Red Hat OpenShift AI Self-Managed 2.10

管理资源

在 OpenShift AI 中管理集群资源、Jupyter 笔记本和数据备份

Last Updated: 2024-06-28

Red Hat OpenShift AI Self-Managed 2.10 管理资源

在 OpenShift AI 中管理集群资源、Jupyter 笔记本和数据备份

法律通告

摘要

在 OpenShift AI 中管理集群资源、Jupyter 笔记本和数据备份。

目录

前言

作为 OpenShift AI 管理员，您可以管理以下资源：

- 仪表板界面，包括导航菜单选项的可见性
- 仪表板中显示的应用程序
- 支持计算密集型数据科学工作的集群资源
- Jupyter 笔记本服务器
- 数据存储备份

您还可以指定是否允许红帽收集有关集群中 OpenShift AI 使用量的数据。

第 1 章 自定义仪表板

OpenShift AI 仪表板提供专为大多数场景而设计的功能。这些功能在 **OdhDashboardConfig** 自定义资源(CR)文件中配置。

要查看 OpenShift AI 仪表板配置文件中选项的描述，请参阅 控制面板配置选项。

作为管理员，您可以自定义仪表板的接口，例如显示或隐藏一些仪表板导航菜单选项。要更改仪表板的默认设置，请编辑 **OdhDashboardConfig** 自定义资源(CR)文件，如 编辑仪表板配置文件 中所述。

1.1. 编辑仪表板配置文件

作为管理员，您可以通过编辑仪表板配置文件来自定义仪表板接口。

先决条件

- 具有 OpenShift Container Platform 集群的集群管理员特权。

流程

1. 以集群管理员身份登录 OpenShift Container Platform 控制台。
2. 在 **Administrator** 视角中，点 **Home → API Explorer**。
3. 在搜索栏中，输入 **OdhDashboardConfig** 以根据 kind 进行过滤。
4. 单击 **OdhDashboardConfig** 自定义资源(CR)以打开资源详情页面。
5. 从 **Project** 列表中选择 **redhat-ods-applications** 项目。
6. 点 **实例** 选项卡。
7. 点 **odh-dashboard-config** 实例打开详情页面。
8. 点 **YAML** 标签。以下是显示默认值的 **OdhDashboardConfig** 文件示例：

```
apiVersion: opendatahub.io/v1alpha
kind: OdhDashboardConfig
metadata:
  name: odh-dashboard-config
spec:
  dashboardConfig:
    enablement: true
    disableBYONImageStream: false
    disableClusterManager: false
    disableISVBadges: false
    disableInfo: false
    disableSupport: false
    disableTracking: true
    disableProjects: true
    disablePipelines: true
    disableModelServing: true
    disableProjectSharing: true
    disableCustomServingRuntimes: false
    disableAcceleratorProfiles: true
```

```
  modelMetricsNamespace: ''
  disablePerformanceMetrics: false
notebookController:
  enabled: true
notebookSizes:
  - name: Small
    resources:
      limits:
        cpu: '2'
        memory: 2Gi
      requests:
        cpu: '1'
        memory: 1Gi
  - name: Medium
    resources:
      limits:
        cpu: '4'
        memory: 4Gi
      requests:
        cpu: '2'
        memory: 2Gi
  - name: Large
    resources:
      limits:
        cpu: '8'
        memory: 8Gi
      requests:
        cpu: '4'
        memory: 4Gi
modelServerSizes:
  - name: Small
    resources:
      limits:
        cpu: '2'
        memory: 8Gi
      requests:
        cpu: '1'
        memory: 4Gi
  - name: Medium
    resources:
      limits:
        cpu: '8'
        memory: 10Gi
      requests:
        cpu: '4'
        memory: 8Gi
  - name: Large
    resources:
      limits:
        cpu: '10'
        memory: 20Gi
      requests:
        cpu: '6'
        memory: 16Gi
groupsConfig:
  adminGroups: 'odh-admins'
```

```
  allowedGroups: 'system:authenticated'
templateOrder:
  - 'ovms'
templateDisablement:
  - 'ovms'
```

9. 编辑您要更改的选项值。
10. 点 **Save** 以应用您的更改，然后点 **Reload** 以确保您的更改同步到集群。

验证

登录到 OpenShift AI 并验证是否应用了仪表板配置。

1.2. 仪表板配置选项

OpenShift AI 仪表板包括一组默认启用的核心功能，它们适用于大多数场景。管理员可以从 OpenShift Container Platform 中的 **OdhDashboardConfig** 自定义资源(CR)配置 OpenShift AI 仪表板。

表 1.1. 仪表板功能配置选项

功能	default	描述
dashboardConfig: 启用	**true**	启用管理员用户将应用程序添加到 OpenShift AI dashboard **Application → Enabled** 页面中。要禁用此功能，请将值设为 **false**。
dashboardConfig: disableInfo	**false**	在 **Applications → Explore** 页面中，当用户点击应用程序标题时，会打开一个信息面板，其中包含有关应用程序的详情。要禁用 **Applications → Explore** 页面中所有应用程序的信息面板，请将值设为 **true**。
dashboardConfig: disableSupport	**false**	当用户点击仪表板工具栏中的 Help 图标时，会显示 **Support** menu 选项。要隐藏此菜单选项，请将值设为 **true**。
dashboardConfig: disableClusterManager	**false**	在仪表板导航菜单中显示 **Settings → Cluster settings** 选项。要隐藏此菜单选项，请将值设为 **true**。
dashboardConfig: disableTracking	**true**	允许红帽收集有关集群中 OpenShift AI 使用量的数据。要启用数据收集，请将值设为 **false**。您还可以在 OpenShift AI 仪表板界面的 **Settings → Cluster settings** 导航菜单中设置这个选项。
dashboardConfig: disableBYONImageStream	**false**	在仪表板导航菜单中显示 **Settings → Notebook** 镜像选项。要隐藏此菜单选项，请将值设为 **false**。
dashboardConfig: disableISVBadges	**false**	在标题上显示标签，指示应用程序是"Red Hat managed"、"Partner managed" 还是 "Self-managed"。要隐藏这些标签，请将值设为 **true**。
dashboardConfig: disableUserManagement	**false**	在仪表板导航菜单中显示 **Settings → User management** 选项。要隐藏此菜单选项，请将值设为 **true**。

