Machine Learning 1.5.1

Securing Cloudera Machine Learning

Date published: 2020-07-16
Date modified: 2023-01-31

https://docs.cloudera.com/

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

Configuring External Authentication with LDAP and SAML

Important: Cloudera recommends you leverage Single Sign-On for users via the CDP Management Console. For instructions on how to configure this, see Configuring LDAP authentication for CDP Private Cloud. If you cannot do this, we recommend contacting Cloudera Support before attempting to use the LDAP or SAML instructions provided in this section.

Cloudera Machine Learning supports user authentication against its internal local database, and against external services such as Active Directory, OpenLDAP-compatible directory services, and SAML 2.0 Identity Providers. By default, Cloudera Machine Learning performs user authentication against its internal local database. This topic describes the signup process for the first user, how to configure authentication using LDAP, Active Directory or SAML 2.0, and an optional workaround that allows site administrators to bypass external authentication by logging in using the local database in case of misconfiguration.

Configuring SAML Authentication

This topic describes how to set up SAML for Single Sign-on authentication for a workspace.

Important: This is not the recommended method to set up SSO. Cloudera recommends you use the CDP management console to set this up: Configuring LDAP authentication for CDP Private Cloud.

Cloudera Machine Learning supports the Security Assertion Markup Language (SAML) for Single Sign-on (SSO) authentication; in particular, between an identity provider (IDP) and a service provider (SP). The SAML specification defines three roles: the principal (typically a user), the IDP, and the SP. In the use case addressed by SAML, the principal (user agent) requests a service from the service provider. The service provider requests and obtains an identity assertion from the IDP. On the basis of this assertion, the SP can make an access control decision—in other words it can decide whether to perform some service for the connected principal.

Note: The user sync feature only works with the SAML IDP provided by the control plane. If a custom SAML IDP is provided then customer has to make sure to turn usersync off. Otherwise, there is a risk that users will be deactivated and therefore causing cron jobs scheduled by users that are deactivated to fail.

The primary SAML use case is called web browser single sign-on (SSO). A user with a user agent (usually a web browser) requests a web resource protected by a SAML SP. The SP, wanting to know the identity of the requesting user, issues an authentication request to a SAML IDP through the user agent. In the context of this terminology, Cloudera Machine Learning operates as a SP.

Cloudera Machine Learning supports both SP- and IDP-initiated SAML 2.0-based SSO. Its Assertion Consumer Service (ACS) API endpoint is for consuming assertions received from the Identity Provider. If your Cloudera Machine Learning domain root were cdsw.*company*.com, then this endpoint would be available at http://cdsw.*company*.com/api/v1/saml/acs. SAML 2.0 metadata is available at http://cdsw.*company*.com/api/v1/saml/metadata for IDP-initiated SSO. Cloudera Machine Learning uses HTTP Redirect Binding for authentication requests and expects to receive responses from HTTP POST Binding.

When Cloudera Machine Learning receives the SAML responses from the Identity Provider, it expects to see at least the following user attributes in the SAML responses:

- The unique identifier or username. Valid attributes are:
  - uid
  - urn:oid:0.9.2342.19200300.100.1.1
- The email address. Valid attributes are:
  - mail
  - email
  - urn:oid:0.9.2342.19200300.100.1.3

- The common name or full name of the user. Valid attributes are:
  - cn
  - urn:oid:2.5.4.3

  In the absence of the cn attribute, Cloudera Machine Learning will attempt to use the following user attributes, if they exist, as the full name of the user:
  - The first name of the user. Valid attributes are:
    - givenName
    - urn:oid:2.5.4.42
  - The last name of the user. Valid attributes are:
    - sn
    - urn:oid:2.5.4.4

Configuration Options

List of properties to configure SAML authentication and authorization in Cloudera Machine Learning.

Cloudera Machine Learning Settings

- Entity ID: Required. A globally unique name for Cloudera Machine Learning as a Service Provider. This is typically the URI.
- NameID Format: Optional. The name identifier format for both Cloudera Machine Learning and Identity Provider to communicate with each other regarding a user. Default: urn:oasis:names:tc:SAML:1.1:nameid-format:emailAddress.
- Authentication Context: Optional. SAML authentication context classes are URIs that specify authentication methods used in SAML authentication requests and authentication statements. Default: urn:oasis:names:tc:SAML:2.0:ac:classes:PasswordProtectedTransport.

Signing SAML Authentication Requests

- CDSW Private Key for Signing Authentication Requests: Optional. If you upload a private key, you must upload a corresponding certificate as well so that the Identity Provider can use the certificate to verify the authentication requests sent by Cloudera Machine Learning. You can upload the private key used for both signing authentication requests sent to Identity Provider and decrypting assertions received from the Identity Provider.
- CML Certificate for Signature Validation: Required if the Cloudera Machine Learning Private Key is set, otherwise optional. You can upload a certificate in the PEM format for the Identity Provider to verify the authenticity of the authentication requests generated by Cloudera Machine Learning. The uploaded certificate is made available at the http://cdsw.*company*.com/api/v1/saml/metadata endpoint.

SAML Assertion Decryption

Cloudera Machine Learning uses the following properties to support SAML assertion encryption & decryption.

- CML Certificate for Encrypting SAML Assertions - Must be configured on the Identity Provider so that Identity Provider can use it for encrypting SAML assertions for Cloudera Machine Learning
- CML Private Key for Decrypting SAML Assertions - Used to decrypt the encrypted SAML assertions.

Identity Provider

- Identity Provider SSO URL: Required. The entry point of the Identity Provider in the form of URI.
- Identity Provider Signing Certificate: Optional. Administrators can upload the X.509 certificate of the Identity Provider for Cloudera Machine Learning to validate the incoming SAML responses.

