Red Hat OpenShift Data Science 1

Getting started with Red Hat OpenShift Data Science

Learn how to work in an OpenShift Data Science environment

Last Updated: 2021-12-14

Red Hat OpenShift Data Science 1 Getting started with Red Hat OpenShift Data Science

Learn how to work in an OpenShift Data Science environment

Legal Notice

Abstract

Log in and start up your notebook server to get started working with your notebooks in JupyterHub.

Table of Contents

PREFACE

This documentation is provided for the Field Trial release of Red Hat OpenShift Data Science.

See the following documents for service and life cycle information related to this Field Trial release:

- OpenShift Data Science Service Definition
- OpenShift Data Science Life Cycle

CHAPTER 1. LOGGING IN TO OPENSHIFT DATA SCIENCE

Log in to OpenShift Data Science from a browser for easy access to JupyterHub and your data science projects.

Procedure

1. Browse to the OpenShift Data Science instance URL and click **Log in with OpenShift**.
   - If you are a data scientist user, your administrator must provide you with the OpenShift Data Science instance URL, for example, **https://rhods-dashboard-redhat-ods-applications.apps.example.abc1.p1.openshiftapps.com/**.
   - If you have access to OpenShift Dedicated, you can browse to the OpenShift Dedicated web console and click the **Application Launcher** () → **Red Hat OpenShift Data Science**.
2. Click the name of your identity provider, for example, **GitHub**.
3. Enter your credentials and click **Log in** (or equivalent for your identity provider).
   If you have not previously authorized the **rhods-dashboard** service account to access your account, the **Authorize Access** page appears prompting you to provide authorization. Inspect the permissions selected by default, and click the **Allow selected permissions** button.

Verification

- OpenShift Data Science opens on the **Enabled applications** page.

Troubleshooting

- If you see **An authentication error occurred** or **Could not create user** when you try to log in:
  - You might have entered your credentials incorrectly. Confirm that your credentials are correct.
  - You might have an account in more than one configured identity provider. If you have logged in with a different identity provider previously, try again with that identity provider.

Additional resources

- Launching JupyterHub and starting a notebook server

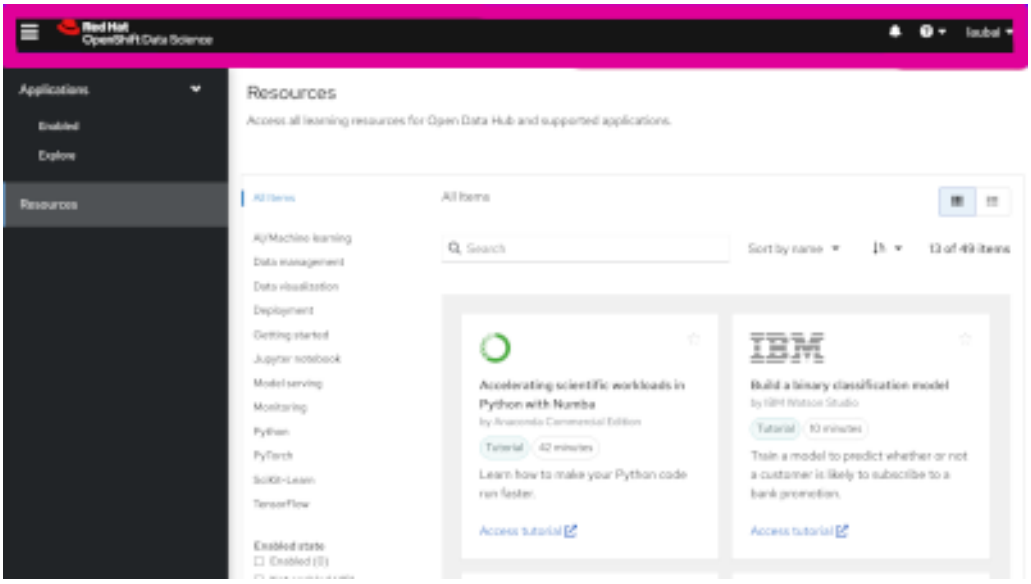
CHAPTER 2. THE OPENSHIFT DATA SCIENCE USER INTERFACE

The Red Hat OpenShift Data Science interface is based on the OpenShift web console user interface.