dashboardConfig: disableProjects	**false**	在仪表板导航菜单中显示 Data Science Projects 选项。要隐藏此菜单选项，请将值设为 **true**。
dashboardConfig: disablePipelines	**false**	在仪表板导航菜单中显示 Data Science Pipelines 选项。要隐藏此菜单选项，请将值设为 **true**。
dashboardConfig: disableModelServing	**false**	在仪表板导航菜单和数据科学项目的组件列表中显示 Model Serving 选项。要从仪表板导航菜单和数据科学项目的组件列表中隐藏 Model Serving，请将值设为 **true**。
dashboardConfig: disableProjectSharing	**false**	允许用户与其他用户共享其数据科学项目的访问权限。要防止用户共享数据科学项目，请将值设为 **true**。
dashboardConfig: disableCustomServingRuntimes	**false**	在仪表板导航菜单 **中显示 Serving** runtime 选项。要隐藏此菜单选项，请将值设为 **true**。
dashboardConfig: disableKServe	**false**	启用选择 KServe 作为 Serving 平台的功能。要禁用此功能，请将值设为 **true**。
dashboardConfig: disableModelMesh	**false**	启用选择 ModelMesh 作为 Serving 平台的功能。要禁用此功能，请将值设为 **true**。
dashboardConfig: disableAcceleratorProfiles	**false**	在仪表板导航菜单中显示 **加速器配置集** 选项。要隐藏此菜单选项，请将值设为 **true**。
dashboardConfig: modelMetricsNamespace	**false**	启用安装 Model Serving Metrics 的 Prometheus Operator 的命名空间。
dashboardConfig: disablePerformanceMetrics	**false**	显示 **Model Serving** 页面中的 **Endpoint Performance** 选项卡。要隐藏此选项卡，请将值设为 **true**。
notebookController: enabled	**true**	控制 Notebook Controller 选项，比如在仪表板中启用它以及哪些部分可见。
notebookSizes		允许您自定义笔记本的名称和资源。使用 Notebook Controller 生成笔记本时出现的下拉菜单中会显示 Kubernetes 风格的大小。注：这些大小必须遵循惯例。例如，请求必须小于限制。
ModelServerSizes		允许您为模型服务器自定义名称和资源。
groupsConfig		控制仪表板功能的访问，如允许用户的生成者，以及管理员用户的集群设置 UI。
templateOrder		指定自定义 Serving Runtime 模板的顺序。当用户创建新模板时，会将其添加到此列表中。

第 2 章 管理仪表板中显示的应用程序

2.1. 将应用程序添加到仪表板中

如果您在 OpenShift Container Platform 集群中安装了应用程序，您可以将应用程序的标题添加到 OpenShift AI 仪表板(Application → **Enabled** 页面)中，使其可以被 OpenShift AI 用户访问。

先决条件

- **具有 OpenShift Container Platform 集群的集群管理员特权。**
- **仪表板配置启用选项被设置为 true （默认值）。请注意，管理员用户可以禁用此功能，如 防止用户将应用程序添加到仪表板 中所述。**

流程

1. **以集群管理员身份登录 OpenShift Container Platform 控制台。**
2. **在 Administrator 视角中，点 Home → API Explorer。**
3. **在 API Explorer 页面中，搜索 OdhApplication kind。**
4. **点 OdhApplication kind 打开资源详情页面。**
5. **在 OdhApplication details 页面中，从Project 列表中选择 redhat-ods-applications 项目。**
6. **点 实例 选项卡。**
7. 单击 Create OdhApplication。
8. **在 Create OdhApplication 页面中，复制以下代码并将其粘贴到 YAML 编辑器中。**

```
apiVersion: dashboard.opendatahub.io/v1
kind: OdhApplication
metadata:
  name: examplename
  namespace: redhat-ods-applications
  labels:
    app: odh-dashboard
    app.kubernetes.io/part-of: odh-dashboard
spec:
  enable:
    validationConfigMap: examplename-enable
  img: >-
    <svg width="24" height="25" viewBox="0 0 24 25" fill="none"
xmlns="http://www.w3.org/2000/svg">
    <path d="path data" fill="#ee0000"/>
    </svg>
  getStartedLink: 'https://example.org/docs/quickstart.html'
  route: exampleroutename
  routeNamespace: examplenamespace
  displayName: Example Name
  kfdefApplications: []
  support: third party support
  csvName: ''
```

```
provider: example
docsLink: 'https://example.org/docs/index.html'
quickStart: ''
getStartedMarkDown: >-
  # Example

  Enter text for the information panel.

description: >-
  Enter summary text for the tile.
category: Self-managed | Partner managed | {org-name} managed
```

9. 修改应用程序代码中的参数。

   提示

   要查看 YAML 文件示例，点 Home → API Explorer，选择 **OdhApplication**，点 Instances 选项卡，选择一个实例，然后点击 YAML 选项卡。

10. 点 Create。此时会出现应用程序详情页面。
11. 登录到 OpenShift AI。
12. 在左侧菜单中，点击 Applications → Explore。
13. 找到应用程序的新标题并点它。
14. 在应用程序的信息窗格中，单击 启用。

