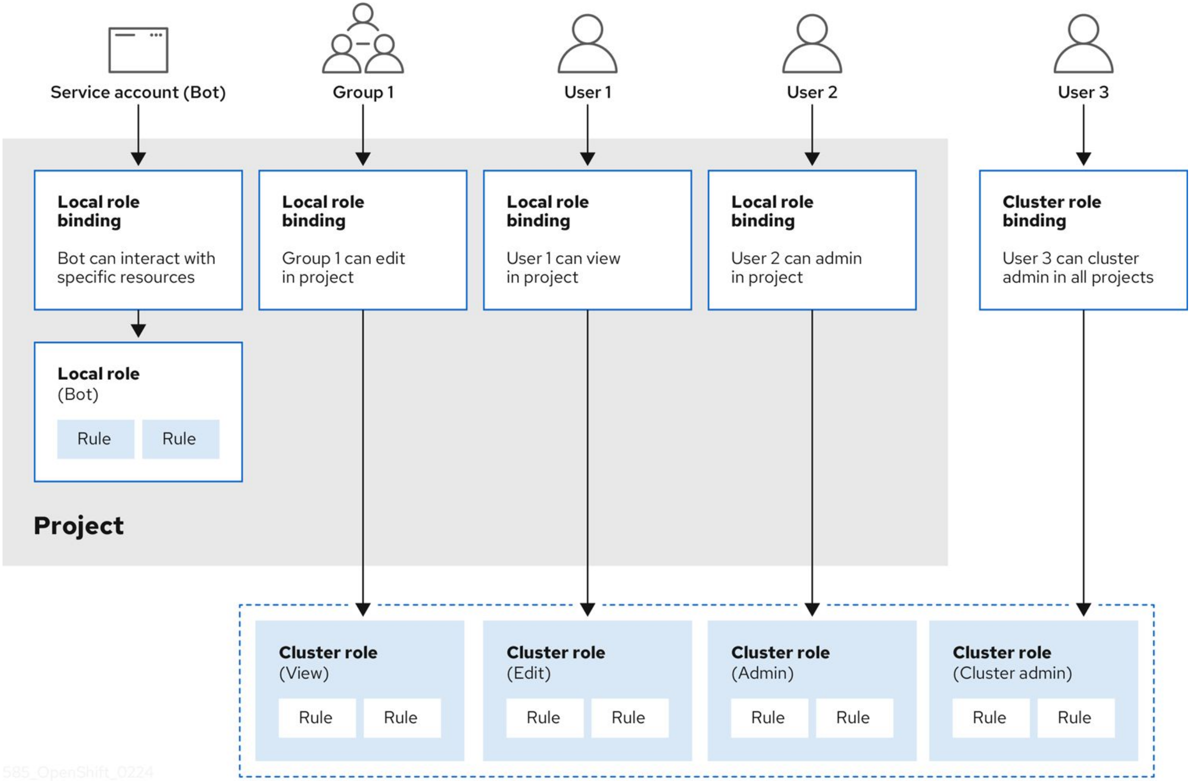
View your existing projects or create new projects.

Name	Workbench	Status	Created
DS test admin	-	-	3/2/2024, 12:22:24 PM

Overview of user types and permissions

- **Data scientists** - Can access and use individual components of RHOAI, such as Jupyter.
- **Administrators** - Have additional permissions to perform these actions:
  - Configure Red Hat OpenShift AI settings
  - Access and manage notebook servers

- Why customize your workbench?
- How can you customize a workbench?
  - Customizing the workbench
  - Creating custom notebook images (don’t dig too deep)
  - Using custom notebook images
- Best practices around custom notebook images



Service account (Bot)

Group 1

User 1

User 2

User 3

Local role binding

Bot can interact with specific resources

Local role binding

Group 1 can edit in project

Local role binding

User 1 can view in project

Local role binding

User 2 can admin in project

Cluster role binding

User 3 can cluster admin in all projects

Local role
(Bot)

Rule Rule

Project

Cluster role
(View)

Rule Rule

Cluster role
(Edit)

Rule Rule

Cluster role
(Admin)

Rule Rule

Cluster role
(Cluster admin)

Rule Rule

Custom notebook images

Custom Workbench Images

- Support
- Why
- How

Custom Workbench Images

- Support
  - Red Hat only supports the images that are provided by default with RHOAI
  - Red Hat also supports your ability to add a custom-built image
  - Red Hat does not provide support for your custom-built images
  - Any issue experienced with a Custom Image should be reproduced in one of the default images

Custom Workbench Images

- Why
  - Customization
    - create an image with the exact packages and versions you require
  - Stability
    - control over the versioning of your custom images
    - (updates, overlap, phase out)
  - Experimental
    - create bleeding edge versions of images
  - Flexibility
    - Jupyter vs VSCode vs R-Studio, etc….

Custom Workbench Images

- How
  - Look at examples
  - Create custom image (on laptop or on cluster) or use pre-built examples
  - Upload/Store image into Container Image repository
  - Add image reference into RHOAI admin interface
  - Test

