



# Red Hat OpenShift AI Cloud Service 1

## Installing the OpenShift AI Cloud Service

Install and uninstall Red Hat OpenShift AI as an add-on to an OpenShift cluster



# Red Hat OpenShift AI Cloud Service 1 Installing the OpenShift AI Cloud Service

---

Install and uninstall Red Hat OpenShift AI as an add-on to an OpenShift cluster

## Legal Notice

Copyright © 2024 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux<sup>®</sup> is the registered trademark of Linus Torvalds in the United States and other countries.

Java<sup>®</sup> is a registered trademark of Oracle and/or its affiliates.

XFS<sup>®</sup> is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL<sup>®</sup> is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js<sup>®</sup> is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack<sup>®</sup> Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

## Abstract

Install and uninstall Red Hat OpenShift AI as an add-on to a Red Hat managed environment, such as Red Hat OpenShift Dedicated and Red Hat OpenShift Service on Amazon Web Services (ROSA).

---

## Table of Contents

<b>CHAPTER 1. ARCHITECTURE OF OPENSIFT AI</b>	<b>3</b>
<b>CHAPTER 2. OVERVIEW OF INSTALLING AND DEPLOYING OPENSIFT AI</b>	<b>5</b>
<b>CHAPTER 3. REQUIREMENTS FOR OPENSIFT AI</b>	<b>6</b>
<b>CHAPTER 4. CONFIGURING AN IDENTITY PROVIDER FOR YOUR OPENSIFT CLUSTER</b>	<b>8</b>
<b>CHAPTER 5. ADDING ADMINISTRATIVE USERS</b>	<b>10</b>
<b>CHAPTER 6. SUBSCRIBING TO THE RED HAT OPENSIFT AI CLOUD SERVICE</b>	<b>11</b>
6.1. SUBSCRIBING TO THE OPENSIFT AI MANAGED CLOUD SERVICE ON AWS OR GCP	11
6.2. SUBSCRIBING TO THE OPENSIFT AI MANAGED CLOUD SERVICE ON RED HAT OPENSIFT SERVICE ON AWS (ROSA)	11
<b>CHAPTER 7. INSTALLING OPENSIFT AI ON YOUR OPENSIFT CLUSTER</b>	<b>13</b>
<b>CHAPTER 8. INSTALLING RED HAT OPENSIFT AI COMPONENTS BY USING THE WEB CONSOLE</b>	<b>15</b>
<b>CHAPTER 9. BACKING UP STORAGE DATA FROM AMAZON EBS</b>	<b>17</b>
<b>CHAPTER 10. BACKING UP STORAGE DATA FROM GOOGLE PERSISTENT DISK</b>	<b>19</b>
<b>CHAPTER 11. UNINSTALLING OPENSIFT AI</b>	<b>21</b>
<b>CHAPTER 12. ACCESSING THE DASHBOARD</b>	<b>23</b>
<b>CHAPTER 13. ENABLING GPU SUPPORT IN OPENSIFT AI</b>	<b>24</b>
<b>CHAPTER 14. TROUBLESHOOTING COMMON INSTALLATION PROBLEMS</b>	<b>26</b>
14.1. THE RED HAT OPENSIFT AI OPERATOR CANNOT BE RETRIEVED FROM THE IMAGE REGISTRY	26
14.2. OPENSIFT AI CANNOT BE INSTALLED DUE TO INSUFFICIENT CLUSTER RESOURCES	26
14.3. THE DEDICATED-ADMINS ROLE-BASED ACCESS CONTROL (RBAC) POLICY CANNOT BE CREATED	27
14.4. OPENSIFT AI DOES NOT INSTALL ON UNSUPPORTED INFRASTRUCTURE	27
14.5. THE CREATION OF THE OPENSIFT AI CUSTOM RESOURCE (CR) FAILS	28
14.6. THE CREATION OF THE OPENSIFT AI NOTEBOOKS CUSTOM RESOURCE (CR) FAILS	28
14.7. THE DEAD MAN'S SNITCH OPERATOR'S SECRET DOES NOT GET CREATED	29
14.8. THE PAGERDUTY SECRET DOES NOT GET CREATED	29
14.9. THE SMTP SECRET DOES NOT EXIST	30
14.10. THE ODH PARAMETER SECRET DOES NOT GET CREATED	30



# CHAPTER 1. ARCHITECTURE OF OPENSIFT AI

Red Hat OpenShift AI is a fully Red Hat managed cloud service that is available as an Add-on to Red Hat OpenShift Dedicated and to Red Hat OpenShift Service on Amazon Web Services (ROSA).

OpenShift AI integrates the following components and services:

- At the service layer:

## OpenShift AI dashboard

A customer-facing dashboard that shows available and installed applications for the OpenShift AI environment as well as learning resources such as tutorials, quick start examples, and documentation. You can also access administrative functionality from the dashboard, such as user management, cluster settings, accelerator profiles, and notebook image settings. In addition, data scientists can create their own projects from the dashboard. This enables them to organize their data science work into a single project.

