CDH Post-Upgrade 1.0.0

Migrating Impala Workloads to Cloudera Data Warehouse on premises

Date published: 2021-2-22
Date modified: 2023-01-18

CLOUDERA

Legal Notice

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

Overview.. 4

CDH/HDP to Cloudera Base on premises..4

Migration options...4

Side-car migration...4

In-place upgrade with new nodes for Cloudera Data Warehouse Data Service...................................... 5

Moving Impala compute workloads from Cloudera Base on premises to Cloudera Data Warehouse Data Service on premises.................................... 6

Workload selection...7

Steps for migrating Impala workloads to Cloudera Data Warehouse Data Service on premises........................ 7

Activate the Cloudera environment in Cloudera Data Warehouse..7

Setup and size an Impala Virtual Warehouse...8

Import custom configurations...11

Create Virtual Warehouses to implement admission control...12

Modify client connections (JDBC and ODBC)... 12

Reference information...13

Guidelines for moving repeated BI workloads...13

Configuration options available in Cloudera Data Warehouse by default..14

Supported authentication.. 14

Setting up ODBC connection from a BI tool...14

Migrating Impala workloads

Cloudera Data Warehouse on premises offers the latest, updated, and managed elastic Impala experience. Migrating the Impala workloads to Cloudera Data Warehouse enables you to leverage auto-scaling, data sharing, data and result caching, and many other powerful capabilities of the Data Lakehouse architecture.

If you are on CDH or HDP platforms, then you must first upgrade your clusters to Cloudera Base on premises before you can install Cloudera Data Services on premises and migrate your Impala compute workloads to Cloudera Data Warehouse. Compute workloads are SQL programs that get generated from either a shell script, a JDBC/ODBC client, thick SQL clients, or Hue.

Moving certain Impala workloads from Cloudera to Cloudera Data Warehouse has the following advantages:

- Ability to scale out the Impala compute layer independently from the storage layer
- Provides better isolation between workloads to prevent interference and maintain SLAs
- Simplifies the management of Impala clusters, by not having you to focus on admission control and tuning configuration parameters
- Automatic installation, setup, start-up, scale-up, and scale-down of workloads running on a Kubernetes backend
- Ability to run different versions of Impala for different types of workloads (For example, operational reporting can remain locked on a given version, while data science users can always upgrade their Virtual Warehouses to the latest versions)

Migrating Impala workloads from CDH/HDP to Cloudera Base on premises

If you are running Impala workloads on CDH or HDP platforms, then you must first upgrade your cluster to CDP Private Cloud Base and then migrate your Impala workloads. If you are already running Impala workloads on CDP Private Cloud Base, then you can skip this step.

Migration choices to migrate workloads to Cloudera Base on premises

To migrate Impala workloads from CDH to Cloudera Base on premises, you can perform a side-car migration. If your goal is to migrate the workloads to Cloudera Data Warehouse on premises, then perform an in-place upgrade of CDH nodes to Cloudera Base on premises and then migrate the Impala workloads to Cloudera Data Warehouse on premises on new nodes.

Each mechanism has common aspects of work, risk mitigation, and successful outcomes expected across all paths from legacy distributions into Cloudera. Both paths include assessing the workloads, testing, validating, and minimizing workload unavailability during the move. Cloudera recommends the in-place upgrade and migration method to migrate Impala workloads from CDH to Cloudera Data Warehouse on premises.

Side-car migration

The side-car migration mechanism is best employed when you have tighter service-level agreements (SLAs) that preclude an extended, multi-hour downtime for your workloads. This process aims to minimize downtime on individual workloads while providing a straightforward roll-back mechanism on a per-workload basis.

For a side-car migration, you must install and configure a new greenfield Cloudera Base on premises cluster on the second set of hardware, consisting of a few dense storage nodes and several compute nodes. The side-car migration breaks down into the following three major phases:

1. Building and configuring the new Cloudera Base on premises cluster
2. Configuring a replication process to provide periodic and consistent snapshots of data, metadata, and accompanying governance policies

3. Deploying the workloads onto the new cluster, testing them, and flipping them into a production state after validation

After you move the workloads, disable them on the legacy cluster. You will temporarily have production workloads running across multiple clusters during the migration period.

Note: For better query performance and more efficient resource utilization by queries and to benefit from the new features, recompute the statistics after migrating your compute workload.

