CDP Public Cloud

CDP Public Cloud Overview

Date published: 2019-08-22
Date modified:

CLOUDERA

https://docs.cloudera.com/

Legal Notice

Contents

CDP Public Cloud

Cloudera Data Platform (CDP) is a hybrid data platform designed for unmatched freedom to choose—any cloud, any analytics, any data. CDP consists of CDP Public Cloud and CDP Private Cloud.

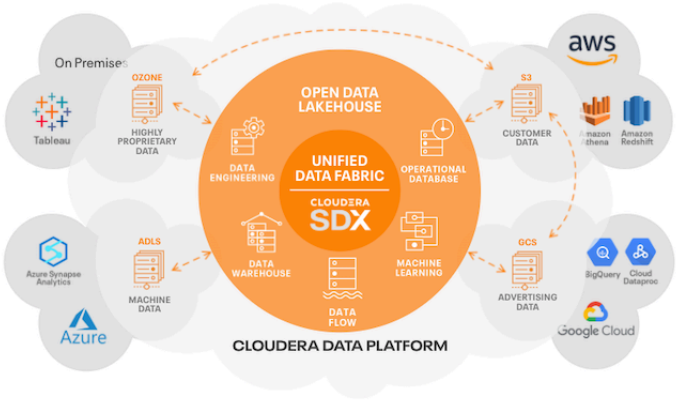
What is CDP

CDP delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security. With CDP you get the value of CDP Private Cloud and CDP Public Cloud for faster time to value and increased IT control.

Cloudera Data Platform provides the freedom to securely move applications, data, and users bi-directionally between the data center and multiple public clouds, regardless of where your data lives. All thanks to modern data architectures:

- A unified data fabric centrally orchestrates disparate data sources intelligently and securely across multiple clouds and on premises.
- An open data lakehouse enables multi-function analytics on both streaming and stored data in a cloud-native object store across hybrid multi-cloud.
- A scalable data mesh helps eliminate data silos by distributing ownership to cross-functional teams while maintaining a common data infrastructure.

With Cloudera Shared Data Experience (Cloudera SDX), CDP offers enterprise-grade security and governance. Cloudera SDX combines enterprise-grade centralized security, governance, and management capabilities with shared metadata and a data catalog, eliminating costly data silos, preventing lock-in to proprietary formats, and eradicating resource contention. Now all users and administrators can enjoy the advantages of a shared data experience.

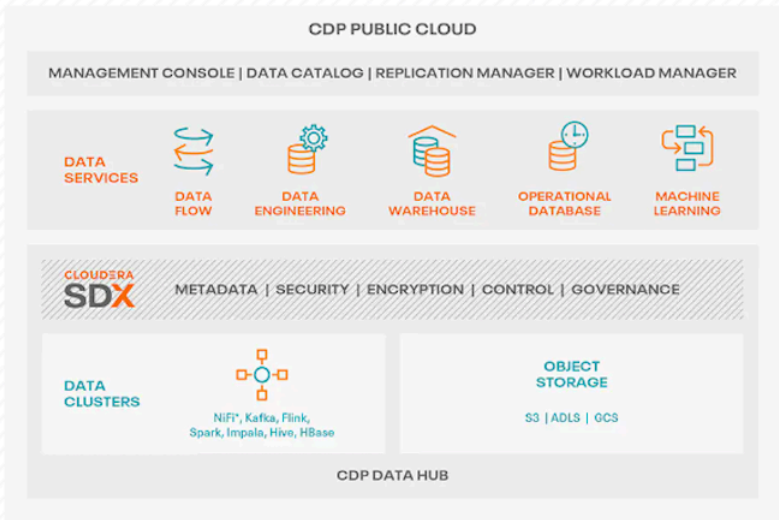


CDP Public Cloud

Create and manage secure data lakes, self-service analytics, and machine learning services without installing and managing the data platform software. CDP Public Cloud services are managed by Cloudera, but unlike other public cloud services, your data will always remain under your control in your workloads and your data will always remain under your control in your cloud account. CDP runs on AWS, Azure and Google Cloud.

CDP Public Cloud lets you:

- Control cloud costs by automatically spinning up workloads when needed, scaling them as the load changes over time and suspending their operation when complete.
- Isolate and control workloads based on user type, workload type, and workload priority.
- Combat proliferating silos and centrally control customer and operational data across multi-cloud and hybrid environments.



CDP Public Cloud use cases

CDP Public Cloud not only offers all the analytics experiences available on-premises, but also includes tools that enable hybrid and multi-cloud use cases. Customers benefit from leveraging a single interface whether deploying on a single cloud provider or using multiple cloud providers.

