

Cloudera AI

## Cloudera AI Workbenches

Date published: 2020-07-16

Date modified: 2025-09-30

# CLOUDERA

<https://docs.cloudera.com/>

# 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>Provisioning Cloudera AI Workbenches.....</b>	<b>5</b>
<b>Configuring User access to Cloudera AI.....</b>	<b>8</b>
Granting Cloudera Users access to Cloudera AI Workbenches.....	9
<b>Granting remote access to Cloudera AI Workbenches.....</b>	<b>9</b>
<b>Accessing Cloudera AI Workbenches via SOCKS Proxy.....</b>	<b>10</b>
<b>Monitoring Cloudera AI Workbenches.....</b>	<b>12</b>
<b>Suspend and resume Cloudera AI Workbenches.....</b>	<b>13</b>
<b>Backing up Cloudera AI Workbenches.....</b>	<b>13</b>
Workbench backup and restore prerequisites.....	14
Backing up an Cloudera AI Workbench.....	19
Restoring a Cloudera AI Workbench.....	19
Restoring to a different environment.....	20
<b>Configuring File Storage Replication on AWS.....</b>	<b>21</b>
Enabling File Storage Replication.....	21
Disabling File Storage Replication.....	23
<b>Performing disaster recovery using failover.....</b>	<b>24</b>
<b>Removing Cloudera AI Workbenches.....</b>	<b>25</b>
<b>Upgrading Cloudera AI Workbenches.....</b>	<b>25</b>
<b>Cloudera AI upgrades using Backup/Restore.....</b>	<b>27</b>
Step 1 : Backing up the workbench.....	28
Step 2: Restoring into a new workbench with a different workbench URL/domain endpoint.....	31
Step 3: Delete the backed-up workbench.....	33
Step 4: Restore into a new workbench with same URL/domain endpoint as backed up workbench.....	33
Step 5: Delete the interim restored workbench.....	35
Frequently Asked Questions.....	35

<b>Tagging disks to avoid garbage collection.....</b>	<b>35</b>
<b>Modifying Resource Group Type.....</b>	<b>36</b>
<b>Modifying workbench persistent volume size.....</b>	<b>37</b>

# Provisioning Cloudera AI Workbenches

This topic describes how to provision Cloudera AI Workbenches.

## Before you begin

The first user to access the Cloudera AI Workbench after it is created must have both the MLAdmin role and the EnvironmentAdmin account role assigned. See *Configuring User Access to Cloudera AI* and *Understanding account roles and resource roles* for information about this resource role.

## Procedure

1. Log in to the Cloudera AI web interface.

On on cloud, log in to <https://console.cdp.cloudera.com> using your corporate credentials or any other credentials that you received from your Cloudera administrator.

2. Click Cloudera AI Workbenches.

3. Click Provision Workbench.

4. Fill out the following fields.

- Workbench Name - Give the Cloudera AI Workbench a name. For example, *USER1\_DEV*. Do not use capital letters in the workbench name.
- Select Environment - From the dropdown, select the environment where the Cloudera AI Workbenches must be provisioned. If you do not have any environments available to you in the dropdown, contact your Cloudera administrator to gain access.



**Note:** You cannot choose an environment when the Environment or associated DataLake and FreeIPA is not in an available or running state.

- Existing NFS - (Azure only) Enter the mount path from the environment creation procedure.
- NFS Protocol version - (Azure only) Specify the protocol version to use when communicating with the existing NFS server.

## 5. Switch the toggle to display Advanced Settings.

### a) CPU Settings - From the dropdown, select the following:

- Instance Type: You must select an instance type that is supported by Cloudera AI, or the associated validation check will fail (See *Other Settings*, below).
- Autoscale Range
- Root Volume Size: If necessary, you can also change the default size of the root volume disk for the nodes in the group.

Click Add CPU Resource Group to add new CPU resource group. Provide the name to the resource group, and select the instance type, autoscale range, and root volume size. This resource groups will be available for the workbenches when running applications, jobs, models, and sessions.

### b) GPU Settings - Click the GPU Instances toggle to enable GPUs for the cluster, and set the following:

- Instance Type: You must select an instance type that is supported by Cloudera AI, or the associated validation check will fail (See *Other Settings*, below).
- Autoscale Range
- Root Volume Size: If necessary, you can also change the default size of the root volume disk for the nodes in the group.

Click Add GPU Resource Group to add new GPU resource group. Provide the name to the resource group, and select the instance type, CPU, GPU, and memory. This resource groups will be available for the workbenches when running applications, jobs, models, and sessions.



**Note:** In addition to the CPU and GPU instances selected here, Cloudera AI also provisions two extra CPU instance groups to run infrastructure pods for the Cloudera AI Workbenches, as follows:

AWS:

- Cloudera AI infrastructure node group: m5.2xlarge, with an autoscale range of 2 to 3.
- Platform infrastructure node group: m5.large, with an autoscale range of 2 to 4.

Azure:

- Cloudera AI infrastructure node group: Standard\_D3s\_v2, with an autoscale range of 2 to 3.
- Platform infrastructure node group: Standard\_D2s\_v3, with an autoscale range of 2 to 4.

These are not configurable by users.

### c) Kubernetes Config - Upload or directly enter the Kubernetes config information.

### d) Network Settings

- Subnets for Worker Nodes: (AWS only) Optionally select one or more subnets to use for Kubernetes worker nodes.
- Subnets for Load Balancer: Optionally select one or more subnets to use for the Load Balancer.
- Load Balancer Source Ranges: (Azure only) Enter a CIDR range of IP addresses allowed to access the cluster.
  - If the Cloudera AI Workbench is provisioned with public access, enter the allowed public IP address range.
  - If the Cloudera AI Workbench is provisioned with private access, enter the allowed private IP address range.



**Note:** When you change the Load Balancer Source Range setting, the changes are propagated to both the deployed Load Balancer in EKS and the underlying Security Group (SG).

- Enable Fully Private Cluster: This Preview Feature provides a simple way to create a secure cluster. Only available in AWS environments in Cloudera.
- Enable Public IP Address for Load Balancer

(AWS only) You can create a load balancer with a public IP address for the private cluster. This is useful in cases where there is no VPN between the Cloudera AI VPC and the customer network. In this case, the connection is over the internet.



**Note:** In this network configuration, to use `kubectl` commands in the private cluster, you need to execute the commands in a network that is peered with the cluster VPN. Enabling the public IP address for the load balancer is not sufficient to allow `kubectl` commands to work.

- Restrict access to Kubernetes API server to authorized IP ranges

You can specify a range of IP addresses in CIDR format that are allowed to access the Kubernetes API server. By default, the Kubernetes API services of Cloudera AI Workbenches are accessible to all public IP addresses (0.0.0.0/0) that have proper credentials.

To specify an address to authorize, enter an address in CIDR format (for example, 1.0.0.0/0) in API Server Authorized IP Ranges, and click the plus (+) icon. In this case, the API server is accessible by the user-provided address as well as control-plane-exit-ips over the public internet.

If the feature is enabled and no IP authorized addresses are specified, then the Kubernetes API server is only accessible by control-plane-exit-ips from the public internet.



**Note:** Both the Amazon EKS and Azure AKS have a quota or upper limit for the maximum number of public endpoint access CIDR ranges per cluster. See the [Amazon EKS service quotas](#) or the [Azure AKS documentation](#) for more details. When the feature is enabled, the Cloudera Control Plane exit IP addresses will be automatically added to the authorized IP ranges for accessing the Kubernetes API server for Cloudera operations, which will use three CIDR blocks against the per-cluster limit.

- Use hostname for a non-transparent proxy

Enter a CIDR range allowed for non-transparent proxy server access to the cluster.

- File Storage Replication (AWS only):
  - Enable File Storage Replication: Enables file storage replication for your workbench by creating a duplicate of the project files in a different location.

#### e) Production Cloudera AI

- Enable Governance - Must be enabled to capture and view information about your Cloudera AI projects, models, and builds from Apache Atlas for a given environment. If you do not select this option, then integration with Atlas will not work.
- Enable Model Metrics - When enabled, stores metrics in a scalable metrics store, enables you to track individual model predictions, and also track and analyze metrics using custom code.

#### f) Other Settings

- Enable TLS - Select this checkbox if you want the workbench to use HTTPS for web communication.
- Enable public Internet access - When enabled, the Cloudera AI Workbench will be available on the public Internet. When disabled, it is assumed that connectivity is achieved through a corporate VPC.
- Enable Monitoring - Administrators (users with the MLAdmin role) can use a Grafana dashboard to monitor resource usage in the provisioned workbench.
- Skip Validation - If selected, validation checks are not performed before a workbench is provisioned. Select this only if validation checks are failing incorrectly.
- Tags - Tags added to cloud infrastructure, compute, and storage resources associated with this Cloudera AI Workbench.

Note that these tags are propagated to your cloud service provider account. See *Related information* for links to AWS and Azure tagging strategies.

- Cloudera AI Static Subdomain - This is a custom name for the workbench endpoint, and it is also used for the URLs of models, applications, and experiments. You can create or restore a workbench to this

same endpoint name, so that external references to the workbench do not have to be changed. Only one workbench with the specific subdomain endpoint name can be running at a time.