In-place upgrade with new nodes for Cloudera Data Warehouse Data Service

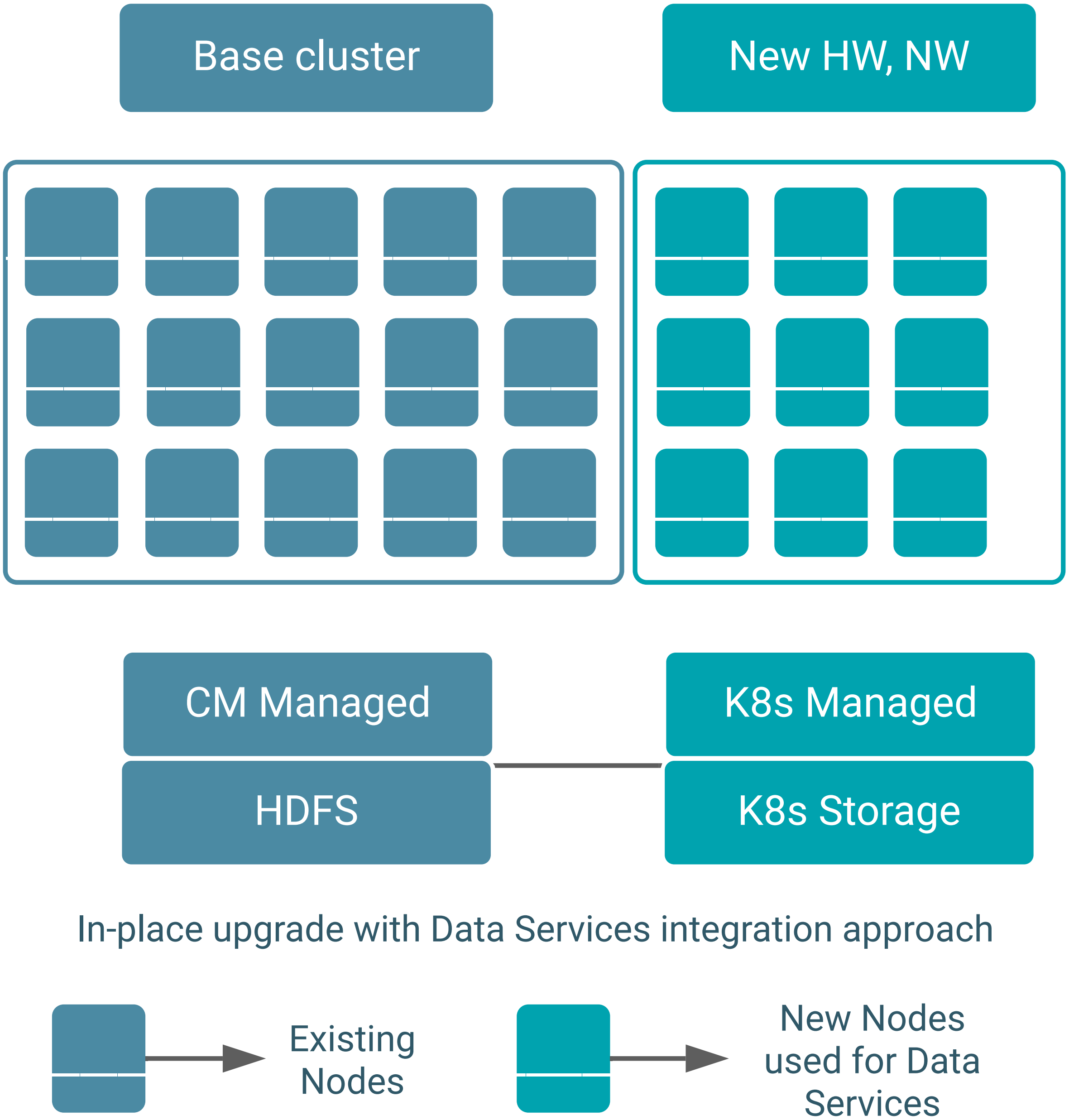
The age and hardware refresh cycle of legacy clusters is an important consideration when deciding on the in-place upgrade strategy. To add new hardware to the base cluster this mechanism is the best choice. Adding new hardware to the base cluster makes it simpler to set up the Cloudera Data Warehouse Data Service, which reduces the time required and lowers risk.

Note: Cloudera recommends this migration path to move from CDH to Cloudera Data Warehouse Data Service on Cloudera on premises.

In-place upgrade and migration is a two-step process. To get to Cloudera Data Warehouse on premises, you must:

- Upgrade an existing CDH cluster to Cloudera Base on premises
- Install Cloudera Data Warehouse Data Service on the new hardware alongside the base cluster and then move your compute workloads from Cloudera Base on premises to Cloudera Data Warehouse Data Service

The following diagram shows running the Cloudera Data Warehouse Data Service on premises on the new hardware alongside the base cluster after an in-place upgrade from CDH. HW stands for hardware and NW stands for new workloads:


Base cluster
New HW, NW
CM Managed
HDFS
K8s Managed
K8s Storage
In-place upgrade with Data Services integration approach
Existing Nodes
New Nodes used for Data Services

To upgrade from CDH to Cloudera Base on premises, see Upgrading CDH 6 to Cloudera Base on premises.

Moving Impala compute workloads from Cloudera Base on premises to Cloudera Data Warehouse Data Service on premises

To migrate the Impala compute workloads to Cloudera Data Warehouse Data Service, you must have completed an in-place upgrade from CDH to Cloudera Base on premises. Review how to identify the workloads to migrate and configuration changes that are needed between Cloudera Base on premises and Cloudera Data Warehouse.

Workload selection

Depending on the size (CPU, Memory, Cache) of your Cloudera Data Warehouse environment, you must choose which workloads should be migrated from base Impala to Cloudera Data Warehouse Impala.

Business Intelligence workloads

If you have teams of data analysts using Business Intelligence (BI) tools and applications, then you can migrate to Cloudera Data Warehouse Data Service on premises to benefit from data and result caching. If you have repeated BI workloads, see Guidelines for moving repeated BI workloads.

Pioneering user base

New features are developed and released in Cloudera Data Warehouse first. If your user base needs to use newer technologies such as Iceberg, Unified Analytics, Data and Result Caching, and so on, then you can consider moving your workloads to Cloudera Data Warehouse Data Service on premises.

Performance-demanding workloads

If your teams are running complex and large queries that require high memory and run for a long time that cause other SLA-driven workloads to be impacted, then you can isolate them out from the Impala service on the base cluster and run them from dedicated Impala Virtual Warehouses to achieve compute isolation.

Kudu workloads

If your workloads depend on streaming datasets that are inserted, updated, or deleted frequently, then you must configure Kudu on the base cluster and you can use Impala to query Kudu on base. Cloudera Data Warehouse supports creating Impala tables in Kudu. For more information, see Configuring Impala Virtual Warehouses to create Impala tables in Kudu in Cloudera Data Warehouse on premises.

Anti-pattern workloads

If your workloads scan extremely large datasets, then you need to consider data locality before moving the workloads from the Impala service on the base cluster to Impala in Cloudera Data Warehouse Data Service. Scanning large data sets over the network can take time. If you plan to move such workloads to Cloudera Data Warehouse, then ensure that the data can be cached for better performance. You can also increase the auto-suspend time for the Impala Virtual Warehouse to make sure that the cache is retained.

