Cloudera Data Warehouse Public Cloud

CDW Top Tasks

Date published: 2024-01-01
Date modified: 2024-12-05

CLOUDERA

Legal Notice

Contents

Exploring a data lake

In Cloudera Data Warehouse, you explore sample airline database tables in your data lake from a Virtual Warehouse. You learn how to load the airline data. You view and query the tables.

Before you begin

- You obtained permissions to access a running environment for creating a Database Catalog and Virtual Warehouse.
- You obtained the DWAdmin role to perform Data Warehouse tasks.
- You logged into the CDP web interface.
- You activated the environment from Cloudera Data Warehouse.
- You set up a Hadoop SQL policy in Ranger to access data in the data lake.

For more information about meeting prerequisites, see Getting started in CDW.

About this task

In this task, you set up a minimal Virtual Warehouse for learning how to explore a data lake. You do not need to configure auto-scaling and optional features to explore a data lake. Plan to delete the Virtual Warehouse after your exploration.

Procedure

1. Navigate to Data Warehouse Database Catalogs New Database Catalog .
2. In New Database Catalog, in Name, specify a Database Catalog name.
3. In Environments, select the name of your environment activated from Cloudera Data Warehouse.
4. Accept default values for the image version and data lake type (SDX).
5. Turn on Load Demo Data to explore sample airline data from Hue, and click Create.
6. Click Virtual Warehouses New Virtual Warehouse .

7. Set up the Virtual Warehouse:

- Specify a Name for the Virtual Warehouse.
- In Type, click the SQL engine you prefer: Hive or Impala.
- Select your Database Catalog and User Group if you have been assigned a user group.