The OpenShift Data Science user interface is divided into several areas:

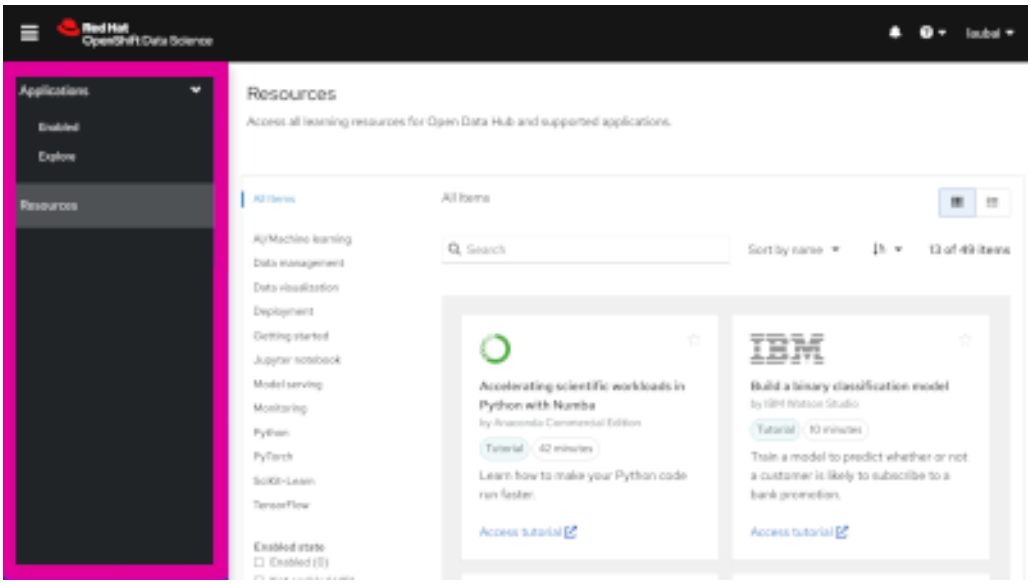
- The global navigation bar, which provides access to useful controls, such as **Help** and **Notifications**.

Figure 2.1. The global navigation bar



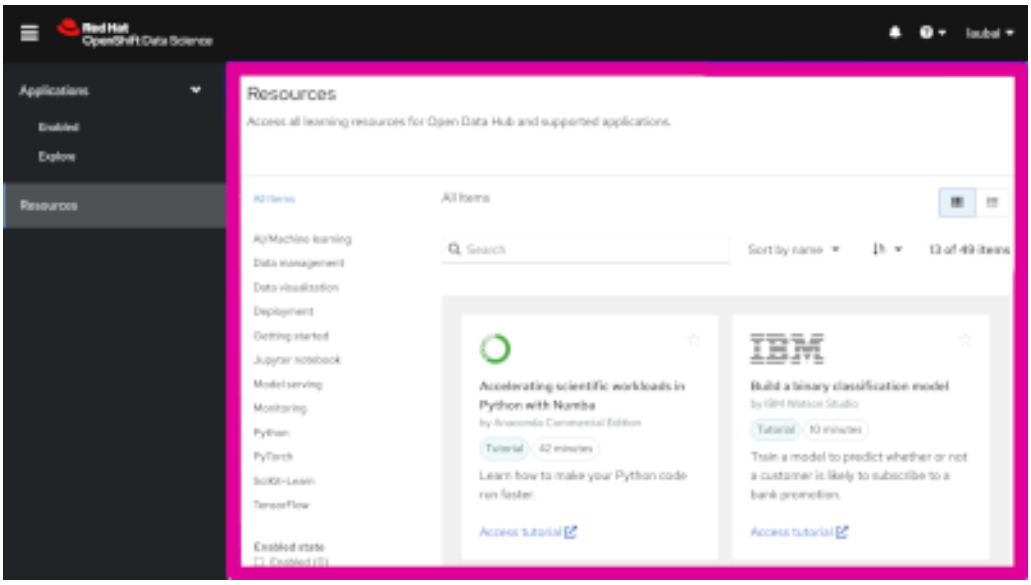
- The side navigation menu, which contains different categories of pages available in OpenShift Data Science.

