

Management Console Overview

Date published: 2019-08-22

Date modified:

CLOUDEXERA

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

Management Console.....	4
Interfaces.....	4
Core concepts.....	5
Environments.....	5
Credentials.....	6
Data Lakes.....	7
Shared resources.....	8
Classic clusters.....	8
CDP accounts.....	8

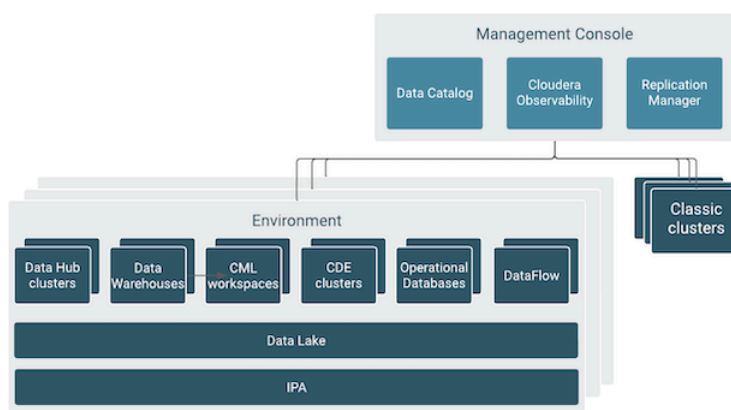
Management Console

The Management Console is an administrative service used by CDP administrators to manage environments, users, and CDP services.

The Management Console allows you to:

- Configure SSO, manage users, and decide who can access which resources.
- Register your existing on-prem Cloudera Distribution of Hadoop (CDH) or Hortonworks Data Platform (HDP) clusters in order to burst your data to the cloud. In the CDP world, these clusters are called classic clusters.
- Register your cloud provider environments (such as your AWS, Azure, or GCP account) in CDP and then launch Data Hub clusters, Data Warehouses, Machine Learning workspaces, Data Engineering and DataFlow clusters, and Operational Databases within these environments and determine which users have access to which resources. These clusters are attached to a Data Lake that runs within the environment and provides security and governance for all attached clusters.
- Utilize services such as Data Catalog, Cloudera Observability, and Replication manager:
 - Data Catalog - A centralized management tool for searching, organizing, securing, and governing data across environments.
 - Cloudera Observability - A centralized management tool for analyzing and optimizing workloads within and across environments.
 - Replication Manager - A centralized management tool for replicating and migrating data and the associated metadata across environments.

The following diagram illustrated the functionality described above:



Related Information

[Interfaces](#)

[Core concepts](#)

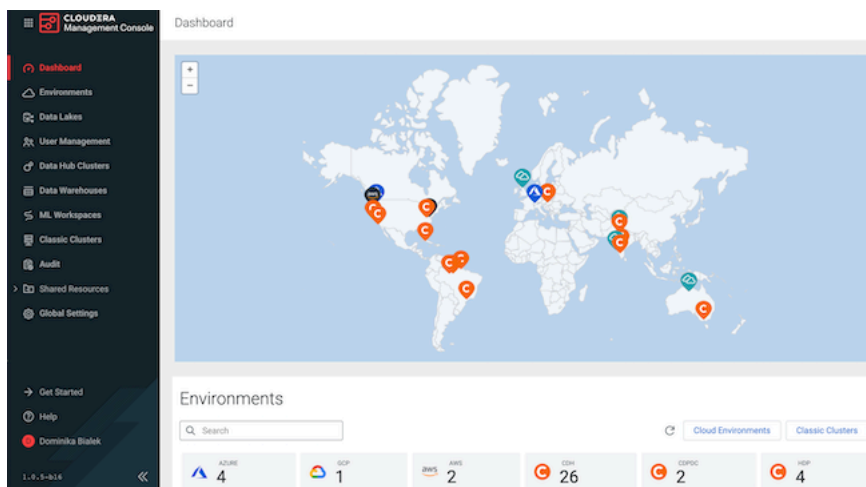
[CDP accounts](#)


Interfaces

There are three basic ways to access and use CDP services: web interface, CLI client, and SDK.

CDP web interface

The CDP web interface provides a web-based, graphical user interface accessible via a browser. As an admin user, you can use the CDP web interface to register environments, manage users, and provision CDP service resources for end users. As an end user, you can use the CDP web interface to access CDP service web interfaces to perform data engineering or data analytics tasks.



The CDP web interface allows you to access multiple CDP services. You can switch between different CDP services by using the  icon in the top left corner.

CDP CLI

If you prefer to work in a terminal window, you can download and configure the CDP client that gives you access to the CDP CLI tool. The CDP CLI allows you to perform the same actions as can be performed from the management console. Furthermore, it allows you to automate routine tasks such as Data Hub cluster creation.

