Cloudera AI

Site Administration

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

https://docs.cloudera.com/

Legal Notice

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.

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

Managing Users

This topic describes how to manage a Cloudera AI Workbench as a site administrator. Site administrators can monitor and manage all user activity across a workbench, add new custom engines, and configure certain security settings.

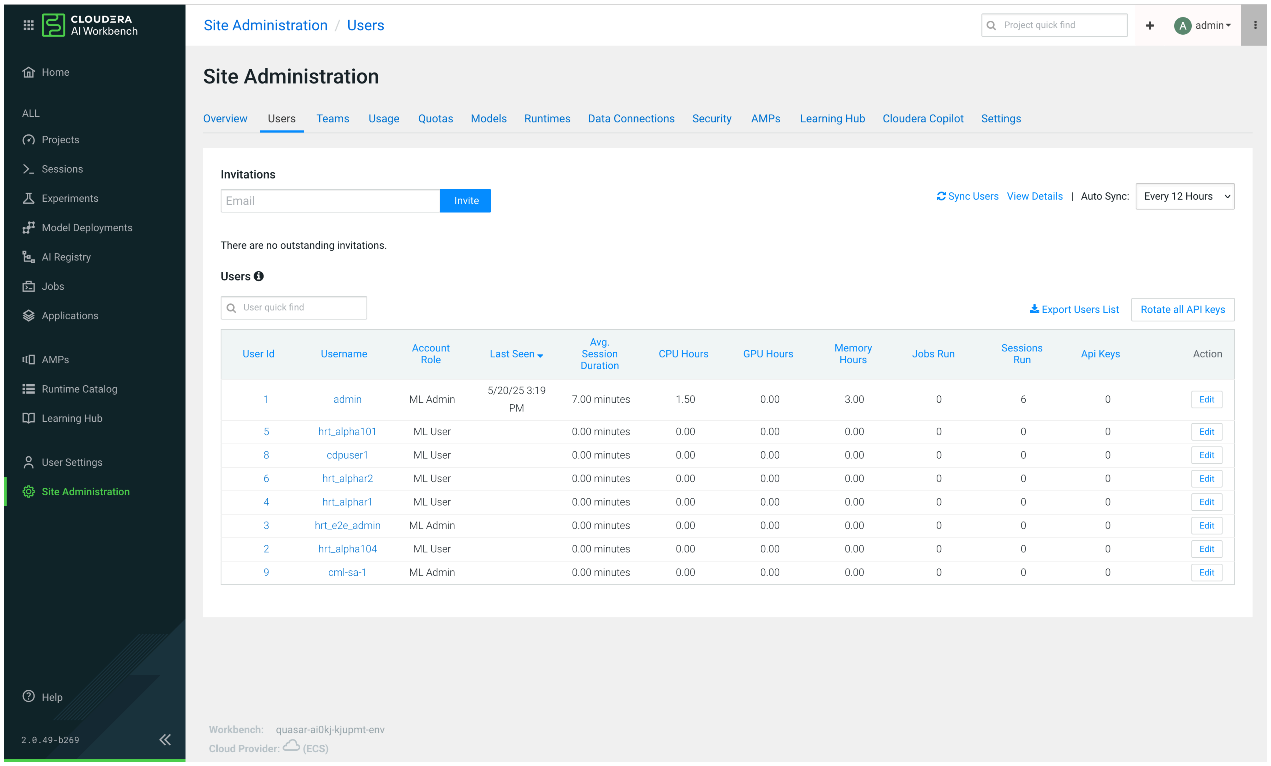
By default, the first user that logs in to a workbench must always be a site administrator. That is, they must have the MLAdmin role granted by a Cloudera PowerUser.

Important: Site administrators have complete access to all activity on the deployment. This includes access to all teams and projects on the deployment, even if they have not been explicitly added as team members or collaborators.

Only site administrators have access to a Site Administration dashboard that can be used to manage the workbench. To access the site administrator dashboard:

1. Go to the Cloudera AI web application and log in as a site administrator.
2. On the left sidebar, click Site Administration. You will see an array of tabs for all the tasks you can perform as a site administrator.

Figure 1: Managing users and teams as a Site Administrator



Site Administration

Invitations

Auto Sync: Every 12 Hours

There are no outstanding invitations.

Users

User Id	Username	Account Role	Last Seen	Avg. Session Duration	CPU Hours	GPU Hours	Memory Hours	Jobs Run	Sessions Run	Api Keys	Action
1	admin	ML Admin	5/20/25 3:19 PM	7.00 minutes	1.50	0.00	3.00	0	6	0	Edit
5	hrt_alpha101	ML User		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit
8	cdpuser1	ML User		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit
6	hrt_alphar2	ML User		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit
4	hrt_alphar1	ML User		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit
3	hrt_e2e_admin	ML Admin		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit
2	hrt_alpha104	ML User		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit
9	cml-sa-1	ML Admin		0.00 minutes	0.00	0.00	0.00	0	0	0	Edit

Workbench: quasar-ai0kj-kjupmt-env
Cloud Provider: (ECS)

Monitoring Users

The Users tab on the **Administrator** dashboard displays the complete list of users. You can see which users are currently active, and when a user last logged in to Cloudera AI. You can search for a user by entering their User ID, Username, or Email in the User quick find box. To modify a user's username, email or permissions, click the Edit button under the **Action** column.

Note: The Disabled checkbox does not have any effect when external authentication is in use.

Synchronizing Users

You can synchronize users within an Cloudera AI Workbench with those users that have been defined access at the Environment level (through the MLAdmin and MLUser roles). Doing so for new users enables you to take administrative actions such as setting Team assignments, defining Project Collaborators, and more, all prior to the new users' first time logging in to the Workbench.

To synchronize users, go to Site Administration Users , and click Sync in the top right corner. This adds any users defined at the Environment level to the workbench, updates any role changes that have been made, and deactivates any users that have been deactivated.

Note: The Administrator shall periodically perform user synchronization to ensure that users who are deactivated on the environment level are also deactivated in Cloudera AI.

To improve the efficiency and usability of the Auto-Synchronization feature, it is enabled by default for users, with a synchronization interval of 12 hours. You have the option to enable or disable the Auto-Synchronization feature as needed. While the synchronization process is in progress, the Sync button is disabled to prevent initiating another synchronization operation.

Synchronizing Groups

Groups of users can be created in the Cloudera Management Console and imported to Cloudera AI. However, changes made in the Cloudera Management Console do not automatically update in Cloudera AI.

To improve the efficiency and usability of the Auto-Synchronization feature, it is enabled by default for teams, with a synchronization interval of 12 hours. You have the option to enable or disable the Auto-Synchronization feature as needed. While the synchronization process is in progress, the Sync button is disabled to prevent initiating another synchronization operation.

Related Information

Cloudera AI email notifications

Creating a Team

Service Accounts

Service accounts are used by machine users that require a user account, without the need of using an account of an actual user.

Like other users, this machine user can be granted necessary permissions and roles, and be added as a collaborator to projects in order to run workloads. Machine users can also create projects and workloads.

Creating a machine user and synchronizing to workbench

The MLAdmin role is required to create machine users.

Procedure

1. In Management Console, go to User Management.
2. In Actions, click Create Machine User.
3. Enter a name for the machine user and click Create.
4. In the Cloudera AI Workbenches UI, find your workbench and in Actions, click Manage Access.

5. Search for the machine user name you just created, and in Update Resource Roles, assign the MLWorkspaceAdmin or MLWorkspaceUser role. Click Update Roles.

   Note: Machine users can alternatively be assigned to environments.

6. Return to the workbench in Cloudera AI, and in Site Administration Users , click Run Sync Now to manually synchronize the users for the workbench.
7. In Site Administration, search for the machine user name.

Synchronizing machine users from the Synced team

You can synchronize machine users that are part of a synced team to your project.

Procedure

1. In Management Console User Management Groups , click Create Group.
2. Enter the name for the group, and click Create.
3. Click Add Members to search for and add group members, including machine users.
4. To add the team (group) to your environment, go to Environments Actions Manage Access .
5. Click Update Role to update the role as follows, and click Update Roles.
   - Environment User: Only users who have read access to the environment are synced. Alternatively, you can assign the Environment User role to the machine user.
   - MLAdmin or MLUser role: only users with either role are synced to Cloudera AI Workbench.
6. Click Synchronize Users and wait for synchronization to complete. Then return to your Cloudera AI Workbench.
7. In Site Administration Teams , select Sync Teams and then choose the group to synchronize.
8. Click Create Team, and the team is created in Cloudera AI.

What to do next

To add members to a synced team, add them in the control plane and synchronize them to Cloudera AI via the Site Administration Teams Sync Teams option. You cannot add users to a group manually in Cloudera AI.

To add the service user as a collaborator of the project, see the instructions in Adding a collaborator and further details in Adding project collaborators.

Running workloads using a service account

You can run various types of workloads using a service account. First, make sure the service account is available in your project.

1. Create a project, or enter an existing project.
2. In Collaborators, add the service account. Specify the Operator or Admin role and click Add.

Run a job with a service account

1. In Jobs, click New Job.
2. For Run Job as, select Service Account and choose the account from the list.
3. Make other settings as needed, and click Create Job.

Run an application with a service account

1. Click New Application.
2. For Run Job as, select Service Account and choose the account from the list.

3. Make other settings as needed, and click Create Application.

Run a model with a service account

1. In Models, click New Model.
2. For Deploy Model as, select Service Account and choose the account from the list.
3. Make other settings as needed, and click Deploy Model.

Authenticating Hadoop for Cloudera AI service accounts

In Cloudera AI, the Kerberos principal for the Service Account may not be the same as your login information. Therefore, ensure you provide the Kerberos identity when you sign in to the Service Account.

Authenticate the Service Account by completing the following procedure:

1. Navigate to your Cloudera AI Workbench.
2. Click Site Administration on the left sidebar.
3. Click the Users tab from the list of tabs displayed under Site Administration.
4. Select the Service Account from the list of users.
5. To authenticate, either provide the password in the Credentials' password field or click Upload Keytab to upload the Keytab file directly.

Once successfully authenticated, Cloudera AI uses stored Service Account credentials to ensure you are secure when running workloads.

Configuring Quotas

This topic describes how to configure CPU, GPU, and memory quotas for users of an Cloudera AI Workbench.

Before you begin

Required Role: MLAdmin

Note: On on premises, the corresponding role is EnvironmentAdmin.

Make sure you are assigned the MLAdmin role in Cloudera. Only users with the MLAdmin role will be logged into Cloudera AI Workbenches with Site Administrator privileges.

There are two types of quota: Default and Custom. Default quotas apply to all users of the workbench. Custom quotas apply to individual users in the workbench, and take precedence over the default quota.

Procedure

1. Log in to the web interface.
2. Click Cloudera AI Workbenches, then open the workbench for which you want to set quotas.
3. Click AdminQuotas.
4. Switch the Default Quotas toggle to ON.

