

Cloudera AI 1.5.4

Using Cloudera AI Registry

Date published: 2020-07-16

Date modified: 2024-12-16

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2026. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

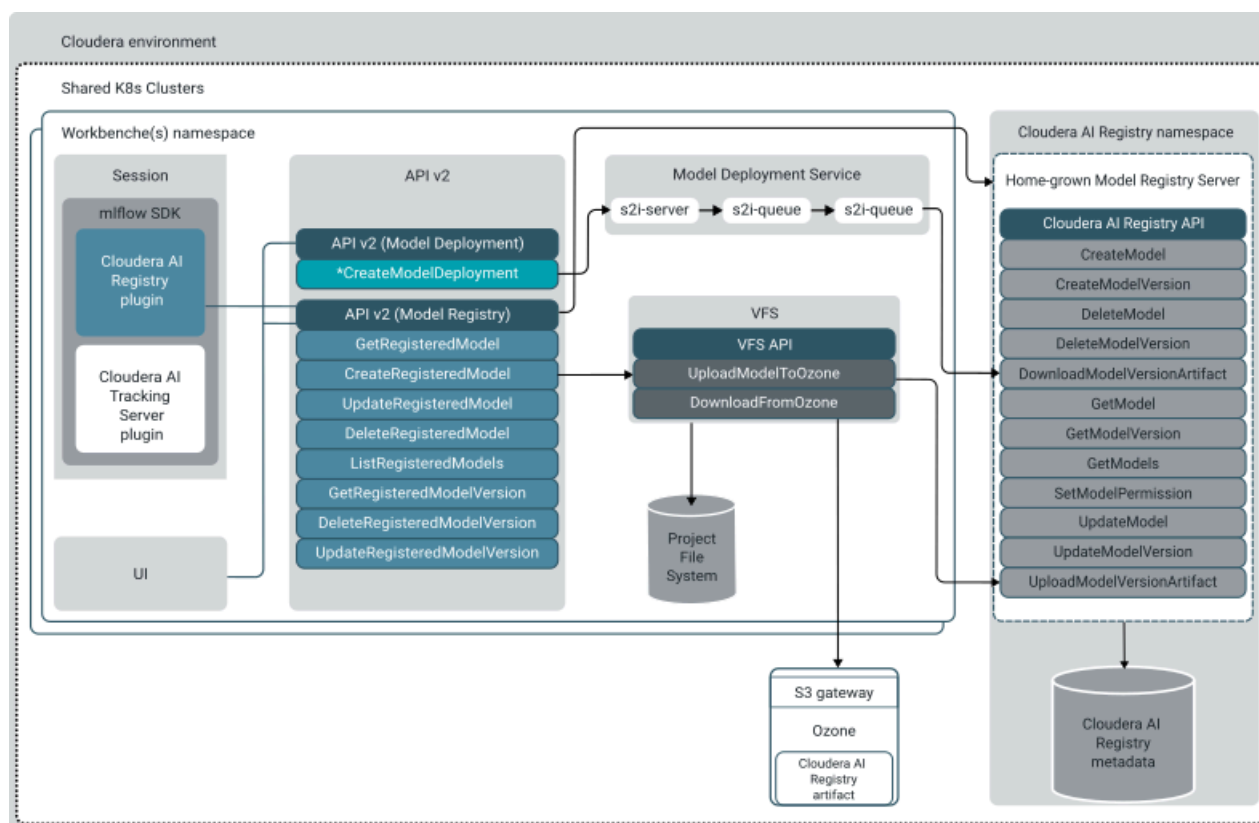
Using Cloudera AI Registry.....	4
Cloudera AI Registry standalone API.....	5
Prerequisites for Cloudera AI Registry standalone API.....	5
Authenticating clients for interacting with Cloudera AI Registry API.....	6
Role-based authorization.....	6
REST Client.....	6
Troubleshooting issues with Cloudera AI Registry API.....	7

Using Cloudera AI Registry

Cloudera AI Registry is the core enabler for MLOps, or DevOps for machine learning.

Cloudera AI Registry stores and manages machine learning models and associated metadata, such as the model's version, dependencies, and performance. The registry enables MLOps and facilitates the development, deployment, and maintenance of machine learning models in a production environment.

Figure 1: Cloudera AI Registry on premises



Cloudera AI Registry includes functionality for the following tasks:

- Storing and organizing different versions of a machine learning model and its associated metadata.
- Tracking the lineage of a model, including who created it, when it was created, and any changes made to it over time.
- Providing APIs for accessing and deploying models, as well as for querying and searching the registry.
- Integrating with CI/CD pipelines and other tools used in the MLOps workflow.

Cloudera AI Registries help organizations improve the quality and reliability of their machine learning models by providing a centralized location for storing and managing models, as well as enabling traceability and reproducibility of model development. They also make deploying and managing models in a production environment easier by providing a single source for model versions and dependencies.