验证

- 在 OpenShift AI 仪表板的左侧菜单中，点 Applications → Enabled 并验证您的应用程序是否可用。

2.2. 防止用户将应用程序添加到仪表板

默认情况下，admin 用户被允许将应用程序添加到 OpenShift AI dashboard Application → Enabled 页面中。

您可以禁止 admin 用户将应用程序添加到仪表板中。

注：Jupyter 标题被默认启用。要禁用它，请参阅 Hiding the default Jupyter application。

前提条件

- 具有 OpenShift Container Platform 集群的集群管理员特权。

流程

1. 以集群管理员身份登录 OpenShift Container Platform 控制台。
2. 打开仪表板配置文件：
   a. 在 Administrator 视角中，点 Home → API Explorer。

b. 在搜索栏中，输入 **OdhDashboardConfig** 以根据 kind 进行过滤。

c. 单击 **OdhDashboardConfig** 自定义资源(CR)以打开资源详情页面。

d. 从 Project 列表中选择 **redhat-ods-applications** 项目。

e. 点 **实例** 选项卡。

f. 点 **odh-dashboard-config** 实例打开详情页面。

g. 点 **YAML** 标签。

3. 在 **spec:dashboardConfig** 部分中，将 **enablement** 的值设置为 **false** 来禁用仪表板用户将应用程序添加到仪表板的能力。

4. 点 **Save** 以应用您的更改，然后点 **Reload** 以确保您的更改同步到集群。

验证

打开 OpenShift AI dashboard Application → Enabled 页面。

2.3. 禁用连接到 OPENSHIFT AI 的应用

您可以禁用应用程序和组件，以便在不再使用它们时（例如，不再需要这些应用程序或应用程序许可证也过期），这些应用程序不再出现在 OpenShift AI 仪表板中。

禁用未使用的应用程序后，您的数据科学家可以从其 OpenShift AI 仪表板中手动删除这些应用程序卡，以便他们能够专注于最有可能使用的应用程序。有关手动删除应用程序标题的更多信息，请参阅 从仪表板中删除禁用的应用程序。

重要

禁用以下应用程序时不要遵循这个步骤：

- Anaconda 专业版.您不能手动禁用 Anaconda Professional 版本。只有在许可证过期时才会自动禁用。

先决条件

- 已登陆到 OpenShift Container Platform Web 控制台。
- 您是 OpenShift Container Platform 中的 **cluster-admins** 用户组的一部分。
- 您已在 OpenShift Container Platform 集群上安装或配置了服务。
- 您要禁用的应用程序或组件被启用并在 Enabled 页面中显示。

流程

1. 在 OpenShift Container Platform web 控制台中切换到 Administrator 视角。
2. 切换到 **redhat-ods-applications** 项目。
3. 点 Operators → Installed Operators。

4. 点您要卸载的 Operator。您可以在 Filter by name 字段中输入关键字，以帮助您更快地找到 Operator。

5. 使用 Operator 界面中的标签页删除任何 Operator 资源或实例。
在安装过程中，一些 Operator 要求管理员使用 Operator 界面中的标签页创建资源或启动进程实例。这些必须在 Operator 可以正确卸载前删除。

6. 在 Operator Details 页面中，点 Actions 下拉菜单并选择 Uninstall Operator。
此时会显示 Uninstall Operator? 对话框。

7. 选择 Uninstall 来卸载 Operator、Operator 部署和 pod。完成后，Operator 会停止运行，不再接收更新。

重要

删除 Operator 不会删除 Operator 的任何自定义资源定义或受管资源。自定义资源定义和受管资源仍然存在，必须手动清理。Operator 部署的任何应用程序以及配置的任何非集群资源都会继续运行，必须手动清理。

验证

- Operator 从其目标集群卸载。
- Operator 不再出现在 Installed Operators 页面中。
- 您的数据科学家将无法使用禁用的应用程序，并在 OpenShift AI 仪表板的 Enabled 页面中被标记为禁用。此操作可能需要几分钟时间在删除 Operator 后进行。

2.4. 显示或隐藏有关启用的应用程序的信息

如果您在 OpenShift Container Platform 集群中安装了另一个应用程序，您可以在 OpenShift AI 仪表板 (Application → Enabled 页面)中添加标题，使其可以被 OpenShift AI 用户访问。

先决条件

- 具有 OpenShift Container Platform 集群的集群管理员特权。

流程

1. 以集群管理员身份登录 OpenShift Container Platform 控制台。
2. 在 Administrator 视角中，点 Home → API Explorer。
3. 在 API Explorer 页面中，搜索 **OdhApplication** kind。
4. 点 **OdhApplication** kind 打开资源详情页面。
5. 在 OdhApplication details 页面中，从 Project 列表中选择 **redhat-ods-applications** 项目。
6. 点 实例 选项卡。
7. 单击 Create OdhApplication。
8. 在 Create OdhApplication 页面中，复制以下代码并将其粘贴到 YAML 编辑器中。

```
apiVersion: dashboard.opendatahub.io/v1
kind: OdhApplication
metadata:
  name: examplename
  namespace: redhat-ods-applications
  labels:
    app: odh-dashboard
    app.kubernetes.io/part-of: odh-dashboard
spec:
  enable:
    validationConfigMap: examplename-enable
  img: >-
    <svg width="24" height="25" viewBox="0 0 24 25" fill="none"
xmlns="http://www.w3.org/2000/svg">
    <path d="path data" fill="#ee0000"/>
    </svg>
  getStartedLink: 'https://example.org/docs/quickstart.html'
  route: exampleroutename
  routeNamespace: examplenamespace
  displayName: Example Name
  kfdefApplications: []
  support: third party support
  csvName: ''
  provider: example
  docsLink: 'https://example.org/docs/index.html'
  quickStart: ''
  getStartedMarkDown: >-
    # Example

    Enter text for the information panel.

  description: >-
    Enter summary text for the tile.
  category: Self-managed | Partner managed | Red Hat managed
```

9. 修改应用程序代码中的参数。

   提示

   要查看 YAML 文件示例，点 Home → API Explorer，选择 **OdhApplication**，点 Instances 选项卡，选择一个实例，然后点击 YAML 选项卡。

10. 点 Create。此时会出现应用程序详情页面。
11. 登录到 OpenShift AI。
12. 在左侧菜单中，点击 Applications → Explore。
13. 找到应用程序的新标题并点它。
14. 在应用程序的信息窗格中，单击 启用。