   This applies a default quota of 2 vCPU and 8 GB memory to each user in the workbench.

   If your workbench was provisioned with GPUs, a default quota of 0 GPU per user applies. If you want users to have access to GPUs, you must modify the default quotas as described in the next step.
5. If you want to change the default quotas, click on Default (per user) .

   Cloudera AI displays the Edit default quota dialog box.
6. Enter the CPU, Memory, and GPU quota values that should apply to all users of the workbench.

7. Click Update.
8. To add a custom quota for a specific user, click Add User.
9. Enter the user name, and enter the quotas for CPU, Memory, and GPU.
10. Click Add.

Results

Enabling and modifying quotas will only affect new workloads. If users have already scheduled workloads that exceed the new quota limits, those will continue to run uninterrupted. If a user is over their limit, they will not be able to schedule any more workloads.

What to do next

To specify the maximum number of replicas in a model deployment, go to Site Administration Settings Model Deployment Settings . The default is 9 replicas, and up to 199 can be set.

Non-user dependent Resource Usage Limiting for workloads

The Resource Usage Limiting feature enables the utilization of CPU and memory independent of the user, in a way that no resource is unnecessarily blocked.

If the Resource Usage Limiting feature is enabled, Administrators can define the standard limits of CPU and memory resources applicable to any started workload, that is, to any started job or session. When a user starts the workload, the standard resource limit set by the administrator will be used for the workload pod’s resource requests. However, the workload pod resource limits will be set based on the Resource Profile selected by the user when starting a workload.

Example for enabled Resource Usage Limiting feature

If the Administrator configures the standard resource limit in 100m CPU and 0.25 GiB memory, the user will be able to create the workload using this configured standard resource, that is 100m CPU and 0.25 GiB memory. However, the resource limits used for workload pods will be the resource profile selected, that is 8 CPU and 16 GiB memory.

Example for disabled Resource Usage Limiting feature

If the Resource Usage Limiting feature is not enabled, the user can create a workload with the capacity configured in the Resource profile, that is, for example, 8 CPU and 16 GiB memory. In this case, even though the workload uses only 200 m CPU and 0.5 GiB memory, it will block the whole capacity allocated in the resource profile. The unused resources in that case are blocked by the workload.

Note: If burstable CPU is enabled, the resource limits for the CPU will not be set from the Resource Profile.

Setting Resource Usage Limiting for workloads

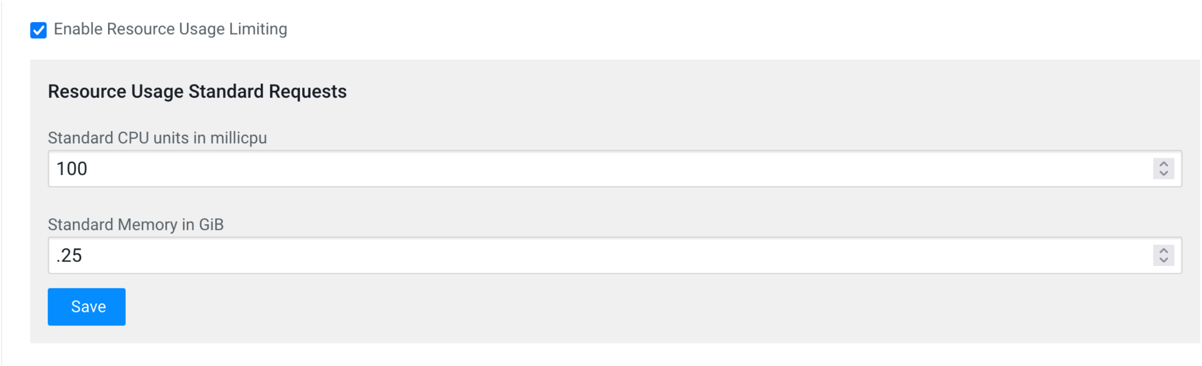
To enable Resource Usage Limiting follow the instructions.

Procedure

1. In the **Cloudera** console, select the **Cloudera AI** tile.
2. Click the **Cloudera AI Workbenches**. The Cloudera AI Workbenches page displays.
3. Click **Site Administration** on the left sidebar. You will see an array of tabs for all the tasks you can perform as a site administrator.
4. Select **Settings**.

5. Select **Enable Resource Usage Limiting**.
6. Define the values for capacity.

   Figure 2: Setting Resource usage limits



☑ Enable Resource Usage Limiting

Resource Usage Standard Requests

Standard CPU units in millicpu

100

Standard Memory in GiB

.25

7. Click **Update**.

Creating Resource profiles

Resource profiles define how many vCPUs and how much memory the product will reserve for a particular workload (for example, session, job, model).

About this task

As a site administrator you can create several different vCPU, GPU, and memory configurations which will be available when launching a session/job. When launching a new session, users will be able to select one of the available resource profiles depending on their project's requirements.

Procedure

1. To create resource profiles, go to the Site Administration Runtime/Engine page.
2. Add a new profile under Resource Profiles.

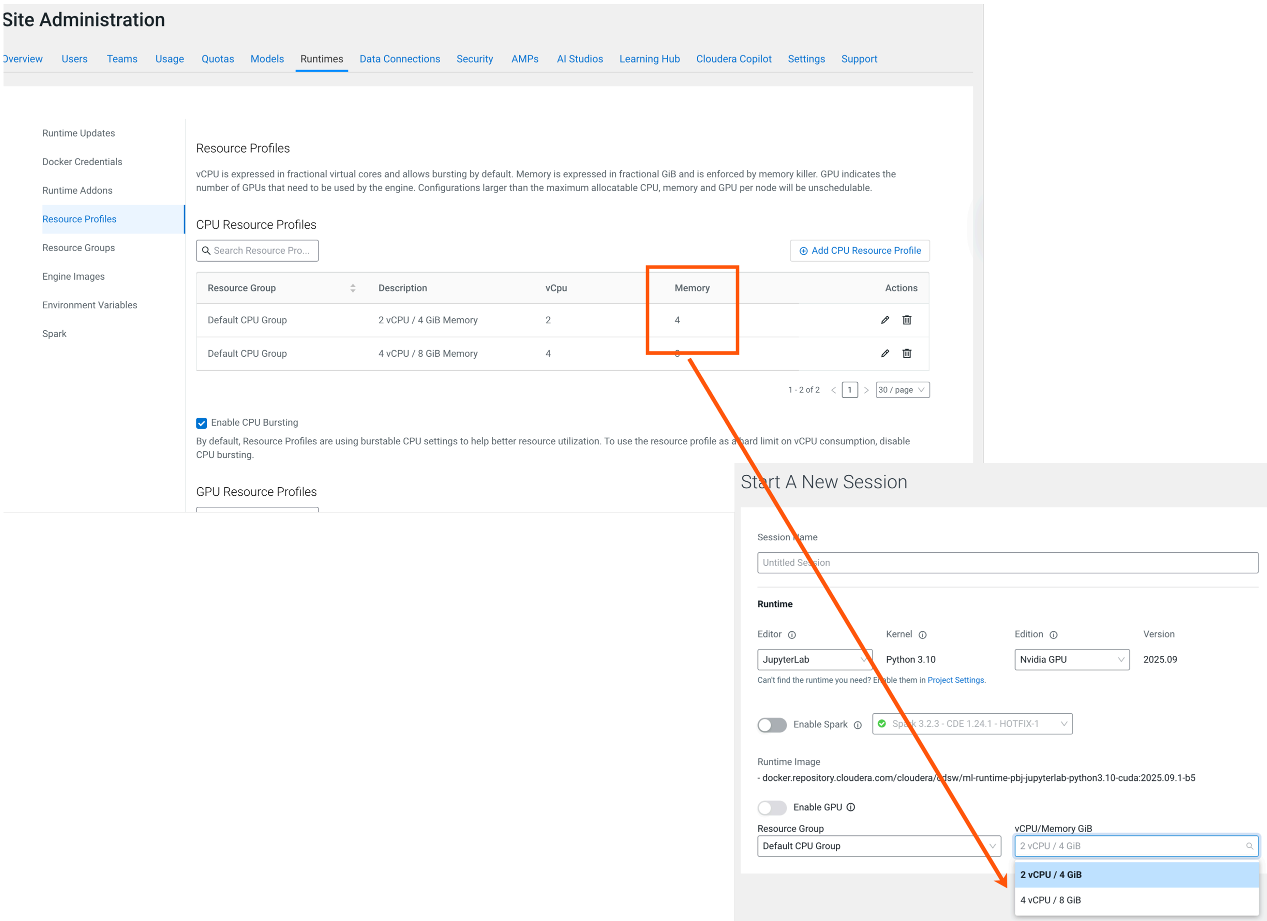
   Cloudera recommends that all profiles include at least 2 GB of RAM to avoid out of memory errors for common user operations.

   You will see the option to add GPUs to the resource profiles only if your Cloudera AI hosts are equipped with GPUs, and you have enabled them for use by setting the relevant properties in cdsw.conf.

Results

If there are two worker nodes and 10 vCPU available overall, if one user tries to establish a session with 8 vCPU, CDSW will not allow it. The memory and CPU must be contiguous (adjacent to each other). When a user spins a session, the pod triggers on a single node and resources on the same node are utilized. This is expected behavior for Kubernetes.

Figure 3: Resource profiles available when launching a session



Site Administration

Resource Profiles

vCPU is expressed in fractional virtual cores and allows bursting by default. Memory is expressed in fractional GiB and is enforced by memory killer. GPU indicates the number of GPUs that need to be used by the engine. Configurations larger than the maximum allocatable CPU, memory and GPU per node will be unschedulable.

CPU Resource Profiles

Resource Group	Description	vCpu	Memory	Actions
Default CPU Group	2 vCPU / 4 GiB Memory	2	4	
Default CPU Group	4 vCPU / 8 GiB Memory	4	8	

☑ Enable CPU Bursting

By default, Resource Profiles are using burstable CPU settings to help better resource utilization. To use the resource profile as a hard limit on vCPU consumption, disable CPU bursting.

GPU Resource Profiles

Start A New Session

Session Name: Untitled Session

Runtime

Editor	Kernel	Edition	Version
JupyterLab	Python 3.10	Nvidia GPU	2025.09

Can't find the runtime you need? Enable them in Project Settings.

Enable Spark: Spark 3.2.3 - CDE 1.24.1 - HOTFIX-1

Runtime Image
- docker.repository.cloudera.com/cloudera/cdsw/ml-runtime-pbj-jupyterlab-python3.10-cuda:2025.09.1-b5

Enable GPU

Resource Group: Default CPU Group

vCPU/Memory GiB: 2 vCPU / 4 GiB

Disable or deprecate Runtime addons

Disable or deprecate a Spark Runtime addon.

