Red Hat OpenShift AI Cloud Service 1

Release notes

Features, enhancements, resolved issues, and known issues associated with this release

Red Hat OpenShift AI Cloud Service 1 Release notes

Features, enhancements, resolved issues, and known issues associated with this release

Legal Notice

Abstract

These release notes provide an overview of new features, enhancements, resolved issues, and known issues in this release of Red Hat OpenShift AI. OpenShift AI is currently available in Red Hat OpenShift Dedicated and Red Hat OpenShift Service on Amazon Web Services (ROSA).

Table of Contents

CHAPTER 1. OVERVIEW OF OPENSHIFT AI

Red Hat OpenShift AI is a platform for data scientists and developers of artificial intelligence and machine learning (AI/ML) applications.

OpenShift AI provides an environment to develop, train, serve, test, and monitor AI/ML models and applications on-premises or in the cloud.

For data scientists, OpenShift AI includes Jupyter and a collection of default notebook images optimized with the tools and libraries required for model development, and the TensorFlow and PyTorch frameworks. Deploy and host your models, integrate models into external applications, and export models to host them in any hybrid cloud environment. You can enhance your data science projects on OpenShift AI by building portable machine learning (ML) workflows with data science pipelines, using Docker containers. You can also accelerate your data science experiments through the use of graphics processing units (GPUs) and Habana Gaudi devices.

For administrators, OpenShift AI enables data science workloads in an existing Red Hat OpenShift or ROSA environment. Manage users with your existing OpenShift identity provider, and manage the resources available to notebook servers to ensure data scientists have what they require to create, train, and host models. Use accelerators to reduce costs and allow your data scientists to enhance the performance of their end-to-end data science workflows using graphics processing units (GPUs) and Habana Gaudi devices.

OpenShift AI offers two distributions:

- A **managed cloud service add-on** for Red Hat OpenShift Dedicated (with a Customer Cloud Subscription for AWS or GCP) or for Red Hat OpenShift Service on Amazon Web Services (ROSA).
  For information about OpenShift AI on a Red Hat managed environment, see Product Documentation for Red Hat OpenShift AI.

- **Self-managed software** that you can install on-premise or on the public cloud in a self-managed environment, such as OpenShift Container Platform.
  For information about 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.

For information about OpenShift AI supported software platforms, components, and dependencies, see Supported configurations.

CHAPTER 2. NEW FEATURES AND ENHANCEMENTS

This section describes new features and enhancements in Red Hat OpenShift AI.

2.1. NEW FEATURES

Distributed workloads

Distributed workloads enable data scientists to use multiple cluster nodes in parallel for faster, more efficient data processing and model training. The CodeFlare framework simplifies task orchestration and monitoring, and offers seamless integration for automated resource scaling and optimal node utilization with advanced GPU support.

Designed for data scientists, the CodeFlare framework enables direct workload configuration from Jupyter Notebooks or Python code, ensuring a low barrier of adoption, and streamlined, uninterrupted workflows. Distributed workloads significantly reduce task completion time, and enable the use of larger datasets and more complex models.

Authorization provider for single-model serving platform

You can now add Authorino as an authorization provider for the single-model serving (KServe) platform. Adding an authorization provider allows you to enable token authorization for models that you deploy on the platform, which ensures that only authorized parties can make inference requests to the models.

2.2. ENHANCEMENTS

Improved Data Science Projects user interface

The Data Science Projects user interface (UI) has been redesigned to simplify accessing and getting started with different project components. The new design includes an updated layout, a more visually-oriented interface, and additional UI text to provide an overview of each project component.

Support for Kubeflow Pipelines v2 in data science pipelines

To keep OpenShift AI updated with the latest features, data science pipelines are now based on KubeFlow Pipelines (KFP) version 2.0 . Data Science Pipelines (DSP) 2.0 is enabled and deployed by default in OpenShift AI. For more information, see Enabling Data Science Pipelines 2.0 .
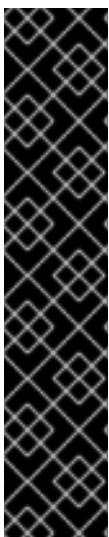
> **IMPORTANT**
>
> Previously, data science pipelines in OpenShift AI were based on KubeFlow Pipelines v1. It is no longer possible to deploy, view, or edit the details of pipelines that are based on DSP 1.0 from the dashboard in OpenShift AI.
>
> DSP 2.0 contains an installation of Argo Workflows. OpenShift AI does not support direct customer usage of this installation of Argo Workflows. To install or upgrade to OpenShift AI with DSP 2.0, ensure that there is no existing installation of Argo Workflows on your cluster.
>
> If you want to use existing pipelines and workbenches with DSP 2.0 after upgrading OpenShift AI, you must update your workbenches to use the 2024.1 notebook image version and then manually migrate your pipelines from DSP 1.0 to 2.0. For more information, see Upgrading to DSP 2.0 .

Updated workbench images

Preinstalled packages for workbench images have been updated with the 2024.1 image versions. You

can optimize your development environment by using the latest workbench image versions. Python packages used in workbench images include advancements in the Python ecosystem, such as PyTorch and TensorFlow. Operating systems supporting the code-server, RStudio, Elyra, Habana, and CUDA workbench images have also received updates for their tools.

CHAPTER 3. TECHNOLOGY PREVIEW FEATURES
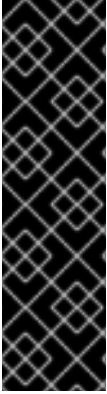
> **IMPORTANT**
>
> This section describes Technology Preview features in Red Hat OpenShift AI. Technology Preview features are not supported with Red Hat production service level agreements (SLAs) and might not be functionally complete. Red Hat does not recommend using them in production. These features 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 .

RStudio Server notebook image

With the **RStudio Server** notebook image, you can access the RStudio IDE, an integrated development environment for R. The R programming language is used for statistical computing and graphics to support data analysis and predictions.

To use the **RStudio Server** notebook image, you must first build it by creating a secret and triggering the **BuildConfig**, and then enable it in the OpenShift AI UI by editing the **rstudio-rhel9** image stream. See Building the RStudio Server notebook images for more information.
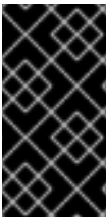
> **IMPORTANT**
>
> **Disclaimer:** Red Hat supports managing workbenches in OpenShift AI. However, Red Hat does not provide support for the RStudio software. RStudio Server is available through rstudio.org and is subject to their licensing terms. You should review their licensing terms before you use this sample workbench.

CUDA – RStudio Server notebook image

With the **CUDA – RStudio Server** notebook image, you can access the RStudio IDE and NVIDIA CUDA Toolkit. The RStudio IDE is an integrated development environment for the R programming language for statistical computing and graphics. With the NVIDIA CUDA toolkit, you can enhance your work by using GPU-accelerated libraries and optimization tools.

To use the **CUDA – RStudio Server** notebook image, you must first build it by creating a secret and triggering the **BuildConfig**, and then enable it in the OpenShift AI UI by editing the **rstudio-rhel9** image stream. See Building the RStudio Server notebook images for more information.

> **IMPORTANT**
>
> **Disclaimer:** Red Hat supports managing workbenches in OpenShift AI. However, Red Hat does not provide support for the RStudio software. RStudio Server is available through rstudio.org and is subject to their licensing terms. You should review their licensing terms before you use this sample workbench.
>
> The **CUDA – RStudio Server** notebook image contains NVIDIA CUDA technology. CUDA licensing information is available in the CUDA Toolkit documentation. You should review their licensing terms before you use this sample workbench.

code-server workbench image

Red Hat OpenShift AI now includes the **code-server** workbench image. See code-server in GitHub for more information.

With the **code-server** workbench image, you can customize your workbench environment by using a variety of extensions to add new languages, themes, debuggers, and connect to additional services. You can also enhance the efficiency of your data science work with syntax highlighting, auto-indentation, and bracket matching.
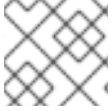
> **NOTE**
>
> Elyra-based pipelines are not available with the **code-server** workbench image.

The **code-server** workbench image is currently available in Red Hat OpenShift AI as a Technology Preview feature.

CHAPTER 4. SUPPORT REMOVALS

This section describes major changes in support for user-facing features in Red Hat OpenShift AI. For information about OpenShift AI supported software platforms, components, and dependencies, see Supported configurations.

4.1. DATA SCIENCE PIPELINES V1 UPGRADED TO V2

Previously, data science pipelines in OpenShift AI were based on KubeFlow Pipelines v1. Data science pipelines are now based on KubeFlow Pipelines v2, which uses a different workflow engine. Data Science Pipelines (DSP) 2.0 is enabled and deployed by default in OpenShift AI. It is no longer possible to deploy, view, or edit the details of pipelines that are based on DSP 1.0 from the dashboard. For more information, see Enabling Data Science Pipelines 2.0 .
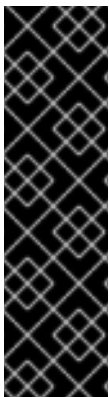
> **IMPORTANT**
>
> DSP 2.0 contains an installation of Argo Workflows. OpenShift AI does not support direct customer usage of this installation of Argo Workflows. To install or upgrade to OpenShift AI with DSP 2.0, ensure that there is no existing installation of Argo Workflows on your cluster.
>
> If you want to use existing pipelines and workbenches with DSP 2.0 after upgrading OpenShift AI, you must update your workbenches to use the 2024.1 notebook image version and then manually migrate your pipelines from DSP 1.0 to 2.0. For more information, see Upgrading to DSP 2.0 .

4.2. REMOVAL OF BIAS DETECTION (TRUSTYAI)

Starting with OpenShift AI 2.7, the bias detection (TrustyAI) functionality has been removed. If you previously had this functionality enabled, upgrading to OpenShift AI 2.7 or later will remove the feature. The default TrustyAI notebook image remains supported.

4.3. VERSION 1.2 NOTEBOOK CONTAINER IMAGES FOR WORKBENCHES ARE NO LONGER SUPPORTED

When you create a workbench, you specify a notebook container image to use with the workbench. Starting with OpenShift AI 2.5, when you create a new workbench, version 1.2 notebook container images are not available to select. Workbenches that are already running with a version 1.2 notebook image continue to work normally. However, Red Hat recommends that you update your workbench to use the latest notebook container image.

4.4. NVIDIA GPU OPERATOR REPLACES NVIDIA GPU ADD-ON

Previously, to enable graphics processing units (GPUs) to help with compute-heavy workloads, you installed the NVIDIA GPU add-on. OpenShift AI no longer supports this add-on.

Now, to enable GPU support, you must install the NVIDIA GPU Operator. To learn how to install the GPU Operator, see NVIDIA GPU Operator on Red Hat OpenShift Container Platform (external).