Custom Workbench Images

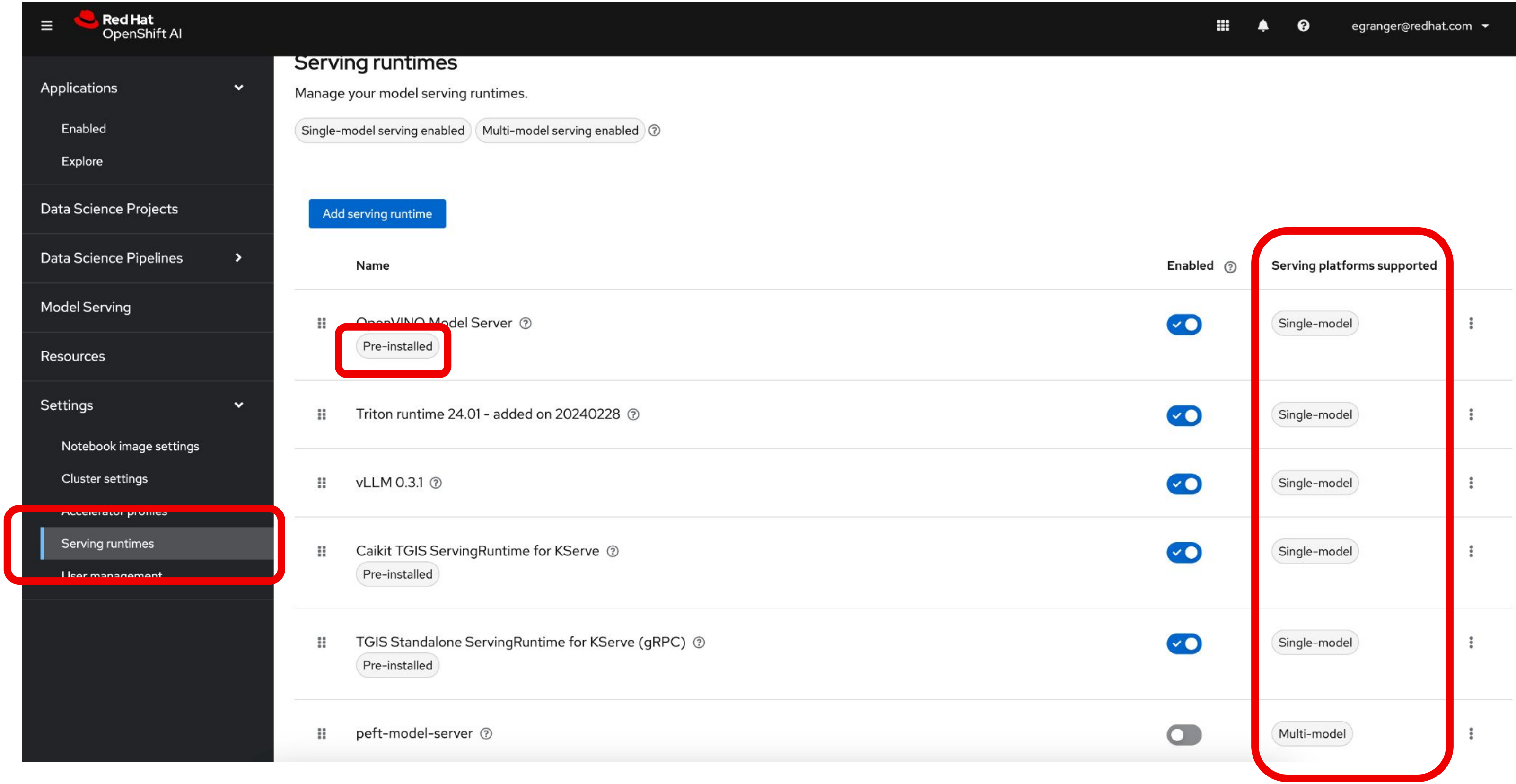
- How
  - There is an infinite ways of creating Workbench Images, therefore, the process cannot be fully documented.
  - Instead, we refer to examples in order to get started:
    - https://ai-on-openshift.io/odh-rhoai/custom-notebooks/
    - https://github.com/opendatahub-io-contrib/workbench-images
    - https://quay.io/repository/opendatahub-contrib/workbench-images?tab=tags

Serving runtimes
Default and Custom

Serving Runtimes

- A Serving Runtime can be either:
  - Default (Provided by Red Hat)
  - Custom (built and deployed by the customer
- For Single-Model serving or for Multi-Model serving
- Serving Runtimes can be managed only by the RHOAI Admins
- Serving Runtimes can be enabled/disabled
- Changes to the runtime definitions do not affect existing Model Servers

Serving Runtimes

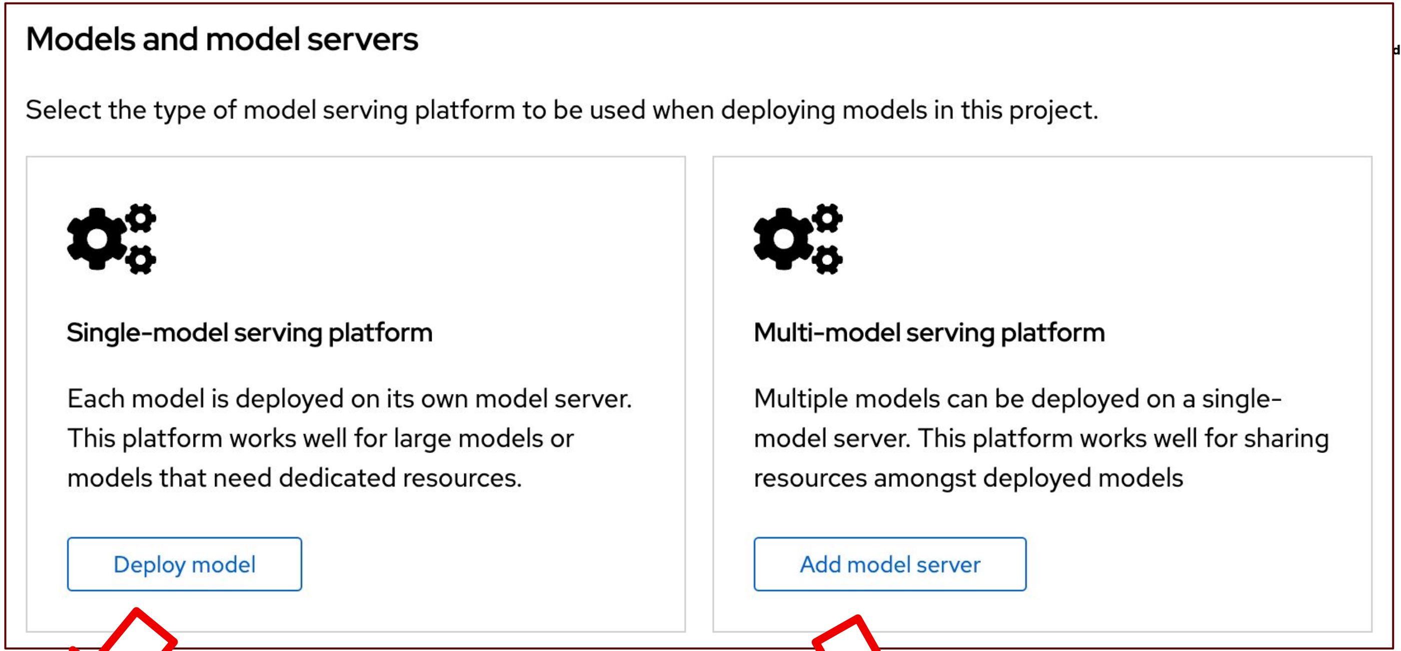


Serving runtimes

Manage your model serving runtimes.