验证

- 在 OpenShift AI 仪表板的左侧菜单中，点 Applications → Enabled 并验证您的应用程序是否可用。

2.5. 隐藏默认的 JUPYTER 应用程序

OpenShift AI 仪表板包括 Jupyter 作为启用的应用程序。

要隐藏 Enabled 应用程序列表中的 Jupyter 标题，请编辑仪表板配置文件。

前提条件

- 具有 OpenShift Container Platform 集群的集群管理员特权。

流程

1. 以集群管理员身份登录 OpenShift Container Platform 控制台。
2. 打开仪表板配置文件：
   a. 在 Administrator 视角中，点 Home → API Explorer。
   b. 在搜索栏中，输入 OdhDashboardConfig 以根据 kind 进行过滤。
   c. 单击 OdhDashboardConfig 自定义资源(CR)以打开资源详情页面。
   d. 从 Project 列表中选择 redhat-ods-applications 项目。
   e. 点 实例 选项卡。
   f. 点 odh-dashboard-config 实例打开详情页面。
   g. 点 YAML 标签。
3. 在 spec:notebookController 部分中，将 enabled 的值设置为 false，以隐藏 Enabled 应用程序列表中的 Jupyter 标题。
4. 点 Save 以应用您的更改，然后点 Reload 以确保您的更改同步到集群。

验证

在 OpenShift AI 仪表板中，选择 Applications> Enabled。您不应该看到 Jupyter 标题。

2.6. 管理员对 JUPYTER 中的常见问题进行故障排除

如果您的用户在与 Jupyter、其笔记本或他们的笔记本服务器相关的 Red Hat OpenShift AI 数据存储中遇到错误，请阅读本节以了解导致问题的原因，以及如何解决这个问题。

如果在此处或发行注记中无法找到相关的信息，请联系红帽支持团队。

2.6.1. 用户在登录到 Jupyter 时收到 404: Page not found 错误

问题

如果您为 OpenShift AI 配置了专用用户组，则可能无法将用户名添加到 OpenShift AI 的默认用户组中。

诊断

检查用户是否是默认用户组的一部分。

1. 查找允许访问 Jupyter 的组名称。
   a. 登陆到 OpenShift Container Platform Web 控制台。
   b. 点 User Management → Groups。
   c. 点用户组的名称，如 rhoai-users。
   此时会出现该组的组详细信息页面。
2. 点组的 Details 选项卡，并确认相关组的 Users 部分包含有权访问 Jupyter 的用户。

解决方案

- 如果用户没有添加到有权访问 Jupyter 的任何组中，请按照 添加用户 来添加它们。
- 如果用户已添加到有访问 Jupyter 的组中，请联系红帽支持。

2.6.2. 用户的笔记本服务器没有启动

问题

托管用户笔记本服务器的 OpenShift Container Platform 集群可能无法访问充足的资源，或者 Jupyter pod 可能出现问题。

诊断

1. 登陆到 OpenShift Container Platform Web 控制台。
2. 删除并重启此用户的笔记本服务器 Pod。
   a. 点 Workloads → Pods，将项目设置为 rhods-notebooks。
   b. 搜索属于此用户的笔记本服务器 pod，例如 jupyter-nb-<username>-*。
   如果笔记本服务器 pod 存在，则笔记本服务器 pod 中可能会出现间歇性失败。

   如果用户的笔记本服务器 pod 不存在，请继续诊断。
3. 根据所选笔记本服务器镜像所需的资源，检查 OpenShift Container Platform 集群中当前可用的资源。
   如果有足够 CPU 和 RAM 的 worker 节点可用于在集群中调度，请继续诊断过程。
4. 检查 Jupyter pod 的状态。

解决方案

- 如果笔记本服务器 pod 出现间歇性失败：
  a. 删除属于用户的笔记本服务器 pod。
  b. 询问用户再次启动其笔记本服务器。
- 如果笔记本服务器没有足够的资源来运行所选笔记本服务器镜像，请在 OpenShift Container Platform 集群中添加更多资源，或者选择较小的镜像大小。
- 如果 Jupyter pod 处于 FAILED 状态：

a. 检索 jupyter-nb114 pod 的日志，并将其发送到红帽支持以进一步评估。

b. 删除 jupyter-nb-* pod。

- 如果没有以前的解决方案，请联系红帽支持。

2.6.3. 用户运行笔记本的 cells 时遇到 database or disk is full 错误或 no space left on device 错误。

问题

用户可能已在其笔记本服务器上耗尽存储空间。

诊断

1. 登录到 Jupyter，并启动属于用户问题的笔记本服务器。如果笔记本服务器没有启动，请按照以下步骤检查用户是否已耗尽存储空间：

   a. 登陆到 OpenShift Container Platform Web 控制台。

   b. 点 Workloads → Pods，将项目设置为 rhods-notebooks。

   c. 点属于该用户的笔记本服务器 pod，例如 jupyter-nb-<idp>-<username>-*。

   d. 点 Logs。如果您看到类似如下的行，用户已超过其可用容量：

   > Unexpected error while saving file: XXXX database or disk is full

解决方案

- 通过扩展其持久性卷来增加用户可用的存储：扩展持久性卷
- 与用户合作找出可以从 /opt/app-root/src 目录中删除的文件，以释放其现有存储空间。

注意

当您使用 JupyterLab 文件探索器删除文件时，文件将移到笔记本的持久性存储中的隐藏 /opt/app-root/src/.local/share/Trash/files 文件夹。要为笔记本释放存储空间，您必须永久删除这些文件。

第 3 章 管理集群资源

3.1. 为集群配置默认 PVC 大小

要配置如何在 OpenShift AI 集群中声明资源，您可以更改集群持久性卷声明(PVC)的默认大小，确保请求的存储与常见存储工作流匹配。PVC 是对集群中的资源请求，还可作为对资源的声明检查。

