Collaborating on Projects with Cloudera Data Science Workbench

Date published: 2020-02-28
Date modified: 2020-12-15

CLOUDERA

Legal Notice

Contents

Collaborating on Projects with Cloudera Data Science Workbench

Cloudera Data Science Workbench supports several collaboration models.

Project Collaborators

If you want to work closely with trusted colleagues on a particular project, you can add them to the project as collaborators.

For instructions, see Adding Collaborators.

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.

Restricting Collaborator and Administrator Access to Active Sessions

By default, the following Cloudera Data Science Workbench users have the ability to execute commands within any active sessions you have created.

About this task

Required Role: Site Administrator

- All Site Administrators
- Users who have been assigned Admin or Contributor privileges for the project where the session is created.
- For team projects, Team Admins have complete access to all team projects and any active sessions running within these projects. Additionally, any team members who have been assigned the Admin or Contributor roles for your projects will also have the ability to execute commands within your active sessions.

Starting with Cloudera Data Science Workbench 1.4.3, site administrators can now restrict this ability by allowing only the user who launched the session to execute commands within their own active sessions. To enable this restriction:

Procedure

1. Log into Cloudera Data Science Workbench with site administrator privileges.
2. Click Admin Security .

3. Under the General section, select the checkbox to enable the Only session creators can execute commands on active sessions property.

When this property is enabled, only the user that creates a session will be able to execute commands in that session. No other users, regardless of their permissions in the team or as project collaborators, will be able to execute commands on active sessions that are not created by them. Even site administrators will not be able to execute commands in other users' active sessions. However, keep in mind that all site administrators still have access to the Site Administrator dashboard and can reverse this change at any time.

Teams

Users who work together on more than one project and want to facilitate collaboration can create a Team. Teams allow streamlined administration of projects.

Team projects are owned by the team, rather than an individual user. Team administrators can add or remove members at any time, assigning each member different permissions.

For more details, see:

- Creating a Team without an Associated LDAP Group
- Creating a Team with an Associated LDAP Group
- Modifying Team Account Settings

Sharing Personal Projects

When you create a project in your personal context, Cloudera Data Science Workbench asks you to assign one of the following visibility levels to the project - Private or Public. Public projects on Cloudera Data Science Workbench grant read-level access to everyone with access to the Cloudera Data Science Workbench application. That means everyone can view the project's files and results, but only those whom you have explicitly added as a collaborator can edit files, run engines, or view the project's environment variables.

You can include a markdown-formatted README.md file in public projects to document your project's purpose and usage.

If you are a project admin, you can set a project's visibility to Public from the Project Settings Options page. For instructions, see Modifying Project Settings.

Forking Projects

You can fork another user's project by clicking Fork on the Project page.

Forking creates a new project under your account that contains all the files, libraries, configuration, and jobs from the original project.

Creating sample projects that other users can fork helps to bootstrap new projects and encourage common conventions.

Note: An issue exists where a timeout might occur when forking large projects.

Collaborating with Git

Cloudera Data Science Workbench provides seamless access to Git projects.

Whether you are working independently, or as part of a team, you can leverage all of benefits of version control and collaboration with Git from within Cloudera Data Science Workbench. Teams that already use Git for collaboration

can continue to do so. Each team member will need to create a separate Cloudera Data Science Workbench project from the central Git repository.

For anything but simple projects, Cloudera recommends using Git for version control. You should work on Cloudera Data Science Workbench the same way you would work locally, and for most data scientists and developers that means using Git.

For more details, see Using Git to Collaborate on Projects.

Sharing Job and Session Console Outputs

Cloudera Data Science Workbench lets you easily share the results of your analysis with one click.

Using rich visualizations and documentation comments, you can arrange your console log so that it is a readable record of your analysis and results. This log continues to be available even after the session stops. This method of sharing allows you to show colleagues and collaborators your progress without your having to spend time creating a report.

To share results from an interactive session, click Share at the top of the console page. From here you can generate a link that includes a secret token that gives access to that particular console output. For jobs results, you can either share a link to the latest job result or a particular job run. To share the latest job result, click the Latest Run link for a job on the Overview page. This link will always have the latest job results. To share a particular run, click on a job run in the job's History page and share the corresponding link.

You can share console outputs with one of the following sets of users.

- All anonymous users with the link - By default, Cloudera Data Science Workbench allows anonymous access to shared consoles. However, site administrators can disable anonymous sharing at any time by going to Admin Security , disabling the Allow anonymous access to shared console outputs checkbox, and clicking Disable anonymous access to confirm.

  Once anonymous sharing has been disabled, all existing publicly shared console outputs will be updated to be viewable only by authenticated users.
- All authenticated users with the link - This means any user with a Cloudera Data Science Workbench account will have access to the shared console.
- Specific users and teams - Click Change to search for users and teams to give access to the shared console. You can also come back to the session and revoke access from a user or team the same way.

Sharing Data Visualizations

If you want to share a single data visualization rather than an entire console, you can embed it in another web page. Click the small circular 'link' button located to the left of most rich visualizations to view the HTML snippet that you can use to embed the visualization.

Sharing Data Visualizations

If you want to share a single data visualization rather than an entire console, you can embed it in another web page.

Procedure

Click the small circular 'link' button located to the left of most rich visualizations to view the HTML snippet that you can use to embed the visualization.

Using Git to Collaborate on Projects

Cloudera Data Science Workbench provides seamless access to Git projects. Whether you are working independently, or as part of a team, you can leverage all of benefits of version control and collaboration with Git from within

Cloudera Data Science Workbench. Teams that already use Git for collaboration can continue to do so. Each team member will need to create a separate Cloudera Data Science Workbench project from the central Git repository.

For anything but simple projects, Cloudera recommends using Git for version control. You should work on Cloudera Data Science Workbench the same way you would work locally, and for most data scientists and developers that means using Git.

Cloudera Data Science Workbench does not include significant UI support for Git, but instead allows you to use the full power of the command line. If you run an engine and open a terminal, you can run any Git command, including `init`, `add`, `commit`, `branch`, `merge` and `rebase`. Everything should work exactly as it does locally, except that you are running on a distributed edge host directly connected to your Apache Hadoop cluster.

Importing a Project From Git

When you create a project, you can optionally supply an HTTPS or SSH Git URL that points to a remote repository. The new project is a clone of that remote repository. You can commit, push and pull your code by running a console and opening a terminal .

Procedure

Using SSH - If you want to use SSH to clone the repo, you will need to first add your personal Cloudera Data Science Workbench SSH key to your GitHub account.

For instructions, see Adding SSH Key to GitHub.

If you see Git commands hanging indefinitely, check with your cluster administrators to make sure that the SSH ports on the Cloudera Data Science Workbench hosts are not blocked.

Linking an Existing Project to a Git Remote

If you did not create your project from a Git repository, you can link an existing project to a Git remote (for example, git@github.com:username/repo.git) so that you can push and pull your code.

About this task
To link to a Git remote:

Procedure

1. Launch a new session.
2. Open a terminal.
3. Enter the following commands:

   Shell

   ```
   git init
   git add *
   git commit -a -m 'Initial commit'
   git remote add origin git@github.com:username/repo.git
   ```

   You can run git status after git init to make sure your .gitignore includes a folder for libraries and other non-code artifacts.