## Model serving

Data scientists can deploy trained machine-learning models to serve intelligent applications in production. After deployment, applications can send requests to the model using its deployed API endpoint.

## Data science pipelines

Data scientists can build portable machine learning (ML) workflows with data science pipelines, using Docker containers. This enables your data scientists to automate workflows as they develop their data science models.

## Jupyter (Red Hat managed)

A Red Hat managed application that allows data scientists to configure their own notebook server environment and develop machine learning models in JupyterLab.

## TrustyAI

Data scientists can review local, global, and time-series explainers for predictive models in both enterprise and data science applications. They can also use tools to detect bias and drift. These capabilities help organizations to deliver higher quality, unbiased AI-based applications.

## Distributed workloads

Data scientists can use multiple nodes in parallel to train machine-learning models or process data more quickly. This approach significantly reduces the task completion time, and enables the use of larger datasets and more complex models.



## IMPORTANT

The distributed workloads feature is currently available in Red Hat OpenShift AI as a Technology Preview feature only. Technology Preview features are not supported with Red Hat production service level agreements (SLAs) and might not be functionally complete. Red Hat does not recommend using them in production. These features provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process.

For more information about the support scope of Red Hat Technology Preview features, see [Technology Preview Features Support Scope](#).

- At the management layer:

## The Red Hat OpenShift AI Operator

A meta-operator that deploys and maintains all components and sub-operators that are part of OpenShift AI.

### Monitoring services

Alertmanager, OpenShift Telemetry, and Prometheus work together to gather metrics from OpenShift AI and organize and display those metrics in useful ways for monitoring and billing purposes. Alerts from Alertmanager are sent to PagerDuty, responsible for notifying Red Hat of any issues with your managed cloud service.

When you install the OpenShift Data Science Add-on in the Cluster Manager, the following new projects are created:

- The **redhat-ods-operator** project contains the Red Hat OpenShift AI Operator.
- The **redhat-ods-applications** project installs the dashboard and other required components of OpenShift AI.
- The **redhat-ods-monitoring** project contains services for monitoring and billing.
- The **rhods-notebooks** project is where notebook environments are deployed by default.

You or your data scientists must create additional projects for the applications that will use your machine learning models.

Do not install independent software vendor (ISV) applications in namespaces associated with OpenShift AI add-ons unless you are specifically directed to do so on the application tile on the dashboard.



## CHAPTER 2. OVERVIEW OF INSTALLING AND DEPLOYING OPENSIFT AI

Red Hat OpenShift AI is a platform for data scientists and developers of artificial intelligence (AI) applications. It provides a fully supported environment that lets you rapidly develop, train, test, and deploy machine learning models on-premises and/or in the public cloud.

OpenShift AI is provided as a managed cloud service add-on for Red Hat OpenShift or as self-managed software that you can install on-premise or in the public cloud on OpenShift.

For information about installing OpenShift AI as self-managed software on your OpenShift cluster in a connected or a disconnected environment, see [Product Documentation for Red Hat OpenShift AI Self-Managed](#).

There are two deployment options for Red Hat OpenShift AI as a managed cloud service add-on:

- **OpenShift Dedicated with a Customer Cloud Subscription on Amazon Web Services or Google Cloud Platform**

OpenShift Dedicated is a complete OpenShift Container Platform cluster provided as a cloud service, configured for high availability, and dedicated to a single customer. OpenShift Dedicated is professionally managed by Red Hat and hosted on Amazon Web Services (AWS) or Google Cloud Platform (GCP). The Customer Cloud Subscription (CCS) model allows Red Hat to deploy and manage clusters into a customer's AWS or GCP account. Contact your Red Hat account manager to get OpenShift Dedicated through a CCS.

- **Red Hat OpenShift Service on AWS (ROSA)**

ROSA is a fully-managed, turnkey application platform that allows you to focus on delivering value to your customers by building and deploying applications. You subscribe to the service directly from your AWS account.

Installing OpenShift AI as a managed cloud service involves the following high-level tasks:

1. Confirm that your OpenShift cluster meets all requirements.
2. Configure an identity provider for your OpenShift cluster.
3. Add administrative users for your OpenShift cluster.
4. Subscribe to the Red Hat OpenShift Data Science Add-on.  
For OpenShift Dedicated with a CCS for AWS or GCP, get a subscription through Red Hat.  
  
For ROSA, get a subscription through the AWS Marketplace.
5. Install the OpenShift Data Science Add-on.
6. Access the OpenShift AI dashboard.
7. Optionally, enable graphics processing units (GPUs) in OpenShift AI to ensure that your data scientists can use compute-heavy workloads in their models.

## CHAPTER 3. REQUIREMENTS FOR OPENSIFT AI

You must meet the following requirements before you can install OpenShift AI on your Red Hat OpenShift Dedicated or Red Hat OpenShift Service on Amazon Web Services (ROSA) cluster.