先决条件

- **您已登陆到 Red Hat OpenShift AI。**

注意

更改 PVC 设置会重启 Jupyter pod，并导致 Jupyter 最多 30 秒。作为临时解决方案，建议您在组织的典型工作日外执行该操作。

流程

1. **在 OpenShift AI 仪表板中点 Settings → Cluster settings。**
2. **在 PVC 大小 下，以 KB 为单位输入新大小。最小值为 1 GiB，最大大小为 16384 GiB。**
3. **点 Save Changes。**

验证

- **使用您配置的默认存储大小创建新的 PVC。**

其他资源

- **了解持久性存储**

3.2. 为集群恢复默认 PVC 大小

要更改 OpenShift AI 集群中使用的资源大小，您可以恢复集群持久性卷声明(PVC)的默认大小。

先决条件

- **您已登陆到 Red Hat OpenShift AI。**
- **您是 OpenShift Container Platform 中的 OpenShift AI 管理员组的一部分。**

流程

1. **在 OpenShift AI 仪表板中点 Settings → Cluster settings。**
2. **点恢复默认恢复默认 PVC 大小为 20GiB。**
3. **点 Save Changes。**

验证

- **创建新的 PVC，其默认存储大小为 20 GiB。**

其他资源

- 了解持久性存储

3.3. 加速器概述

如果使用大型数据集，您可以使用加速器来优化 OpenShift AI 中数据科学模型的性能。通过加速器，您可以扩展工作、缩短延迟并提高生产力。您可以在 OpenShift AI 中使用加速器来协助数据科学家在以下任务中：

- 自然语言处理(NLP)
- Inference
- 培训深层网络
- 数据清理和数据处理

OpenShift AI 支持以下加速器：

- NVIDIA 图形处理单元(GPU)
  - 要在模型中使用计算密集型工作负载，您可以在 OpenShift AI 中启用 NVIDIA 图形处理单元(GPU)。
  - 要在 OpenShift 中启用 GPU，您必须安装 NVIDIA GPU Operator。
- Habana Gaudi 设备(HPU)
  - Habana 是 Intel 公司，提供用于深入学习工作负载的硬件加速器。您可以使用与笔记本中提供的 Habana Gaudi 设备关联的 Habana 库和软件。
  - 在 OpenShift AI 中启用 Habana Gaudi 设备前，您必须先安装必要的依赖项和 HabanaAI Operator 版本，该版本与部署中的 HabanaAI 工作台镜像匹配。有关如何为 Habana Gaudi 设备启用 OpenShift 环境的更多信息，请参阅 OpenShift 的 HabanaAI Operator v1.10 和 HabanaAI Operator v1.13。
  - 您可以在内部或使用 AWS 实例中的 AWS DL1 计算节点启用 Habana Gaudi 设备。

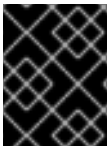
在 OpenShift AI 中使用加速器前，您的 OpenShift 实例必须包含关联的加速器配置文件。对于部署到部署的加速器，您必须为上下文中的加速器配置加速器配置集。您可以从 OpenShift AI 仪表板上的 Settings → Accelerator profile 页面创建加速器配置集。如果您的部署包含已经配置了关联加速器配置集的现有加速器，则在升级到最新版本的 OpenShift AI 后会自动创建加速器配置集。

其他资源

- HabanaAI Operator v1.10 for OpenShift
- 用于 OpenShift 的 HabanaAI Operator v1.13
- Habana，Intel 公司
- Amazon EC2 DL1 实例
- lspci(8) - Linux man page

3.3.1. 在 OpenShift AI 中启用 GPU 支持

另外，为了确保数据科学家可以在其模型中使用计算密集型工作负载，您可以在 OpenShift AI 中启用图形处理单元(GPU)。



重要

如果您在断开连接的自管理的环境中使用 OpenShift AI，请参阅在 OpenShift AI 中启用 GPU 支持。

先决条件

- 已登陆到 OpenShift Container Platform 集群。
- 在 OpenShift Container Platform 集群中具有 cluster-admin 角色。

流程

1. 要在 OpenShift 集群上启用 GPU 支持，请按照 NVIDIA 文档中的 Red Hat OpenShift Container Platform 上的 NVIDIA GPU Operator 的说明进行操作。
2. 删除 migration-gpu-status ConfigMap。
   a. 在 OpenShift Container Platform web 控制台中切换到 Administrator 视角。
   b. 将项目设置为 All Projects 或 redhat-ods-applications，以确保您可以看到适当的 ConfigMap。
   c. 搜索 migration-gpu-status ConfigMap。
   d. 点操作菜单 (⋮)，并从列表中选择 Delete ConfigMap。
   此时会出现 Delete ConfigMap 对话框。
   e. 检查对话框，并确认您删除正确的 ConfigMap。
   f. 点击 Delete。
3. 重启仪表板 replicaset。
   a. 在 OpenShift Container Platform web 控制台中切换到 Administrator 视角。
   b. 点 Workloads → Deployments。
   c. 将项目设置为 All Projects 或 redhat-ods-applications，以确保您可以看到适当的部署。
   d. 搜索 rhods-dashboard 部署。
   e. 点操作菜单(HBAC)，然后从列表中选择 Restart Rollout。
   f. 等待 Status 列指出 rollout 中的所有 pod 都完全重启。

验证

- NVIDIA GPU Operator 会出现在 OpenShift Container Platform Web 控制台的 Operators → Installed Operators 页面中。

- 重置 migration-gpu-status 实例存在于 AcceleratorProfile 自定义资源定义 (CRD) 详情页面上的 Instances 选项卡中。

安装 NVIDIA GPU Operator 后，创建一个加速器配置集，如 使用加速器配置集 中所述。

3.3.2. 启用 Habana Gaudi 设备

在 OpenShift AI 中使用 Habana Gaudi 设备前，您必须安装必要的依赖项并部署 HabanaAI Operator。

先决条件

- 已登陆到 OpenShift Container Platform。
- 在 OpenShift Container Platform 中具有 cluster-admin 角色。