Figure 2.2. The side navigation menu



- The main display area, which displays the current page and shares space with any drawers currently displaying information, such as notifications or quick start guides.

Figure 2.3. The main display area



2.1. GLOBAL NAVIGATION

There are four items in the top navigation:

- The **Toggle side navigation menu** button ( ) toggles whether or not the side navigation is displayed.
- The **Notifications** button ( ) opens and closes the **Notifications** drawer, letting you read current and previous notifications in more detail.
- The **Help** menu ( ) provides a link to create a ticket with Red Hat Support and access the OpenShift Data Science documentation.
- The **User** menu displays the name of the currently logged-in user and provides access to the **Log out** button.

2.2. SIDE NAVIGATION

There are three main sections in the side navigation:

Applications → Enabled

The **Enabled** page displays applications that are enabled and ready to use on OpenShift Data Science. JupyterHub is the only application installed by default. This page is the default landing page for OpenShift Data Science.

Click the **Launch** button on an application card to open the application interface in a new tab. Some applications also have **Quick start** links so that you have direct access to the quick start tour for that application.

Applications → Explore

The **Explore** page displays applications that are available for use with OpenShift Data Science. Click on a card for more information about the application or to access the **Enable** button. The **Enable** button is only visible if your administrator has purchased and enabled an application at the OpenShift Dedicated level.

Resources

The **Resources** page displays learning resources such as documentation, how-to material, and quick start tours. You can filter visible resources using the options displayed on the left, or enter terms into the search bar.

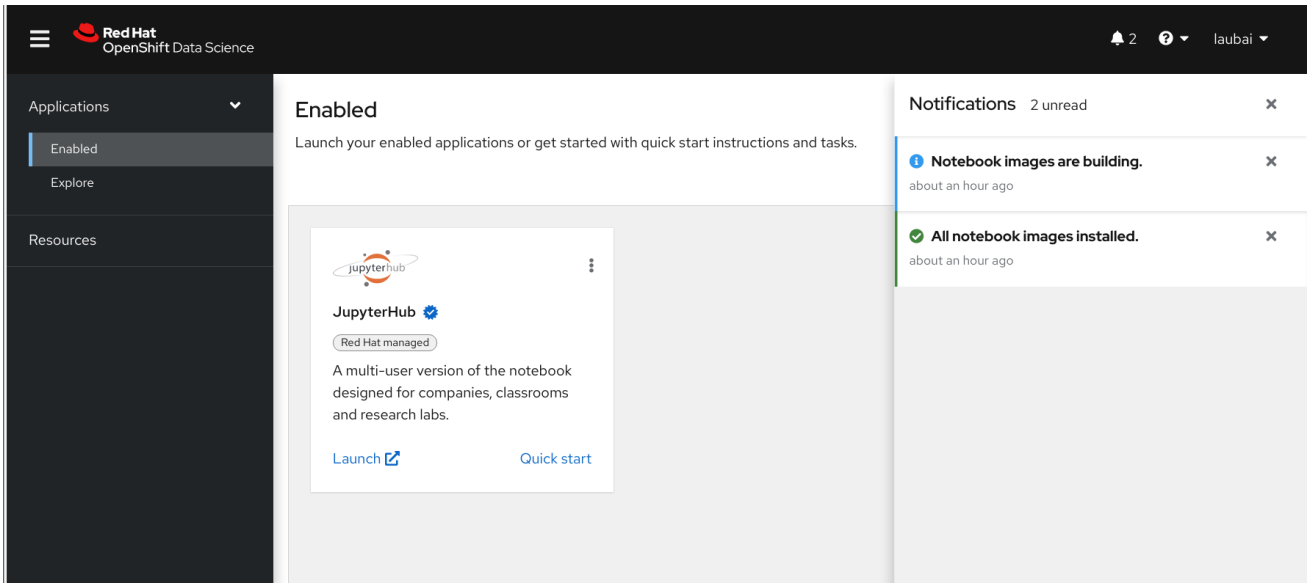
CHAPTER 3. NOTIFICATIONS IN OPENSHIFT DATA SCIENCE

Red Hat OpenShift Data Science displays notifications when important events happen in the cluster.

