Red Hat OpenShift AI Self-Managed 2.6

Managing resources

Learn to manage cluster resources, Jupyter notebooks, and data backup in Red Hat OpenShift AI

Last Updated: 2024-02-08

Red Hat OpenShift AI Self-Managed 2.6 Managing resources

Learn to manage cluster resources, Jupyter notebooks, and data backup in Red Hat OpenShift AI

Legal Notice

Abstract

Learn to manage cluster resources, Jupyter notebooks, and data backup in Red Hat OpenShift AI.

Table of Contents

PREFACE

As an OpenShift AI administrator, you can manage the following resources:

- Cluster resources to support compute-intensive data science work.
- Jupyter notebook servers.
- Data storage backup.

You can also specify whether to allow Red Hat to collect data about OpenShift AI usage in your cluster.

CHAPTER 1. MANAGING CLUSTER RESOURCES

1.1. CONFIGURING THE DEFAULT PVC SIZE FOR YOUR CLUSTER

To configure how resources are claimed within your OpenShift AI cluster, you can change the default size of the cluster’s persistent volume claim (PVC) ensuring that the storage requested matches your common storage workflow. PVCs are requests for resources in your cluster and also act as claim checks to the resource.

Prerequisites

- You have logged in to Red Hat OpenShift AI.

> **NOTE**
>
> Changing the PVC setting restarts the Jupyter pod and makes Jupyter unavailable for up to 30 seconds. As a workaround, it is recommended that you perform this action outside of your organization’s typical working day.

Procedure

1. From the OpenShift AI dashboard, click **Settings → Cluster settings**.
2. Under **PVC size**, enter a new size in gibibytes. The minimum size is 1 GiB, and the maximum size is 16384 GiB.
3. Click **Save changes**.

Verification

- New PVCs are created with the default storage size that you configured.

Additional resources

- Understanding persistent storage

1.2. RESTORING THE DEFAULT PVC SIZE FOR YOUR CLUSTER

To change the size of resources utilized within your OpenShift AI cluster, you can restore the default size of your cluster’s persistent volume claim (PVC).

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are part of the administrator group for OpenShift AI in OpenShift Container Platform.

Procedure

1. From the OpenShift AI dashboard, click **Settings → Cluster settings**.
2. Click **Restore Default** to restore the default PVC size of 20GiB.

3. Click **Save changes**.

Verification

- New PVCs are created with the default storage size of 20 GiB.

Additional resources

- Understanding persistent storage

1.3. ENABLING GPU SUPPORT IN OPENSHIFT AI

Optionally, to ensure that your data scientists can use compute-heavy workloads in their models, you can enable graphics processing units (GPUs) in OpenShift AI.

> **IMPORTANT**
>
> If you are using OpenShift AI in a disconnected self-managed environment, see Enabling GPU support in OpenShift AI instead.

Prerequisites

- You have logged in to your OpenShift Container Platform cluster.
- You have the **cluster-admin** role in your OpenShift Container Platform cluster.

Procedure

1. To enable GPU support on an OpenShift cluster, follow the instructions here: NVIDIA GPU Operator on Red Hat OpenShift Container Platform in the NVIDIA documentation.
2. Delete the **migration-gpu-status** ConfigMap.
    a. In the OpenShift Container Platform web console, switch to the **Administrator** perspective.
    b. Set the **Project** to **All Projects** or **redhat-ods-applications** to ensure you can see the appropriate ConfigMap.
    c. Search for the **migration-gpu-status** ConfigMap.
    d. Click the action menu (⋮) and select **Delete ConfigMap** from the list.
    The **Delete ConfigMap** dialog appears.
    e. Inspect the dialog and confirm that you are deleting the correct ConfigMap.
    f. Click **Delete**.
3. Restart the dashboard replicaset.
    a. In the OpenShift Container Platform web console, switch to the **Administrator** perspective.
    b. Click **Workloads → Deployments**.
    c. Set the **Project** to **All Projects** or **redhat-ods-applications** to ensure you can see the appropriate deployment.

d. Search for the **rhods-dashboard** deployment.

e. Click the action menu (⋮) and select **Restart Rollout** from the list.

f. Wait until the **Status** column indicates that all pods in the rollout have fully restarted.