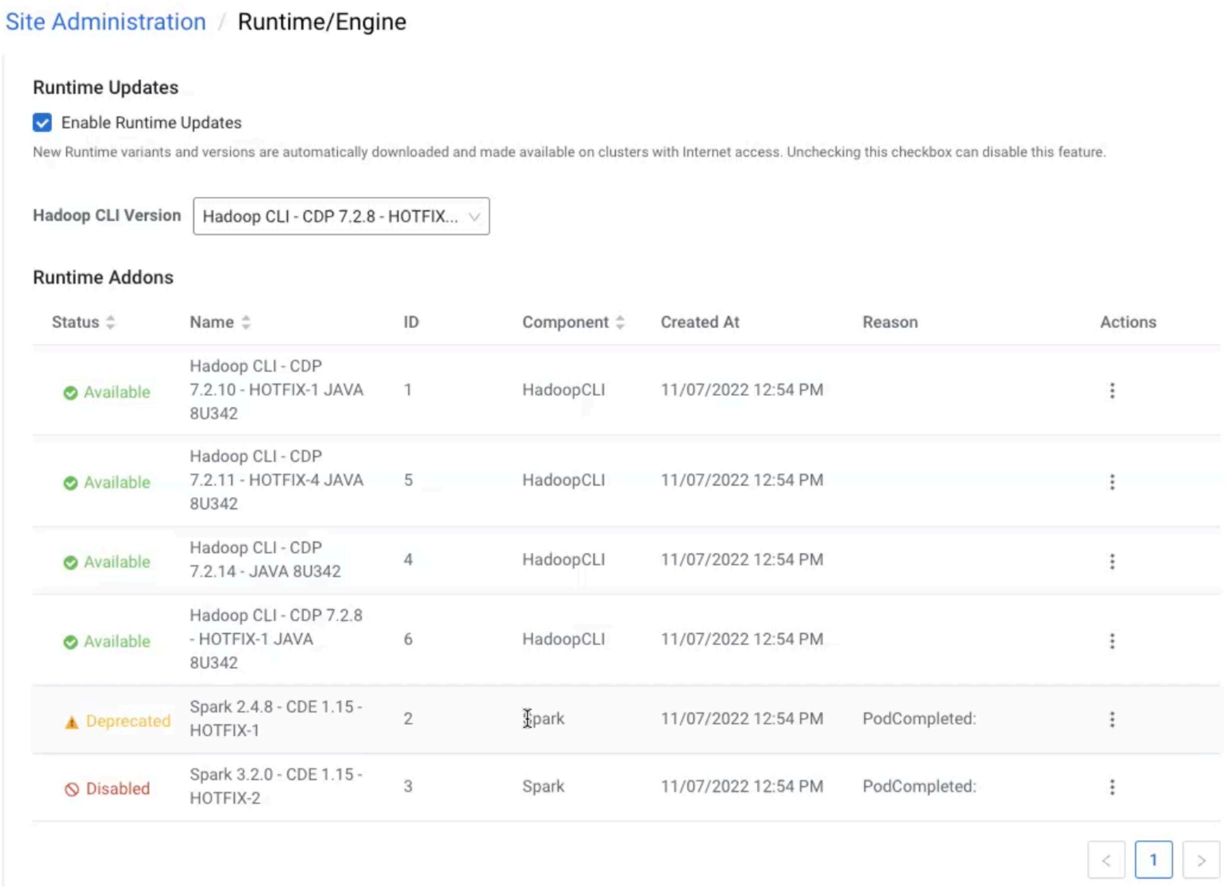
About this task

You can disable or deprecate any Spark Runtime addon from the Runtime/Engine tab of Site Administration.

Procedure

1. Select Site Administration in the left Navigation bar.
2. Select the Runtime/Engine tab.

3. Select Disabled or Deprecated from Actions next to any *SPARK* addon.



Site Administration / Runtime/Engine

Runtime Updates

☑ Enable Runtime Updates

New Runtime variants and versions are automatically downloaded and made available on clusters with Internet access. Unchecking this checkbox can disable this feature.

Hadoop CLI Version: Hadoop CLI - CDP 7.2.8 - HOTFIX...

Runtime Addons

Status	Name	ID	Component	Created At	Reason	Actions
✓ Available	Hadoop CLI - CDP 7.2.10 - HOTFIX-1 JAVA 8U342	1	HadoopCLI	11/07/2022 12:54 PM		⋮
✓ Available	Hadoop CLI - CDP 7.2.11 - HOTFIX-4 JAVA 8U342	5	HadoopCLI	11/07/2022 12:54 PM		⋮
✓ Available	Hadoop CLI - CDP 7.2.14 - JAVA 8U342	4	HadoopCLI	11/07/2022 12:54 PM		⋮
✓ Available	Hadoop CLI - CDP 7.2.8 - HOTFIX-1 JAVA 8U342	6	HadoopCLI	11/07/2022 12:54 PM		⋮
▲ Deprecated	Spark 2.4.8 - CDE 1.15 - HOTFIX-1	2	Spark	11/07/2022 12:54 PM	PodCompleted:	⋮
⦸ Disabled	Spark 3.2.0 - CDE 1.15 - HOTFIX-2	3	Spark	11/07/2022 12:54 PM	PodCompleted:	⋮

Note: You can also return the status to Available using Actions.

Onboarding Business Users

There are two procedures required for adding Business Users to Cloudera AI. First, an Administrator ensures the Business User has the correct permissions, and second, a Project Owner adds the Business User as a Collaborator.

Before you begin

Make sure the user is already assigned in your external identity provider, such as LDAP.

About this task

The Admin user performs these steps:

Procedure

1. Navigate to Workbench Manage access .
2. Select the applicable user or group.
3. Select the MLWorkspaceBusinessUsers role for the selected user or group.
4. Navigate to Site Administration Users in the Cloudera AI Workbench and click Synchronize Users.

What to do next

Add the ML Business User as a Collaborator to a Project.

Related Information

Adding a collaborator

Adding a collaborator

Project owners can add collaborators to a project.

About this task

Complete the following steps as a Project owner:

Procedure

1. In the Cloudera console, click the Cloudera AI tile.

   The Home page displays.
2. Select the required Workbench.

   The Cloudera AI Workbench page displays.
3. Click Projects in the left navigation pane and select the required project.
4. Go to Collaborators, and enter the user ID in the Search box.
5. Choose the User ID, and click Add. The user or team is added with their role displayed.

Results

When the Business User logs in, they can access the Applications within this project.

User roles

Users in Cloudera AI are assigned one or more of the following roles.

There are two categories of roles: environment resource roles, which apply to a given Cloudera environment, and workbench resource roles, which apply to a single workbench. To use workbench resource roles, you may need to upgrade the workbench or create a new workbench.

If a user has more than one role, then the role with the highest level of permissions takes precedence. If a user is a member of a group, it may gain additional roles through that membership.

Environment resource roles

- MLAdmin: It grants a Cloudera user the ability to create and delete Cloudera AI Workbenches within a given Cloudera environment. MLAdmins also have Administrator level access to all the workbenches provisioned within this environment. They can run workloads, monitor, and manage all user activity on these workbenches. This user also needs the account-level role of IAMViewer, in order to access the environment Manage Access page. To create or delete workbenches, this user also needs the EnvironmentAdmin role.
- MLUser: It grants a Cloudera user the ability to view Cloudera AI Workbenches provisioned within a given Cloudera environment. MLUsers are also able to run workloads on all the workbenches provisioned within this environment.

Workbench resource roles

Workbench roles are for users who are granted access to only a single workbench.

- MLWorkspaceAdmin: It grants permission to manage all Cloudera AI workloads and settings inside a specific workbench. To perform resource role assignment, the IAMViewer role is also needed. A user with this role can administer the workbench, but is not able to utilize Cloudera APIs that modify a workbench (for example, creating or upgrading workbenches).
- MLWorkspaceBusinessUser: It grants permission to view shared Cloudera AI applications inside a specific workbench.
- MLWorkspaceUser: It grants permission to run Cloudera AI workloads inside a specific workbench.

Using the workbench resource roles

A power user or account administrator must assign the first MLWorkspaceAdmin to a workbench. Subsequently, if the MLWorkspaceAdmin also has the IAMViewer role, they can assign resource roles to the workbench.

An MLAdmin (an environment resource role) is not automatically able to assign workbench resource roles on the Manage access page. A role such as MLWorkspaceAdmin is needed to do this.

You can check the permissions for a given resource role by clicking the Information icon by each resource role shown in User Management, on the Resources tab for a user, or in a Cloudera user profile.

Note: Any user that lists users or assigns resource roles also needs the account-level role of IAMViewer.

Business Users and Cloudera AI

A user is considered a Business User in Cloudera AI if they are assigned the MLWorkspaceBusinessUser role on the workbench resource role assignment. Inside the workbench, a Business User is able to access and view applications, but does not have privileges to access any other workloads in the workbench.

Logging in as a Business User

When you log in as a Business User, the only page you see is the Applications page. The page shows any applications associated with any projects that you have been added to as a Collaborator, even though you do not have rights to access the other assets associated with those projects.

In order for applications to appear in your view, contact the Project Owner to add you as a Collaborator to the project. If you have not been added to any projects, or none of the projects that you have been added to have applications, the Applications page displays the message, You currently don’t have any applications.

Managing your Personal Account

You can edit personal account settings such as email, SSH keys and Hadoop credentials.

About this task

You can also access your personal account settings by clicking Account settings in the upper right-hand corner drop-down menu. This option will always take you to your personal settings page, irrespective of the context you are currently in.

Procedure

1. Sign in to Cloudera AI.
2. From the upper right drop-down menu, switch context to your personal account.

3. Click Settings.

   Profile

   You can modify your name, email, and bio on this page.

   Teams

   This page lists the teams you are a part of and the role assigned to you for each team.

   SSH Keys

   Your public SSH key resides here. SSH keys provide a useful way to access to external resources such as databases or remote Git repositories. For instructions, see *SSH Keys*.

Related Information

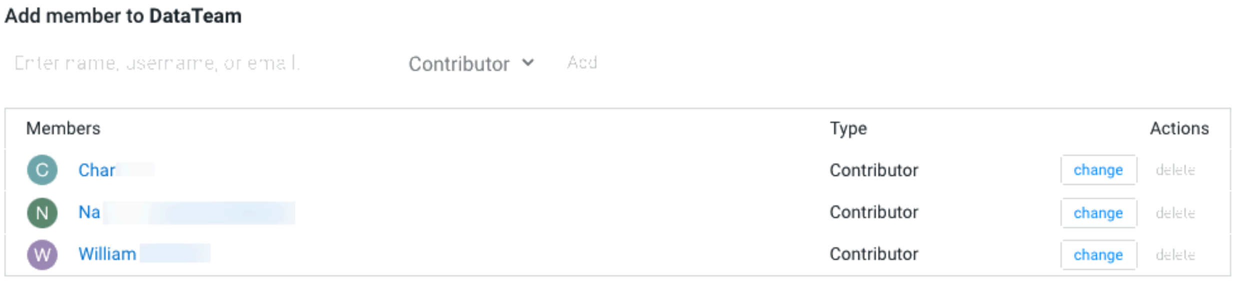
SSH Keys

Creating a Team

Users who work together on more than one project and want to facilitate collaboration can create a Team. Teams enable you to efficiently manage the users assigned to projects.