Other things that you must consider while assessing the workloads are tenant modeling and sizing followed by performance testing. This ensures that queries return correct results and that performance is acceptable.

Steps for migrating Impala workloads to Cloudera Data Warehouse Data Service on premises

After you upgrade from CDH to Cloudera Base on premises you can retain your compute workload on Cloudera Base on premises and continue to use Impala as your query engine. To get to the latest Impala version with the full feature set, the ideal choice available to you is to move your compute workload from Cloudera Base on premises to Cloudera Data Warehouse on premises

Before you begin, acquire basic information about the Cloudera platform and the interfaces.

Activate the Cloudera environment in Cloudera Data Warehouse

Before you can create a Database Catalog to use with a Virtual Warehouse, you must activate a Cloudera environment in Cloudera Data Warehouse. Activating an environment causes Cloudera to connect to the Kubernetes cluster, which provides the computing resources for the Database Catalog. In addition, activating an environment enables the

Cloudera Data Warehouse service to use the existing data lake that was set up for the environment, including all data, metadata, and security.

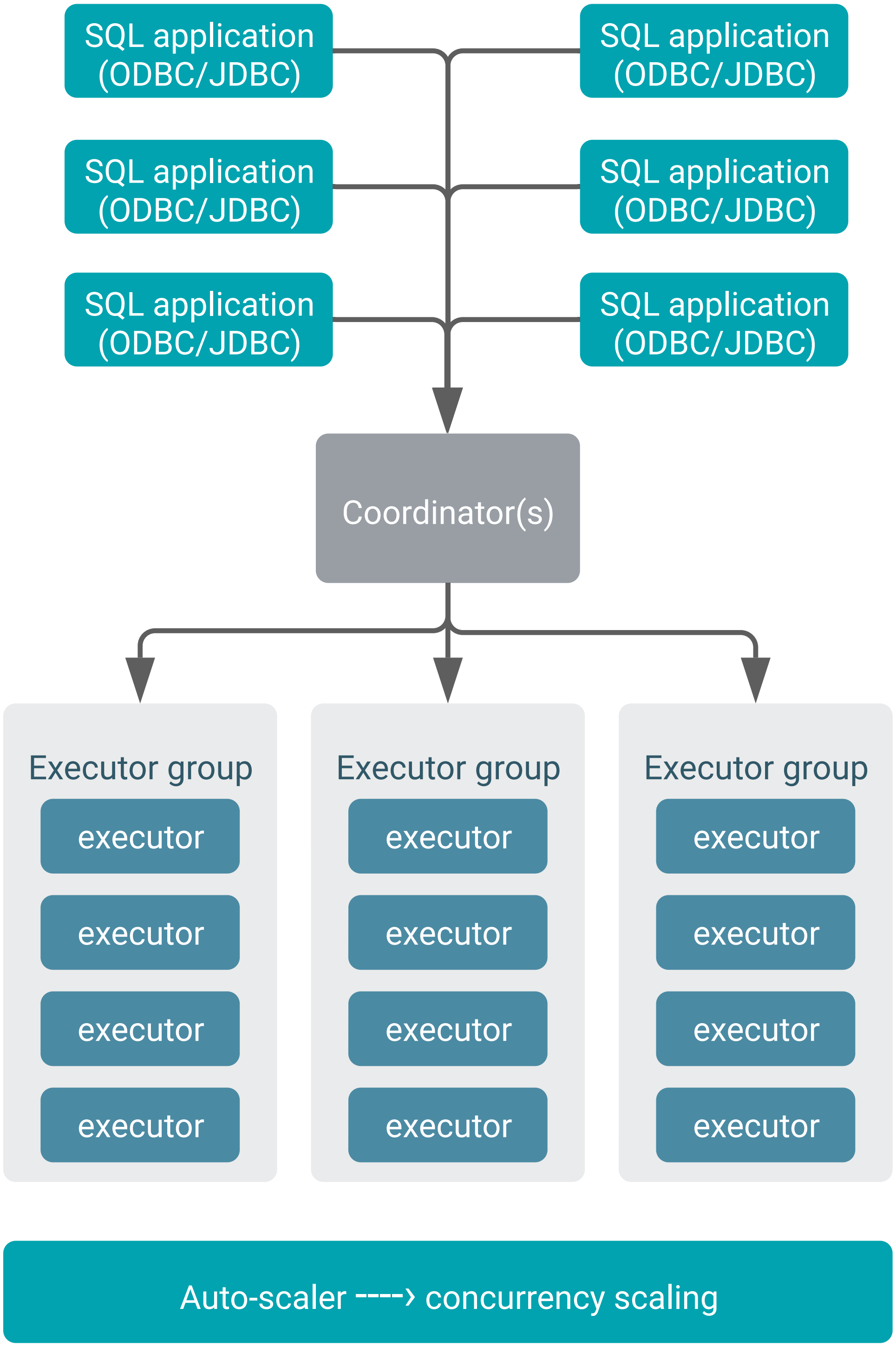
To activate an environment, see Activating OpenShift and Embedded Container Service environments. Also, review the sizing information to plan resources for your environment. See How to use the Cloudera Data Services on premises sizing spreadsheet.

Setup and size an Impala Virtual Warehouse

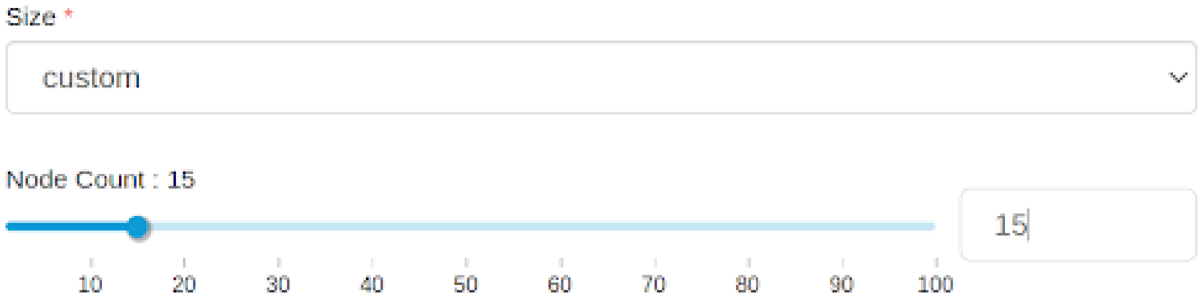
After activating the environment, you must create and configure an Impala Virtual Warehouse. Learn about the configurations and options that Cloudera Data Warehouse (CDW) provides to size the Virtual Warehouse.