**Note:** The endpoint name can have a maximum of 15 characters, using alphanumerics and hyphen or underscore only, and must start and end with an alphanumeric character.

- Maximize IOPS and throughput of the root volumes - (AWS only) If selected, the root volumes attached to the worker nodes in AWS will have its IOPS and throughput set to the maximum values, that is, 16000 IOPS and 1000 MIB/s throughput. This will incur additional charges from AWS.

6. Click Provision Workbench.

### Results

It can take up to an hour for a Cloudera AI Workbench to be provisioned and installed. Once the status changes to show that the workbench has been successfully provisioned, click on the workbench name to go to the web application.

Note that the domain name for the provisioned workbench is randomly generated and cannot be changed.

### What to do next

Grant users access to this Cloudera AI Workbench using the instructions at *Configuring User access to Cloudera AI*.

### Related Information

[Configuring User access to Cloudera AI](#)

[Understanding account roles and resource roles](#)

[Best Practices for Tagging AWS Resources](#)

[AWS EKS cluster endpoint access control](#)

[AWS Elastic Kubernetes Service endpoints and quotas](#)

[Create an AKS cluster with API service authorized IP ranges enabled](#)

[Use tags to organize your Azure resources and management hierarchy](#)

[Use a non-transparent proxy with Cloudera AI on AWS environments](#)

## Configuring User access to Cloudera AI

This topic describes how to grant users/groups access to an environment so that they can provision and/or list Cloudera AI Workbenches within that environment. The same users will also be granted Single Sign-on (SSO) access to the workbenches. In addition, if a user needs to provision, upgrade, or delete an Cloudera AI Workbench, they also need the account-level EnvironmentAdmin role assigned. For more information, see [Understanding account roles and resource roles](#).



**Note:** This topic applies to Cloudera AI on cloud releases.

Required Role: PowerUser

There are two Cloudera user roles associated with the Cloudera AI service: MLAdmin and MLUser. A Cloudera PowerUser will need to assign these roles to users who require access to the Cloudera AI service within an environment.

- MLAdmin - This role grants a Cloudera user/group the ability to create and delete Cloudera AI Workbenches within a given Cloudera environment. MLAdmins will also have Site Administrator level access to all the workbenches provisioned within this environment. That is, they can run workloads, monitor, and manage all user activity on these workbenches.
- MLUser - This role grants a Cloudera user/group the ability to list Cloudera AI Workbenches provisioned within a given Cloudera environment. MLUsers will also be able to run workloads on all the workbenches provisioned within this environment.



For instructions, see *Granting Cloudera Users Access to Cloudera AI Service*.

### Related Information

[Understanding account roles and resource roles](#)

## Granting Cloudera Users access to Cloudera AI Workbenches

This topic describes how to grant the MLAdmin and MLUser roles to users/groups that must be allowed to provision/list and access Cloudera AI Workbenches within a specific environment.



**Note:** This topic applies only to on cloud releases.

### Procedure

1. Log in to the Cloudera AI web interface.
2. For a specific environment, grant the MLAdmin and MLUser roles to users/groups that must be allowed to provision/list and access Cloudera AI Workbenches within that environment. In addition, if a user needs to provision, upgrade, or delete an Cloudera AI Workbench, they also need the account-level EnvironmentAdmin role assigned. For more information, see [Understanding account roles and resource roles](#).
  - a) Click Environments.
  - b) Search for the environment and navigate to the environment's Clusters page.
  - c) Expand the Actions dropdown and click Manage Access.
  - d) Search for the user or group that requires access to the Cloudera AI service in this environment and assign one of the following roles to each user/group:
    - MLAdmin - This role grants a Cloudera user/group the ability to create and delete Cloudera AI Workbenches within a given Cloudera environment. MLAdmins will also have Site Administrator level access to all the workbenches provisioned within this environment. That is, they can run workloads, monitor, and manage all user activity on these workbenches.
  - OR
  - MLUser - This role grants a Cloudera user/group the ability to list Cloudera AI Workbenches provisioned within a given Cloudera environment. MLUsers will also be able to run workloads on all the workbenches provisioned within this environment.
  - e) Click Update Roles.
  - f) If necessary, search for and assign the EnvironmentAdmin role in the same way.



**Note:** The first user to log in to an Cloudera AI Workbench must always be a Site Admin (that is, a user with the MLAdmin role assigned to them). If a user assigned the MLUser role attempts to access the workbench first, the web application will display an error.

### Related Information

[Understanding account roles and resource roles](#)

## Granting remote access to Cloudera AI Workbenches

This topic shows you how to allow specific users remote access to the underlying cluster that powers an Cloudera AI Workbench.

### About this task



**Note:** This topic applies to AWS public cloud. On Azure public cloud, a user with the MLAdmin role can download the `kubeconfig` file, and this file alone grants access to any user who has it.

Required Role: MLAdmin

### Before you begin

As part of this process, you will be required to enter the user's Amazon Resource Name (ARN). Make sure you have access to this information before you begin. Either get the ARN from the user OR look up a user's ARN in your AWS account. For the latter, go to your organisation's AWS Account Identity and Access Management (IAM) Users and lookup the user. The ARN is available on their Summary page.

If you are using the AWS CLI, you can run the following command to get the ARN:

```
aws sts get-caller-identity
```

```
#Sample output
{
  "UserId": "ABCDE12345FGHIJKLMNOP6789",
  "Account": "8888888888888888",
  "Arn": "arn:aws:iam::888888888888:user/<USERNAME>"
}
```

### Procedure

1. Log in to the Cloudera AI web interface.
2. Click Cloudera AI Workbenches.
3. Click Actions to expand the dropdown menu.
4. Click Manage Remote Access.
5. Enter the user's ARN, or select the user's name.
6. Click Grant Access.

To remove access for a user, in the Actions column, click Revoke Access next to the user's name.

7. Click Download Kubeconfig.

### What to do next

Send the downloaded Kubernetes config file to the user who has been granted access. To be able to connect to the EKS cluster, they will need to have aws-iam-authenticator installed.

### Related Information

[Installing aws-iam-authenticator \(AWS Documentation\)](#)

## Accessing Cloudera AI Workbenches via SOCKS Proxy

This topic describes how to configure a SOCKS proxy to access Cloudera AI Workbenches on non-publicly routable VPCs. A SOCKS proxy server allows your web browser to connect directly and securely to your Cloudera AI Workbenches without exposing their ports outside the subnet.

### About this task



**Note:** This topic applies to on cloud releases.

## Procedure

1. In the non routable VPC, create an EC2 instance for your SOCKS server (for example, *MY-EC2-SOCKS-SERVER*) with a public IP and an SSH key-pair (for example, *MY-KEY-FILE.PEM*).

Use the AWS documentation to create the EC2 instance: [Getting Started with Amazon EC2 Instances](#)

Depending on whether you want multiple users to share the SOCKS server or have everyone create their own server, pick the SSH key pair for the instance accordingly. More information is available in the AWS documentation: [Amazon EC2 Key Pairs](#).

2. Set up a SOCKS proxy server with SSH to access the EC2 instance, *MY-EC2-SOCKS-SERVER*.

```
nohup ssh -i
    "my-key-file.pem" -CND 8157
    ec2-user@<public ip for my-ec2-socks-server> &
```

- nohup (optional) is a POSIX command to ignore the HUP (hangup) signal so that the proxy process is not terminated automatically if the terminal process is later terminated.
- *MY-KEY-FILE.pem* is the private key you used to create the EC2 instance where the SOCKS server is running.
- C sets up compression.
- N suppresses any command execution once established.
- D 8157 sets up the SOCKS 5 proxy on the port. (The port number 8157 in this example is arbitrary, but must match the port number you specify in your browser configuration in the next step.)
- ec2-user is the AMI username for the EC2 instance. The AMI username can be found in the details for the instance displayed in the AWS Management Console on the Instances page under the Usage Instructions tab.
- <*PUBLIC IP FOR MY-EC2-SOCKS-SERVER*> is the public IP address of the EC2 instance running the SOCKS server.
- & (optional) causes the SSH connection to run as an operating system background process, independent of the command shell. (Without the &, you leave your terminal open while the proxy server is running and use another terminal window to issue other commands.)

3. Configure Your Browser to Use the Proxy. This example uses Google Chrome.

By default, Google Chrome uses system-wide proxy settings on a per-profile basis. To get around that you can start Chrome using the command line and specify the following:

- The SOCKS proxy port to use (must be the same value used in step 1)
- The profile to use (this example creates a new profile)

This creates a new profile and launches a new instance of Chrome that does not interfere with any currently running instance.

- Linux

```
/usr/bin/google-chrome \
--user-data-dir="$HOME/chrome-with-proxy" \
--proxy-server="socks5://localhost:8157"
```

- MacOS

```
"/Applications/Google Chrome.app/Contents/MacOS/Google Chrome" \
--user-data-dir="$HOME/chrome-with-proxy" \
--proxy-server="socks5://localhost:8157"
```

- Windows

```
"C:\Program Files (x86)\Google\Chrome\Application\chrome.exe" ^
--user-data-dir="%USERPROFILE%\chrome-with-proxy" ^
--proxy-server="socks5://localhost:8157"
```

### Results

You shall now be able to navigate to any Cloudera AI Workbench in the browser launched using SOCKS proxy.