About this task

Team projects are owned by the team, rather than an individual user. Team administrators, contributors, or operators can add or remove members at any time, assigning each member different permissions. A team cannot be deleted and at least one member must be there in the team.



Add member to DataTeam

Enter name, username, or email. Contributor Add

Members	Type	Actions
C Char [redacted]	Contributor	change delete
N Na [redacted]	Contributor	change delete
W William [redacted]	Contributor	change delete

Procedure

1. In Site Administration Teams , select New Team.
2. Enter the name of the team.
3. Select Local or Synced Team.

   Local Teams are created and managed directly within the Cloudera AI Workbench and are not visible in the Cloudera Management Console. In contrast, Synced Teams are mapped to Cloudera groups defined in the **User Management** section of the Cloudera Management Console. Only groups explicitly assigned to the environment or to the Cloudera AI Workbench are displayed in the Synced Teams list.
4. If Synced Team is selected, choose a group name and role under Add Groups and click Add. You can add multiple groups and roles using the Add option.
5. Enter a Description, if needed.

6. Add or invite team members. Team members can have one of the following privilege levels:
   - Viewer - The Viewer has read-only access to team projects. The Viewere cannot create new projects within the team but can be added to existing ones.
   - Operator - The Operator has read-only access to team projects. Additionally, Operators can start and stop existing jobs in the projects that they have access to.
   - Contributor - The Contributor has write-level access to all team projects to all team projects with Team or Public visibility. The Contributor can create new projects within the team. They can also be added to existing team projects.
   - Admin - The Administrator has complete access to all team projects, can add new team members, and modify team account information. The creator of the team is assigned the Administrator privilege, and can also assign other team members the Administrator privilege. Each team must have at least one Administrator user.
7. Select Create Team.
8. Select Sync Teams to update the teams with information in the Cloudera Management Console.

Related Information

Synchronizing group membership

Managing a Team Account

Team administrators can modify account information, add or invite new team members, and view/edit privileges of existing members.

Procedure

1. From the upper right drop-down menu, switch context to the team account.
2. Click Settings to open up the Account Settings dashboard.
3. Modify any of the following settings:

   Profile

   Modify the team description on this page.

   Members

   You can add new team members on this page, and modify privilege levels for existing members.

   SSH Keys

   The team's public SSH key resides here. Team SSH keys provide a useful way to give an entire team access to external resources such as databases. For instructions, see *SSH Keys*. Generally, team SSH keys should not be used to authenticate against Git repositories. Use your personal key instead.

Related Information

SSH Keys

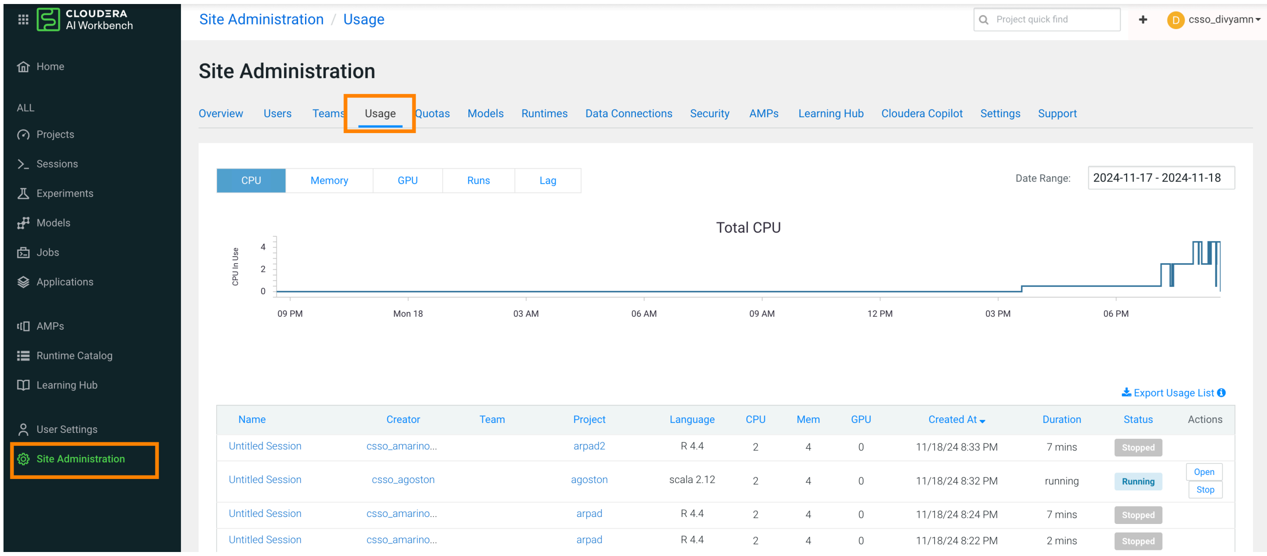
Monitoring Cloudera AI activity

This topic describes how to monitor user activity on an Cloudera AI Workbench.

Required Role: Site Administrator

The Admin Overview tab displays basic information about your deployment, such as the number of users signed up, the number of teams and projects created, memory used, and some average job scheduling and run times. You can also see the version of Cloudera AI you are currently running.

The Admin Activity tab of the dashboard displays the following time series charts. These graphs should help site administrators identify basic usage patterns, understand how cluster resources are being utilized over time, and how they are being distributed among teams and users.



Name	Creator	Team	Project	Language	CPU	Mem	GPU	Created At	Duration	Status	Actions
Untitled Session	csso_amarino...		arpad2	R 4.4	2	4	0	11/18/24 8:33 PM	7 mins	Stopped	
Untitled Session	csso_agoston		agoston	scala 2.12	2	4	0	11/18/24 8:32 PM	running	Running	Open Stop
Untitled Session	csso_amarino...		arpad	R 4.4	2	4	0	11/18/24 8:24 PM	7 mins	Stopped	
Untitled Session	csso_amarino...		arpad	R 4.4	2	4	0	11/18/24 8:22 PM	2 mins	Stopped	

Important: The graphs and numbers on the Admin Activity page do not account for any resources used by active models on the deployment. For that information, go to Admin Models page.

- **CPU** - This is the total number of CPUs requested by sessions running at this time.

  Note that code running inside an n-CPU session, job, experiment or model replica can access at least n-CPUs worth of CPU time. Each user pod can utilize all of its host's CPU resources except the amount requested by other user workloads or Cloudera AI application components. For example, a 1-core Python session can use more than 1 core if other cores have not been requested by other user workloads or Cloudera AI application components.
- **Memory** - This is the total memory (in GiB) requested by sessions running at this time.
- **GPU** - This is the total number of GPUs requested by sessions running at this time.
- **Runs** - This is the total number of sessions and jobs running at this time.
- **Lag** - It depicts session scheduling and startup times.

  - **Scheduling Duration:** The amount of time it took for a session pod to be scheduled on the cluster.
  - **Starting Duration:** The amount of time it took for a session to be ready for user input. This is the amount of time since a pod was scheduled on the cluster until code could be executed.

The Export Sessions List provides a CSV export file of the columns listed in the table. It is important to note that the exported duration column is in seconds for a more detailed output.

Tracked user events

The tables on this page describe the user events that are logged by Cloudera AI.

Table 1: Database Columns

When you query the user_events table, the following information can be returned:

Information	Description
id	The ID assigned to the event.
user_id	The UUID of the user who triggered the event.
ipaddr	The IP address of the user or component that triggered the event. 127.0.0.1 indicates an internal component.

Information	Description
user agent	The user agent for this action, such as the web browser. For example: `Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36`
event_name	The event that was logged. The tables on this page list possible events.
description	This field contains the model name and ID, the user type (NORMAL or ADMIN), and the username.
created_at	The date (YYYY-MM-DD format) and time (24-hour clock) the event occurred .

Table 2: Events Related to Engines

Event	Description
engine environment vars updated	-
engine mount created	-
engine mount deleted	-
engine mount updated	-
engine profile created	-
engine profile deleted	-
engine profile updated	-

Table 3: Events Related to Experiments

Event	Description
experiment run created	-
experiment run repeated	-
experiment run cancelled	-

Table 4: Events Related to Files

Event	Description
file downloaded	-
file updated	-
file deleted	-
file copied	-
file renamed	-
file linked	The logged event indicates when a symlink is created for a file or directory.
directory uploaded	-

Table 5: Events Related to Models

Event	Description
model created	-
model deleted	-

Table 6: Events Related to Jobs

Event	Description
job created	-
job started	-
stopped all runs for job	-
job shared with user	-
job unshared with user	-
job sharing updated	The logged event indicates when the sharing status for a job is changed from one of the following options to another: • All anonymous users with the link • All authenticated users with the link • Specific users and teams

Table 7: Events Related to Licenses

Event	Description
license created	-
license deleted	-

Table 8: Events Related to Projects

Event	Description
project created	-
project updated	-
project deleted	-
collaborator added	-
collaborator removed	-
collaborator invited	-

Table 9: Events Related to Sessions

Event	Description
session launched	-
session terminated	-
session stopped	-
session shared with user	-
session unshared with user	-
update session sharing status	The logged event indicates when the sharing status for a session is changed from one of the following options to another: • All anonymous users with the link • All authenticated users with the link • Specific users and teams

Table 10: Events Related to Admin Settings

site config updated	The logged event indicates when a setting on the Admin Settings page is changed.

Table 11: Events Related to Teams

Event	Description
add member to team	-
delete team member	-
update team member	-

Table 12: Events Related to Users

Event	Description
forgot password	-
password reset	-
update user	If the logged event shows that a user is banned, that means that the user account has been deactivated and does not count toward the license.
user signup	-
user login	The logged event includes the authorization method, LDAP/SAML or local.
user logout	-
ldap/saml user creation	The logged event indicates when a user is created with LDAP or SAML.

Monitoring user events

This topic shows you how to query the PostgresSQL database that is embedded within the Cloudera AI deployment to monitor or audit user events.

About this task

Querying the PostgresSQL database that is embedded within the Cloudera AI deployment requires root access to the Cloudera AI Master host.

Procedure

1. SSH to the Cloudera AI Master host and log in as root.

   For example, the following command connects to cdsw-master-host. as root:

   ```
   ssh root@cdsw-master-host.yourcdswdomain.com
   ```

2. Get the name of the database pod:

   ```
   kubectl get pods -l role=db
   ```

   The command returns information similar to the following example:

   ```
   NAME                    READY   STATUS    RESTARTS   AGE
   db-86bbb69b54-d5q88     1/1     Running   0          4h46m
   ```

3. Enter the following command to log into the database as the sense user:

   ```
    kubectl exec <database pod> -ti -- psql -U sense
   ```

   For example, the following command logs in to the database on pod db-86bbb69b54-d5q88:

   ```
   kubectl exec db-86bbb69b54-d5q88 -ti -- psql -U sense
   ```

   You are logged into the database as the sense user.
4. Run queries against the user_events table.

