Upgrading Cloudera AI Inference service

Date published: 2026-01-19
Date modified:

Legal Notice

Contents

Prerequisites for upgrading Cloudera AI Inference service

Before upgrading Cloudera AI Inference service, you must fulfill all prerequisites.

Check the following requirements, permissions and roles for upgrading Cloudera AI Inference service:

Roles and permissions

You must have the following roles:

- MLAdmin role
- EnvironmentAdmin role

Service state

Check the statuses of the following services:

- The Cloudera AI Inference service must be in the Ready state.
- The Upgrade option is available only when a higher Cloudera AI Inference service version is supported.

Note:

If the service is in a failed, started, or any in-progress state, wait until the operation completes or resolve the issue before proceeding with the upgrade.

Infrastructure requirements

You must meet the following infrastructure requirements:

- The Kubernetes cluster is healthy and reachable.
- Network connectivity between the on-premises cluster and the configured custom Docker Registry (if any) is stable.
- The Kubeconfig credentials with appropriate permissions are configured and up to date.
- Sufficient resources, CPU, GPU, and memory, are available in the cluster.

Storage requirements

You must meet the following storage requirements:

- The configured storage backend, which is NFS, S3-compatible, or equivalent, is accessible.
- Storage credentials are valid.

Manually Upgrading Cloudera AI Inference service using the UI

In cases where a Cloudera AI Inference service upgrade is available but no Kubernetes upgrade is required, you can upgrade the Cloudera AI Inference service directly.

About this task

Warning: The model endpoint interruptions and application downtime are required during the upgrade.

You can upgrade an existing Cloudera AI Inference service instance when a higher supported version becomes available. Upgrading allows you to take advantage of new features, performance improvements, security updates, and supported component versions.

Upgrade your Cloudera AI Inference service in the following cases:

- A new version of Cloudera AI Inference service is released.
- Security patches or critical bug fixes are available.
- New features or enhancements are required.
- Alignment with the latest supported Kubernetes, Istio, or serving components is needed.

Procedure

1. In the **Management Console**, click the **Cloudera AI** tile.

   The **Cloudera AI** page displays.
2. Click **AI Inference Services** under **ADMINISTRATION** on the left navigation menu.
3. Select the Cloudera AI Inference service instance wou want to upgrade and click ⋮ next to the instance.
4. Select the Upgrade action.

   The upgrade confirmation dialog displays.
5. Click OK to initiate the upgrade.

Manually Upgrading Cloudera AI Inference service using the CLI

In cases where a Cloudera AI Inference service upgrade is available but no Kubernetes upgrade is required, you can upgrade the Cloudera AI Inference service directly.

You can upgrade an existing Cloudera AI Inference service instance when a higher supported version becomes available. Upgrading allows you to take advantage of new features, performance improvements, security updates, and supported component versions.

Upgrade your Cloudera AI Inference service in the following cases:

- A higher version of Cloudera AI Inference service is released.
- Security patches or critical bug fixes are available.
- New features or enhancements are required.
- Alignment with the latest supported Kubernetes, Istio, or serving components is needed.

1. Run the following command to list all inference services and locate the appcrn for your specific instance:

   ```
   cdp ml list-ml-serving-apps
   ```
2. Run the upgrade command using the CRN identified in the previous step:

   ```
   cdp ml upgrade-ml-serving-app --app-crn <app crn in double quotes>
   ```

   For example:

   ```
   cdp ml upgrade-ml-serving-app --app-crn "appcrn": "crn:cdp:ml:us-west-1:
   230228-5d32-4adf-936c-a71364d41ea:mlserving:5eaac17c-df0a-4aa8-bfd5-c07d
   6fdafc74"
   ```

Automatic upgrade of Cloudera AI Inference service during Compute Cluster upgrade

Cloudera AI Inference service upgrade is tied directly to the lifecycle of the Compute Cluster. If Cloudera AI Inference service upgrade is available when you perform a Kubernetes externalized cluster upgrade, it will be applied automatically.

1. Compute Cluster upgrade: Upgrade your Compute Cluster using the Cloudera Management Console for your specific cloud provider.
   - Upgrading Compute Clusters - AWS
   - Upgrading Compute Clusters - Azure
2. Automatic Service update: Once the Compute Cluster upgrade is successfully completed, Cloudera AI Inference service gets automatically upgraded.