CDP Private Cloud Data Services 1.5.4

CDP Private Cloud Data Services Overview

Date published: 2023-12-16
Date modified: 2024-05-30

CLOUDERA

https://docs.cloudera.com/

Legal Notice

Contents

About CDP Private Cloud Data Services

CDP Private Cloud Data Services works on top of CDP Private Cloud Base and is the on-premises offering of Cloudera Data Platform (CDP) that brings many of the benefits of the public cloud deployments to on-premises CDP deployments. CDP Private Cloud Data Services lets you deploy and use the Cloudera Data Warehouse (CDW), Cloudera Machine Learning (CML), and Cloudera Data Engineering (CDE) Data Services.

Building on Apache Spark, CDE 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. CDE powers consistent, repeatable, and automated data engineering workflows.

CDW 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 CDW, brings high-performance, self-service reporting and analytics to your business – simplifying data management for data practitioners and administrators.

CML, 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 CDP Private Cloud Base 7.1.8 or higher clusters using Cloudera Manager version 7.7.3 or higher.

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 CDP Public and Private Cloud, these data services open up avenues for a hybrid data lifecycle.

Benefits of CDP Private Cloud Data Services

CDP Private Cloud Data Services 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, CDP Private Cloud Data Services brings both agility and predictable performance to analytic applications. CDP Private Cloud Data Services inherits unified security, governance, and metadata management through Cloudera Shared Data Experience (SDX) from the CDP Private Cloud Base cluster.

The three main benefits of CDP Private Cloud Data Services are simplified multitenancy and isolation, simplified deployment of applications, and better utilization of infrastructure.

Simplified multitenancy and isolation

The containerized deployment of applications in CDP Private Cloud 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

CDP Private Cloud 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, CDP Private Cloud enables you to rapidly perform such deployments.

Better utilization of infrastructure

Similar to CDP Public Cloud, CDP Private Cloud 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 CDP Private Cloud ensures that your on-premises infrastructure is utilized optimally.

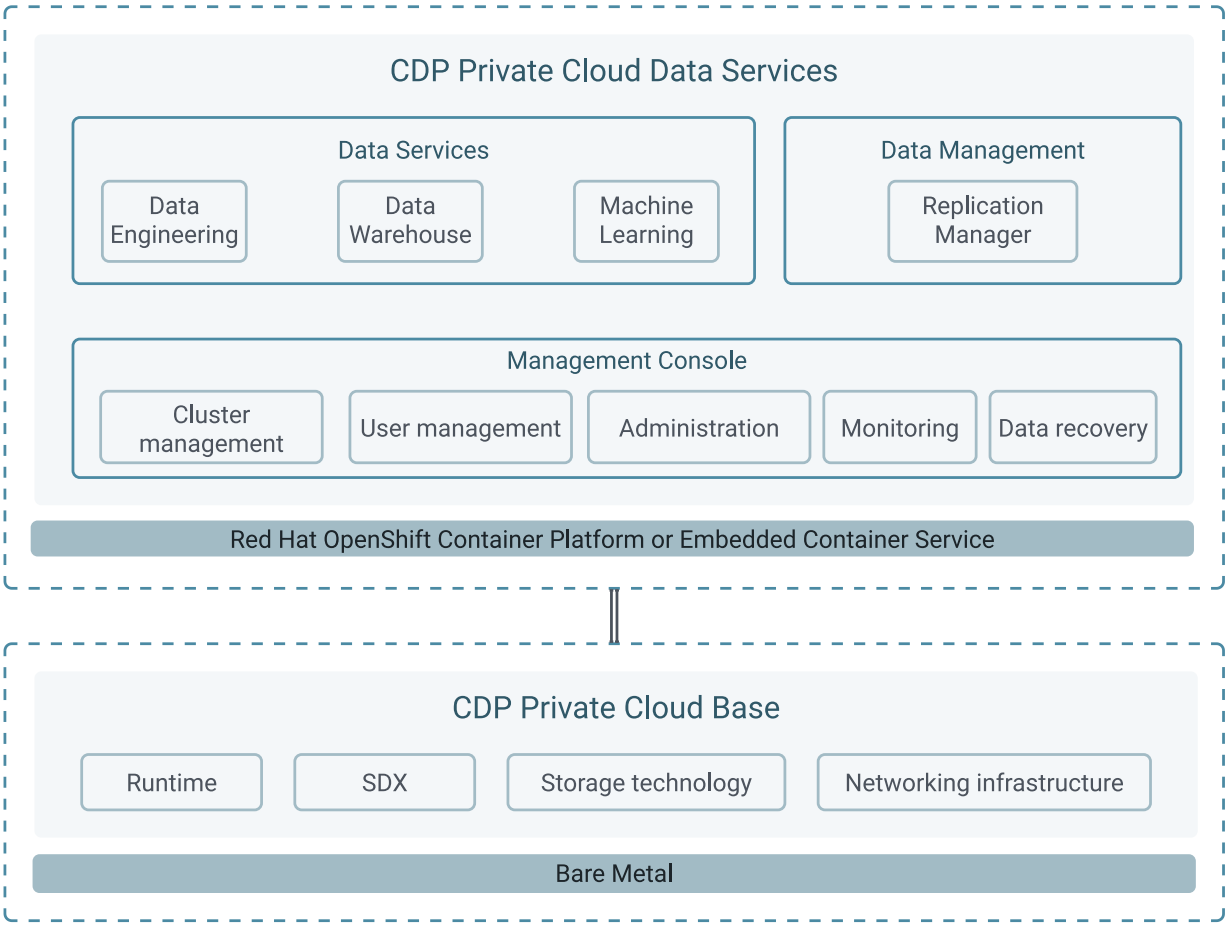
Deployment and high-level architecture

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

To deploy CDP Private Cloud Data Services you need a CDP Private Cloud Base cluster, along with container-based clusters that run the Data Services. You can either use a dedicated RedHat OpenShift container cluster or deploy an Embedded Container Service (ECS) container cluster.

The Private Cloud deployment process involves configuring 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 Management Console, and easily scale them up or down as required.



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

- 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

CDP Private Cloud Data Services glossary

The CDP Private Cloud Data Services documentation uses terms related to enterprise data cloud and cloud computing.

CDP CLI:

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

CDP Private Cloud Base:

An on-premises version of Cloudera Data Platform. 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 CDP Private Cloud Base. It includes approximately 50 open-source projects providing data management tools within CDP.

Control Plane service:

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

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. Allows users to define and tag assets using custom rules to identify and classify assets for their business requirements. 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.

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 CDP Private Cloud deployment, the Data Lake services are hosted on the CDP Private Cloud Base cluster. In addition, the Data Lake services are shared between multiple workloads.

Data Service:

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

Data Warehouse:

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

Environment:

A logical entity that represents the association of your Private Cloud user account with compute resources using which you can provision and manage workloads such as Data Warehouse and Machine Learning.

Machine Learning:

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 Machine Learning workspace can be created from the Management Console and accessed by end users (data scientists).

Management Console:

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

Replication Manager:

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