New Virtual Warehouse

Name *

    my-vw

Type *

    HIVE   IMPALA

Database Catalog *

    qe-dc-dwx-hive-eodo

Availability Zone ⓘ

    -- select an option --

✓ Enable SSO ⓘ

User Groups ⓘ

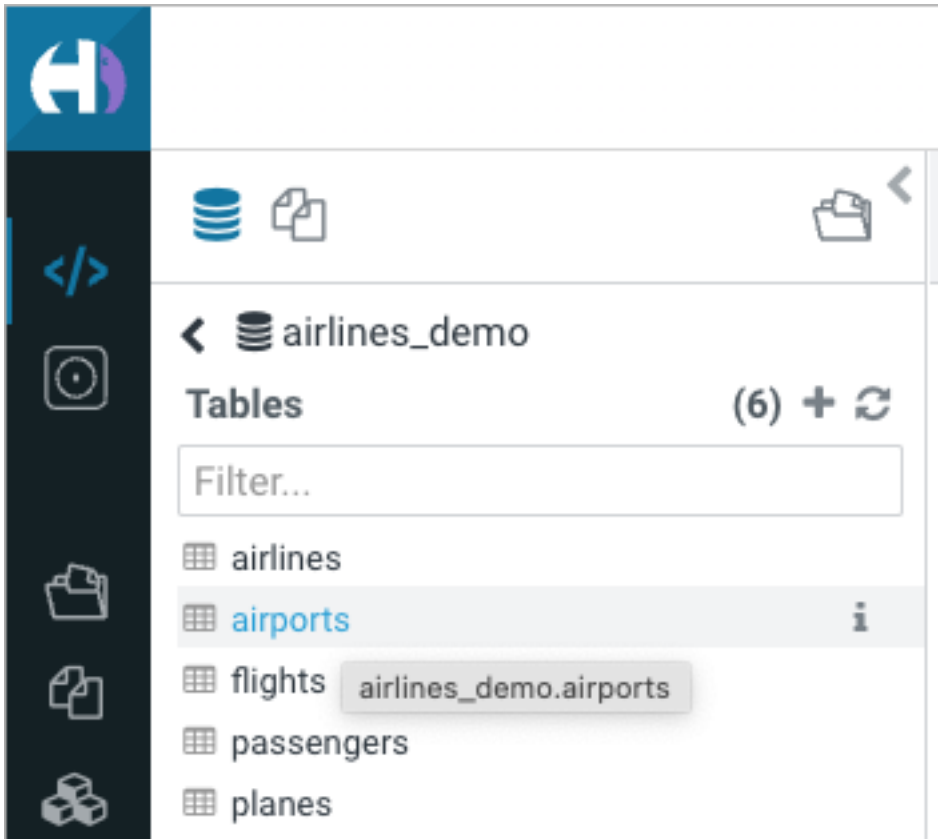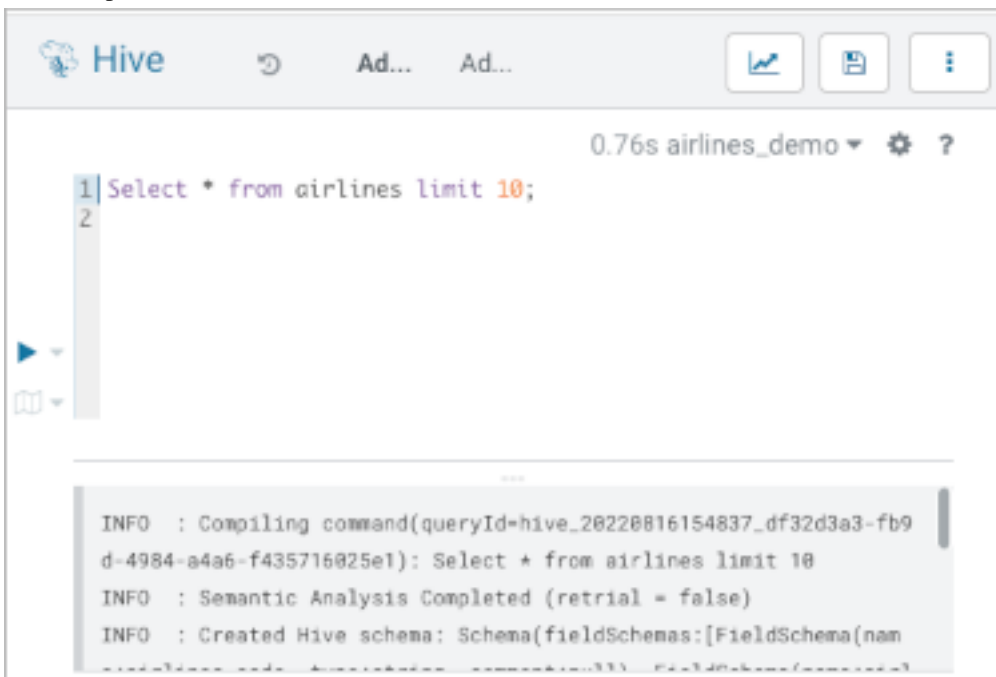   Select Groups

Tagging ⓘ

- In Size, select the number of executors, for example xsmall-2Executors.
- Accept default values for other settings.

8. Click Create.

9. After your Virtual Warehouse starts running, click Hue, and expand Tables to explore available data.

10. Explore data lake contents by running queries.

For example, select all data from the airlines table.

Creating a Virtual Warehouse

A Virtual Warehouse is an instance of compute resources in Public Cloud that is equivalent to an on-prem cluster. You learn how to create a new Virtual Warehouse in Cloudera Data Warehouse (CDW) Public Cloud.

About this task

A Virtual Warehouse provides access to the data in tables and views in the data lake your Database Catalog uses. A Virtual Warehouse can access only the Database Catalog you select during creation of the Virtual Warehouse.

In this task and subtasks, you configure Virtual Warehouse features, including performance-related features for production workloads, such as the Virtual Warehouse size and auto-scaling. These features are designed to manage huge workloads in production, so if you are evaluating CDW, or just learning, simply accept the default values. This task covers the bare minimum configurations.

In subtopics, you see details about how to configure features for production workloads, such as Hive query isolation and Impala catalog high availability.

Before you begin

- You obtained permissions to access a running environment for creating a Virtual Warehouse.
- You obtained the DWAdmin role to perform Data Warehouse tasks.
- You logged into the CDP web interface.
- Your activated the environment from Cloudera Data Warehouse.

For more information about meeting these prerequisites, see Getting started in CDW.

Procedure

1. Navigate to   Data Warehouses Virtual Warehouses Create Virtual Warehouse .
2. In New Virtual Warehouse, specify a Name.