   For example, run the following query to view the most recent user event:

   ```
   select * from user_events order by created_at DESC LIMIT 1
   ```

   The command returns information similar to the following:

   ```

   id           | 3658
   user_id      | 273
   ipaddr       | ::ffff:127.0.0.1
   user_agent   | node-superagent/2.3.0
   event_name   | model created
   description | {"model":"Simple Model 1559154287-ex5yn","modelId":"50","
   userType":"NORMAL","username":"DonaldBatz"}
   created_at   | 2019-05-29 18:24:47.65449
   ```
5. Optionally, you can export the user events to a CSV file for further analysis:

   a) Copy the user_events table to a CSV file:

   ```
   copy user_events to '/tmp/user_events.csv' DELIMITER ',' CSV HEADER;
   ```

   b) Find the container that the database runs on:

   ```
   docker ps | grep db-86bbb
   ```

   The command returns output similar to the following:

   ```
   c56d04bbd58 c230b2f564da "docker-entrypoint..." 7 days ago Up 7 days k8s
   _db_db-86bbb69b54-fcfm6_default_8b2dd23d-88b9-11e9-bc34-0245eb679f96_0
   ```

   The first entry is the container ID.

   c) Copy the user_events.csv file out of the container into a temporary directory on the Master host:

   ```
   docker cp <container ID>:/tmp/user_events.csv /tmp/user_events.csv
   ```

   For example:

   ```
   docker cp 8c56d04bbd58:/tmp/user_events.csv /tmp/user_events.csv
   ```

   d) Use SCP to copy /tmp/user_events.csv from the Cloudera AI Master host to a destination of your choice.

   For example, run the following command on your local machine to copy user_events.csv to a local directory named events:

   ```
   scp root@cdsw-master-host.yourcdswdomain.com:/tmp/user_events.csv /local
   /directory/events/
   ```

What to do next

For information about the different user events, see *Tracked User Events*.

Related Information

Tracked user events

Collecting project size information

Configure your project size information to be available in your Cloudera AI Workbench.

About this task

This configuration enables a scheduled job to update the Cloudera AI Workbench database project table daily with the size of the project.

Procedure

1. Configure Cloudera AI to collect the project sizes for each project daily, go to: Cloudera AI UI Site Administration Settings Collect Project Sizes .
2. View the project sizes by running the following command:

```
# export KUBECONFIG=<kubeconfig_file>
# kubectl exec -it $(kubectl get pods  -n [***CLOUDERA AI NAMESPACE***] -
l app=db -o jsonpath='{.items[*].metadata.name}') -c db -n [***CLOUDERA AI
 NAMESPACE***] -- psql -P pager=off -U sense -c " select p.id, p.name,u.us
ername, u.email,u.type,u.name,p.size from projects p, users u where p.us
er_id=u.id"
```

Results

Enabling this will update Cloudera AI Workbench with the project size information on the last accessed date of the project.

Monitoring active Models across the Workbench

This topic describes how to monitor all active models currently deployed on your workbench.

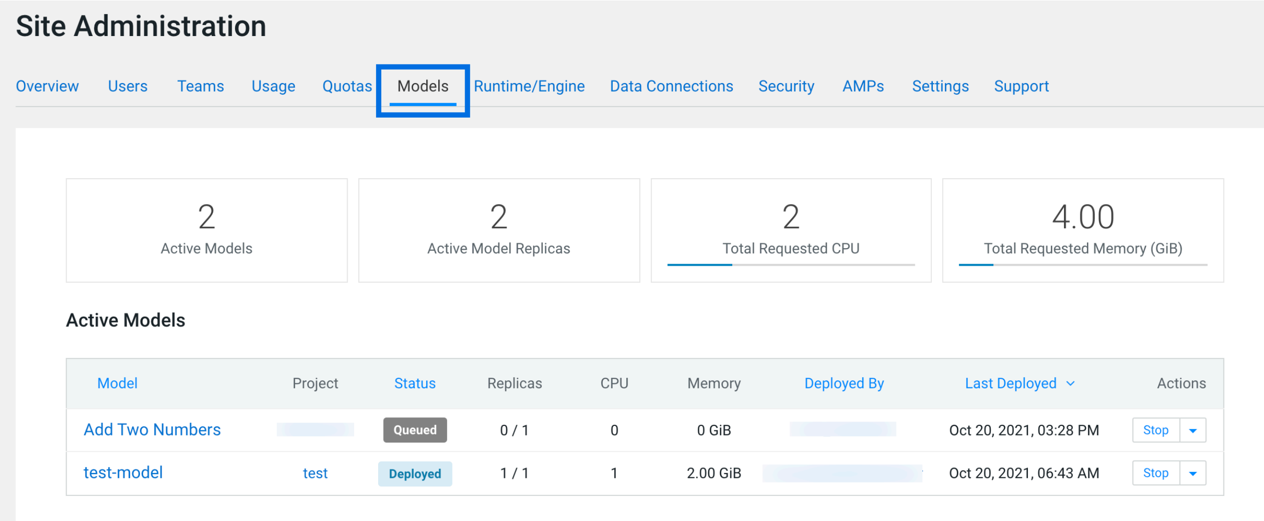
What is an active Model?

A model that is in the Deploying, Deployed, or Stopping stages is referred to as an active model.

Monitoring all active Models across the Workbench

Required Role: Site Administrator

To see a complete list of all the models that have been deployed on a deployment, and review resource usage across the deployment by models alone, go to Admin Models . On this page, site administrators can also Stop/Restart/Rebuild any of the currently deployed models.



2 Active Models

2 Active Model Replicas

2 Total Requested CPU

4.00 Total Requested Memory (GiB)

Active Models

Model	Project	Status	Replicas	CPU	Memory	Deployed By	Last Deployed	Actions
Add Two Numbers	[redacted]	Queued	0 / 1	0	0 GiB	[redacted]	Oct 20, 2021, 03:28 PM	Stop
test-model	test	Deployed	1 / 1	1	2.00 GiB	[redacted]	Oct 20, 2021, 06:43 AM	Stop

Monitoring and alerts

Cloudera AI leverages Cloudera Monitoring based on Prometheus and Grafana to provide dashboards that allow you to monitor how CPU, memory, storage, and other resources are being consumed by your Cloudera AI Workbenches.

Prometheus is an internal data source that is auto-populated with resource consumption data for each deployment. Grafana is the monitoring dashboard that allows you to create visualizations for resource consumption data from Prometheus. By default, Cloudera AI provides three Grafana dashboards: K8 Cluster, K8s Containers, and K8s Node. You can extend these dashboards or create more panels for other metrics. For more information, see the Grafana documentation.

Related Information

Grafana documentation

Application polling endpoint

The Cloudera AI server periodically polls applications for their status. The default polling endpoint is the root endpoint (/), but a custom polling endpoint can be specified if the server or other application has difficulty with the default endpoint.

When creating or modifying an application, you can specify a new value for the CDSW_APP_POLLING_END POINT environmental variable. Just replace the default value / that is shown. For more information, see *Analytical Applications*.

You can also set the environmental value in Project Settings Advanced . In this case, any setting made here can be overridden by settings in a given application. However, settings made in Project Settings Advanced also apply when polling sessions.

Related Information

Analytical Applications

Choosing default engine

This topic describes how to choose a default engine for creating projects.

Before you begin

Required Role: MLAdmin

Note: On on premises, the corresponding role is EnvironmentAdmin.

Make sure you are assigned the MLAdmin role in Cloudera. Only users with the MLAdmin role will be logged into Cloudera AI Workbenches with Site Administrator privileges.

There are two types of default engines: ML Runtimes and Legacy Engines. However, legacy engines are deprecated in the current release and project settings default to ML Runtime.

Legacy engines Engines contain the machinery necessary to run sessions using all four interpreter options that Cloudera AI currently supports (Python 2, Python 3, R and Scala) and other support utilities (C and Fortran compilers, LaTeX, etc.). ML Runtimes are thinner and more lightweight than legacy engines. Rather than supporting multiple programming languages in a single engine, each Runtime variant supports a single interpreter version and a subset of utilities and libraries to run the user’s code in Sessions, Jobs, Experiments, Models, or Applications.

Procedure

1. Log in to the web interface.
2. Click Cloudera AI Workbenches , then open the workbench for which you want to set Default Engine.
3. Click Admin Runtime/Engine .
4. Choose the Default Engine you would like to use as the default for all newly created projects in this workbench.

   Note: Legacy Engines are deprecated in this release and Cloudera recommends using Runtime.
5. Modify the remaining information on the page:
   - Resource Profiles listed in the table are selectable resource options for both legacy Engines and ML Runtime (for example, when starting a Session or Job)
   - The remaining information on the page applies to site-level settings specific for legacy Engines.

Controlling User access to features

Cloudera AI provides Site Administrators with the ability to restrict or hide specific functionality that non-Site Administrator users have access to in the UI. For example, a site administrator can hide the models and experiments features from the Cloudera AI Workbench UI.

The settings on this page can be configured through the Security and Settings tabs on the Administration page.

Table 13: Security Tab

Property	Description
Allow remote editing	Disable this property to prevent users from connecting to the Cloudera AI deployment with cdswctl and using local IDEs, such as PyCharm.

Property	Description
Allow only session creators to run commands on active sessions	By default, a user's permission to active sessions in a project is the same as the user's permission to that project, which is determined by the combination of the user's permission as a project collaborator, the user's permission in the team if this is a team project, and whether the user is a Site Administrator. By checking this checkbox, only the user that created the active session will be able to run commands in that session. No other users, regardless of their permissions in the team or as project collaborators, will be able to run commands on active sessions that are not created by them. Even Site Administrators will not be able to run commands in other users' active sessions.
Allow console output sharing	Disable this property to remove the Share button from the project workbench and workbench UI as well as disable access to all shared console outputs across the deployment. Note that re-enabling this property does not automatically grant access to previously shared consoles. You will need to manually share each console again
Allow anonymous access to shared console outputs	Disable this property to require users to be logged in to access shared console outputs.
Allow file upload/download through UI	Use this checkbox to show/hide file upload/download UI in the project workbench. When disabled, Cloudera AI API will forbid request of downloading file(s) as attachment. Note that the backend API to upload/edit/read the project files are intact regardless of this change in order to support basic Cloudera AI functionality such as file edit/read.