Single-model serving enabled | Multi-model serving enabled

Name	Enabled	Serving platforms supported
OpenVINO Model Server Pre-installed	✓	Single-model
Triton runtime 24.01 - added on 20240228	✓	Single-model
vLLM 0.3.1	✓	Single-model
Caikit TGIS ServingRuntime for KServe Pre-installed	✓	Single-model
TGIS Standalone ServingRuntime for KServe (gRPC) Pre-installed	✓	Single-model
peft-model-server	○	Multi-model



Models and model servers

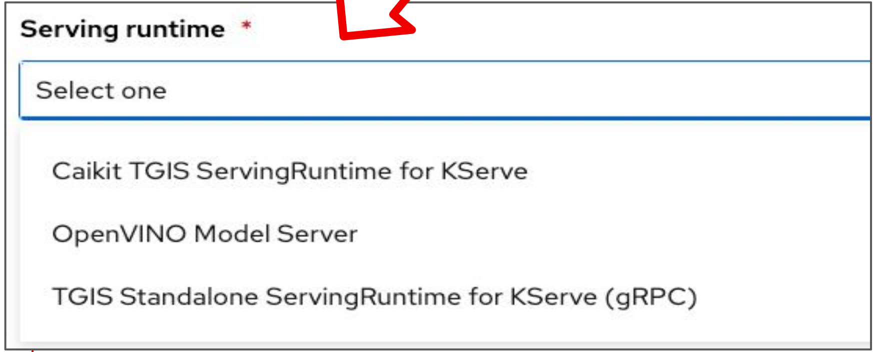
Select the type of model serving platform to be used when deploying models in this project.

Single-model serving platform

Each model is deployed on its own model server. This platform works well for large models or models that need dedicated resources.

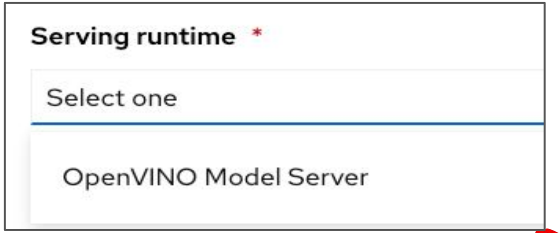
Multi-model serving platform

Multiple models can be deployed on a single-model server. This platform works well for sharing resources amongst deployed models



Serving runtime *

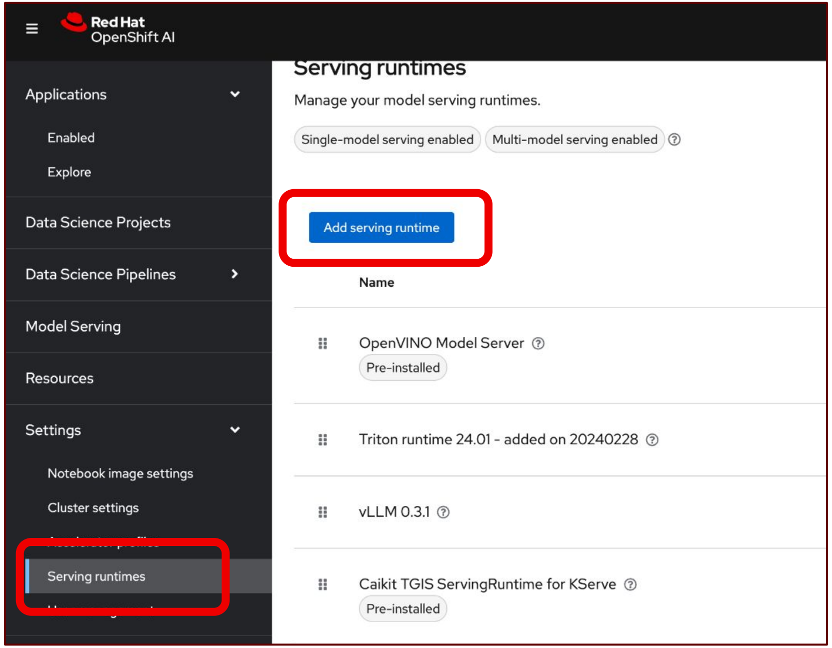
Select one
Caikit TGIS ServingRuntime for KServe
OpenVINO Model Server
TGIS Standalone ServingRuntime for KServe (gRPC)



Serving runtime *

Adding a Custom Serving Runtimes

- Locate an example
  - https://redhatquickcourses.github.io/rhods-deploy/rhods-deploy/1.33/chapter1/section3.html
  - or
  - [Custom Serving Runtime (Triton) – AI on OpenShift](#)
- Navigate to the right location:

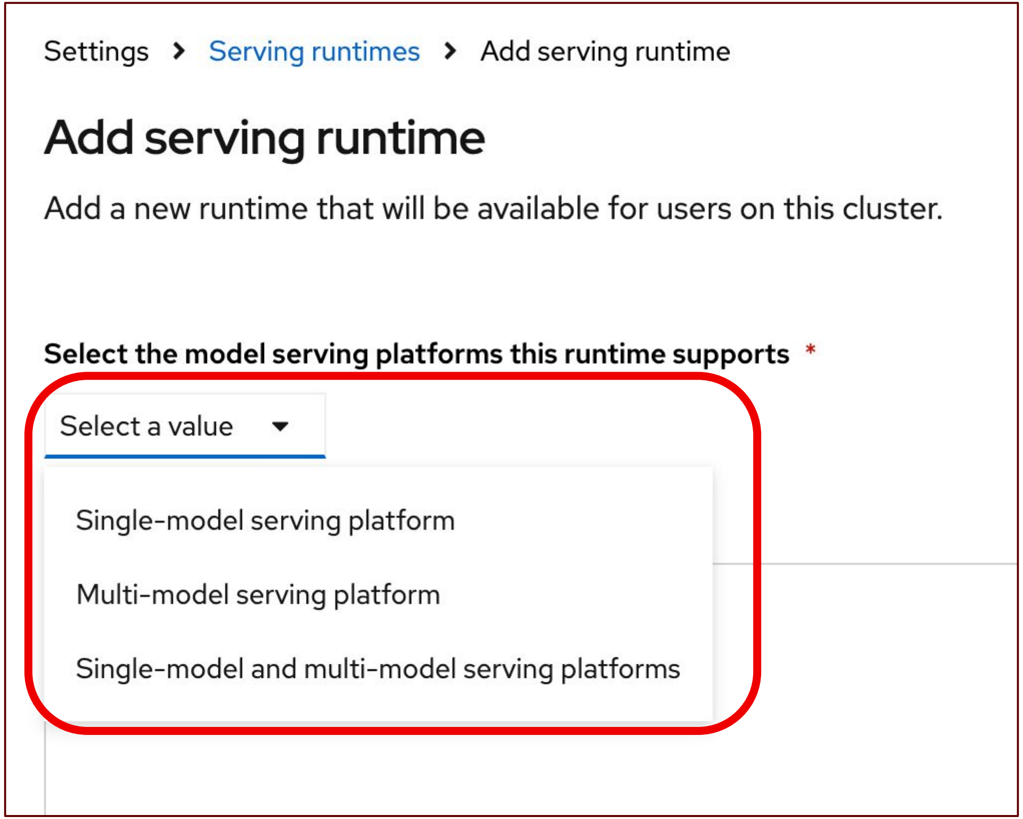


Serving runtimes

Manage your model serving runtimes.

Name
OpenVINO Model Server Pre-installed
Triton runtime 24.01 - added on 20240228
vLLM 0.3.1
Caikit TGIS ServingRuntime for KServe Pre-installed

Serving Runtimes



Settings › Serving runtimes › Add serving runtime

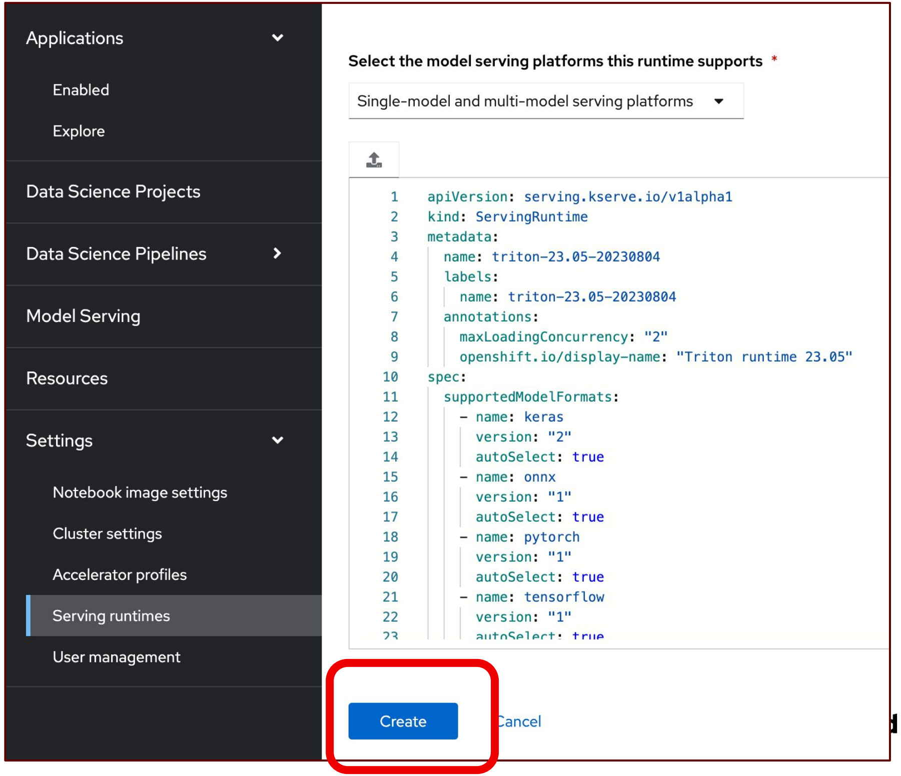
Add serving runtime

Add a new runtime that will be available for users on this cluster.

Select the model serving platforms this runtime supports *

Select a value
Single-model serving platform
Multi-model serving platform
Single-model and multi-model serving platforms

Serving Runtimes



Select the model serving platforms this runtime supports *

Single-model and multi-model serving platforms

```yaml
apiVersion: serving.kserve.io/v1alpha1
kind: ServingRuntime
metadata:
  name: triton-23.05-20230804
  labels:
    name: triton-23.05-20230804
  annotations:
    maxLoadingConcurrency: "2"
    openshift.io/display-name: "Triton runtime 23.05"
spec:
  supportedModelFormats:
    - name: keras
      version: "2"
      autoSelect: true
    - name: onnx
      version: "1"
      autoSelect: true
    - name: pytorch
      version: "1"
      autoSelect: true
    - name: tensorflow
      version: "1"
      autoSelect: true
```