The Cloudera AI Registry integrates MLFlow and maintains compatibility with the open source ecosystem.

Limitations

- Upgrade to the General Availability (GA) version of Cloudera AI Registry might not be supported. Alternatively, upgrade to the GA version of Cloudera AI Registry might require reinstalling Cloudera AI Registry which could

result in loss of Cloudera AI Registry data configured with the technical preview (TP) version of Cloudera AI Registry.

Cloudera AI Registry standalone API

You can use standalone Cloudera AI Registry API to communicate with the Cloudera AI Registry using the REST client or CLI client.

After the release of Cloudera AI Registry General Availability (GA), the Cloudera AI Registry API has been exposed through Cloudera AI Workbench APIv2. So, a Cloudera AI Workbench must be present in the same Cloudera environment to communicate with the Cloudera AI Registry. For more information, see [REST API](#).

The Cloudera AI Registry standalone API supports the following functionalities:

- GET/PATCH/DELETE for the model and model version
- GET a curated list of NGC models
- Import external model from [NVIDIA NGC](#) or [HuggingFace](#) to Cloudera AI Registry through the POST method

Currently, the Cloudera AI Registry Standalone API does not support uploading the models through POST method from the local machine.

Cloud Platforms

Cloudera AI Registry API is available only on AWS and Azure.

API definition

The Swagger definition is available in the [Cloudera AI API documentation](#).

Prerequisites for Cloudera AI Registry standalone API

To set up the Cloudera AI Registry standalone API, configure the Cloudera AI Inference service and import pretrained Models.

Prerequisites for Cloudera AI Inference service

Cloudera AI Registry is a prerequisite for Cloudera AI Inference service because the Cloudera AI Inference service needs to deploy the models that are stored in the Cloudera AI Registry.

- To use the Cloudera AI Inference service, the latest Cloudera AI Registry must be present in the same Cloudera environment before the Cloudera AI Inference service is created.
- If there is an older Cloudera AI Registry in the environment that is created before May 14, 2024, follow the *Upgrade Cloudera AI Registry* instructions to upgrade the Cloudera AI Registry to the latest version before you create the Cloudera AI Inference service.
- If the Cloudera AI Registry is re-created, upgraded, or cert-renewed while the Cloudera AI Inference service is present, then follow the steps listed in the *Manually updating Cloudera AI Registry configuration* topic to ensure that the configuration of Cloudera AI Registry and Cloudera AI Inference service are synchronized.

Prerequisites to import pretrained models

You must add the URL details to allow them in the firewall rules.

NVIDIA GPU Cloud (NGC)

Add the following URL details so they can be allowed in the firewall's rules.

- prod.otel.kaizen.nvidia.com (NVIDIA open telemetry)
- api.ngc.nvidia.com
- files.ngc.nvidia.com

Hugging Face

Add the following URL details so they can be allowed in the firewall's rules.

- huggingface.co
- cdn-lfs.huggingface.co
- *.cloudfront.net (CDN)



Note: If required, you must allow more URLs based on your requirements.

Authenticating clients for interacting with Cloudera AI Registry API

Clients that interact with the Cloudera AI Registry Standalone API and with model endpoints must obtain a JSON Web Token (JWT) from the Cloudera control plane, which must be passed as a Bearer token in HTTP requests sent to the serving API and endpoints.

To obtain JWT, run the following Cloudera CLI command:

```
$ CDP_TOKEN=$(cdp iam generate-workload-auth-token --workload-name DE | jq -r '.token')
```

Then pass CDP_TOKEN in the HTTP request header as follows

```
$ curl -H "Authorization: Bearer ${CDP_TOKEN}" <URL>
```

The token obtained using this method expires in one hour.

Role-based authorization

Cloudera AI Registry implements role-based access control. It allows access for users with the following IAM roles:

- MLUser
- MLAdmin (admin user)

For more information about the access control for the registered models, see [Model access control](#).

REST Client

You need the Domain information to use the REST client to interact with the registry. To obtain the Domain information, perform the following:

1. In the **Cloudera** console, click the **Cloudera AI** tile.
2. Click AI Registries in the left navigation menu. The AI Registries page displays.

- Click on the name of the Cloudera AI Registry to display the Cloudera AI Registry information. The Domain name is displayed in the Details tab.