- **A subscription for Red Hat OpenShift Dedicated or a subscription for ROSA**

You can deploy Red Hat OpenShift Dedicated on your Amazon Web Services (AWS) or Google Cloud Platform (GCP) account by using the [Customer Cloud Subscription on AWS](#) or [Customer Cloud Subscription on GCP](#) model. Note that while Red Hat provides an option to install OpenShift Dedicated on a Red Hat cloud account, if you want to install OpenShift AI then you must install OpenShift Dedicated on your own cloud account.

Contact your Red Hat account manager to purchase a new Red Hat OpenShift Dedicated subscription. If you do not yet have an account manager, complete the form at <https://cloud.redhat.com/products/dedicated/contact/> to request one.

You can subscribe to Red Hat OpenShift Service on AWS (ROSA) directly from your AWS account or by contacting your Red Hat account manager.

- **A Red Hat customer account**

Go to OpenShift Cluster Manager (<http://console.redhat.com/openshift>) and log in or register for a new account.

- **Cluster administrator access to your OpenShift cluster**

Use an existing cluster or create a new cluster by following the steps in the relevant documentation:

- [Creating an OpenShift Dedicated cluster](#)
- [Creating a ROSA cluster with STS](#)

- **An OpenShift Dedicated or ROSA cluster configuration that meets the following configuration requirements.**

At least 2 worker nodes with at least 8 CPUs and 32 GiB RAM available for OpenShift AI to use when you install the Add-on. If this requirement is not met, the installation process fails to start and an error is displayed.

When you create a new cluster, select **m6a.2xlarge** for the computer node instance type to satisfy the requirements.

For an existing ROSA cluster, you can get the compute node instance type by using this command:

```
rosa list machinepools --cluster=cluster-name
```

You cannot alter a cluster's compute node instance type, but you can add an additional machine pool or modify the default pool to meet the minimum requirements. However, the minimum resource requirements must be met by a single machine pool in the cluster.

For more information, see the relevant documentation:

- [Creating a machine pool in OpenShift Dedicated](#)
- [OpenShift AI Service Definition](#)
- [Creating a machine pool in ROSA](#)

- [Prepare your environment](#) (ROSA)
- **For a ROSA cluster, select an access management strategy**

For installing OpenShift AI on a ROSA cluster, decide whether you want to install on a ROSA cluster that uses AWS Security Token Service (STS) or one that uses AWS Identity and Access Management (IAM) credentials. See [Install ROSA Classic clusters](#) for advice on deploying a ROSA cluster with or without AWS STS.
- **Install the Red Hat OpenShift Pipelines Operator**

OpenShift AI supports data science pipelines. A pipeline is a collection of Task resources that are arranged in a specific order of execution. By using Red Hat OpenShift AI pipelines, you can standardize and automate machine learning workflows to automate the build and deployment of your data science models. Before you can use pipelines with OpenShift AI, install the Red Hat OpenShift Pipelines Operator as described in [Installing OpenShift Pipelines](#).
- **Install KServe dependencies**

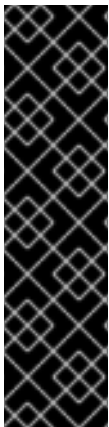
To support KServe components, you must install dependent Operators, including the Red Hat OpenShift Serverless and Red Hat OpenShift Service Mesh Operators. For more information, see [Serving large language models](#).

## CHAPTER 4. CONFIGURING AN IDENTITY PROVIDER FOR YOUR OPENSHIFT CLUSTER

Configure an identity provider for your OpenShift Dedicated or Red Hat OpenShift Service on Amazon Web Services (ROSA) cluster to manage users and groups.

Red Hat OpenShift AI supports the same authentication systems as Red Hat OpenShift Dedicated and ROSA. Check the appropriate documentation for your cluster for more information.

- [Supported identity providers on OpenShift Dedicated](#)
- [Supported identity providers on ROSA](#)



### IMPORTANT

Adding more than one OpenShift Identity Provider can create problems when the same user name exists in multiple providers.

When **mappingMethod** is set to **claim** (the default mapping method for identity providers) and multiple providers have credentials associated with the same user name, the first provider used to log in to OpenShift is the one that works for that user, regardless of the order in which identity providers are configured.

Refer to [Identity provider parameters](#) in the OpenShift Dedicated documentation for more information about mapping methods.

### Prerequisites

- Credentials for OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
- An existing OpenShift Dedicated cluster.

### Procedure

1. Log in to OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
2. Click **Clusters**. The **Clusters** page opens.
3. Click the name of the cluster to configure.
4. Click the **Access control** tab.
5. Click **Identity providers**.
6. Click **Add identity provider**.
  - a. Select your provider from the **Identity Provider** list.
  - b. Complete the remaining fields relevant to the identity provider that you selected. See [Configuring identity providers](#) for more information.
7. Click **Confirm**.

### Verification

- The configured identity providers are visible on the **Access control** tab of the **Cluster details** page.

#### Additional resources

- [Configuring identity providers](#)
- [Syncing LDAP groups](#)

## CHAPTER 5. ADDING ADMINISTRATIVE USERS

Before you can install and configure OpenShift AI for your data scientist users, you must define administrative users. Only administrative users can install and configure OpenShift AI.