Size

The size of a Virtual Warehouse controls the “Executor group” size as shown in the following diagram. This controls the number of Impala executor pods that a single query runs on. Select a size that is appropriate for the amount of data to be scanned or aggregated and the amount of data cache that is needed overall. By default, each executor pod and coordinator pod caches 300 GB data and scratches 300 GB data.


SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
Coordinator(s)
Executor group
executor
executor
executor
executor
Executor group
executor
executor
executor
executor
Executor group
executor
executor
executor
executor
Auto-scaler ----› concurrency scaling

You can select a standard size (2, 10, 20, or 40, set a custom size of, say, 15 nodes in an executor group) as shown in the following image:


Size *

custom

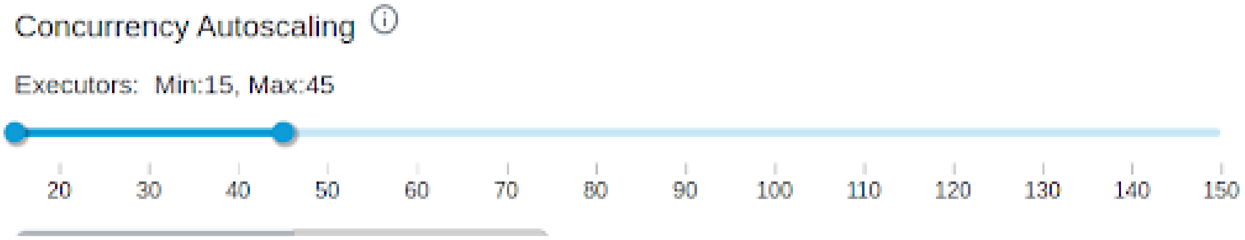
Node Count : 15

15

Tip: If your query is running slow and scans take a lot of time, then increase the size of the executor group.

Concurrency-based auto-scaling

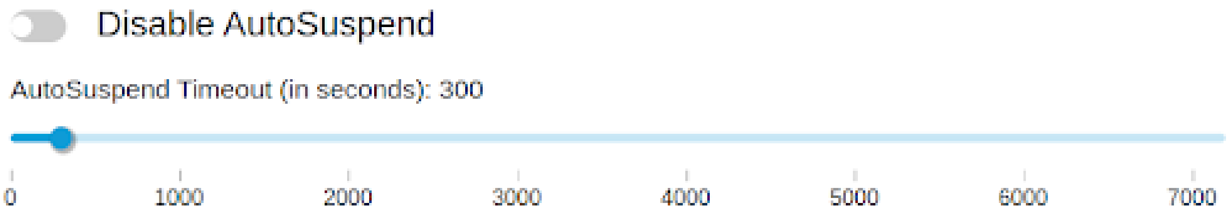
Once you have selected the executor group size, concurrency-based autoscaling allows the Virtual Warehouse to scale up when concurrent queries are seen. This triggers scales up the Virtual Warehouse size to run the concurrent query. You can scale up the Virtual Warehouse in proportion to its size. For example, if you set the Virtual Warehouse size equal to 15 nodes, then the Virtual Warehouse scales up in multiples of 15 nodes, namely, 15, 30, 45, 60, and so on. You can control the scale up ratio by controlling the maximum number of executors as shown in the following image:


Concurrency Autoscaling

Executors: Min:15, Max:45

Auto-suspend a Virtual Warehouse

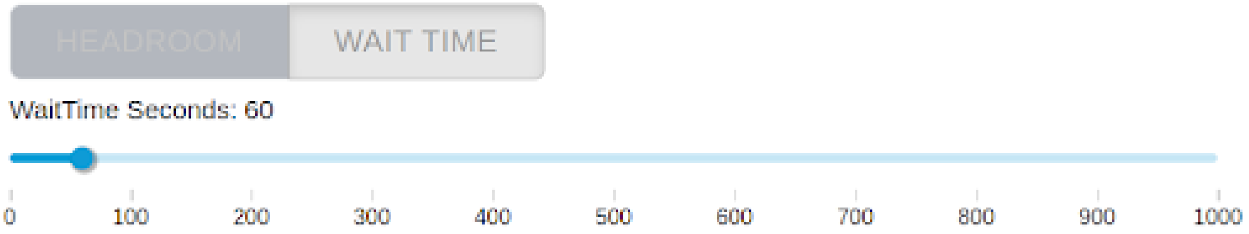
Virtual Warehouses automatically stop after a period of inactivity. You can disable it or set the amount of time before the Virtual Warehouse suspends itself as shown in the following image:


Disable AutoSuspend

AutoSuspend Timeout (in seconds): 300

Wait time

Wait time is used to set the amount of time a query stays in a queue before a new executor group is spawned, as shown in the following image:



WaitTime Seconds: 60

Import custom configurations

You must import custom configurations that you used for optimizing performance, mitigating the pressure on the network, avoiding resource usage spikes and out-of-memory conditions, to keep long-running queries or idle sessions from tying up cluster resources, and so on.

