

Data Warehouse Overview

Date published: 2024-01-01

Date modified: 2024-08-15

CLOUDERA

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

Service overview.....4

Service architecture.....4

Key features.....7

About SQL AI Assistant.....8

Cloudera Data Warehouse - Service overview and components

Cloudera Data Warehouse (CDW) Data Service is a containerized application for creating highly performant, independent, self-service data warehouses in the cloud which can be scaled dynamically and upgraded independently. Learn more about the service architecture, and how CDW enables data practitioners and IT administrators to achieve their goals.

Cloudera Data Platform (CDP) enables you to implement powerful modern data architectures such as Data Mesh, Data Fabric, and Data Lakehouse. CDP supports a Data Lakehouse architecture by pre-integrating and unifying the capabilities of Data Warehouses and Data Lakes, to support data engineering, business intelligence, and machine learning – all on a single platform. 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 both for data practitioners and administrators.

CDW provides tight integration with the other Cloudera Data Services providing data ingestion, data engineering, machine learning, and data visualization.

CDW leverages Apache Iceberg, Apache Impala, Hive ACID, and Hive table format to provide broad coverage, enabling the best optimized set of capabilities to each workload.

For data practitioners, CDW provides consistent quick response times with high concurrency, easy data exploration and business intelligence on any enterprise data across the Data Lake. CDW also supports streamlined application development with open standards, open file and table formats, and standard APIs. Through tenant isolation, CDW can process workloads that do not interfere with each other, so everyone meets report timelines while controlling costs.

For administrators, CDW simplifies administration by making multi-tenancy secure and manageable and reduces cloud costs. Virtual Warehouses can be provisioned on-demand, using self-service, and de-provisioned when idle. Administrators also benefit from the ability to independently upgrade the Virtual Warehouses and Database Catalogs. Additionally, CDW allows the choice of the version for the Hive or Impala Virtual Warehouse, and Hue you want to use.

Related Information

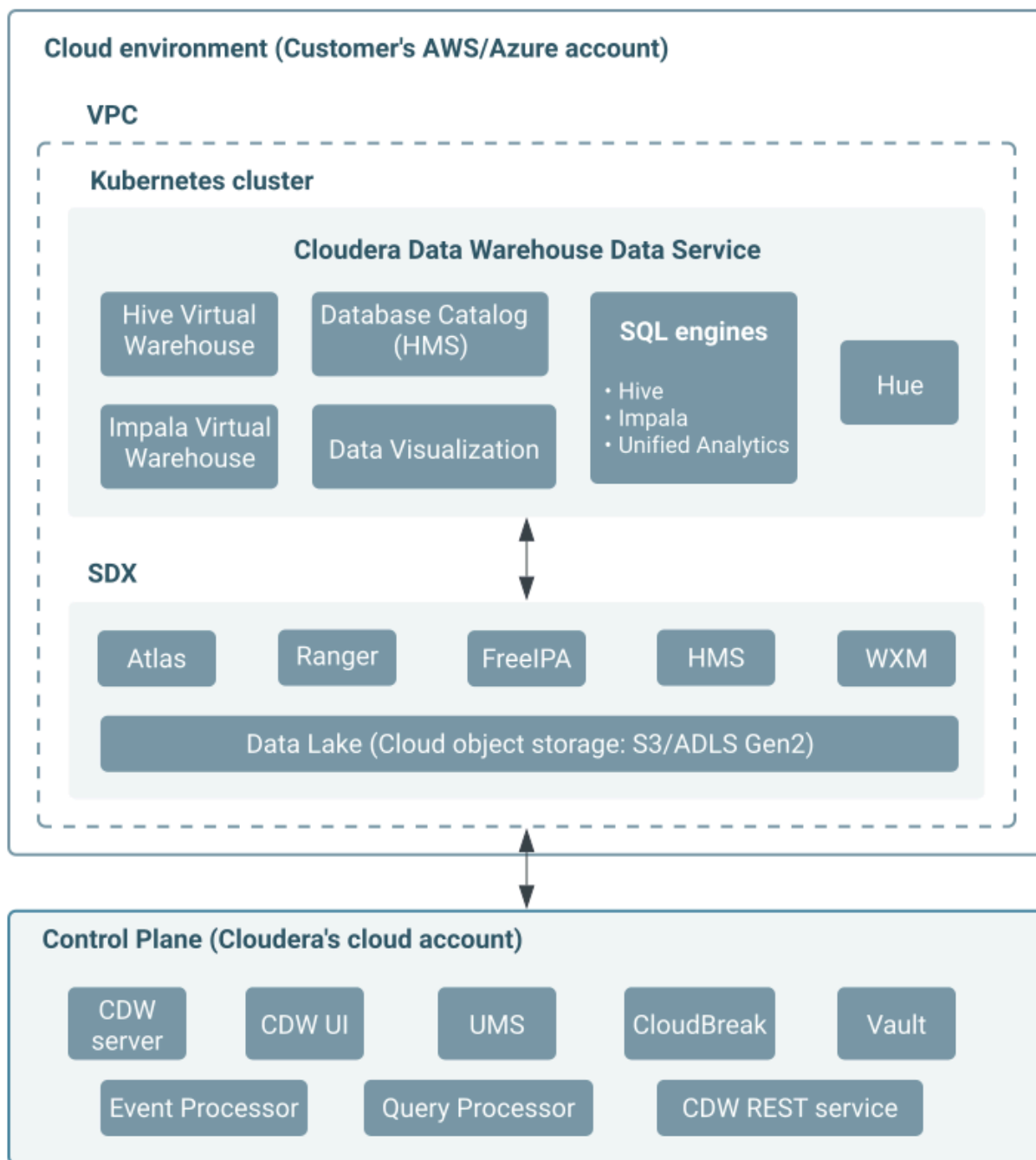
[Adding a new Virtual Warehouse](#)

