Cloudera Data Warehouse Private Cloud 1.5.3

Supporting Cloudera Data Warehosue

Date published: 2020-08-17
Date modified: 2024-03-27

https://docs.cloudera.com/

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

Monitoring Data Warehouse service resources with Grafana dashboards..... 4
Connecting to Grafana dashboards in Private Cloud..... 5
Limitations of Grafana in Cloudera Data Warehouse Private Cloud..... 6

Troubleshooting..... 6
Locate logs..... 6
Hive diagnostic bundles..... 7
Generate diagnostic bundles..... 8
Impala queries fail..... 9
Debug Impala warehouse..... 13
Kerberos authentication failure..... 16

CDP CLI for Data Warehouse..... 17

Runtime documentation..... 17

List of labels for third-party integration..... 18

Monitoring Data Warehouse service resources with Grafana dashboards

Grafana is visualization and analytics software that enables the development of dashboards to monitor metrics data. You can access pre-built Grafana dashboards to monitor Virtual Warehouses and your compute cluster in Cloudera Data Warehouse (CDW).

You connect to prebuilt dashboards to view metrics of CDW operations. Cloudera provides prebuilt Grafana dashboards for Hive, Impala, and Hue dashboards of metrics data, charts, and other visuals.

Using Grafana, CDP metrics are centralized in a single spot, stored in the Prometheus database, and monitored by Prometheus. Your workload databases are not involved in any way. You can immediately view the following pre-built dashboards:

Hive dashboards

The Hive dashboards cover the following operations of the Hive SQL engine in CDW:

- Auto-scaling
- Hive metastore
- HiveServer
- The Hive service itself (Hive-Home)
- LLAP

Impala dashboards

The Impala dashboards include the following operations of the Impala SQL engine in CDW:

- Catalog server
- Coordinator
- Executor
- Statestore
- The Impala service itself

The following screenshot shows the available scratch and cache disk utilization graphs for the Impala Virtual Warehouse:

You can view dashboard metrics for different time periods by selecting the period of interest from the time range dropdown in the horizontal navigation.

On the Embedded Container Service (ECS) platform, you can view the CPU, memory, network usage, and disk input-output for each CDW node using the [***ENVIRONMENT-NAME***]-Nodes option. You can also expand the individual dashboards to see more details, as described in the following table:

Dashboard name	Description	Available metrics
CPU	CPU utilization per node	• Usage per node • Usage per user • Usage per system • Idle time • IO wait
Memory	Memory utilization per node	• Usage per node • Buffer cache • Page cache • Total, used, and available
Network	Number of bytes and packets sent and received	• Network transmitted • Network received • Network transmitted by an interface • Network received by an interface
Disk	Disk bytes read and written	• Bytes written • Bytes read • IO wait time

Note: Node-level metrics are currently available only for ECS environments.

Related Information
Grafana documentation

Connecting to Grafana dashboards in Private Cloud

You can access Grafana from the Management Console and view various dashboards related to the Data Warehouse service.

Before you begin

You must have an activated CDW environment to view dashboards for CDW service resources in Grafana.

Procedure

1. Log in to the Management Console as an Administrator.
2. Go to the **Dashboard** page and click Monitoring Dashboard.
3. In the Grafana web interface, click in the left navigation menu, and select Manage.
   A list of available monitoring dashboards is displayed:

CDW service area	Dashboard topics
CDP Control Plane	Alerts generated by Management Console, pod status, including count, restarts, CPU usage, memory usage, and container memory and CPU usage
Data Warehouse compute auto-scaling	Auto-scaling
Hive	Hive MetaStore (HMS), HiveServer2, Hive service (Hive-Home), and several dashboards for LLAP
Impala	Impala components: catalog server, coordinator, executor, statestore, and the overall Impala service (Impala-Home)
Hue	overall Hue service (Hue-Home)

CDW service area	Dashboard topics
Overview	Kubernetes alerts, pod status, pod CPU usage, pod memory usage, app CPU usage, app memory usage, container memory usage, container CPU usage
Nodes	CPU, memory, network usage and disk IO metrics at the node level for a given environment

Limitations of Grafana in Cloudera Data Warehouse Private Cloud

Learn about the Grafana capabilities in Cloudera Data Warehouse (CDW) that Cloudera does not support. Grafana in CDW is intended for use by cluster operations professionals who are familiar with monitoring tools, interpreting metrics, and performing maintenance.

Supported features

- Viewing and organizing Grafana dashboards.

Unsupported features

Storing metrics longer than 15 days, or consuming more than 90GB of disk space is not supported. Metrics older than 15 days are deleted. If the stored metrics consume more than 90GB of disk space, metrics will be deleted regardless of the number of days stored.

Troubleshooting issues in CDW Private Cloud

Get help and resources for troubleshooting issues in Cloudera Data Warehouse (CDW) on Private Cloud.

Note: The timestamp in the filenames of the files inside a diagnostic bundle is as per the time zone configured in Cloudera Manager. However, if you directly list the files in HDFS, the timestamp in the filenames is as per UTC, but the file contents are as per the time zone configured in Cloudera Manager.

Locating Cloudera Data Warehouse Private Cloud logs

Learn how you can access logs for Cloudera Data Warehouse (CDW) Private Cloud.

