

Cloudera Data Science Workbench Security Guide

Date published: 2020-02-28

Date modified:



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

Cloudera Data Science Workbench Security Guide.....	4
Security Model.....	4
Wildcard DNS Subdomain Requirement.....	5
Authentication.....	5
Authorization.....	6
Cluster Authorization.....	6
User Role Authorization.....	6
Access Control for Teams and Projects.....	6
Wire Encryption.....	10
External Communications.....	10
Internal Communications.....	10
Cloudera Data Science Workbench Gateway Host Security.....	11
Base Engine Image Security.....	11

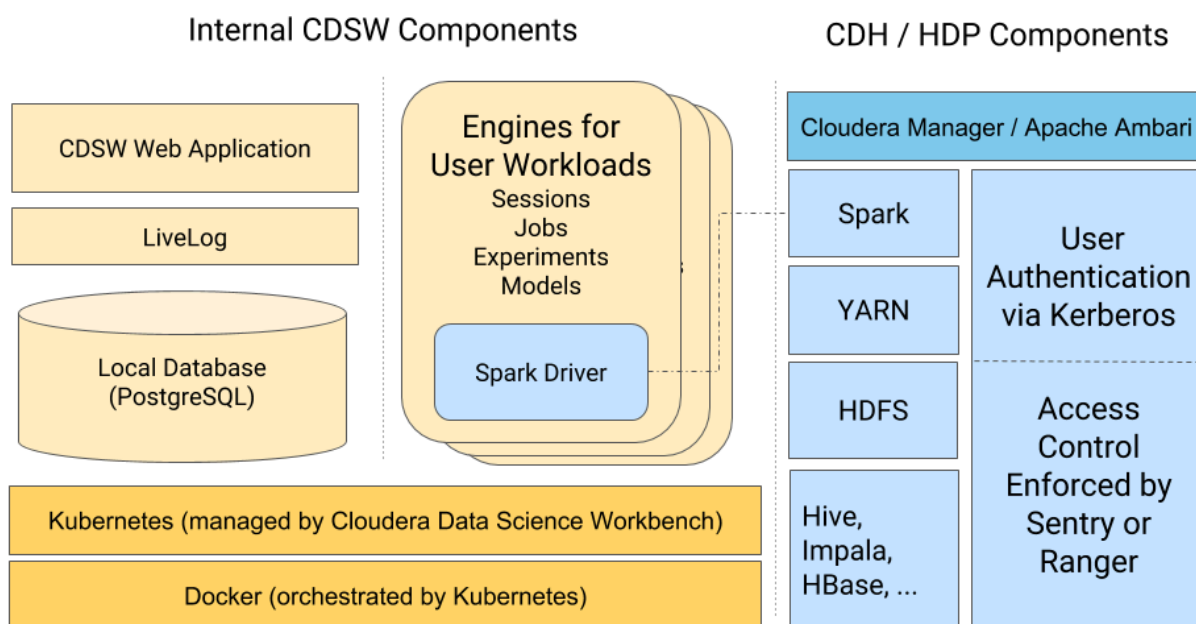
Cloudera Data Science Workbench Security Guide

This topic provides an overview of the Cloudera Data Science Workbench security model and describes how Cloudera Data Science Workbench leverages the [security and governance capabilities of your Cloudera Enterprise cluster](#) to deliver a secure, robust, collaborative data science platform for the enterprise.

Security Model

Cloudera Data Science Workbench uses Kubernetes to schedule and manage Docker containers.

All Cloudera Data Science Workbench components (web application, PostgreSQL database, LiveLog, and so on) run inside Docker containers. These are represented on the left-hand side of the diagram. Similarly, the environments that users operate in (via sessions, jobs, experiments, models), also run within isolated Docker containers that we call engines.



This architecture allows Cloudera Data Science Workbench to leverage the isolation properties of Docker to achieve notable security benefits.