Verification

- The NVIDIA GPU Operator appears on the **Operators** → **Installed Operators** page in the OpenShift Container Platform web console.

- The reset **migration-gpu-status** instance is present in the **Instances** tab on the **AcceleratorProfile** custom resource definition (CRD) details page.

After installing the NVIDIA GPU Operator, create an accelerator profile as described in Working with accelerator profiles.

1.4. ALLOCATING ADDITIONAL RESOURCES TO OPENSHIFT AI USERS

As a cluster administrator, you can allocate additional resources to a cluster to support compute-intensive data science work. This support includes increasing the number of nodes in the cluster and changing the cluster’s allocated machine pool.

For more information about allocating additional resources to an OpenShift Container Platform cluster, see Manually scaling a compute machine set .

CHAPTER 2. MANAGING JUPYTER NOTEBOOK SERVERS

2.1. ACCESSING THE JUPYTER ADMINISTRATION INTERFACE

You can use the Jupyter administration interface to control notebook servers in your Red Hat OpenShift AI environment.

Prerequisite

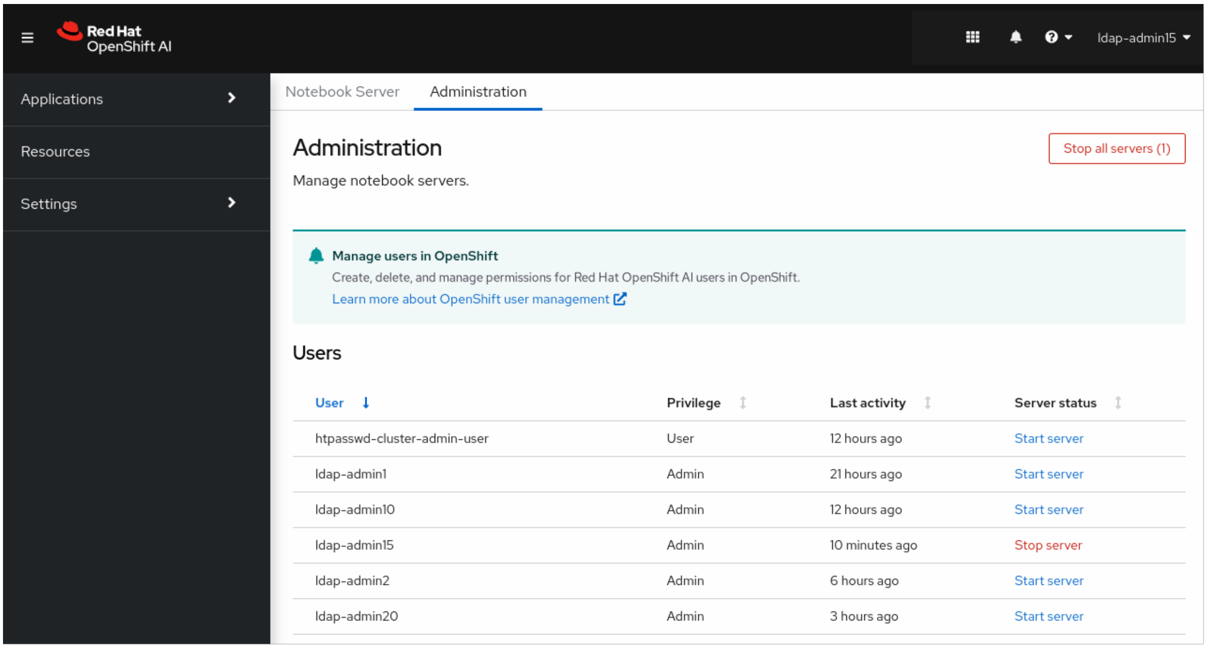
- You are part of the OpenShift Container Platform administrator group. For more information, see Adding administrative users for OpenShift Container Platform .

Procedure

- To access the Jupyter administration interface from OpenShift AI, perform the following actions:
  i. In OpenShift AI, in the **Applications** section of the left menu, click **Enabled**.
  ii. Locate the Jupyter tile and click **Launch application**.
  iii. On the page that opens when you launch Jupyter, click the **Administration** tab.
  The **Administration** page opens.
- To access the Jupyter administration interface from JupyterLab, perform the following actions:
  i. Click **File → Hub Control Panel**.
  ii. On the page that opens in OpenShift AI, click the **Administration** tab.
  The **Administration** page opens.

