

Managing Projects in Cloudera Data Science Workbench

Date published: 2020-02-28

Date modified: 2020-12-15

The Cloudera logo is displayed in a bold, orange, sans-serif font. The word "CLOUDERA" is written in all caps, with a stylized 'E' that has a horizontal bar extending to the right.

Legal Notice

© Cloudera Inc. 2024. 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.

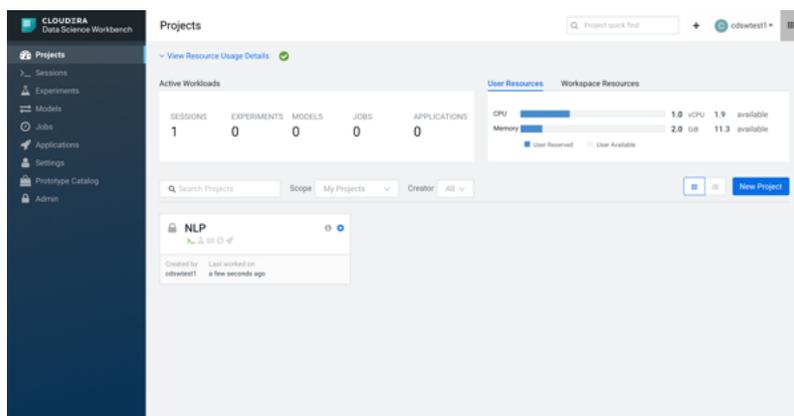
Contents

Managing Projects in Cloudera Data Science Workbench.....	4
Creating a Project with Legacy Engine Variants.....	4
Creating a Project with ML Runtimes Variants.....	5
Adding Collaborators.....	7
Modifying Project Settings.....	7
Managing Files.....	8
Disabling Project File Uploads and Downloads.....	9
Custom Template Projects.....	9
Deleting a Project.....	10

Managing Projects in Cloudera Data Science Workbench

Projects form the heart of Cloudera Data Science Workbench. They hold all the code, configuration, and libraries needed to reproducibly run analyses. Each project is independent, ensuring users can work freely without interfering with one another or breaking existing workloads.

Access the Projects page by clicking Projects in the navigation panel. The Projects page gives you a quick summary of project information.



- Active Workloads - If there are active workloads running, this section describes the number of Sessions, Experiments, Models, Jobs, and Applications that are running.
- Resource Usage Details - A collapsible section that displays resource usage.
 - Active Workloads - If there are active workloads running, this section describes the number of Sessions, Experiments, Models, Jobs, and Applications that are running.
 - User Resources and Workspace Resources
 - Click on the User Resources tab to see the CPU and memory resource usage for the user. The maximum usage of the vCPU and GB is calculated based on whether or not you have a quota. If you have a quota, the maximum usage will be based on your quota. If you don't have a quota, the maximum usage will be what is available on the cluster. If you have a GPU, you'll also see the GPU usage.
 - Click on the Workspace Resources tab to see usage overall.
- Search Projects - Enter a term for keyword search across Project names.
- Scope - An additional filter only viewable by Administrators.
 - Selecting My Projects displays only the Projects that you have created or are a Collaborator of.
 - Selecting All Projects displays all Projects on the Workspace.
- Creator - An additional filter to only display Projects created by a specified user.
- Projects View Selector - A setting that enables you to display Projects in a summary card-based view or a detailed table-based view.

Creating a Project with Legacy Engine Variants

To create a Cloudera Data Science Workbench project:

Procedure

1. Go to Cloudera Data Science Workbench and on the left sidebar, click Projects.
2. Click New Project.

3. If you are a member of a team, from the drop-down menu, select the Account under which you want to create this project. If there is only one account on the deployment, you will not see this option.
4. Enter a Project Name.
5. Select Project Visibility from one of the following options.
 - Private - Only project collaborators can view or edit the project.
 - Team - If the project is created under a team account, all members of the team can view the project. Only explicitly-added collaborators can edit the project.
 - Public - All authenticated users of Cloudera Data Science Workbench will be able to view the project. Collaborators will be able to edit the project.
6. Under Initial Setup, you can either create a blank project, or select one of the following sources for your project files.
 - Built-in Templates - Template projects contain example code that can help you get started with the Cloudera Data Science Workbench. They are available in R, Python, PySpark, and Scala. Using a template project is not required, but it helps you start using the Cloudera Data Science Workbench right away.

Custom Templates - Starting with version 1.3, site administrators can add template projects that are customized for their organization's use-cases. For details, see [Custom Template Projects](#).
 - Local - If you have an existing project on your local disk, use this option to upload compressed files or folders to Cloudera Data Science Workbench.
 - Git - If you already use Git for version control and collaboration, you can continue to do so with the Cloudera Data Science Workbench. Specifying a Git URL will clone the project into Cloudera Data Science Workbench. If you use a Git SSH URL, your personal private SSH key will be used to clone the repository. This is the recommended approach. However, you must add the public SSH key from your personal Cloudera Data Science Workbench account to the remote Git hosting service before you can clone the project. Specify your username and password in the URL as follows:

```
http://username:password@server/path/project.git
```

7. Click Create Project.

After the project is created, you can see your project files and the list of jobs defined in your project.



