

## Using Model Hub

Date published: 2020-07-16

Date modified: 2025-05-29

# CLOUDERA

# Legal Notice

© Cloudera Inc. 2025. 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

<b>Using Model Hub.....</b>	<b>4</b>
Role-based authorization in Model Hub.....	4
Model Hub in air-gapped installations.....	4
Importing models from NVIDIA NGC.....	5
Importing models from Hugging Face (Technical Preview).....	7
Downloading and uploading Model Repositories for an air-gapped environment.....	10

## Using Model Hub

You can easily import the models listed in the Model Hub into the Registered Models page and then deploy it using the Cloudera AI Inference service. This streamlines the workflow of developers working on AI use cases by simplifying the process of discovering, deploying, and testing models.

Model Hub is a catalog of top-performing LLM and generative AI models. The Model Hub page displays the list of different models along with their source type, tags, and description. You can import models listed on the Model Hub page and deploy it from the Registered Models page.

### Related Information

[Upgrading Cloudera AI Registry](#)

## Role-based authorization in Model Hub

Model Hub implements role-based access control.

Users need to have EnvironmentUser role along with one of the following roles in the environment tied to AI Registry that the model is being imported to:

- MLAdmin (admin user)
- MLUser

Without these roles, if you try to import models, the Failed to Fetch error or something similar is displayed.

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

### Related Information

[Model access control](#)

## Model Hub in air-gapped installations

For air-gapped installations using a [proxy setup](#), you must whitelist the following URLs in your firewall rules to ensure the Model Hub functions correctly:

### Hugging Face

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

### Nvidia NGC Catalog

- xfiles.ngc.nvidia.com

### GitHub (for latest Model Hub catalog)

- github.com/cloudera



**Note:** In case of any errors, the system will fall back to a static Model Hub catalog, which may not be the latest.

You must ensure that these URLs are properly whitelisted to maintain access to up-to-date model catalogs.

## Importing models from NVIDIA NGC

You can import the NVIDIA NGC Catalog models listed in the Model Hub page and deploy it from the Registered Models page.

### Before you begin

To import models, you must add the following **NVIDIA GPU Cloud (NGC)** URL details so they can be allowed in the firewall's rules.


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

### Procedure

1. In the **Cloudera** console, click the **Cloudera AI** tile.  
The **Cloudera AI Workbenches** page displays.
2. Click **Model Hub** under **AI Hub** in the left navigation menu.  
The **Model Hub** page displays. The page lists different models along with its source type, tags, and description.

### 3. Click Import on the model you want to import.

The **Import Model** page displays.



Llama3 Instruct

Details

Llama 3 is a large language AI model comprising a collection of models capable of generating text and code in response to prompts. Meta developed and released the Meta Llama 3 family of large language models (LLMs), a collection of pretrained and instruction tuned generative text models in 8 and 70B sizes. Token counts refer to pretraining data only. Both the 8 and 70B versions use Grouped-Query Attention (GQA) for improved inference scalability. The Llama 3 instruction tuned models are optimized for dialogue use cases and outperform many of the available open source chat models on common industry benchmarks. Further, in developing these models, we took great care to optimize helpfulness and safety.

Tags

Llama

Meta

Chat

Large Language Model

TensorRT-LLM

Language Generation

NeMo

NVIDIA Validated

\* Select Registry

aws

aws-ml-model-registry:ml-model-459-74d

▼

\* Select Model Size

Llama 3 70B Instruct

▼

\* Select Optimization

Llama 3 70B Instruct A100 FP16 LoRA Throughput

▼

PROFILE	PRECISION	GPU	GPU DEVICE
Throughput	FP16	A100	20b2:10de
NIM VERSION	FEAT_LORA	FEAT_LORA_MAX_RANK	COUNT
1.0.0	true	32	4

\* Enter Name or Select Model ⓘ

lim-sales

▼

⚠ Warning

IMPORTANT: Please read the following before proceeding. This AI model is a third party software package that is not validated or maintained by Cloudera, Inc. ("Cloudera"). By configuring and launching this AI model, you will cause such third party software to be downloaded and installed into your environment direct from an external website. If you do not wish to download and install the third party software packages, do not configure, launch or otherwise use this AI model. By configuring, launching or otherwise using the AI model, you acknowledge the foregoing statement and agree that Cloudera is not responsible or liable in any way for the third party software packages.