Notification messages are displayed in the lower left corner of the Red Hat OpenShift Data Science interface when they are triggered.

If you miss a notification message, click the **Notifications** button () to open the **Notifications** drawer and view unread messages.

Figure 3.1. The Notifications drawer



CHAPTER 4. LAUNCHING JUPYTERHUB AND STARTING A NOTEBOOK SERVER

Launch JupyterHub and start a notebook server to start working with your notebooks.

Prerequisites

- You have logged in to Red Hat OpenShift Data Science.
- You know the names and values you want to use for any environment variables in your notebook server environment, for example, **AWS_SECRET_ACCESS_KEY**.
- If you want to work with a very large data set, work with your administrator to proactively increase the storage capacity of your notebook server.

Procedure

1. Locate the **JupyterHub** card on the **Enabled applications** page.
2. Click **Launch**.
   a. If prompted, select your identity provider.
   b. Enter your credentials and click **Log in** (or equivalent for your identity provider).
   If you see **Error 403: Forbidden**, you are not in the default user group or the default administrator group for OpenShift Data Science. Contact your administrator so that they can add you to the correct group using Adding users for OpenShift Data Science .

   If you have not previously authorized the **jupyterhub-hub** service account to access your account, the **Authorize Access** page appears prompting you to provide authorization. Inspect the permissions selected by default, and click the **Allow selected permissions** button.
3. Start a notebook server.
   This is not required if you have previously launched JupyterHub.
   a. Select the **Notebook image** to use for your server.
   b. If the notebook image contains multiple versions, select the version of the notebook image from the **Versions** section.

NOTE

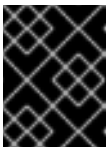
When a new version of a notebook image is released, the previous version remains available and supported on the cluster. This gives you to time to migrate your work to the latest version of the notebook image.

Notebook images can take up to 40 minutes to install. Notebooks images that have not finished installing are not available for you to select. If an installation of a notebook image has not completed, an alert is displayed.

   c. Select the **Container size** for your server.
   d. Optional: Select and specify values for any new **Environment variables**.

For example, if you plan to integrate with Red Hat OpenShift Streams for Apache Kafka, create environment variables to store your Kafka bootstrap server and the service account username and password here.

The interface stores these variables so that you only need to enter them once. Example variable names for common environment variables are automatically provided for frequently integrated environments and frameworks, such as Amazon Web Services (AWS).



IMPORTANT

Ensure that you select the **Secret** checkbox for any variables with sensitive values that must be kept private, such as passwords.

e. Click **Start server**.
The **Starting server** progress indicator appears. Depending on the deployment size and resources you requested, starting the server can take up to several minutes. After the server starts, the JupyterLab interface opens.

Verification

- The JupyterLab interface opens in a new tab.

Additional resources

- Options for notebook server environments
- Troubleshooting common problems in JupyterHub

Troubleshooting

- If you see the "Unable to load notebook server configuration options" error message, contact your administrator so that they can review the logs associated with your JupyterHub pod and determine further details about the problem.

4.1. OPTIONS FOR NOTEBOOK SERVER ENVIRONMENTS

When you launch JupyterHub for the first time, or after stopping your notebook server, you must select server options in the **Start a notebook server** wizard so that the software and variables that you expect are available on your server. This section explains the options available in the **Start a notebook server** wizard in detail.

The **Start a notebook server** page is divided into several sections:

Notebook image

Specifies the container image that your notebook server is based on. Different notebook images have different packages installed by default. See Notebook image options for details.

Deployment size

Specifies the compute resources available on your notebook server.
Container size controls the number of CPUs, the amount of memory, and the maximum request capacity of the container.

Environment variables

Specifies the name and value of variables to be set on the notebook server. Setting environment

variables during server startup means that you do not need to define them in the body of your notebooks, or with the JupyterHub command line interface. See Recommended environment variables for a list of reserved variable names for each item in the **Environment variables** list.

