Cloudera AI

Managing Cloudera Copilot

Date published: 2020-07-16
Date modified: 2025-10-31

https://docs.cloudera.com/

Legal Notice

Contents

Cloudera Copilot Overview

Learn how to configure and use Cloudera Copilot with Cloudera AI Inference service and Amazon Bedrock models.

Cloudera Copilot is an AI-powered coding assistant designed for seamless integration within JupyterLab ML Runtimes. With its chat interface and comprehensive code completion features, Cloudera Copilot enhances the development experience for machine learning projects. It offers compatibility with model endpoints deployed in Cloudera AI Inference service as well as Amazon Bedrock models, providing developers with flexibility and efficiency in their workflows.

Cloudera AI Inference service vs Amazon Bedrock

Cloudera AI Inference service model endpoints may be a good choice if you are concerned about proprietary data being sent to a third-party service provider. With Cloudera AI Inference service, you can run your own models and ensure that your proprietary data stays within your cloud deployment.

If you are not concerned about proprietary data being sent to a third-party service provider, and you do not expect high volumes of Cloudera Copilot usage, then using Amazon Bedrock models may be a simpler and more cost-effective solution.

Configuring Cloudera AI Inference service and Amazon Bedrock to set up Cloudera Copilot

To use Cloudera Copilot, as a Site Administrator, set up Cloudera AI Inference service model endpoints or configure the credentials in Amazon Bedrock depending on where you want to deploy your custom model.

For Cloudera AI Inference

Cloudera AI Inference service is a production-grade serving environment for traditional, generative AI, and LLM models. It is designed to handle the challenges of production deployments, such as high availability, fault tolerance, and scalability. Follow the steps to configure Cloudera AI Inference service model endpoints to use Cloudera Copilot.

1. Configure authentication, and authorization, and import a model from NGC.
2. Create a Cloudera AI Inference service instance.
3. Create a model endpoint using UI.

No additional credentials need to be configured to use Cloudera AI Inference service model endpoints with Cloudera Copilot.

For Amazon Bedrock

To use Cloudera Copilot, as a Site Administrator, you can configure the credentials to use Amazon Bedrock models.

1. Generate a pair of Access and Secret keys through AWS IAM. For more information, see Manage access keys for IAM users.
2. In the **Cloudera** console, click the **Cloudera AI** tile.

   The **Cloudera AI Workbenches** page displays.
3. Click on the workbench name.

   The **Workbenches Home** page displays.
4. Click **Site Administration** in the left navigation menu.

   The Site Administration page displays.

5. Click Settings Environment Variables and add the following obtained in Step 1:
   - AWS_SECRET_ACCESS_KEY
   - AWS_ACCESS_KEY_ID
   - AWS_DEFAULT_REGION

Configuring Cloudera Copilot

After setting up credentials, you must make configuration changes in the Cloudera AI UI before using Cloudera Copilot.

Choosing a model

Models vary in accuracy, and cost. Larger models will provide more accurate responses but will cost more. For Cloudera AI Inference service models, larger models require more expensive GPU hardware to run on, while in Amazon Bedrock, larger models will cost more per prompt.