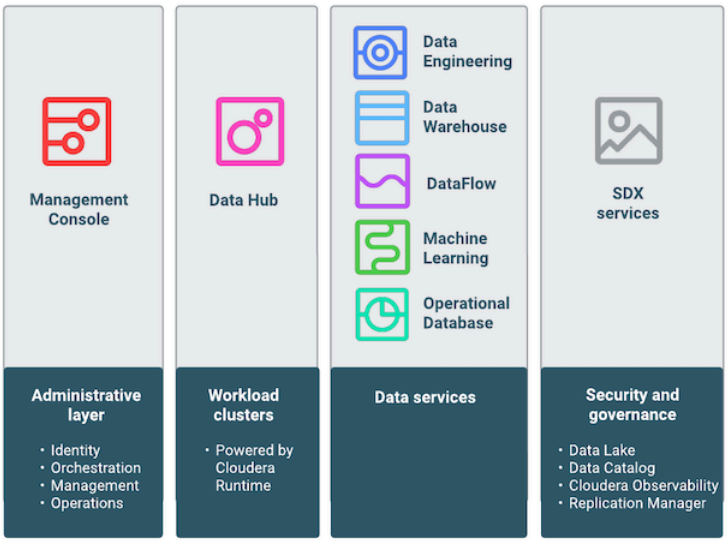
Some of the common use cases are:

- Ingest large volumes of data in real time with streaming solutions that can further enrich the data sets and curate production data that can be made available for practitioners downstream. Ensure real time data flow control to help manage transfer of data between various sources and destinations.
- Leverage object stores as centralized storage to bring together various datasets, enrich and analyze using analytical engines in CDP and generate a comprehensive understanding of the various entities in your value chain, thereby operationalizing a Customer 360 use case.
- Collect metrics from a variety of systems in both process or discrete manufacturing to ensure deviations are captured, modeled in real time and alerts are sent out for course-correction before it's too late.
- For companies looking to optimize cloud costs, run your workloads in the right place at the right time to reduce compute and storage costs and avoid lock-in with cloud providers.

CDP Public Cloud services

CDP Public Cloud consists of a number of cloud services designed to address specific enterprise data cloud use cases.

This includes Data Hub powered by Cloudera Runtime, data services (Data Warehouse, Machine Learning, Data Engineering, and DataFlow), the administrative layer (Management Console), and SDX services (Data Lake, Data Catalog, Replication Manager, and Cloudera Observability).



Administrative layer

Management Console is a general service used by CDP administrators to manage, monitor, and orchestrate all of the CDP services from a single pane of glass across all environments. If you have deployments in your data center as well as in multiple public clouds, you can manage them all in one place - creating, monitoring, provisioning, and destroying services.

Workload clusters

Data Hub is a service for launching and managing workload clusters powered by Cloudera Runtime (Cloudera's new unified open source distribution including the best of CDH and HDP). This includes a set of cloud optimized built-in templates for common workload types as well as a set of options allowing for extensive customization based on your enterprise's needs.

Data Hub provides complete workload isolation and full elasticity so that every workload, every application, or every department can have their own cluster with a different version of the software, different configuration, and running on different infrastructure. This enables a more agile development process.

Since Data Hub clusters are easy to launch and their lifecycle can be automated, you can create them on demand and when you don't need them, you can return the resources to the cloud.

Data services

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

DataFlow is a cloud-native universal data distribution service powered by Apache NiFi that lets developers connect to any data source anywhere with any structure, process it, and deliver to any destination. It offers a flow-based low-code development paradigm that aligns best with how developers design, develop, and test data distribution pipelines.

Data Warehouse is a service for creating and managing self-service data warehouses for teams of data analysts. This service makes it easy for an enterprise to provision a new data warehouse and share a subset of the data with a specific team or department. The service is ephemeral, allowing you to quickly create data warehouses and terminate them once the task at hand is done.

Machine Learning is a service for creating and managing self-service Machine Learning workspaces. This 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.

Operational Database is a service for self-service creation of an operational database. Operational Database is a scale-out, autonomous database powered by Apache HBase and Apache Phoenix. You can use it for your low-latency and high-throughput use cases with the same storage and access layers that you are familiar with using in CDH and HDP.

Security and governance

Shared Data Experience (SDX) is a suite of technologies that make it possible for enterprises to pull all their data into one place to be able to share it with many different teams and services in a secure and governed manner. There are four discrete services within SDX technologies: Data Lake, Data Catalog, Replication Manager, and Cloudera Observability.

Data Lake is a set of functionality for creating safe, secure, and governed data lakes which provides a protective ring around the data wherever that's stored, be that in cloud object storage or HDFS. Data Lake functionality is subsumed by the Management Console service and related Cloudera Runtime functionality (Ranger, Atlas, Hive MetaStore).