About this task

When you generate logs using the Collect Diagnostic Bundles option from the environment, they are written to a partition on the Hive sys.logs table and are stored in the following location on HDFS:

```
/warehouse/[***ENVIRONMENT-NAMEabcde***]/[***DATABASE-CATALOG***]/warehouse/
tablespace/external/hive/sys.db/logs
```

“abcde” is a random 5-character string that is appended to the environment name.

These partitions are retained for 7 days by default.

Procedure

1. Log in to the OpenShift or Experiences Compute Service (ECS) cluster and determine the location of the sys.logs table by running the following query:

   ```
   DESCRIBE FORMATTED sys.logs;
   ```

   This SQL statement returns information about the location of the table which contains the logs.
2. Use the location obtained in Step 1 to locate the CDW Private Cloud logs on the OpenShift or ECS clusters.

Downloading Hive diagnostic bundles in Data Warehouse Private Cloud

You can download diagnostic bundles for troubleshooting a Hive Virtual Warehouse in Cloudera Data Warehouse (CDW) Private Cloud. The diagnostic bundles contain log files for the sidecar containers that support Hive components and for the components themselves. These diagnostic bundles are stored on HDFS in the form of ZIP files.

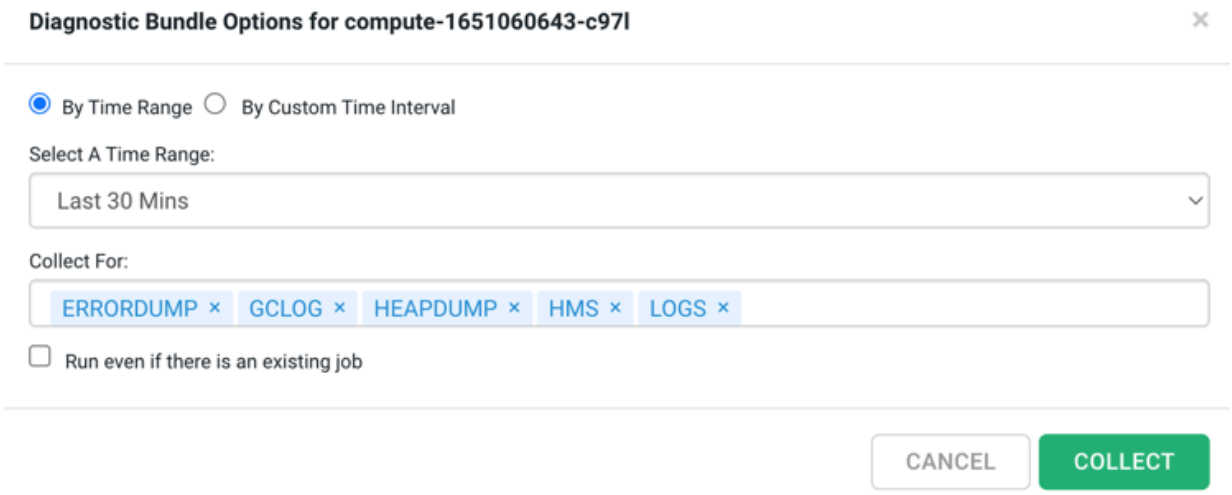
About this task

The log files are generated when you run some workloads on your Hive Virtual Warehouse.

Note: The timestamp in the filenames of the files inside a diagnostic bundle is as per the time zone configured in Cloudera Manager. However, if you directly list the files in HDFS, the timestamp in the filenames is as per UTC, but the file contents are as per the time zone configured in Cloudera Manager.

Procedure

1. Log in to the Data Warehouse service as a DWAdmin.
2. Go to a Hive Virtual Warehouse and click ⋮ Collect Diagnostic Bundle .
   The options for generating the diagnostic bundles are displayed as shown in the following image:



3. Select the time period for which you want to generate the logs.
   - Select the By Time Range option to generate logs from last 30 minutes, one hour, 12 hours, or 24 hours.
   - Select By Custom Time Interval option to generate logs for a specific time period based on your requirement.

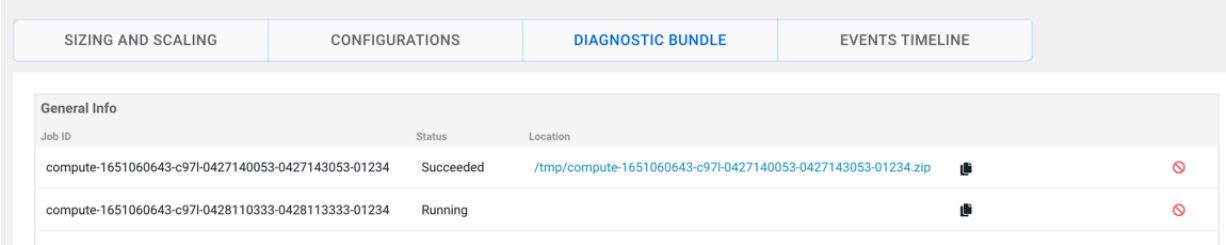
   Note: You must set the time range as per the UTC timezone.
4. Select the categories for which you want to generate the logs by selecting the options from the Collect For section.
   By default, ERRORDUMP, GCLOG, HEAPDUMP, HMS, LOGS, CRINFO, K8S-RESOURCE-INFO are selected. Click X to remove the ones you do not need.
   - ERRORDUMP contains exceptions from the containers
   - CGLOG contains JVM garbage collector-related logs
   - HEAPDUMP contains JVM heapdump
   - HMS contains sidecar container logs that support the metastore
   - LOGS contains logs of Hive, Coordinator, and Executor processes and their supporting containers
5. Select the Run even if there is an existing job option to trigger another diagnostic bundle creation when one job is running.