CDP SDK

You can use the CDP SDK for Java to integrate CDP services with your applications. Use the CDP SDK to connect to CDP services, create and manage clusters, and run jobs from your Java application or other data integration tools that you may use in your organization.

Core concepts

The following concepts are key to understanding the Management Console service and CDP in general:

Environments

In CDP, an environment is a logical subset of your cloud provider account including a specific virtual network. You can register as many environments as you require.

In the on premise world, your clusters run on machines in your data center and are accessible on your corporate network. In contrast, when you launch clusters in the cloud, your cloud provider, such as AWS, Azure, or Google Cloud, provides all the infrastructure including private networks and VMs. You provision a private network in a selected region, configure access settings, and then provision resources into that private network. All these resources are accessible as part of your cloud provider account and can be shared by users within your organization.

The “environment” concept of CDP is closely related to the private network in your cloud provider account. Registering an environment provides CDP with access to your cloud provider account and identifies the resources in your cloud provider account that CDP services can access or provision. Once you’ve registered an environment in CDP, you can start provisioning CDP resources such as clusters, which run on the physical infrastructure in an AWS, Azure, or Google Cloud data center.

You may want to register multiple environments corresponding to different regions that your organization would like to use.

For more information about environments, refer to cloud provider specific documentation linked below.

Related Information

[Introduction to AWS environments](#)

[Introduction to Azure environments](#)

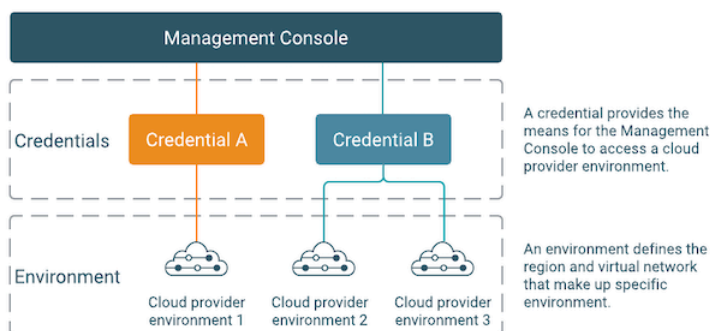
[Introduction to Google Cloud environments](#)

Credentials

A credential allows CDP to authenticate with your cloud provider account and obtain authorization to provision cloud provider resources on your behalf.

The authentication and authorization process varies depending on the cloud provider, but is typically done by assigning a specific role (with a specific set of permissions) that can be assumed by CDP, allowing it to perform certain actions within your cloud provider account.

A credential is a core component of an environment, providing access to the region and virtual network that make up the environment and allowing CDP to provision resources within that environment. Credentials are managed separately from environments, because you can reuse the same credential across multiple environments if needed. For example, the following diagram presents a scenario where one credential (credential A) is used by a single environment (cloud provider environment 1) but another credential (credential B) is used by multiple environments (cloud provider environment 2 and 3). In this case, it is implied that cloud provider environment 2 and 3 must represent the same AWS, Azure, Google Cloud account, but may correspond to different regions and/or VPCs/subnets.



Related Information

[Role-based credential on AWS](#)

[App-based credential on Azure](#)

[Provisioning credential for Google Cloud](#)

Data Lakes

In CDP, a Data Lake is a service for creating a protective ring of security and governance around your data, whether the data is stored in cloud object storage or HDFS.

When you register an environment in CDP, a Data Lake is automatically deployed for that environment. The Data Lake runs in the virtual network of the environment and provides security and governance layer for the environment's workload resources, such as Data Hub clusters.



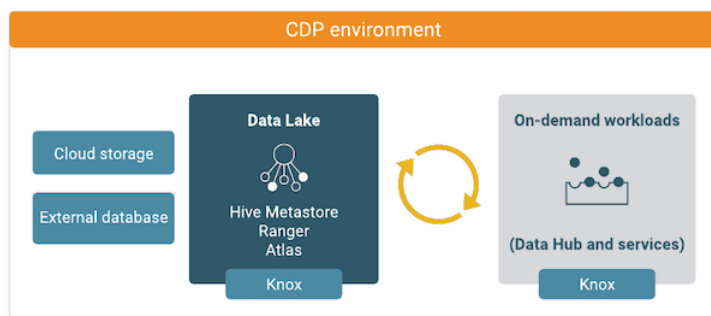
Note: Currently, the Data Lake automatically deployed is scaled for light duty. In the future, you'll have more options for scaling your Data Lake.

The Data Lake provides a way for you to create, apply, and enforce user authentication and authorization and to collect audit and lineage metadata from across multiple ephemeral workload clusters. When you start a workload cluster in the context of a CDP environment, the workload cluster is automatically "attached" with the security and governance infrastructure of the Data Lake. "Attaching" your workload resources to the Data Lake instance allows the attached cluster workloads to access data and run in the security context provided by the Data Lake.