4.5. KUBEFLOW NOTEBOOK CONTROLLER REPLACES JUPYTERHUB

In OpenShift AI 1.15 and earlier, JupyterHub was used to create and launch notebook server environments. In OpenShift AI 1.16 and later, JupyterHub is no longer included, and its functionality is replaced by Kubeflow Notebook Controller.

This change provides the following benefits:

- Users can now immediately cancel a request, make changes, and retry the request, instead of waiting 5+ minutes for the initial request to time out. This means that users do not wait as long when requests fail, for example, when a notebook server does not start correctly.

- The architecture no longer prevents a single user from having more than one notebook server session, expanding future feature possibilities.

- The removal of the PostgreSQL database requirement allows for future expanded environment support in OpenShift AI.

However, this update also creates the following behavior changes:

- For IT Operations administrators, the notebook server administration interface does not currently allow login access to data scientist users' notebook servers. This is planned to be added in future releases.

- For data scientists, the JupyterHub interface URL is no longer valid. Update your bookmarks to point to the OpenShift AI Dashboard.

The JupyterLab interface is unchanged and data scientists can continue to use JupyterLab to work with their notebook files as usual.

CHAPTER 5. RESOLVED ISSUES

The following notable issues are resolved in Red Hat OpenShift AI.

RHOAIENG-5067 – Model server metrics page does not load for a model server based on the ModelMesh component

Previously, data science project names that contained capital letters or spaces could cause issues on the model server metrics page for model servers based on the ModelMesh component. The metrics page might not have received data correctly, resulting in a **400 Bad Request** error and preventing the page from loading. This issue is now resolved.

RHOAIENG-4966 – Self-signed certificates in a custom CA bundle might be missing from the **odh-trusted-ca-bundle** configuration map

Previously, if you added a custom certificate authority (CA) bundle to use self-signed certificates, sometimes the custom certificates were missing from the **odh-trusted-ca-bundle** ConfigMap, or the non-reserved namespaces did not contain the **odh-trusted-ca-bundle** ConfigMap when the ConfigMap was set to **managed**. This issue is now resolved.

RHOAIENG-4938 (previously documented as **RHOAIENG-4327**) – Workbenches do not use the self-signed certificates from centrally configured bundle automatically

There are two bundle options to include self-signed certificates in OpenShift AI, **ca-bundle.crt** and **odh-ca-bundle.crt**. Previously, workbenches did not automatically use the self-signed certificates from the centrally configured bundle and you had to define environment variables that pointed to your certificate path. This issue is now resolved.

RHOAIENG-4572 – Unable to run data science pipelines after install and upgrade in certain circumstances

Previously, you were unable to run data science pipelines after installing or upgrading OpenShift AI in the following circumstances:

- You installed OpenShift AI and you had a valid CA certificate. Within the **default-dsci** object, you changed the **managementState** field for the **trustedCABundle** field to **Removed** post-installation.

- You upgraded OpenShift AI from version 2.6 to version 2.8 and you had a valid CA certificate.

- You upgraded OpenShift AI from version 2.7 to version 2.8 and you had a valid CA certificate.

This issue is now resolved.

RHOAIENG-4524 – BuildConfig definitions for RStudio images contain occurrences of incorrect branch

Previously, the **BuildConfig** definitions for the **RStudio** and **CUDA – RStudio** workbench images pointed to the wrong branch in OpenShift AI. This issue is now resolved.

RHOAIENG-3963 – Unnecessary managed resource warning

Previously, when you edited and saved the **OdhDashboardConfig** custom resource for the **redhat-ods-applications** project, the system incorrectly displayed a **Managed resource** warning message. This issue is now resolved.

RHOAIENG-2542 – Inference service pod does not always get an Istio sidecar

Previously, when you deployed a model using the single-model serving platform (which uses KServe), the **istio-proxy** container could be missing in the resulting pod, even if the inference service had the **sidecar.istio.io/inject=true** annotation. This issue is now resolved.

RHOAIENG-1666 – The Import Pipeline button is prematurely accessible

Previously, when you imported a pipeline to a workbench that belonged to a data science project, the **Import Pipeline** button was accessible before the pipeline server was fully available. This issue is now resolved.

RHOAIENG-673 (previously documented as RHODS-12946) – Cannot install from PyPI mirror in disconnected environment or when using private certificates

In disconnected environments, Red Hat OpenShift AI cannot connect to the public-facing PyPI repositories, so you must specify a repository inside your network. Previously, if you were using private TLS certificates and a data science pipeline was configured to install Python packages, the pipeline run would fail. This issue is now resolved.

RHOAIENG-3355 – OVMS on KServe does not use accelerators correctly

Previously, when you deployed a model using the single-model serving platform and selected the **OpenVINO Model Server** serving runtime, if you requested an accelerator to be attached to your model server, the accelerator hardware was detected but was not used by the model when responding to queries. This issue is now resolved.

RHOAIENG-2869 – Cannot edit existing model framework and model path in a multi-model project

Previously, when you tried to edit a model in a multi-model project using the **Deploy model** dialog, the **Model framework** and **Path** values did not update. This issue is now resolved.

RHOAIENG-2724 – Model deployment fails because fields automatically reset in dialog

Previously, when you deployed a model or edited a deployed model, the **Model servers** and **Model framework** fields in the "Deploy model" dialog might have reset to the default state. The **Deploy** button might have remained enabled even though these mandatory fields no longer contained valid values. This issue is now resolved.

RHOAIENG-2099 – Data science pipeline server fails to deploy in fresh cluster

Previously, when you created a data science pipeline server on a fresh cluster, the user interface remained in a loading state and the pipeline server did not start. This issue is now resolved.

RHOAIENG-1199 (previously documented as ODH-DASHBOARD-1928) – Custom serving runtime creation error message is unhelpful

Previously, when you tried to create or edit a custom model-serving runtime and an error occurred, the error message did not indicate the cause of the error. The error messages have been improved.

RHOAIENG-556 – ServingRuntime for KServe model is created regardless of error

Previously, when you tried to deploy a KServe model and an error occurred, the **InferenceService** custom resource (CR) was still created and the model was shown in the **Data Science Projects** page, but the status would always remain unknown. The KServe deploy process has been updated so that the ServingRuntime is not created if an error occurs.

RHOAIENG-548 (previously documented as ODH-DASHBOARD-1776) – Error messages when user does not have project administrator permission

Previously, if you did not have administrator permission for a project, you could not access some features, and the error messages did not explain why. For example, when you created a model server in an environment where you only had access to a single namespace, an **Error creating model server** error message appeared. However, the model server is still successfully created. This issue is now resolved.

RHOAIENG-66 - **Ray dashboard route deployed by CodeFlare SDK exposes self-signed certs instead of cluster cert**

Previously, when you deployed a Ray cluster by using the CodeFlare SDK with the **openshift_oauth=True** option, the resulting route for the Ray cluster was secured by using the **passthrough** method and as a result, the self-signed certificate used by the OAuth proxy was exposed. This issue is now resolved.

RHOAIENG-12 - **Cannot access Ray dashboard from some browsers**

In some browsers, users of the distributed workloads feature might not have been able to access the Ray dashboard because the browser automatically changed the prefix of the dashboard URL from **http** to **https**. This issue is now resolved.

RHODS-6216 - **The ModelMesh oauth-proxy container is intermittently unstable**

Previously, ModelMesh pods did not deploy correctly due to a failure of the ModelMesh **oauth-proxy** container. This issue occurred intermittently and only if authentication was enabled in the ModelMesh runtime environment. This issue is now resolved.

RHOAIENG-535 - **Metrics graph showing HTTP requests for deployed models is incorrect if there are no HTTP requests**

Previously, if a deployed model did not receive at least one HTTP request for each of the two data types (success and failed), the graphs that show HTTP request performance metrics (for all models on the model server or for the specific model) rendered incorrectly, with a straight line that indicated a steadily increasing number of failed requests. This issue is now resolved.

RHOAIENG-1467 - **Serverless net-istio controller pod might hit OOM**

Previously, the Knative **net-istio-controller** pod (which is a dependency for KServe) might continuously crash due to an out-of-memory (OOM) error. This issue is now resolved.

RHOAIENG-1899 (previously documented as RHODS-6539) - **The Anaconda Professional Edition cannot be validated and enabled**

Previously, you could not enable the Anaconda Professional Edition because the dashboard's key validation for it was inoperable. This issue is now resolved.

RHOAIENG-2269 - **(Single-model) Dashboard fails to display the correct number of model replicas**

Previously, on a single-model serving platform, the **Models and model servers** section of a data science project did not show the correct number of model replicas. This issue is now resolved.

RHOAIENG-2270 - **(Single-model) Users cannot update model deployment settings**

Previously, you couldn't edit the deployment settings (for example, the number of replicas) of a model you deployed with a single-model serving platform. This issue is now resolved.

RHODS-8865 - **A pipeline server fails to start unless you specify an Amazon Web Services (AWS) Simple Storage Service (S3) bucket resource**

Previously, when you created a data connection for a data science project, the **AWS_S3_BUCKET** field was not designated as a mandatory field. However, if you attempted to configure a pipeline server with a data connection where the **AWS_S3_BUCKET** field was not populated, the pipeline server failed to start successfully. This issue is now resolved. The **Configure pipeline server** dialog has been updated to include the **Bucket** field as a mandatory field.

RHODS-12899 - OpenVINO runtime missing annotation for NVIDIA GPUs

Previously, if a user selected the **OpenVINO model server (supports GPUs)** runtime and selected an NVIDIA GPU accelerator in the model server user interface, the system could display a unnecessary warning that the selected accelerator was not compatible with the selected runtime. The warning is no longer displayed.

RHOAIENG-84 - Cannot use self-signed certificates with KServe

Previously, the single-model serving platform did not support self-signed certificates. This issue is now resolved. To use self-signed certificates with KServe, follow the steps described in Working with certificates.

RHOAIENG-164 - Number of model server replicas for Kserve is not applied correctly from the dashboard

Previously, when you set a number of model server replicas different from the default (1), the model (server) was still deployed with 1 replica. This issue is now resolved.

RHOAIENG-288 - Recommended image version label for workbench is shown for two versions

Most of the workbench images that are available in OpenShift AI are provided in multiple versions. The only recommended version is the latest version. In Red Hat OpenShift AI 2.4 and 2.5, the **Recommended** tag was erroneously shown for multiple versions of an image. This issue is now resolved.