Benefits of the Docker Isolation Model - Docker containers share the underlying host operating system, but are isolated from the rest of the host and from each other. Each container gets its own:

- **Isolated File System:** The Docker container does not see the host file system, but instead sees only the filesystem provided by the container and any host volumes that you have explicitly mounted into the container. This means a user launching a Cloudera Data Science Workbench session will only have access to the project files, and any specific host volumes you have chosen to mount into the session engine. They will not otherwise have access to the underlying host filesystem.
- **Isolated Process Namespace:** Docker containers cannot affect any processes running either on the host operating system or in other containers. Cloudera Data Science Workbench creates a new container each time a session/job/experiment/model is launched. This means user workloads can run in complete isolation from each other.



Note: Running CDSW as a non-root machine user is not, and will not, be supported. This issue has been resolved in Private Cloud.

There are no security concerns Cloudera is aware of. The root user is required for the Kubernetes and Docker systems and their ability to read/write to the file system, however all user sessions and user Docker containers run as the CDSW user, not as the root user, so end users are not able to ever use root.

For more details on the Docker security model, see [Docker Security Overview](#).

Wildcard DNS Subdomain Requirement

When you first set up Cloudera Data Science Workbench, you are asked to create a wildcard DNS entry for the Cloudera Data Science Workbench domain. Cloudera Data Science Workbench uses these wildcard subdomains (*.cdsw.<your_domain>.com) to route HTTP requests to engines and services launched by users.

Every time users launch workloads (session/job/experiment/model) on Cloudera Data Science Workbench, a new engine is created for each workload. These engines are isolated Docker containers where users can run code. Each engine is assigned its own unique, randomly-generated ID, which is saved to the CDSW_ENGINE_ID environment variables. This ID is also used to create a unique subdomain for each engine. These subdomains are of the form: <CDSW_ENGINE_ID>.cdsw.<your_domain>.com.

Assigning a unique subdomain to each engine allows Cloudera Data Science Workbench to:

- Securely expose interactive session services, such as visualizations, the terminal, and web UIs such as TensorBoard, Shiny, Plotly, and so on;
- Prevent cross-site scripting (XSS) attacks by securely isolating user-generated content from the Cloudera Data Science Workbench application.

It is important to note that because there is no limit to the number of workloads (i.e. engines) users can launch, Cloudera Data Science Workbench requires the ability to randomly generate large numbers of engine IDs (and their subdomains) on-demand. Therefore, creating a wildcard DNS subdomain is essential for Cloudera Data Science Workbench to function successfully.

Additionally, if you want to enable TLS for your deployment, your TLS certificate will need to include both, the Cloudera Data Science Workbench domain, as well as the wildcard for all first-level subdomains. This is required so that your browser can trust communications with the <CDSW_ENGINE_ID>.cdsw.<your_domain>.com subdomains.

Authentication

Authentication occurs in various forms in the Cloudera Data Science Workbench application.

These are:

- **User Login:** Occurs when users log in to the Cloudera Data Science Workbench Web UI and authenticate themselves to the application. Cloudera Data Science Workbench works with the following authentication backends: the local CDSW database, LDAP/AD, and SAML. Using either [LDAP](#) or [SAML](#) is recommended, as it eases the administrative burden of managing identity in Cloudera Data Science Workbench.
- **SSH Key Authentication:** Each user account is assigned an [SSH key pair](#) for use in sessions. This SSH key pair can be used to authenticate to an external version control system such as Git. Only the public key appears in the UI; the private key is loaded into user sessions when launched.
- **Hadoop Cluster Authentication:** Authentication to the underlying CDH/HDP cluster is handled via Kerberos. To authenticate themselves to the cluster, users must provide a Kerberos principal and keytab/password. These credentials are stored in the internal Cloudera Data Science Workbench database. Note that Cloudera Data Science Workbench user sessions are only provided with the Kerberos credential-cache file which does not store the user's password. For more details, see [Hadoop Authentication with Kerberos for Cloudera Data Science Workbench](#).
- **API Authentication:** Each user account is assigned an API Key that can be used for authentication when communicating with the Cloudera Data Science Workbench API. For more details, see [Cloudera Data Science Workbench Jobs API](#).