6. Click Collect.

   The following message is displayed: Collection of Diagnostic Bundle for compute-1651060643-c97l initiated. Please go to details page for more information.

7. Go to the Virtual Warehouses details page by clicking ⋮ Edit .
8. Go to the **DIAGNOSTIC BUNDLE** tab.

   The jobs that have been triggered for generating the diagnostic bundles are displayed, as shown in the following image:



9. Click on the link in the Location column to download the diagnostic bundle to your computer.

Generating and downloading diagnostic bundles

Cloudera Data Warehouse (CDW) collects diagnostic data on workload logs, such as Impala Coordinator, Statefulset, CatalogD logs and stores it in the tmp directory on HDFS. You can download the logs using the Hue File Browser from the base cluster.

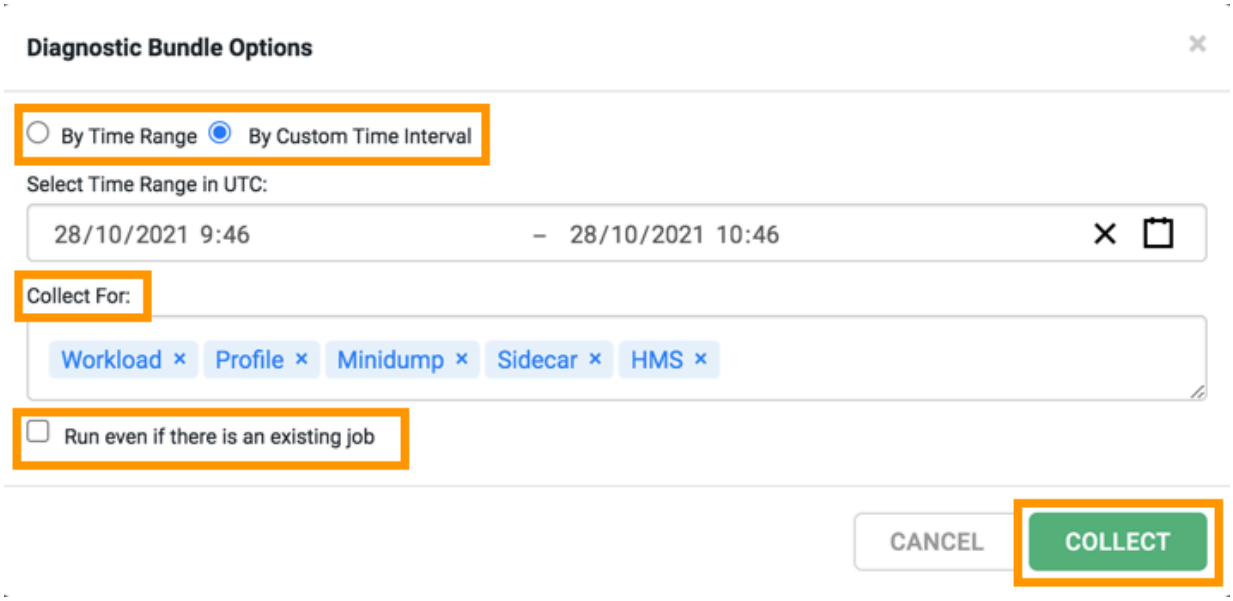
About this task

During the lifetime of a cluster, logs are continously written to the following directory on HDFS: [**WAREHOUSE-DIR**]/warehouse/tablespace/external/hive/sys.db/. When you click Collect Diagnostic Bundle from the CDW web interface, CDW collects the logs for the specified time interval and for the services that you select. These logs are compressed in a ZIP file format and stored in the tmp directory.

Procedure

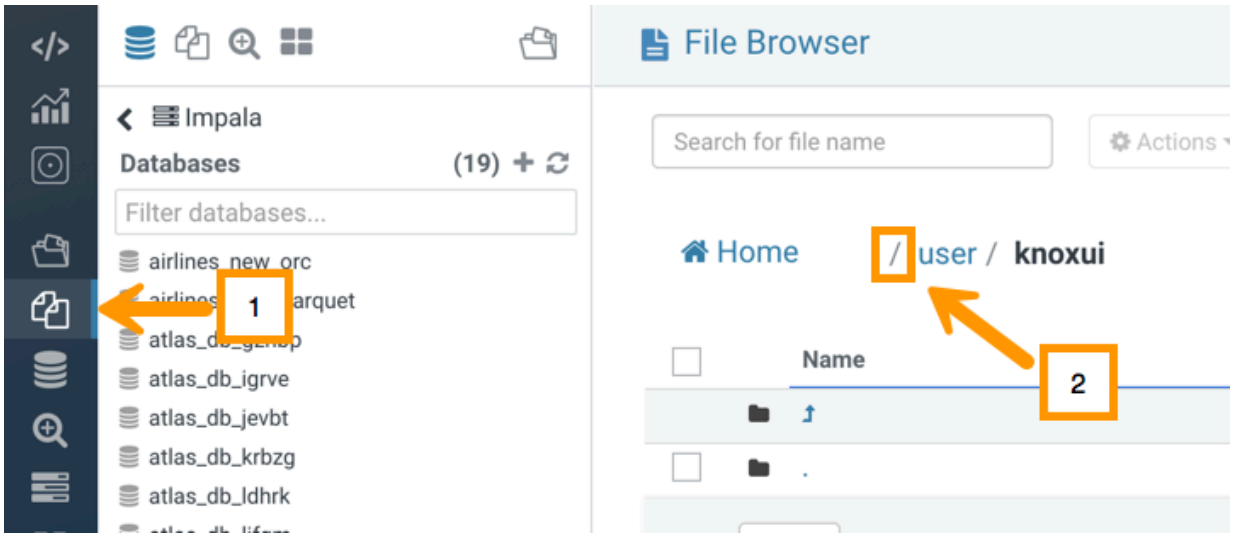
1. Log in to the Data Warehouse service as a DWAdmin.
2. Click ⋮ Collect Diagnostic Bundle on the Virtual Warehouse for which you want to collect the logs .