Accelerators and GPUs

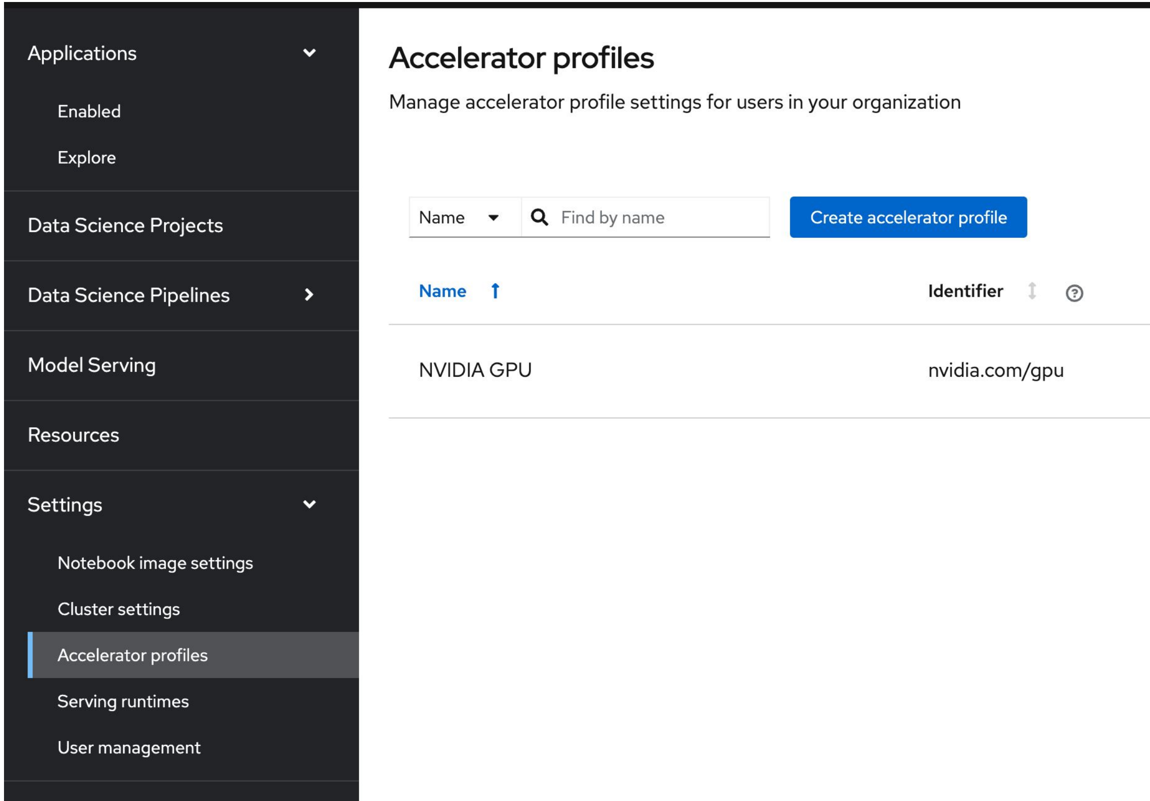
Red Hat

Accelerator profiles

- Why
  - There are various types of accelerators
  - e.g. Nvidia GPUs, Intel Gaudi, etc…
  - We need a way to surface these resources to the users
  - We need to differentiate, for example, different types of GPUs, and land the pods on the right machines
- How
  - Accelerator Profiles are managed by RHOAI admins.