### Prerequisites

- Credentials for OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
- An existing OpenShift Dedicated or Red Hat OpenShift Service on AWS (ROSA) cluster with an identity provider configured.

### Procedure

1. Log in to OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
2. Click **Clusters**. The **Clusters** page opens.
3. Click the name of the cluster to configure.
4. Click the **Access control** tab.
5. Click **Cluster Roles and Access**.
6. Under **Cluster administrative users** click the **Add user** button. The **Add cluster user** popover appears.
7. Enter the user name in the **User ID** field.
8. Select an appropriate **Group** for the user.



### IMPORTANT

If this user needs to use existing groups in an identity provider to control OpenShift AI access, select **cluster-admins**.

For more information about these user types, see [Managing administration roles and users](#) in the OpenShift Dedicated documentation or [Default cluster roles](#) in the ROSA documentation.

9. Click **Add user**.

### Verification

- The user name and selected group are visible in the list of **Cluster administrative users**.

### Additional resources

- [OpenShift Dedicated Cluster administration](#)
- [ROSA Cluster administration](#)

## CHAPTER 6. SUBSCRIBING TO THE RED HAT OPENSIFT AI CLOUD SERVICE

You can subscribe to the Red Hat OpenShift AI managed cloud service in the following ways:

- Subscribe through Red Hat if you have a Red Hat OpenShift Dedicated cluster deployed with a Customer Cloud Subscription (CCS) on Amazon Web Services (AWS) or Google Cloud Platform (GCP).
- Subscribe through the AWS Marketplace if you have a Red Hat OpenShift Service on AWS (ROSA) cluster.



### NOTE

You can also purchase Red Hat OpenShift AI as self-managed software. To purchase a new subscription, contact your Red Hat account manager. If you do not yet have an account manager, complete the form at <https://www.redhat.com/en/contact> to request one.

### 6.1. SUBSCRIBING TO THE OPENSIFT AI MANAGED CLOUD SERVICE ON AWS OR GCP

For a Red Hat OpenShift Dedicated cluster that is deployed on AWS or GCP, contact your Red Hat account manager to purchase a new subscription. If you do not yet have an account manager, complete the form at <https://cloud.redhat.com/products/dedicated/contact/> to request one.

#### Prerequisite

- You have worked with Red Hat Sales to enable a private offer of OpenShift AI, follow these steps to accept your offer and deploy the solution.

#### Procedure

1. Visit your Private Offer with the URL link provided by your Red Hat Sales representative.
2. Click **Accept Terms** to subscribe to the AMI Private Offer named **Openshift Data Science from AWS Marketplace**.
3. After accepting the offer terms, click **Continue to Configuration**.

### 6.2. SUBSCRIBING TO THE OPENSIFT AI MANAGED CLOUD SERVICE ON RED HAT OPENSIFT SERVICE ON AWS (ROSA)

For a ROSA cluster, you can subscribe to the OpenShift AI managed cloud service through the Amazon Web Services (AWS) Marketplace.

#### Prerequisites

- Access to a ROSA cluster, including permissions to view and install add-ons.
- An AWS account with permission to view and subscribe to offerings in the AWS marketplace.

## Procedure

1. In the AWS Console, navigate to the AWS Marketplace. For example:
  - a. Click the help icon and then select Getting Started Resource Center.
  - b. Select AWS Marketplace > Browse AWS Marketplace.
2. In the top **Search** field, type: **Red Hat OpenShift Data Science**
3. Select one of the two options depending on the geographical location of the billing address for your AWS account (note that this location might differ from the geographical location of the cluster):
  - Europe, the Middle East, and Africa (EMEA region)
  - North America and regions outside EMEA
4. Click **Continue to Subscribe**
5. Click **Continue to Configuration** and then select the appropriate fulfillment options. Note that some selectors might have only one option.
6. Click **Continue to Launch**.
7. Link your AWS account with your Red Hat account to complete your registration:
  - a. In the AWS Marketplace console, navigate to the **Manage Subscriptions** page.
  - b. On the **Red Hat OpenShift Data Science** tile, click **Set up product**
  - c. On the top banner, click **Set up account**  
This link takes you to the Red Hat Hybrid console.
  - d. If you are not already logged in, log in.
  - e. Review and then accept the terms and agreements.
  - f. Click **Connect accounts**.

## Verification

The Data Science product page opens.



## CHAPTER 7. INSTALLING OPENSIFT AI ON YOUR OPENSIFT CLUSTER

You can use Red Hat OpenShift Cluster Manager to install Red Hat OpenShift AI as an Add-on to your Red Hat OpenShift cluster.