3. On the **Diagnostic Bundle Options** dialog box, select the time interval and the type of logs you want to collect and click COLLECT.



4. To view the status of the job and to obtain the HDFS location where the logs are stored, select Edit from the Virtual Warehouse options menu and go to the DIAGNOSTIC BUNDLE tab.

   The logs are collected and bundled under the /tmp/[***VIRTUAL-WAREHOUSE-ID-TIMESTAMP***].zip directory.
5. To access and download the logs, open the Hue service from the base cluster.
6. Go to the Hue File Browser and click the forward slash (/) before the user directory as shown in the following image:



The tmp directory is displayed. You can access and download the logs to your computer by clicking Download.

Impala queries fail

Condition

Impala queries running with high concurrency fail on Embedded Container Service (ECS) with the following errors: Invalid or unknown query handle and Invalid session id.

Cause

Impala queries might fail because a single ECS server may not be able to handle the load. To resolve this issue, enable ECS High Availability and increase the ECS server replicas. This process is called promoting the ECS agents to servers. You must promote only one ECS agent at a time. This procedure is explained using an example where you promote the ECS agent on agent1.example.com and then promote the ECS agent on agent2.example.com.

Solution

Procedure

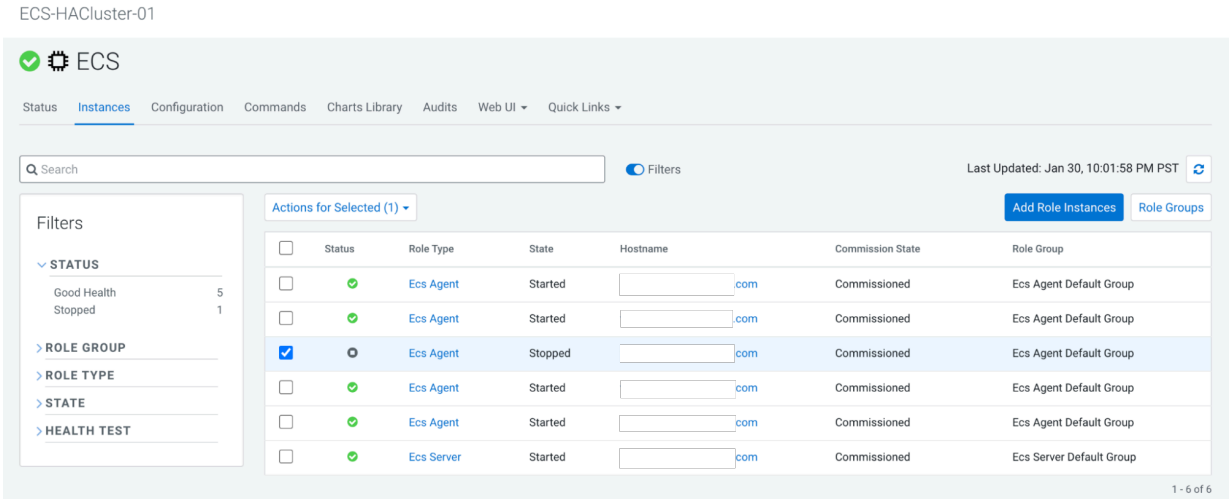
1. Prepare the agent node for promotion by running the following commands on the command line of your ECS server host.

```
sudo /var/lib/rancher/rke2/bin/kubectl --kubeconfig=/etc/rancher/rke2/rke2.yaml get nodes
```