RHOAIENG-293 - Deprecated ModelMesh monitoring stack not deleted after upgrading from 2.4 to 2.5

In Red Hat OpenShift AI 2.5, the former ModelMesh monitoring stack was no longer deployed because it was replaced by user workload monitoring. However, the former monitoring stack was not deleted during an upgrade to OpenShift AI 2.5. Some components remained and used cluster resources. This issue is now resolved.

RHOAIENG-343 - Manual configuration of OpenShift Service Mesh and OpenShift Serverless does not work for KServe

If you installed OpenShift Serverless and OpenShift Service Mesh and then installed Red Hat OpenShift AI with KServe enabled, KServe was not deployed. This issue is now resolved.

RHOAIENG-517 - User with edit permissions cannot see created models

A user with edit permissions could not see any created models, unless they were the project owner or had admin permissions for the project. This issue is now resolved.

RHOAIENG-804 - Cannot deploy Large Language Models with KServe on FIPS-enabled clusters

Previously, Red Hat OpenShift AI was not yet fully designed for FIPS. You could not deploy Large Language Models (LLMs) with KServe on FIPS-enabled clusters. This issue is now resolved.

RHOAIENG-908 - Cannot use ModelMesh if KServe was previously enabled and then removed

Previously, when both ModelMesh and KServe were enabled in the **DataScienceCluster** object, and you subsequently removed KServe, you could no longer deploy new models with ModelMesh. You could continue to use models that were previously deployed with ModelMesh. This issue is now resolved.

RHOAIENG-2184 - Cannot create Ray clusters or distributed workloads

Previously, users could not create Ray clusters or distributed workloads in namespaces where they have **admin** or **edit** permissions. This issue is now resolved.

ODH-DASHBOARD-1991 - ovms-gpu-ootb is missing recommended accelerator annotation

Previously, when you added a model server to your project, the **Serving runtime** list did not show the **Recommended serving runtime** label for the NVIDIA GPU. This issue is now resolved.

RHOAIENG-807 - Accelerator profile toleration removed when restarting a workbench

Previously, if you created a workbench that used an accelerator profile that in turn included a toleration, restarting the workbench removed the toleration information, which meant that the restart could not complete. A freshly created GPU-enabled workbench might start the first time, but never successfully restarted afterwards because the generated pod remained forever pending. This issue is now resolved.

DATA-SCIENCE-PIPELINES-OPERATOR-294 - Scheduled pipeline run that uses data-passing might fail to pass data between steps, or fail the step entirely

A scheduled pipeline run that uses an S3 object store to store the pipeline artifacts might fail with an error such as the following:

> Bad value for --endpoint-url "cp": scheme is missing. Must be of the form http://<hostname>/ or https://<hostname>/

This issue occurred because the S3 object store endpoint was not successfully passed to the pods for the scheduled pipeline run. This issue is now resolved.

RHODS-4769 - GPUs on nodes with unsupported taints cannot be allocated to notebook servers

GPUs on nodes marked with any taint other than the supported *nvidia.com/gpu* taint could not be selected when creating a notebook server. This issue is now resolved.

RHODS-6346 - Unclear error message displays when using invalid characters to create a data science project

When creating a data science project's data connection, workbench, or storage connection using invalid special characters, the following error message was displayed:

> the object provided is unrecognized (must be of type Secret): couldn't get version/kind; json parse error: unexpected end of JSON input ({"apiVersion":"v1","kind":"Sec ...)

The error message failed to clearly indicate the problem. The error message now indicates that invalid characters were entered.

RHODS-6950 - Unable to scale down workbench GPUs when all GPUs in the cluster are being used

In earlier releases, it was not possible to scale down workbench GPUs if all GPUs in the cluster were being used. This issue applied to GPUs being used by one workbench, and GPUs being used by multiple workbenches. You can now scale down the GPUs by selecting **None** from the **Accelerators** list.

RHODS-8939 – Default shared memory for a Jupyter notebook created in a previous release causes a runtime error

Starting with release 1.31, this issue is resolved, and the shared memory for any new notebook is set to the size of the node.

For a Jupyter notebook created in a release earlier than 1.31, the default shared memory for a Jupyter notebook is set to 64 MB and you cannot change this default value in the notebook configuration.

To fix this issue, you must recreate the notebook or follow the process described in the Knowledgebase article How to change the shared memory for a Jupyter notebook in Red Hat OpenShift AI .

RHODS-9030 – Uninstall process for OpenShift AI might become stuck when removing **kfdefs** resources

The steps for uninstalling the OpenShift AI managed service are described in Uninstalling OpenShift AI.

However, even when you followed this guide, you might have seen that the uninstall process did not finish successfully. Instead, the process stayed on the step of deleting **kfdefs** resources that were used by the Kubeflow Operator. As shown in the following example, **kfdefs** resources might exist in the **redhat-ods-applications**, **redhat-ods-monitoring**, and **rhods-notebooks** namespaces:

```
$ oc get kfdefs.kfdef.apps.kubeflow.org -A

NAMESPACE              NAME                              AGE
redhat-ods-applications    rhods-anaconda                    3h6m
redhat-ods-applications    rhods-dashboard                   3h6m
redhat-ods-applications    rhods-data-science-pipelines-operator  3h6m
redhat-ods-applications    rhods-model-mesh                  3h6m
redhat-ods-applications    rhods-nbc                         3h6m
redhat-ods-applications    rhods-osd-config                  3h6m
redhat-ods-monitoring      modelmesh-monitoring              3h6m
redhat-ods-monitoring      monitoring                        3h6m
rhods-notebooks            rhods-notebooks                   3h6m
rhods-notebooks            rhods-osd-config                  3h5m
```

Failed removal of the **kfdefs** resources might have also prevented later installation of a newer version of OpenShift AI. This issue no longer occurs.

RHODS-9764 – Data connection details get reset when editing a workbench

When you edited a workbench that had an existing data connection and then selected the **Create new data connection** option, the edit page might revert to the **Use existing data connection** option before you had finished specifying the new connection details. This issue is now resolved.

RHODS-9583 – Data Science dashboard did not detect an existing OpenShift Pipelines installation

When the OpenShift Pipelines Operator was installed as a global operator on your cluster, the OpenShift AI dashboard did not detect it. The OpenShift Pipelines Operator is now detected successfully.

ODH-DASHBOARD-1639 – Wrong TLS value in dashboard route

Previously, when a route was created for the OpenShift AI dashboard on OpenShift, the **tls.termination** field had an invalid default value of **Reencrypt**. This issue is now resolved. The new value is **reencrypt**.

ODH-DASHBOARD-1638 – Name placeholder in Triggered Runs tab shows Scheduled run name

Previously, when you clicked **Pipelines** > **Runs** and then selected the **Triggered** tab to configure a triggered run, the example value shown in the **Name** field was **Scheduled run name**. This issue is now resolved.

ODH-DASHBOARD-1547 – "We can't find that page" message displayed in dashboard when pipeline operator installed in background

Previously, when you used the **Data Science Pipelines** page of the dashboard to install the OpenShift Pipelines Operator, when the Operator installation was complete, the page refreshed to show a "We can't find that page" message. This issue is now resolved. When the Operator installation is complete, the dashboard redirects you to the **Pipelines** page, where you can create a pipeline server.

ODH-DASHBOARD-1545 – Dashboard keeps scrolling to bottom of project when Models tab is expanded

Previously, on the **Data Science Projects** page of the dashboard, if you clicked the **Deployed models** tab to expand it and then tried to perform other actions on the page, the page automatically scrolled back to the **Deployed models** section. This affected your ability to perform other actions. This issue is now resolved.

NOTEBOOKS-156 – Elyra included an example runtime called Test

Previously, Elyra included an example runtime configuration called **Test**. If you selected this configuration when running a data science pipeline, you could see errors. The **Test** configuration has now been removed.

RHODS-9622 – Duplicating a scheduled pipeline run does not copy the existing period and pipeline input parameter values

Previously, when you duplicated a scheduled pipeline run that had a periodic trigger, the duplication process did not copy the configured execution frequency for the recurring run or the specified pipeline input parameters. This issue is now resolved.

RHODS-8932 – Incorrect cron format was displayed by default when scheduling a recurring pipeline run

When you scheduled a recurring pipeline run by configuring a cron job, the OpenShift AI interface displayed an incorrect format by default. It now displays the correct format.

RHODS-9374 – Pipelines with non-unique names did not appear in the data science project user interface

If you launched a notebook from a Jupyter application that supported Elyra, or if you used a workbench, when you submitted a pipeline to be run, pipelines with non-unique names did not appear in the **Pipelines** section of the relevant data science project page or the **Pipelines** heading of the data science pipelines page. This issue has now been resolved.

RHODS-9329 – Deploying a custom model-serving runtime could result in an error message

Previously, if you used the OpenShift AI dashboard to deploy a custom model-serving runtime, the deployment process could fail with an **Error retrieving Serving Runtime** message. This issue is now resolved.

RHODS-9064 – After upgrade, the Data Science Pipelines tab was not enabled on the OpenShift AI dashboard

When you upgraded from OpenShift AI 1.26 to OpenShift AI 1.28, the **Data Science Pipelines** tab was not enabled in the OpenShift AI dashboard. This issue is resolved in OpenShift AI 1.29.

[RHODS-9443](#) **– Exporting an Elyra pipeline exposed S3 storage credentials in plain text**

In OpenShift AI 1.28.0, when you exported an Elyra pipeline from JupyterLab in Python DSL format or YAML format, the generated output contained S3 storage credentials in plain text. This issue has been resolved in OpenShift AI 1.28.1. However, after you upgrade to OpenShift AI 1.28.1, if your deployment contains a data science project with a pipeline server and a data connection, you must perform the following additional actions for the fix to take effect:

1. Refresh your browser page.

2. Stop any running workbenches in your deployment and restart them.

Furthermore, to confirm that your Elyra runtime configuration contains the fix, perform the following actions:

1. In the left sidebar of JupyterLab, click **Runtimes** ( ).

2. Hover the cursor over the runtime configuration that you want to view and click the **Edit** button
   ( ).
   The **Data Science Pipelines runtime configuration** page opens.

3. Confirm that **KUBERNETES_SECRET** is defined as the value in the **Cloud Object Storage Authentication Type** field.

4. Close the runtime configuration without changing it.

[RHODS-8460](#) **– When editing the details of a shared project, the user interface remained in a loading state without reporting an error**

When a user with permission to edit a project attempted to edit its details, the user interface remained in a loading state and did not display an appropriate error message. Users with permission to edit projects cannot edit any fields in the project, such as its description. Those users can edit only components belonging to a project, such as its workbenches, data connections, and storage.

The user interface now displays an appropriate error message and does not try to update the project description.