    Note: The fully qualified domain name of your Virtual Warehouse, which includes the Virtual Warehouse name plus the environment name must not exceed 64 characters; otherwise, Hue cannot load.

3. Select the Hive or Impala type of Virtual Warehouse you want.

    Virtual Warehouses can use Hive or Impala as the underlying SQL execution engine. Typically, Hive is used to support complex reports and enterprise dashboards. Impala is used to support interactive, ad-hoc analysis.
4. Select the Environment and Database Catalog for this Virtual Warehouse.
5. In AWS environments only, accept the default availability zone, or select an availability zone, such as us-east-1c.

    The default behavior is to randomly select an availability zone from the list of configured availability zones for the associated environment. Generally, it is fine to accept the default. All compute resources will run in the selected zone.

    Note: Selection of the zone is not an option in Azure environments.

6. Select the Compute Instance Types for the Virtual Warehouse based on your workload.

    For more information, see *Supported Compute Instance Types*.
7. Select the Hive Image Version or the Impala Image Version version, and the Hue Image version you want to use, or accept the default version (latest) at the top of the drop-down menus.
8. Select the Size of the Virtual Warehouse as described in the next subtopic.

    The "AutoSuspend Timeout" and "Concurrency Autoscaling" controls appear.
9. Configure auto-scaling as described in the subsequent subtopic.

10. In Authentication, select Enable SSO to enable single sign-on to your Virtual Warehouse, and in User Groups, select a user group set up in advance to access endpoints. If you do not have a user group set up for SSO, do not select Enable SSO.

    In Management Console User Management you set up a user group, required for enabling SSO, that identifies the users authorized to access to this Virtual Warehouse.

11. Enter keys and values for Tagging the Virtual Warehouse.

12. Accept default values for other settings, or change the values to suit your use case, and click Create Virtual Warehouse to create the new Virtual Warehouse.

    Click the tooltip ⓘ for information about settings.

    When you create a Virtual Warehouse, a cluster is created in your cloud provider account. This cluster has two buckets. One bucket is used for managed data and the other is used for external data.

How to size a Virtual Warehouse

When you create a Virtual Warehouse, you need to carefully set the size of your Virtual Warehouse. The size of the Virtual Warehouse you select during Virtual Warehouse creation determines the number of executors and concurrent queries the Virtual Warehouse can run.

Before creating a Virtual Warehouse and setting the Virtual Warehouse size, learn about critical "Warehouse sizing requirements". If you misconfigure the size of your Virtual Warehouse, you can use one of the following methods, depending on the Impala or Hive type of Virtual Warehouse, to correct the size:

- Impala: Edit your Virtual Warehouse configuration to correct the size.
- Hive: Delete and recreate the Virtual Warehouse to correct the size.

When you create a Virtual Warehouse, you select one of the following Virtual Warehouse sizes:

Virtual Warehouse Size	Number of Executors
XSMALL	2
SMALL	10
MEDIUM	20
LARGE	40
Custom	Enter a value between '1' and '100'

If you are evaluating CDW, or just learning, XSMALL is the recommended size. For production workloads, choose a size based on the following factors:

- The number of executors you typically use for clusters in an on-premises deployment.
- The complexity of your queries and the size of the data sets that they access.

Large warehouses with more executors can cache more data than small warehouses. Caching enhances performance.

Auto-suspend timeout in CDW

Auto-suspend enables you to handle resources when the auto-scaler has scaled back to the last executor group. You can control the time that the original warehouse executor group idles after all other groups scale down and release their executors. The JDBC endpoint lives on to respond to queries from the result cache or statistics, but expensive executors no longer run.

When the auto-scaler increases or decreases the number of executor groups in the Virtual Warehouse, it can take several minutes before the scaling up or down takes effect. This slight delay is caused by the time required by your cloud provider to provision clusters.

When no queries are sent to an executor group, resources scale down and executors are released. When all executor groups are scaled back, when executors are idle, and after a period of idle time (Auto-suspend Timeout), the Virtual Warehouse is suspended.

You set an Auto-suspend Timeout to configure how long a Virtual Warehouse idles before shutting down.

Auto-suspend timeout is independent of the auto-scaling process and only applies to the original Virtual Warehouse and not to any additional warehouses that are created as a result of auto-scaling.

