Red Hat
OpenShift
Data Science

Red Hat OpenShift Data Science

Develop, train, test, and deploy intelligent cloud-native applications

Highlights

- Develop with your choice of tools without worrying about infrastructure.
- Reduce friction and collaborate on a common platform to bring data scientists, developers, and IT operations together.
- Accelerate the delivery of intelligent applications and reduce time to market.
- Empower your data scientists by giving them a choice of applications and services from a vast partner ecosystem.

Embrace intelligent applications, faster

Artificial intelligence (AI), machine learning (ML), and deep learning (DL) are having a profound influence on application modernization efforts across myriad businesses and industries. Diverse organizations need to derive strategic value and new insights from their data, fostering expanding use of intelligent cloud-native applications and DevOps methodologies. This brave new world can be complex, with implications for everyone—from developers, to data scientists, to operations staff. Traditional approaches can present challenges:

- Getting started can be daunting, from keeping rapidly evolving tools and application services current and consistent, to provisioning hardware resources like graphic processing units (GPUs), to scaling intelligent applications.
- Popular cloud platforms offer scale and attractive, integrated environments and toolsets, but they can effectively lock in users with restrictive toolchains and limited deployment options.
- Having different platforms for application developers and data scientists can complicate collaboration and hinder the speed of development.
- Deploying intelligent applications at scale can be difficult, especially if separate development and production platforms are involved.

As a managed cloud service offering, Red Hat® OpenShift® Data Science gives data scientists and developers a powerful AI/ML platform for building and deploying intelligent applications. Organizations can experiment with a choice of tools, collaborate on a common platform, and accelerate speed to market—all within one common platform. OpenShift Data Science combines the self-service environment that data scientists and developers want with the confidence enterprise IT demands.

Having a trusted foundation reduces friction throughout the life cycle. OpenShift Data Science offers a robust platform, a broad ecosystem of popular certified tools and familiar workflows for deploying models into production. With these advantages, teams can collaborate with less friction and get intelligent applications into the market more efficiently, ultimately delivering greater value for the business.

Rapidly develop, train, test, and deploy

OpenShift Data Science is based on the community Open Data Hub project and Operate First. Open Data Hub demonstrates an AI/ML platform on Red Hat OpenShift with upstream efforts such as Apache Kafka and Kubeflow. Operate First brings open source concepts to operations, letting developers and operators collaborate to infuse operational excellence, without proprietary lock-in. OpenShift Data Science provides a subset of Open Data Hub tools in a fully supported cloud service, managed on Amazon Web Services (AWS) with optional independent software vendor (ISV) offerings.

Experiment with a choice of tools

With OpenShift Data Science, data scientists can experiment and discover new ways of bringing insights into the business. As a fully managed cloud service, data scientists can develop, train, and test machine learning models before they deploy. Teams get access to advanced tools delivered in an integrated experience. Data scientists can use their familiar tools or access a growing technology partner ecosystem for deeper AI/ML expertise—all without being burdened with a prescriptive toolchain. Rather than waiting for IT to provision necessary resources, they get on-demand infrastructure with a click rather than an IT ticket.

Collaborate on a common platform

OpenShift Data Science builds on an open source architecture designed for machine learning workloads and development workflows. It narrows the gaps between data science and DevOps, reducing the pain of handoffs on the way to production. Data scientists collaborate in real time in Jupyter notebooks. Developers integrate container-ready models into intelligent applications with less friction. IT worries less about governance, with no need to chase down rogue cloud-platform accounts.

Accelerate speed to market for intelligent applications

OpenShift Data Science brings machine learning models from early pilots into intelligent applications with greater speed on a shared, consistent platform. Data scientists can start fast with their choice of tools and access to self-service infrastructure. The service connects every machine learning life cycle stage with deeper AI capabilities through its software partner ecosystem, offering a wide range of certified tools with specialty AI/ML expertise. You can deploy models to hybrid cloud environments, gaining the flexibility of running workloads wherever you need them, with no commercial cloud lock-in.