When you connect to the Cloudera AI Workbench, the browser actually connects to the proxy server, which performs the required SSH tunneling.

## Monitoring Cloudera AI Workbenches

This topic shows you how to monitor resource usage on your Cloudera AI Workbenches.

### About this task

Cloudera AI leverages Prometheus and Grafana to provide a dashboard that allows you to monitor how CPU, memory, storage, and other resources are being consumed by Cloudera AI Workbenches. Prometheus is an internal data source that is auto-populated with resource consumption data for each workbench. Grafana is a monitoring dashboard that allows you to create visualizations for resource consumption data from Prometheus.

Each Cloudera AI Workbench has its own Grafana dashboard.

### Before you begin

Required Role: MLAdmin

You need the MLAdmin role to view the Workbench details page.

### Procedure

1. Log in to the Cloudera AI web interface.
2. Click Cloudera AI Workbenches.
3. For the workbench you want to monitor, click **Actions Open Grafana**.

### Results

Cloudera AI provides you with several default Grafana dashboards:

- K8s Cluster: Shows cluster health, deployments, and pods
- K8s Containers: Shows pod info, cpu and memory usage
- K8s Node: Shows node CPU and memory usage, disk usage and network conditions
- Models: Shows response times, requests per second, CPU and memory usage for model replicas.

You might choose to add new dashboards or create more panels for other metrics. For more information, see the *Grafana documentation*.

### What to do next



**Note:** Prometheus captures data for the previous two weeks.

### Related Information

[Grafana documentation](#)

[Monitoring and Alerts](#)

## Suspend and resume Cloudera AI Workbenches

Cloud consumption costs are a pain point for many public cloud users. The Cloudera AI Suspend feature allows users to scale down the Kubernetes pods running on Cloudera AI infra and CPU/GPU nodes for a given Cloudera AI Workbench. When the resume operation is performed on the suspended workbench, the suspended pods scale up.

### About this task

A suspended Cloudera AI Workbench has all its autoscaling node groups, except the Platform Infra node group, shrunk to zero instances, thereby saving compute instance costs for the duration the workbench is suspended. However, Kubernetes pods running on Platform Infra nodes continue to run when a workbench is suspended.

When a workbench is suspended, you cannot access the workbench URL, and all associated models, applications, sessions, and jobs also become unavailable. The suspend operation terminates sessions and jobs, so the suspend should be started only after those operations have finished. When the workbench is resumed, models and applications automatically resume operation at the same URLs as before.



**Note:** Make sure that disks are tagged to avoid garbage collection during backup, restore, upgrade, or suspend operations on Cloudera AI Workbenches. For more information, see *Tagging disks to avoid garbage collection*.

### Procedure

1. To suspend a Cloudera AI Workbench, in the workbenches UI, select **Actions Suspend Workbench** for the workbench to suspend. Then click OK to start the suspend process.
2. To resume a Cloudera AI Workbench, in the workbenches UI, select **Actions Resume Workbench** for the workbench to resume. Then click OK to start the resume process.

### Related Information

[Tagging disks to avoid garbage collection](#)

## Backing up Cloudera AI Workbenches

Cloudera AI makes it easy to create machine learning projects, jobs, experiments, ML models, and applications in workbenches. The data and metadata of these artifacts are stored in different types of storage systems in the cloud.

You can backup an Cloudera AI Workbench, and restore it to a new workbench later. The backup preserves all files, models, applications and other assets in the workbench (files are not backed up by Cloudera AI automatically for external NFS-based workbenches). All workbench backups can be viewed in the Workbench Backup Catalog UI.

The Backup and Restore feature gives you the ability to backup all of your data (except files in external NFS-backed workbenches) to protect your machine learning artifacts against disasters. If your Cloudera AI Workbench is backed up, this feature lets you restore the saved data into a new Cloudera AI Workbench so that you can recover your Cloudera AI artifacts as they were saved in the desired backup. The Backup and Restore feature gives the administrator the ability to take “on-demand” backups of Cloudera AI Workbenches. Core services running in the workbench are shut down during the backup process to ensure consistency in the backup data. It is recommended that backups are taken during off-peak hours to minimize user impacts.

The time required to complete backing up a workbench depends on the amount of data to copy. The backup process copies data from both EBS volumes and EFS. In general, the time taken to backup EFS is more significant than for EBS. Due to the incremental nature of backups, the first backup always takes the longest amount of time. Subsequent backups should complete faster as they are built on top of the initial backup copy. For this reason, we recommend that Cloudera AI Workbenches be backed up regularly.

The time to backup EFS is highly dependent on the amount of data, and on the nature and number of files. It is also affected by available bandwidth in the AWS cloud backend. We have seen first-time backup of a 600 GB EFS file system taking around 10 hours. If you have much more than 600 GB on your EFS file system, the default backup timeout of 12 hours may not be long enough. In such cases, we recommend you take your first backup with a lower timeout, such as 2 hours. The Cloudera AI Control Plane may abort the backup due to the timeout expiry. However, the Control Plane does not cancel the underlying backup jobs. You can monitor these backup jobs on the AWS Backup console, and if all eventually complete successfully, you can initiate the backup operation again from the Cloudera AI Control Plane. This should complete in a relatively shorter time, and you will have a good backup copy to restore from if necessary.

There is currently no restriction on the number of backups one can take, and the backup snapshots are retained indefinitely in the backup service vault of the underlying cloud platform. Cloudera AI Workbench backup details are stored in the Workbench Backup Catalog UI in the Cloudera AI control plane, and these entries may be listed, viewed, deleted or restored as desired.

Restoring a backup creates a new Cloudera AI Workbench wherein the restored data is automatically imported. All the projects, jobs, applications, etc., that were in existence during the backup are automatically available in the new workbench. Restoring a Cloudera AI backup provisions a new cluster, and then launches restore jobs to create storage volumes from the backup snapshots. The restore process takes longer than a regular workbench provisioning operation due to the extra work in copying data from backup to the new storage volumes. While backups are incremental, restores are always full-copy restores. The time to restore is dominated by EFS restoration, which takes at least as long as the time to backup the file system. The restored workbench is always created with the latest Cloudera AI software version, which may be different from the Cloudera AI version of the original workbench that was backed up.



**Note:** At this time, the Cloudera AI Workbench Backup and Restore feature is available on AWS, both through the UI and CLI. On Azure, this feature is only available through CLI.



**Note:** Make sure that disks are tagged to avoid garbage collection during backup, restore, upgrade, or suspend operations on Cloudera AI Workbenches. For more information, see *Tagging disks to avoid garbage collection*.



**Note:** A restored workbench from the workbench backup is considered a new separate workbench independent of the original workbench. Users' roles in the Control Plane from the original or backup workbench are not copied and assigned to the restored workbenches by the restore process to avoid security concerns and have to be moved after restore manually if intended.

### Related Information

[Tagging disks to avoid garbage collection](#)

## Workbench backup and restore prerequisites

To backup and restore workbenches, check that the following prerequisites are satisfied.

### AWS backup service opt-in

Login to your AWS account and navigate to the AWS Backup Service console. Make sure the AWS region matches the region where you have your Cloudera AI Workbench. Click on Settings in the navigation pane, and in the Service opt-in table, ensure that EBS and EFS services are enabled for protection by the AWS Backup service, as shown here.

EBS	✔ Enabled
EC2	✔ Enabled
EFS	✔ Enabled

For additional information about this feature, see [July 2022: Cloudera Customer Advisory: The new feature Cloudera AI Backup and Restore on AWS requires changes to IAM permissions in their cross account roles](#) (requires login).

### AWS Cloudera cross-account role permissions

1. Install the Backup IAM Policy. On the AWS console, navigate to the IAM service and click on Policies Create Policy . Click on the JSON tab, and replace the default text with the contents of the following JSON file. ([Click here to download the file](#))

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "backup:*",
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": "backup-storage:*",
      "Resource": "*"
    },
    {
      "Action": [
        "elasticfilesystem:DescribeFilesystems",
        "elasticfilesystem:Backup",
        "elasticfilesystem:DescribeTags"
      ],
      "Resource": "arn:aws:elasticfilesystem:*:file-system/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ec2:DescribeSnapshots",
        "ec2:DescribeVolumes",
        "ec2:describeAvailabilityZones",
        "ec2:DescribeVpcs",
        "ec2:DescribeAccountAttributes",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeSubnets",
        "ec2:DescribePlacementGroups",
        "ec2:DescribeInstances",
        "ec2:DescribeTags"
      ],
      "Effect": "Allow",
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "ec2:CreateTags",
        "ec2:DeleteSnapshot"
      ],
      "Resource": "arn:aws:ec2:*:snapshot/*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "ec2:DeleteSnapshot",
        "ec2:CreateSnapshot"
      ],
      "Resource": [
        "arn:aws:ec2:*:snapshot/*",
```