If a query is executed on a suspended Virtual Warehouse, then the query coordinator queues the query. When a queued query is detected, an executor group is immediately added (a scale-up occurs) to run the query on the Virtual Warehouse.

This feature can help lower cloud costs.

Auto-suspend differs from manually stopping a Virtual Warehouse. Sending a query to a stopped Virtual Warehouse returns an HTTP 503 "no healthy upstream" error, no scale-up occurs, and the query does not execute.

Auto-suspend is enabled by default.

While creating a Virtual Warehouse, set Auto-suspend Timeout to the number of seconds you want the Virtual Warehouse to idle before shutting down. For example, if you set the timeout to 300 seconds, the Virtual Warehouse suspends itself after 300 seconds to save compute resource expense. The suspended Virtual Warehouse restarts as soon as you run a query.

To disable auto-suspend, select the Disable Auto-suspend option. You can disable the auto-suspend option both while creating a Virtual Warehouse or after creating it.

Configuring auto-scaling

When you create a Virtual Warehouse, you set auto-scaling to increase and decrease resources according to demand. Auto-scaling relinquishes resources when demand decreases to limit unnecessary cloud expenses. Auto-scaling increases resources to speed performance.

Auto-scaling is designed for huge workloads in production, so if you are evaluating CDW, or just learning, simply accept the default values. Later, you can edit the Virtual Warehouse to tune auto-scaling.

Before configuring or tuning auto-scaling, you need to understand the auto-scaling process to make good configuration decisions. As a Hive Virtual Warehouse user, choose the method of auto-scaling and see the relevant configuration subtopic (below):

- Concurrency for BI-type queries
- Isolation for ETL-type queries

To configure auto-scaling in a Hive or Impala Virtual Warehouse, see the topics below:

Configuring Hive VW concurrency auto-scaling

Configuring the Hive Virtual Warehouse to use concurrency auto-scaling is critical for controlling cloud expenses.

Before you begin

- You are familiar with the auto-scaling process.
- You are creating a Hive Virtual Warehouse for running BI-type queries.
- In Cloudera Data Warehouse, you added a Hive Virtual Warehouse, configured the size of the Hive Virtual Warehouse, and configured auto-suspend as described in previous topics.
- You obtained the DWAdmin role.

About this task

In this task, first you select the number of executors for your virtual cluster.

- Minimum executors

  The fewest executors you think you will need to provide sufficient resources for your queries. The number of executors needed for a cloud workload is analogous to the number of nodes needed for an on-premises workload.
- Maximum number

  The maximum number of executors you will need. This setting limits prevents running an infinite number of executors and runaway costs.

Consider the following factors when configuring the minimum and maximum number of executors:

- Number of concurrent queries
- Complexity of queries
- Amount of data scanned by the queries
- Number of queries

Next, you configure when your cluster should scale up and down based on one of the following factors:

- Headroom: The number of available coordinators that trigger auto-scaling. For example, if Desired Free Capacity is set to 1 on an XSMALL-sized Virtual Warehouse, which has 2 executors, when there is less than one free coordinator (2 queries are concurrently executing), the warehouse auto-scales up and an additional executor group is added.
- Wait Time: How long queries wait in the queue to execute. A query is queued if it arrives on HiveServer and no coordinator is available. For example, if WaitTime Seconds is set to 10, queries are waiting to execute in the queue for 10 seconds. The warehouse auto-scales up and adds an additional executor group.

Auto-scaling based on wait time is not as predictable as auto-scaling based on headroom. The scaling based on wait time might react to non-scalable factors to spin up clusters. For example, query wait times might increase due to inefficient queries, not query volume.

When headroom or wait time thresholds are exceeded, the Virtual Warehouse adds executor groups until the maximum setting for Minimum and Maximum has been reached.

Procedure

1. In Concurrency Autoscaling properties, limit auto-scaling by setting the minimum and maximum number of executor nodes that can be added.

2. Choose when your cluster auto-scales up based on one of the following choices:

   - Headroom
   - Wait Time

3. Click Apply Changes.

Configuring Hive query isolation auto-scaling

You can configure your Hive Virtual Warehouse to add dedicated executors to run scan-heavy, data-intensive queries, also known as ETL queries. You learn how to enable a Virtual Warehouse auto-scaling feature and set a query isolation parameter in the Hive configuration.