Table 4.1. Notebook image options

Image name	Preinstalled packages
Minimal Python	• Python 3.8.8 • JupyterLab 3.0.14 • Notebook 6.0.3
PyTorch	• Python 3.8.8 • JupyterLab 3.0.16 • Notebook 6.4.4 • PyTorch 1.8.1 • Tensorboard 1.15.0 • Boto3 1.17.11 • Kafka-Python 2.0.2 • Matplotlib 3.4.1 • Numpy 1.19.2 • Pandas 1.2.4 • Scikit-learn 0.24.1 • Scipy 1.6.2
Standard Data Science	• Python 3.8.6 • JupyterLab 3.0.16 • Notebook 6.4.0 • Boto3 1.17.11 • Kafka-Python 2.0.2 • Matplotlib 3.4.2 • Pandas 1.2.5 • Numpy 1.21.0 • Scikit-learn 0.24.2 • Scipy 1.7.0

Image name	Preinstalled packages
TensorFlow	• Python 3.8.8 • JupyterLab 3.0.16 • Notebook 6.4.4 • TensorFlow 2.4.1 • Tensorboard 2.4.1 • Boto3 1.17.11 • Kafka-Python 2.0.2 • Matplotlib 3.4.1 • Numpy 1.19.2 • Pandas 1.2.4 • Scikit-learn 0.24.1 • Scipy 1.6.2

Table 4.2. Recommended environment variables

Environment variable option	Recommended variable names

Environment variable option	Recommended variable names
AWS	• **AWS_ACCESS_KEY_ID** specifies your Access Key ID for Amazon Web Services. • **AWS_SECRET_ACCESS_KEY** specifies your Secret access key for the account specified in **AWS_ACCESS_KEY_ID**.

Additional resources

- Launching JupyterHub and starting a notebook server

CHAPTER 5. TUTORIALS FOR DATA SCIENTISTS

To help you get started quickly, you can access learning resources for Red Hat OpenShift Data Science and its supported applications. These resources are available on the **Resources** tab of the Red Hat OpenShift Data Science user interface.

Table 5.1. Tutorials

Resource Name	Description
Accelerating scientific workloads in Python with Numba	Make Python data science code work faster.
Build a binary classification model	Train a model to predict whether a customer is likely to subscribe to a bank promotion.
Build interactive visualizations and dashboards in Python	Build simple and complex figures. Work with billions of data points, add interactive behavior, widgets, and controls. Deploy full dashboards and applications.
Data analysis in Python with Pandas	Learn how to use pandas, a data analysis library for Python.
Machine learning in Python with Scikit-learn	Build machine learning models with scikit-learn for supervised learning, unsupervised learning, and classification problems.
Python tools for data visualization	An open platform to help you choose open-source (OSS) Python data visualization tools.
Run a Python notebook to generate results in IBM Watson OpenScale	Run a Python notebook to create, train, and deploy a machine learning model.
Run an AutoAI experiment to build a model	Learn how to build a binary classification model for a marketing campaign with IBM Watson Machine Learning in IBM Cloud Pak for Data.
Scalable computing in Python with Dask	Learn Dask and parallel data analysis. Analyze medium-sized data sets in parallel.
Using Jupyter notebooks in IBM Watson Studio	The basics on working with Jupyter notebooks in IBM Watson Studio.
Various types of model servings	A series of tutorials highlighting models, natural language, image processing classifiers, and KubeFlow integration.
What is Anaconda?	Learn about the Anaconda Distribution and why Python is so frequently used for data science tasks.