### Prerequisites

- A subscription to the Red Hat OpenShift Data Science Add-on, as described in [Subscribing to the Red Hat OpenShift Data Science Add-on](#).
- If you purchased the Red Hat OpenShift Data Science Add-on for ROSA by using the AWS Marketplace, you have associated your AWS account with your Red Hat account as described in [Subscribing to the Red Hat OpenShift Data Science Add-on through the AWS Marketplace \(ROSA only\)](#).
- Credentials for Red Hat OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
- Administrator access to the OpenShift cluster.
- To support KServe components, you installed the dependent Operators, including the Red Hat OpenShift Serverless and Red Hat OpenShift Service Mesh Operators. For more information, see [Serving large language models](#).



### NOTE

For information about the lifecycle associated with Red Hat OpenShift AI, see [Red Hat OpenShift AI Life Cycle](#).

### Procedure

1. Log in to OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
2. Click **Clusters**.  
The **Clusters** page opens.
3. Click the name of the cluster you want to install OpenShift AI on.  
The **Details** page for the cluster opens.
4. Click the **Add-ons** tab and locate the **Red Hat OpenShift Data Science** tile.



### NOTE

If there is a **Prerequisites not met** warning message, click the **Prerequisites** tab. Note down the error message. If the error message states that you require a new machine pool, or that more resources are required, take the appropriate action to resolve the problem. You might need to add more resources to your cluster, or increase the size of your default machine pool. To increase your cluster's resources, contact your infrastructure administrator. For more information about increasing the size of your machine pool, see [Allocating additional resources to OpenShift AI users](#).

5. Select a **Subscription type**:

If you obtained your RHODS subscription through your Red Hat account manager, select **Standard** and then skip to Step 7.

If you obtained your RHODS subscription directly from the AWS Marketplace, select **Marketplace** and then continue to Step 6.

6. For a Marketplace subscription, select your AWS account number from the list.



#### NOTE

If your AWS account number is not in the list, you might need to link your Red Hat and AWS accounts, as described in [Subscribing to the Red Hat OpenShift Data Science Add-on through the AWS Marketplace \(ROSA only\)](#).

7. Click **Install**. The **Configure Red Hat OpenShift Data Science** pane appears.
8. In the **Notification email** field, enter any email addresses that you want to receive important alerts about the state of Red Hat OpenShift AI, such as outage alerts.
9. Click **Install**.

#### Verification

- In OpenShift Cluster Manager, on the **Add-ons** tab for the cluster, confirm that the OpenShift Data Science tile shows one of the following states:
  - **Installing** - installation is in progress; wait for this to change to **Installed**. This takes around 30 minutes.
  - **Installed** - installation is complete; verify that the **View in console** button is visible.
- In OpenShift Dedicated, click **Home** → **Projects** and confirm that the following project namespaces are visible and listed as **Active**:
  - **redhat-ods-applications**
  - **redhat-ods-monitoring**
  - **redhat-ods-operator**
  - **rhods-notebooks**

## CHAPTER 8. INSTALLING RED HAT OPENSIFT AI COMPONENTS BY USING THE WEB CONSOLE

The following procedure shows how to use the OpenShift Dedicated web console to install specific components of Red Hat OpenShift AI on your cluster.



### IMPORTANT

The following procedure describes how to create and configure a **DataScienceCluster** object to install Red Hat OpenShift AI components as part of a *new* installation. However, if you upgraded from version 1 of OpenShift AI (previously OpenShift Data Science), the upgrade process automatically created a default **DataScienceCluster** object. If you upgraded from a previous minor version, the upgrade process used the settings from the previous version's **DataScienceCluster** object. To inspect the **DataScienceCluster** object and change the installation status of Red Hat OpenShift AI components, see [Updating the installation status of Red Hat OpenShift AI components by using the web console](#).

### Prerequisites

- To support the KServe component, you installed dependent Operators, including the Red Hat OpenShift Serverless and Red Hat OpenShift Service Mesh Operators. For more information, see [Serving large language models](#).
- Red Hat OpenShift AI is installed as an Add-on to your Red Hat OpenShift cluster.
- You have cluster administrator privileges for your OpenShift Dedicated cluster.

### Procedure

1. Log in to the OpenShift Dedicated web console as a cluster administrator.
2. In the web console, click **Operators** → **Installed Operators** and then click the Red Hat OpenShift AI Operator.
3. Create a **DataScienceCluster** object to install OpenShift AI components by performing the following actions:
  - a. Click the **Data Science Cluster** tab.
  - b. Click **Create DataScienceCluster**.
  - c. For **Configure via**, select **YAML view**.  
An embedded YAML editor opens showing a default custom resource (CR) for the **DataScienceCluster** object.
  - d. In the **spec.components** section of the CR, for each OpenShift AI component shown, set the value of the **managementState** field to either **Managed** or **Removed**. These values are defined as follows:

#### Managed

The Operator actively manages the component, installs it, and tries to keep it active. The Operator will upgrade the component only if it is safe to do so.

#### Removed

The Operator actively manages the component but does not install it. If the component is already installed, the Operator will try to remove it.