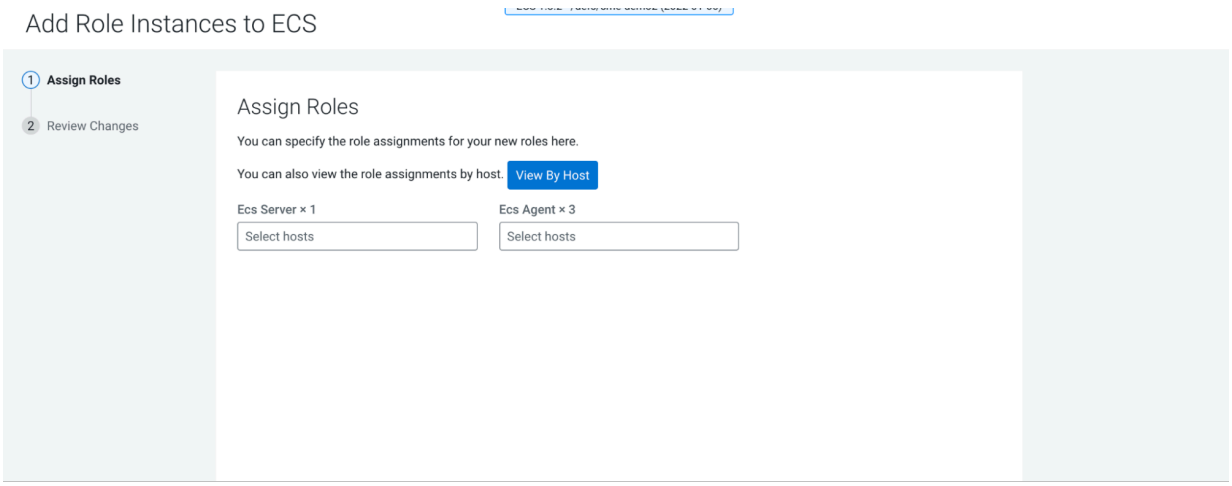
```
sudo /var/lib/rancher/rke2/bin/kubectl --kubeconfig=/etc/rancher/rke2/rke2.yaml drain agent1.example.com --ignore-daemonsets --delete-emptydir-data
```

Note: This may take a few minutes.

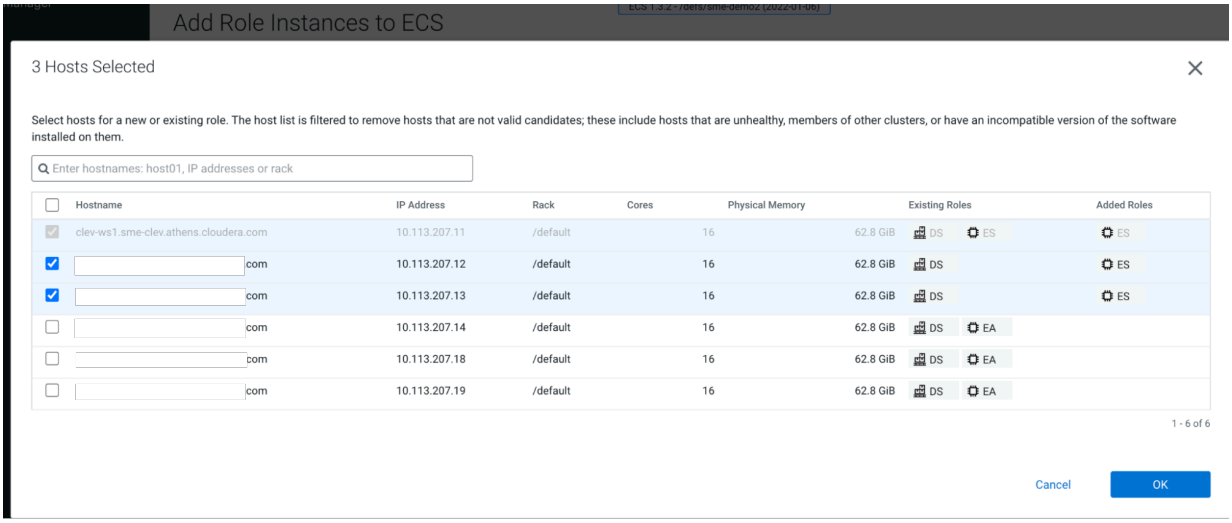
2. In Cloudera Manager, navigate to ECS Cluster ECS . Stop the ECS Agent running on agent1 and then delete the agent by selecting the respective option from the Actions for Selected drop-down menu.