A Data Lake cluster includes Apache Knox. Knox provides a protected gateway for access to Data Lake UIs. Knox is also installed on all workload clusters, providing a protected gateway for access to cluster UIs.

While workloads can be short-lived, the security policies around your data schema are long-running and shared for all workloads. The Data Lake instance provides consistent and available security policy definitions that are available for current and future ephemeral workloads. All information related to metadata, policies, and audits is stored on external locations (external databases and cloud storage).

The Data Lake stores its metadata, policies, and audits in external databases and cloud storage, reducing the resource footprint on the cluster.



The following technologies provide capabilities for the Data Lake:

Component	Technology	Description
Schema	Apache Hive Metastore	Provides Hive schema (tables, views, and so on). If you have two or more workloads accessing the same Hive data, you need to share schema across these workloads.
Authorization Policies	Apache Ranger	Defines security policies around Hive schema. If you have two or more users accessing the same data, you need security policies to be consistently available and enforced.
Audit Tracking	Apache Ranger	Audits user access and captures data access activity for the workloads.
Governance	Apache Atlas	Provides metadata management and governance capabilities.
Security Gateway	Apache Knox	Supports a single workload endpoint that can be protected with SSL and enabled for authentication to access to resources.

Related Information

[Data lakes](#)

Shared resources

CDP allows you to manage and reuse certain resources across environments.

Shared resources include:

Resource	Description
Credentials	A credential allows your CDP environment to authenticate with your cloud provider account and to obtain authorization to provision cloud provider resources for Data Hub clusters and other resources provisioned via CDP. There is one credential per environment.
Cluster definitions	A cluster definition defines cloud provider specific Data Hub cluster settings such as instance types, storage, and so on. Cluster definitions are specified in JSON format. Each Data Hub cluster is based on a single cluster definition. Each cluster definition references a single cluster template.
Cluster templates	A cluster template is a blueprint that defines cluster topology, including the number of host groups and all components and sub-components installed on each host group. Cluster templates are specified in JSON format. Each Data Hub cluster is based on a single cluster template.
Image catalogs	By default, Data Hub includes an image catalog with default images. If necessary, you can customize a default image and register it as part of a custom image catalog. These images are used for creating Data Hub clusters.
Recipes	A recipe is a custom script that can be uploaded and used for running a specific task on Data Hub or Data Lake clusters. You can upload and select multiple recipes to have them executed on a Data Hub or Data Lake cluster at a specific time.

Shared resources only need to be registered once and then can be reused for multiple environments. Since shared resources are not attached to a specific workload cluster, their lifespan is not limited to the lifespan of the cluster.

Related Information

[Credentials](#)

[Cluster definitions](#)

[Cluster templates](#)

[Recipes](#)

Classic clusters

Classic clusters are on-prem Cloudera Distribution of Hadoop (CDH) or CDP Data Center (CDP-DC) clusters registered in CDP.

You can register your existing CDH or CDP-DC classic clusters in order to burst or migrate a workload to their public cloud environment by replicating the data and creating a Data Hub cluster to host the workload.

Related Information

[Classic clusters](#)

CDP accounts

A CDP account (sometimes called a CDP tenant) is the management console where users log in to access their services. A CDP account contains one or more CDP environments.

CDP Accounts

Physically, all management consoles run in a shared infrastructure (the Cloudera Control Plane). Each management console has a unique ID and is logically isolated from other CDP accounts. For example, if you have two CDP accounts, you cannot view the resources in account 2 if you are logged into account 1. Each CDP account, including sub-accounts, has a separate invoice.

A CDP account can have one or more identity providers. Each identity provider is connected to one SAML provider, which in turn is connected to an Active Directory. A CDP account can be connected to multiple Azure Active Directory organizations (via identity provider) and multiple subscriptions. Cloudera creates one CDP account per customer; you can create additional CDP accounts through a manual approvals process.

For example, the "Cloudera CDP Demos" account is connected to:

- 3 identity providers, each connected to an LDAP Server via SAML:
 - Corporate OKTA, so that anybody with a cloudera.com email can log in (as long as they are entitled).
 - An LDAP server used for workshops, so that we can provision temporary users without having to go through corporate account lifecycle management.
 - An LDAP server hosting demo users, so that we can create new personas on demand for demo scenarios.
- Multiple Azure subscriptions; one per budget item (Sales, Product Management, Services, Tradeshow Demos etc.).
- This allows different groups to host their own demos and pay for them using their own cloud account.

CDP Environments

A CDP environment exists inside a CDP account, and each CDP account can have many environments. An environment is a logical container for a Data Lake cluster and any workloads that are attached to that cluster, such as Data Hubs, Data Warehouses, Machine Learning environments, etc. All services running within an environment share the same metadata and use the same Data Lake.

Each environment can be associated with only one cloud account. This association is called a CDP credential and points to either:

- A service principal for a specific Azure subscription
- A cross-account role for a specific AWS account