### IMPORTANT

- To learn how to install the KServe component, which is used by the single model serving platform to serve large language models, see [Serving large language models](#).
- The CodeFlare and KubeRay components are Technology Preview features only. Technology Preview features are not supported with Red Hat production service level agreements (SLAs) and might not be functionally complete. Red Hat does not recommend using them in production. These features provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process. For more information about the support scope of Red Hat Technology Preview features, see [Technology Preview Features Support Scope](#).
- To learn how to configure the distributed workloads feature that uses the CodeFlare and KubeRay components, see [Configuring distributed workloads](#).

4. Click **Create**.

### Verification

- On the **DataScienceClusters** page, click the **default-dsc** object and then perform the following actions:
  - Select the **YAML** tab.
  - In the **installedComponents** section, confirm that the components you installed have a status value of **true**.
- In the OpenShift Dedicated web console, click **Workloads** → **Pods** and then perform the following actions:
  - In the **Project** list at the top of the page, select the **redhat-ods-applications** project.
  - In the project, confirm that there are running pods for each of the OpenShift AI components that you installed.

# CHAPTER 9. BACKING UP STORAGE DATA FROM AMAZON EBS

Red Hat recommends that you back up the data on your persistent volume claims (PVCs) regularly. Backing up your data is particularly important before deleting a user and before uninstalling OpenShift AI, as all PVCs are deleted when OpenShift AI is uninstalled.

## Prerequisites

- You have credentials for Red Hat OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
- You have administrator access to the OpenShift Dedicated cluster.
- You have credentials for the Amazon Web Services (AWS) account that the OpenShift Dedicated cluster is deployed under.

## Procedure

1. Determine the IDs of the persistent volumes (PVs) that you want to back up.
  - a. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
  - b. Click **Home** → **Projects**.
  - c. Click the **rhods-notebooks** project.  
The **Details** page for the project opens.
  - d. Click the **PersistentVolumeClaims** in the **Inventory** section.  
The **PersistentVolumeClaims** page opens.
  - e. Note the ID of the persistent volume (PV) that you want to back up.



### NOTE

The persistent volumes (PV) that you make a note of are required to identify the correct EBS volume to back up in your AWS instance.

2. Locate the EBS volume containing the PVs that you want to back up.  
See [Amazon Web Services documentation: Create Amazon EBS snapshots](#) for more information.
  - a. Log in to AWS (<https://aws.amazon.com>) and ensure that you are viewing the region that your OpenShift Dedicated cluster is deployed in.
  - b. Click **Services**.
  - c. Click **Compute** → **EC2**.
  - d. Click **Elastic Block Storage** → **Volumes** in the side navigation.  
The **Volumes** page opens.
  - e. In the search bar, enter the ID of the persistent volume (PV) that you made a note of earlier.  
The **Volumes** page reloads to display the search results.

- f. Click on the volume shown and verify that any **kubernetes.io/created-for/pvc/namespace** tags contain the value **rhods-notebooks**, and any **kubernetes.io/created-for/pvc/name** tags match the name of the persistent volume that the EC2 volume is being used for, for example, **jupyter-nb-user1-pvc**.
3. Back up the EBS volume that contains your persistent volume (PV).
    - a. Right-click on the volume that you want to back up and select **Create Snapshot** from the list.  
The **Create Snapshot** page opens.
    - b. Enter a **Description** for the volume.
    - c. Click **Create Snapshot**.  
The snapshot of the volume is created.
    - d. Click **Close**.

### Verification

- The snapshot that you created is visible on the **Snapshots** page in AWS.

### Additional resources

- [Amazon Web Services documentation: Create Amazon EBS snapshots](#)

# CHAPTER 10. BACKING UP STORAGE DATA FROM GOOGLE PERSISTENT DISK

Red Hat recommends that you back up the data on your persistent volume claims (PVCs) regularly. Backing up your data is particularly important before deleting a user and before uninstalling OpenShift AI, as all PVCs are deleted when OpenShift AI is uninstalled.

## Prerequisites

- You have credentials for Red Hat OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
- You have administrator access to the OpenShift Dedicated cluster.
- You have credentials for the Google Cloud Platform (GCP) account that the OpenShift Dedicated cluster is deployed under.

## Procedure

1. Determine the IDs of the persistent volumes (PVs) that you want to back up.
  - a. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
  - b. Click **Home** → **Projects**.
  - c. Click the **rhods-notebooks** project.  
The **Details** page for the project opens.
  - d. Click the **PersistentVolumeClaims** in the **Inventory** section.  
The **PersistentVolumeClaims** page opens.
  - e. Note the ID of the persistent volume (PV) that you want to back up.  
The persistent volume (PV) IDs are required to identify the correct persistent disk to back up in your GCP instance.
2. Locate the persistent disk containing the PVs that you want to back up.
  - a. Log in to the Google Cloud console (<https://console.cloud.google.com>) and ensure that you are viewing the region that your OpenShift Dedicated cluster is deployed in.
  - b. Click the navigation menu (≡) and then click **Compute Engine**.
  - c. From the side navigation, under **Storage**, click **Disks**.  
The **Disks** page opens.
  - d. In the **Filter** query box, enter the ID of the persistent volume (PV) that you made a note of earlier.  
The **Disks** page reloads to display the search results.
  - e. Click on the disk shown and verify that any **kubernetes.io/created-for/pvc/namespace** tags contain the value **rhods-notebooks**, and any **kubernetes.io/created-for/pvc/name** tags match the name of the persistent volume that the persistent disk is being used for, for example, **jupyterhub-nb-user1-pvc**.
3. Back up the persistent disk that contains your persistent volume (PV).
  - a. Select **CREATE SNAPSHOT** from the top navigation