OpenShift Data Science

Figure 1 illustrates how the model operation life cycle integrates with the initial offering of OpenShift Data Science as a common platform. This cloud service is available on Red Hat OpenShift Dedicated (on AWS) and Red Hat OpenShift Service on AWS. It provides a core data science workflow as a Red Hat managed service, with the opportunity for increased capabilities and collaboration through

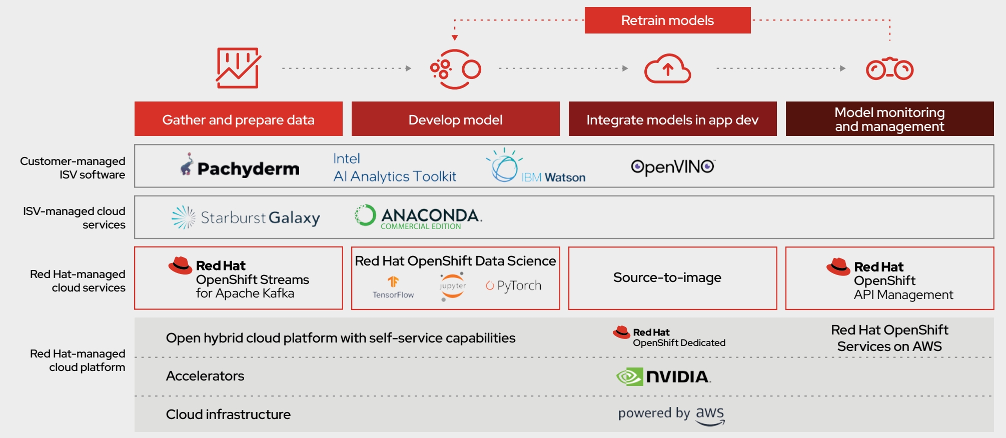
Retrain models
Gather and prepare data
Develop model
Integrate models in app dev
Model monitoring and management
Customer-managed ISV software
Pachyderm
Intel AI Analytics Toolkit
IBM Watson
OpenVINO
ISV-managed cloud services
Starburst Galaxy
ANACONDA COMMERCIAL EDITION
Red Hat-managed cloud services
Red Hat OpenShift Streams for Apache Kafka
Red Hat OpenShift Data Science
TensorFlow
jupyter
PyTorch
Source-to-image
Red Hat OpenShift API Management
Red Hat-managed cloud platform
Open hybrid cloud platform with self-service capabilities
Red Hat OpenShift Dedicated
Red Hat OpenShift Services on AWS
Accelerators
NVIDIA.
Cloud infrastructure
powered by aws

Figure 1. OpenShift Data Science release components

ISV-certified software. Models are either hosted on OpenShift cloud service or exported for integration into an intelligent application.

Core tools and capabilities provided with OpenShift Data Science offer a solid foundation:

- **Jupyter notebooks.** Data scientists can conduct exploratory data science in JupyterLab with access to core AI/ML libraries and frameworks, including TensorFlow and PyTorch.
- **Source-to-image (S2I).** Models can be published as endpoints via S2I for integration into intelligent applications and can be rebuilt and redeployed based on changes to the source notebook.
- **Optimized inference.** Deep learning models can be converted into optimized inference engines to accelerate experiments.

Red Hat provides Jupyter notebook images for Tensorflow and PyTorch as a part of the service, making it less complex for teams to embrace these powerful technologies without starting from scratch. For consistency and flexibility, the Jupyter spawner can deploy an organization's custom images to the data science teams, incorporating preferred libraries, tools, and languages. The service also includes the Git plug-in to JupyterLab, making it require less time to integrate with Git directly from the JupyterLab interface. Other common analytics packages provided as a part of the service simplify operation and make it easier to get started with the right tools for your project, including Pandas, scikit-learn, and NumPy.

As a managed cloud service, Red Hat provides site reliability engineering (SRE) support for the underlying OpenShift application platform and the OpenShift Data Science service. This support allows you to focus on your business analytics, not the underlying platform. Red Hat maintains high availability for the Red Hat OpenShift Data Science service, including the underlying Red Hat OpenShift managed cloud services environment. All updates, upgrades, and compatibility are managed as a part of the service, eliminating the need to track potentially complex compatibility matrices between analytics tools.

Tools for the complete model life cycle

OpenShift Data Science provides the services and software to let organizations successfully deploy their models and move them to production (Figure 2). In addition to OpenShift Data Science, this process is integrated with Red Hat OpenShift Streams for Apache Kafka and Red Hat OpenShift API Management.