[RHODS-8482](#) **– Data science pipeline graphs did not display node edges for running pipelines**

If you ran pipelines that did not contain Tekton-formatted **Parameters** or **when** expressions in their YAML code, the OpenShift AI user interface did not display connecting edges to and from graph nodes. For example, if you used a pipeline containing the **runAfter** property or **Workspaces**, the user interface displayed the graph for the executed pipeline without edge connections. The OpenShift AI user interface now displays connecting edges to and from graph nodes.

[RHODS-8923](#) **– Newly created data connections were not detected when you attempted to create a pipeline server**

If you created a data connection from within a Data Science project, and then attempted to create a pipeline server, the **Configure a pipeline server** dialog did not detect the data connection that you created. This issue is now resolved.

RHODS-8461 – **When sharing a project with another user, the OpenShift AI user interface text was misleading**

When you attempted to share a Data Science project with another user, the user interface text misleadingly implied that users could edit all of its details, such as its description. However, users can edit only components belonging to a project, such as its workbenches, data connections, and storage. This issue is now resolved and the user interface text no longer misleadingly implies that users can edit all of its details.

RHODS-8462 – **Users with "Edit" permission could not create a Model Server**

Users with "Edit" permissions can now create a Model Server without token authorization. Users must have "Admin" permissions to create a Model Server with token authorization.

RHODS-8796 – **OpenVINO Model Server runtime did not have the required flag to force GPU usage**

OpenShift AI includes the OpenVINO Model Server (OVMS) model-serving runtime by default. When you configured a new model server and chose this runtime, the **Configure model server** dialog enabled you to specify a number of GPUs to use with the model server. However, when you finished configuring the model server and deployed models from it, the model server did not actually use any GPUs. This issue is now resolved and the model server uses the GPUs.

RHODS-8861 – **Changing the host project when creating a pipeline ran resulted in an inaccurate list of available pipelines**

If you changed the host project while creating a pipeline run, the interface failed to make the pipelines of the new host project available. Instead, the interface showed pipelines that belong to the project you initially selected on the **Data Science Pipelines** > **Runs** page. This issue is now resolved. You no longer select a pipeline from the **Create run** page. The pipeline selection is automatically updated when you click the **Create run** button, based on the current project and its pipeline.

RHODS-8249 – **Environment variables uploaded as ConfigMap were stored in Secret instead**

Previously, in the OpenShift AI interface, when you added environment variables to a workbench by uploading a **ConfigMap** configuration, the variables were stored in a **Secret** object instead. This issue is now resolved.

RHODS-7975 – **Workbenches could have multiple data connections**

Previously, if you changed the data connection for a workbench, the existing data connection was not released. As a result, a workbench could stay connected to multiple data sources. This issue is now resolved.

RHODS-7948 – **Uploading a secret file containing environment variables resulted in double-encoded values**

Previously, when creating a workbench in a data science project, if you uploaded a YAML-based secret file containing environment variables, the environment variable values were not decoded. Then, in the resulting OpenShift secret created by this process, the encoded values were encoded again. This issue is now resolved.

RHODS-6429 – **An error was displayed when creating a workbench with the Intel OpenVINO or Anaconda Professional Edition images**

Previously, when you created a workbench with the Intel OpenVINO or Anaconda Professional Edition images, an error appeared during the creation process. However, the workbench was still successfully created. This issue is now resolved.

RHODS-6372 - Idle notebook culler did not take active terminals into account

Previously, if a notebook image had a running terminal, but no active, running kernels, the idle notebook culler detected the notebook as inactive and stopped the terminal. This issue is now resolved.

RHODS-5700 - Data connections could not be created or connected to when creating a workbench

When creating a workbench, users were unable to create a new data connection, or connect to existing data connections.

RHODS-6281 - OpenShift AI administrators could not access Settings page if an admin group was deleted from cluster

Previously, if a Red Hat OpenShift AI administrator group was deleted from the cluster, OpenShift AI administrator users could no longer access the **Settings** page on the OpenShift AI dashboard. In particular, the following behavior was seen:

- When an OpenShift AI administrator user tried to access the **Settings → User management** page, a "Page Not Found" error appeared.

- Cluster administrators *did not* lose access to the **Settings** page on the OpenShift AI dashboard. When a cluster administrator accessed the **Settings → User management** page, a warning message appeared, indicating that the deleted OpenShift AI administrator group no longer existed in OpenShift. The deleted administrator group was then removed from **OdhDashboardConfig**, and administrator access was restored.

This issue is now resolved.

RHODS-1968 - Deleted users stayed logged in until dashboard was refreshed

Previously, when a user's permissions for the Red Hat OpenShift AI dashboard were revoked, the user would notice the change only after a refresh of the dashboard page.

This issue is now resolved. When a user's permissions are revoked, the OpenShift AI dashboard locks the user out within 30 seconds, without the need for a refresh.

RHODS-6384 - A workbench data connection was incorrectly updated when creating a duplicated data connection

When creating a data connection that contained the same name as an existing data connection, the data connection creation failed, but the associated workbench still restarted and connected to the wrong data connection. This issue has been resolved. Workbenches now connect to the correct data connection.

RHODS-6370 - Workbenches failed to receive the latest toleration

Previously, to acquire the latest toleration, users had to attempt to edit the relevant workbench, make no changes, and save the workbench again. Users can now apply the latest toleration change by stopping and then restarting their data science project's workbench.

RHODS-6779 - Models failed to be served after upgrading from OpenShift AI 1.20 to OpenShift AI 1.21

When upgrading from OpenShift AI 1.20 to OpenShift AI 1.21, the **modelmesh-serving** pod attempted to pull a non-existent image, causing an image pull error. As a result, models were unable to be served using the model serving feature in OpenShift AI. The **odh-openvino-servingruntime-container-v1.21.0-15** image now deploys successfully.

RHODS-5945 – **Anaconda Professional Edition could not be enabled in OpenShift AI**

Anaconda Professional Edition could not be enabled for use in OpenShift AI. Instead, an **InvalidImageName** error was displayed in the associated pod's **Events** page. Anaconda Professional Edition can now be successfully enabled.

RHODS-5822 – **Admin users were not warned when usage exceeded 90% and 100% for PVCs created by data science projects.**

Warnings indicating when a PVC exceeded 90% and 100% of its capacity failed to display to admin users for PVCs created by data science projects. Admin users can now view warnings about when a PVC exceeds 90% and 100% of its capacity from the dashboard.

RHODS-5889 – **Error message was not displayed if a data science notebook was stuck in "pending" status**

If a notebook pod could not be created, the OpenShift AI interface did not show an error message. An error message is now displayed if a data science notebook cannot be spawned.

RHODS-5886 – **Returning to the Hub Control Panel dashboard from the data science workbench failed**

If you attempted to return to the dashboard from your workbench Jupyter notebook by clicking on **File → Log Out**, you were redirected to the dashboard and remained on a "Logging out" page. Likewise, if you attempted to return to the dashboard by clicking on **File → Hub Control Panel**, you were incorrectly redirected to the **Start a notebook server** page. Returning to the Hub Control Panel dashboard from the data science workbench now works as expected.

RHODS-6101 – **Administrators were unable to stop all notebook servers**

OpenShift AI administrators could not stop all notebook servers simultaneously. Administrators can now stop all notebook servers using the **Stop all servers** button and stop a single notebook by selecting **Stop server** from the action menu beside the relevant user.

RHODS-5891 – **Workbench event log was not clearly visible**

When creating a workbench, users could not easily locate the event log window in the OpenShift AI interface. The **Starting** label under the **Status** column is now underlined when you hover over it, indicating you can click on it to view the notebook status and the event log.

RHODS-6296 – **ISV icons did not render when using a browser other than Google Chrome**

When using a browser other than Google Chrome, not all ISV icons under **Explore** and **Resources** pages were rendered. ISV icons now display properly on all supported browsers.

RHODS-3182 – **Incorrect number of available GPUs was displayed in Jupyter**

When a user attempts to create a notebook instance in Jupyter, the maximum number of GPUs available for scheduling was not updated as GPUs are assigned. Jupyter now displays the correct number of GPUs available.

RHODS-5890 - When multiple persistent volumes were mounted to the same directory, workbenches failed to start

When mounting more than one persistent volume (PV) to the same mount folder in the same workbench, creation of the notebook pod failed and no errors were displayed to indicate there was an issue.

RHODS-5768 - Data science projects were not visible to users in Red Hat OpenShift AI

Removing the **[DSP]** suffix at the end of a project's **Display Name** property caused the associated data science project to no longer be visible. It is no longer possible for users to remove this suffix.

RHODS-5701 - Data connection configuration details were overwritten

When a data connection was added to a workbench, the configuration details for that data connection were saved in environment variables. When a second data connection was added, the configuration details are saved using the same environment variables, which meant the configuration for the first data connection was overwritten. At the moment, users can add a maximum of one data connection to each workbench.

RHODS-5252 - The notebook Administration page did not provide administrator access to a user's notebook server

The notebook **Administration** page, accessed from the OpenShift AI dashboard, did not provide the means for an administrator to access a user's notebook server. Administrators were restricted to only starting or stopping a user's notebook server.

RHODS-2438 - PyTorch and TensorFlow images were unavailable when upgrading

When upgrading from OpenShift AI 1.3 to a later version, PyTorch and TensorFlow images were unavailable to users for approximately 30 minutes. As a result, users were unable to start PyTorch and TensorFlow notebooks in Jupyter during the upgrade process. This issue has now been resolved.

RHODS-5354 - Environment variable names were not validated when starting a notebook server

Environment variable names were not validated on the **Start a notebook server** page. If an invalid environment variable was added, users were unable to successfully start a notebook. The environmental variable name is now checked in real-time. If an invalid environment variable name is entered, an error message displays indicating valid environment variable names must consist of alphabetic characters, digits, _, -, or ., and must not start with a digit.

RHODS-4617 - The Number of GPUs drop-down was only visible if there were GPUs available

Previously, the **Number of GPUs** drop-down was only visible on the **Start a notebook server** page if GPU nodes were available. The **Number of GPUs** drop-down now also correctly displays if an autoscaling machine pool is defined in the cluster, even if no GPU nodes are currently available, possibly resulting in the provisioning of a new GPU node on the cluster.

RHODS-5420 - Cluster admin did not get administrator access if it was the only user present in the cluster

Previously, when the cluster admin was the only user present in the cluster, it did not get Red Hat OpenShift administrator access automatically. Administrator access is now correctly applied to the cluster admin user.

RHODS-4321 - Incorrect package version displayed during notebook selection

The **Start a notebook server** page displayed an incorrect version number (11.4 instead of 11.7) for the CUDA notebook image. The version of CUDA installed is no longer specified on this page.