```

        "arn:aws:ec2:*:*:volume/*"
    ],
    {
        "Action": [
            "ec2:DeleteSnapshot"
        ],
        "Effect": "Allow",
        "Resource": "*",
        "Condition": {
            "ForAnyValue:StringEquals": {
                "aws:CalledVia": [
                    "backup.amazonaws.com"
                ]
            }
        }
    },
    {
        "Action": [
            "tag:GetTagKeys",
            "tag:GetTagValues",
            "tag:GetResources"
        ],
        "Effect": "Allow",
        "Resource": "*"
    },
    {
        "Action": [
            "iam:ListRoles",
            "iam:GetRole"
        ],
        "Effect": "Allow",
        "Resource": "*"
    },
    {
        "Effect": "Allow",
        "Action": "iam:PassRole",
        "Resource": [
            "arn:aws:iam:*:*:role/*"
        ],
        "Condition": {
            "StringLike": {
                "iam:PassedToService": "backup.amazonaws.com"
            }
        }
    },
    {
        "Action": [
            "kms:ListKeys",
            "kms:DescribeKey",
            "kms:GenerateDataKey",
            "kms:ListAliases"
        ],
        "Effect": "Allow",
        "Resource": "*"
    },
    {
        "Action": [
            "kms:CreateGrant"
        ],
        "Effect": "Allow",
        "Resource": "*",
        "Condition": {
            "ForAnyValue:StringEquals": {

```



```

        "kms:EncryptionContextKeys": "aws:backup:backup-vault"
      },
      "Bool": {
        "kms:GrantIsForAWSResource": true
      },
      "StringLike": {
        "kms:ViaService": "backup.*.amazonaws.com"
      }
    },
    {
      "Effect": "Allow",
      "Action": "iam:CreateServiceLinkedRole",
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "iam:AWSServiceName": "backup.amazonaws.com"
        }
      }
    }
  ]
}

```

Save this policy as cml-backup-policy.

2. Install the Restore Policy. On the AWS console, navigate to the IAM service and click on Policies Create Policy . Click on the JSON tab, and replace the default text with the contents of the following JSON file. ([Click here to download the file](#))

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "ec2:CreateVolume",
        "ec2:DeleteVolume"
      ],
      "Resource": [
        "arn:aws:ec2:*::snapshot/*",
        "arn:aws:ec2:*::volume/*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "ec2:CreateTags"
      ],
      "Resource": "arn:aws:ec2:*::snapshot/*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "ec2:DescribeSnapshots",
        "ec2:DescribeVolumes"
      ],
      "Resource": "*"
    },
    {
      "Effect": "Allow",
      "Action": [
        "elasticfilesystem:Restore",
        "elasticfilesystem:CreateFilesystem",
        "elasticfilesystem:DescribeFilesystems",

```

```

        "elasticfilesystem:DeleteFilesystem",
        "elasticfilesystem:TagResource"
    ],
    "Resource": "arn:aws:elasticfilesystem:*:*:file-system/*"
  },
  {
    "Effect": "Allow",
    "Action": "kms:DescribeKey",
    "Resource": "*"
  },
  {
    "Effect": "Allow",
    "Action": [
      "kms:Decrypt",
      "kms:Encrypt",
      "kms:GenerateDataKey",
      "kms:ReEncryptTo",
      "kms:ReEncryptFrom"
    ],
    "Resource": "*",
    "Condition": {
      "StringLike": {
        "kms:ViaService": [
          "ec2.*.amazonaws.com",
          "elasticfilesystem.*.amazonaws.com"
        ]
      }
    }
  },
  {
    "Effect": "Allow",
    "Action": "kms:CreateGrant",
    "Resource": "*",
    "Condition": {
      "Bool": {
        "kms:GrantIsForAWSResource": "true"
      }
    }
  }
]
}

```

Save this as `cml-restore-policy`.

3. Set up the Trust Relationship. AWS Backup service needs to be able to assume the AWS cross-account role that is used by the Cloudera control plane to manage AWS cloud resources. To enable this, add the following trust relationship to your AWS cross-account role's Trust relationships (navigate to the IAM service console, then find your cross-account role by clicking on Roles).

```

{
  "Effect": "Allow",
  "Principal": {
    "Service": "backup.amazonaws.com"
  },
  "Action": "sts:AssumeRole"
}

```

4. Attach the Backup and Restore policies to the Cross-Account Role. While still on the configuration page of your cross-account role in the IAM console, click on the Permissions tab. Click Attach policies to attach the `cml-back-up-policy` and `cml-restore-policy` policies created above. This step ensures that the AWS Backup service will have the necessary permissions to call the EBS and EFS services on behalf of the cross-account role to manage backups.

## Backing up an Cloudera AI Workbench

Backing up an Cloudera AI Workbench preserves all files, models, applications, and other assets in the workbench, although files in external NFS-backed workbenches are not backed up by Cloudera AI automatically.

### Procedure

1. In the Workbenches UI, find the workbench to back up. The workbench must be in the Installation completed state, otherwise backup is disabled.
2. Enter the workbench, and manually stop all workloads (sessions, jobs, applications, and models).  
For external NFS backed workbenches, manually back up the configured external NFS data to another location. This manual backup of the NFS data will be used when this particular backup is restored in future. Ignore this step if the workbench is configured with internal NFS, as internal NFS data is backed up and restored automatically by Cloudera AI.
3. In the Actions menu for that workbench, select Backup Workbench.
4. In the Backup Workbench modal, enter a Backup Name to identify the workbench, and enter an appropriate timeout value.
5. Click Backup to start the process.

### Results

The workbench shuts down, and the backup process begins. The workbench state changes to reflect the ongoing backup progress. If necessary, click Cancel to cancel the backup process. The backup process can take some time to complete, depending on the amount of data to copy.



**Note:** The default timeout is 12 hours. The estimated time to complete a backup (from the cloud provider) is now periodically added to the event logs.

### What to do next

Monitoring event logs

You can monitor the progress of the backup process by checking the event logs. In the Actions menu for the workbench, click View Event Logs, and then on the Events & Logs tab, click View Event Logs again for the latest backup event.

When the backup process completes, the workbench enters the Installation completed state again.

If there were issues during backup, appropriate error messages will be displayed in the event logs. However the workbench will recover from failure and will be reverted back to the original state when backup was triggered.

## Restoring a Cloudera AI Workbench

Restoring a backup creates a new Cloudera AI Workbench, and recreates all of the projects, jobs, applications and so on in the original workbench.

### About this task



**Note:** Restoring a workbench is a non-reversible operation. The restore process overwrites the existing workspace with older backup data. Any data in the running workbench that is not backed up will be lost. To save the current state, take a new backup before proceeding with the restore operation.

Restoring a workbench with multiple CPU and GPU resource groups

During a restore operation, the following is the behavior if you have multiple CPU and GPU resource groups:

- If the backed-up workbench originally had only a single CPU and/or single GPU resource group, the user is restricted to provisioning only single CPU and GPU resource groups during the restore process. The UI will not show the option for multiple groups.
- If the backed-up workbench already contained multiple CPU/GPU resource groups, the customer is free to restore to a multi CPU/GPU environment.
- Rerunning Workloads: If you re-run an existing workload created prior to the upgrade, it will continue to use the resources it was previously assigned. If you want that workload to run on a newly created Resource Group, you must explicitly select the new Resource Group when re-running or re-creating the workload.

### Procedure

1. In the Workbench Backups UI, find the workbench to restore. You can search for the workbench name or CRN. There can be multiple backups for a given workbench.
2. Enter the workbench, and manually stop all workloads (sessions, jobs, applications, and models).  
For external NFS backed workbenches, copy the manual backup of external NFS data (corresponding to this particular backup) to the configured external NFS export. Ignore this step if the workbench is configured with internal NFS, as internal NFS data is backed up and restored automatically by Cloudera AI.
3. Look for the backup to restore, and click Restore. The restore process starts, and the workplace states changes to Creating Workbench.
4. Provision a new workbench that is in the same Cloudera environment as the original workbench.

### Results

The restore process can take some time, depending on the amount of data to copy. When it is complete, you can find the restored workbench in the Workbenches UI.



**Note:** If there is an issue during the restore process, the event log will show the relevant error messages. In case of error, the workbench will not recover from the failure automatically and will not revert back to the original state prior to the restore operation.

### What to do next

#### Monitoring event logs

You can monitor the progress of the backup process by checking the event logs. In the Actions menu for the workbench, click View Event Logs, and then on the Events & Logs tab, click View Event Logs again for the latest backup event.

When the backup process completes, the workbench enters the installation completed state again.

If there were issues during backup, appropriate error messages will be displayed in the event logs. However the workbench will recover from failure and will be reverted back to the original state when backup was triggered.

## Restoring to a different environment

A backup can be restored to a different Cloudera environment, as long as it is within the same AWS account and region. Make sure the following requirements are met:

- Environment roles must be within the same AWS account and region.
- The target environment must have the necessary restore-related permissions, entitlements, and trust relationships.
- Within the environment where the backup is stored, the user must have the `ml/listWorkspaceBackups` permission.
- Within the environment where the workbench will be restored, the user must have the `ml/createWorkspace` permission.

# Configuring File Storage Replication on AWS

File Storage Replication is designed to enhance business continuity, disaster recovery, and operational efficiency for your Cloudera AI workbench. It ensures that an up-to-date copy of your critical project files is maintained. You can manage file storage replication features using the UI or CDP CLI.

