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

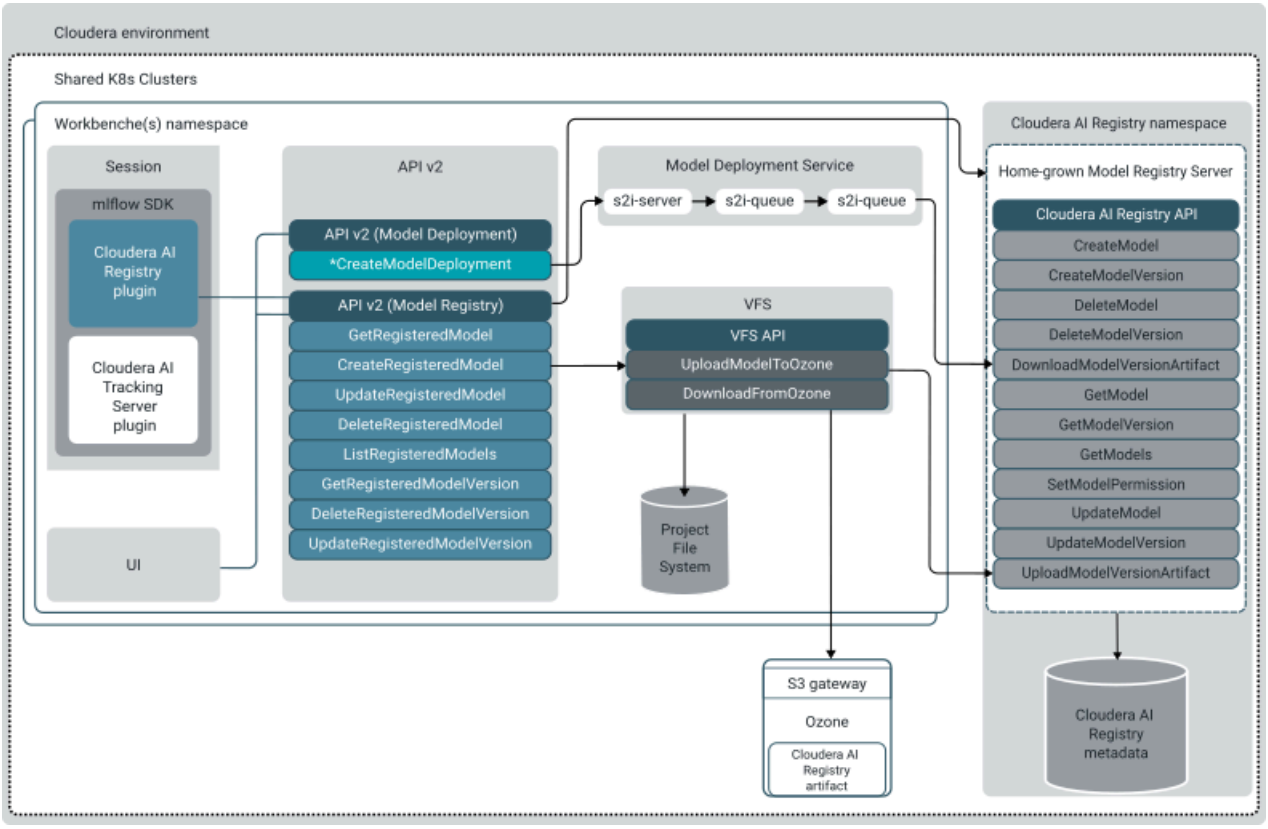
Contents

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)

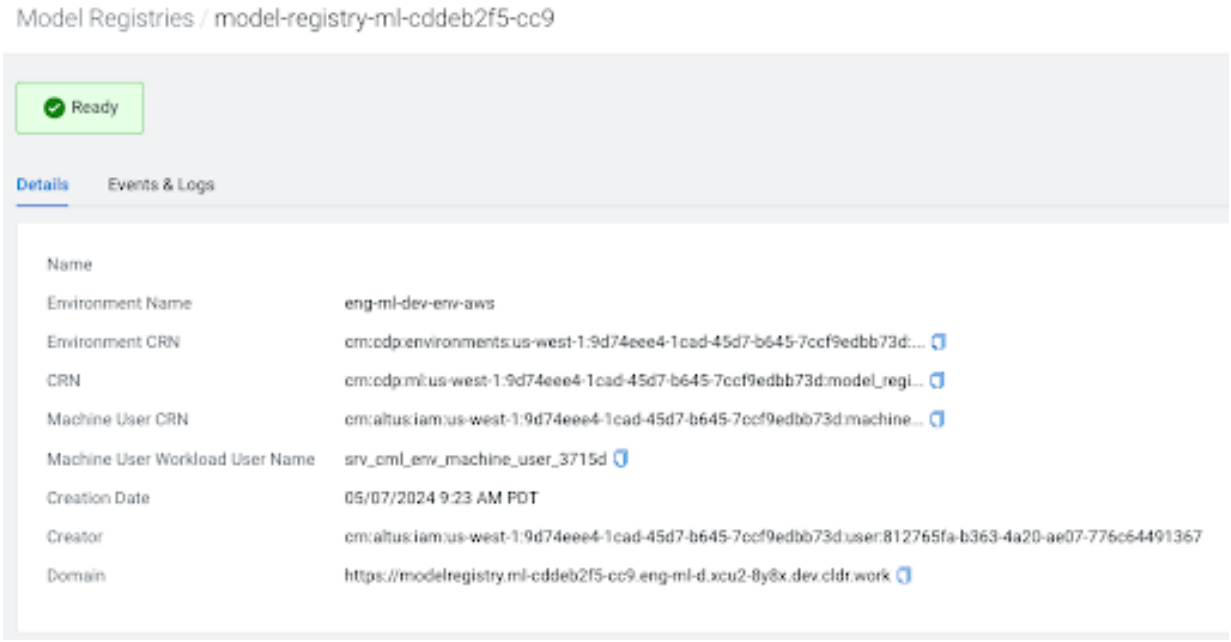
For more information about the access control for the registered models, see Model access control.