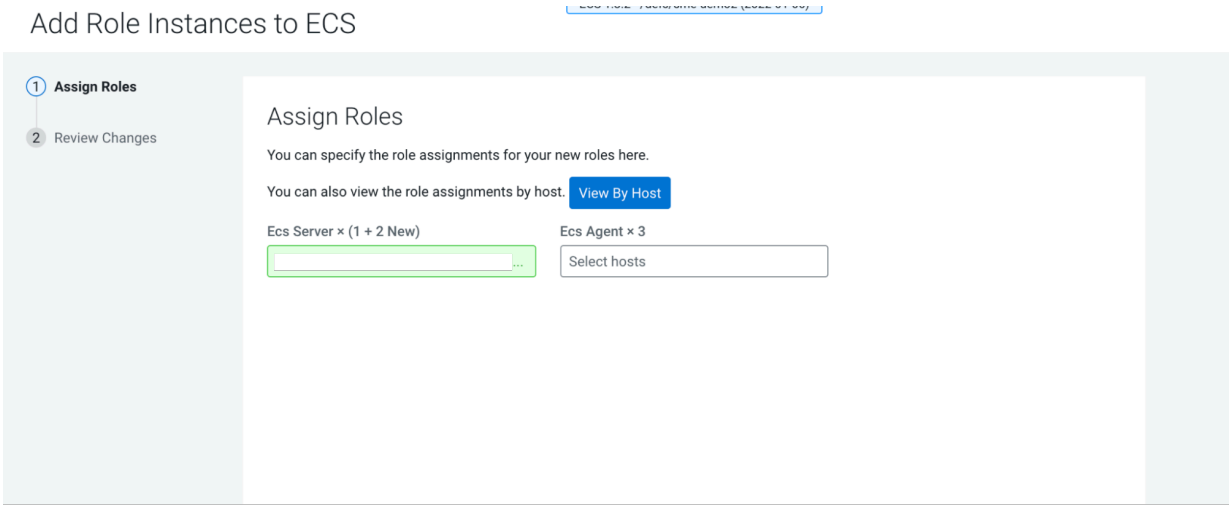


3. In Cloudera Manager, navigate to ECS Cluster ECS and click Add Role Instances.

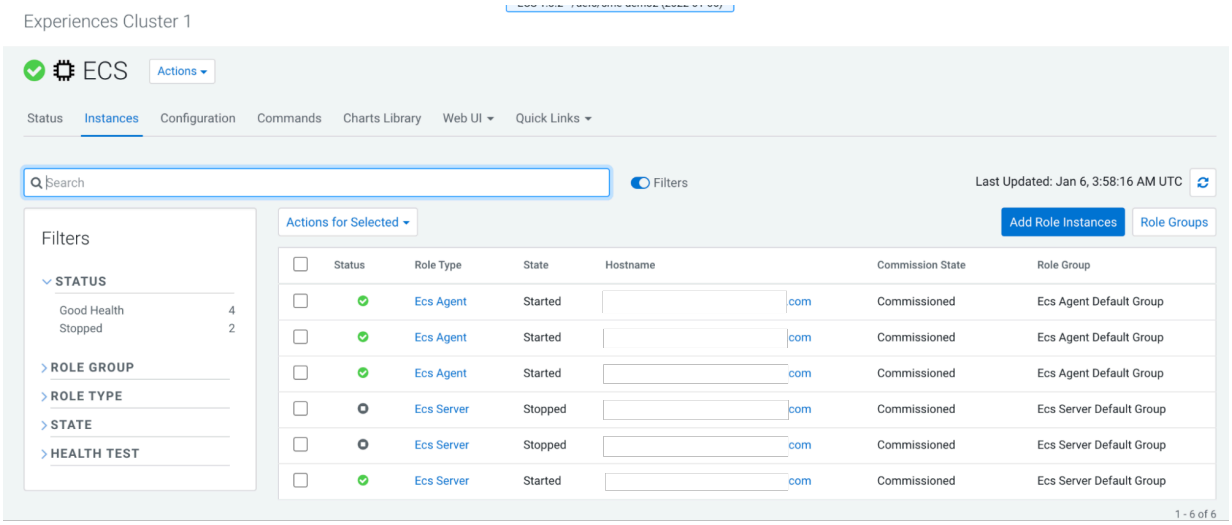


4. Add the available host agent1 as an ECS server in the Add Role Instances to ECS pop-up. Click Ok.

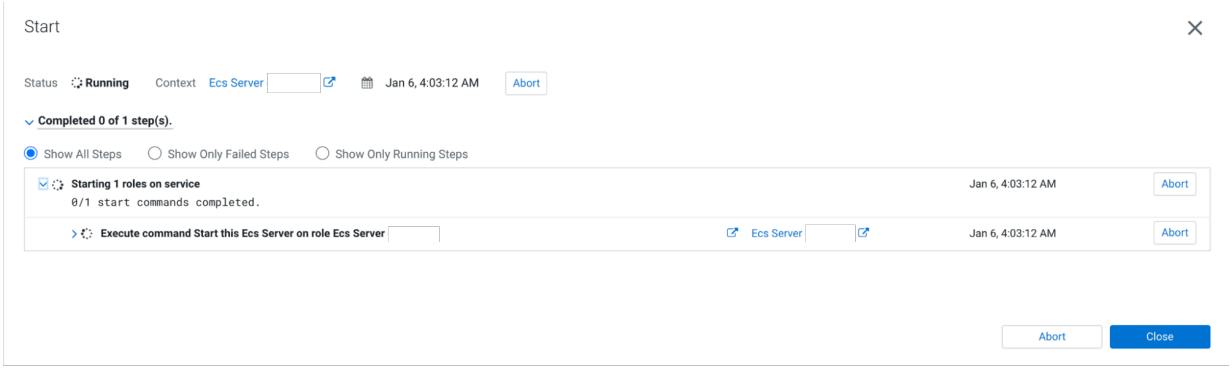




5. Click Continue.



6. Start the new ECS server from ECS Instances view. For example, start ECSServer on agent1.



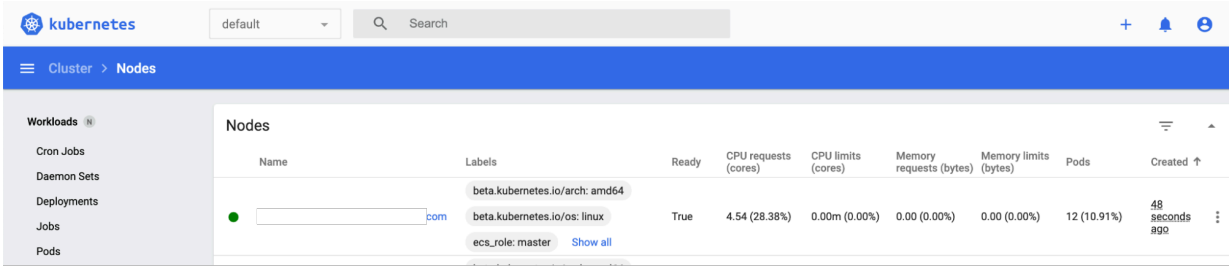
7. On the command line, uncordon the node by running the following command:

```
sudo /var/lib/rancher/rke2/bin/kubectl --kubeconfig=/etc/rancher/rke2/rke2.yaml uncordon agent1.example.com
```

8. Confirm the node's status from webUI or the command line by running the following command:

```
sudo /var/lib/rancher/rke2/bin/kubectl --kubeconfig=/etc/rancher/rke2/rke2.yaml get nodes
```

Note: Do not proceed until node status is Ready. This may take several minutes.



