

## Setting Up Data Lake Access

Date published: 2020-07-16

Date modified: 2025-05-29

# CLOUDERA

# Legal Notice

© Cloudera Inc. 2025. 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

|                                    |          |
|------------------------------------|----------|
| <b>Setup Data Lake Access.....</b> | <b>4</b> |
|------------------------------------|----------|

# Setup Data Lake Access

Cloudera AI can access data tables stored in an AWS or Microsoft Azure Data Lake. As a Cloudera AI Admin, follow this procedure to set up the necessary permissions.

## About this task

The instructions apply to Data Lakes on both AWS and Microsoft Azure. Follow the instructions that apply to your environment.

## Procedure

### 1. Cloud Provider Setup

Make sure the prerequisites for AWS or Azure are satisfied (see the Related Topics for AWS environments and Azure environments). Then, create a Cloudera environment as follows.

- a) For environment logs, create an S3 bucket or ADLS Gen2 container.



**Note:** For ADLS Gen2, create a dedicated container for logs, and a dedicated container for data, within the same account.

- b) For environment storage, create an S3 bucket or ADLS Gen2 container.
- c) For AWS, create AWS policies for each S3 bucket, and create IAM roles (simple and instance profiles) for these policies.
- d) For Azure, create managed identities for each of the personas, and create roles to map the identities to the ADLS permissions.

For detailed information on S3 or ADLS, see Related information.

## 2. Environment Setup

In Cloudera, set up paths for logs and native data access to the S3 bucket or ADLS Gen2 container.

In the Environment Creation wizard, set the following:

The screenshot shows the 'Logs Storage and Audits' configuration page in the Cloudera Environment Setup wizard. It includes three main configuration sections, each with a dropdown menu and a help icon (question mark):

- Instance Profile:** A dropdown menu showing 'MLX\_DEV\_DATA LAKE\_LOG\_ROLE'. Above it is the instruction 'Provide an existing location where log files will be stored.' and a link 'Click here to refresh instance profiles from the cloud provider.'
- Logs Location Base:** A dropdown menu showing 's3a://fooenv/logs'. Above it is the instruction 'Select an Instance Profile\*'.
- Ranger Audit Role:** A dropdown menu showing 'arn:aws:iam::886883559913:role/mlx-dev-prod-env\_RANGER\_AUD'.

### a) Logs Storage and Audits


1. Instance Profile - The IAM role or Azure identity that is attached to the master node of the Data Lake cluster. The Instance Profile enables unauthenticated access to the S3 bucket or ADLS container for logs.
2. Logs Location Base - The location in S3 or ADLS where environment logs are saved.



**Note:** The instance profile or Azure identity must refer to the same logs location base in S3 or ADLS.

3. Ranger Audit Role - The IAM role or Azure identity that has S3 or ADLS access to write Ranger audit events. Ranger uses Hadoop authentication, therefore it uses IDBroker to access the S3 bucket or ADLS container, rather than using Instance profiles or Azure identities directly.

### b) Data Access




## Data Access


Provide an existing location where workload data will be stored.

Select an Instance Profile\*


[Click here](#) to refresh instance profiles from the cloud provider.



Storage Location Base\*



Data Access Role\*



ID Broker Mappings

You may optionally provide mappings for users or groups.

|                |   |           |   |                                       |
|----------------|---|-----------|---|---------------------------------------|
| User or Group: | <input type="text" value="ml-data-scientists"/> | Role Arn: | <input type="text" value="arn:aws:iam::886883559913:ro"/> | <input type="button" value="Remove"/> |
| User or Group: | <input type="text" value="ml-data-engineers"/>  | Role Arn: | <input type="text" value="arn:aws:iam::886883559913:ro"/> | <input type="button" value="Remove"/> |
| User or Group: | <input type="text" value="ml-dl-admins"/>       | Role Arn: | <input type="text" value="arn:aws:iam::886883559913:ro"/> | <input type="button" value="Remove"/> |