Table 5.2. Quick start guides

Resource Name	Description
Connecting to Red Hat OpenShift Streams for Apache Kafka	Connect to Red Hat Streams for Apache Kafka from a Jupyter notebook.
Creating a Jupyter notebook	To support your data science work, you can create a Jupyter notebook from an existing notebook container image to access its resources and properties.
Creating an Anaconda-enabled Jupyter notebook	With an Anaconda-enabled Jupyter notebook, you can access and use Anaconda packages, curated for security and compatibility.
Deploying a model with IBM Watson Studio	Import a notebook in IBM Watson Studio, build a model with AutoAI, and then deploy it.
Deploying a sample Python application using Flask and OpenShift	Deploy a Python model using Flask and OpenShift.
Intel® oneAPI AI Analytics Toolkit (AI Kit) Notebook	Learn how to use the Intel oneAPI AI Analytics Toolkit.
Launch a SKLearn model and update model by canarying	Perform a canary promotion of a Scikit-Learn model.
Monitor a deployed model for drift	Monitor drift for an image classifier model.
Optimized Inference Notebook	Learn how to use the OpenVINO toolkit.
Securing a deployed model using Red Hat OpenShift API Management	Learn how to protect a model service API using Red Hat OpenShift API Management.
See outlier scores for predictions to deployed model	View outlier scores for predictions to a deployed image classifier model.
See predictions and explanations for a deployed SKLearn model	View predictions and explanations for deployed income classifier model.

Table 5.3. How to guides

Resource Name	Description
How to choose between notebook runtime environment options	Understand the options available when you configure a notebook run-time environment in IBM Watson Studio.

Resource Name	Description
How to clean, shape, and visualize my data	Clean and shape tabular data using IBM Watson Studio data refinery.
How to create a connection to access data	Create connections to various data sources across the platform in IBM Watson Studio.
How to create a deployment space for deploying my model	Create a deployment space for IBM Watson Machine Learning.
How to create a notebook in IBM Watson Studio	Create a basic Jupyter notebook in IBM Watson Studio.
How to create a project in Watson Studio	Create an analytics project in IBM Watson Studio.
How to create a project that integrates with Git	How to add assets from a Git repository into a project in IBM Watson Studio.
How to install Python packages on your notebook server	Install additional Python packages into your notebook server.
How to load data in a notebook	Learn how to integrate data sources into a Jupyter notebook using IBM Watson Studio.
How to perform operational tasks	Best practices for managing a Seldon Deploy cluster.
How to serve a model using OpenVINO Model Server	Deploy optimized models with OpenVINO Model Server using OpenVINO custom resources.
How to set up IBM Watson OpenScale	Learn how to track and measure outcomes from models with IBM Watson OpenScale.
How to update notebook server settings	Learn how to restart your notebook server with a different notebook image.
How to use data from Amazon S3 buckets	Connect to Data in S3 Storage using environment variables.
How to view installed packages on your notebook server	You can view an alphabetical list of the Python packages that are installed on your notebook server along with the version of the packages.

5.1. ACCESSING TUTORIALS

You can access learning resources for Red Hat OpenShift Data Science and supported applications.

Prerequisites

- Ensure that you have logged in to Red Hat OpenShift Data Science.
- You have logged in to the OpenShift Dedicated web console.

Procedure

1. On the Red Hat OpenShift Data Science home page, click **Resources**.
   The **Resources** page opens.
2. Click **Access Tutorial** on the relevant card.

Verification

- You can view and access the learning resources for Red Hat OpenShift Data Science and supported applications.

Additional resources

- Common questions

CHAPTER 6. ENABLING SERVICES CONNECTED TO OPENSHIFT DATA SCIENCE

You must enable SaaS-based services, such as Red OpenShift Streams for Apache Kafka and Anaconda, before using them with Red Hat OpenShift Data Science. On-cluster services are enabled automatically.

For most services, the service endpoint is available on the service's tile on the **Enabled** page of OpenShift Data Science. Certain services cannot be accessed directly from their tiles, for example, OpenVINO and Anaconda provide notebook images for use in JupyterHub and do not provide an endpoint link from their tile. Additionally, for services such as OpenShift Streams for Apache Kafka, it may be useful to store these endpoint URLs as environment variables for easy reference in a notebook environment.

To help you get started quickly, you can access the service's learning resources and documentation on the **Resources** page, or by clicking the relevant link on the service's tile on the **Enabled** page.

Prerequisites

- You have logged in to OpenShift Data Science.
- Your administrator has installed or configured the service on your OpenShift Dedicated cluster.