About this task

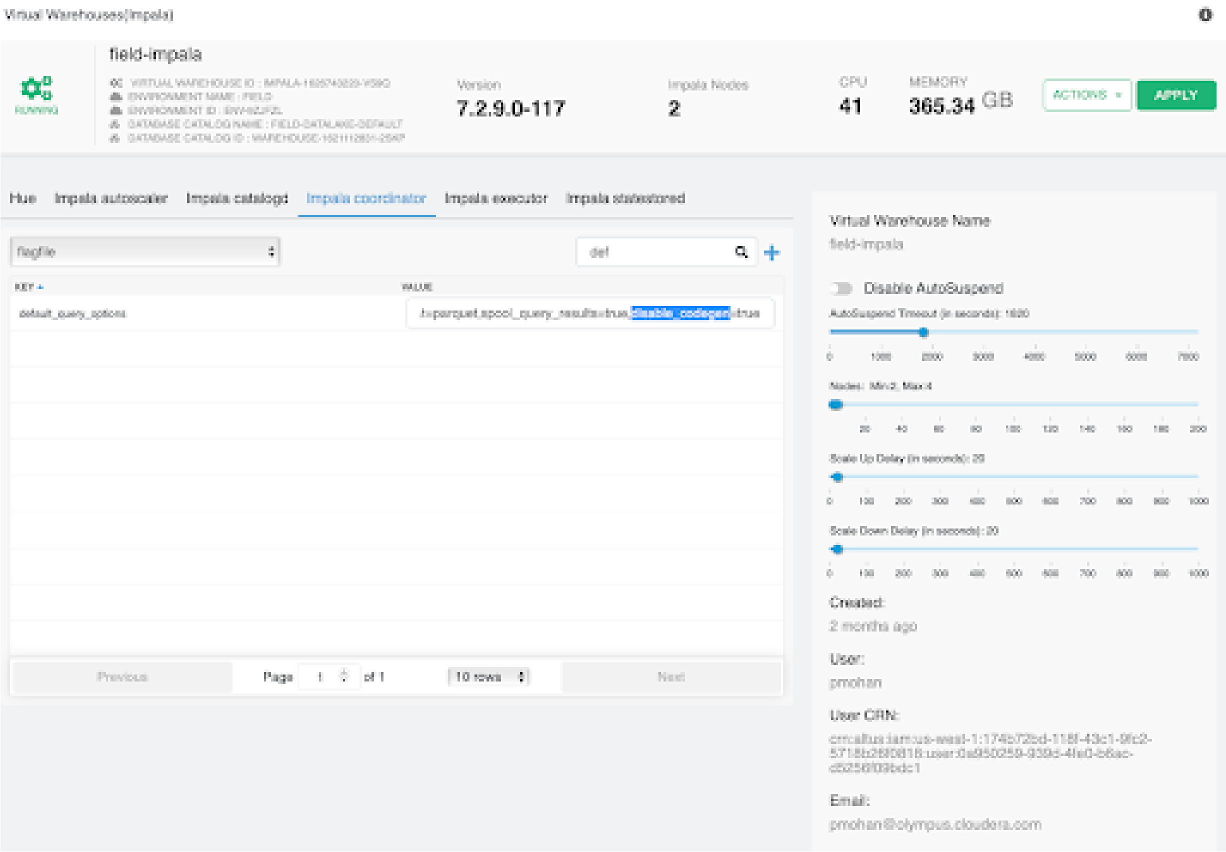
Following is a list of configurations that are typically modified and should be copied to the Cloudera Data Warehouse Data Service from the base cluster:

- default_file_format and default_transactional_type
- CONVERT_LEGACY_HIVE_PARQUET_UTC_TIMESTAMPS and USE_LOCAL_TZ_FOR_UNIX_TIMESTAMP_CONVERSIONS (especially for older datasets to avoid Hive vs Impala timestamp issues)
- Runtime filters for performance: RUNTIME_BLOOM_FILTER_SIZE, RUNTIME_FILTER_MAX_SIZE, RUNTIME_FILTER_MIN_SIZE, RUNTIME_FILTER_MODE, and RUNTIME_FILTER_WAIT_TIME_MS
- PARQUET_FALLBACK_SCHEMA_RESOLUTION (to handle any column order changes)
- TIMEZONE (because CDW uses the UTC timezoone)
- COMPRESSION_CODEC (based on your need; LZ4 is recommended)
- SCHEDULE_RANDOM_REPLICA and REPLICA_PREFERENCE (to avoid hotspotting)
- EXEC_TIME_LIMIT_S, IDLE_SESSION_TIMEOUT, and QUERY_TIMEOUT_S
- DISABLE_CODEGEN_ROWS_THRESHOLD, EXEC_SINGLE_NODE_ROWS_THRESHOLD, and BROADCAST_BYTES_LIMIT
- SCRATCH_LIMIT if set to higher on CDP Private Cloud Base needs to be limited to a value between 300 and 600 GB
- Enable ALLOW_ERASURE_CODED_FILES and disable DECIMAL_V2 only if needed
- MEM_LIMIT (automatically selected between a range of 2 G to 50 G)
  - You can set MEM_LIMIT in the default query options, if needed
  - Administrators can tweak the Impala's admission control configurations to set a reasonable range for minimum and maximum query memory limit

Procedure

1. Log in to the Cloudera Data Warehouse service as a DWAdmin.
2. Select the Impala Virtualk Warehouse that you want to modify and click ⋮ Edit CONFIGURATIONS Impala Coordinator and select flagfile from the drop-down list.