- **Per-Model Authentication:** Each model is assigned a unique Access Key. This access key serves as an authentication token that allows users to make calls to the model once it has been deployed. For details, see [Models Access Key](#).

Authorization

This section describes how Cloudera Data Science Workbench handles authorization for users attempting to connect to the attached CDH/HDP cluster, and how Cloudera Data Science Workbench users and teams are granted access to projects.

Cluster Authorization

Cloudera Data Science Workbench connects to your cluster like any other client. It relies on your cluster's built-in security layer to enforce strong authentication and authorization for connections from Cloudera Data Science Workbench.

Cloudera Data Science Workbench users attempting to connect to a secure CDH or HDP cluster must first authenticate themselves using their Kerberos. Once a user has been successfully authenticated, Cloudera Data Science Workbench then depends on your cluster's built-in authorization tools (such as HDFS ACLs, Sentry, Ranger, and so on) to enforce their own existing access control rules when a Cloudera Data Science Workbench user attempts to access data on the cluster.

For example, let's assume you are running a secure CDH cluster with Impala (secured by Sentry). On this cluster, only the `impala-admin` group has been granted the required Sentry privileges to modify data in Impala. If a Cloudera Data Science Workbench user runs a Python job that attempts to modify data in an Impala table, Sentry privileges will be enforced. That is, if the user is not a member of the `impala-admin` group, the access request will be denied.

Similarly, if a Cloudera Data Science Workbench user is attempting to submit Spark jobs on secure HDP clusters, Ranger policies (via the respective HDFS and YARN plugins) will be enforced to decide whether the user has the permission to access the requested directories in HDFS, and whether they can submit jobs to the Spark queue.

This means as long as your cluster components have been secured with access control rules (via ACLs or Sentry or Ranger) that forbid specific users/groups from performing certain actions on the cluster, then all access attempts from Cloudera Data Science Workbench users will also be subject to those rules.

User Role Authorization

Cloudera Data Science Workbench has one major specialized user role across the deployment: site administrators.

Site administrators are superusers who have complete access to all activity on your Cloudera Data Science Workbench deployment. This includes all the configuration settings, projects (private and public), and workloads running on the deployment.

When LDAP or SAML based authentication is used, it is possible to restrict the users who can log into Cloudera Data Science Workbench based on their LDAP/SAML group membership. Additionally, you can specify LDAP/SAML groups that should automatically be granted site administrator privileges when they log in to Cloudera Data Science Workbench.

Apart from site administrators, every other Cloudera Data Science Workbench user is granted access to projects either as a project collaborator or as part of a team. Continue reading to find out how these project and team permissions interact with each other.

Access Control for Teams and Projects

When a team or project is created, the Team/Project Admin role is assigned to the user who created it. Other Team/Project Admins can be assigned later but there must always be at least one user assigned as Admin for the team or project. Team and project administrators then decide what level of access other users are granted per-team or per-project.

Project Access Levels

Users who are explicitly added to a project are referred to as project collaborators. Project collaborators can be assigned one of the following 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.

Team Access Levels

Users who are explicitly added to a team are referred to as team members. Team members can be assigned one of the following levels of access:

- Viewer - Read-only access to team projects. Cannot create new projects within the team but can be added to existing ones.
- Operator - Read-only access to team projects. Cannot create new projects within the team but can be added to existing ones. Additionally, Operators can start and stop existing jobs in the projects that they have access to.
- Contributor - Write-level access to all team projects to all team projects with Team or Public visibility. Can create new projects within the team. They can also be added to existing team projects.
- Admin - Has complete access to all team projects, can add new team members, and modify team account information.

Project Visibility Levels

Projects can be created either in your personal context, or in a team context. Furthermore, projects can be created with one of the following visibility levels:

- Private - Private projects can be created either in your personal context, or in a team context. They can only be accessed by project collaborators.
- Team - Team projects can only be created in a team context. They can be viewed by all members of the team.
- Public - Public projects can be created either in your personal context, or in a team context. They can be viewed by all authenticated Cloudera Data Science Workbench users.