- a. Select **CREATE SNAPSHOT** from the top navigation.  
The **Create a snapshot** page opens.
- b. Enter a unique **Name** for the snapshot.
- c. Under **Source disk**, verify the persistent disk you want to back up is displayed.
- d. Change any optional settings as needed.
- e. Click **CREATE**.  
The snapshot of the persistent disk is created.

### Verification

- The snapshot that you created is visible on the **Snapshots** page in GCP.

### Additional resources

- [Google Cloud documentation: Create and manage disk snapshots](#)



## CHAPTER 11. UNINSTALLING OPENSIFT AI

You can use Red Hat OpenShift Cluster Manager to safely uninstall Red Hat OpenShift AI from your OpenShift cluster.

### Prerequisites

- Credentials for Red Hat OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
- Administrator access to the OpenShift cluster.
- For AWS clusters, you have backed up the EBS volume containing your Persistent Volume Claims (PVCs). See [Amazon Web Services documentation: Create Amazon EBS snapshots](#) for more information.
- For GCP clusters, you have backed up the persistent disk containing your Persistent Volume Claims (PVCs). See [Google Cloud documentation: Create and manage disk snapshots](#) for more information.

### Procedure

1. Log in to Red Hat OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
2. Click **Clusters**.  
The **Clusters** page opens.
3. Click the name of the cluster that hosts the instance OpenShift AI to uninstall.  
The **Details** page for the cluster opens.
4. Click the **Add-ons** tab and locate the **Red Hat OpenShift Data Science** tile.
5. Click **Uninstall**.  
This process takes approximately 30 minutes to complete. Do not manually delete any resources while uninstalling OpenShift AI, as this can interfere with the uninstall process.

OpenShift AI is uninstalled and any persistent volume claims (PVCs) associated with your OpenShift AI instance are deleted. However, any user groups for OpenShift AI that you previously created remain on your cluster.

### Verification

- In Red Hat OpenShift Cluster Manager, on the **Add-ons** tab for the cluster, confirm that the OpenShift Data Science tile does not show the **Installed** state.
- In your OpenShift cluster, click **Home** → **Projects** and confirm that the following project namespaces are not visible:
  - **redhat-ods-applications**
  - **redhat-ods-monitoring**
  - **redhat-ods-operator**

### Additional resources

- [Amazon Web Services documentation: Create Amazon EBS snapshots](#)

- [Google Cloud documentation: Create and manage disk snapshots](#)
- [Deleting users and user resources](#)


## CHAPTER 12. ACCESSING THE DASHBOARD

After you have installed OpenShift AI and added users, you can access the URL for your OpenShift AI console and share the URL with the users to let them log in and work on their models.

### Prerequisites

- You have installed OpenShift AI on your OpenShift Dedicated or Red Hat OpenShift Service on Amazon Web Services (ROSA) cluster.
- You have added at least one user to the user group for OpenShift AI as described in [Adding users](#).

### Procedure

1. Log in to OpenShift web console.
2. Click the application launcher (  ).
3. Right-click on **Red Hat OpenShift AI** and copy the URL for your OpenShift AI instance.
4. Provide this instance URL to your data scientists to let them log in to OpenShift AI.

### Verification

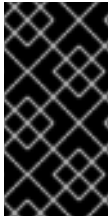
- Confirm that you and your users can log in to OpenShift AI by using the instance URL.

### Additional resources

- [Logging in to OpenShift AI](#)
- [Adding users](#)

## CHAPTER 13. 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

The NVIDIA GPU Add-on is no longer supported. Instead, enable GPUs by installing the NVIDIA GPU Operator. If your deployment has a previously-installed NVIDIA GPU Add-on, before you install the NVIDIA GPU Operator, use Red Hat OpenShift Cluster Manager to uninstall the NVIDIA GPU Add-on from your cluster.

### Prerequisites

- You have logged in to your OpenShift cluster.
- You have the **cluster-admin** role in your OpenShift 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 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 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 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](#).

## CHAPTER 14. TROUBLESHOOTING COMMON INSTALLATION PROBLEMS

If you are experiencing difficulties installing the Red Hat OpenShift Data Science Add-on, read this section to understand what could be causing the problem, and how to resolve the problem.

If you cannot see the problem here or in the release notes, contact Red Hat Support.

