

## Cloudera Data Services on premises Overview

Date published: 2023-12-16

Date modified: 2025-06-06

# 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

About Cloudera Data Services on premises..... 4

Benefits.....4

Deployment and high-level architecture..... 5

Glossary.....6

## About Cloudera Data Services on premises

Cloudera Data Services on premises works on top of Cloudera Base on premises and is the on-premises offering of Cloudera that brings many of the benefits of the public cloud deployments to Cloudera on premises. Cloudera Private Cloud Data Services lets you deploy and use the Cloudera Data Warehouse, Cloudera AI, and Cloudera Data Engineering Data Services.

Building on Apache Spark, Cloudera Data Engineering is an all-inclusive data engineering toolset that enables orchestration automation with Apache Airflow, advanced pipeline monitoring, visual troubleshooting, and comprehensive management tools to streamline ETL processes across enterprise analytics teams. Cloudera Data Engineering powers consistent, repeatable, and automated data engineering workflows.

Cloudera Data Warehouse enables you to create highly-performant, independent, self-service data warehouses for teams of business analysts. Cloudera's support for an open data lakehouse, centered on Cloudera Data Warehouse, brings high-performance, self-service reporting and analytics to your business – simplifying data management for data practitioners and administrators.

Cloudera AI, Cloudera's platform for machine learning and AI unifies self-service data science and data engineering in a single, portable service for multi-function analytics on data. It optimizes ML workflows across your business with native and robust tools for deploying, serving, and monitoring models.

Replication Manager enables you to copy and migrate HDFS data, Hive external tables, and Ozone data between Cloudera Base on premises 7.1.8 or higher clusters using Cloudera Manager version 7.7.3 or higher.

Cloudera Data Catalog is a service within Cloudera Data Platform that enables you to understand, manage, secure, and govern data assets across the enterprise. Data Catalog helps you discern data lying in your Data Lake (Base cluster).

Providing a consistent user experience to data practitioners and IT admins across Cloudera on cloud and Cloudera on premises, these data services open up avenues for a hybrid data lifecycle.

## Benefits of Cloudera Data Services on premises

Cloudera Data Services on premises disaggregates compute and storage, which allows independent scaling of compute and storage clusters. The Data Services provide containerized analytic applications that scale dynamically and can be upgraded independently. Through the use of containers deployed on Kubernetes, Cloudera Data Services on premises brings both agility and predictable performance to analytic applications. Cloudera Data Services on premises inherits unified security, governance, and metadata management through Cloudera Shared Data Experience (SDX) from the Cloudera Private Cloud Base cluster.

The three main benefits of Cloudera Data Services on premises are simplified multitenancy and isolation, simplified deployment of applications, and better utilization of infrastructure.

### Simplified multitenancy and isolation

The containerized deployment of applications in Cloudera on premises ensures that each application is sufficiently isolated and can run independently from others on the same Kubernetes infrastructure, in order to eliminate resource contention. Such a deployment also helps in independently upgrading applications based on your requirements. In addition, all these applications can share a common Data Lake instance.

### Simplified deployment of applications

Cloudera on premises ensures a much faster deployment of applications with a shared Data Lake compared to monolithic clusters where separate copies of security and governance data would be required for each separate application. In situations where you need to provision applications on an arbitrary basis, for example, to deploy test

applications or to allow for self-service, transient workloads without administrative or operations overhead, Cloudera on premises enables you to rapidly perform such deployments.

### Better utilization of infrastructure

Similar to Cloudera on cloud, Cloudera on premises enables you to provision resources in real time when deploying applications. In addition, the ability to scale or suspend applications on a need basis in Cloudera on premises ensures that your on-premises infrastructure is utilized optimally.