It is important to remember that irrespective of the visibility level of the project, site administrators will always have complete Admin-level access to all projects on Cloudera Data Science Workbench. Additionally, depending on the visibility level of the project, and the context in which it was created, a few other users/team members might also have Contributor or Admin-level access to your project by default.

Use the following table to find out who might have default access to your projects on Cloudera Data Science Workbench.

Project Visibility	Access Levels for Cloudera Data Science Workbench Users
Private Visibility	Private Projects Created in Personal Context The following user roles will have access to private projects in your personal context: Admin Access • Site Administrators • Project Admins Contributor Access • Collaborators explicitly added to the project and given Contributor access. Operator Access • Operators explicitly added to the project and given Operator access. Viewer Access • Viewers explicitly added to the project and given Viewer access.
	Private Projects Created in a Team Context For private projects created within a team context, project-level permissions granted by Project Admins will take precedence over team-level permissions. The only exception to this rule are users who are Team Admins. Team Admins will always have Admin-level access to all projects within their team context, irrespective of the access level granted to them per-project. The following user roles will have access to private projects created within a team context: Admin Access • Site Administrators • Project Admins • Team Admins Contributor Access • Collaborators explicitly added to the project and given Contributor access. Operator Access • Operators explicitly added to the project and given Operator access. Viewer Access • Viewers explicitly added to the project and given Viewer access.

Project Visibility	Access Levels for Cloudera Data Science Workbench Users
Team Visibility	Team Projects Projects with Team visibility can only be created in a team context. For team projects, both team access levels and project access levels must be taken into consideration to determine who has access to these projects. Points to note: - Team members do not need to be explicitly added as project collaborators to have access to a team project. While you can explicitly invite specific team members to collaborate on your project, it is important to remember that all team members will have some level of access to your project. - The project Collaborators page does not list all team members; it only lists those you have explicitly added as collaborators. However, team members will still have access to all team projects. By default, their level of access to the projects is the same as their level of access to the team. - Project Admins cannot downgrade access levels for team members. If project-level permissions don't match up to team-level permissions, team permissions will take precedence. For example, if you add a Team Contributor as a collaborator to a team project, but only give them Project Viewer permission, the user will still have Contributor-level access to the project. Similarly, Team Admins will always have Admin-level access to all projects within their team context, irrespective of the access level granted to them per-project. - Project Admins also cannot upgrade access levels for team members with Viewer-level access to the team. That is, Team Viewers cannot be given Contributor or Admin access to any team projects. The following user roles will have access to team projects on Cloudera Data Science Workbench: Admin Access - Site Administrators - Team Admins - Project Admins Contributor Access - All team members with Contributor access. Operator Access - All team members with Operator access. Viewer Access - All team members with Viewer access.
Public Visibility	Public Projects Created in Personal Context The following user roles will have access to public projects on Cloudera Data Science Workbench: Admin Access - Site Administrators - Project Admins Contributor Access - Collaborators explicitly added to the project and given Contributor access. Operator Access - Operators explicitly added to the project and given Operator access. Viewer Access - All authenticated CDSW users.

Project Visibility	Access Levels for Cloudera Data Science Workbench Users
	Public Projects Created in a Team Context The following user roles will have access to public projects created in team contexts: Admin Access • Site Administrators • Project Admins • Team Admins Contributor Access • All team members with Contributor access. The team/project access rules and nuances described in the Team section apply here as well. Operator Access • All team members with Operator access. Viewer Access • All authenticated CDSW users.



Note: Restricting Access to Active Sessions

Users with Admin or Contributor-level permissions on projects have access to all of the project's active sessions and can execute commands within these active sessions. Cloudera Data Science Workbench (1.4.3 and higher) includes a feature that allows site administrators to restrict this ability by allowing only session creators to execute commands within their own active sessions. For details on how to enable this, see [Restricting Access to Active Sessions](#).