Note: As part of the project filesystem, Cloudera Data Science Workbench also creates the following .git ignore file.

```
R
node_modules
*.pyc
.*
!.gitignore
```

8. (Optional) To work with team members on a project, use the instructions in the following section to add them as collaborators to the project.

Creating a Project with ML Runtimes Variants

Projects create an independent working environment to hold your code, configuration, and libraries for your analysis. This topic describes how to create a project with ML Runtimes variants in Cloudera Data Science Workbench.

Procedure

1. Go to Cloudera Data Science Workbench and on the left sidebar, click Projects.
2. Click New Project.
3. If you are a member of a team, from the drop-down menu, select the Account under which you want to create this project. If there is only one account on the deployment, you will not see this option.

4. Enter a Project Name.
5. Select Project Visibility from one of the following options.
 - Private - Only project collaborators can view or edit the project.
 - Team - If the project is created under a team account, all members of the team can view the project. Only explicitly-added collaborators can edit the project.
 - Public - All authenticated users of Cloudera Machine Learning will be able to view the project. Collaborators will be able to edit the project.
6. Under Initial Setup, you can either create a blank project, or select one of the following sources for your project files.
 - Blank - The project will contain no information from a template, local file, or Git.
 - Template - Template projects contain example code that can help you get started with Cloudera Machine Learning. They are available in R, Python, PySpark, and Scala. Using a template project is not required, but it helps you start using Cloudera Machine Learning right away.



Note: Templates are designed for Engines and might not work with ML Runtimes.

- ML Prototype - You can choose a prototype that will include components to create a complete project.
 - Git - If you already use Git for version control and collaboration, you can continue to do so with Cloudera Machine Learning. Specifying a Git URL will clone the project into Cloudera Machine Learning. If you use a Git SSH URL, your personal private SSH key will be used to clone the repository. This is the recommended approach. However, you must add the public SSH key from your personal Cloudera Machine Learning account to the remote Git hosting service before you can clone the project.
7. Click Create Project. After the project is created, you can see your project files and the list of jobs defined in your project.

Note that as part of the project filesystem, Cloudera Machine Learning also creates the following `.gitignore` file.

```
R
node_modules
*.pyc
.*
!.gitignore
```

8. Set or verify the ML Runtimes settings for the project.

Within the selected project, you can modify the default engine configuration:

- a) In the left navigation bar, click Project Settings.
- b) Select the Runtime/Engine tab.
- c) Next to Default Engine, select ML Runtime.



Note: The ability to switch the default engine is intended to support migration of projects from Engines to Runtimes. Cloudera does not recommend making this change frequently.

Once switched to ML Runtimes, all modules (Sessions, Jobs, Models, etc.) for the project will configure using ML Runtimes instead of Legacy Engines. Setting Engine parameters for the project will no longer be possible.

Existing and running instances (for example, Jobs, Models or Applications) previously configured with a particular Engine configuration will keep their configuration until you change the related settings. If you want to change the engine configuration for existing and running instances, you will need to update those based on the new, Runtime-based settings.



Note: Note that site administrators now have the ability to select ML Runtimes as a default for newly created projects. This setting can be found within the Admin menu below the Runtime/Engine tab.

Adding Collaborators

If you want to work closely with colleagues on a particular project, use the following steps to add them to the project.

Procedure

1. Navigate to the project overview page.
2. Click Team to open the Collaborators page.
3. Search for collaborators by either name or email address and click Add.

For a project created under your personal account, anyone who belongs to your organization can be added as a collaborator. For a project created under a team account, you can only add collaborators that already belong to the team. If you want to work on a project that requires collaborators from different teams, create a new team with the required members, and then create a project under that account. If your project was created from a Git repository, each collaborator must create the project from the same central Git repository.

You can grant project collaborators one of three levels of access:

- Viewer - Read-only access to code, data, and results.
- Operator - Read-only access to code, data, and results. Additionally, Operators can start and stop existing jobs in the projects that they have access to.
- Contributor - Can view, edit, create, and delete files and environmental variables, run sessions/experiments/jobs/models and execute code in running jobs. Additionally, Contributors can set the default engine for the project.
- Admin - Has complete access to all aspects of the project. This includes the ability to add new collaborators, and delete the entire project.



Note:

Collaborating Securely on Projects

Before adding project collaborators, you must remember that assigning the Contributor or Admin role to a project collaborator is the same as giving them write access to your data in CDH. This is because project contributors and project administrators have write access to all your project code (including any library code that you might not be actively inspecting). For example, a contributor/admin could modify project file(s) to insert code that deletes some data on the CDH cluster. The next time you launch a session and run the same code, it will appear as though you deleted the data yourself.