Table 14: Settings Tab

Property	Description
Require invitation to sign up	Enable this property to send email invitations to users when you add them to a group. To send email, an SMTP server must first be configured in Settings Email .
Allow users to create public projects	Disable this property to restrict users from creating new public projects. Site Administrators will have to create any new public projects.
Allow Legacy Engine users to use the Python 2 kernel	Enable this property to allow Legacy Engine users to select the Python 2 kernel when creating a job. Python 2 is disabled by default.
Allow users to create projects	Disable this property to restrict users from creating new projects. Site Administrators will have to create any new projects.
Allow users to create teams	Disable this property to restrict users from creating new teams. Site Administrators will have to create any new teams.
Allow users to run experiments	Disable this property to hide the Experiments feature in the UI. Note that this property does not affect any active experiments. It will also not stop any experiments that have already been queued for execution.
Allow users to create models	Disable this property to hide the Models feature in the UI. Note that this property does not affect any active models. In particular, if you do not stop active models before hiding the Models feature, they continue to serve requests and consume computing resources in the background.
Allow users to create applications	Disable this property to hide the Applications feature in the UI. Note that this property does not affect any active applications. In particular, if you do not stop active applications before hiding the feature, they continue to serve requests and consume computing resources in the background.

Setting custom Spark configurations at workbench-level

Administrators can configure custom Spark settings at the Cloudera AI Workbench level. These configurations will then be applied to all projects and newly launched Spark sessions within that workbench. Non-administrator users can view the applied configurations from Cloudera AI 1.5.5 SP1, but cannot modify them at this level.

About this task

Understanding Spark configuration layers and precedence

Spark configuration layers applied to workbenches have the following hierarchy and precedence:

1. Project-level configurations that are set in the spark-defaults.conf file: These are configurations set within specific project files and have the highest precedence, overriding any workbench-level defaults. For details on setting project-level defaults, see Spark configuration files.
2. Custom workbench level: These are the custom settings you can configure in this topic, applied by administrators at the workbench level. The configurations in the workbench defaults are applied unless overridden in the custom workbench level.
3. Workbench defaults level: These are the default Spark configurations applied by the Cloudera AI Workbench system to all Spark sessions. Users can view these defaults, which are displayed in an uneditable textbox.
4. Base cluster level: If Spark pushdown is enabled the Cloudera Base on premises cluster-level Spark configuration details are visible in an uneditable textbox.

Note: This feature is available from Cloudera AI 1.5.5 SP1 or higher versions.

Procedure

1. In the Cloudera console, click the Cloudera AI tile.

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

   The workbench **Home** page displays.
3. Click Site Administration in the left navigation pane.
4. Select Runtimes tab.
5. On the Runtimes page, scroll down to find the Spark Configuration section.

   If Spark Pushdown is not enabled, you can see two main text areas under the Spark on Kubernetes tab:

   - The Cloudera AI Workbench Defaults area displays the default Spark configurations applied by Cloudera AI. This section is not editable.
   - The Custom workbench level for spark on Kubernetes workloads area is the editable textbox in which you can enter your custom Spark properties.

   Note:

   This feature is available from Cloudera AI 1.5.5 SP1 or higher versions.

   If Spark Pushdown is enabled, select the Spark Pushdown to base cluster tab, where you can see three main text areas:

   - The Base Cluster uneditable textbox, which displays Spark configuration details if Spark pushdown is enabled.
   - The Cloudera AI Workbench Defaults area displays the default Spark configurations applied by Cloudera AI. This section is not editable.
   - The Custom workbench level for spark Pushdown workloads area is the editable textbox in which you can enter your custom Spark properties.

6. Enter the desired custom Spark properties in the Custom workbench level for spark on Kubernetes workloads textbox. Each property must be on a new line, typically in the `key=value` format, similar to a spark-defaults.conf file.
7. Click Save Spark Configuration located at the bottom right of the section.

Enabling Spark pushdown configuration

Administrators must enable the Spark pushdown configuration option to grant users access to Spark pushdown within project configurations.

Procedure

1. Enable users to configure projects to use Spark pushdown so that setting up portforwards and configurations for running spark workloads on yarn is an available option.
   a) In the Cloudera console, click the Cloudera AI tile.

      The Cloudera AI Workbenches page displays.
   b) Select Site Administration in the left navigation pane.

      The Site Administration page displays.
   c) Select the Settings tab.
   d) On the Settings page scroll down to the Feature Flags section.
   e) Select the checkbox for the Allow users to enable Spark Pushdown Configuration for Projects option.

      When this option is enabled, users can configure projects to use Spark pushdown that enables setting up portforwards and configurations for running spark workloads on yarn.

      Spark pushdown is enabled.
2. Enable portforwarding rules and default spark configurations to allow spark job executors to be scheduled in Yarn in the base cluster.

   To enable Spark Pushdown for project users, you must have a Contributor, Operator, or Administrator role in the project.

   This is a project-specific setting to enable Spark pushdown for all newly launched workloads in the project. Each project that intends to use the Cloudera Base on premises cluster Yarn for Spark workloads must enable this setting.

   a) In the Cloudera console, click the Cloudera AI tile.

      The Cloudera AI Workbenches page displays.
   b) Select the required workbench.
   c) Click Projects from the left navigation tree.
   d) From the list of projects, select the one to modify.
   e) Select Project Settings to open the Project Settings dashboard.
   f) Select the **General** tab.
   g) Select the Enable Spark Pushdown option

Enabling AI Studios

Administrators must enable the AI Studios option to allow users to build and deploy AI-powered applications.

Procedure

1. In the Cloudera console, click the Cloudera AI tile.

   The Cloudera AI Workbenches page displays.
2. Select the required workbench.

   The Home page of the workbench is displayed.
3. Select Site Administration in the left navigation pane.

   The Site Administration page displays.
4. Select the Settings tab.
5. On the Settings page scroll down to the Feature Flags section.
6. Select the checkbox for the Enable AI Studios (Technical Preview) option.
   The AI Studios feature is enabled.

Configuring Job Retry settings

The Job Retry feature is available from Cloudera AI 1.5.5 SP1 or higher releases. Job retry runs are designed to operate asynchronously, ensuring they do not disrupt the normal flow of a job run. These retries are executed concurrently to maintain efficiency.

About this task

The Administrator can define default values for the Job Retry parameters and only the Administrator can configure a hard limit on the maximum number of job retry runs that can be executed alongside normal job runs. This setting must only be enabled if you want to manage and limit resource usage for job retry runs.

Procedure

1. In the Cloudera console, click the Cloudera AI tile.

   The Cloudera AI Workbenches page displays.
2. Click on the name of the workbench.

   The workbench Home page displays.
3. Select Site Administration in the left Navigation pane.
4. Select the Settings tab.
5. Select Job Retry Configuration Limit Concurrent Retries .
6. Enable Limit Concurrent Retries by selecting the checkbox.

   Enabling this option sets a limit to how many job retry runs (at maximum) can be active at the same time.
7. Define the limit value for Maximum Concurrent Retry Limit.

   The Maximum Concurrent Retry Limit specifies the maximum number of job retry runs that can execute concurrently across the entire workbench, regardless of the total number of jobs running.

   If the maximum limit value defined as Maximum Concurrent Retry Limit is reached, any additional job retry runs are rescheduled until the number of active retry runs falls below the limit.

   Enable this hard limit only if job retry runs are consuming excessive resources, otherwise, avoid setting a hard limit.

   Administrators can set this value if the Limit Concurrent Retires option is enabled.

8. Under Default Settings for all jobs, select Enable Retry to enable a retry run for the job.

   Define the following parameters for Job Retry:

   - Maximum Retry – The maximum number of retry attempts which can be triggered for a single job run in case of continuous failure of retry job runs.

     The minimum value is 1.
   - Retry Delay (minutes) – The delay between two consecutive retry job runs for a failed instance of the run.

     The minimum value is 1 minute.
   - Retry Conditions – Different options can be configured to control the terminal states of a job run that trigger a retry. The Retry process completes as soon as at least one (or more) option is selected.

     Select at least one of the following criteria if Retry is enabled, but you can select any combination of the following Retry Conditions options:

     - Script Failure – Runs the Retry process for user script failures if the user script exits with a non-zero exit code after the execution of the script.
     - System Failure – Runs the Retry process for any kind of system- or engine-related failures not including user script failures.
     - Timed-out Runs – Runs the Retry process for timed-out job runs.
     - Skipped Runs – Runs the Retry process for skipped job runs.
9. Click on Update to save the settings.

Related Information

Creating a job

Job retry parameters

Cloudera AI email notifications

Cloudera AI allows you to send email notifications when you add collaborators to a project, share a project with a colleague, and for job status updates (email recipients are configured per-job). This topic shows you how to specify email address for such outbound communications.

Note that email notifications are not currently enabled by default. Emails are not sent when you create a new project. Email preferences cannot currently be configured at an individual user level.

Option 1: If your existing corporate SMTP server is accessible from the VPC where your Cloudera AI Workbench is running, you can continue to use that server. Go to the Admin Settings tab to specify an email address for outbound invitations and job notifications.

Option 2: If your existing SMTP solution cannot be used, consider using an email service provided by your cloud provider service. For example, Amazon provides Amazon Simple Email Service (Amazon SES).

Mail relay hosts often limit the authenticated sender reply address. Make sure to select a No reply email which you are allowed to use, otherwise email sending may fail.

Web session timeouts

You can set web sessions to time out and require the user to log in again. This time limit is not based on activity, it is the maximum time allowed for a web session.

You can set timeout limits for Users and Admin Users in Site AdministrationSecurity.

- User Web Browser Timeout (minutes) - This timeout sets the default maximum length of time that a web browser session can remain inactive. You remain logged in if you are actively using the session. If you are not active, then after a 5-minute warning, you are automatically logged out. Any changes to the setting take effect for any subsequent user logins.

- Admin User Web Browser Timeout (minutes) - This timeout sets the default maximum length of time that a web browser session for an Admin user can remain inactive. You remain logged in if you are actively using the session. If you are not active, then after a 5-minute warning, you are automatically logged out. Any changes to the setting take effect for any subsequent Admin user logins.

Project garbage collection

Marks orpaned files for deletion from a project and cleans up projects that are marked for deletion.

Procedure

1. Click Site Administration Settings.
2. Scroll to Project Garbage Collection.

   Click Garbage Collect Projects to permanently delete projects marked for deletion.

   Click Clean Up Orphaned Projects to mark orphaned projects for deletion.

Results

Orphaned project files are marked for deletion. All files marked for deletion are permanently deleted when you click Garbage Collect Projects.

How to make base cluster configuration changes

When you make base cluster configuration changes, you need to restart the base cluster to propagate those changes.

In general, as Administrator you perform the following steps:

1. Make the necessary configuration changes in the base cluster.
2. Restart the base cluster.
3. In the on premises compute cluster, perform the specific Kubernetes commands below to restart the ds-cdh-client pods for Cloudera AI.

For ECS:

1. Access Cloudera Manager.
2. Navigate to the Containerized Cluster Cloudera Embedded Container Service Web UI: Clusters Your embedded Cluster Cloudera Embedded Container Service Web UI Cloudera Embedded Container Service web UI
3. Select the namespace of your Cloudera AI Workbench on the top left dropdown.
4. Navigate to Workloads Deployments.
5. Locate ds-cdh-client in the list and perform Restart from the breadcrumbs on the right.