☒ By checking this box, you confirm that you have read and agreed to the [Use Policy](#) and [License Agreement](#) for this model.

Import


View On NVIDIA ↗

Cancel

### 4. In the Select Registry drop-down list, select the AI registry to which you want to import the model.

5. In the Select Parameter drop-down list, select the model variant.



**Note:** After you select the parameter, the [View On NVIDIA](#)  option is enabled. Click on it to view the model details of the selected model on the *NVIDIA* website.

6. In the Select Optimization drop-down list, select the model variant that is optimized for your use case and domain. The details of the chosen optimization profile is displayed.
7. In the Enter Name or Select Model field, select a name from the existing list or enter a new name for the model you are importing.
8. If displayed, read the User Policy and License Agreement, and click the checkbox if you agree.
9. Click Import.

The **Model Hub** page is displayed with the message that the model is imported successfully.

## Results

You can click Registered Models in the left navigation menu to view the newly imported model.

## Importing models from Hugging Face (Technical Preview)

You can import the Hugging Face models listed on the Model Hub page into your Model Registry. If your preferred model is not listed on the Model Hub page, you can [import it from the Registered Models](#) page. After you import the model, the newly imported model will be listed on the Registered Models page.



**Note:** This feature is in Technical Preview and not recommended for production deployments. Cloudera recommends that you try this feature in test or development environments.

## Procedure

1. In the **Cloudera** console, click the **Cloudera AI** tile.  
The **Cloudera AI Workbenches** page displays.
2. Click **Model Hub** under **AI Hub** in the left navigation menu.  
The **Model Hub** page displays. The page lists different models along with its source type, tags, and description.

3. Click Import on the model you want to import.



The **Import Model** page displays.

# Import Model

 StarCoder


Details

The StarCoder models are 15.5B parameter models trained on 80+ programming requests excluded. The model uses Multi Query Attention, a context window of 8 Middle objective on 1 trillion tokens.

Tags

- bigCode
- StarCoder
- Code Generation
- Text Generation
- Multilingual support

\* Select AI Registry

 eng-ml-dev-env-azure model-registry-ml-a857b6bd-4a6

\* Select Model Size

StarCoder

\* Select Optimization

starcoder

A10G	A100	L40S
2	1	1

\* Enter Name or Select Model 

google-startcoder

4. In the Select AI Registry drop-down list, select the model registry to which you want to import the model.
5. In the Select Model Size drop-down list, select the model size.
6. In the Select Optimization drop-down list, select the optimization profile. It displays the recommended GPU counts for the specific GPU of the optimization profile.
7. In the Enter Name or Select Model field, select a name from the existing list or enter a new name for the model you are importing.
8. For the gated models of Hugging Face, an access token is required. The Hugging Face Token field is displayed only if it requires a token. Enter the access token obtained from the Hugging Face website if this field is displayed.



**Important:** Consent on Hugging Face and Access Token Required

Some models like gated models on Hugging Face require an access token. To obtain access tokens from the Hugging Face website, follow the instructions displayed on the Import Model page.

In addition to access tokens, some models on Hugging Face require you to log in to your Hugging Face account, navigate to the model, and accept an agreement. After accepting the agreement, your existing Hugging Face token will work. Follow the instructions displayed on the Import Model page to know how to accept the agreement on *Hugging Face*.

9. Click Import. The Model Hub page displays a message that the model import has been triggered successfully along with a button to view the status of that import process.

### Results

You can click Model Registry in the left navigation menu to view the newly imported model.

## Downloading and uploading Model Repositories for an air-gapped environment

An air-gapped environment is physically isolated from the internet and external networks, preventing the transmission or reception of data online. As a result, enabling the download of Model Repositories in such environments requires the Administrator to perform additional steps.

To use Models from NVIDIA NGC and Hugging Face, the Administrator must download Model artifacts from these sources on specially networked hosts. The artifacts must then be manually transferred, uploaded to the object storage utilized by the Cloudera AI Registry and Cloudera AI Inference service. Following that, the available Models are ready to be used. This solution is an alternative to accessing Model Hub in an air-gapped environment.

### Related Information

[Prerequisites for downloading and uploading Model artifacts in air-gapped environment](#)

[Downloading Model Repositories for an air-gapped environment](#)

[Uploading Model Repositories for an air-gapped environment](#)

[Creating the Model entry in Cloudera AI Registry in air-gapped environment](#)

[Importing Model to Cloudera AI Registry in air-gapped environment](#)