## Key Features:

- **Persistent Storage:** Your Cloudera AI workbench utilizes AWS EFS for persistent storage of project files.
- **Availability Zone Redundancy:** Replication within the same region across different Availability Zones helps reduce single points of failure. For AWS EFS, continuous replication can be implemented across Availability Zones within the same region.
- **One-Way Continuous Replication:** AWS EFS provides one-way continuous replication from a source file system to a destination file system, automatically synchronizing all data and metadata changes.
- **Read-Only Replica:** During active replication, the replicated file system is available in read-only mode to prevent modifications that could conflict with the replication process.
- **Minimal Performance Impact:** The replication process is designed to have minimal impact on the performance of your production file systems.
- **Simple Management:** Replication can be easily configured and monitored through the UI or CDP CLI.



**Note:** Duplicating workbench project files to a different location incurs additional cloud costs.

## Enabling File Storage Replication

File storage replication is not enabled by default. You can enable it either during the initial provisioning of a new workbench or by modifying the settings of an existing workbench or during a workbench restore.

### During provisioning a Cloudera AI Workbench

You can select the Enable File Storage Replication option when provisioning a workbench to enable file storage replication.

#### For Using UI

1. In the Cloudera console, click the Cloudera AI tile.  
The **Cloudera AI Workbenches** page is displayed.
2. Click AI Workbench Backups in the left navigation pane.
3. Provision a new Cloudera AI Workbench as explained in the [Provisioning a new Workbench](#) section.
4. Ensure that you select the Enable File Storage Replication option. This action automatically creates a replica of the file system in the same region, thereby enhancing data durability and availability.

#### For Using CDP CLI

When using CDP CLI to provision a workbench, set the `enableFileSystemReplica` value to `true` within the JSON configuration.

```
{
  "environmentName": "eng-ml-dev-env-aws",
  "workspaceName": "createReplicass",
  "disableTLS": false,
  "usePublicLoadBalancer": false,
  "privateCluster": false,
  ...
}
```

```

    },
    "subnetsForLoadBalancers": [],
    "skipValidation": true,
    "disableSSO": false,
    "enableFileSystemReplica": true,
    "xEntitlements": [
      "ML_FILESYSTEM_REPLICA"
    ]
  }
}

```

### Enabling file storage replication in an existing workbench

You can modify the settings of an existing workbench to enable file storage replication.

#### For Using UI

1. In the Cloudera console, click the Cloudera AI tile.  
The **Cloudera AI Workbenches** page is displayed.
2. Click AI Workbench Backups in the left navigation pane.
3. From the Actions menu of the workbench, select View Workbench Details.
4. Select the Enable File Storage Replication option. This action automatically creates a replica of the file system in the same region, thereby enhancing data durability and availability.

The progress of the data replication can be monitored in the Replication Status field. Upon completion of the replication, the replication file system ID, last synchronization time, and replication status are displayed.

#### For Using CDP CLI

Use the createFileReplica method:

```
cdp ml create-file-replica --workspace-crn <value> --profile int
```

```

{
  "workspaceCrn": "sample:CRN",
  "xEntitlements": [
    "ML_FILESYSTEM_REPLICA"
  ]
}

```

### During Workbench restore with replication

When restoring a workbench from a backup or snapshot, you can choose to enable file storage replication. This creates a replica of the restored file system, maintaining high availability and resilience.

#### For Using UI

1. In the Cloudera console, click the Cloudera AI tile.  
The **Cloudera AI Workbenches** page is displayed.
2. Click AI Workbench Backups in the left navigation pane.
3. In the Workbench Backups UI, locate the workbench for which you want to enable file storage replication.  
You can search by workbench name or CRN. A given workbench can have multiple backups.
4. Click Restore. The **Provision Workbench from Backup** window is displayed.
5. In the **Provision Workbench from Backup** window, provide a name for your workbench.
6. In Select Environment, choose your AWS environment. The Advanced Options toggle button is displayed.
7. Toggle the Advanced Options button.

8. Select the Enable File Storage Replication checkbox.
9. Click Provision Workbench.

### For Using CDP CLI

Include the "enableFileSystemReplica": true line within the newWorkspaceParameters object when restoring a workbench.

```
{
  "newWorkspaceParameters": {
    "environmentName": "eng-ml-dev-env-aws",
    "workspaceName": "cus_ws_2_restore",
    "disableTLS": false,
    "usePublicLoadBalancer": false,
    "privateCluster": false,
    "enableMonitoring": true,
    "enableGovernance": false,
    "enableModelMetrics": true,
    "enableFileSystemReplica": true,
    "loadBalancerIPWhitelists": [],
    "whitelistAuthorizedIPRanges": false,
    "authorizedIPRanges": [
    ],
    "provisionK8sRequest": {
      ...
    }
  }
}
```

## Disabling File Storage Replication

You can disable file storage replication on a workbench if redundancy is no longer required.

### For Using UI

1. In the Cloudera console, click the Cloudera AI tile.  
The **Cloudera AI Workbenches** page is displayed.
2. From the Actions menu of the workbench, select View Workbench Details.
3. Toggle the File Storage Replication button. A confirmation pop-up message will appear.
4. Click Yes to disable file storage replication.

The status of replication deletion can be monitored in the Replication Status field.

### For Using CDP CLI

Use the CDP CLI deleteFileReplica method.

```
cdp ml delete-file-replica --workspace-crn <value> --profile int
```

```
{
  "workspaceCrn": "sample:CRN",
  "xEntitlements": [
    "ML_FILESYSTEM_REPLICA"
  ]
}
```

## Performing disaster recovery using failover

Failover in Cloudera AI is a manual disaster recovery process that switches operations to a backup file system if the primary one fails, ensuring continued access to your workbenches. This active-passive solution relies on a CDP CLI command to move your workbench from the unavailable primary system to the replicated secondary system.

Failover is a critical reliability mechanism in file systems and databases that automatically switches operations to a backup system when the primary system fails or becomes unavailable. Cloudera AI supports a failover disaster recovery solution to ensure that Cloudera AI Workbench can be recovered in the event of a catastrophe.

The failover and failback process must be executed manually using a CDP CLI command.

Cloudera AI supports an active-passive disaster recovery solution that can span two file systems. If the primary AWS EFS becomes unavailable, Cloudera AI can be forced to fail over to the secondary (backup) file system. During normal operation, the primary file system is writable, while the backup file system is read-only due to the replication configuration in place.

Run the following CDP CLI command to perform a failover.

During the failover, your workbench is moved from an existing file system to the replication file system that has been created.

```
cdp ml fail-over-file-system --workspace-crn [***WORKSPACE_CRN***]
[--x-entitlements [X_ENTITLEMENTS ...]] [--delete-primary-storage] [--no-delete-primary-storage]
[--cli-input-json CLI_INPUT_JSON]
[--generate-cli-skeleton] [help]
```



**Note:** The `cdp ml fail-over-file-system` command requires the `--workspace-crn` argument.

For example,

```
cdp ml fail-over-file-system --workspace-crn
crn:cdp:ml:us-west-1:csdb-ccce-4f8d-a581-830970ba9808:workspace:678a6b
13-69cc-34ff-a111-f934552afabf --profile int
```

When you run the command, the following failover tasks take place. In the following example, the primary file system is referred to as **F1** and the backup file system as **F2**.

1. [Suspends the workbench.](#)
2. Access points and mount targets are deleted from F1.
3. When the replication configuration is deleted, F2 becomes read-write.
4. Access points and mount targets are created on the replica (F2).
5. F2 is mounted by performing a Helm installation.
6. Workbench is scaled up and brought back to its normal operating state.

Now, F2 becomes the primary file system. From the backend, another replica of F2 (F3) is created. This makes F2 writable and F3 the new replica file system.

Using the CDP CLI, you can delete the corrupted (failed) file system F1 by setting the `deletePrimaryStorage` flag to `true`.



## Removing Cloudera AI Workbenches

This topic describes how to remove an existing Cloudera AI Workbench and clean up any cloud resources associated with the workbench. Currently, only Cloudera users with both the MLAdmin role and the EnvironmentAdmin account role can remove workbenches.

### Procedure

1. Log in to the Cloudera AI web interface.
2. Click Cloudera AI Workbenches.
3. Click on the Actions icon and select Remove Workbench.
  - a) Remove EFS Storage - This option is enabled by default. If you want to retain project files on EFS, disable this property.
  - b) Force Delete - This property is not required by default. You should first attempt to remove your workbench with this property disabled.

Enabling this property will delete the workbench from Cloudera but does not guarantee that the underlying cloud resources used by the workbench will be cleaned up properly. Go to your cloud service provider account to make sure that the cloud resources have been successfully deleted.

When manually cleaning up resources, make sure that the following types of shared resources are not deleted:

AWS:

- VPCs
- Subnets
- Storage (S3 buckets and bucket entries)
- AWS IAM roles

Microsoft Azure

- Virtual networks
- Subnets
- ADLS storage
- Azure resource groups (RGs named <liftie-id> and MC\_<liftie-id>\_<azure-region>)

4. Click OK to confirm.



**Note:** On Azure public cloud, you also need to delete NFS storage after removing the workbench, if the NFS service is no longer needed.

## Upgrading Cloudera AI Workbenches

This topic describes how to upgrade existing Cloudera AI Workbenches. Currently, only Cloudera users with both the MLAdmin role and the EnvironmentAdmin account role can create, upgrade, or remove workbenches.