Debugging Impala Virtual Warehouses

You can use the Catalog Web UI, Coordinator Web UI, and the StateStore Web UI to debug Impala Virtual Warehouses in Cloudera Data Warehouse (CDW).

Table level events

In addition to global metrics described below, the following table metrics are available for debugging an Impala Virtual Warehouse:

- avg-events-process-duration
- events-consuming-delay-ms

avg-events-process-duration metric

This metric represents the sum of the time for processing all events. This metric is helpful to identify the average duration of processed events on the table and to identify which tables are causing the event-processor to lag behind. As a temporary workaround, you can disable event processing on that table. You can set the metric collection period to 1 minute, 5 minutes, and 15 minutes duration:

- avg-events-process-duration-1min-rate

  Exponentially weighted moving average (EWMA) of number of events processed in last 1 min
- avg-events-process-duration-5min-rate

  Exponentially weighted moving average (EWMA) of number of events processed in last 5 min
- avg-events-process-duration-1min-rate

  Exponentially weighted moving average (EWMA) of number of events processed in last 15 min

Metric output looks something like this:

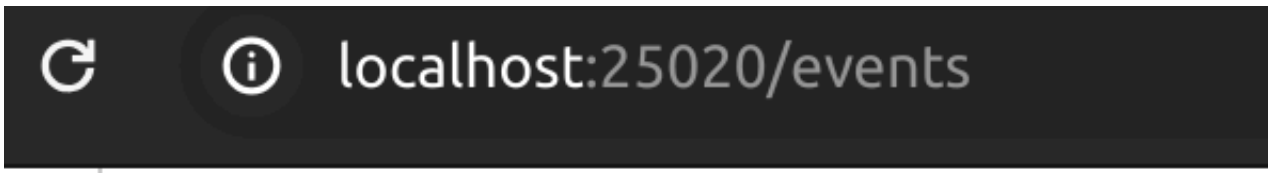
```
events-process-duration:
   Count: 2
   Mean rate: 0.0109
   1 min. rate: 0.0066
   5 min. rate: 0.0048
   15 min. rate: 0.002
   Min (msec): 24
   Max (msec): 33
   Mean (msec): 28.9089
   Median (msec): 33.4825
   75th-% (msec): 33.4825
   95th-% (msec): 33.4825
   99th-% (msec): 33.4825

hms-load-tbl-schema:
   Count: 1
   Mean rate: 0.0055
   1 min. rate: 0.0108
   5 min. rate: 0.1116
   15 min. rate: 0.1647
```

events-consuming-delay-ms metric

This metric represents the time difference between creating an event in the metastore and processing an event. Using this metric, you can gauge how long the event processor is lagging.

Metric output looks something like this:




```
    Mean (msec): 67.9801
    Median (msec): 78.3973
    75th-% (msec): 78.3973
    95th-% (msec): 78.6436
    99th-% (msec): 78.6436

events-consuming-delay:
    Count: 12
    Mean rate: 0.0342
    1 min. rate: 0.0017
    5 min. rate: 0.0146
    15 min. rate: 0.0095
    Min (msec): 2000
    Max (msec): 10000
    Mean (msec): 3472.5061
    Median (msec): 2000
    75th-% (msec): 4000
    95th-% (msec): 8000
    99th-% (msec): 10000
```

About this task

The Impala daemons (impalad, statestored, and catalogd) debug Web UIs, which can be used in CDP Runtime by using Cloudera Manager, is also available in the CDW service. In CDW service, the following Web UIs are provided:

- Impala Catalog Web UI

  This UI provides the same type of information as the Catalog Server Web UI in Cloudera Manager. It includes information about the objects managed by the Impala Virtual Warehouse. For more information about this debug Web UI, see Debug Web UI for Catalog Server.
- Impala Coordinator Web UI

  This UI provides the same type of information as the Impala Daemon Web UI in Cloudera Manager. It includes information about configuration settings, running and completed queries, and associated performance and resource usage for queries. For information about this debug Web UI, see Debug Web UI for Impala Daemon.
- Impala StateStore Web UI

  This UI provides the same type of information as the StateStore Web UI in Cloudera Manager. It includes information about memory usage, configuration settings, and ongoing health checks that are performed by the Impala statestored daemon. For information about this debug Web UI, see Debug Web UI for StateStore.
- Impala Autoscaler Web UI

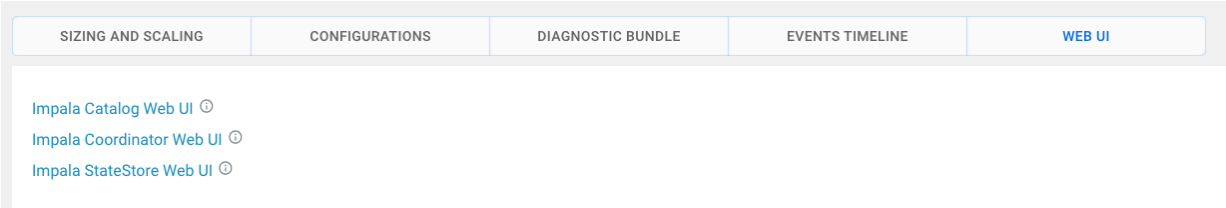
  This UI gives you insight into Autoscaler operations, accessing log messages, and resetting the log level. The autoscaler Web UI includes information about the queries queued and running, executor groups, suspended calls, scale up/down calls, the autoscaler config, and the autoscaler logs.