Verification

- You can see the Jupyter administration interface.



Administration

Manage notebook servers.

Manage users in OpenShift

Create, delete, and manage permissions for Red Hat OpenShift AI users in OpenShift.

Learn more about OpenShift user management

Users

User	Privilege	Last activity	Server status
htpasswd-cluster-admin-user	User	12 hours ago	Start server
ldap-admin1	Admin	21 hours ago	Start server
ldap-admin10	Admin	12 hours ago	Start server
ldap-admin15	Admin	10 minutes ago	Stop server
ldap-admin2	Admin	6 hours ago	Start server
ldap-admin20	Admin	3 hours ago	Start server

2.2. STARTING NOTEBOOK SERVERS OWNED BY OTHER USERS

Administrators can start a notebook server for another existing user from the Jupyter administration interface.

Prerequisites

- You are part of the OpenShift Container Platform administrator group. For more information, see Adding administrative users for OpenShift Container Platform .
- You have launched the Jupyter application, as described in Launching Jupyter and starting a notebook server.

Procedure

1. On the page that opens when you launch Jupyter, click the **Administration** tab.
2. On the **Administration** tab, perform the following actions:
   a. In the **Users** section, locate the user whose notebook server you want to start.
   b. Click **Start server** beside the relevant user.
   c. Complete the **Start a notebook server** page.
   d. Optional: Select the **Start server in current tab** checkbox if necessary.
   e. Click **Start server**.
      After the server starts, you see one of the following behaviors:
      - If you previously selected the **Start server in current tab** checkbox, the JupyterLab interface opens in the current tab of your web browser.
      - If you did not previously select the **Start server in current tab** checkbox, the **Starting server** dialog box prompts you to open the server in a new browser tab or in the current tab.
        The JupyterLab interface opens according to your selection.

Verification

- The JupyterLab interface opens.

Additional resources

- Options for notebook server environments

2.3. ACCESSING NOTEBOOK SERVERS OWNED BY OTHER USERS

Administrators can access notebook servers that are owned by other users to correct configuration errors or to help them troubleshoot problems with their environment.

Prerequisites

- You are part of the OpenShift Container Platform administrator group. For more information, see Adding administrative users for OpenShift Container Platform .

- You have launched the Jupyter application, as described in Launching Jupyter and starting a notebook server.

- The notebook server that you want to access is running.

Procedure

1. On the page that opens when you launch Jupyter, click the **Administration** tab.

2. On the **Administration** page, perform the following actions:

    a. In the **Users** section, locate the user that the notebook server belongs to.

    b. Click **View server** beside the relevant user.

    c. On the **Notebook server control panel** page, click **Access notebook server**.

Verification

- The user’s notebook server opens in JupyterLab.

2.4. STOPPING NOTEBOOK SERVERS OWNED BY OTHER USERS

Administrators can stop notebook servers that are owned by other users to reduce resource consumption on the cluster, or as part of removing a user and their resources from the cluster.

Prerequisites

- If you are using specialized OpenShift AI groups, you are part of the administrator group (for example, **rhoai-admins**). If you are not using specialized groups, you are part of the OpenShift Container Platform administrator group. For more information, see Adding administrative users for OpenShift Container Platform.

- You have launched the Jupyter application, as described in Launching Jupyter and starting a notebook server.

- The notebook server that you want to stop is running.

Procedure

1. On the page that opens when you launch Jupyter, click the **Administration** tab.

2. Stop one or more servers.

    - If you want to stop one or more specific servers, perform the following actions:

        i. In the **Users** section, locate the user that the notebook server belongs to.

        ii. To stop the notebook server, perform one of the following actions:

            - Click the action menu (⋮) beside the relevant user and select **Stop server**.

            - Click **View server** beside the relevant user and then click **Stop notebook server**. The **Stop server** dialog box appears.

        iii. Click **Stop server**.

- If you want to stop all servers, perform the following actions:
  i. Click the **Stop all servers** button.
  ii. Click **OK** to confirm stopping all servers.

Verification

- The **Stop server** link beside each server changes to a **Start server** link when the notebook server has stopped.