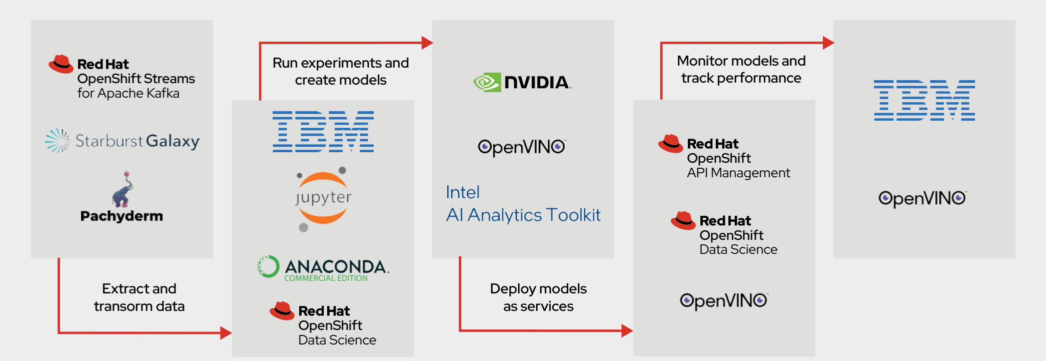
Red Hat OpenShift Streams for Apache Kafka
Starburst Galaxy
Pachyderm
Extract and transorm data
Run experiments and create models
IBM
jupyter
ANACONDA COMMERCIAL EDITION
Red Hat OpenShift Data Science
NVIDIA
OpenVINO
Intel AI Analytics Toolkit
Deploy models as services
Monitor models and track performance
Red Hat OpenShift API Management
Red Hat OpenShift Data Science
OpenVINO
IBM
OpenVINO

Figure 2. A typical model operationalization life cycle

Easing adoption, the Red Hat OpenShift Data Sciences dashboard provides a central place to discover and access all applications and documentation. Smart start tutorials offer best practice guidance for common components and integrated partner software and are available directly from the dashboard to help data scientists learn and get started faster. The following sections describe the principal analytics tools included with Red Hat OpenShift Data Science.

Starburst

Starburst accelerates analytics by making it fast and easy for your teams to capitalize on your data to improve how the business functions. Delivered as a self-managed product or a fully managed service, Starburst democratizes data access, bringing more comprehensive insights to data consumers. Starburst is built on open source Trino (formerly known as PrestoSQL), the premiere massively parallel processing (MPP) SQL engine. Built and operated by Trino experts and the creators of Presto, Starburst gives you the freedom to interrogate diverse data sets wherever they are located without having to move your data.

Starburst integrates with the scalable cloud storage and computing services provided by Red Hat OpenShift with a more stable, security-focused, efficient, and cost-effective way to query all your enterprise data. Benefits include:

- **Automation.** Starburst and Red Hat OpenShift operators provide auto-configuration, auto-tuning, and auto-management of clusters.
- **High availability and graceful scaledown.** The Red Hat OpenShift load balancer can keep services like the Trino coordinator in an always-on state.
- **Elastic scalability.** Red Hat OpenShift can automatically scale the Trino worker cluster based on query load.

Anaconda Commercial Edition

Anaconda Commercial Edition provides curated access to an extensive set of data science packages for use in Jupyter projects, with prebuilt Jupyter images available directly from the Red Hat OpenShift Data Sciences dashboard. Anaconda Commercial Edition gives organizations access to the world's most popular open source package distribution and management experience, optimized for commercial use, including:

- Open source innovation, with more than 7,500 Anaconda-curated data science and ML packages in Anaconda's premium repository.
- Content trust features, such as Conda signature verification, that help you keep vulnerabilities and unreliable software out of your data science and ML pipelines.
- Confidence with uptime service-level agreements (SLAs) and support you can depend on for production workflows.
- Full compliance for commercial usage under the Anaconda terms of service.

IBM Watson Studio

IBM Watson Studio[1] lets you build, run, and manage AI models at scale with Watson Machine Learning and Watson OpenScale. The platform combines open source frameworks like PyTorch, TensorFlow, and scikit-learn with IBM and its ecosystem tools for code-based and visual data science. The platform works with Jupyter notebooks, JupyterLab, command-line interfaces (CLIs), and Python languages.

IBM Watson helps operationalize AI and advances trust from principal to practice. Transparent processes provide insight into AI-led decisions. IBM Watson allows data privacy, compliance, and security across highly regulated industries and supports an open, diverse ecosystem promoting responsible use of AI. IBM Watson Studio delivers:

- AutoAI and AutoML to automatically build model pipelines, prepare data and select model types, and generate and rank model pipelines.
- Advanced data refinery to cleanse and shape data with a graphics flow editor.
- Integrated visual tooling through IBM SPSS Modeler to prepare data quickly and develop models visually.
- Model training and development to build experiments quickly with optimized pipelines.
- Embedded decision optimization to combine predictive and prescriptive models.
- Model management and monitoring of quality, fairness, and drift metrics.
- Model export as Python Jupyter Notebook.