3. Update the configurations in the default_query_options field as shown in the following image:



Virtual Warehouses(Impala)

field-impala

VIRTUAL WAREHOUSE ID : IMPALA-1626743829-V59Q
ENVIRONMENT NAME : FIELD
ENVIRONMENT ID : ENV-9ZJFZL
DATABASE CATALOG NAME : FIELD-DATALAKE-DEFAULT
DATABASE CATALOG ID : WAREHOUSE-1621112831-2SKP

Version: 7.2.9.0-117
Impala Nodes: 2
CPU: 41
MEMORY: 365.34 GB

flagfile

def

KEY	VALUE
default_query_options	...=parquet,spool_query_results=true,disable_codegen=true

Virtual Warehouse Name: field-impala

Disable AutoSuspend

AutoSuspend Timeout (in seconds): 1800

Nodes: Min:2, Max:4

Scale Up Delay (in seconds): 20

Scale Down Delay (in seconds) : 20

Created: 2 months ago

User: pmohan

User CRN: crn:altus:iam:us-west-1:174b72bd-118f-43c1-9fc2-5718b2680818:user:0a950259-939d-4fe0-b6ac-d5256f09bdc1

Email: pmohan@olympus.cloudera.com

Similarly, make the required changes for the Impala executor (Impala executor flagfile default_query_options .

4. Click APPLY.
5. Restart the Virtual Warehouse.

Related Information

Admission Control Architecture for Cloudera

Managing Resources in Impala

Create Virtual Warehouses to implement admission control

There is only one queue in Cloudera Data Warehouse's Virtual Warehouses that is used to run all queries in a first-in-first-out (FIFO) order. If you have multiple queues on the base cluster that you are migrating to Cloudera Data Warehouse Impala, then you must create a Virtual Warehouse for each queue or a request pool, so that you can isolate the compute environments for each of those user groups.

Related Information

Adding a new Virtual Warehouse

Modify client connections (JDBC and ODBC)

Impala in Cloudera Data Warehouse provides a JDBC/ODBC endpoint that can be used to configure all BI tools and applications for each Virtual Warehouse. In Cloudera Data Warehouse, different user groups will likely get their own Virtual Warehouse, each of which has its own unique JDBC/ODBC URL. Make sure that you point your BI clients to the corresponding URL.

In the CDH environment, you had access to one monolithic Impala cluster and you used one JDBC/ODBC URL to connect your Impala clients to the cluster. All your client applications used that one URL. In Cloudera Data Warehouse, you must direct your individual client applications to their own Virtual Warehouse JDBC/ODBC URLs. The URL for each Virtual Warehouse is unique. Therefore, you can recreate a Virtual Warehouse with the same name so that the URL remains the same.

Following is a list of changes for the clients between the base cluster and Cloudera Data Warehouse Data Service on premises:

- Impala in Cloudera Data Warehouse uses the port 443 and communicates over the http protocol, whereas, Impala on the base cluster uses the binary 21050 protocol.
- Impala in Cloudera Data Warehouse uses LDAP authentication, whereas, Impala on the base cluster uses Kerberos and Knox. Therefore, you must specify a username and password.

  Kerberos authentication will be available in Cloudera Data Warehouse Data Service on premises in a future release.
- Impala in Cloudera Data Warehouse uses the latest Simba drivers. Download the latest JDBC driver from the Cloudera Downloads page, or alternatively, on the Virtual Warehouses page. Simba drivers are backward-compatible.
- For granting access to the Impala-shell, click ⋮ Copy Impala shell command on the Virtual Warehouse tile to copy the command line for invocating the impala-shell. This command line contains all the required parameters, including the default user ID for the Virtual Warehouse (this is the Virtual Warehouse's admin user) as required for LDAP, and the network address of the Impala coordinator.

  The Impala-shell command line in Cloudera Data Warehouse contains the following parameters that are not present in the Impala-shell command line on the base cluster:

  - --protocol='hs2-http': The HiveServer2 protocol that impala-shell uses to speak to the Impala daemon
  - --ssl: Used to enable TLS/SSL for the Impala-shell. This parameter is always present in the command line.
  - Uses the 443 port by changing the endpoint to the Virtual Warehouse's coordinator endpoint inside the Virtual Warehouse's coordinator pod. For example: coordinator-default-impala.dw-demo.ylcu-atmi.cloudera.site:443
  - -l: Used to specify LDAP user name and password for LDAP-based authentication instead of Kerberos. LDAP authentication uses the LDAP server defined on the **Authentication** page of the Cloudera Management Consoleon premises.

  Following is a sample Impala shell command:

  ```
  impala-shell --protocol='hs2-http' --strict_hs2_protocol --ssl -i coordi
  nator-default-impala.dw-demo.ylcu-atmi.cloudera.site:443' -u [***USERNAM
  E***] -l
  ```

Related Information

Setting up ODBC connection from a BI tool

Reference information

Review guidelines and related information while migrating Impala workloads to Cloudera Data Warehouse Data Service on premises.

Guidelines for moving repeated BI workloads

Review the guidelines for moving Repeated BI workloads.

- Avoid selecting workloads that download data.
- Ensure that the datasets are compressed Parquet tables and not in text or ORC format.
- Select workloads that are an output of BI queries submitted from any BI tools and repeated multiple times.

  You can find the repeated queries under simple query profile analysis.
- A small to medium-sized VW is recommended for BI workloads.
- Redirect the BI tool to point to Cloudera Data Warehouse for the workloads.
- Check that the performance of the queries in Cloudera Data Warehouse is faster.

- Set proper values for run-time filters, EXEC_TIME_LIMIT_S property, and so on as recommended for TPC-DS queries.
- Set query timeouts and session timeouts to avoid non-responsive queries.
- Look at the peak spilled metrics in the query profiles and depending on the value increase the data cache size and lower the scratch space (SCRATCH_LIMIT) to cache more data for better performance.
- Increase the value of the EXEC_SINGLE_NODE_ROWS_THRESHOLD property to at least 5000 for small query optimization.
- Schedule weekly statistic data collection for better query performance.

Configuration options available in Cloudera Data Warehouse by default

Certain customization that you embraced in a CDH environment need not be imported to Cloudera Data Warehouse, due to its containerized and compute-isolated architecture and available default configurations.

The following lis a ist of customization options that are set by default in Cloudera Data Warehouse and need not be configured:

- Admission control in Cloudera Data Warehouse is tuned by default because of tenant isolation using multiple Virtual Warehouses, and auto-scaling.
- Impala multithreading and remote data cache settings can be ignored because Cloudera Data Warehouse provides them as default configurations.