[RHODS-5001](#) - **Admin users could add invalid tolerations to notebook pods**

An admin user could add invalid tolerations on the **Cluster settings** page without triggering an error. If a invalid toleration was added, users were unable to successfully start notebooks. The toleration key is now checked in real-time. If an invalid toleration name is entered, an error message displays indicating valid toleration names consist of alphanumeric characters, -, _, or ., and must start and end with an alphanumeric character.

[RHODS-5100](#) - **Group role bindings were not applied to cluster administrators**

Previously, if you had assigned cluster admin privileges to a group rather than a specific user, the dashboard failed to recognize administrative privileges for users in the administrative group. Group role bindings are now correctly applied to cluster administrators as expected.

[RHODS-4947](#) - **Old Minimal Python notebook image persisted after upgrade**

After upgrading from OpenShift AI 1.14 to 1.15, the older version of the Minimal Python notebook persisted, including all associated package versions. The older version of the Minimal Python notebook no longer persists after upgrade.

[RHODS-4935](#) - **Excessive "missing x-forwarded-access-token header" error messages displayed in dashboard log**

The **rhods-dashboard** pod's log contained an excessive number of "missing x-forwarded-access-token header" error messages due to a readiness probe hitting the **/status** endpoint. This issue has now been resolved.

[RHODS-2653](#) - **Error occurred while fetching the generated images in the sample Pachyderm notebook**

An error occurred when a user attempted to fetch an image using the sample Pachyderm notebook in Jupyter. The error stated that the image could not be found. Pachyderm has corrected this issue.

[RHODS-4584](#) - **Jupyter failed to start a notebook server using the OpenVINO notebook image**

Jupyter's **Start a notebook server** page failed to start a notebook server using the OpenVINO notebook image. Intel has provided an update to the OpenVINO operator to correct this issue.

[RHODS-4923](#) - **A non-standard check box displayed after disabling usage data collection**

After disabling usage data collection on the **Cluster settings** page, when a user accessed another area of the OpenShift AI dashboard, and then returned to the **Cluster settings** page, the **Allow collection of usage data** check box had a non-standard style applied, and therefore did not look the same as other check boxes when selected or cleared.

[RHODS-4938](#) - **Incorrect headings were displayed in the Notebook Images page**

The **Notebook Images** page, accessed from the **Settings** page on the OpenShift AI dashboard, displayed incorrect headings in the user interface. The **Notebook image settings** heading displayed as **BYON image settings**, and the **Import Notebook images** heading displayed as **Import BYON images**. The correct headings are now displayed as expected.

[RHODS-4818](#) - **Jupyter was unable to display images when the NVIDIA GPU add-on was installed**

The **Start a notebook server** page did not display notebook images after installing the NVIDIA GPU add-on. Images are now correctly displayed, and can be started from the **Start a notebook server** page.

RHODS-4797 - **PVC usage limit alerts were not sent when usage exceeded 90% and 100%**

Alerts indicating when a PVC exceeded 90% and 100% of its capacity failed to be triggered and sent. These alerts are now triggered and sent as expected.

RHODS-4366 - **Cluster settings were reset on operator restart**

When the OpenShift AI operator pod was restarted, cluster settings were sometimes reset to their default values, removing any custom configuration. The OpenShift AI operator was restarted when a new version of OpenShift AI was released, and when the node that ran the operator failed. This issue occurred because the operator deployed ConfigMaps incorrectly. Operator deployment instructions have been updated so that this no longer occurs.

RHODS-4318 - **The OpenVINO notebook image failed to build successfully**

The OpenVINO notebook image failed to build successfully and displayed an error message. This issue has now been resolved.

RHODS-3743 - **Starburst Galaxy quick start did not provide download link in the instruction steps**

The Starburst Galaxy quick start, located on the **Resources** page on the dashboard, required the user to open the **explore-data.ipynb notebook**, but failed to provide a link within the instruction steps. Instead, the link was provided in the quick start's introduction.

RHODS-1974 - **Changing alert notification emails required pod restart**

Changes to the list of notification email addresses in the Red Hat OpenShift AI Add-On were not applied until after the **rhods-operator** pod and the **prometheus-*** pod were restarted.

RHODS-2738 - **Red Hat OpenShift API Management 1.15.2 add-on installation did not successfully complete**

For OpenShift AI installations that are integrated with the Red Hat OpenShift API Management 1.15.2 add-on, the Red Hat OpenShift API Management installation process did not successfully obtain the SMTP credentials secret. Subsequently, the installation did not complete.

RHODS-3237 - **GPU tutorial did not appear on dashboard**

The "GPU computing" tutorial, located at Gtc2018-numba, did not appear on the **Resources** page on the dashboard.

RHODS-3069 - **GPU selection persisted when GPU nodes were unavailable**

When a user provisioned a notebook server with GPU support, and the utilized GPU nodes were subsequently removed from the cluster, the user could not create a notebook server. This occurred because the most recently used setting for the number of attached GPUs was used by default.

RHODS-3181 - **Pachyderm now compatible with OpenShift Dedicated 4.10 clusters**

Pachyderm was not initially compatible with OpenShift Dedicated 4.10, and so was not available in OpenShift AI running on an OpenShift Dedicated 4.10 cluster. Pachyderm is now available on and compatible with OpenShift Dedicated 4.10.

RHODS-2160 – Uninstall process failed to complete when both OpenShift AI and OpenShift API Management were installed

When OpenShift AI and OpenShift API Management are installed together on the same cluster, they use the same Virtual Private Cluster (VPC). The uninstall process for these Add-ons attempts to delete the VPC. Previously, when both Add-ons are installed, the uninstall process for one service was blocked because the other service still had resources in the VPC. The cleanup process has been updated so that this conflict does not occur.

RHODS-2747 – Images were incorrectly updated after upgrading OpenShift AI

After the process to upgrade OpenShift AI completed, Jupyter failed to update its notebook images. This was due to an issue with the image caching mechanism. Images are now correctly updating after an upgrade.

RHODS-2425 – Incorrect TensorFlow and TensorBoard versions displayed during notebook selection

The **Start a notebook server** page displayed incorrect version numbers (2.4.0) for TensorFlow and TensorBoard in the TensorFlow notebook image. These versions have been corrected to TensorFlow 2.7.0 and TensorBoard 2.6.0.

RHODS-24339 – Quick start links did not display for enabled applications

For some applications, the **Open quick start** link failed to display on the application tile on the **Enabled** page. As a result, users did not have direct access to the quick start tour for the relevant application.

RHODS-2215 – Incorrect Python versions displayed during notebook selection

The **Start a notebook server** page displayed incorrect versions of Python for the TensorFlow and PyTorch notebook images. Additionally, the third integer of package version numbers is now no longer displayed.

RHODS-1977 – Ten minute wait after notebook server start fails

If the Jupyter leader pod failed while the notebook server was being started, the user could not access their notebook server until the pod restarted, which took approximately ten minutes. This process has been improved so that the user is redirected to their server when a new leader pod is elected. If this process times out, users see a 504 Gateway Timeout error, and can refresh to access their server.

CHAPTER 6. KNOWN ISSUES

This section describes known issues in Red Hat OpenShift AI and any known methods of working around these issues.

RHOAIENG-6701 – **Users without cluster administrator privileges cannot access the job submission endpoint of the Ray dashboard**

Users of the distributed workloads feature who do not have cluster administrator privileges for OpenShift might not be able to access or use the job submission endpoint of the Ray dashboard.

RHOAIENG-6649 – **An error is displayed when viewing a model on a model server that has no external route defined**

If you use the dashboard to deploy a model on a model server that does not have external routes enabled, a *t.components is undefined* error message might be shown while the model creation is in progress.

RHOAIENG-6646 – **An error is displayed when viewing the Model Serving page during an upgrade**

If you try to use the dashboard to deploy a model while an upgrade of OpenShift AI is in progress, a *t.status is undefined* error message might be shown.

Workaround

Wait until the upgraded OpenShift AI Operator is ready and then refresh the page in your browser.

RHOAIENG-6578 – **Models that show authorization tokens do not actually require them after deletion of Authorino Operator**

If you added Authorino as an authorization provider for the single-model serving platform, but later delete the Authorino Operator, the OpenShift AI dashboard might still show tokens for models that had token authorization enabled. However, authorization is no longer enabled, and you are able to make successful inference requests to the models without specifying the tokens in your requests.

RHOAIENG-6486 – **Pod labels, annotations, and tolerations cannot be configured when using the Elyra JupyterLab extension with the TensorFlow 2024.1 notebook image**

When using the Elyra JupyterLab extension with the TensorFlow 2024.1 notebook image, you cannot configure pod labels, annotations, or tolerations from an executed pipeline. This is due to a dependency conflict with the kfp and tf2onnx packages.

Workaround

If you are working with the TensorFlow 2024.1 notebook image, after you have completed your work, change the assigned workbench notebook image to the Standard Data Science 2024.1 notebook image.

In the **Pipeline properties** tab in the Elyra JupyterLab extension, set the Tensorflow runtime image as the default runtime image for the pipeline node individually, along with the relevant pod label, annotation or toleration, for each pipeline node.

RHOAIENG-6435 – **Distributed workloads resources are not included in Project metrics**

When you click **Distributed Workloads Metrics** > **Project metrics** and view the **Requested resources** section, the **Requested by all projects** value currently excludes the resources for distributed workloads that have not yet been admitted to the queue.

RHOAIENG-6409 - **Cannot save parameter** errors appear in pipeline logs for successful runs

When you run a pipeline more than once with Data Science Pipelines 2.0, **Cannot save parameter** errors appear in the pipeline logs for successful pipeline runs. You can safely ignore these errors.

RHOAIENG-6376 - **Pipeline run creation fails after setting pip_index_urls in a pipeline component to a URL that contains a port number and path**

When you create a pipeline and set the **pip_index_urls** value for a component to a URL that contains a port number and path, compiling the pipeline code and then creating a pipeline run results in the following error:

> ValueError: Invalid IPv6 URL

Workaround

1. Create a new pip server using only **protocol://hostname**, and update the **pip_index_urls** value for the component with the new server.

2. Recompile the pipeline code.

3. Create a new pipeline run.

RHOAIENG-6343 - **Some components are set to Removed after installing OpenShift AI**

After you install OpenShift AI, the **managementState** field for the **codeflare**, **kueue**, and **ray** components is incorrectly set to **Removed** instead of **Managed** in the **DataScienceCluster** custom resource.