Before you begin

- You are familiar with the auto-scaling process.
- You are creating a Hive Virtual Warehouse for running ETL-type queries.
- In Cloudera Data Warehouse, you added a Hive Virtual Warehouse, configured the size of the Hive Virtual Warehouse, and configured auto-suspend as described in previous topics.
- You obtained the DWAdmin role.

Note: You can enable query isolation only while creating a Virtual Warehouse. On existing warehouses, you can only disable the query isolation option.

About this task

In this task, first you configure the same auto-scaling properties as described in the previous topic Configuring Hive VW concurrency auto-scaling, and then you enable query isolation. Next, you set the hive.query.isolation.scan.size.threshold configuration parameter.
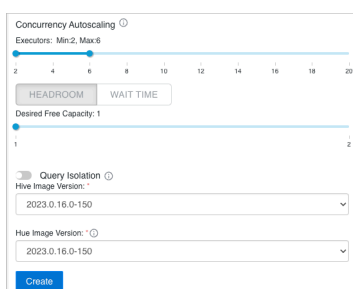
Procedure

1. In Concurrency Autoscaling properties, limit auto-scaling by setting the minimum and maximum number of executor nodes that can be added.

2. Choose when your cluster auto-scales up based on one of the following choices:
   - HEADROOM
   - WAIT TIME

3. In Concurrency Autoscaling, select Enable Query Isolation.

   Two configuration options appear: Max Concurrent Isolated Queries and Max Nodes Per Isolated Query.

4. In Max Concurrent Isolated Queries, set the maximum number of isolated queries that can run concurrently in their own dedicated executor nodes.

   Select this number based on the scan size of the data for your average scan-heavy, data-intensive query.

5. In Max Nodes Per Isolated Query, set how many executor nodes can be spawned for each isolated query.

6. Click Create Virtual Warehouse.

7. In the Data Warehouse service **Overview** page, click Virtual Warehouses, and find your Virtual Warehouse. Click
   ⋮ Edit Configurations Hiveserver2 .

8. Select hive-site from the Configuration files drop-down menu and set hive.query.isolation.scan.size.threshold parameter to limit how much data can be scanned.

   For example, set 400GB.

9. Click Apply Changes.

Configuring Impala VW auto-scaling

Configuring the Impala Virtual Warehouse to use concurrency auto-scaling is critical for controlling cloud expenses.

Before you begin

- You are familiar with the auto-scaling process.
- You are creating a Virtual Warehouse for running BI-type queries.
- In Cloudera Data Warehouse, you added an Impala Virtual Warehouse, configured the size of the Impala Virtual Warehouse, and configured auto-suspend as described in previous topics.

- You obtained the DWAdmin role.

About this task

In this task, you configure the following properties:

- Scale Up Delay: Sets the length of time in seconds to wait before adding more executors if queries wait in the queue.
- Scale Down Delay: Sets the length of time in seconds to wait before removing executors if executor groups are idle.

The time to auto-scale up or down is affected by the underlying Kubernetes configuration.

By default Impala Virtual Warehouses can run 3 large queries per executor group. Executors can handle more queries that are simpler and that do not utilize concurrency on the executor.

.

Procedure

1. In Scale Up Delay, set the seconds to wait before adding more executors if queries wait in the queue.
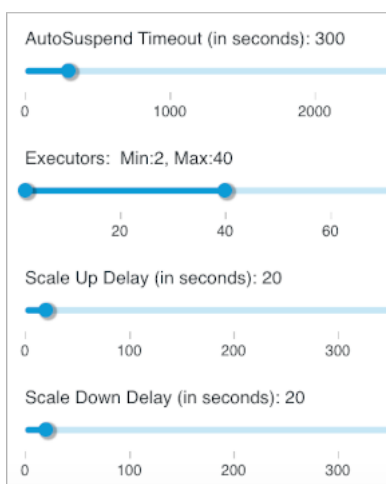
2. In Scale Down Delay, set the seconds to wait before removing executors when executor groups are idle.
3. Click Create Virtual Warehouse.