Accelerators and GPUs

- Accelerator Profiles



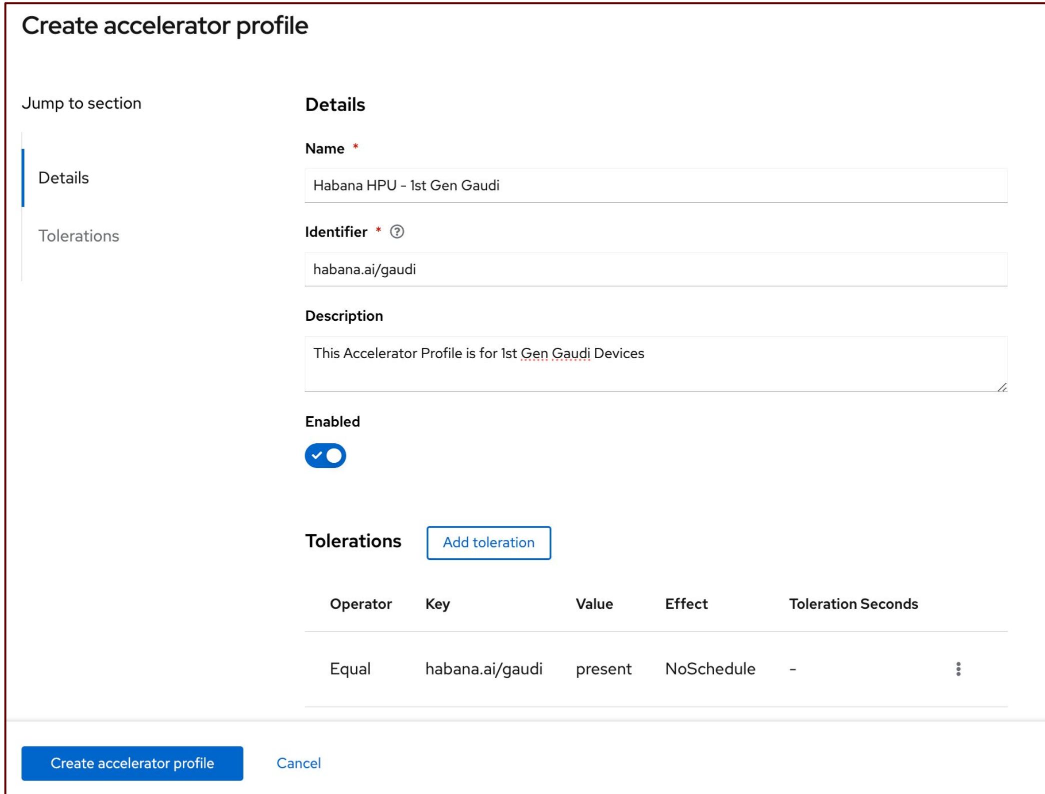
Accelerator profiles

Manage accelerator profile settings for users in your organization

Name	Identifier
NVIDIA GPU	nvidia.com/gpu

Accelerators and GPUs

- Creating a new profile



Create accelerator profile

Jump to section

Details

Name *

Habana HPU - 1st Gen Gaudi

Identifier *

habana.ai/gaudi

Description

This Accelerator Profile is for 1st Gen Gaudi Devices

Enabled

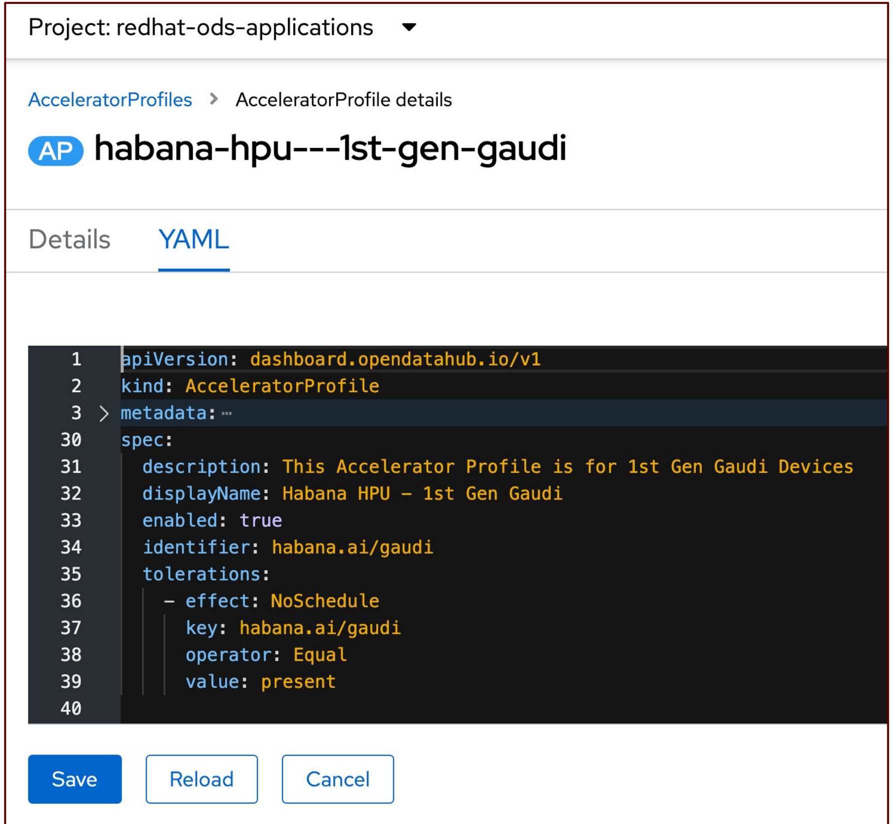
✓

Tolerations

Operator	Key	Value	Effect	Toleration Seconds
Equal	habana.ai/gaudi	present	NoSchedule	-

Accelerators and GPUs

- Equivalent YAML

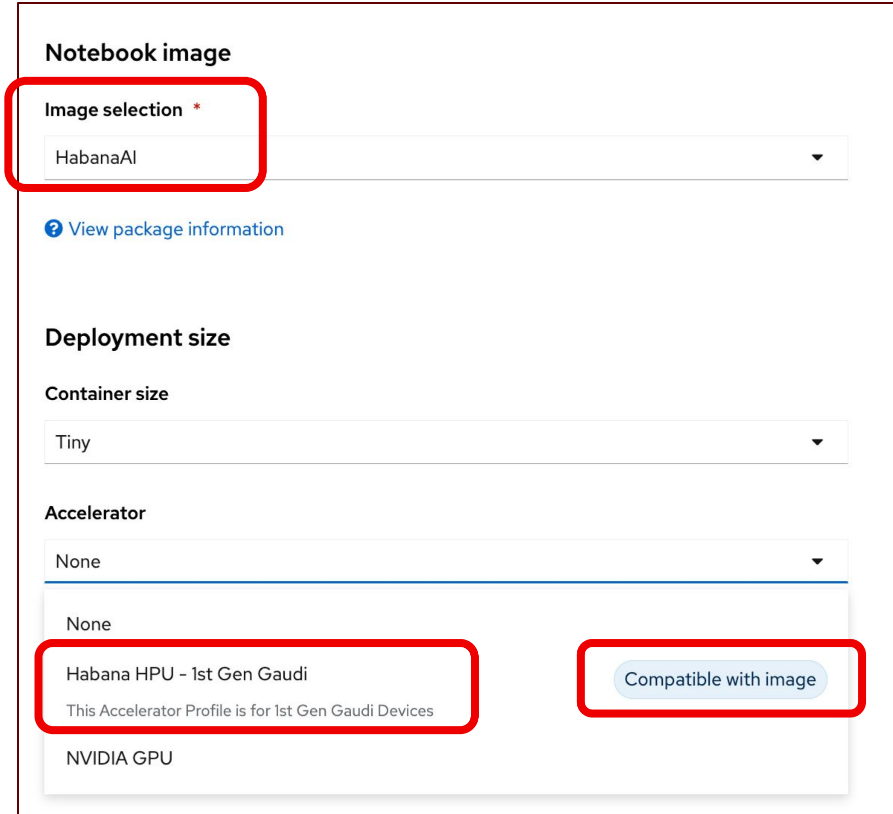


habana-hpu---1st-gen-gaudi

```yaml
apiVersion: dashboard.opendatahub.io/v1
kind: AcceleratorProfile
metadata:…
spec:
  description: This Accelerator Profile is for 1st Gen Gaudi Devices
  displayName: Habana HPU – 1st Gen Gaudi
  enabled: true
  identifier: habana.ai/gaudi
  tolerations:
    – effect: NoSchedule
      key: habana.ai/gaudi
      operator: Equal
      value: present

```

Accelerators and GPUs



Notebook image

Image selection *

HabanaAI

View package information

Deployment size

Container size

Tiny

Accelerator

None

None

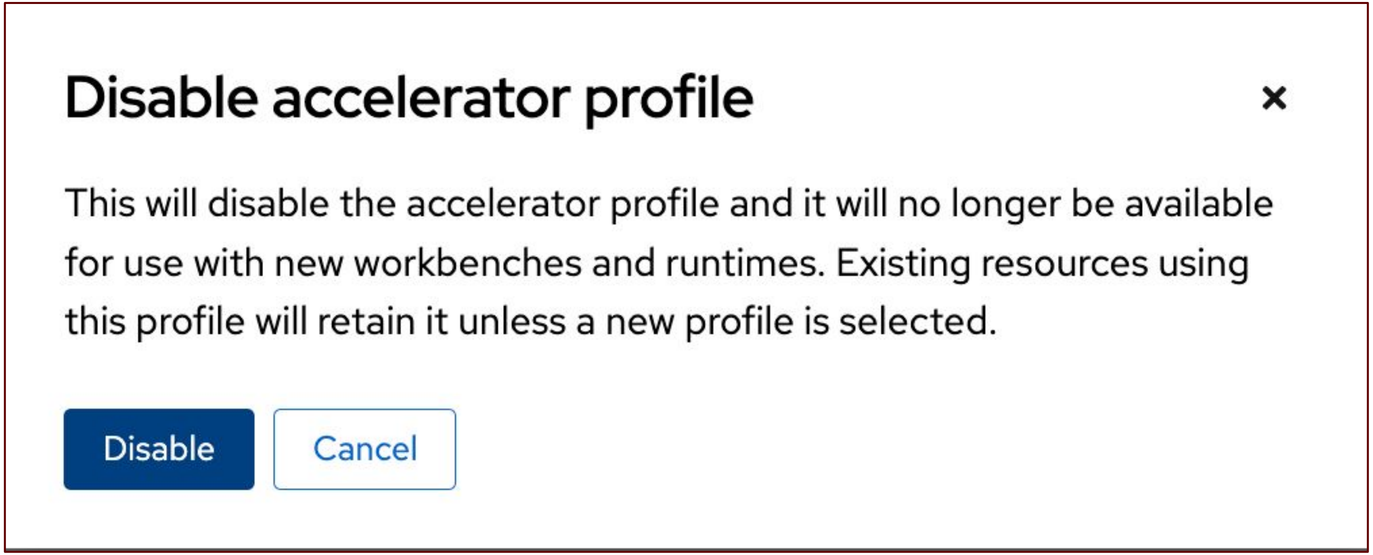
Habana HPU - 1st Gen Gaudi

This Accelerator Profile is for 1st Gen Gaudi Devices

Compatible with image

NVIDIA GPU

Accelerators and GPUs



Disable accelerator profile

This will disable the accelerator profile and it will no longer be available for use with new workbenches and runtimes. Existing resources using this profile will retain it unless a new profile is selected.