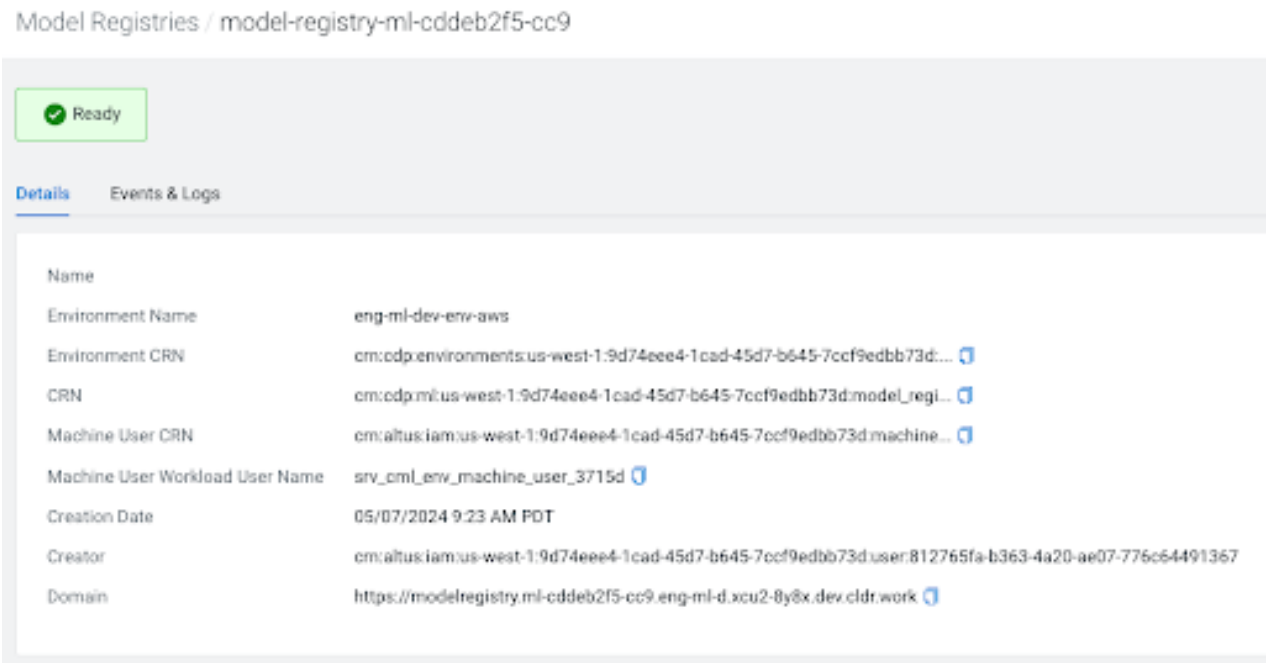
REST Client

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.

3. 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.





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

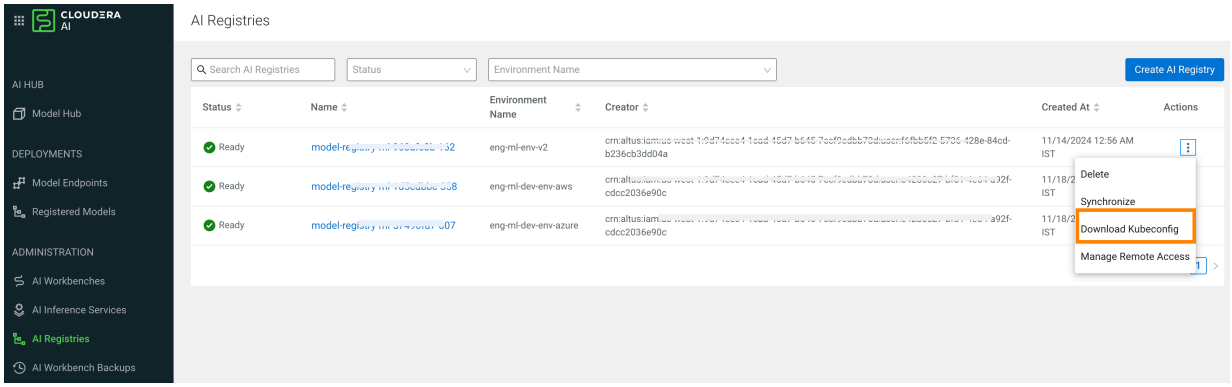
To debug errors that occurred on the Cloudera AI Registry server, you can access the logs found in the API v2 pod.

Access logs from Cloudera AI Registry Kubernetes cluster.

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

1. 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**.



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
```