流程

1. 要在 OpenShift AI 中启用 Habana Gaudi 设备，请按照 OpenShift 的 HabanaAI Operator 中的说明操作。
2. 在 OpenShift AI 仪表板中，点 Settings → Accelerator profiles。
   此时会出现 加速器配置文件 页面，显示现有的加速器配置文件。要启用或禁用现有的加速器配置集，请在包含相关加速器配置集的行中点 Enable 列中的切换。
3. 点 Create accelerator profile。
   Create accelerator 配置集 对话框将打开。
4. 在 Name 字段中输入 Habana Gaudi 设备的名称。
5. 在 Identifier 字段中输入唯一字符串，用于标识 Habana Gaudi 设备，例如habana.ai/gaudi。
6. 可选：在 Description 字段中输入 Habana Gaudi 设备的描述。
7. 要在创建后立即为 Habana Gaudi 设备启用或禁用加速器配置文件，请点击 Enable 列中的切换。
8. 可选：添加容限来调度具有匹配污点的 pod。
   a. 点 Add toleration。
      此时会打开 Add toleration 对话框。
   b. 在 Operator 列表中，选择以下选项之一：
      - equal - 键/值/effect 参数必须匹配。这是默认值。
      - exists - key/effect 参数必须匹配。您必须保留一个空 value 参数，该参数与 any 匹配。
   c. 在 Effect 列表中，选择以下选项之一：
      - None
      - NoSchedule - 与污点不匹配的新 pod 不会调度到该节点上。该节点上现有的 pod 会保留。
      - PreferNoSchedule - 与污点不匹配的新 pod 可能会调度到该节点上，但调度程序会尝试。该节点上现有的 pod 会保留。

- NoExecute - 与污点不匹配的新 pod 无法调度到该节点上。节点上没有匹配容限的现有 pod 将被移除。

d. 在 Key 字段中，输入 toleration 键**habana.ai/gaudi**。key 是任意字符串，最多 253 个字符。key 必须以字母或数字开头，可以包含字母、数字、连字符、句点和下划线。

e. 在 Value 字段中输入容限值。该值是任意字符串，最多 63 个字符。value 必须以字母或数字开头，可以包含字母、数字、连字符、句点和下划线。

f. 在 Toleration Seconds 部分中，选择以下选项之一来指定 pod 保持与具有节点状况的节点绑定的时长。

- 永久地 - Pod 保持永久绑定到节点。
- Custom 值 - 输入值（以秒为单位），以定义 pod 保持与具有节点状况的节点绑定的时长。

g. 点击 Add。

9. 点 Create accelerator profile。

验证

- 从 Administrator 视角中，以下 Operator 会出现在 Operators → Installed Operators 页面中。
  - HabanaAI
  - 节点功能发现(NFD)
  - 内核模块管理(KMM)
- 加速器 列表 在启动笔记本服务器页面 中显示 Habana Gaudi 加速器。选择加速器后，会出现 Number of accelerators 字段，您可以使用它来选择笔记本服务器的加速器数量。
- Accelerator 配置集会出现在 Accelerator 配置集 页面中
- 加速器配置集会出现在 **AcceleratorProfile** 自定义资源定义(CRD)的详细信息页面上。

其他资源

- 用于 OpenShift 的 HabanaAI Operator v1.10。
- 用于 OpenShift 的 HabanaAI Operator v1.13。

3.4. 为 OPENSHIFT AI 用户分配其他资源

作为集群管理员，您可以将其他资源分配给集群以支持计算密集型数据科学工作。此支持包括增加集群中的节点数量并更改集群分配的机器池。

有关为 OpenShift Container Platform 集群分配其他资源的更多信息，请参阅 手动扩展计算机器集。

第 4 章 管理 JUPYTER 笔记本服务器

4.1. 访问 JUPYTER 管理界面

您可以使用 Jupyter 管理界面控制 Red Hat OpenShift AI 环境中的笔记本服务器。

前提条件

- **您是 OpenShift Container Platform 管理员组的一部分。如需更多信息，请参阅为 OpenShift Container Platform 添加管理用户。**

流程

- **要从 OpenShift AI 访问 Jupyter 管理界面，请执行以下操作：**
  i. **在 OpenShift AI 中，在左侧菜单的 Applications 部分，单击 Enabled。**
  ii. **找到 Jupyter 标题并点 Launch application。**
  iii. **在启动 Jupyter 时打开的页面中，点 Administration 选项卡。Administration 页面将打开。**
- **要从 JupyterLab 访问 Jupyter 管理界面，请执行以下操作：**
  i. **点 File → Hub Control Panel。**
  ii. **在 OpenShift AI 中打开的页面中，点 Administration 选项卡。Administration 页面将打开。**

验证

- **您可以看到 Jupyter 管理界面。**

4.2. 启动由其他用户拥有的笔记本服务器

管理员可以从 Jupyter 管理界面为另一现有用户启动笔记本服务器。

先决条件

- **您是 OpenShift Container Platform 管理员组的一部分。如需更多信息，请参阅为 OpenShift Container Platform 添加管理用户。**
- **您已启动 Jupyter 应用程序，如 启动 Jupyter 笔记本服务器 中所述。**

流程

1. **在启动 Jupyter 时打开的页面中，点 Administration 选项卡。**
2. **在 Administration 选项卡中，执行以下操作：**
   a. **在 Users 部分中，找到您要启动其笔记本服务器的用户。**
   b. **点相关用户旁的 Start server。**

c. 完成 Start abook server页面。

d. 可选： 如果需要，选择 Start server in current tab。

e. 单击 Start server。
服务器启动后，您会看到以下行为之一：

- 如果您之前选择了 Start server in current tab，则 JupyterLab 界面会在 Web 浏览器的当前标签页中打开。
- 如果您之前没有选中 Start server in current tab，则启动服务器对话框会提示您在新浏览器标签页或当前标签页中打开服务器。
JupyterLab 接口根据您的选择打开。