Pachyderm

Organizations need data management solutions that facilitate everything from laptop experiments to important enterprise deployments. Pachyderm lets data science teams build and scale containerized, data-driven ML pipelines with a guaranteed data lineage provided by automatic data versioning. Engineered to solve real-world data science problems, Pachyderm provides the data foundation that allows teams to automate and scale their ml life cycle while guaranteeing reproducibility. With use cases that range from unstructured data to data warehouses, natural language processing, video and image ETL, financial services, and life sciences, Pachyderm provides:

- Automated data versioning that gives teams a high-performance way to keep track of all data changes.
- Data-driven containerized pipelines that speed data processing while lowering compute costs.
- An immutable data lineage that provides a fixed record for all activities and assets in the ML life cycle.
- The Pachyderm Console which provides intuitive visualization of your directed acyclic graph (DAG) and aids with debugging and reproducibility.
- Jupyter notebook support with Pachyderm's JupyterLab Mount Extension for a point-and-click interface to Pachyderm-versioned data.
- Enterprise administration with robust tools for deploying and administering Pachyderm at scale across different teams within the organization.

[1] *IBM Watson Studio and Watson Machine Learning are a part of IBM's Cloud Pak for Data offering.*

NVIDIA accelerated data science

Scalable data processing, data analytics, machine learning training, and inference all represent highly resource-intensive computational tasks. NVIDIA software makes it possible to accelerate all aspects of end-to-end data science by taking advantage of the parallel processing capabilities of GPUs. Scaling on-premise GPU resources or configuring Kubernetes provisioning to use them should not distract data scientists who would rather be extracting value from their data.

Diverse organizations already use NVIDIA solutions for machine learning and a host of other services. OpenShift Data Science reduces the complexity to stand up GPU-enabled hardware to accelerate resource-intensive data science experiments. With OpenShift Data Science, organizations can employ Amazon Elastic Computing (EC2) instances powered by NVIDIA GPUs on demand, growing or shrinking computational resources as needed.

Intel OpenVINO toolkit

The Intel Distribution of the OpenVINO toolkit accelerates developing and deploying high-performance DL inference applications on Intel platforms. The toolkit lets you build, optimize, tune, and run comprehensive AI inferencing using the included model optimizer along with runtime and development tools.

- **Build.** Developers can use the Open Model Zoo to find open source, pretrained, and preoptimized models ready for inference, or they can use their own DL models.
- **Optimize.** The Model Optimizer can convert the model to an Intermediate Representation (IR), resulting in a pair of files describing the network topology and containing the model's weights and biases.
- **Deploy.** The Inference Engine can output results on multiple processors, accelerators, and environments with a write-once, deploy-anywhere efficiency.

Intel® AI Analytics Toolkit

The Intel AI Analytics Toolkit gives data scientists, AI developers, and researchers familiar Python tools and frameworks to accelerate end-to-end data science and analytics pipelines on Intel architectures. The components use oneAPI libraries for low-level compute optimizations. This toolkit maximizes performance from preprocessing through ML and provides interoperability for efficient model development.

Using the Intel AI Analytics Toolkit, you can:

- Deliver high-performance, DL training on Intel XPUs and integrate fast inference into your AI development workflow with Intel-optimized, DL frameworks for TensorFlow and PyTorch, pre-trained models, and low-precision tools.
- Achieve drop-in acceleration for data preprocessing and ML workflows with compute-intensive Python packages, Modin, scikit-learn, and XGBoost, optimized for Intel.
- Gain direct access to analytics and AI optimizations from Intel to ensure that your software works together uninterrupted.

Conclusion

With OpenShift Data Science, organizations can experiment, collaborate, and ultimately accelerate their intelligent application journey. The cloud-based, add-on service managed by Red Hat simplifies and accelerates experimentation for data scientists, a modern containerized AI/ML platform, and the convenience and scalability of AWS. Self-service for developers and data scientists speeds innovation on an application platform already used and fully trusted by enterprise IT. Unlike competing approaches, data scientists can choose tooling with no restrictive toolchain, bringing new data insights without forcing arbitrary limitations.

About Red Hat

Red Hat helps customers standardize across environments, develop cloud-native applications, and integrate, automate, secure, and manage complex environments with award-winning support, training, and consulting services.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

North America
1 888 REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com