2.5. STOPPING IDLE NOTEBOOKS

You can reduce resource usage in your OpenShift AI deployment by stopping notebook servers that have been idle (without logged in users) for a period of time. This is useful when resource demand in the cluster is high. By default, idle notebooks are not stopped after a specific time limit.

> **NOTE**
>
> If you have configured your cluster settings to disconnect all users from a cluster after a specified time limit, then this setting takes precedence over the idle notebook time limit. Users are logged out of the cluster when their session duration reaches the cluster-wide time limit.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are part of the administrator group for OpenShift AI in OpenShift Container Platform.

Procedure

1. From the OpenShift AI dashboard, click **Settings → Cluster settings**.
2. Under **Stop idle notebooks**, select **Stop idle notebooks after**.
3. Enter a time limit, in **hours** and **minutes**, for when idle notebooks are stopped.
4. Click **Save changes**.

Verification

- The **notebook-controller-culler-config** ConfigMap, located in the **redhat-ods-applications** project on the **Workloads → ConfigMaps** page, contains the following culling configuration settings:
  - **ENABLE_CULLING**: Specifies if the culling feature is enabled or disabled (this is **false** by default).
  - **IDLENESS_CHECK_PERIOD**: The polling frequency to check for a notebook’s last known activity (in minutes).
  - **CULL_IDLE_TIME**: The maximum allotted time to scale an inactive notebook to zero (in minutes).

- Idle notebooks stop at the time limit that you set.

2.6. CONFIGURING A CUSTOM NOTEBOOK IMAGE

In addition to notebook images provided and supported by Red Hat and independent software vendors (ISVs), you can configure custom notebook images that cater to your project's specific requirements.

Red Hat supports you in adding custom notebook images to your deployment of OpenShift AI and ensuring that they are available for selection when creating a notebook server. However, Red Hat does not support the contents of your custom notebook image. That is, if your custom notebook image is available for selection during notebook server creation, but does not create a usable notebook server, Red Hat does not provide support to fix your custom notebook image.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are assigned the **cluster-admin** role in OpenShift Container Platform.
- Your custom notebook image exists in an image registry and is accessible.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Notebook image settings**.
   The **Notebook image settings** page appears. Previously imported notebook images are displayed. To enable or disable a previously imported notebook image, on the row containing the relevant notebook image, click the toggle in the **Enable** column.

   > NOTE
   >
   > If you have already configured an accelerator identifier for a notebook image, you can specify a recommended accelerator for the notebook image by creating an associated accelerator profile. To do this, click **Create profile** on the row containing the notebook image and complete the relevant fields. If the notebook image does not contain an accelerator identifier, you must manually configure one before creating an associated accelerator profile.

2. Click **Import new image**. Alternatively, if no previously imported images were found, click **Import image**.
   The **Import Notebook images** dialog appears.
3. In the **Image location** field, enter the URL of the repository containing the notebook image. For example: **quay.io/my-repo/my-image:tag**, **quay.io/my-repo/my-image@sha256:xxxxxxxxxxxxx**, or **docker.io/my-repo/my-image:tag**.
4. In the **Name** field, enter an appropriate name for the notebook image.
5. Optional: In the **Description** field, enter a description for the notebook image.
6. Optional: From the **Accelerator identifier** list, select an identifier to set its accelerator as recommended with the notebook image. If the notebook image contains only one accelerator identifier, the identifier name displays by default.
7. Optional: Add software to the notebook image. After the import has completed, the software is added to the notebook image's meta-data and displayed on the Jupyter server creation page.

a. Click the **Software** tab.

b. Click the **Add software** button.

c. Click **Edit** (✎).

d. Enter the **Software** name.

e. Enter the software **Version**.

f. Click **Confirm** (✓) to confirm your entry.

g. To add additional software, click **Add software**, complete the relevant fields, and confirm your entry.

8. Optional: Add packages to the notebook images. After the import has completed, the packages are added to the notebook image's meta-data and displayed on the Jupyter server creation page.

   a. Click the **Packages** tab.

   b. Click the **Add package** button.

   c. Click **Edit** (✎).

   d. Enter the **Package** name.

   e. Enter the package **Version**.

   f. Click **Confirm** (✓) to confirm your entry.

   g. To add an additional package, click **Add package**, complete the relevant fields, and confirm your entry.