Workaround

1. Log in to the OpenShift web console as a cluster administrator.

2. Click **Operators → Installed Operators** and then click the Red Hat OpenShift AI Operator.

3. Click the **Data Science Cluster** tab.

4. Select the **default-dsc** object.

5. Click the **YAML** tab and in the **spec.components** section, set the value of the **managementState** field for **codeflare**, **kueue**, and **ray** to **Managed** as shown:

```
spec:
 components:
   "codeflare": {
     "managementState": "Managed"
   },
   "kueue": {
     "managementState": "Managed"
    },
   "ray": {
     "managementState": "Managed"
   },
```

6. Click **Save**.

RHOAIENG-6317 – **An error is displayed when viewing pipeline run pod logs in the dashboard**

When using the log viewer in the OpenShift AI dashboard to view pipeline run pod logs, a **pods not found** error message might be shown.

Workaround

> Follow the steps in the Data Science Pipelines workaround for how to view pipeline run pod logs in the Red Hat OpenShift AI dashboard knowledgebase article.

RHOAIENG-5314 – **Data science pipeline server fails to deploy in fresh cluster due to network policies**

When you create a data science pipeline server on a fresh cluster, the user interface remains in a loading state and the pipeline server does not start. A "Pipeline server failed" error message might be displayed.

Workaround

1. Log in to the OpenShift web console as a cluster administrator.

2. Click **Networking** > **NetworkPolicies**.

3. Click the **Project** list and select your project.

4. Click the **Create NetworkPolicy** button.

5. For **Configure via**, select **YAML view** and define the network policy as shown:

   ```
   kind: NetworkPolicy
   apiVersion: networking.k8s.io/v1
   metadata:
     name: allow-from-redhat-ods-app-to-mariadb
   spec:
     podSelector:
       matchLabels:
         app: mariadb-pipelines-definition
     ingress:
       - ports:
           - protocol: TCP
             port: 3306
         from:
           - podSelector:
               matchLabels:
                 app.kubernetes.io/name: data-science-pipelines-operator
             namespaceSelector:
               matchLabels:
                 kubernetes.io/metadata.name: redhat-ods-applications
     policyTypes:
       - Ingress
   ```

6. Click **Create**.

RHOAIENG-4812 – **Distributed workload metrics exclude GPU metrics**

In this release of OpenShift AI, the distributed workload metrics exclude GPU metrics.

RHOAIENG-4570 – Existing Argo Workflows installation conflicts with install or upgrade

Data Science Pipelines (DSP) 2.0 contains an installation of Argo Workflows. OpenShift AI does not support direct customer usage of this installation of Argo Workflows. To install or upgrade OpenShift AI with DSP 2.0, ensure that there is no existing installation of Argo Workflows on your cluster. For more information, see Enabling Data Science Pipelines 2.0 .

Workaround

> Remove the existing Argo Workflows installation or set **datasciencepipelines** to **Removed**, and then proceed with the installation or upgrade.

RHOAIENG-3913 – Red Hat OpenShift AI Operator incorrectly shows **Degraded** condition of **False** with an error

If you have enabled the KServe component in the DataScienceCluster (DSC) object used by the OpenShift AI Operator, but have not installed the dependent Red Hat OpenShift Service Mesh and Red Hat OpenShift Serverless Operators, the **kserveReady** condition in the DSC object correctly shows that KServe is not ready. However, the **Degraded** condition incorrectly shows a value of **False**.

Workaround

> Install the Red Hat OpenShift Serverless and Red Hat OpenShift Service Mesh Operators, and then recreate the DSC.

RHOAIENG-4252 – Data science pipeline server deletion process fails to remove **ScheduledWorkFlow** resource

The pipeline server deletion process does not remove the **ScheduledWorkFlow** resource. As a result, new DataSciencePipelinesApplications (DSPAs) do not recognize the redundant **ScheduledWorkFlow** resource.

Workaround

1. Delete the pipeline server. For more information, see Deleting a pipeline server.

2. In the OpenShift command-line interface (CLI), log in to your cluster as a cluster administrator and perform the following command to delete the redundant **ScheduledWorkFlow** resource.

   ```
   $ oc -n <data science project name> delete scheduledworkflows --all
   ```

RHOAIENG-4240 – Jobs fail to submit to Ray cluster in unsecured environment

When running distributed data science workloads from notebooks in an unsecured OpenShift cluster, a **ConnectionError: Failed to connect to Ray** error message might be shown.

Workaround

> In the **ClusterConfiguration** section of the notebook, set the **openshift_oauth** option to **True**.

RHOAIENG-3981 – In unsecured environment, the functionality to wait for Ray cluster to be ready gets stuck

When running distributed data science workloads from notebooks in an unsecured OpenShift cluster, the functionality to wait for the Ray cluster to be ready before proceeding (**cluster.wait_ready()**) gets stuck even when the Ray cluster is ready.

Workaround

Perform one of the following actions:

- In the **ClusterConfiguration** section of the notebook, set the **openshift_oauth** option to **True**.

- Instead of using the **cluster.wait_ready()**, functionality, you can manually check the Ray cluster availability by opening the Ray cluster Route URL. When the Ray dashboard is available on the URL, then the cluster is ready.

[RHOAIENG-3025](#) – OVMS expected directory layout conflicts with the KServe StoragePuller layout

When you use the OpenVINO Model Server (OVMS) runtime to deploy a model on the single-model serving platform (which uses KServe), there is a mismatch between the directory layout expected by OVMS and that of the model-pulling logic used by KServe. Specifically, OVMS requires the model files to be in the **/<mnt>/models/1/** directory, while KServe places them in the **/<mnt>/models/** directory.

Workaround

Perform the following actions:

1. In your S3-compatible storage bucket, place your model files in a directory called **1**/, for example, **/<s3_storage_bucket>/models/1/<model_files>**.

2. To use the OVMS runtime to deploy a model on the single-model serving platform, choose one of the following options to specify the path to your model files:

   - If you are using the OpenShift AI dashboard to deploy your model, in the **Path** field for your data connection, use the **/<s3_storage_bucket>/models/** format to specify the path to your model files. Do not specify the **1**/ directory as part of the path.

   - If you are creating your own **InferenceService** custom resource to deploy your model, configure the value of the **storageURI** field as **/<s3_storage_bucket>/models/**. Do not specify the **1**/ directory as part of the path.

KServe pulls model files from the subdirectory in the path that you specified. In this case, KServe correctly pulls model files from the **/<s3_storage_bucket>/models/1/** directory in your S3-compatible storage.

[RHOAIENG-3018](#) – OVMS on KServe does not expose the correct endpoint in the dashboard

When you use the OpenVINO Model Server (OVMS) runtime to deploy a model on the single-model serving platform, the URL shown in the **Inference endpoint** field for the deployed model is not complete. To send queries to the model, you must add the **/v2/models/_<model-name>_/infer** string to the end of the URL. Replace **_<model-name>_** with the name of your deployed model.

[RHOAIENG-3378](#) – Internal Image Registry is an undeclared hard dependency for Jupyter notebooks spawn process

Before you can start OpenShift AI notebooks and workbenches, you must first enable the internal, integrated container image registry in OpenShift. Attempts to start notebooks or workbenches without first enabling the image registry will fail with an "InvalidImageName" error.

You can confirm whether the image registry is enabled for a cluster by using the following command:

> $ oc get pods -n openshift-image-registry

Workaround

> Enable the internal, integrated container image registry in OpenShift. See Image Registry Operator in OpenShift for more information about how to set up and configure the image registry.

RHOAIENG-2759 – Model deployment fails when both secured and regular model servers are present in a project

When you create a second model server in a project where one server is using token authentication, and the other server does not use authentication, the deployment of the second model might fail to start.

RHOAIENG-2602 – "Average response time" server metric graph shows multiple lines due to ModelMesh pod restart

The **Average response time**server metric graph shows multiple lines if the ModelMesh pod is restarted.

RHOAIENG-2585 – UI does not display an error/warning when UWM is not enabled in the cluster

Red Hat OpenShift AI does not correctly warn users if User Workload Monitoring (UWM) is **disabled** in the cluster. UWM is necessary for the correct functionality of model metrics.

Workaround

> Manually ensure that UWM is enabled in your cluster, as described in Enabling monitoring for user-defined projects.

RHOAIENG-2555 – Model framework selector does not reset when changing Serving Runtime in form

When you use the **Deploy model** dialog to deploy a model on the single-model serving platform, if you select a runtime and a supported framework, but then switch to a different runtime, the existing framework selection is not reset. This means that it is possible to deploy the model with a framework that is not supported for the selected runtime.

Workaround

> While deploying a model, if you change your selected runtime, click the **Select a framework** list again and select a supported framework.

RHOAIENG-2468 – Services in the same project as KServe might become inaccessible in OpenShift

If you deploy a non-OpenShift AI service in a data science project that contains models deployed on the single-model serving platform (which uses KServe), the accessibility of the service might be affected by the network configuration of your OpenShift cluster. This is particularly likely if you are using the OVN-Kubernetes network plugin in combination with host network namespaces.

Workaround

> Perform one of the following actions:

> - Deploy the service in another data science project that does not contain models deployed on the single-model serving platform. Or, deploy the service in another OpenShift project.

> - In the data science project where the service is, add a network policy to accept ingress traffic to your application pods, as shown in the following example:

```
apiVersion: networking.k8s.io/v1
kind: NetworkPolicy
metadata:
  name: allow-ingress-to-myapp
spec:
  podSelector:
    matchLabels:
      app: myapp
  ingress:
    - {}
```

RHOAIENG-2312 – Importing numpy fails in **code-server** workbench

Importing **numpy** in your **code-server** workbench fails.

Workaround

1. In your **code-server** workbench, from the **Activity bar**, select the menu icon (☰) > **View** > **Command Palette** to open the Command Palette.
   In Firefox, you can use the F1 keyboard shortcut to open the command palette.

2. Enter **python: s**.

3. From the drop-down list, select the **Python: Select interpreter** action.

4. In the **Select Interpreter** dialog, select **Enter interpreter path…**.

5. Enter **/opt/app-root/bin/python3** as the interpreter path and press **Enter**.

6. From the drop-down list, select the new Python interpreter.

7. Confirm that the new interpreter (**app-root**) appears on the **Status bar**. The selected interpreter persists if the workbench is stopped and started again, so the workaround should need to be performed only once for each workbench.

RHOAIENG-2228 – The performance metrics graph changes constantly when the interval is set to 15 seconds

On the **Endpoint performance** tab of the model metrics screen, if you set the **Refresh interval** to 15 seconds and the **Time range** to 1 hour, the graph results change continuously.

RHOAIENG-2183 – Endpoint performance graphs might show incorrect labels

In the **Endpoint performance** tab of the model metrics screen, the graph tooltip might show incorrect labels.

RHOAIENG-1919 – Model Serving page fails to fetch or report the model route URL soon after its deployment

When deploying a model from the OpenShift AI dashboard, the system displays the following warning message while the **Status** column of your model indicates success with an **OK**/green checkmark.