For OCP:

Access the openshift cluster with oc or kubectl, and scale the deployment of ds-cdh-client down and back up. Use the following commands.

1. oc scale deployment/ds-cdh-client --namespace <ml-namespace> --replicas 0
2. oc scale deployment/ds-cdh-client --namespace <ml-namespace> --replicas 1

Ephemeral storage

Ephemeral storage space is scratch space that a Cloudera AI session, job, application or model can use. This feature helps in better scheduling of Cloudera AI pods, and provides a safety valve to ensure runaway computations do not consume all available scratch space on the node.

By default, each user pod in Cloudera AI is allocated 0 GB of scratch space, and it is allowed to use up to 10 GB. These settings can be applied to an entire site, or on a per-project basis.

How Spark uses ephemeral storage in Cloudera AI

Spark drivers and executors write shuffle files, spilled RDD/DataFrame blocks, broadcast variables, and task logs under directories referenced by SPARK_LOCAL_DIRS.

On Kubernetes these paths are mounted as one emptyDir volume per pod; emptyDir is wiped as soon as the pod terminates, so the data is **ephemeral**.

If this volume fills up, the kubelet evicts the pod and Spark surfaces errors such as:

- java.io.IOException: No space left on device
- org.apache.spark.shuffle.MetadataFetchFailedException

This is followed by a Kubernetes event similar to Evicted: The node was low on resource:#ephemeral#storage

How does the CML UI map to Kubernetes resources

Table 15: Mapping to Kubernetes Resources

CML field	Pod spec element	What it does
Ephemeral Storage (GB) – Request	resources.requests.ephemeral-storage	Scheduler bin#packing & cluster#autoscaler logic
Ephemeral Storage (GB) – Max	resources.limits.ephemeral-storage	Hard ceiling; usage ##limit # pod eviction

Both the driver and every executor inherit the values you set here (or an override in Project#Settings###Advanced).

Sizing guidelines for common Spark workloads

Table 16: Sizing Guidelines

Work#load pattern	Rule of thumb (across all executors)	Rationale
SQL/ETL with light aggregations	##1#×#largest input size	Minimal shuffle spill
Joins, `groupByKey`, heavy shuffle	2#–#3#×#largest input size	Shuffle writes often exceed input volume
ML pipelines with .cache() / .persist()	Cached dataset size × #replicas	Cached blocks are duplicated

Quick workflow: Start with a generous limit, run once, open Spark UI##Executors##Shuffle Spill#(Disk), and set the per#pod limit to peak#spill ÷##executors

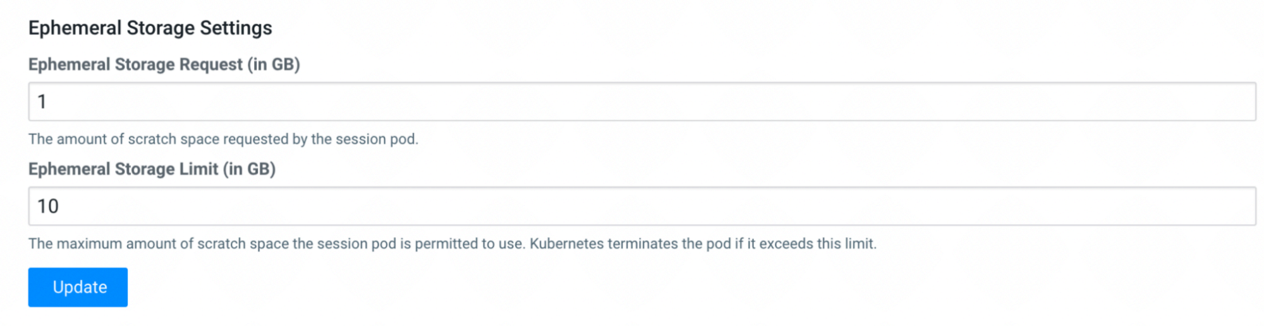
Tips to reduce Spark’s scratch#disk footprint

Table 17:

Goal	Knob	Notes
Fewer shuffle bytes	spark.sql.shuffle.partitions (closer to number of executors) and spark.sql.adaptive.enabled=true	Adaptive Query Execution coalesces partitions on the fly
Eliminate shuffle joins	Broadcast the small side: '/*+ BROADCAST(t) */'	Keeps data in RAM when feasible
Compress spill data	Ensure spark.shuffle.compress=true (default)	Small CPU cost, large disk savings
Use RAM#backed volumes (SSD#less nodes)	spark.kubernetes.local.dirs.tmpfs=true and raise spark.kubernetes.{driver,executor}.memoryOverheadFactor	Mounts emptyDir as tmpfs
Persist scratch across pod restarts	Mount a PVC at /spark-local with spark.kubernetes.executor.volumes.persistentVolumeClaim.<name>.mount.path	Gives Spark a dedicated disk

Change site-wide ephemeral storage configuration

In Site Administration Settings Advanced, you can see the fields to change the ephemeral storage request (minimum) and maximum limit.



Ephemeral Storage Settings

Ephemeral Storage Request (in GB)

1

The amount of scratch space requested by the session pod.

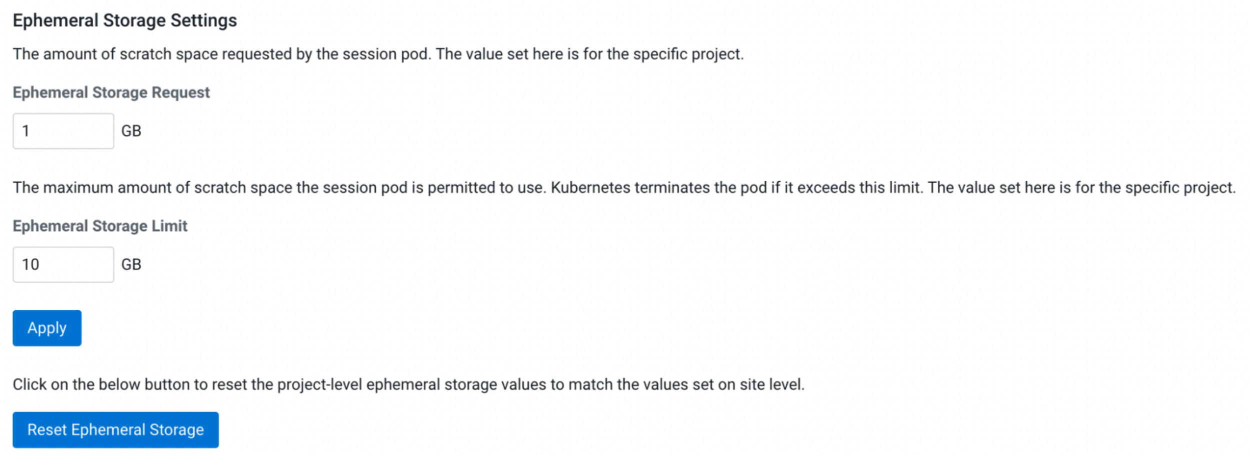
Ephemeral Storage Limit (in GB)

10

The maximum amount of scratch space the session pod is permitted to use. Kubernetes terminates the pod if it exceeds this limit.

Override Site-wide ephemeral storage configuration

If you want to customize the ephemeral storage settings, you can do so on a per-project basis. Open your project, then click on Project Settings Advanced and adjust the ephemeral storage parameters.



Ephemeral Storage Settings

The amount of scratch space requested by the session pod. The value set here is for the specific project.

Ephemeral Storage Request

1 GB

The maximum amount of scratch space the session pod is permitted to use. Kubernetes terminates the pod if it exceeds this limit. The value set here is for the specific project.

Ephemeral Storage Limit

10 GB

Click on the below button to reset the project-level ephemeral storage values to match the values set on site level.

NTP proxy setup on Cloudera AI

Cloudera AI requires specific proxy configurations to manage workbench connections efficiently in an air-gapped setup with restricted outbound connections. This setup ensures seamless access to external resources while adhering to network security and management policies.

Depending on your cluster platform, whether it is RKE2 for Cloudera Embedded Container Service clusters or OpenShift Container Platform, specific configurations and deployment methods apply. Consider the detailed instructions for configuring proxy settings in Cloudera Embedded Container Service server and agent configurations, as well as enabling cluster-wide proxies in OpenShift Container Platform environments. Also consider the provided sample configurations and guidelines for configuring proxy servers, including specifying Classless Inter-domain Routing (CIDR) ranges to exclude from proxy routing and updating proxy server allowlists.

Related Information

Installing a non-transparent proxy in a Cloudera AI environment

Enabling proxies in ECS and OCP environment

Installing in air gap environment

Proxy setting best practices

Updating proxy configuration in an existing workbench

In order to enable new proxy configuration details update the new values under the Cloudera AI Workbench namespace and restart all deployments after the updates.

About this task

Updating proxy configuration in an existing workbench is applicable from Cloudera AI 1.5.4 Cummulative Hotfix 1 (CHF1).

Procedure

1. Update the new values for setting No Proxy under the no_proxy and NO_PROXY sections in the cml_proxy_config configmap under the Cloudera AI Workbench namespace.
2. Restart all deployments under the Cloudera AI namespace to apply all configuration changes. Kubernetes' rolling update mechanism facilitates this process while minimizing service disruption.

Export Usage List

You can export a list of sessions, jobs, workers, and experiments. You can either download a complete list of workloads or you can filter the workloads by date to download a more concise list. Timestamps in the list are given in Coordinated Universal Time (UTC).

Procedure

1. Select Site Administration in the navigation pane.
2. Select the Usage tab.
3. If you want a list of workloads specific to a date range, you can filter the list of workloads by setting the Date Range.
4. Select Export Usage List to download the list of workloads.

Results

The list downloads to your computer as a .csv file.

Disable addons

As a on premises Administrator, you shall ensure that Spark Runtime Addons used on your site are compatible with the base cluster version. In practice, this means you should disable the incompatible versions that may be installed.

1. Go to the Cloudera Management Console, and determine the base cluster version.
2. In each Cloudera AI Workbench, in Site Administration Runtime,
   - Set the Hadoop addon as default that has the base cluster version in its name.
   - Keep those Spark addons enabled that have the base cluster version in their name and disable other Spark Addons.
   - If some workloads have been configured to use a disabled Spark Addon, the affected workloads must be reconfigured to use an enabled Spark Addon. This can happen in the event a workbench is upgraded.

Optimizing performance for scalability

Optimize performance and scalability in Cloudera AI on premises by utilizing the Dashboards Archive feature, introduced in Cloudera AI on premises 1.5.5 CHF1, alongside configuring the livelog retention period and performing database cleanup, introduced in Cloudera AI on premises 1.5.5 SP1. The Dashboards Archive feature is designed to enhance system performance, manage data retention effectively, and provide a clear distinction between active and historical data within the dashboards system. Additionally, further performance optimization can be achieved by configuring the livelog retention period and cleaning up databases in Cloudera AI on premises.