9. Click **Import**.

Verification

- The notebook image that you imported is displayed in the table on the **Notebook image settings** page.

- Your custom notebook image is available for selection on the **Start a notebook server**page in Jupyter.

Additional resources

- Managing image streams

- Understanding build configurations

CHAPTER 3. BACKING UP DATA

3.1. BACKING UP STORAGE DATA

It is a best practice to back up the data on your persistent volume claims (PVCs) regularly.

Backing up your data is particularly important before you delete a user and before you uninstall OpenShift AI, as all PVCs are deleted when OpenShift AI is uninstalled.

See the documentation for your cluster platform for more information about backing up your PVCs.

Additional resources

- Understanding persistent storage

CHAPTER 4. USAGE DATA COLLECTION

Red Hat OpenShift AI administrators can choose whether to allow Red Hat to collect data about OpenShift AI usage in their cluster. Collecting this data allows Red Hat to monitor and improve our software and support. For further details about the data Red Hat collects, see Usage data collection notice for OpenShift AI.

Usage data collection is enabled by default when you install OpenShift AI on your OpenShift Container Platform cluster except when clusters are installed in a disconnected environment.

See Disabling usage data collection for instructions on disabling the collection of this data in your cluster. If you have disabled data collection on your cluster, and you want to enable it again, see Enabling usage data collection for more information.

4.1. USAGE DATA COLLECTION NOTICE FOR OPENSHIFT AI

In connection with your use of this Red Hat offering, Red Hat may collect usage data about your use of the software. This data allows Red Hat to monitor the software and to improve Red Hat offerings and support, including identifying, troubleshooting, and responding to issues that impact users.

What information does Red Hat collect?

Tools within the software monitor various metrics and this information is transmitted to Red Hat. Metrics include information such as:

- Information about applications enabled in the product dashboard.
- The deployment sizes used (that is, the CPU and memory resources allocated).
- Information about documentation resources accessed from the product dashboard.
- The name of the notebook images used (that is, Minimal Python, Standard Data Science, and other images.).
- A unique random identifier that generates during the initial user login to associate data to a particular username.
- Usage information about components, features, and extensions.

Third Party Service Providers

Red Hat uses certain third party service providers to collect the telemetry data.

Security

Red Hat employs technical and organizational measures designed to protect the usage data.

Personal Data

Red Hat does not intend to collect personal information. If Red Hat discovers that personal information has been inadvertently received, Red Hat will delete such personal information and treat such personal information in accordance with Red Hat’s Privacy Statement. For more information about Red Hat’s privacy practices, see Red Hat’s Privacy Statement.

Enabling and Disabling Usage Data

You can disable or enable usage data by following the instructions in Disabling usage data collection or Enabling usage data collection.

4.2. ENABLING USAGE DATA COLLECTION

Red Hat OpenShift AI administrators can choose whether to allow Red Hat to collect data about OpenShift AI usage in their cluster. Usage data collection is enabled by default when you install OpenShift AI on your OpenShift Container Platform cluster except when clusters are installed in a disconnected environment. If you have disabled data collection previously, you can re-enable it by following these steps.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are part of the administrator group for OpenShift AI in OpenShift Container Platform except when clusters are installed in a disconnected environment.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Cluster settings**.
2. Locate the **Usage data collection**section.
3. Select the **Allow collection of usage data**checkbox.
4. Click **Save changes**.

Verification

- A notification is shown when settings are updated: **Settings changes saved.**

Additional resources

- Usage data collection notice for OpenShift AI

4.3. DISABLING USAGE DATA COLLECTION

Red Hat OpenShift AI administrators can choose whether to allow Red Hat to collect data about OpenShift AI usage in their cluster. Usage data collection is enabled by default when you install OpenShift AI on your OpenShift Container Platform cluster except when clusters are installed in a disconnected environment.

You can disable data collection by following these steps.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are part of the administrator group for OpenShift AI in OpenShift Container Platform except when clusters are installed in a disconnected environment.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Cluster settings**.
2. Locate the **Usage data collection**section.
3. Deselect the **Allow collection of usage data**checkbox.

4. Click **Save changes**.

Verification

- A notification is shown when settings are updated: **Settings changes saved.**

Additional resources

- Usage data collection notice for OpenShift AI