1. Instance Profile - The IAM role or Azure identity that is attached to the IDBroker node of the Data Lake cluster. IDBroker uses this profile to assume roles on behalf of users and get temporary credentials to access S3 buckets or ADLS containers.
2. Storage Location Base - The S3 or ADLS location where data pertaining to the environment is saved.
3. Data Access Role - The IAM role or Azure identity that has access to read or write environment data. For example, Hive creates external tables by default in the Cloudera environments, where metadata is stored in HMS running in the Data Lake. The data itself is stored in S3 or ADLS. As Hive uses Hadoop authentication, it uses IDBroker to access S3 or ADLS, rather than using Instance profiles or Azure identities. Hive uses the data access role for storage access.



**Note:** The data access role must have permissions to access the S3 or ADLS storage location.

4. ID Broker Mappings - These specify the mappings between the Cloudera user or groups to the AWS IAM roles or Azure roles that have appropriate S3 or ADLS access. This setting enables IDBroker to get appropriate S3 or ADLS credentials for the users based on the role mappings defined.

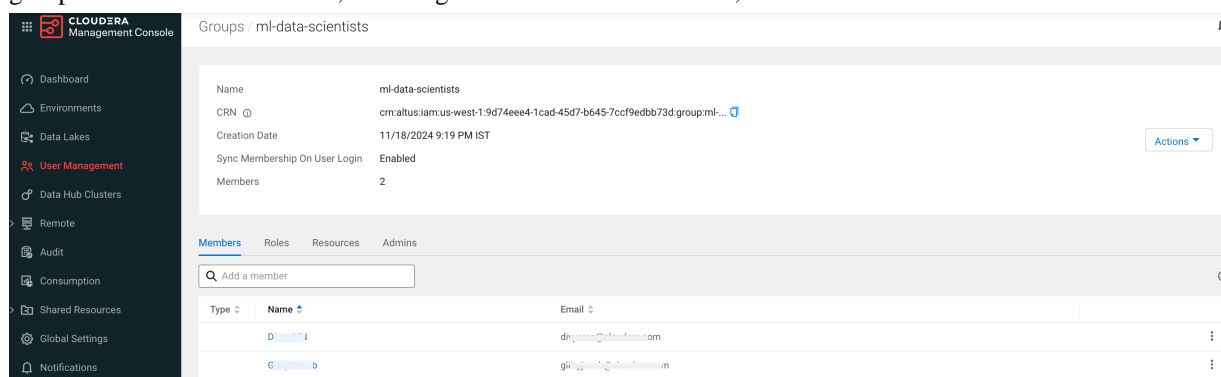


**Note:** There is no limit to the number of mappings that one can define but each user can only be assigned to one of the role mappings.

This completes installation of the environment.

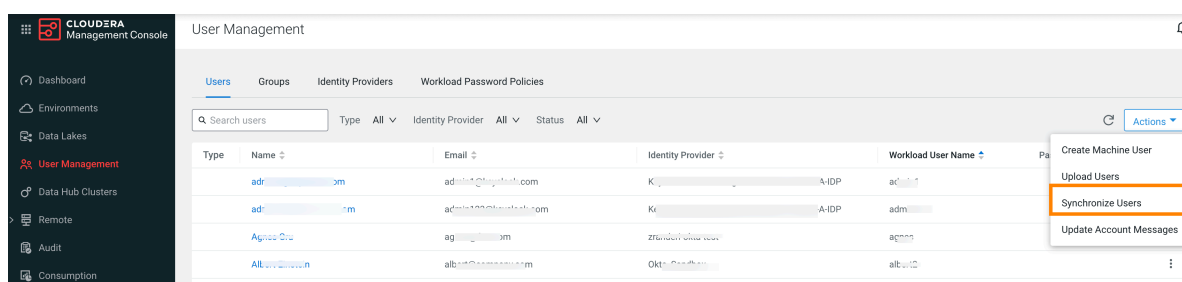
### 3. User Group Mappings

In Cloudera, you can assign users to groups to simplify permissions management. For example, you could create a group called ml-data-scientists, and assign two individual users to it, as shown here. .