Historical workload cleanup settings

For optimizing performance you can configure livelog retention period and clean up databases from Cloudera AI on premises 1.5.5 SP1. Enable these features in Site Administration Settings under the Historical Workload Cleanup Settings option to enhance and optimize performance.

About this task

Cloudera AI offers enhanced flexibility for managing persistent volume (PV) size. You can configure the livelog retention period and enable a periodic database cleanup feature. The retention period is easily customizable in the Site Administration Settings and defaults to 180 days, allowing you to better adapt to your specific storage needs.

To configure livelog retention and clean up databases, follow the steps:

Procedure

1. In the Cloudera console, click the Cloudera AI tile.

   The Cloudera AI Workbenches page displays.
2. Select the required workbench.

   The Home page of the workbench is displayed.
3. Select Site Administration in the left navigation pane.

   The Site Administration page displays.
4. Select the Settings tab.
5. Navigate to the Historical Workload Cleanup Settings section of the Settings page.

6. Add the number of days for livelog retention period in the Livelog Retention Period (Days) field.

   The default value is 180 days.
7. Select the Clean DB entries checkbox to enable periodic cleanup of database tables.

   The Cleanup DB entries checkbox in the Site Administration Settings allows Administrators to enable periodic cleanup of database tables, including dashboards, dashboard_pods, model_deployments, and user_events, for entries older than the livelog retention period.

   Note:

   The automatic cleanup process does not include the Dashboards Archive tables. If you are using the Dashboards Archive feature, you will need to manually clean up the dashboards_archive table. For details on the Dashboard Archive feature, see Optimized queries with Dashboards Archive table.
8. Select the Update button.

Optimized queries with Dashboards Archive table

The Dashboards Archive feature introduces a mechanism from 1.5.5 CHF1 to optimize performance, manage data retention, and clarify the distinction between active and historical data in the dashboards system.

By archiving older records into a dedicated dashboards_archive table, the size of the primary dashboards table is significantly reduced. This leads to faster query execution for active dashboards, enhancing overall system performance. This improvement is particularly beneficial if you manage large volumes of Apache Spark jobs, for example.

Archived records are stored in the dashboards_archive table, ensuring that historical data is not lost and remains accessible when needed. Active dashboards are retained in the dashboards table. This separation creates a clear distinction between current and historical data, simplifying data management.

APIs referencing the dashboards table do not automatically query the dashboards_archive table. You must run reports before initiating the archive process to ensure all required data is captured.

Features or services that rely on the dashboards table for metrics, such as active dashboards, running jobs, or user activity, do not include archived records in their calculations, so metrics reflect only the current state of active dashboards.

The following APIv2 endpoints do not account for archived records:

- batchListProjects
- getApplication
- getExperiment
- getExperimentRun
- getExperimentRunMetrics
- getJob
- getJobRun
- getProject
- listAllExperiments
- listApplications
- listExperimentRuns
- listExperiments
- listJobRuns
- listJobs
- listProjects

Host name required by Learning Hub

Learning Hub requires internet access to link to the displayed content. Learning Hub cannot be supported on a fully airgapped cluster.

The following domain must be added to the allow list so that links from the content will work:

- *.raw.githubusercontent.com

Configuring HTTP Headers for Cloudera AI

This topic provides guidence on customizing the HTTP headers that are accepted by Cloudera AI.

Required Role: Site Administrator

These properties are available under Site Administration Security .

Enable Cross-Origin Resource Sharing (CORS)

Most modern browsers enforce the Same-Origin Policy, which restricts how a document or a script from one origin can interact with a resource from another origin. When the Enable cross-origin resource sharing property is enabled on Cloudera AI, web servers add the Access-Control-Allow-Origin: * HTTP header to their HTTP responses, allowing web applications from different domains to access the Cloudera AI API through browsers.

This property is disabled by default.

If this property is disabled, web applications from different domains will not be able to programmatically communicate with the Cloudera AI API through browsers.

Enable HTTP security headers

When the Enable HTTP security headers property is enabled, the following HTTP headers are included in HTTP responses from servers:

- X-XSS-Protection
- X-DNS-Prefetch-Control
- X-Frame-Options
- X-Download-Options
- X-Content-Type-Options

This property is enabled by default.

Disabling this property might expose your Cloudera AI deployment to vulnerabilities, such as clickjacking, cross-site scripting (XSS), or other injection attacks.

Enable HTTP Strict Transport Security (HSTS)

Note: If TLS/SSL is not enabled, configuring this property will have no impact on your browser.

When both the TLS/SSL and the Enable HTTP Strict Transport Security (HSTS) property are enabled, Cloudera AI instructs your browser not to load sites using HTTP. Additionally, all attempts to access Cloudera AI using HTTP will automatically be converted to HTTPS.

This property is disabled by default.

To revert back to HTTP, use the following steps:

1. Deactivate the Enable HTTP Strict Transport Security (HSTS) checkbox to disable HSTS and restart Cloudera AI.
2. Load the Cloudera AI web application in each browser to clear the respective browser's HSTS setting.
3. Disable TLS/SSL across the cluster.

By following these instructions, you can prevent users from being locked out of their accounts due to issues caused by browser caching.

Enable HTTP security headers

When Enable HTTP security headers is enabled, the following HTTP headers will be included in HTTP responses from servers:

- X-XSS-Protection
- X-DNS-Prefetch-Control
- X-Frame-Options
- X-Download-Options
- X-Content-Type-Options

This property is enabled by default .

Disabling this property could leave your Cloudera AI deployment vulnerable to clickjacking, cross-site scripting (XSS), or any other injection attacks.

Enable HTTP Strict Transport Security (HSTS)

Note: Without TLS/SSL enabled, configuring this property will have no effect on your browser.

When both TLS/SSL and this property (Enable HTTP Strict Transport Security (HSTS)) are enabled, Cloudera AI will inform your browser that it should never load the site using HTTP. Additionally, all attempts to access Cloudera AI using HTTP will automatically be converted to HTTPS.

This property is disabled by default .

If you ever need to downgrade to back to HTTP, use the following sequence of steps: First, deactivate this checkbox to disable HSTS and restart Cloudera AI. Then, load the Cloudera AI web application in each browser to clear the respective browser's HSTS setting. Finally, disable TLS/SSL across the cluster. Following this sequence shall help tto avoid a situation where users get locked out of their accounts due to browser caching.

Enable Cross-Origin Resource Sharing (CORS)

Most modern browsers implement the Same-Origin Policy, which restricts how a document or a script loaded from one origin can interact with a resource from another origin. When the Enable cross-origin resource sharing property is enabled on Cloudera AI, web servers will include the Access-Control-Allow-Origin: * HTTP header in their HTTP responses. This gives web applications on different domains permission to access the Cloudera AI API through browsers.

This property is disabled by default .

If this property is disabled, web applications from different domains will not be able to programmatically communicate with the Cloudera AI API through browsers.

SSH Keys

This topic describes the different types of SSH keys used by Cloudera AI, and how you can use those keys to authenticate to an external service such as GitHub.

Personal key

Cloudera AI automatically generates an SSH key pair for your user account. You can rotate the key pair and view your public key on your user settings page. It is not possible for anyone to view your private key.

Every console you run has your account's private key loaded into its SSH-agent. Your consoles can use the private key to authenticate to external services, such as GitHub. For instructions, see *Adding an SSH Key to GitHub*.

Related Information

Adding an SSH Key to GitHub

Team key

Team SSH keys provide a useful way to give an entire team access to external resources such as databases or GitHub repositories (as described in the next section).

Like Cloudera AI users, each Cloudera AI team has an associated SSH key. You can access the public key from the team's account settings. Click Account, then select the team from the drop-down menu at the upper right corner of the page.

When you launch a console in a project owned by a team, you can use that team's SSH key from within the console.

Adding an SSH key to GitHub

Cloudera AI creates a public SSH key for each account. You can add this SSH public key to your GitHub account if you want to use password-protected GitHub repositories to create new projects or collaborate on projects.

Procedure

1. Sign in to Cloudera AI.
2. Go to the upper right drop-down menu and switch context to the account whose key you want to add. This could be a individual user account or a team account.
3. On the left sidebar, click User Settings.
4. Go to the Outbound SSH tab and copy the User Public SSH Key.
5. Sign in to your GitHub account and add the Cloudera AI key copied in the previous step to your GitHub account. For instructions, refer the GitHub documentation on Adding a new SSH key to your GitHub account.

Creating an SSH tunnel

You can use your SSH key to connect Cloudera AI to an external database or cluster by creating an SSH tunnel.

About this task

In some environments, external databases and data sources reside behind restrictive firewalls. A common pattern is to provide access to these services using a bastion host with only the SSH port open. Cloudera AI provides a convenient way to connect to such resources using an SSH tunnel.

If you create an SSH tunnel to an external server in one of your projects, then all engines that you run in that project are able to connect securely to a port on that server by connecting to a local port. The encrypted tunnel is completely transparent to the user and code.

Procedure

1. Open the Project Settings page.
2. Open the Tunnels tab.
3. Click New Tunnel.
4. Enter the server IP Address or DNS hostname.
5. Enter your username on the server.
6. Enter the local port that should be proxied, and to which remote port on the server.

What to do next

On the remote server, configure SSH to accept password-less logins using your individual or team SSH key. Often, you can do so by appending the SSH key to the file /home/username/.ssh/authorized_keys.

Hadoop authentication for Cloudera AI Workbenches

Cloudera AI does not assume that your Kerberos principal is always the same as your login information. Therefore, you will need to make sure Cloudera AI knows your Kerberos identity when you sign in.

About this task

This procedure is required if you want to run Spark workloads in an Cloudera AI Workbench. This is also required if connecting Cloudera Data Visualization running in Cloudera AI to an Impala instance using Kerberos for authentication.

Procedure

1. Navigate to your Cloudera AI Workbench.
2. Go to the top-right dropdown menu, click Account settings Hadoop Authentication .
3. To authenticate, either enter your password or click Upload Keytab to upload the keytab file directly.

Results

Once successfully authenticated, Cloudera AI uses your stored credentials to ensure you are secure when running workloads.

Cloudera AI and outbound network access

Cloudera AI expects access to certain external networks. See the related information *Configuring proxy hosts for Cloudera AI Workbench connections* for further information.

Note: The outbound network access destinations listed in *Configuring proxy hosts for Cloudera AI Workbench connections* are only the minimal set required for Cloudera installation and operation. For environments with limited outbound internet access due to using a firewall or proxy, access to Python or R package repositories such as Python Package Index or CRAN may need to be whitelisted if your use cases require installing packages from those repositories. Alternatively, you may consider creating mirrors of those repositories within your environment.

Related Information

Configuring proxy hosts for Cloudera AI workbench connections