Configuring Impala coordinator shutdown

To cut cloud expenses, you need to know how to configure Impala coordinators to automatically shutdown during idle periods. You need to know how to prevent unnecessary restarts. Monitoring programs that periodically connect to Impala can cause unnecessary restarts.

About this task

When you create a Virtual Warehouse, you can configure Impala coordinators to automatically shutdown during idle periods. The coordinator start up can last several minutes, so clients connected to the Virtual Warehouse can time out.

Before you begin

- Update impyla, jdbc, impala shell clients if used to connect to Impala.

- When you create a Virtual Warehouse of SQL engine type Impala, and then you enable Unified Analytics, you can configure Impala coordinator shutdown only if you also configure Active-Passive mode.

Procedure

1. Follow instructions for "Creating a Virtual Warehouse".
2. Select a size for the Virtual Warehouse.
3. Do not select the Disable AutoSuspend option.

   The Impala coordinator does not automatically shutdown unless the Impala executors are suspended.

4. Select the Allow Shutdown of Coordinator option.

After Impala executors have been suspended, the Impala coordinator waits for the time period specified by the Trigger Shutdown Delay before shutting down.

For example, if AutoSuspend Timeout = 300 seconds and Trigger Shutdown Delay=150 seconds, after 300 seconds of inactivity Impala executors suspend, and then 150 seconds later, the Impala coordinator shuts down.

5. Accept default values for other settings, or change the values to suit your use case, and click Create Virtual Warehouse.

Click the tooltip ⓘ for information about a setting.

Configuring Impala coordinator high availability

A single Impala coordinator might not handle the number of concurrent queries you want to run or provide the memory your queries require. You can configure multiple active coordinators to resolve or mitigate these problems. You can change the number of active coordinators later.

About this task

You can configure up to five active-active Impala coordinators to run in an Impala Virtual Warehouse. When you create an Impala Virtual Warehouse, CDW provides you an option to configure Impala coordinator and Database Catalog high availability, described in the next topic. You can choose one of the following options:

Disabled

       Disables Impala coordinator and Database Catalog high availability

Active-passive

       Runs multiple coordinators (one active, one passive) and Database Catalogs (one active, one passive)

Active-active

Runs multiple coordinators (both active) and Database Catalogs (one active, one passive)

By using two coordinators in an active-passive mode, one coordinator is active at a time. If one coordinator goes down, the passive coordinator becomes active.

If you select the Impala coordinators to be in an active-active mode, the client software uses a cookie to keep a virtual connection to a particular coordinator. When a coordinator disappears for some reason, perhaps due to a coordinator shutting down, then the client software may print the error "Invalid session id" before it automatically reconnects to a new coordinator.

Using active-active coordinators, you can have up to five coordinators running concurrently in active-active mode with a cookie-based load-balancing.

An Impala Web UI is available for each coordinator which you can use for troubleshooting purposes.

Clients who connect to your Impala Virtual Warehouse using multiple coordinators must use the latest Impala shell. The following procedure covers these tasks.

Procedure

1. Follow instructions for "Creating Virtual Warehouse".
2. Select the number of executors you need from the Size dropdown menu.

   A number of additional options are displayed, including High availability (HA).