Procedure

1. On the OpenShift Data Science home page, click **Explore**.
   The **Explore** page opens.
2. Click the card of the service that you want to enable.
3. Click **Enable** on the drawer for the service.
4. If prompted, enter the service's key and click **Connect**.
5. Clink **Enable** to confirm service enablement.

Verification

- The service that you enabled appears on the **Enabled** page.
- The service endpoint is displayed on the service's tile on the **Enabled** page

CHAPTER 7. SUPPORT REQUIREMENTS AND LIMITATIONS

Review this section to understand the requirements for Red Hat support and any limitations to Red Hat support of Red Hat OpenShift Data Science.

7.1. SUPPORTED BROWSERS

Red Hat OpenShift Data Science supports the latest version of the following browsers:

- Google Chrome
- Mozilla Firefox
- Safari

7.2. SUPPORTED SERVICES

Red Hat OpenShift Data Science supports the following services:

Table 7.1. Supported services

Service Name	Description
Anaconda Commercial Edition	A toolkit containing a distribution of Python and R. Anaconda simplifies package management and contains many important data science packages and libraries.
IBM Watson Studio	A software platform that allows you to build, run, and manage AI models.
Intel® oneAPI AI Analytics Toolkit	A set of integrated AI software tools built using Intel oneAPI. Intel oneAPI provides drop-in acceleration for end-to-end data science and machine learning pipelines on Intel architectures with familiar Python libraries and frameworks.
JupyterHub	A Red Hat managed application that allows you to configure a notebook server environment and develop machine learning models in JupyterLab.  **IMPORTANT** While every effort is made to make Red Hat OpenShift Data Science resilient to OpenShift node failure, upgrades, and similarly disruptive operations, individual users' notebook environments can be interrupted during these events. If an OpenShift node restarts or becomes unavailable, any user notebook environment on that node is restarted on a different node. When this occurs, any ongoing process executing in the user's notebook environment is interrupted, and the user needs to re-execute it when their environment becomes available again. Due to this limitation, Red Hat recommends that processes for which interruption is unacceptable are not executed in the JupyterHub notebook server environment on OpenShift Data Science. For example, rather than serve a data science model from within a notebook server on OpenShift Data Science, we recommend leveraging the model serving capabilities of Seldon instead.

Service Name	Description
Red Hat OpenShift API Management	A Red Hat managed service that makes it easy to secure, share, and control access to APIs for services, applications, and enterprise systems across public and private clouds.
Red Hat OpenShift Streams for Apache Kafka	A Red Hat managed cloud service that allows you to create, discover, and connect your data science projects to real-time data streams.
OpenVINO	The Open Visual Inference and Neural Network Optimization toolkit optimizes the performance of neural network inference on Intel hardware.
Seldon Deploy	A framework that allows you to deploy machine learning models quickly and efficiently at scale.
Starburst Galaxy (Beta)	A SQL-based massively parallel processing (MPP) query engine allowing you to run analytics on data wherever it is stored, reducing the time required to access the data.

7.3. SUPPORTED PACKAGES

Notebook server images in Red Hat OpenShift Data Science are installed with Python 3.8 by default. See the table in Options for notebook server environments for a complete list of packages and versions included in these images.

You can install packages that are compatible with Python 3.8 on any notebook server that has the binaries required by that package. If the required binaries are not included on the notebook server image you want to use, contact Red Hat Support to request that the binary be considered for inclusion.

You can install packages on a temporary basis by using the **pip install** command. You can also provide a list of packages to the **pip install** command using a **requirements.txt** file. See Installing Python packages on your notebook server for more information.

You must re-install these packages each time you start your notebook server.

You can remove packages by using the **pip uninstall** command.

Additional resources

- Installing Python packages on your notebook server
- Options for notebook server environments

CHAPTER 8. COMMON QUESTIONS

In addition to documentation, Red Hat provides a number of "how to" documents that answer common questions a data scientist might have as they work.