[CDP identity management](#)

Cloudera Data Warehouse service architecture

Administrators and IT teams can get a high-level view of the Cloudera Data Warehouse service components and how they are integrated within the CDP stack.

The Data Warehouse service is composed of the Data Warehouse Control Plane (CRUD application), Database Catalogs (storage prepared for use with a Virtual Warehouse) and Virtual Warehouses (compute environments that can access a Database Catalog) and they are decoupled by design. Multiple Virtual Warehouses of differing sizes and types can be configured to operate on the same Database Catalog, providing workload diversity and isolation on the same data at the same time. Cloudera Data Warehouse is also integrated with Cloudera Data Visualization for visualizing data and obtaining insights.



Cloudera Data Warehouse Control Plane

The Cloudera Data Warehouse Control Plane is a CRUD application that serves every operation that you can take on the Data Warehouse UI or by using the CDP CLI. For example, editing a Data Warehouse environment, deactivating the environment, upgrading a Virtual Warehouse, suspending (stopping), and resuming (starting or restarting) Data Warehouse entities. The Cloudera Data Warehouse Control Plane version is displayed on the bottom left corner of the UI and automatically reflects the latest version when Cloudera publishes a release. For example 1.9.2-b657.

Cloudera Data Warehouse Environment

A Cloudera Data Warehouse Environment is a cluster that is deployed on your cloud infrastructure (AWS or Azure). After you register your account with Cloudera on the Management Console, the environment is made available in the Data Warehouse service. You must then activate your environment in the Data Warehouse service. The Data Warehouse environment inherits the version of the Data Warehouse Control Plane when it is created.

The Environment version is displayed on the **Environment Details** page on the Data Warehouse UI. For example, 1.9.2-b657. The Data Warehouse Environment version is not automatically upgraded. You must deactivate and reactivate the environment using the backup and restore process to upgrade to the latest version. Reactivating an environment also updates the Kubernetes (AKS or EKS) version.

Database Catalog

A Database Catalog is a logical collection of table and view metadata, security permissions, and other information. Behind each Database Catalog is a Hive metastore (HMS) that collects your definitions about data in cloud storage. An object store in a secure data lake contains all the actual data for your environment. A Database Catalog includes transient user and workload contexts from the Virtual Warehouse and governance artifacts that support functions such as auditing. Multiple Virtual Warehouses can query a Database Catalog. An environment can have multiple Database Catalogs.

When you activate an environment from the Data Warehouse, a Database Catalog is created automatically and named after your environment. The environment shares a default HMS with services, such as Cloudera Data Engineering (CDE), CDW, Cloudera Machine Learning (CML) to some extent, and Data Hub templates, such as Data Mart. Consequently, the same objects and data sets are accessible from CDW or any Data Hubs created in the environment by virtue of using the same HMS. Queries and query history saved in the Hue database are stored in the Database Catalog and not deleted when you delete a Virtual Warehouse.

The Database Catalog version is displayed on the Database Catalog Details page. For example, 2024.0.18.1-1. This is the Runtime version of the HMS. The Database Catalogs are automatically upgraded when you deactivate and reactivate your Data Warehouse Environment. You can also upgrade the Database Catalog independently without reactivating the Environment. Data Warehouse displays an Upgrade Now option when a new version is available.