Data Catalog is a service for searching, organizing, securing, and governing data within the enterprise data cloud. Data Catalog is used by data stewards to browse, search, and tag the content of a data lake, create and manage authorization policies (by file, table, column, row, and so on), identify what data a user has accessed, and access the lineage of a particular data set.

Replication Manager is a service for copying, migrating, snapshotting, and restoring data between environments within the enterprise data cloud. This service is used by administrators and data stewards to move, copy, backup, replicate, and restore data in or between data lakes. This can be done for backup, disaster recovery, or migration purposes, or to facilitate dev/test in another virtual environment.

Cloudera Observability is a service for analyzing and optimizing workloads within the enterprise data cloud. This service is used by database and workload administrators to troubleshoot, analyze, and optimize workloads in order to improve performance and/or cost.

The following table illustrates which services are available on supported cloud providers:

	CDP service	AWS	Azure	GCP
Data services	Data Engineering	GA	GA	Not available
	DataFlow	GA	GA	Only DataFlow Functions are available on GCP. Other DataFlow features are not available.
	Data Hub	GA	GA	GA
	Data Warehouse	GA	GA	Not available
	Machine Learning	GA	GA	Not available
	Operational Database	GA	GA	GA (with HDFS) Preview (with GCS)
Control Plane services	Data Catalog	GA	GA	GA
	Management Console	GA	GA	GA
	Replication Manager	GA	GA	Not available
	Cloudera Observability	GA	GA	GA

Related Information

Management Console
Data Hub
Data Engineering
DataFlow
Data Warehouse
Machine Learning

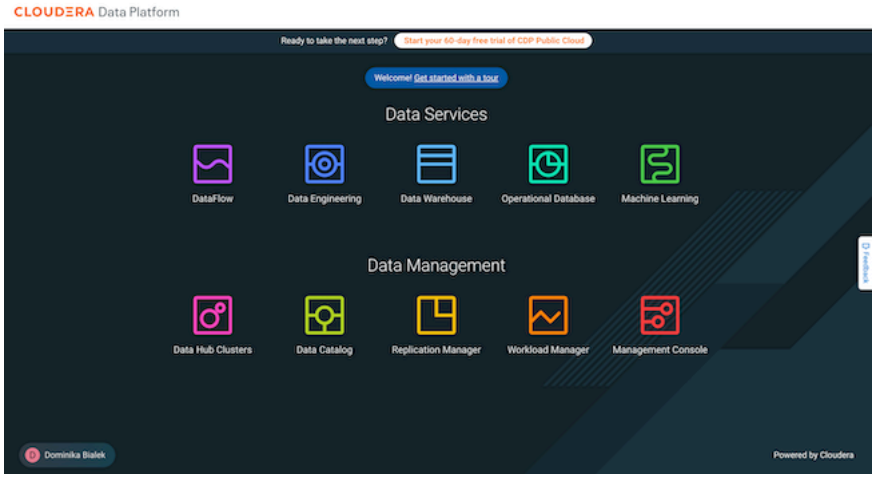
Data Catalog
Replication Manager
Cloudera Observability

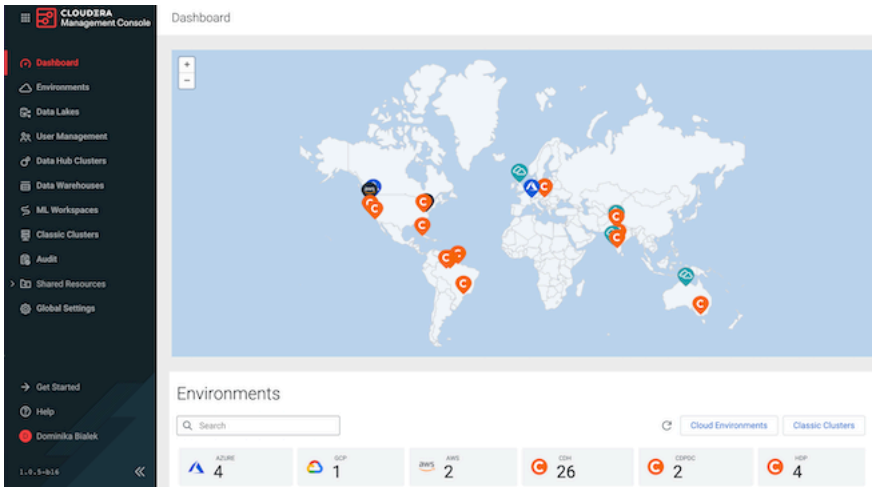
CDP Public Cloud interfaces

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

Web interface

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





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 web console. Furthermore, it allows you to automate routine tasks such as cluster creation.

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.

Related Information

Supported Browsers Policy

CDP Public Cloud glossary

CDP Cloud documentation uses terminology related to enterprise data cloud and cloud computing.