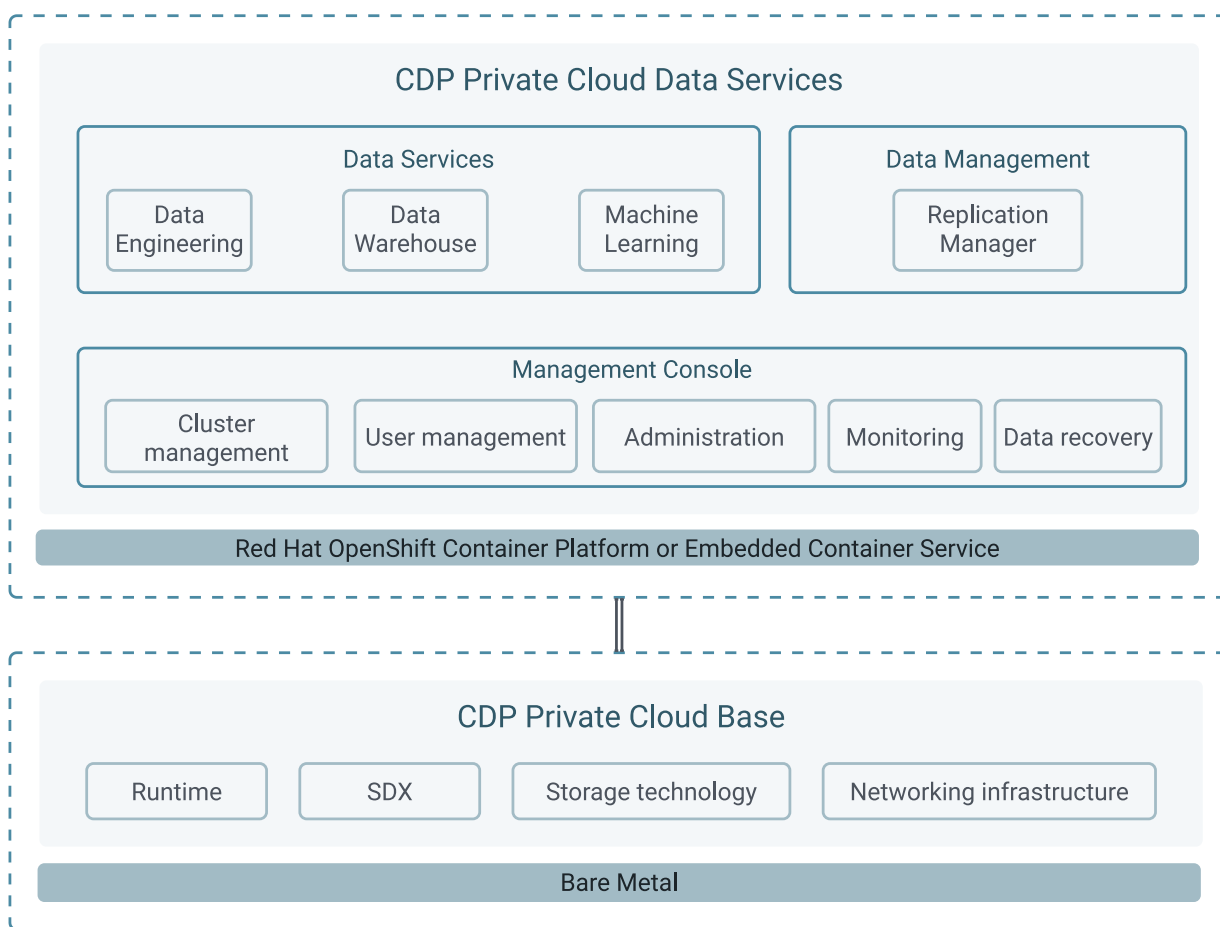
## Deployment and high-level architecture

A Cloudera Data Services on premises deployment includes an Environment, a Data Lake, the Cloudera Management Console, and Data Services (Data Warehouse, Cloudera AI, Data Engineering). Other tools and utilities include Replication Manager, Data Recovery Service, CDP CLI, and monitoring using Grafana.

To deploy Cloudera Data Services on premises you need a Cloudera Base on premises cluster, along with container-based clusters that run the Data Services. You can either use a dedicated RedHat OpenShift container cluster or deploy an Cloudera Embedded Container Service container cluster.

The on premises deployment process involves configuring Cloudera Management Console, registering an environment by providing details of the Data Lake configured on the Base cluster, and then creating the workloads.