  Cloudera Machine Learning extracts the Identity Provider SSO URL and Identity Provider Signing Certificate information from the uploaded Identity Provider Metadata file. Cloudera Machine Learning also expects all

Identity Provider metadata to be defined in a <md:EntityDescriptor> XML element with the namespace "urn:oasis:names:tc:SAML:2.0:metadata", as defined in the SAMLMeta-xsd schema.

For on-premises deployments, you must provide a certificate and private key, generated and signed with your trusted Certificate Authority, for Cloudera Machine Learning to establish secure communication with the Identity Provider.

Authorization

When you're using SAML 2.0 authentication, you can use the following properties to restrict the access to Cloudera Machine Learning to certain groups of users:

- SAML Attribute Identifier for User Role: The Object Identifier (OID) of the user attribute that will be provided by your identity provider for identifying a user's role/affiliation. You can use this field in combination with the following SAML User Groups property to restrict access to Cloudera Machine Learning to only members of certain groups.

  For example, if your identity provider returns the OrganizationalUnitName user attribute, you would specify the OID of the OrganizationalUnitName, which is urn:oid:2.5.4.11, as the value for this property.
- SAML User Groups: A list of groups whose users have access to Cloudera Machine Learning. When this property is set, only users that are successfully authenticated AND are affiliated to at least one of the groups listed here, will be able to access Cloudera Machine Learning.

  For example, if your identity provider returns the OrganizationalUnitName user attribute, add the value of this attribute to the SAML User Groups list to restrict access to Cloudera Machine Learning to that group.

  If this property is left empty, all users that can successfully authenticate themselves will be able to access Cloudera Machine Learning.
- SAML Full Administrator Groups: A list of groups whose users are automatically granted the site administrator role on Cloudera Machine Learning.

  The groups listed under SAML Full Administrator Groups do not need to be listed again under the SAML User Groups property.

Configuring HTTP Headers for Cloudera Machine Learning

This topic explains how to customize the HTTP headers that are accepted by Cloudera Machine Learning.

Required Role: Site Administrator

These properties are available under the site administrator panel at Admin Security .

Important: Any changes to the following properties require a full restart of Cloudera Machine Learning. To do so, run cdsw restart on the master host.

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 Machine Learning, 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 Machine Learning 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 Machine Learning API through browsers.

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 Machine Learning 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 Machine Learning will inform your browser that it should never load the site using HTTP. Additionally, all attempts to access Cloudera Machine Learning 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 Machine Learning. Then, load the Cloudera Machine Learning web application in each browser to clear the respective browser's HSTS setting. Finally, disable TLS/SSL across the cluster. Following this sequence should help avoid a situation where users get locked out of their accounts due to 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 Machine Learning 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 Machine Learning will inform your browser that it should never load the site using HTTP. Additionally, all attempts to access Cloudera Machine Learning 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 Machine Learning. Then, load the Cloudera Machine Learning web application in each browser to clear the respective browser's HSTS setting. Finally, disable TLS/SSL across the cluster. Following this sequence should help 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 Machine Learning, 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 Machine Learning 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 Machine Learning API through browsers.

SSH Keys

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

Personal Key

Cloudera Machine Learning 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 #unique_12.

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 Machine Learning users, each Cloudera Machine Learning 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 Machine Learning 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 Machine Learning.
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 Machine Learning 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 Machine Learning 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 Machine Learning 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.

Autoscaling Workloads with Kubernetes

Autoscaling on Private Cloud

CML on Private Cloud supports application autoscaling on multiple fronts. Additional compute resources are utilized when users self-provision sessions, run jobs, and utilize other compute capabilities. Within a session, users can also leverage the worker API to launch resources necessary to host TensorFlow, PyTorch, or other distributed applications. Spark on Kubernetes scales up to any number of executors as requested by the user at runtime.

Restricting User-Controlled Kubernetes Pods

Cloudera Machine Learning includes three properties that allow you to control the permissions granted to user-controlled Kubernetes pods.

Required Role: Site Administrator

An example of a user-controlled pod is the engine pod, which provides the environment for sessions, jobs, etc. These pods are launched in a per-user Kubernetes namespace. Since the user has the ability to launch arbitrary pods, these settings restrict what those pods can do.

They are available under the site administrator panel at Admin Security under the Control of User-Created Kubernetes Pods section.

Do not modify these settings unless you need to run pods that require special privileges. Enabling any of these properties puts CML user data at risk.

Allow privileged pod containers

Pod containers that are "privileged" are extraordinarily powerful. Processes within such containers get almost the same privileges that are available to processes outside the container.

If this property is enabled, a privileged container could potentially access all data on the host.

This property is disabled by default .

Allow pod containers to mount unsupported volume types

The volumes that can be mounted inside a container in a Kubernetes pod are already heavily restricted. Access is normally denied to volume types that are unfamiliar, such as GlusterFS, Cinder, Fibre Channel, etc. If this property is enabled, pods will be able to mount all unsupported volume types.

This property is disabled by default .

Hadoop Authentication for ML Workspaces

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

About this task

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

Procedure

1. Navigate to your ML workspace.
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 Machine Learning uses your stored credentials to ensure you are secure when running workloads.

CML and outbound network access

Cloudera Machine Learning expects access to certain external networks. See the related information *Configuring proxy hosts for CML workspace connections* for further information.

Note: The outbound network access destinations listed in *Configuring proxy hosts for CML workspace connections* are only the minimal set required for CDP 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 CML workspace connections