### 14.1. THE RED HAT OPENSIFT AI OPERATOR CANNOT BE RETRIEVED FROM THE IMAGE REGISTRY

#### Problem

When attempting to retrieve the Red Hat OpenShift AI Operator from the image registry, an **Failure to pull from quay** error message appears. The Red Hat OpenShift AI Operator might be unavailable for retrieval in the following circumstances:

- The image registry is unavailable.
- There is a problem with your network connection.
- Your cluster is not operational and is therefore unable to retrieve the image registry.

#### Diagnosis

Check the logs in the **Events** section in OpenShift Dedicated for further information about the **Failure to pull from quay** error message.

#### Resolution

- To resolve this issue, contact Red Hat support.

### 14.2. OPENSIFT AI CANNOT BE INSTALLED DUE TO INSUFFICIENT CLUSTER RESOURCES

#### Problem

When attempting to install OpenShift AI, an error message appears stating that installation prerequisites have not been met.

#### Diagnosis

1. Log in to OpenShift Cluster Manager (<https://console.redhat.com/openshift/>).
2. Click **Clusters**.  
The **Clusters** page opens.
3. Click the name of the cluster you want to install OpenShift AI on.  
The **Details** page for the cluster opens.
4. Click the **Add-ons** tab and locate the **Red Hat OpenShift Data Science** tile.
5. Click **Install**. The **Configure Red Hat OpenShift Data Science** pane appears.

6. If the installation fails, click the **Prerequisites** tab.
7. Note down the error message. If the error message states that you require a new machine pool, or that more resources are required, take the appropriate action to resolve the problem.

### Resolution

- You might need to add more resources to your cluster, or increase the size of your machine pool. To increase your cluster's resources, contact your infrastructure administrator. For more information about increasing the size of your machine pool, see [Nodes](#) and [Allocating additional resources to OpenShift AI users](#).

## 14.3. THE DEDICATED-ADMINS ROLE-BASED ACCESS CONTROL (RBAC) POLICY CANNOT BE CREATED

### Problem

The Role-based access control (RBAC) policy for the dedicated-admins group in the target project cannot be created. This issue occurs in unknown circumstances.

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-<random string>** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Attempt to create the RBAC policy for dedicated admins group in \$target\_project failed.** error message.

### Resolution

- Contact Red Hat support.

## 14.4. OPENSIFT AI DOES NOT INSTALL ON UNSUPPORTED INFRASTRUCTURE

### Problem

Customer deploying on an environment not documented as being supported by the RHODS operator.

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.

3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-<random string>** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Deploying on \$infrastructure, which is not supported. Failing Installation** error message.

### Resolution

Before proceeding with a new installation, ensure that you have a fully supported environment on which to install OpenShift AI. For more information, see [Requirements for OpenShift AI](#).

## 14.5. THE CREATION OF THE OPENSIFT AI CUSTOM RESOURCE (CR) FAILS

### Problem

During the installation process, the OpenShift AI Custom Resource (CR) does not get created. This issue occurs in unknown circumstances.

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-<random string>** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Attempt to create the ODH CR failed.** error message.

### Resolution

Contact Red Hat support.

## 14.6. THE CREATION OF THE OPENSIFT AI NOTEBOOKS CUSTOM RESOURCE (CR) FAILS

### Problem

During the installation process, the OpenShift AI Notebooks Custom Resource (CR) does not get created. This issue occurs in unknown circumstances.

### Diagnosis



1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-`<random string>`** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Attempt to create the RHODS Notebooks CR failed.** error message.

### Resolution

Contact Red Hat support.

## 14.7. THE DEAD MAN'S SNITCH OPERATOR'S SECRET DOES NOT GET CREATED

### Problem

An issue with Managed Tenants SRE automation process causes the Dead Man's Snitch operator's secret to not get created.

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-`<random string>`** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Dead Man Snitch secret does not exist.** error message.

### Resolution

Contact Red Hat support.

## 14.8. THE PAGERDUTY SECRET DOES NOT GET CREATED

### Problem

An issue with Managed Tenants SRE automation process causes the PagerDuty's secret to not get created.

### Diagnosis

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-`<random string>`** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Pagerduty secret does not exist** error message.

### Resolution

Contact Red Hat support.

## 14.9. THE SMTP SECRET DOES NOT EXIST

### Problem

An issue with Managed Tenants SRE automation process causes the SMTP secret to not get created.

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-`<random string>`** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: SMTP secret does not exist** error message.

### Resolution

Contact Red Hat support.

## 14.10. THE ODH PARAMETER SECRET DOES NOT GET CREATED

### Problem

An issue with the OpenShift Data Science Add-on's flow could result in the ODH parameter secret to not get created.

### Diagnosis

1. In the OpenShift Dedicated web console, change into the **Administrator** perspective.
2. Click **Workloads** → **Pods**.
3. Set the **Project** to **All Projects** or **redhat-ods-operator**.
4. Click the **rhods-operator-<random string>** pod.  
The **Pod details** page appears.
5. Click **Logs**.
6. Select **rhods-deployer** from the drop-down list
7. Check the log for the **ERROR: Addon managed odh parameter secret does not exist.** error message.

### Resolution

Contact Red Hat support.