The currently available "how to" documents are linked here:

- How to update notebook server settings
- https://access.redhat.com/documentation/en-us/red_hat_openshift_data_science/1/html/working_with_notebooks/working-with-notebooks-on-openshift-data-science_nb-server#installing-python-packages-on-your-notebook-server_nb-serverHow to install python packages on your notebook server
- How to view installed packages on your notebook server
- How to use data from Amazon S3 buckets
- How to use data from Kafka

CHAPTER 9. TROUBLESHOOTING COMMON PROBLEMS IN JUPYTERHUB

If you are seeing errors in Red Hat OpenShift Data Science related to JupyterHub, your notebooks, or your notebook server, read this section to understand what could be causing the problem.

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

9.1. I SEE A 403: FORBIDDEN ERROR WHEN I LOG IN TO JUPYTERHUB

Problem

Your user name might not be added to the default user group or the default administrator group for OpenShift Data Science. Contact your administrator so that they can add you to the correct group/s.

Diagnosis

Check whether the user is part of either the default user group or the default administrator group.

1. Find the names of groups allowed access to JupyterHub.
   a. Log in to OpenShift Dedicated web console.
   b. Click **Workloads → ConfigMaps** and click on the **rhods-groups-config** ConfigMap to open it.
   c. Click on the YAML tab and check the values for **admin_groups** and **allowed_groups**. These are the names of groups that have access to JupyterHub.

```
data:
  admin_groups: rhods-admins
  allowed_groups: rhods-users
```

2. Click **User management → Groups** and click on the name of each group to see its members.

Resolution

- If the user is not added to any of the groups allowed access to JupyterHub, follow Adding users for OpenShift Data Science to add them.
- If the user is already added to a group that is allowed to access JupyterHub, contact Red Hat Support.

9.2. MY NOTEBOOK SERVER DOES NOT START

Problem

The OpenShift Dedicated cluster that hosts your notebook server might not have access to enough resources, or the JupyterHub pod may have failed. Contact your administrator so that they can perform further checks.

Diagnosis

1. Log in to OpenShift Dedicated web console.

2. Delete and restart the notebook server pod for this user.

    a. Click **Workloads → Pods** and set the **Project** to **rhods-notebooks**.

    b. Search for the notebook server pod that belongs to this user exists, for example, **jupyterhub-nb-username-***.
    If the notebook server pod exists, an intermittent failure may have occurred in the notebook server pod.

    If the notebook server pod for the user does not exist, continue with diagnosis.

3. Check the resources currently available in the OpenShift Dedicated cluster against the resources required by the selected notebook server image.
If worker nodes with sufficient CPU and RAM are available for scheduling in the cluster, continue with diagnosis.

4. Check the state of the JupyterHub pod.

Resolution

- If there was an intermittent failure of the notebook server pod:

    a. Delete the notebook server pod that belongs to the user.

    b. Ask the user to start their notebook server again.

- If the notebook server does not have sufficient resources to run the selected notebook server image, either add more resources to the OpenShift Dedicated cluster, or choose a smaller image size.

- If the JupyterHub pod is in a **FAILED** state:

    a. Retrieve the logs for the **jupyterhub-*** pod and send them to Red Hat Support for further evaluation.

    b. Delete the **jupyterhub-*** pod.

WARNING

Ensure that you delete the correct pod. Do not delete the **jupyterhub-db-*** pod by mistake.

- If none of the previous resolutions apply, contact Red Hat Support.

9.3. I SEE A DATABASE OR DISK IS FULL ERROR OR A NO SPACE LEFT ON DEVICE ERROR WHEN I RUN MY NOTEBOOK CELLS

Problem

You might have run out of storage space on your notebook server. Contact your administrator so that they can perform further checks.

Diagnosis

1. Log in to JupyterHub and start the notebook server that belongs to the user having problems. If the notebook server does not start,
2. Check whether the user has run out of storage space.
    a. Log in to OpenShift Dedicated web console.
    b. Click **Workloads → Pods** and set the **Project** to **rhods-notebooks**.
    c. Click the notebook server pod that belongs to this user, for example, **jupyterhub-nb-username-***.
    d. Click **Logs**. The user has exceeded their available capacity if you see lines similar to the following:

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

Resolution

- Increase the user's available storage by expanding their persistent volume: Expanding persistent volumes
- Work with the user to identify files that can be deleted from the **/opt/app-root/src** directory to free up their existing storage space.