Model Registries / model-registry-ml-cddeb2f5-cc9

The screenshot shows the Cloudera AI Registry console interface. At the top, there is a green 'Ready' status indicator. Below it, there are two tabs: 'Details' (selected) and 'Events & Logs'. The 'Details' tab displays a list of key-value pairs for the model registry:

Name	
Environment Name	eng-ml-dev-env-aws
Environment CRN	cm:odp:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:...
CRN	cm:odp:ml:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:model_regi...
Machine User CRN	cm:altus:iam:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:machine...
Machine User Workload User Name	srv_cm_l_env_machine_user_3715d
Creation Date	05/07/2024 9:23 AM PDT
Creator	cm:altus:iam:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:user:812765fa-b363-4a20-ae07-776c64491367
Domain	https://modelregistry.ml-cddeb2f5-cc9.eng-ml-d.xcu2-8y8x.dev.cldr.work

Model Registries / model-registry-ml-cddeb2f5-cc9

This screenshot is identical to the one above, showing the Cloudera AI Registry console interface with the 'Ready' status and the 'Details' tab selected. The details table is the same as in the previous screenshot.

Troubleshooting issues with Cloudera AI Registry API

Learn about some of the recommended series of steps to perform when troubleshooting issues related to the Cloudera AI Registry API.

Debugging the model import failure

To debug errors that occurred on the Cloudera AI Registry server, you can access the logs found in the API v2 pod.


Access logs from Cloudera AI Registry Kubernetes cluster.

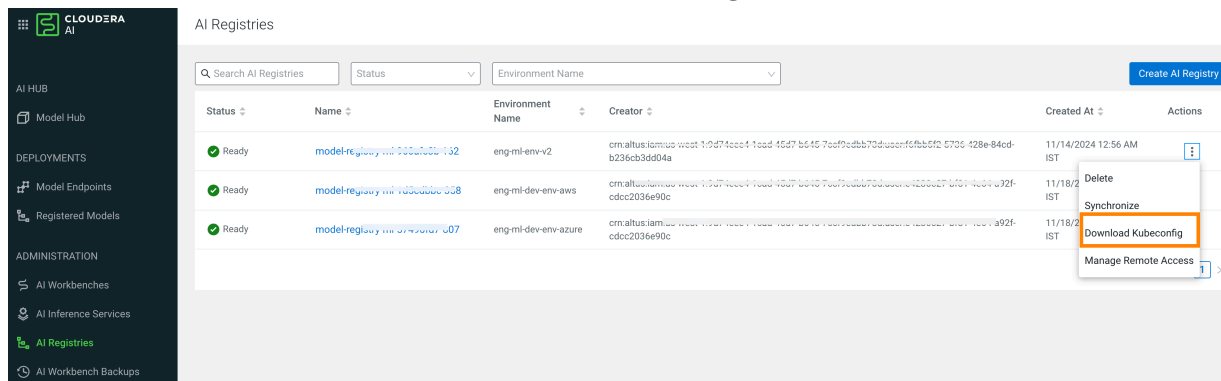
You can obtain the kubeconfig for the Cloudera AI Registry cluster.

- In the **Cloudera** console, click the **Cloudera AI** tile.

2. Click AI Registries in the left navigation menu. The AI Registries page displays.

3.

In the **Actions** menu, click  and select **Download Kubeconfig**.



The screenshot shows the Cloudera AI Registry console interface. On the left is a navigation menu with categories: AI HUB, DEPLOYMENTS, and ADMINISTRATION. The 'AI Registries' option is highlighted. The main content area displays a table of AI Registries with columns for Status, Name, Environment Name, Creator, Created At, and Actions. A dropdown menu is open for the first registry, showing options: Delete, Synchronize, Download Kubeconfig (highlighted with an orange box), and Manage Remote Access.

Status	Name	Environment Name	Creator	Created At	Actions
Ready	modelregistry-wl-932af52c-132	eng-mi-env-v2	cm.allius.komus-west-1:5d74cc1-1aad-46d7-b546-7a07f9eabb70d-us-east-1:bb57d-0700-428e-84cd-b236cb3d904a	11/14/2024 12:56 AM IST	[Info] [Delete] [Synchronize] [Download Kubeconfig] [Manage Remote Access]
Ready	modelregistry-wl-7d8e08be-058	eng-mi-dev-env-aws	cm.allius.komus-west-1:5d74cc1-1aad-46d7-b546-7a07f9eabb70d-us-east-1:bb57d-0700-428e-84cd-b236cb3d904a	11/18/2 IST	[Info] [Delete] [Synchronize] [Download Kubeconfig] [Manage Remote Access]
Ready	modelregistry-wl-07490101-007	eng-mi-dev-env-azure	cm.allius.komus-west-1:5d74cc1-1aad-46d7-b546-7a07f9eabb70d-us-east-1:bb57d-0700-428e-84cd-b236cb3d904a	11/19/2 IST	[Info] [Delete] [Synchronize] [Download Kubeconfig] [Manage Remote Access]

In AWS, you need to add your identity under Manage Remote Access to access the Kubernetes cluster.

You must add your identity under Manage Remote Access. For information on granting remote access, see *Granting Remote Access to Cloudera AI Workbench*. After the kubeconfig is set up, run the following `kubectl` command to get logs for the Cloudera AI Registry pod:

```
kubectl logs <model registry pod name> -n mlx
```