Managing RHOAI resources

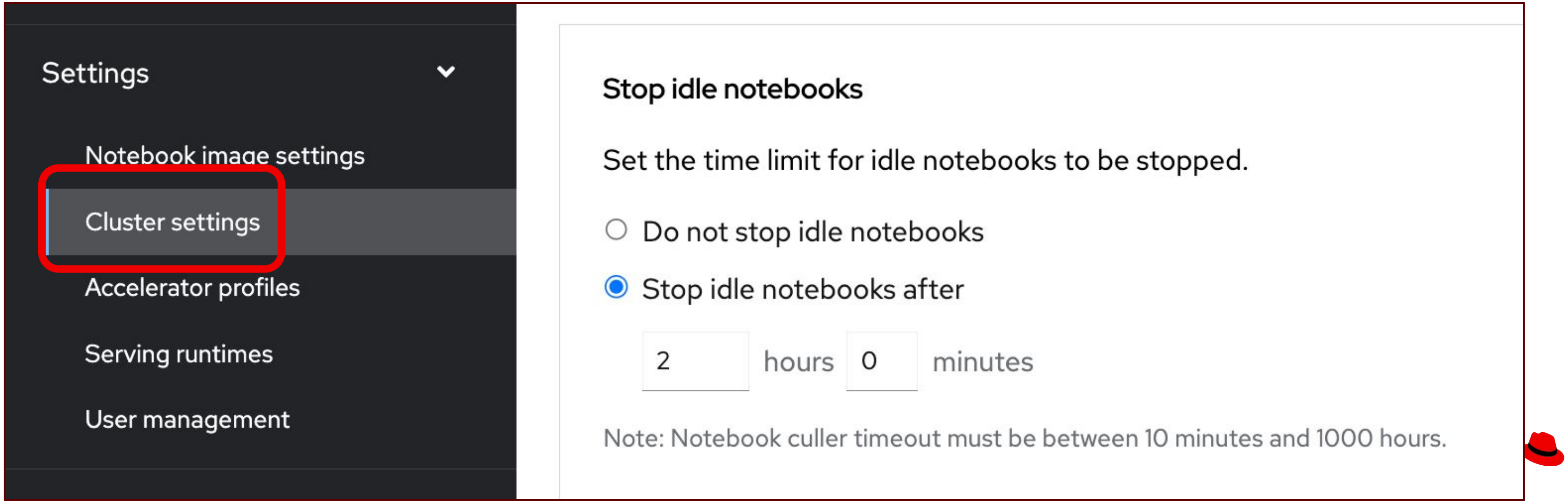
Idle Culler and other admin Settings

Idle Culler

- What
  - The idle culler is a mechanism that will stop Notebook servers left "idle"
  - "idle" in this context means that the notebook pod is running, but no browser is connected to it
- Why
  - Notebooks/Workbenches are supposed to be interactive environments
  - But if too many are left running all the time, they can consume all the cluster's resources
  - Idle Culler ensures regular cleanup of leftover workbenches

Idle Culler

- How
  - Default value is "do not stop..."



Stop idle notebooks

Set the time limit for idle notebooks to be stopped.

○ Do not stop idle notebooks

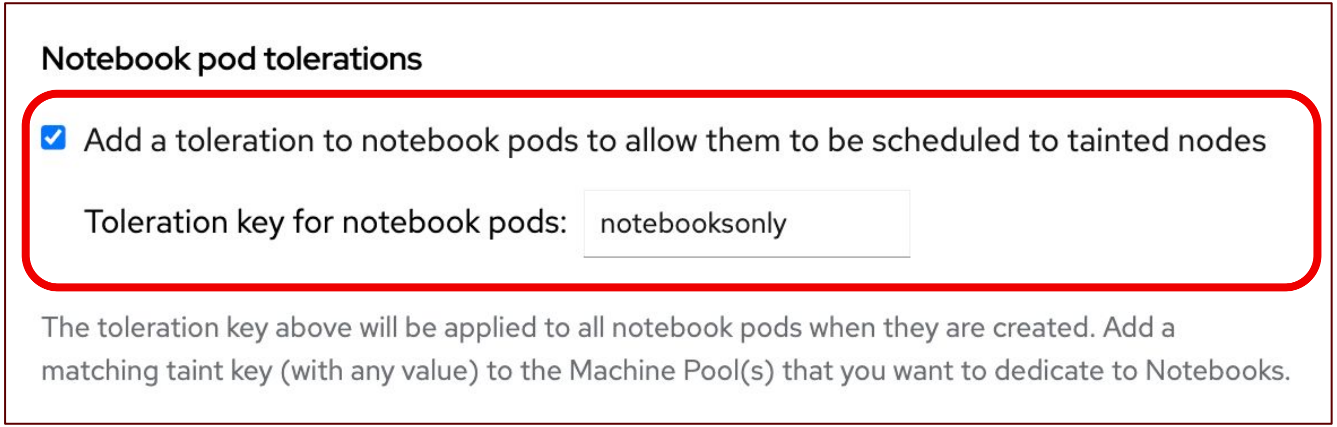
● Stop idle notebooks after

2 hours 0 minutes

Note: Notebook culler timeout must be between 10 minutes and 1000 hours.

