Red Hat OpenShift Data Science 1

Introduction to Red Hat OpenShift Data Science
Abstract

Red Hat OpenShift Data Science is a platform for data scientists and developers of artificial intelligence and machine learning applications.
# Table of Contents

CHAPTER 1. OVERVIEW OF OPENSFIFT DATA SCIENCE ...................................................... 3

CHAPTER 2. SUPPORTED CONFIGURATIONS ................................................................. 4

CHAPTER 3. PRODUCT FEATURES ................................................................................. 5
  3.1. FEATURES FOR DATA SCIENTISTS ............................................................... 5
  3.2. FEATURES FOR IT OPERATIONS ADMINISTRATORS ........................................ 5

CHAPTER 4. TRY IT ........................................................................................................ 7

CHAPTER 5. GET IT ......................................................................................................... 8
CHAPTER 1. OVERVIEW OF OPENSSHIFT DATA SCIENCE

Red Hat OpenShift Data Science is a platform for data scientists and developers of artificial intelligence (AI) and machine learning (ML) applications. It is built from a curated set of components from the Open Data Hub open source project.

OpenShift Data Science provides a fully supported environment that lets you develop, train, test, deploy and monitor your machine learning models on-premises and/or in the cloud.

OpenShift Data Science provides a wide range of tools for data scientists. It comes with several Jupyter notebook images as default, which include popular frameworks such as PyTorch and TensorFlow. Moreover, you have the flexibility to expand this collection by incorporating your own custom notebook images. If you require more than just notebooks, you can also take advantage of Data Science Pipelines to parallelize and schedule your workloads efficiently.

Once your model is ready for deployment, OpenShift Data Science offers model serving capabilities. You can use the default OpenVINO runtime or create your own custom runtime to serve your models. By utilizing the Data Science Projects feature, you can organize these artifacts in a meaningful way, facilitating efficient collaboration among team members.

Furthermore, OpenShift Data Science provides convenient access to GPUs, ensuring that data scientists have the accelerators they require.

For IT operations administrators, OpenShift Data Science provides a secure and supported environment that can run on top of any OpenShift cluster whether that cluster is in the cloud, on-premise, or even fully disconnected.
CHAPTER 2. SUPPORTED CONFIGURATIONS

OpenShift Data Science is supported in two configurations:

- **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 Data Science on a Red Hat managed environment, see [Product Documentation for Red Hat OpenShift Data Science](#).

- **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 Data Science as self-managed software on your OpenShift cluster in a connected or a disconnected environment, see [Product Documentation for Red Hat OpenShift Data Science self-managed](#).
CHAPTER 3. PRODUCT FEATURES

Red Hat OpenShift Data Science provides several features for data scientists and IT operations administrators.

3.1. FEATURES FOR DATA SCIENTISTS

Containers
While tools like JupyterLab already offer intuitive ways for data scientists to develop models on their machines, there are always inherent complexities involved with collaboration and sharing work. Moreover, using specialized hardware such as powerful GPUs can be very expensive when you have to buy and maintain your own. The Jupyter environment that is included with OpenShift Data Science lets you take your development environments anywhere you need it to be. Because all of the workloads are run as containers, collaboration is as easy as sharing an image with your team members, or even simply adding it to the list of default containers that they can use. GPUs and large amounts of memory suddenly become a lot more accessible, too, since you are no longer limited by what your laptop can support.

Integration with third-party machine learning tools
We have all run into situations where our favorite tools or services don’t play well with one another. OpenShift Data Science is designed with flexibility in mind. You can use a wide range of open source and third-party tools with OpenShift Data Science. These tools support the complete machine learning lifecycle, from data engineering and feature extraction to model deployment and management.

Collaboration on notebooks with Git
Use Jupyter’s Git interface to work collaboratively with others, and keep good track of the changes to your code.

Securely built notebook images
Choose from a default set of notebook images that are pre-configured with the tools and libraries that you need for model development. Software stacks, especially those involved in machine learning, tend to be complex systems. There are numerous modules and libraries in the Python ecosystem that can be used, so determining which versions of what libraries to use can be very challenging. OpenShift Data Science comes with many packaged notebook images that have been built with insight from data scientists and recommendation engines. You can start new projects quickly on the right foot without worrying about downloading unproven and possibly insecure images from random upstream repositories.

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

Data science pipelines
OpenShift Data Science supports data science pipelines for a mature and efficient way of running your data science workloads. You can standardize and automate machine learning workflows that enable you to develop and deploy your data science models.

Model serving
As a data scientist, you can deploy your trained machine-learning models to serve intelligent applications in production. Deploying or serving a model makes the model’s functions available as a service endpoint that can be used for testing or integration into applications. You have much control over how this serving is performed.

3.2. FEATURES FOR IT OPERATIONS ADMINISTRATORS
Manage users with an identity provider

OpenShift Data Science supports the same authentication systems as your OpenShift cluster. By default, OpenShift Data Science is accessible to all users listed in your identity provider and those users do not need a separate set of credentials to access OpenShift Data Science. Optionally, you can limit the set of users who have access by creating an OpenShift group that specifies a subset of users. You can also create an OpenShift group that identifies the list of users who have administrator access to OpenShift Data Science.

Manage resources with OpenShift

Use your existing OpenShift knowledge to configure and manage resources for your OpenShift Data Science users.

Control Red Hat usage data collection

Choose whether to allow Red Hat to collect data about OpenShift Data Science usage in your cluster. Usage data collection is enabled by default when you install OpenShift Data Science on your OpenShift cluster.

Apply autoscaling to your cluster to reduce usage costs

Use the cluster autoscaler to adjust the size of your cluster to meet its current needs and optimize costs.

Manage resource usage by stopping idle notebooks

Reduce resource usage in your OpenShift Data Science deployment by automatically stopping notebook servers that have been idle for a period of time.

Implement model-serving runtimes

OpenShift Data Science provides support for model-serving runtimes. A model-serving runtime provides integration with a specified model server and the model frameworks that it supports. By default, OpenShift Data Science includes the OpenVINO Model Server runtime. However, if this runtime doesn’t meet your needs (for example, if it doesn’t support a particular model framework), you can add your own custom runtimes.

Install in a disconnected environment

OpenShift Data Science self-managed supports installation in a disconnected environment. Disconnected clusters are on a restricted network, typically behind a firewall and unable to reach the Internet. In this case, clusters cannot access the remote registries where Red Hat provided OperatorHub sources reside. In this case, you deploy the OpenShift Data Science Operator to a disconnected environment by using a private registry in which you have mirrored (copied) the relevant images.
CHAPTER 4. TRY IT

Data scientists and developers can try OpenShift Data Science and access tutorials and activities in the Red Hat Developer sandbox environment.

IT operations administrators can try OpenShift Data Science in your own cluster with a 60-day product trial.
CHAPTER 5. GET IT

Managed cloud service
You have the following options for subscribing to OpenShift Data Science as a managed service:

- For OpenShift Dedicated, subscribe through Red Hat.
- For Red Hat OpenShift Service on Amazon Web Services (ROSA), subscribe through Red Hat or subscribe through the AWS Marketplace.

Self-managed software
To get Red Hat OpenShift Data Science as self-managed software, sign up for it with your Red Hat account team.