Virtual Warehouses

A Virtual Warehouse is an instance of compute resources running in Kubernetes to execute the queries. From a Virtual Warehouse, you access tables and views of your data in a Database Catalog's Data Lake. Virtual Warehouses bind compute and storage by executing authorized queries on tables and views through the Database Catalog. Virtual Warehouses can scale automatically, and ensure performance even with high concurrency. All JDBC/ODBC compliant tools connect to the virtual warehouse to run queries. Virtual Warehouses also expose HS2-compatible endpoints for CLI tools such as Beeline, Impala-Shell, and Impyla.

The Virtual Warehouse version is displayed on the **Virtual Warehouse Details** page. For example, 2024.0.18.1-1. This is the Runtime version of the underlying SQL engines (Hive, Impala) and Iceberg service that is distributed with a particular release. The automatic or semi-automatic backup and restore procedure recreates the Virtual Warehouses with the latest version however this is not the recommended way to upgrade the Virtual Warehouses. You can upgrade the Virtual Warehouses independently without reactivating the Environment. Data Warehouse displays an Upgrade Now option when a new version is available.



Note: At times, you might be required to upgrade your Database Catalogs when you upgrade the Virtual Warehouses. A warning indicating this is displayed on the Data Warehouse UI.

Data Visualization

In addition to Database Catalogs and Virtual Warehouses that you use to access your data, CDW integrates Data Visualization for building graphic representations of data, dashboards, and visual applications based on CDW data, or other data sources you connect to. You, and authorized users, can explore data across the entire CDP data lifecycle using graphics, such as pie charts and histograms. You can arrange visuals on a dashboard for collaborative analysis.

Key features of Cloudera Data Warehouse Public Cloud

The Cloudera Data Warehouse (CDW) service provides data warehouses that can be automatically configured and isolated. CDW optimizes existing workloads when you move to the cloud. You scale resources up and down to meet your business demands, and save costs by suspending and resuming resources automatically. Data warehouses comply with your Data Lake security requirements.

Automatically configured and isolated

Each data warehouse can be automatically configured for you by the CDW service, but you can adjust certain settings to suit your needs. Individual warehouses are completely isolated, ensuring that the users have access to only their data and eliminating the problem of "noisy neighbors." *Noisy neighbors* are workloads that monopolize system resources and interfere with the queries from other tenants. With Cloudera Data Warehouse, you can easily offload noisy neighbor workloads to their own Virtual Warehouse instance so other tenants have access to enough compute resources for their workloads to complete and meet their SLAs.

This capability to isolate individual warehouses is equally useful for "VIP workloads." *VIP workloads* are crucial workloads that must have resources to complete immediately and as quickly as possible without waiting in a queue. You can run these VIP workloads in their own warehouse to ensure they get the resources they need to complete as soon as possible.

Optimized for your workloads

Data warehouses are automatically optimized for your workloads. This includes pre-configuring the software and creating the different caching layers, which means that you need not engage in complex capacity planning or tuning. You create a Virtual Warehouse that specifies a SQL engine:

- Hive for data warehouses that support complex reports and enterprise dashboards.
- Impala for Data Marts that support interactive, ad-hoc analysis.

You choose the Virtual Warehouse instance size and adjust thresholds for auto-scaling.

Auto-scaling

Auto-scaling enables both scaling up and scaling down of Virtual Warehouse instances so they can meet your varying workload demands and save costs on cloud resources when they are not needed.

Auto-scaling provides the following benefits:

- Service availability: Clusters are ready to accept queries "24 x 7".
- Auto-scaling based on query wait-time: Queries start executing within the number of seconds that you specify and cluster resources are added or shut down to meet demand.
- Auto-scaling based on number of concurrent queries running on the system: "Infinite scaling" means that the number of concurrent queries can go from 10 to 100 in minutes.
- Cost guarantee: You can configure auto-scaling upper limits, which determine how large a compute cluster can grow. Since compute costs increase as cluster size increases, having a way to configure upper limits gives administrators a method to stay within a budget.

Auto-suspend and resume

You have the capability to set an AutoSuspend Timeout when creating a Virtual Warehouse. This sets the maximum time a Virtual Warehouse idles before shutting down. For example, if you set this to 60 seconds, then if the Virtual Warehouse is idle for 60 seconds, it suspends itself so you do not have to pay for unused compute resources. The first time a new query is run against an auto-suspend Virtual Warehouse, it restarts. This feature helps you maintain a tight control on your cloud spend while ensuring availability to run your workloads.

Security compliance

Your Database Catalogs and Virtual Warehouses automatically inherit the same security restrictions that are applicable to your CDP environment. There is no need to specify the security setup again for each Database Catalog or Virtual Warehouse. For more information, see CDP identity management. It discusses integration with Apache Knox and your LDAP provider which uses FreeIPA Identity Management.