3. Select the Enabled (Active-Active) option from the High availability (HA) drop-down menu.
4. Select the number of coordinators you need from the Number of Active Coordinators drop-down menu ranging from 2 to 5.
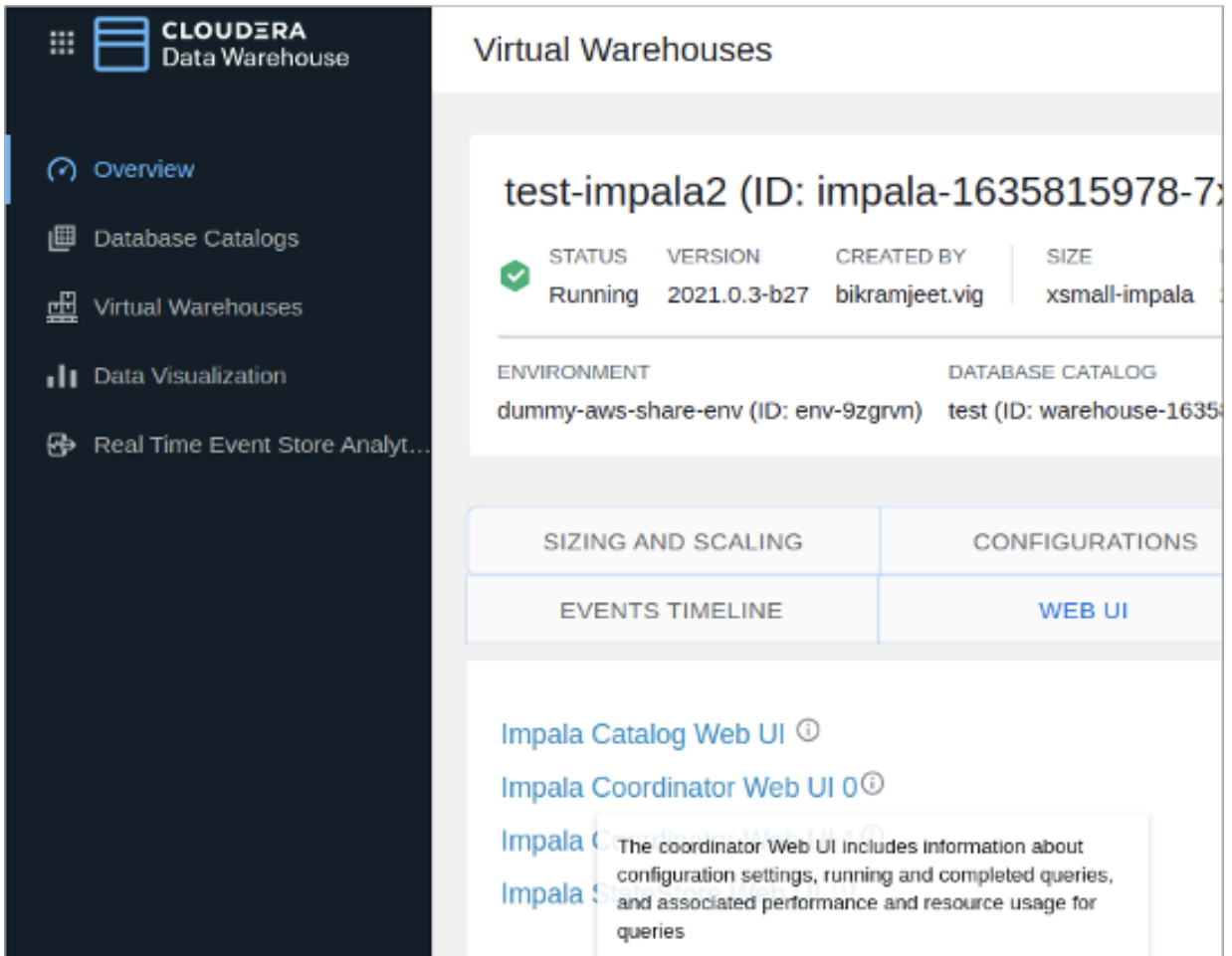
   You can edit an existing Impala Virtual Warehouse to change the number of active coordinators.

   ⚠ **Important:** Do not decrease the number of active-active coordinators you set up initially; otherwise, the Virtual Warehouse may shut down immediately. If clients are running queries on the Virtual Warehouse, the queries could fail.

5. Change values for other settings as needed, click Create Virtual Warehouse, and wait for the Impala Virtual Warehouse to be in the running state.

   Click ⓘ to learn more about the setting.

6.
Go to Cloudera Data Warehouse Overview Impala Virtual Warehouse ⋮ Edit Web UI , and then click each Impala Coordinator Web UI *n* link to get information about the coordinator.

7.
Go to Overview Impala Virtual Warehouse ⋮   and select the Copy Impala shell Download command option.

The following command is copied to your clipboard:

```
pip install impala-shell==4.1.0
```

8. Provide the command to clients who want to connect to the Impala Virtual Warehouse with multiple coordinators using the Impala shell.

9. Instruct the client user to update impyla to version compatible with CDW, as listed in Version information for Cloudera Data Warehouse Public Cloud components.

For example, installing/updating impyla 0.18a2, is required to connect to your Virtual Warehouse active-active coordinators in CDW 2021.0.3-b27.