Existing Cloudera AI Workbenches periodically should be upgraded. Upgrading the workbench upgrades the Cloudera AI software version to the current version, and may also upgrade cluster software. In case the underlying Kubernetes software must be upgraded, a warning banner displays, notifying you that you shall upgrade the workbench promptly.

- During an upgrade, any running models and applications shut down, but they automatically restart after the upgrade is complete.

- To upgrade Kubernetes, only use the upgrade method provided in Cloudera AI. Do not upgrade Kubernetes directly in the cloud console or through the CLI. Follow the instructions here to upgrade Kubernetes. If there is some error, then repeat the instructions. This applies to both Microsoft Azure and AWS.
- You should back up your workbench before starting the upgrade. For more information, see [Backing up Cloudera AI Workbenches](#).

### When is an upgrade necessary?

Cloud service providers define their generally available version of Kubernetes based on their Kubernetes version support policies. For AKS refer to [Supported Kubernetes versions in Azure Kubernetes Service \(AKS\)](#) and for EKS refer to [Amazon EKS Kubernetes release calendar](#).

Cloud service providers may have different deprecation policies for Kubernetes versions:

- For AWS deprecation policy, refer their FAQ section in [Amazon EKS version support and FAQ](#).
- For Azure, refer to the [Azure Kubernetes FAQ](#).

If any Kubernetes version used in your Cloudera AI Workbenches is deprecated by the cloud providers and Cloudera AI upgrades are enabled, the warning banner displays.

ACTION REQUIRED: A new Cloudera AI version is available and it is highly recommended to upgrade to the latest version as soon as possible. To perform an upgrade, select Upgrade Workbench from the Actions menu.

In order to avoid unplanned service interruption caused by the automatic Kubernetes upgrade by EKS and continue to receive support from AKS for your Cloudera AI Workbenches on Azure, it is important to make sure that your Cloudera AI Workbenches are using supported Kubernetes versions. Upgrading a Cloudera AI Workbench will automatically upgrade the Kubernetes to a supported version. We recommend our users to upgrade the Cloudera AI Workbenches promptly when the warning banner appears.



### What type of upgrades does Cloudera AI Support?

In-place Cloudera AI upgrades

Upgrades are done in-place on the existing Cloudera AI Workbenches. This may involve a Kubernetes upgrade (if there is an upgrade available) followed by upgrading the Cloudera AI software.



**Note:** Make sure to backup your workbench before starting the upgrade process. For more information, see [Backing up Cloudera AI Workbenches](#).

1. In the Cloudera console, click the **Cloudera AI** tile.  
The Cloudera AI Workbenches page displays.
2. For a given workbench, click  from the Actions menu and select Upgrade Workbench.
3. Click OK to confirm.
4. In case of an upgrade failure, click  from the Actions menu and select Retry Upgrade Workbench.



**Note:** The upgrade process is estimated to take approximately two to four hours. In case of upgrade retry scenarios, execution will resume from the point of failure in the last upgrade attempt. This is especially beneficial for recovering from interrupted or failed tasks, as it avoids restarting the entire upgrade workflow from the beginning.

Upgrades through Cloudera AI backup & restore

If a Cloudera AI Workbench upgrade from a specific version could not be validated due to Kubernetes version deprecations on cloud providers or is deemed risky, in-place upgrades will be disabled for these versions.

In such cases, depending on the version of Cloudera AI either the upgrade button is disabled or the in-place upgrade pre-flight check will fail, with a failure message pops up that says:

```
In-place upgrades from <EXISTING_VERSION> are not supported. Follow the documentation for the backup based upgrade steps.
```

Here, <EXISTING\_VERSION> is the version number of your workbench.

In this case, it is recommended to go with Backup/Restore to upgrade to the latest Cloudera AI version, essentially performing a workbench upgrade with all your previous data in place. Refer to [Cloudera AI Upgrades using Backup/Restore](#) for more information.



**Note:** Make sure that disks are tagged to avoid garbage collection during backup, restore, upgrade, or suspend operations on Cloudera AI Workbenches. For more information, see [Tagging disks to avoid garbage collection](#).

### Related Information

[Backing up Cloudera AI Workbenches](#)

[Cloudera AI upgrades using Backup/Restore](#)

[Supported Kubernetes versions in Azure Kubernetes Service \(AKS\)](#)

[Azure Kubernetes FAQ](#)

[Amazon EKS Kubernetes release calendar](#)

[Amazon EKS version support and FAQ](#)

[Tagging disks to avoid garbage collection](#)

## Cloudera AI upgrades using Backup/Restore

Cloudera strongly recommends following the Cloudera AI release cadence by upgrading to every version soon after they are released. Following this process ensures that the Cloudera AI Workbench is up to date with the latest security and bug fixes as well to benefit from new feature development. This document will take you through some considerations to be aware of before performing an upgrade, options you have when performing the upgrade, and the steps to complete the upgrade.

### Before you begin

If a Cloudera AI Workbench upgrade from a specific version could not be validated due to Kubernetes version EOL or is deemed risky, in-place upgrades will be disabled for these versions.

In-place upgrades will be disabled from Cloudera AI versions if the underlying Kubernetes versions are deprecated or going to be deprecated very soon. In such cases, depending on the version of Cloudera AI either the upgrade button is disabled or the in-place upgrade pre-flight check will fail, with a failure message pops up that says: In-place upgrades from <existing\_version> are not supported. Follow the documentation for the backup based upgrade steps.

In this case, it is recommended to go with Cloudera AI Backup/Restore to upgrade to the latest Cloudera AI version, essentially performing a workbench upgrade with all your previous data in place. Since a restore always installs the latest Cloudera AI version, it essentially performs a workbench upgrade with all your existing workbench data intact. Backup/Restore is the recommended path to upgrade when a Cloudera AI Workbench cannot be reliably in-place upgraded from its current version.

### Prerequisites

Backup/Restore on AWS


For AWS, Backup/Restore functionality is GA, and is usable from the UI and CDP CLI. The documentation is already available in [Backing up Cloudera AI Workbenches](#). Refer to the documentation for prerequisites for using Backup/Restore on AWS.

Backup/Restore on Azure

Currently, Backup/Restore in Azure is available only through the CDP CLI. Additionally, the Backup/Restore feature does not perform a backup of NFS.

Steps to upgrade workbenches using Backup/Restore

There are five major steps to go through to upgrade older workbenches to the current version. Make sure to go through the following steps in order.

 **Note:** Make sure that disks are tagged to avoid garbage collection during backup, restore, upgrade, or suspend operations on Cloudera AI Workbenches. For more information, see *Tagging disks to avoid garbage collection*.

Related Information

- Upgrading Cloudera AI Workbenches
- Tagging disks to avoid garbage collection

Step 1 : Backing up the workbench

After Step 1, Backing up the workbench, you shall follow steps 2 through 5 in order to restore the workbench.

Backing up AWS workbench

For information on backing up workbenches, see [Backing up Cloudera AI Workbenches](#).

Cloudera AI Workbenches

Status	Version	Workbench	Environment	Region	Creation Date	Cloud Provider	Actions
Ready	2.0.47	qu...	eng...	us-west-2	11/15/2024 5:53 PM IST	aws AWS	<div><div>View Workbench Details</div><div>View Event Logs</div><div>Manage Access</div><div>Manage Remote Access</div><div>Download Kubeconfig</div><div>Open Grafana</div><div>Retry Install Workbench</div><div>Upgrade Workbench</div><div>Suspend Workbench</div><div>Backup Workbench</div><div>Remove Workbench</div></div>
Ready	2.0.47	ns...	eng...	us-west-2	11/15/2024 3:45 PM IST	aws AWS	
Removing Workbench	2.0.47	qua...	eng...	us-west-2	11/15/2024 2:54 PM IST	aws AWS	
Suspended	2.0.46	yun...	eng...	us-west-2	11/15/2024 2:29 PM IST	aws AWS	
Suspended	2.0.46	yun...	eng...	us-west-2	11/15/2024 2:27 PM IST	aws AWS	
Ready	2.0.46	e...	ai...	us-west-2	11/15/2024 1:26 PM IST	aws AWS	
Ready	2.0.47	ns...	eng...	us-west-2	11/15/2024 12:07 PM IST	aws AWS	
Ready	2.0.47	cm...	eng...	us-west-2	11/15/2024 11:15 AM IST	aws AWS	
Suspended	2.0.47	sg...	eng...	us-west-2	11/15/2024 10:58 AM IST	aws AWS	
Suspended	2.0.47	rar...	eng...	us-west-2	11/15/2024 10:56 AM IST	aws AWS	

## Backup Workbench



- The backup process will **Shut Down** the Workbench.
- A backup vault will be created, unless one already exists.
- The backup process will take several minutes.
- You can provide timeout in minutes for the backup operation.
- The default timeout for backup operation is 12 hours.
- Note that the workbench will be **unavailable** for end-users for up to 12 hours if backed up.

\* Add **Backup Name** to help identify this backup

Backup Timeout in Minutes (optional)

☐ Skip Validation

Note: The workbench **CRN** is the only unique identifier of the workbench so you may want to take note of it. You may need it to identify the backup.