Wire Encryption

Provides information on encryption for external and internal communications.

External Communications

Cloudera Data Science Workbench uses HTTP and WebSockets (WS) to support interactive connections to the Cloudera Data Science Workbench web application. However, these connections are not secure by default.

For secure, encrypted communication, Cloudera Data Science Workbench can be configured to use a TLS termination proxy to handle incoming connection requests. The termination proxy server will decrypt incoming connection requests and forward them to the Cloudera Data Science Workbench web application.

The Cloudera Data Science Workbench documentation describes two different approaches to TLS termination: [internal and external TLS termination](#). Both provide a secure TLS connection between users and Cloudera Data Science Workbench. If you require more control over the TLS protocol and cipher suite, we recommend external termination. Both approaches require TLS certificates that list both, the Cloudera Data Science Workbench domain, as well as a wildcard for all first-level subdomains. For example, if the Cloudera Data Science Workbench domain is `cdsw.<your_domain>.com`, then the TLS certificate must include both `cdsw.<your_domain>.com` and `*.cdsw.<your_domain>.com`.

Browser Security

Cloudera Data Science Workbench also allows you to customize the HTTP headers accepted by Cloudera Data Science Workbench. The list of security headers enabled by default can be found in the documentation here: [HTTP Headers](#). Disabling these features could leave your Cloudera Data Science Workbench deployment vulnerable to clickjacking, cross-site scripting (XSS), or any other injection attacks.

Internal Communications

Internal communications between some Cloudera Data Science Workbench components are protected by mutually authenticated TLS.

The underlying Kubernetes cluster has a root Certificate Authority (CA) that is used to validate certificates for internal Kubernetes components. For details, refer the Kubernetes documentation here: [TLS certificates](#).

Cloudera Data Science Workbench Gateway Host Security

The Cloudera Data Science Workbench master host stores all the critical, stateful, persistent data for a Cloudera Data Science Workbench deployment.

This data includes your deployment secrets, such as, Kerberos credentials, encrypted passwords, SSH and API keys, and so on. While the Cloudera Data Science Workbench worker hosts do not store the same secrets, they also store sensitive information. Therefore, protecting the master and worker hosts is extremely important. Cloudera recommends the following security best practices for all the Cloudera Data Science Workbench hosts:

- Disable untrusted SSH access to the Cloudera Data Science Workbench hosts. Cloudera Data Science Workbench assumes that users only access the gateway hosts through the web application. Users with SSH access to a Cloudera Data Science Workbench host can gain full access to the cluster, including access to other users' workloads. Therefore, untrusted (non-sudo) SSH access to Cloudera Data Science Workbench hosts must be disabled to ensure a secure deployment.
- Tightly control root access via sudo.
- Uninstall or disable any other unnecessary services running on the Cloudera Data Science Workbench gateway hosts.
- Keep the hosts' operating system updated to avoid security vulnerabilities.
- Monitor user login activity on the system.

Host Mounts

Cloudera Data Science Workbench allows site administrators to expose part of the host's file system into users' engine containers at runtime. This is done using the host mounts feature. It is worth noting that directories mounted using this feature are then available to all projects across the deployment. Use this feature with great care and ensure that no sensitive information is mounted.

Base Engine Image Security

The base engine image is a Docker image that contains all the building blocks needed to launch a Cloudera Data Science Workbench session and run a workload.

It consists of kernels for Python, R, and Scala, and some common third-party libraries and packages that can be used to run common data analytics operations. Additionally, to provide a flexible, collaborative environment, data science teams can install their own preferred data science packages, just as they would on their local computers, to run workloads on data in the associated Hadoop cluster.

This base image itself is built and shipped along with Cloudera Data Science Workbench. Cloudera Data Science Workbench strives to ship a base engine image free from security vulnerabilities. Our engine images are regularly scanned for potential security vulnerabilities and we have been incorporating the recommendations from these scans into the product. For organizations that require more control over the contents of the engines used by Cloudera Data Science Workbench users, we recommend building your own [customized engine images](#).