Required role: EnvironmentAdmin

Before you begin

Procedure

1. In the CDW UI on the Overview page, locate the Impala Virtual Warehouse for which you want to view the debug UIs, and select Edit from the options menu on the tile. This launches the details page for this Virtual Warehouse.
2. In the **Virtual Warehouse** details page, select the WEB UI tab on the right. The list of debug Web UI links are displayed as shown in the following image:



3. Click a Web UI link corresponding to an Impala daemon that you want to debug.

   You are prompted to enter your workload user name and password.

Results

After you are authenticated, you can view the debug Web UI and use the information to help you troubleshoot issues with your Impala Virtual Warehouse.

Resolving Kerberos authentication failure

Condition

When you use impala-shell or a JDBC connection with Kerberos as the authentication mechanism, an unauthorized response is received from Impala Virtual Warehouse and the following warning message is visible in the logs of the impala coordinator pod: W0530 12:08:09.118422 21760 authentication.cc:783] Failed to authenticate request

from <ip-address>:57978 via SPNEGO: Not authorized: Unspecified GSS failure. Minor code may provide more information: Request ticket server hive/dwx-env-<env-name>.cdp.local@ROOT.HWX.SITE kvno 2 found in keytab but not with enctype des3-hmac-sh.

Cause

DES and DES3 encryption types are deprecated. This error occurs when these deprecated encryption types are in the list of enabled Kerberos encryption types, but the Impala Virtual Warehouse pods are running on a cluster where these insecure encryption types are not supported by the operating system running on the cluster nodes.

Solution

To resolve this problem, you can disable using DES and DES3 encryption types in the Kerberos configuration. On test or development environments, you can include and use deprecated encryption types such as “rc4-hmac”, but you must omit DES and DES3 from the list of the allowed encryption types. Modify the Kerberos configuration of the CDP Private Cloud Base cluster by setting appropriate values for the default_tgs_enctypes, default_tkt_enctypes, and permitted_enctypes parameters in the libdefaults block of the /etc/krb5.conf file.

For example:

```
default_tgs_enctypes = rc4-hmac aes256-cts aes128-cts
default_tkt_enctypes = rc4-hmac aes256-cts aes128-cts
permitted_enctypes = rc4-hmac aes256-cts aes128-cts
allow_weak_crypto = true
```

Note: The allow_weak_crypto setting is also required to use the deprecated encryption types.

Cloudera recommends that you do not use deprecated encryption types in production environments. Use stronger encryption types such as AES 256 and AES 128. For example:

```
default_tgs_enctypes = aes256-cts aes128-cts
default_tkt_enctypes = aes256-cts aes128-cts
permitted_enctypes = aes256-cts aes128-cts
```

After modifying the Kerberos configuration of the CDP Private Cloud Base cluster, go to the CDW web interface and refresh the Dtabase Catalog and the Virtual Warehouse by clicking ⋮ Refresh . This copies configurations from the base cluster to CDW.

CDP CLI for Data Warehouse

CDP CLI allows you to manage users, environments and other entities in your Private Cloud deployment. After you setup and configure CDP CLI on your clusters, you can use the various commands and sub-commands that are available.

Commands for Cloudera Data Warehouse are available here. To see whether a command or a sub-command is available for Private Cloud, check the "Form Factor" section.

Runtime documentation for Data Warehouse

Data Warehouse Runtime provides tightly integrated Apache Hive, Apache Impala, and Hue services for Cloudera Data Warehouse (CDW). In this section, you can read about how to use Apache Hive and Apache Impala SQL from clients that connect to your Virtual Warehouse in CDP Private Cloud and how to use Iceberg. Also covered is the Hue interactive query editor, which you can open from a Virtual Warehouse to run SQL queries.

Related Information

Data Warehouse Runtime documentation
Runtime component versions for CDW Private Cloud

List of labels for third-party integration

You can integrate third-party apps and services with CDW by using node labels. By tagging and labeling Cloudera Data Warehouse (CDW) entities and providing the Helm charts of the services you want to integrate, the third-party applications can write Kubernetes injectors to add custom services to CDW without modifying the CDW Docker images. These services then run as sidecar containers to the CDW pods.

Pod labels for labeling CDW entities

The following table provides a list of labels with accepted values and examples:

Label key	Description	Accepted values
cdw.cloudera.com/application	The application name	• database-catalog • impala • hive • hue
cdw.cloudera.com/component	The component name	• metastore • coordinator • executor • admission-controller • catalog • statestore • hiveserver2 • frontend • backend
cdw.cloudera.com/app-version	The CDP Runtime version for Hive, Impala, and Hue	See the CDW component version information
cdw.cloudera.com/cdw-version	The CDW version	See the CDW component version information

Related Information

Version information for Cloudera Data Warehouse Private Cloud components