CRN    crn:cdp:ml:us-west-1:9d74... 

Cancel

Backup

The time required to backup or restore AWS based workbenches mainly depends on the size of EFS (File System for projects storage) associated with the workbench. To get the EFS ID associated with the workbench, click on View Workbench Details from the UI and note the Filesystem ID. The size associated with the EFS can be retrieved from the AWS console using the Filesystem ID.

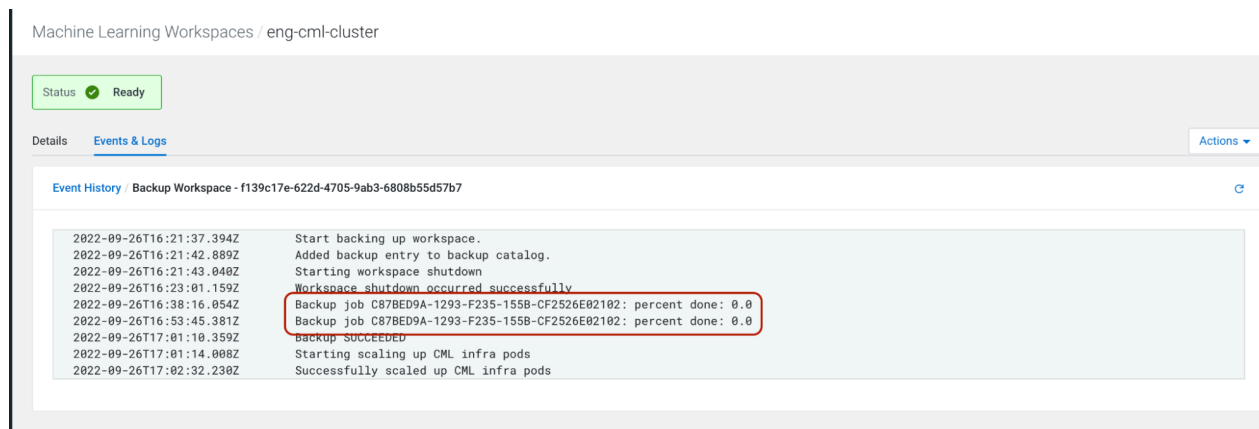
The Backup timeout for storage volumes (EFS and EBS) is by default set to 12 hours, but the user can customize the timeout. While clicking on the Backup Workbench, there will be an option to specify Backup Timeout (in minutes), to accommodate users with large EFS size.

Similarly, if the user suspects that the EFS is too large to be restored within the default 12 hours, custom restore timeouts can be set in the advanced settings during restoring an AWS workbench from a backup snapshot through the UI.

### Tracking running backups on AWS

For AWS based backup jobs, Cloudera AI prints out backup job completion percentage for all backup jobs associated with a backup snapshot at an interval of 15 minutes as part of event logs. However, the completion percentage is an estimate returned by AWS APIs, and Cloudera AI just surfaces the same. Cloudera AI is not running any mechanisms/heuristics to calculate the completion percentage for a particular backup.

From our experience, we have seen that AWS-provided completion percentages can vary wildly, jump abruptly and can be downright misleading. Cloudera AI advises to take the percentage numbers with a grain of salt.



### Canceling long running backups on AWS

Backups generally take a long time if you have lots of data in EFS. This is expected behavior. However if the backup is taking longer than expected, you can cancel the backup jobs from AWS dashboard and Cloudera AI will detect this in a while, and will fail the backup.

To cancel the corresponding AWS backup jobs from AWS dashboard:

- Go to **AWS Backup Jobs Backup Jobs** and identify the running backups associated with the backup snapshot. The backups should have started at approximately the same time you triggered the backup from the Cloudera AI console.
- Abort all such backup jobs.

You can retry the backups again using the above mentioned steps for backing up the workbench.

## Backing up Azure workbench

### Prerequisites

There are a few prerequisites to Azure Backup/Restore:

- If your environment is configured with a pre-existing resource group, then Cloudera AI backup service would use the same resource group for taking snapshots of Azure Disks. Else, please ensure that you have a resource group created in your Azure Account with the nomenclature `cml-snapshots-<azure_region>`. For example, if your Azure workbench resides in the `westus2` region, there should be a resource group present named `cml-snapshots-westus2`.
- Please refer to Azure documentation for roles needed to perform a backup: [Use Azure role-based access control to manage Azure Backup recovery points](#).

### Suspending the workbench

Suspend the workbench to ensure correctness of data in NFS on Azure during backup. Suspend the workbench by clicking on the **Suspend Workbench** option for the workbench and wait for the suspend operation to complete successfully.

Since the workbench is now in a suspended state, it is now guaranteed that no writes/mutations are happening on the NFS or Azure disks associated with the workbench.

### Invoking Backup

Once you have the prerequisites sorted, please note the workbench CRN from the View Workbench Details page, and run the following command from ClouderaCLI to initiate workbench backup. Please replace the values in brackets with your own values.

```
$ cdp ml backup-workspace --workspace-crn <crn:cdp:ml:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:workspace:792f8cfc-ba33-428d-9c80-b9bc6e799ce9> --backup-name <name-of-backup-for-upgrade>
```

```
---
{
  "backupCrn": "crn:cdp:ml:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:workspace_backup:b6cee77e-9e38-4e30-9d72-481088f43de0"
}
```

Please note the backup CRN returned from the CLI call, as it is the backup snapshot which will be used to restore into a new workbench.

Use these variables to restore into a new workbench:

- backupCRN: The CRN returned in the response of CLI call for backup-workbench
- existingNFS: The existing NFS server path can be retrieved from View Workbench Details Filesystem ID .
- existingNFSVersion: The existing NFS version can be retrieved from View Workbench Details NFS Protocol version .



**Note:** Please do not perform any operations with the existing NFS in use with the suspended workbench as we will attach the same NFS in the new (restored) workbench.

## Step 2: Restoring into a new workbench with a different workbench URL/domain endpoint

Restore into a new workbench with useStaticSubdomain set to false in the Cloudera CLI. This brings up a workbench with a different URL/domain endpoint from the original workbench that was backed up. This step is needed to ensure that we can safely validate that restoration of the workbench is successful before executing Step 4 below to delete the original workbench. If any of the following steps fail, please contact your customer support representative.

### Restore on AWS

For information on this step, see [Restore an Cloudera AI Workbench](#).

The screenshot shows the Cloudera AI Backup Catalog interface. On the left, the 'AI Workbench Backups' option is highlighted in the sidebar. The main area displays a table of backups with columns for Workbench, CRN, Environment, and Backup Vault. A detailed view of a backup is shown on the right, including the backup status (Ready), name, date, creator, and version. The 'Restore' button is highlighted with an orange box.

The UI for restore is quite similar to the Provision Workbench UI and shall be familiar.

## Restore on Azure

To restore into a new workbench from the backup taken above, please run the following Cloudera CLI command. The workbench provisioning parameters of the request needs to be configured according to your needs. Please ensure that no “write operations” are undertaken on this restored workbench since we will be using the same NFS in Step 5. This is to ensure that there is no state mismatch between the restored Azure disks and the NFS.

```
$ cdp ml restore-workspace --cli-input-json '{
  "newWorkspaceParameters": {
    "environmentName": "eng-ml-dev-env-azure",
    "workspaceName": "new-workspace",
    "disableTLS": false,
    "usePublicLoadBalancer": false,
    "enableMonitoring": true,
    "enableGovernance": true,
    "enableModelMetrics": true,
    "whitelistAuthorizedIPRanges": false,
    "existingNFS": "<existingNFS>",
    "nfsVersion": "<existingNFSVersion>",
    "provisionK8sRequest": {
      "instanceGroups": [
        {
          "instanceType": "Standard_DS3_v2",
          "rootVolume": {
            "size": 128
          },
          "autoscaling": {
            "minInstances": 1,
            "maxInstances": 10
          }
        }
      ],
      "environmentName": "eng-ml-dev-env-azure",
      "tags": [],
      "network": {
        "topology": {
          "subnets": []
        }
      }
    },
    "backupCrn": "<backupCRN>",
    "useStaticSubdomain": false
  },
  "workspaceCrn": "crn:cdp:ml:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:workspace:081ee5d2-4e82-487c-9404-2537a0ab4019"
}
```

Wait for the restore operation to succeed.

After restore, please login to the newly created workbench and verify that all projects from the original workbenches are available. Do not launch any sessions or applications, create new projects, or otherwise make any changes to the workbench, as that can make changes to the NFS file system that will be incompatible with what will be restored in step 4 below.



### Related Information

[Restoring a Cloudera AI Workbench](#)

## Step 3: Delete the backed-up workbench

If Step 2 completes successfully, then delete the old (original) workbench.

After all is validated, and you confirm that the projects are in place, delete the original backed up workbench. To do so, from the UI, select **Remove Workbench**.

## Step 4: Restore into a new workbench with same URL/domain endpoint as backed up workbench

If Step 2 and 3 complete successfully, then restore into a new workbench with the same URL/endpoint as the backed-up workbench.

Since a restored workbench is a brand new workbench with data from an older workbench, a restored workbench gets a new subdomain by default. This means that any endpoints (for models, applications, etc.) that you were using from the old workbench is not valid.