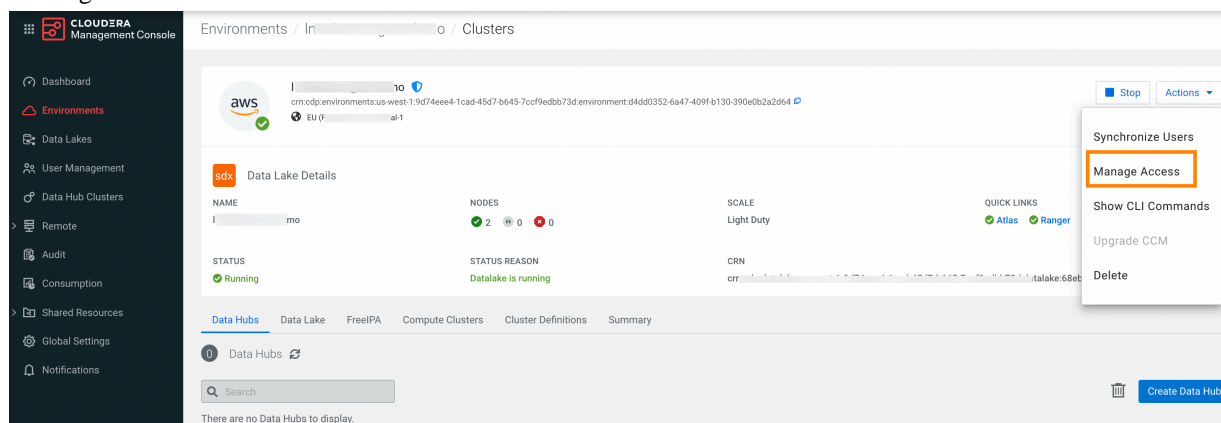
#### a. Sync users

Whenever you make changes to user and group mappings, make sure to sync the mappings with the authentication layer. In User Management > Actions, click Sync Users, and select the environment.



### 4. IDBroker

IDBroker allows an authenticated and authorized user to exchange a set of credentials or a token for cloud vendor access tokens. You can also view and update the IDBroker mappings at this location. IDBroker mappings can be accessed through Environments > Manage Access. Click on the IDBroker Mappings tab. Click **Edit** to edit or add mappings. When finished, synchronize the mappings to push the settings from Cloudera to the IDBroker instance running inside the Data Lake of the environment.



At this point, Cloudera resources can access the AWS S3 buckets or Azure ADLS storage.

## 5. Ranger

To get admin access to Ranger, users need the EnvironmentAdmin role, and that role must be synced with the environment.

- a. Click Environments > Env > Actions > Manage Access > Add User
- b. Select EnvironmentAdmin resource role.
- c. Click Update Roles
- d. On the Environments page for the environment, in Actions, select Synchronize Users to FreeIPA.

The permissions are now synchronized to the Data Lake, and you have admin access to Ranger.

Update permissions in Ranger

- a. In Environments > Env > Data Lake Cluster, click Ranger.
- b. Select the Hadoop SQL service, and check that the users and groups have sufficient permissions to access databases, tables, columns, and urls.

For example, a user can be part of these policies:

- all - database,table,column
- all - url

This completes all configuration needed for Cloudera AI to communicate with the Data Lake.

## 6. Cloudera AI User Setup

Now, Cloudera AI is able to communicate with the Data Lake. There are two steps to get the user ready to work.

- a. In Environments > Environment name > Actions > Manage Access > Add user, the Admin selects MLUser resource role for the user.
- b. The User logs into the workbench in Cloudera AI Workbench > Workbench name, click Launch Workbench.

The user can now access the workbench.

### Related Information

[AWS environments](#)

[Azure environments](#)