Supported authentication

In CDH, the list of supported authentication modes for Impala clients included Kerberos. However, Kerberos authentication is not supported for Impala clients in Cloudera Data Warehouse. To connect to Impala that is running in Cloudera Data Warehouse you can use your workload userID and password.

Security change - Apache Knox authentication not supported in Cloudera Data Warehouse

Knox proxy that is generally used to extend the reach of Apache™ Hadoop® services to users outside of a Hadoop cluster is not supported within Cloudera Data Warehouse. Open the ports for on premises with Data Services as a gateway for communicating with AWS services.

Switch to LDAP auth for Impala-shell and Impyla clients

You can use impala-shell and impyla to query Impala Virtual Warehouses in Cloudera Data Warehouse; however, Kerberos is not supported with impala-shell and impyla in Cloudera Data Warehouse. To query Impala Virtual Warehouse, you must switch to LDAP authentication. To switch to LDAP, upgrade impala-shell to the latest version by running the following command: pip2 install impala-shell. You can specify the following connection string when starting impala-shell to control how shell commands are executed. You can specify options on the command line or in the impala-shell configuration file:

```
Impala-shell --protocol='hs2-http' --ssl -i "<impalad=host-name>:443" -u <us
er-name> -l
```

where,

- hs2-http is the HiveServer2 protocol that impala-shell uses to speak to the Impala daemon
- -l enables LDAP authentication
- --ssl enables TLS/SSL for impala-shell

Setting up ODBC connection from a BI tool

Describes how to connect to Impala Virtual Warehouses using ODBC with your BI tool, with Tableau as an example.

Before you begin

Before you can use Tableau with Impala Virtual Warehouses, you must have created a Database Catalog that is populated with data. You have the option to populate your Database Catalog with sample data when you create it. You must also create an Impala Virtual Warehouse, which is configured to connect to the Database Catalog that is populated with data.

Procedure

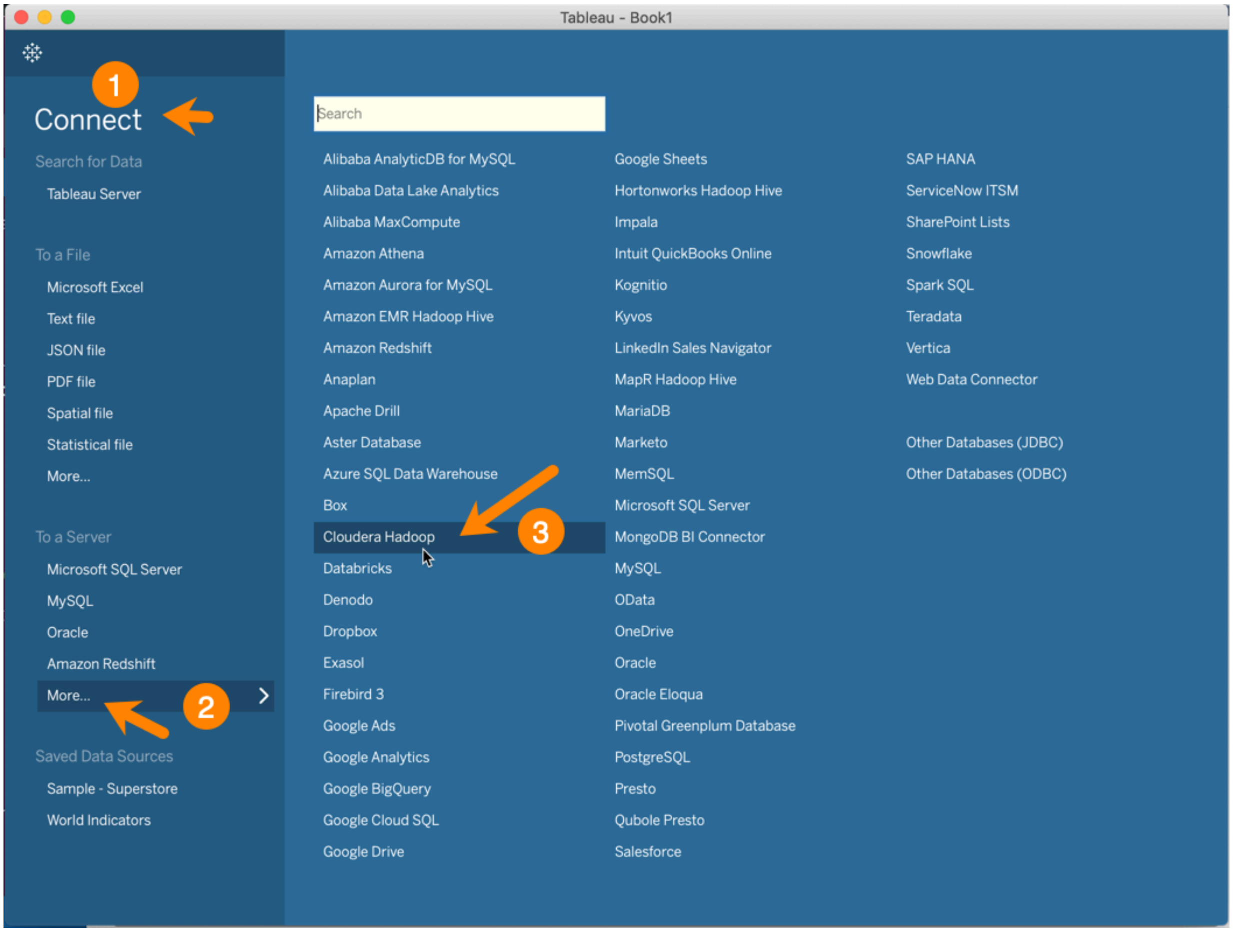
1. Download the latest version of the Impala ODBC driver from Downloads page or alternatively, on the **Virtual Warehouses** page, click the options menu for the warehouse you want to connect to your BI tool, and select Download JDBC/ODBC Driver and install it.
2. Install the driver on the local host where you intend to use Tableau Desktop.
3. Log in to the web interface and navigate to the **Data Warehouse** service.
4. Click Virtual Warehouse in the left navigation panel.
5. On the **Virtual Warehouses** page, click for the Impala warehouse you want to connect to with Tableau, and select Copy JDBC URL:

   This copies the JDBC URL to your system clipboard.
6. Paste the copied JDBC URL into a text file. It should look similar to the following:

   ```
   jdbc:hive2://<your-virtual-warehouse>.<your-environment>.<dwx.company.co
   m>/default;transportMode=http;httpPath=cliservice;ssl=true;retries=3
   ```
7. From the text file where you just pasted the URL, copy the host name from the JDBC URL to your system clipboard. For example, in the URL shown in Step 6, the host name is:

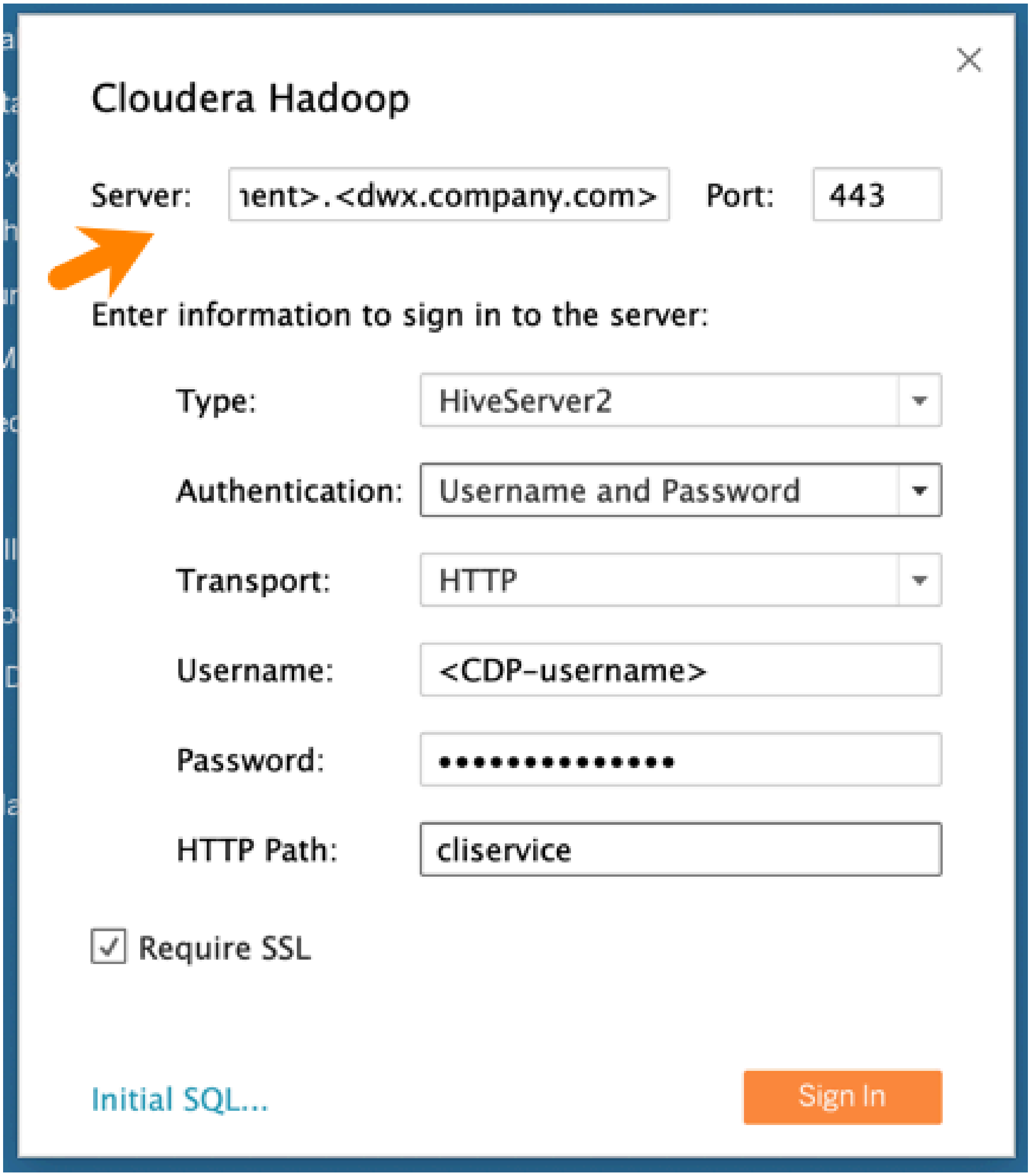
   ```
   <your-virtual-warehouse>.<your-environment>.<dwx.company.com>
   ```

8. Start Tableau and navigate to ConnectMore…Cloudera Hadoop :



This launches the Hadoop dialog box.

9. In the Tableau Hadoop dialog box, paste the host name you copied to your clipboard in Step 7 into the Server field:



Cloudera Hadoop

Server: ɿent>.<dwx.company.com> Port: 443

Enter information to sign in to the server:

Type: HiveServer2

Authentication: Username and Password

Transport: HTTP

Username: <CDP-username>

Password: ••••••••••••••

HTTP Path: cliservice

☑ Require SSL

10. Then in the Tableau Hadoop dialog box, set the following options:

- Port: 443
- Type: HiveServer2
- Authentication: Username and Password
- Transport: HTTP
- Username: Username to connect to the CDP Data Warehouse service.
- Password: Password to connect to the CDP Data Warehouse service.
- HTTP Path: cliservice
- Require SSL: Make sure this is selected.

11. Click Sign In.