Language models vs Embedding models

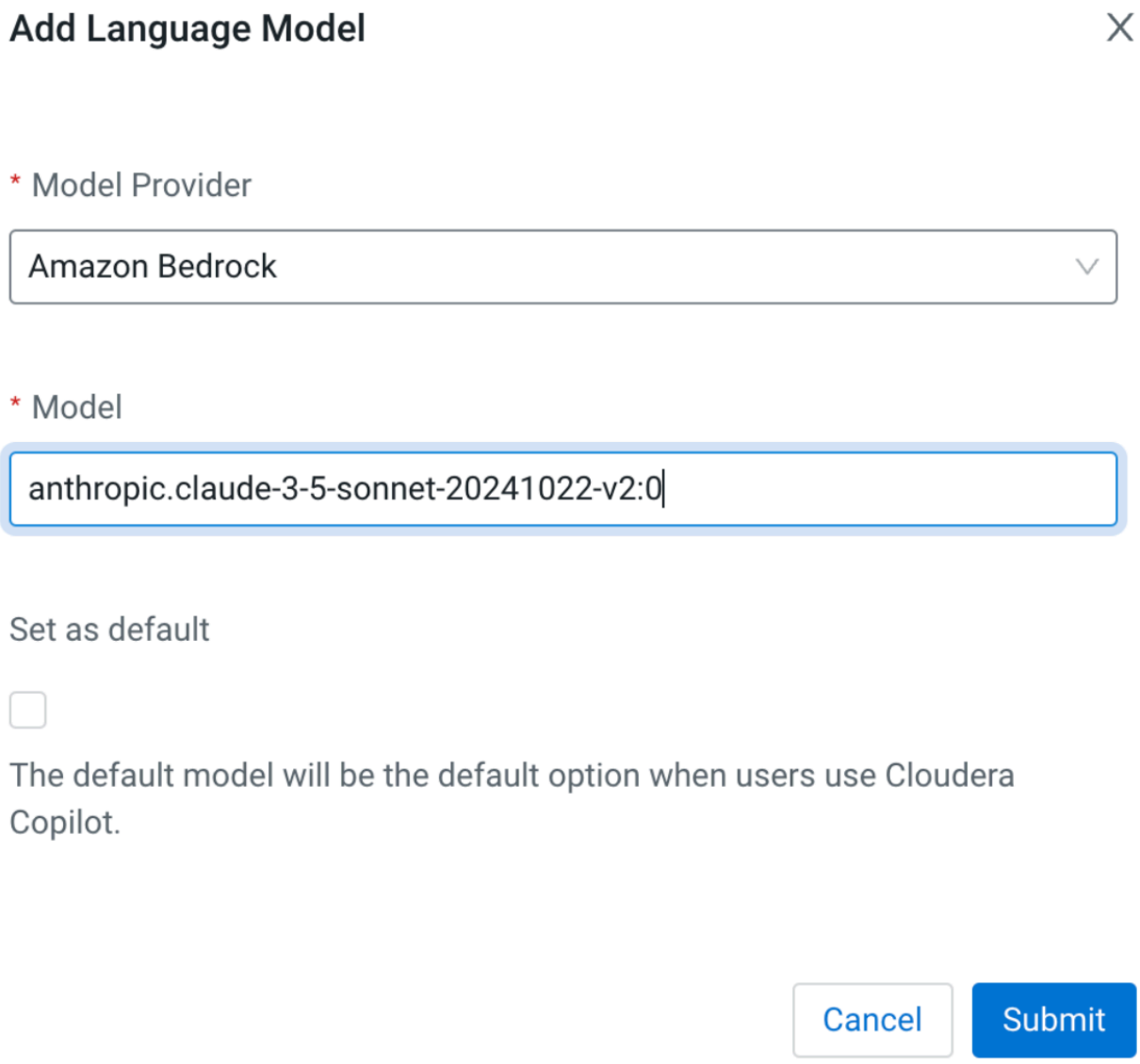
Cloudera Copilot supports the following model types:

- Language models: These are used for code completion, debugging, and chat.
- Embedding models: These are used for Retrieval Augmented Generation (RAG) use cases. This allows you to augment language model responses with specific information that a language model is not aware of. For example, you can provide internal company documents that map company acronyms to their definitions.

Procedure

1. In the **Cloudera** console, click the **Cloudera AI** tile.
   The **Cloudera AI Workbenches** page displays.
2. Click on the workbench name.
   The **Workbenches Home** page displays.
3. Click **Site Administration** in the left navigation menu.
   The **Site Administration** page displays.
4. Click Settings , and select the Enable Cloudera Copilot checkbox under Feature Flags.
   A new navigation tab Cloudera Copilot appears at the top of the **Site Administration** page.
5. Click the Cloudera Copilot tab.
   The **Cloudera Copilot** page displays.
6. Select the type of model you want to add: Language Model or Embedding Model.

7. Click Add Model button.



8. Select a model provider from the Model Provider dropdown list.
9. In the Model field, provide the model name:
   - For Bedrock models: Select a model name from the Model dropdown list.
   - For Cloudera AI Inference service models, provide the model endpoint and the `model_id` as the model name string. You can get the model endpoint and `model_id` information from the Model Endpoint details page.
     - Example Model Endpoint: *HTTPS://CAII-PROD-LONG-RUNNING.ENG-ML-L.VNU8-SQZE.YOURCOMPANY.SITE/NAMESPACES/SERVING-DEFAULT/ENDPOINTS/LLAMA-31-8B-INSTRUCT-2XA10G/V1/CHAT/COMPLETIONS*
     - Example model ID: *PHWQ-GQMD-4KOS-PERD*

   The model that you add for the first time is selected as the default model automatically and the deselect option is disabled. This is to enforce that there is always one default model. When you add more models, you can choose one of them to be the default model. You can have one default language model, and one default embedding model.
10. Click Add.

Results

The model appears under Models or Third Party Models depending on the model provider type you selected.