Toleration for notebooks

- Why
  - If you want to dedicate some nodes to notebooks
  - Very useful for auto-scaling



Notebook pod tolerations

☑ Add a toleration to notebook pods to allow them to be scheduled to tainted nodes

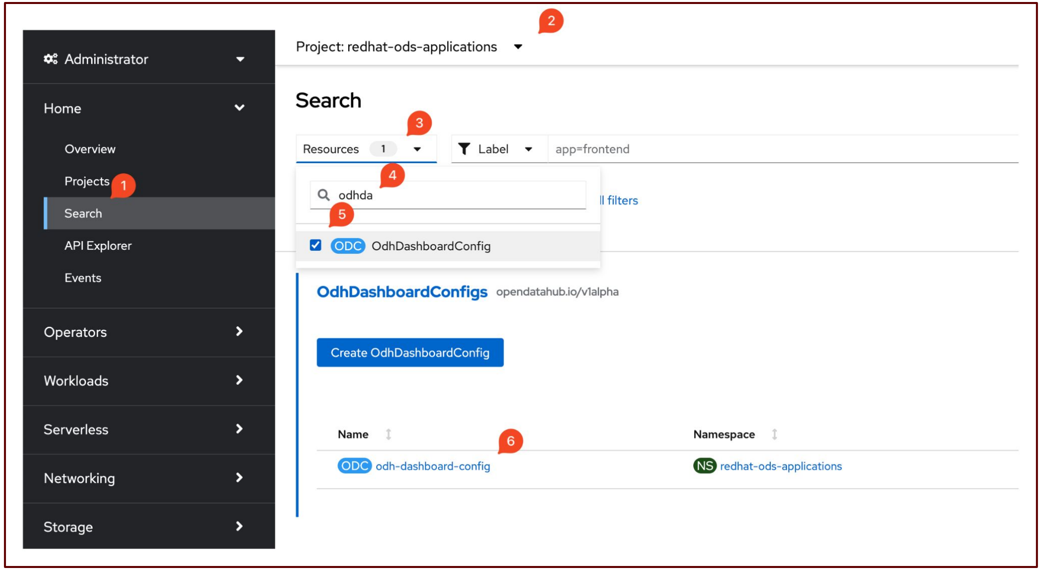
Toleration key for notebook pods: notebooksonly

The toleration key above will be applied to all notebook pods when they are created. Add a matching taint key (with any value) to the Machine Pool(s) that you want to dedicate to Notebooks.

Managing Container Sizes

Managing Container Sizes

- What
  - When launching Workbenches and Model Servers, users are prompted for a Container Size (S/M/L/XL)
- RHOAI admins can control the sizes available to the users



Project: redhat-ods-applications

app=frontend

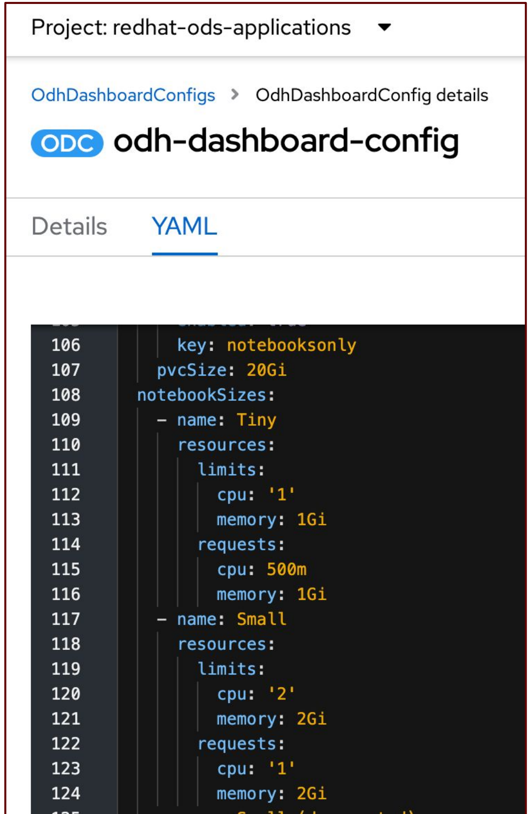
odhda

☑ ODC OdhDashboardConfig

OdhDashboardConfigs opendatahub.io/v1alpha

Name	Namespace
ODC odh-dashboard-config	NS redhat-ods-applications

Managing Container Sizes



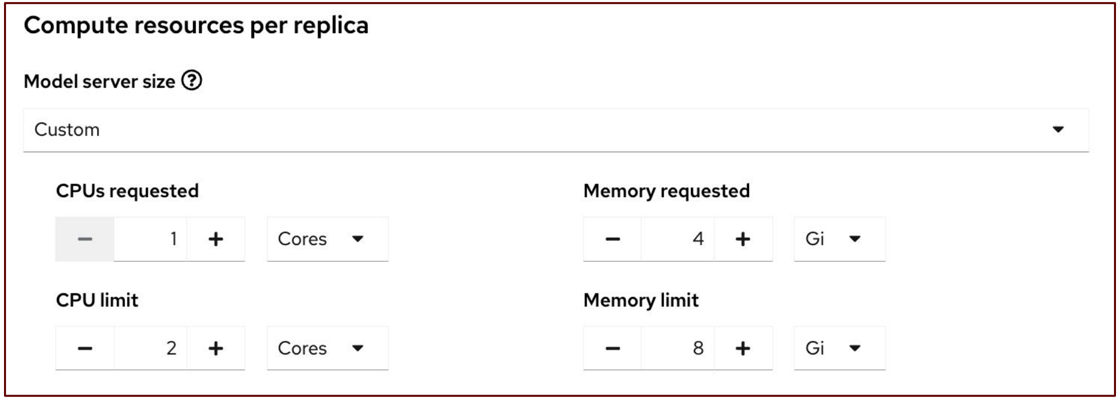
Project: redhat-ods-applications

OdhDashboardConfigs > OdhDashboardConfig details

ODC odh-dashboard-config

```
      key: notebooksonly
    pvcSize: 20Gi
  notebookSizes:
    - name: Tiny
      resources:
        limits:
          cpu: '1'
          memory: 1Gi
        requests:
          cpu: 500m
          memory: 1Gi
    - name: Small
      resources:
        limits:
          cpu: '2'
          memory: 2Gi
        requests:
          cpu: '1'
          memory: 2Gi
```

- Notebook sizes and Model Servers sizes are managed independently
- Model Servers Sizes can also be "custom":



Compute resources per replica

Model server size ?

Custom

CPUs requested: 1 Cores

Memory requested: 4 Gi

CPU limit: 2 Cores

Memory limit: 8 Gi

Managing Container Sizes

- Note:
  - When changing a size, existing Workbenches and Model Servers remain unchanged
  - To avoid "unknown" sizes, leave a (deprecated) placeholder in place for a little while

end of section

in linkedin.com/company/red-hat

▶ youtube.com/user/RedHatVideos

f facebook.com/redhatinc

𝕏 twitter.com/RedHat

Red Hat