CDP (Cloudera Data Platform) Public Cloud - CDP Public Cloud is a cloud service platform that consists of a number of services. It enables administrators to deploy CDP service resources and allows end users to process and analyze data by using these resources.

CDP CLI - Provides a command-line interface to access and manage CDP services and resources.

CDP web console - The web interface for accessing and manage CDP services and resources.

Cloudera Observability (data service) - A CDP data service used by database and workload administrators to troubleshoot, analyze and optimize workloads in order to improve performance and/or cost.

Cloudera Runtime - The open source software distribution within CDP that is maintained, supported, versioned, and packaged by Cloudera. Cloudera Runtime combines the best of CDP and HDP. Cloudera Runtime 7.0.0 is the first version.

Cluster - Also known as compute cluster, workload cluster, or Data Hub cluster. The cluster created by using the Data Hub service for running workloads. A cluster makes it possible to run one or more Cloudera Runtime components on some number of VMs and is associated with exactly one data lake.

Cluster definition - A reusable cluster template in JSON format that can be used for creating multiple Data Hub clusters with identical cloud provider settings. Data Hub includes a few built-in cluster definitions and allows you to save your own cluster definitions. A cluster definition is not synonymous with a blueprint, which primarily defines Cloudera Runtime services.

Cluster Repair - A feature in CDP Management Console that enables you to select specific nodes within a node group for a repair operation. This feature reduces the downtime incurred when only a subset of the nodes are unhealthy.

Cluster template - A reusable cluster template in JSON format that can be used for creating multiple Data Hub clusters with identical Cloudera Runtime settings. It primarily defines the list of Cloudera Runtime services included and how their components are distributed on different host groups. Data Hub includes a few built-in blueprints and allows you to save your own blueprints. A blueprint is not synonymous with a cluster definition, which primarily defines cloud provider settings.

Control Plane - A Cloudera operated cloud service that includes services like Management Console, Cloudera Observability, Replication Manager and Data Catalog. These services interact with your account in Amazon Web Services (AWS), Microsoft Azure, and Google Cloud to provision and manage compute infrastructure that you can use to manage the lifecycle of data stored in your cloud account. In addition, the Control Plane can interface with your on-premises and Private Cloud infrastructure to support hybrid cloud deployments.

Credential - Allows an administrator to configure access from CDP to a cloud provider account so that CDP can communicate with that account and provision resources within it. There is one credential per environment.

Data Catalog (data service) - A CDP data service used by data stewards to browse, search, and tag the content of a data lake, create and manage authorization policies, identify what data a user has accessed, and access the lineage of a particular data set.

DataFlow (data service) - A CDP data service that enables you to import and deploy your data flow definitions efficiently, securely, and at scale.

Data Lake - A single logical store of data that provides a mechanism for storing, accessing, organizing, securing, and managing that data.

Data Lake cluster - A special cluster type that implements the Cloudera Runtime services (such as HMS, Ranger, Atlas, and so on) necessary to implement a data lake that further provides connectivity to a particular cloud storage service such as S3 or ADLS.

Data Hub (service) - A CDP service that administrators use to create and manage clusters powered by Cloudera Runtime.

Data Warehouse (data service) - A CDP data service for creating and managing self-service data warehouses for teams of data analysts.

Data warehouse - The output of the Data Warehouse service. Users access data warehouses via standard business intelligence tools such as JDBC or Tableau

Environment - A logical environment defined with a specific virtual network and region in a customer’s cloud provider account. One can refer to a "CDP environment" once a cloud provider virtual network, cloud storage, and other cloud provider artifacts present in a customer's AWS, Azure, or GCP account have been registered in CDP. CDP service components such as Data Hub clusters, Data Warehouse clusters, and so on, run in an environment.

Image catalog - Defines a set of images that can be used for provisioning Data Hub cluster. Data Hub includes a built-in image catalog with a set of built-in base and prewarmed images and allows you to register your own image catalog.

Machine Learning (data service) - A CDP data service that administrators use to create and manage Machine Learning workspaces and that allows data scientists to do their machine learning.

Machine Learning workspace - The output of the Machine Learning service. Each workspace corresponds to a single cluster that can be accessed by end users.

Management Console (data service) - A CDP data service that allows an administrator to manage environments, users, and services; and download and configure the CLI.

Operational Database (data service) - A CDP data service that administrators use to create and manage scale-out, autonomous database powered by Apache HBase and Apache Phoenix.

Recipe - A reusable script that can be used to perform a specific task on a specific resource.

Replication Manager (data service) - A CDP data service used by administrators and data stewards to move, copy, backup, replicate, and restore data in or between data lakes.

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 Hub, Data Warehouse, Machine Learning.

Shared resources - A set of resources such as cloud credentials, recipes (custom scripts), and other that can be reused across multiple environments.