10. Inform the client that to connect over ODBC to an HA-configured Impala Virtual Warehouse that uses active-active coordinators, you must append impala.session.id to the HTTPAuthCookies connector configuration option of the Cloudera ODBC driver.

Table 1: HTTPAuthCookies

Key Name	Value	Required
HTTPAuthCookies	impala.auth,JSESSIONID,KNOXSESSIONID,impala.session.id	Yes

Configuring Impala catalog high availability

By default, the Impala Virtual Warehouse runs a single instance of the catalog. The catalog stores databases, tables, resource usage information, configuration settings, and other objects managed by Impala. You can optionally configure running an additional Impala catalog instance. One catalog instance operates in active mode, the other in passive mode. The passive instance serves as a backup and takes over if the active instance goes down.

About this task

You enable catalog high availability when you create a New Virtual Warehouse. You cannot turn on, but you can turn off catalog high availability after creating an Impala Virtual Warehouse.

When you create an Impala Virtual Warehouse, you use the same UI dropdown to configure Impala coordinator, covered in the previous topic, and Database Catalog high availability. You can choose one of the following options:

- Disabled

  Disables Impala coordinator and Database Catalog high availability

- Active-passive

  Runs multiple coordinators (one active, one passive) and Database Catalogs (one active, one passive)

- Active-active coordinators

  Runs multiple coordinators (both active) and Database Catalogs (one active, one passive)

Before you begin
You must obtain the DWAdmin role.

Procedure

1. Follow instructions for Adding a new Virtual Warehouse.
2. In Size, select the number of executors, for example xsmall-2Executors.

   A number of additional options appear, including High availability.

3. In High availability (HA), accept the default Enabled (Active-Passive) or Enabled (Active-Active).

   Either option enables Database Catalog high availability in active-passive mode.

4. Accept default values for other settings, or change the values to suit your use case.

Click the tooltip ⓘ for information about settings.

5. Click Create Virtual Warehouse.

Resizing the Hive Virtual Warehouse size

The size of the Hive Virtual Warehouse you select during Virtual Warehouse creation determines the number of executors and concurrent queries the Virtual Warehouse can run. You need to know how to change the size of the Virtual Warehouse upward or downward to tune performance and manage cost.

About this task

You cannot change the size of a Hive Virtual Warehouse, but you can handle incorrect sizing in the following ways.

• You can delete the Virtual Warehouse, and then recreate it in a different size.
• You can change the auto-scaling thresholds to change the effective size of the Virtual Warehouse based on demand. The actual size does not change, but increases or decreases in resources occurs automatically.

This task assumes you have two Virtual Warehouses that you decide are incorrectly sized for some reason. You correct the sizing of one by deleting and recreating the Virtual Warehouse. You correct the effective sizing of the other by changing auto-scaling thresholds.

Before you begin

• You obtained the DWAdmin role.

Procedure

First Virtual Warehouse: Replace this Virtual Warehouse

1. Log in to the CDP web interface, navigate to Data Warehouse Overview , note the name of the Virtual Warehouse you want to modify, and note which Database Catalog it is configured to access.

2. Click the options ⋮ of the Virtual Warehouse you want to delete, and select Delete.

3. Go to the Virtual Warehouses tab and click New Virtual Warehouse.

4. Set up the new Virtual Warehouse:

• Type the same Name for the new Virtual Warehouse as you used for the old Virtual Warehouse.

> **Note:** The fully qualified domain name of your Virtual Warehouse, which includes the Virtual Warehouse name plus the environment name must not exceed 64 characters; otherwise, Hue cannot load.

• In Type, click the SQL engine you prefer: Hive or Impala.
• Select your Database Catalog and User Group if you have been assigned a user group.
• In Size, select the number of executors, for example xsmall-2Executors.
• Accept default values for other settings, or change the values to suit your use case.

5. Click Create Virtual Warehouse.

Second Virtual Warehouse: Change the Auto-Scaling Thresholds

6. In Data Warehouse Overview , click the options ⋮ of the other Virtual Warehouse, a Hive Virtual Warehouse for example, to change auto-scaling thresholds, and select Edit.

7. Go to the Sizing And Scaling tab and in Concurrency Autoscaling, slide the control to change the Max number of executors.

8. Click Apply Changes.