验证

- 此时会打开 JupyterLab 接口。

4.3. 访问其他用户拥有的笔记本服务器

管理员可以访问由其他用户拥有的笔记本服务器，以更正配置错误或帮助他们对其环境进行故障排除。

先决条件

- 您是 OpenShift Container Platform 管理员组的一部分。如需更多信息，请参阅为 OpenShift Container Platform 添加管理用户。
- 您已启动 Jupyter 应用程序，如 启动 Jupyter 笔记本服务器 中所述。
- 您要访问的笔记本服务器正在运行。

流程

1. 在启动 Jupyter 时打开的页面中，点 Administration 选项卡。
2. 在 Administration 页面中，执行以下操作：
   a. 在 Users 部分中，找到笔记本服务器所属的用户。
   b. 点相关用户旁的 View server。
   c. 在笔记本服务器控制面板页面中，点 Access notebook server。

验证

- 用户的笔记本服务器在 JupyterLab 中打开。

4.4. 停止其他用户拥有笔记本服务器

管理员可以停止由其他用户拥有的笔记本服务器，以降低集群上的资源消耗，或作为从集群删除用户及其资源的一部分。

先决条件

- 如果您使用专用的 OpenShift AI 组，则作为管理员组的一部分（例如 rhoai-admins）。如果不使用专用组，则作为 OpenShift Container Platform 管理员组的一部分。如需更多信息，请参阅为 OpenShift Container Platform 添加管理用户。
- 您已启动 Jupyter 应用程序，如 启动 Jupyter 笔记本服务器 中所述。
- 要停止的笔记本服务器正在运行。

流程

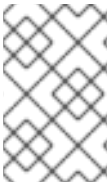
1. 在启动 Jupyter 时打开的页面中，点 Administration 选项卡。
2. 停止一个或多个服务器。
   - 如果要停止一个或多个特定服务器，请执行以下操作：
     i. 在 Users 部分中，找到笔记本服务器所属的用户。
     ii. 要停止笔记本服务器，请执行以下操作之一：
       - 点相关用户旁边的操作菜单(WWN)，然后选择 Stop server。
       - 点相关用户旁边的 View server，然后点 Stop notebook server。此时会出现 Stop server 对话框。
     iii. 点 Stop server。
   - 如果要停止所有服务器，请执行以下操作：
     i. 点 Stop all servers 按钮。
     ii. 点 OK 以确认停止所有服务器。

验证

- 当笔记本服务器停止后，每个服务器都会将 Stop server 链接更改为 Start server 链接。

4.5. 停止闲置的 NOTEBOOK

您可以通过停止闲置（无需登录用户）的笔记本服务器来减少 OpenShift AI 部署中的资源使用量。当集群中资源需求很高时，这非常有用。默认情况下，闲置 Notebook 不会在特定时间限制后停止。



注意

如果您将集群设置配置为在指定时间限制后从集群断开所有用户，那么这个设置优先于闲置 Notebook 时间限制。当用户的会话持续时间达到集群范围时间限制时，会在集群注销。

先决条件

- 您已登陆到 Red Hat OpenShift AI。
- 您是 OpenShift Container Platform 中的 OpenShift AI 管理员组的一部分。

流程

1. 在 OpenShift AI 仪表板中点 Settings → Cluster settings。
2. 在 Stop idle notebooks 下，选择 Stop idle notebooks after。
3. 以小时和分钟为单位输入时间限制，用于指定空闲 notebooks 被停止的时间。
4. 点 Save Changes。

验证

- notebook-controller-culler-config ConfigMap 位于 Workloads → ConfigMaps 页面中的 redhat-ods-applications 项目中，包含以下 culling 配置设置：
  - ENABLE_CULLING ：指定是否启用或禁用 culling 功能（默认为 false ）。
  - IDLENESS_CHECK_PERIOD ：轮询频率，以检查笔记本的最后已知活动（以分钟为单位）。
  - CULL_IDLE_TIME ：将不活跃笔记本扩展为零（以分钟为单位）的最大分配时间。
- 闲置笔记本会在您设置的时间限制停止。

4.6. 配置自定义笔记本镜像

除了由红帽和独立软件供应商(ISV)提供和支持的笔记本镜像外，您还可以配置针对项目特定要求的自定义笔记本镜像。

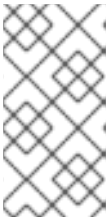
红帽支持您在 OpenShift AI 部署中添加自定义笔记本镜像，并确保它们可用于在创建笔记本服务器时进行选择。但是，红帽不支持您的自定义笔记本镜像的内容。也就是说，如果您的自定义笔记本镜像可在笔记本服务器创建过程中选择，但没有创建可用的笔记本服务器，红帽不提供支持修复您的自定义笔记本镜像。

先决条件

- 您已登陆到 Red Hat OpenShift AI。
- 在 OpenShift Container Platform 中分配了 cluster-admin 角色。
- 您的自定义笔记本镜像存在于镜像 registry 中，可访问。
- 您可以访问 Settings → Notebook images 仪表板导航菜单选项。

流程

1. 在 OpenShift AI 仪表板中，点 Settings → Notebook images。
   此时会出现 Notebook images 页面。以前导入的 notebook 镜像会被显示。要启用或禁用之前导入的 notebook 镜像，请在包含相关 notebook 镜像的行中点 Enable 列中的切换。



注意

如果您已经为笔记本镜像配置了加速器标识符，您可以通过创建关联的加速器配置集为笔记本镜像指定推荐的加速器。要做到这一点，请点击包含 notebook 镜像所在行上的 Create profile，并完成相关字段。如果笔记本镜像不包含加速器标识符，您必须在创建关联的加速器配置集前手动配置。