Failed to get endpoint for this deployed model. routes.rout.openshift.io"<model_name>" not found

■

Workaround

Refresh your browser page.

RHOAIENG-1452 – The Red Hat OpenShift AI Add-on gets stuck

The Red Hat OpenShift AI Add-on uninstall does not delete OpenShift AI components after being triggered via OCM APIs.

Workaround

Manually delete the remaining OpenShift AI resources as follows:

1. Delete the **DataScienceCluster** CR.

2. Wait until all pods are deleted from the **redhat-ods-applications** namespace.

3. If Serverless was set to **Managed** in the **DataScienceCluster** CR, wait until all pods are deleted from the **knative-serving** namespace.

4. Delete the **DSCInitialization** CR.

5. If Service Mesh was set to **Managed** in the **DSCInitialization** CR, wait until all pods are deleted from the **istio-system** namespace.

6. Uninstall the Red Hat OpenShift AI Operator.

7. Wait until all pods are deleted from the **redhat-ods-operator** namespace and the **redhat-ods-monitoring** namespace.

RHOAIENG-880 – Default pipelines service account is unable to create Ray clusters

You cannot create Ray clusters using the default pipelines Service Account.

Workaround

Authenticate using the CodeFlare SDK, by adding the following lines to the pipeline code:

```
from codeflare_sdk.cluster.auth import TokenAuthentication
  auth = TokenAuthentication(
    token=openshift_token, server=openshift_server
  )
  auth_return = auth.login()
```
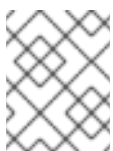
> **NOTE**
>
> If your cluster uses self-signed certificates, include **ca-cert-path=<path>** in the **TokenAuthentication** parameter list.

RHOAIENG-404 – No Components Found page randomly appears instead of Enabled page in OpenShift AI dashboard

A No Components Found page might appear when you access the Red Hat OpenShift AI dashboard.

Workaround

Refresh the browser page.

RHOAIENG-2541 - KServe controller pod experiences OOM because of too many secrets in the cluster

If your OpenShift cluster has a large number of secrets, the KServe controller pod might continually crash due to an out-of-memory (OOM) error.

Workaround

Reduce the number of secrets in the OpenShift cluster until the KServe controller pod becomes stable.

RHOAIENG-1128 - Unclear error message displays when attempting to increase the size of a Persistent Volume (PV) that is not connected to a workbench

When attempting to increase the size of a Persistent Volume (PV) that is not connected to a workbench, an unclear error message is displayed.

Workaround

Verify that your PV is connected to a workbench before attempting to increase the size.

RHOAIENG-545 - Cannot specify a generic default node runtime image in JupyterLab pipeline editor

When you edit an Elyra pipeline in the JupyterLab IDE pipeline editor, and you click the **PIPELINE PROPERTIES** tab, and scroll to the **Generic Node Defaults** section and edit the **Runtime Image** field, your changes are not saved.

Workaround

Define the required runtime image explicitly for each node. Click the **NODE PROPERTIES** tab, and specify the required image in the **Runtime Image** field.

RHOAIENG-497 - Removing DSCI Results In OpenShift Service Mesh CR Being Deleted Without User Notification

If you delete the **DSCInitialization** resource, the OpenShift Service Mesh CR is also deleted. A warning message is not shown.

RHOAIENG-307 - Removing the DataScienceCluster deletes all OpenShift Serverless CRs

If you delete the **DataScienceCluster** custom resource (CR), all OpenShift Serverless CRs (including knative-serving, deployments, gateways, and pods) are also deleted. A warning message is not shown.

RHOAIENG-282 - Workload should not be dispatched if required resources are not available

Sometimes a workload is dispatched even though a single machine instance does not have sufficient resources to provision the RayCluster successfully. The **AppWrapper** CRD remains in a **Running** state and related pods are stuck in a **Pending** state indefinitely.

Workaround

Add extra resources to the cluster.

RHOAIENG-131 - gRPC endpoint not responding properly after the InferenceService reports as Loaded

When numerous **InferenceService** instances are generated and directed requests, Service Mesh Control Plane (SMCP) becomes unresponsive. The status of the **InferenceService** instance is **Loaded**, but the call to the gRPC endpoint returns with errors.

Workaround

Edit the **ServiceMeshControlPlane** custom resource (CR) to increase the memory limit of the Istio egress and ingress pods.

RHOAIENG-130 – Synchronization issue when the model is just launched

When the status of the KServe container is **Ready**, a request is accepted even though the TGIS container is not ready.

Workaround

Wait a few seconds to ensure that all initialization has completed and the TGIS container is actually ready, and then review the request output.

RHOAIENG-3115 – Model cannot be queried for a few seconds after it is shown as ready

Models deployed using the multi-model serving platform might be unresponsive to queries despite appearing as **Ready** in the dashboard. You might see an "Application is not available" response when querying the model endpoint.

Workaround

Wait 30-40 seconds and then refresh the page in your browser.

RHOAIENG-1619 (previously documented as DATA-SCIENCE-PIPELINES-165) – Poor error message when S3 bucket is not writable

When you set up a data connection and the S3 bucket is not writable, and you try to upload a pipeline, the error message **Failed to store pipelines** is not helpful.

Workaround

Verify that your data connection credentials are correct and that you have write access to the bucket you specified.

RHOAIENG-1207 (previously documented as ODH-DASHBOARD-1758) – Error duplicating OOTB custom serving runtimes several times

If you duplicate a model-serving runtime several times, the duplication fails with the **Serving runtime name "<name>" already exists** error message.

Workaround

Change the **metadata.name** field to a unique value.

RHOAIENG-1204 (previously documented as ODH-DASHBOARD-1771) – JavaScript error during Pipeline step initializing

Sometimes the pipeline **Run details** page stops working when the run starts.

Workaround

Refresh the page.

RHOAIENG-1203 (previously documented as ODH-DASHBOARD-1781) – Missing tooltip for Started Run status

Data science pipeline runs sometimes don't show the tooltip text for the status icon shown.

Workaround

For more information, view the pipeline **Run details** page and see the run output.

RHOAIENG-1201 (previously documented asODH-DASHBOARD-1908) – Cannot create workbench with an empty environment variable

When creating a workbench, if you click **Add variable** but do not select an environment variable type from the list, you cannot create the workbench. The field is not marked as required, and no error message is shown.

RHOAIENG-1196 (previously documented asODH-DASHBOARD-2140) – Package versions displayed in dashboard do not match installed versions

The dashboard might display inaccurate version numbers for packages such as JupyterLab and Notebook. The package version number can differ in the image if the packages are manually updated.

Workaround

To find the true version number for a package, run the **pip list** command and search for the package name, as shown in the following examples:

```
$ pip list | grep jupyterlab
jupyterlab                3.5.3
$ pip list | grep notebook
notebook                  6.5.3
```

RHOAIENG-582 (previously documented asODH-DASHBOARD-1335) – Rename Edit permission to Contributor

The term *Edit* is not accurate:

- For *most* resources, users with the **Edit** permission can not only edit the resource, they can also create and delete the resource.

- Users with the **Edit** permission cannot edit the project.

The term *Contributor* more accurately describes the actions granted by this permission.

RHOAIENG-432 (previously documented asRHODS-12928) – Using unsupported characters can generate Kubernetes resource names with multiple dashes

When you create a resource and you specify unsupported characters in the name, then each space is replaced with a dash and other unsupported characters are removed, which can result in an invalid resource name.

RHOAIENG-226 (previously documented asRHODS-12432) – Deletion of the notebook-culler ConfigMap causes Permission Denied on dashboard

If you delete the **notebook-controller-culler-config** ConfigMap in the **redhat-ods-applications** namespace, you can no longer save changes to the **Cluster Settings** page on the OpenShift AI dashboard. The save operation fails with an **HTTP request has failed** error.

Workaround

Complete the following steps as a user with **cluster-admin** permissions:

1. Log in to your cluster by using the **oc** client.

2. Enter the following command to update the **OdhDashboardConfig** custom resource in the **redhat-ods-applications** application namespace:

```
$ oc patch OdhDashboardConfig odh-dashboard-config -n redhat-ods-applications --
type=merge -p '{"spec": {"dashboardConfig": {"notebookController.enabled": true}}}'
```

RHOAIENG-133 – Existing workbench cannot run Elyra pipeline after notebook restart

If you use the Elyra JupyterLab extension to create and run data science pipelines within JupyterLab, and you configure the pipeline server *after* you created a workbench and specified a notebook image within the workbench, you cannot execute the pipeline, even after restarting the notebook.

Workaround

1. Stop the running notebook.

2. Edit the workbench to make a small modification. For example, add a new dummy environment variable, or delete an existing unnecessary environment variable. Save your changes.

3. Restart the notebook.

4. In the left sidebar of JupyterLab, click **Runtimes**.

5. Confirm that the default runtime is selected.

RHOAIENG-52 – Token authentication fails in clusters with self-signed certificates

If you use self-signed certificates, and you use the Python **codeflare-sdk** in a notebook or in a Python script as part of a pipeline, token authentication will fail.

RHOAIENG-11 – Separately installed instance of CodeFlare Operator not supported

In Red Hat OpenShift AI, the CodeFlare Operator is included in the base product and not in a separate Operator. Separately installed instances of the CodeFlare Operator from Red Hat or the community are not supported.

Workaround

Delete any installed CodeFlare Operators, and install and configure Red Hat OpenShift AI, as described in the Red Hat Knowledgebase solution How to migrate from a separately installed CodeFlare Operator in your data science cluster.

RHODS-12798 – Pods fail with "unable to init seccomp" error

Pods fail with **CreateContainerError** status or **Pending** status instead of **Running** status, because of a known kernel bug that introduced a **seccomp** memory leak. When you check the events on the namespace where the pod is failing, or run the **oc describe pod** command, the following error appears:

```
runc create failed: unable to start container process: unable to init seccomp: error loading seccomp
filter into kernel: error loading seccomp filter: errno 524
```

Workaround

Increase the value of **net.core.bpf_jit_limit** as described in the Red Hat Knowledgebase solution Pods failing with error loading seccomp filter into kernel: errno 524 in OpenShift 4 .

KUBEFLOW-177 - Bearer token from application not forwarded by OAuth-proxy

You cannot use an application as a custom workbench image if its internal authentication mechanism is based on a bearer token. The OAuth-proxy configuration removes the bearer token from the headers, and the application cannot work properly.

NOTEBOOKS-210 - A notebook fails to export as a PDF file in Jupyter

When you export a notebook as a PDF file in Jupyter, the export process fails with an error.