To maintain the endpoints that were configured with the older workbench, check the option `useStaticSubdomain` in the restore payload to provision the new restored workbench with the same URL as the older one. Additionally, `Use Static Subdomain` is also provided as a checkbox in the Restore UI.

### Restore on AWS

To restore a workbench, see [Restore as Cloudera AI Workbench](#). While restoring, please ensure that `Use Static Subdomain` is checked in the restore UI.

## Provision Workbench from Backup

☐ Enable Fully Private Cluster ⓘ  
☐ Enable Public IP Address for Load Balancer ⓘ  
☐ Restrict access to Kubernetes API server to authorized IP ranges ⓘ  
☐ Use hostname for a non-transparent proxy ⓘ  
**Production Machine Learning**  
☐ Enable Governance ⓘ  
☒ Enable Model Metrics ⓘ  
**Other Settings**  
☒ Enable TLS ⓘ  
☒ Enable Monitoring ⓘ  
☐ Skip Validation ⓘ  
 Tags ⓘ  

- +

☐ Use Static Subdomain ⓘ

Restore Timeout in Minutes ⓘ

## Restore on Azure

To restore into a new workbench from the backup taken above, please run the following Cloudera CLI command. You need to tune various parameters of the request to suit workbench configuration needs.

```
$ cdp ml restore-workspace --cli-input-json '{
  "newWorkspaceParameters": {
    "environmentName": "eng-ml-dev-env-azure",
    "workspaceName": "new-workspace",
    "disableTLS": false,
    "usePublicLoadBalancer": false,
    "enableMonitoring": true,
    "enableGovernance": true,
    "enableModelMetrics": true,
    "whitelistAuthorizedIPRanges": false,
    "existingNFS": "<existingNFS>",
    "nfsVersion": "<existingNFSVersion>",
    "provisionK8sRequest": {
      "instanceGroups": [
        {
          "instanceType": "Standard_DS3_v2",
          "rootVolume": {
            "size": 128
          },
          "autoscaling": {
            "minInstances": 1,
            "maxInstances": 10
          }
        }
      ]
    }
  }
}
```

```

    }
  },
  "environmentName": "eng-ml-dev-env-azure",
  "tags": [],
  "network": {
    "topology": {
      "subnets": []
    }
  },
},
"backupCrn": "<backupCRN>",
"useStaticSubdomain": true
}
,

--
{
  "workspaceCrn": "crn:cdp:ml:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:workspace:081ee5d2-4e82-487c-9404-2537a0ab4019"
}

```

### Related Information

[Restoring a Cloudera AI Workbench](#)

## Step 5: Delete the interim restored workbench

The upgraded workbench which was restored from the backup, in order to check the sanity of the restored workbench in Step 2, can now be safely deleted. Please identify the workbench from your control plane UI and delete the same.

## Frequently Asked Questions

Some frequently asked questions about upgrading workbenches with the Backup/Restore feature.

### How long does it take for a Cloudera AI Workbench to be backed-up and/ or restored?

Cloudera AI relies on cloud provider's native services for backup/ restore. Time consumed for backup and restore depends on multiple factors such as infrastructure, network latency, data size, file structure, number of files etc and can vary across workbenches. For internal test parameters, the backup of 600GB data took approximately 10 hours on AWS.

### What happens to customizations done on Kubernetes Clusters during Restore?

Cloudera AI does not support applying customizations during Backup and Restore. All customizations will have to be applied through automation or manually post Cloudera AI Workbench Restore.

## Tagging disks to avoid garbage collection

During backup, restore, upgrade, or suspend operations on Cloudera AI Workbenches, EBS or Azure disks can be left in a temporarily detached or unmanaged state when the associated EKS or AKS cluster is still running. If there is a garbage collection script running, and it is not properly configured, the disks used by the workbench can be deleted unintentionally. If the disk is not backed up, then the data will be lost.

Garbage collection scripts need to check for the following tags and ignore disks that are tagged with the corresponding values.

Cloud	Tag key	Tag value
Azure	k8s-azure-created-by	kubernetes-azure-dd
AWS	kubernetes.io/cluster/liftie-*	owned

## Modifying Resource Group Type

You can easily add, edit, or delete the CPU and GPU resources of Cloudera AI Workbenches, which is beneficial for optimizing performance and cost.

Key Benefits of Modifying Instance Groups:

1. **Scalability and Flexibility:** Scale up or down to meet user workload needs, handle peak traffic or save costs during off-peak periods.
2. **Cloud Provider Compatibility:** We understand that cloud providers may retire or end-of-service (EOS) certain instance types. Our Modify Resource Group feature takes this into account and allows you to seamlessly adapt to changes in the cloud provider's offerings, ensuring your Cloudera AI Workbench stays up-to-date.

Currently, Instance Group modification is only supported for CPU and GPU Worker resource groups.



**Note:** As cloud platforms do not support heterogeneous instance types within a single instance group, the current Modify Instance Group Type workflow involves deleting the existing instance group and recreating it with the desired instance type. However, it's important to note that this process may disrupt user workloads running in the user namespace of the Cloudera AI Workbench, including user-created sessions, jobs, models, and applications.

### Modify the Resource Group from the Cloudera CLI

The following example shows how to modify the instance group from the command line.

```
cdp ml modify-cluster-instance-group --workspace-crn <workspace-crn>
  --instance-group-name <instance-group-name> -instance-type <instance-type>
e>
```

### Modify the Resource Group from the UI

1. Go to the Workbench Details page.
2. Navigate to the Workbench Instances section on the Workbench Details page.

## 3. Click Add CPU Resource

## Add CPU Resource Group

\* Enter Resource Group Name

Default CPU Group

Instance Type

m5.2xlarge

8 CPU

32 GiB

Autoscale Range



0

50

Root Volume Size

128

Add

Cancel

Group.

4. Specify the Resource Group name, instance type, autoscale range, and root volume size.
5. Click Save.
6. Click Add GPU Resource Group.
7. Specify the Resource Group name, instance type, autoscale range, and root volume size.
8. Click Save.
9. Click Edit to edit the resource group name, instance type, and autoscale range.
10. Click Delete to delete a resource group. A confirmation box will appear. Click OK button to delete the chosen resource group.



**Note:** During the modification process, there may be errors such as Bad Gateway or 404 Page not found. In this case, the applications can fail and may need to be restarted.

## Modifying workbench persistent volume size


You can increase the size of a Persistent Volume Claim (PVC) associated with your Cloudera AI Workbench for enhanced data size requirements. You can modify the size either using the Cloudera AI user interface (UI) or CDP CLI.

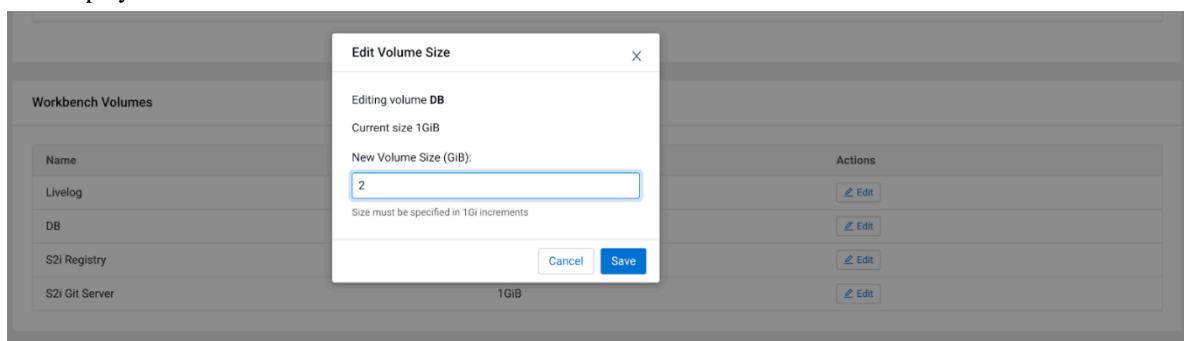


**Note:** Decreasing the volume size is not allowed.

The feature to change persistent volume size is currently unavailable for private clusters.

### For Using UI

1. In the Cloudera console, click the Cloudera AI tile.  
The **Cloudera AI Workbenches** page displays.
2. In the Cloudera AI Workbenches page, click  from the Actions menu next to the desired Cloudera AI Workbench.
3. Click View Workbench Details. The Workbench Details page displays.
4. Scroll down to the Workbench Volumes section.
5. Click the Edit button next to the PVC for which you want to increase the size. The Edit Volume Size dialog box displays.



6. Enter the new volume size in the New Volume Size (GiB) field. You can increase the volume size in increments of 1 GiB.
7. Click Save.

### For Using CDP CLI

Modify the block volume size using the following command:

```
cdp ml modify-block-volume --workspace-crn [***workspace-crn***] --block-volume-specifications persistentVolumeClaim=[***pvc-name***],namespace=[***pvc-namespace***],size=[***pvc-size***]
```

Example:

```
cdp ml modify-block-volume --workspace-crn crn:cdp:ml:us-west-1:9d74eee4-1cad-45e7-we45-7ccf9edbb73d:workspace:a92327e8-8ec3-412c-bad5-174c247 -block-volume-specifications persistentVolumeClaim=my-pvc,namespace=my-namespace,size=600GiB
```