2. 点 Import new image。或者，如果未找到之前导入的镜像，请点 Import image。此时会出现 Import Notebook 镜像 对话框。
3. 在 Image location 字段中，输入包含 notebook 镜像的存储库的 URL。例如：quay.io/my-repo/my-image:tag,quay.io/my-repo/my-image@sha256:xxxxxxxxxxxxx, 或 docker.io/my-repo/my-image:tag。
4. 在 Name 字段中，为 notebook 输入相应的名称。
5. 可选：在 Description 字段中输入 notebook 镜像的描述。
6. 可选：在 加速器标识符 列表中，选择一个标识符，用于推荐使用笔记本镜像设置其加速器。如果笔记本镜像只包含一个加速器标识符，则默认显示标识符名称。
7. 可选：在 notebook 镜像中添加软件。导入完成后，软件将添加到 notebook 镜像的 meta-data 中，并显示在 Jupyter 服务器创建页面中。
   a. 点 Software 选项卡。
   b. 点添加软件按钮。
   c. 点 Edit ( )。
   d. 输入软件名称。
   e. 输入软件版本。
   f. 点 Confirm ( )确认您的条目。
   g. 要添加其他软件，请点 Add software，填写相关字段并确认您的条目。
8. 可选：在 notebook 镜像中添加软件包。导入完成后，软件包将添加到 notebook 镜像的 meta-data 中，并显示在 Jupyter 服务器创建页面中。
   a. 点 Packages 选项卡。
   b. 点 Add package 按钮。
   c. 点 Edit ( )。
   d. 输入软件包名称。
   e. 输入软件包版本。
   f. 点 Confirm ( )确认您的条目。
   g. 要添加附加软件包，请点 Add package，完成相关字段并确认您的条目。
9. 点 Import。

验证

- 您导入的笔记本镜像会在 Notebook 镜像页面的表中显示。
- 在 Jupyter 的 启动一个笔记本服务器 页面中可选择您的自定义笔记本镜像。

其他资源

- 管理镜像流
- 了解构建配置

第 5 章 备份数据

5.1. 备份存储数据

最佳实践是定期备份持久性卷声明(PVC)中的数据。

在删除用户前和卸载 OpenShift AI 之前备份您的数据非常重要，因为在卸载 OpenShift AI 时，所有 PVC 都会被删除。

有关备份 PVC 的更多信息，请参阅集群平台的文档。

其他资源

- **了解持久性存储**

第 6 章 使用数据收集

Red Hat OpenShift AI 管理员可以选择是否允许红帽收集有关集群中 OpenShift AI 使用情况的数据。收集这些数据可让红帽监控并改进我们的软件和支持。有关红帽收集的数据的详细信息，请参阅 OpenShift AI 的使用数据收集通知。

当您在 OpenShift Container Platform 集群上安装 OpenShift AI 时，使用数据收集功能会被默认启用，除非在断开连接的环境中安装集群。

有关在集群中禁用此数据收集的说明，请参阅禁用使用数据收集。如果您在集群中禁用了数据收集，并且希望再次启用它，请参阅 启用使用数据收集。

6.1. OPENSHIFT AI 的使用数据收集通告

在您使用此红帽产品时，红帽可能会收集您使用软件的使用数据。这些数据可让红帽监控软件并改进红帽产品和支持，包括识别、故障排除和响应影响用户的问题。

红帽收集哪些信息？

软件中的工具监控各种指标，这些信息会转移到红帽。指标包括如下信息：

- 有关产品仪表板中启用的应用程序的信息。
- 使用的部署大小（即分配的 CPU 和内存资源）。
- 有关从产品仪表板访问的文档资源的信息。
- 使用的笔记本镜像的名称（即 Minimal Python、Standard Data Science 和其他镜像）。
- 在初始用户登录过程中生成的随机识别符，用于将数据与特定用户名相关联。
- 有关组件、功能和扩展的使用情况信息。

第三方服务提供商

红帽使用某些第三方服务提供商收集遥测数据。

安全性

红帽采用技术和组织措施来保护使用数据。

个人数据

红帽公司不会收集个人信息。如果红帽发现个人信息被意外地收到，红帽将根据红帽的隐私声明删除此类个人信息并处理此类个人信息。有关红帽隐私实践的更多信息，请参阅 红帽隐私声明。

启用和禁用使用数据

您可以按照禁用使用 数据收集或 启用使用数据收集 中的说明禁用或启用使用数据。

6.2. 启用使用数据收集

Red Hat OpenShift AI 管理员可以选择是否允许红帽收集有关集群中 OpenShift AI 使用情况的数据。当您在 OpenShift Container Platform 集群上安装 OpenShift AI 时，使用数据收集功能会被默认启用，除非在断开连接的环境中安装集群。如果您之前禁用了数据收集，您可以按照下列步骤重新启用它。

先决条件

- 您已登陆到 Red Hat OpenShift AI。

- 您是 OpenShift Container Platform 中的 OpenShift AI 管理员组的一部分，除非在断开连接的环境中安装集群。

流程

1. 在 OpenShift AI 仪表板中点 Settings → Cluster settings。
2. 找到使用数据收集部分。
3. 选择 Allow collection usage data复选框。
4. 点 Save Changes。

验证

- 更新设置时会显示通知：Settings changes saved.

其他资源

- OpenShift AI 的使用数据收集通告

6.3. 禁用使用数据收集

Red Hat OpenShift AI 管理员可以选择是否允许红帽收集有关集群中 OpenShift AI 使用情况的数据。当您在 OpenShift Container Platform 集群上安装 OpenShift AI 时，使用数据收集功能会被默认启用，除非在断开连接的环境中安装集群。

您可以按照以下步骤禁用数据收集。

先决条件

- 您已登陆到 Red Hat OpenShift AI。
- 您是 OpenShift Container Platform 中的 OpenShift AI 管理员组的一部分，除非在断开连接的环境中安装集群。

流程

1. 在 OpenShift AI 仪表板中点 Settings → Cluster settings。
2. 找到使用数据收集部分。
3. 取消选中 Allow collection usage data复选框。
4. 点 Save Changes。

验证

- 更新设置时会显示通知：Settings changes saved.

其他资源

- OpenShift AI 的使用数据收集通告