The following security controls are inherited from your CDP environment:

- Authentication: Ensures that all users have proven their identity before accessing the Cloudera Data Warehouse service or any created Database Catalogs or Virtual Warehouses.
- Authorization: Ensures that only users who have been granted adequate permissions are able to access the Cloudera Data Warehouse service and the data stored in the tables.
- Dynamic column masking: If rules are set up to mask certain columns when queries run, based on the user executing the query, then these rules also apply to queries executed in the Virtual Warehouses.
- Row-level filtering: If rules are set up to filter certain rows from being returned in the query results, based on the user executing the query, then these same rules also apply to queries executed in the Virtual Warehouses.

Tenant isolation

The multitenant storage technique in CDW offers increased security over the storage method used in earlier releases. The earlier releases based all storage access in CDW on a single EC2 instance role. Tenant isolation offers users independence on several levels:

- Isolation reduces contention
- Workloads do not interfere with each other
- Each tenant can choose a version of an independent deployment
- Independent upgrades limit the scope of changes

Related Information

[CDP identity management](#)

About the Hue SQL AI Assistant

Learn about the AI models and services that Hue uses to run the SQL AI Assistant and its limitations. Review what data is shared with the LLM models before you start using the SQL AI Assistant with Hue.

A SQL AI Assistant has been integrated into Hue with the capability to leverage the power of Large Language Models (LLMs) for various SQL tasks. It helps you to create, edit, optimize, fix, and succinctly summarize queries using natural language and makes SQL development faster, easier, and less error-prone. SQL AI assistant is available with the Hue image version 2023.0.16.0 and higher on CDW Public Cloud. Both Hive and Impala dialects are supported.

AI models and services that Hue uses

The SQL AI Assistant supports various LLMs and hosting services. The models run on cloud infrastructure, and the AI Assistant can be configured to use them remotely. Cloudera has tested with GPT running in Open AI, Microsoft Azure, and Amazon Bedrock. The following service-model combinations are supported:

Service Provider	Model	Model Versions
OpenAI	OpenAI GPT	<ul style="list-style-type: none">• gpt-3.5-turbo• gpt-3.5-turbo-16k Current GPT version is 3.5 turbo. You can configure GPT 4 for better results.

Service Provider	Model	Model Versions
Microsoft Azure	OpenAI GPT	<ul style="list-style-type: none"> gpt-3.5-turbo gpt-3.5-turbo-16k
Amazon Bedrock	Anthropic Claude	<ul style="list-style-type: none"> anthropic.claude-v1 anthropic.claude-v2
Amazon Bedrock	Amazon Titan	<ul style="list-style-type: none"> amazon.titan-text-express-v1

**Note:**

You must have access to the Hugging Face to download the required sentence transformer model, and ensure your system can connect to the internet, specifically, huggingface.co. Since these models aren't pre-bundled, they must be downloaded during setup. For more information, see [Hugging Face](#)

For better results, Cloudera recommends you to use the SQL AI assistant with the Azure OpenAI service. This ensures that the models run in your Virtual Private Cloud (VPC) network.

The SQL AI Assistant uses a Retrieval Augmented Generation (RAG)-based architecture for augmenting results. It uses the sentence-transformer library for semantic search, and Hue can be configured with any of the [pre-trained models](#) for better multi-lingual support. By default, “all-MiniLM-L6-v2” models are used.

Embedding Model	Language Support
all-MiniLM-L6-v2	English
distiluse-base-multilingual-cased-v1	Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, and Turkish.

What data is shared with the LLM models

The following details are shared with the LLMs:

- Everything that a user inputs
- Dialect in use
- Table details such as table name, column names, column data types and related keys, partitions, and constraints that the logged-in user has access to.
- Three sample rows from the tables (as per the best practices specified in [Evaluating the Text-to-SQL Capabilities of Large Language Models](#))

Limitations**Non-deterministic nature**

LLMs are non-deterministic, which means you cannot guarantee the same output for the same input every time, and it can lead to different responses to similar queries.

Ambiguity

LLMs may struggle to handle ambiguous queries or contexts. SQL queries often rely on specific and unambiguous language, but LLMs can misinterpret or generate ambiguous SQL queries, leading to incorrect results.

Hallucinations

In the context of LLMs, hallucination refers to a phenomenon where these models generate text or responses that are incorrect, nonsensical, or fabricated. Occasionally you might see incorrect identifiers or literals in the response.

Related Information

[About setting up the SQL AI Assistant in CDW](#)

[Starting the SQL AI Assistant in Hue](#)