

Cloudera AI ..

Upgrading Cloudera AI Inference service

Date published: 2026-01-19

Date modified:

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

Prerequisites for upgrading Cloudera AI Inference service.....	4
Manually Upgrading Cloudera AI Inference service using the UI.....	4
Manually Upgrading Cloudera AI Inference service using the CLI.....	5
Automatic upgrade of Cloudera AI Inference service during Compute Cluster upgrade.....	6

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 you 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-c07d6fdafc74"
```

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.