Platform Managers and Administrators can rapidly provision and deploy the data services through the Cloudera Management Console, and easily scale them up or down as required.



You can install Cloudera Base on premises on virtual machines or bare-metal hardware. Cloudera Base on premises provides the following components and services that are used by Cloudera Data Services on premises:

- SDX Data Lake cluster for security, metadata, and governance
- HDFS and Ozone for storage
- Powerful and open-source Cloudera Runtime services such as Ranger, Atlas, Hive Metastore (HMS), and so on
- Networking infrastructure that supports network traffic between storage and compute environments

## Cloudera Data Services on premises glossary

The Cloudera Data Services on premises documentation uses terms related to enterprise data cloud and cloud computing.

### **CDP CLI:**

A command-line interface to access and manage Cloudera services and resources.

### **Cloudera Base on premises:**

An on-premises version of Cloudera. It combines the best of Cloudera Enterprise Data Hub (CDH) and Hortonworks Data Platform Enterprise (HDP) along with new features and enhancements across the stack.

### **Cloudera Manager for Data Services:**

A tool to install, manage, monitor, and configure your clusters and services using the Cloudera Manager Admin Console web application or the Cloudera Manager API.

### **Cloudera Runtime:**

The core open source software distribution within Cloudera Base on premises. It includes approximately 50 open-source projects providing data management tools within Cloudera.

### **Cloudera Control Plane service:**

Cloudera service that includes services like Cloudera Management Console, Replication Manager, Data Recovery Service, Service Discovery Service, and so on. These services interact with your environment on Cloudera Embedded Container Service or OCP to provision and manage compute infrastructure that you can use to manage the lifecycle of data stored on HDFS or Ozone.

### **Cloudera Data Catalog:**

This data service allows the Data Stewards, Business analysts, and Data Administrators to curate different datasets by adding multiple assets in it.

Users can scan and profile information for a data lake to gather metadata about schema and identify sensitive information. The metadata is gathered by the profilers and stored in Cloudera Apache Atlas to be retrieved during search and discovery operations. Users are allowed to define and tag assets using custom rules to identify and classify assets for their business requirements and discover metadata in **Asset details** page about a given asset and also tag and group various assets to create a dataset. Users can search and manage datasets based on various attributes.

### **Cloudera Data Engineering:**

This data service allows you to create, manage, and schedule Apache Spark jobs without the overhead of creating and maintaining Spark clusters. You can define virtual clusters with a range of CPU and memory resources, and the cluster scales up and down as needed to execute your Spark workloads, helping to control your cloud costs.

### **Data Lake:**

Creates a protective ring of security and governance around your data. For a Cloudera on premises deployment, the Data Lake services are hosted on the Cloudera Base on premises cluster. In addition, the Data Lake services are shared between multiple workloads.

### **Data Service:**

A defined subset of Cloudera functionality that enables a Cloudera user to solve a specific problem related to their data lake (process, analyze, predict, and so on). Example services: Cloudera Data Engineering, Cloudera Data Warehouse, Cloudera AI.

**Cloudera Data Warehouse:**

This data service enables an enterprise to provision a new Cloudera Data Warehouse and share a subset of the data with a specific team or department. A Cloudera Data Warehouse cluster can be created from the Cloudera Management Console and accessed by end users (data analysts).

**Environment:**

A logical entity that represents the association of your on premises user account with compute resources using which you can provision and manage workloads such as Cloudera Data Warehouse and Cloudera AI.

**Cloudera AI:**

This data service enables teams of data scientists to develop, test, train, and ultimately deploy machine learning models for building predictive applications all on the data under management within the enterprise data cloud. A Cloudera AI Workbench can be created from the Cloudera Management Console and accessed by end users (data scientists).

**Cloudera Management Console:**

The user interface for administering Cloudera Data Services on premises. As a Cloudera administrator, you can use Cloudera Management Console to manage environments, data lakes, environment resources, and users across all Cloudera Data Services on premises.

**Replication Manager:**

This data service allows you to copy and migrate HDFS data, Hive external tables, and Ozone data between Cloudera Base on premises 7.1.8 or higher clusters using Cloudera Manager version 7.7.3 or higher.