Additionally, project collaborators also have access to all actively running sessions and jobs. This means that a malicious user can easily impersonate you by accessing one of your active sessions. Therefore, it is extremely important to restrict project access to trusted collaborators only. Note that Cloudera Data Science Workbench 1.4.3 introduces a new feature that allows site administrators to restrict this ability by allowing only session creators to execute commands within their own active sessions. For details, see [Restricting Access to Active Sessions](#).

For these reasons, Cloudera recommends using Git to collaborate securely on shared projects. This will also help avoid file modification conflicts when your team is working on more elaborate projects.

For more information on collaborating effectively, see [Collaborating on Projects with Cloudera Data Science Workbench](#).

Modifying Project Settings

Project contributors and administrators can modify aspects of the project environment such as the engine or ML Runtimes used to launch sessions, the environment variables, and create SSH tunnels to access external resources.

Procedure

1. Switch context to the account where the project was created.

2. Click Projects.
3. From the list of projects, select the one you want to modify.
4. Click Settings to open up the Project Settings dashboard.

Options

Modify the project name and its privacy settings on this page.

Runtime/Engine

Cloudera Data Science Workbench ensures that your code is always run with the specific engine version you selected. You can select to use either Legacy Engines or Machine Learning Runtimes. For legacy engines only, you can also select the engine version and add third-party editors here. For advanced use cases, Cloudera Data Science Workbench projects can use custom Docker images for their projects. Site administrators can include images in the allowlist for use in projects, and project administrators can use this page to select which of these whitelisted images is installed for their projects. For an example, see [Customized Engine Images](#).

Environment - If there are any environmental variables that should be injected into all the engines running this project, you can add them to this page. For more details, see [Engine Environment Variables](#).

Tunnels

In some environments, external databases and data sources reside behind restrictive firewalls. Cloudera Data Science Workbench provides a convenient way to connect to such resources using your SSH key. For instructions, see [SSH Tunnels](#).

Delete Project

This page can only be accessed by project administrators. Remember that deleting a project is irreversible. All files, data, sessions, and jobs will be lost.

Managing Files

Cloudera Data Science Workbench allows you to move, rename, copy, and delete files within the scope of the project where they live.

About this task

You can also upload new files to a project, or download project files. Files can only be uploaded within the scope of a single project. Therefore, to access a script or data file from multiple projects, you will need to manually upload it to all the relevant projects.



Note: For use cases beyond simple projects, Cloudera strongly recommends using Git to manage your projects using version control.

Procedure

1. Switch context to the account where the project was created.
2. Click Projects.
3. From the list of projects, click on the project you want to modify. This will take you to the project overview.

4. Click Files.

Upload Files to a Project

Click Upload. Select Files or Folder from the dropdown, and choose the files or folder you want to upload from your local filesystem.

In addition to uploading files or a folder, you can upload a .tar file of multiple files and folders. After you select and upload the .tar file, you can use a terminal session to extract the contents:

- a. On the project overview page, click Open Workbench and select a running session or create a new one.
- b. Click Terminal access.
- c. In the terminal window, extract the contents of the .tar file:

```
tar -xvf <file_name>.tar.gz
```

The extracted files are now available for the project.

Download Project Files

Click Download to download the entire project in a .zip file. To download only a specific file, select the checkbox next to the file(s) to be download and click Download.

You can also use the checkboxes to Move, Rename, or Delete files within the scope of this project.

Disabling Project File Uploads and Downloads

By default, all Cloudera Data Science Workbench users are allowed to upload and download files to/from a project. Version 1.5 introduces a new feature flag that allows site administrators to hide the UI features that let users upload and download project files.

About this task

This feature flag only removes the relevant features from the Cloudera Data Science Workbench UI. It does not disable the ability to upload and download files through the backend web API.



Note: You cannot disable file upload and download when using the Jupyter Notebook.

Procedure

1. Go to Admin Security .
2. Under the File Upload/Download section, disable the Allow file upload/download through UI checkbox.

Custom Template Projects

Site administrators can add template projects that have been customized for their organization's use-cases. These custom project templates can be added in the form of a Git repository.

To add a new template project, go to Admin Settings . Under the Project Templates section, provide a template name, the URL to the project's Git repository, and click Add.

The added templates will become available in the Template tab on the Create Project page. Site administrators can add, edit, or delete custom templates, but not the built-in ones. However, individual built-in templates can be disabled using a checkbox in the Project Templates table at Admin Settings .

Deleting a Project

Deleting a project is an irreversible action. All files, data, and history related to the project will be lost. This includes any jobs, sessions or models you created within the project.

Procedure

1. Go to the project Overview page.
2. On the left sidebar, click Settings.
3. Go to the Delete Project.
4. Click Delete Project and click OK to confirm.