RHOAIENG-1210 (previously documented asODH-DASHBOARD-1699) - Workbench does not automatically restart for all configuration changes

When you edit the configuration settings of a workbench, a warning message appears stating that the workbench will restart if you make any changes to its configuration settings. This warning is misleading because in the following cases, the workbench does not automatically restart:

- Edit name

- Edit description

- Edit, add, or remove keys and values of existing environment variables

Workaround

Manually restart the workbench.

RHOAIENG-1208 (previously documented asODH-DASHBOARD-1741) - Cannot create a workbench whose name begins with a number

If you try to create a workbench whose name begins with a number, the workbench does not start.

Workaround

Delete the workbench and create a new one with a name that begins with a letter.

RHOAIENG-1205 (previously documented as RHODS-11791) - Usage data collection is enabled after upgrade

If you previously had the **Allow collection of usage data** option deselected (that is, disabled), this option becomes selected (that is, enabled) when you upgrade OpenShift AI.

Workaround

Manually reset the **Allow collection of usage data** option. To do this, perform the following actions:

1. In the OpenShift AI dashboard, in the left menu, click **Settings → Cluster settings**. The **Cluster Settings** page opens.

2. In the **Usage data collection**section, deselect **Allow collection of usage data**.

3. Click **Save changes**.

KUBEFLOW-157 - Logging out of JupyterLab does not work if you are already logged out of the OpenShift AI dashboard

If you log out of the OpenShift AI dashboard before you log out of JupyterLab, then logging out of JupyterLab is not successful. For example, if you know the URL for a Jupyter notebook, you are able to open this again in your browser.

Workaround

Log out of JupyterLab before you log out of the OpenShift AI dashboard.

RHODS-9789 – Pipeline servers fail to start if they contain a custom database that includes a dash in its database name or username field

When you create a pipeline server that uses a custom database, if the value that you set for the **dbname** field or **username** field includes a dash, the pipeline server fails to start.

Workaround

Edit the pipeline server to omit the dash from the affected fields.

RHOAIENG-580 (previously documented as RHODS-9412) – Elyra pipeline fails to run if workbench is created by a user with edit permissions

If a user who has been granted edit permissions for a project creates a project workbench, that user sees the following behavior:

- During the workbench creation process, the user sees an **Error creating workbench** message related to the creation of Kubernetes role bindings.

- Despite the preceding error message, OpenShift AI still creates the workbench. However, the error message means that the user will not be able to use the workbench to run Elyra data science pipelines.

- If the user tries to use the workbench to run an Elyra pipeline, Jupyter shows an **Error making request** message that describes failed initialization.

  Workaround

  A user with administrator permissions (for example, the project owner) must create the workbench on behalf of the user with edit permissions. That user can then use the workbench to run Elyra pipelines.

RHOAIENG-583 (previously documented as RHODS-8921 and RHODS-6373) – You cannot create a pipeline server or start a workbench when cumulative character limit is exceeded

When the cumulative character limit of a data science project name and a pipeline server name exceeds 62 characters, you are unable to successfully create a pipeline server. Similarly, when the cumulative character limit of a data science project name and a workbench name exceeds 62 characters, workbenches fail to start.

Workaround

Rename your data science project so that it does not exceed 30 characters.

RHODS-7718 – User without dashboard permissions is able to continue using their running notebooks and workbenches indefinitely

When a Red Hat OpenShift AI administrator revokes a user's permissions, the user can continue to use their running notebooks and workbenches indefinitely.

Workaround

When the OpenShift AI administrator revokes a user's permissions, the administrator should also stop any running notebooks and workbenches for that user.

RHOAIENG-1157 (previously documented asRHODS-6955) – An error can occur when trying to edit a workbench

When editing a workbench, an error similar to the following can occur:

> Error creating workbench
> Operation cannot be fulfilled on notebooks.kubeflow.org "workbench-name": the object has been modified; please apply your changes to the latest version and try again

RHOAIENG-1132 (previously documented asRHODS-6383) – An ImagePullBackOff error message is not displayed when required during the workbench creation process

Pods can experience issues pulling container images from the container registry. If an error occurs, the relevant pod enters into an **ImagePullBackOff** state. During the workbench creation process, if an **ImagePullBackOff** error occurs, an appropriate message is not displayed.

Workaround

Check the event log for further information on the **ImagePullBackOff** error. To do this, click on the workbench status when it is starting.

RHOAIENG-1152 (previously documented asRHODS-6356) – The notebook creation process fails for users who have never logged in to the dashboard

The dashboard's notebook **Administration** page displays users belonging to the user group and admin group in OpenShift. However, if an administrator attempts to start a notebook server on behalf of a user who has never logged in to the dashboard, the server creation process fails and displays the following error message:

> Request invalid against a username that does not exist.

Workaround

Request that the relevant user logs into the dashboard.

RHODS-5763 – Incorrect package version displayed during notebook selection

The **Start a notebook server**page displays an incorrect version number for the Anaconda notebook image.

RHODS-5543 – When using the NVIDIA GPU Operator, more nodes than needed are created by the Node Autoscaler

When a pod cannot be scheduled due to insufficient available resources, the Node Autoscaler creates a new node. There is a delay until the newly created node receives the relevant GPU workload. Consequently, the pod cannot be scheduled and the Node Autoscaler's continuously creates additional new nodes until one of the nodes is ready to receive the GPU workload. For more information about this issue, see the Red Hat Knowledgebase solution When using the NVIDIA GPU Operator, more nodes than needed are created by the Node Autoscaler.

Workaround

Apply the **cluster-api/accelerator** label in **machineset.spec.template.spec.metadata**. This causes the autoscaler to consider those nodes as unready until the GPU driver has been deployed.

RHOAIENG-1137 (previously documented as RHODS-5251) – Notebook server administration page shows users who have lost permission access

If a user who previously started a notebook server in Jupyter loses their permissions to do so (for example, if an OpenShift AI administrator changes the user's group settings or removes the user from a permitted group), administrators continue to see the user's notebook servers on the server **Administration** page. As a consequence, an administrator is able to restart notebook servers that belong to the user who's permissions were revoked.

RHODS-4799 – Tensorboard requires manual steps to view

When a user has TensorFlow or PyTorch notebook images and wants to use TensorBoard to display data, manual steps are necessary to include environment variables in the notebook environment, and to import those variables for use in your code.

Workaround

When you start your notebook server, use the following code to set the value for the TENSORBOARD_PROXY_URL environment variable to use your OpenShift AI user ID.

```
import os
os.environ["TENSORBOARD_PROXY_URL"]= os.environ["NB_PREFIX"]+"/proxy/6006/"
```

RHODS-4718 – The Intel® oneAPI AI Analytics Toolkits quick start references nonexistent sample notebooks

The Intel® oneAPI AI Analytics Toolkits quick start, located on the **Resources** page on the dashboard, requires the user to load sample notebooks as part of the instruction steps, but refers to notebooks that do not exist in the associated repository.

RHODS-4627 – The CronJob responsible for validating Anaconda Professional Edition's license is suspended and does not run daily

The CronJob responsible for validating Anaconda Professional Edition's license is automatically suspended by the OpenShift AI operator. As a result, the CronJob does not run daily as scheduled. In addition, when Anaconda Professional Edition's license expires, Anaconda Professional Edition is not indicated as disabled on the OpenShift AI dashboard.

RHOAIENG-1141 (previously documented as RHODS-4502) – The NVIDIA GPU Operator tile on the dashboard displays button unnecessarily

GPUs are automatically available in Jupyter after the NVIDIA GPU Operator is installed. The **Enable** button, located on the NVIDIA GPU Operator tile on the **Explore** page, is therefore redundant. In addition, clicking the **Enable** button moves the NVIDIA GPU Operator tile to the **Enabled** page, even if the Operator is not installed.

RHOAIENG-1135 (previously documented as RHODS-3985) – Dashboard does not display Enabled page content after ISV operator uninstall

After an ISV operator is uninstalled, no content is displayed on the **Enabled** page on the dashboard. Instead, the following error is displayed:

```
Error loading components
HTTP request failed
```

Workaround

Wait 30-40 seconds and then refresh the page in your browser.

RHODS-3984 - Incorrect package versions displayed during notebook selection

In the OpenShift AI interface, the **Start a notebook server page**displays incorrect version numbers for the JupyterLab and Notebook packages included in the oneAPI AI Analytics Toolkit notebook image. The page might also show an incorrect value for the Python version used by this image.

Workaround

When you start your oneAPI AI Analytics Toolkit notebook server, you can check which Python packages are installed on your notebook server and which version of the package you have by running the **!pip list** command in a notebook cell.

RHODS-2956 - Error can occur when creating a notebook instance

When creating a notebook instance in Jupyter, a **Directory not found** error appears intermittently. This error message can be ignored by clicking **Dismiss**.

RHOAING-1147 (previously documented as RHODS-2881) - Actions on dashboard not clearly visible

The dashboard actions to revalidate a disabled application license and to remove a disabled application tile are not clearly visible to the user. These actions appear when the user clicks on the application tile's **Disabled** label. As a result, the intended workflows might not be clear to the user.

RHOAIENG-1134 (previously documented as RHODS-2879) - License revalidation action appears unnecessarily

The dashboard action to revalidate a disabled application license appears unnecessarily for applications that do not have a license validation or activation system. In addition, when a user attempts to revalidate a license that cannot be revalidated, feedback is not displayed to state why the action cannot be completed.

RHOAIENG-2305 (previously documented as RHODS-2650) - Error can occur during Pachyderm deployment

When creating an instance of the Pachyderm operator, a webhook error appears intermittently, preventing the creation process from starting successfully. The webhook error is indicative that, either the Pachyderm operator failed a health check, causing it to restart, or that the operator process exceeded its container's allocated memory limit, triggering an Out of Memory (OOM) kill.

Workaround

Repeat the Pachyderm instance creation process until the error no longer appears.

RHODS-2096 - IBM Watson Studio not available in OpenShift AI

IBM Watson Studio is not available when OpenShift AI is installed on OpenShift Dedicated 4.9 or higher, because it is not compatible with these versions of OpenShift Dedicated. Contact Marketplace support for assistance manually configuring Watson Studio on OpenShift Dedicated 4.9 and higher.

RHODS-1888 - OpenShift AI hyperlink still visible after uninstall

When the OpenShift AI Add-on is uninstalled from an OpenShift Dedicated cluster, the link to the OpenShift AI interface remains visible in the application launcher menu. Clicking this link results in a "Page Not Found" error because OpenShift AI is no longer available.

CHAPTER 7. PRODUCT FEATURES

Red Hat OpenShift AI provides a rich set of features for data scientists and IT operations administrators. To learn more, see Introduction to Red Hat OpenShift AI .