

## Managing the Cloudera Data Science Workbench Service in Cloudera Manager

Date published: 2020-02-28

Date modified:

**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

## Managing the Cloudera Data Science Workbench Service in Cloudera

<b>Manager.....</b>	<b>4</b>
Adding the Cloudera Data Science Workbench Service.....	4
Roles Associated with the Cloudera Data Science Workbench Service.....	4
Accessing Cloudera Data Science Workbench from Cloudera Manager.....	5
Configuring Cloudera Data Science Workbench Properties.....	5
Starting, Stopping, and Restarting the Service.....	5
Checking the Status of the CDSW Service.....	6
Managing Cloudera Data Science Workbench Worker Hosts.....	6
Health Tests.....	6
Tracking Disk Usage on the Application Block Device.....	6
Create a Chart to Track Disk Usage on the Application Block Device.....	6
Create a Trigger to Notify Cluster Administrators when Free Space Runs Low.....	7
Creating Diagnostic Bundles.....	8

# Managing the Cloudera Data Science Workbench Service in Cloudera Manager

This topic describes how to configure and manage Cloudera Data Science Workbench using Cloudera Manager. The contents of this topic only apply to CSD-based deployments.

If you installed Cloudera Data Science Workbench using the RPM, the Cloudera Data Science Workbench service will not be available to you in Cloudera Manager.

## Adding the Cloudera Data Science Workbench Service

Cloudera Data Science Workbench is available as an add-on service for Cloudera Manager.

To install Cloudera Data Science Workbench, you require the following files: a CSD JAR file that contains all the configuration needed to describe and manage the new Cloudera Data Science Workbench service, and the Cloudera Data Science Workbench parcel.

To install this service, first download and copy the CSD file to the Cloudera Manager Server host. Then use Cloudera Manager to distribute the Cloudera Data Science Workbench parcel to the relevant gateway hosts. You can then use Cloudera Manager's Add Service wizard to add the Cloudera Data Science Workbench service to your cluster.

For the complete set of instructions, see <https://docs.cloudera.com/cdsw/1.10.0/installation/topics/cdsw-install.html>.

## Roles Associated with the Cloudera Data Science Workbench Service

This topic defines the roles associated with the Cloudera Data Science Workbench service.

### Master

Runs the Kubernetes master components on the CDSW master host.

The Master role must only be assigned to the Cloudera Data Science Workbench master host.

### Worker

Runs the Kubernetes worker/host components on the CDSW worker hosts.

The Worker role must be assigned to all Cloudera Data Science Workbench worker hosts. Do not assign the Master and Worker roles to the same host. Even if you are running a single-host proof-of-concept deployment, the single Master host will be able to run user workloads just as a worker host can.

### Docker Daemon

Runs underlying Docker processes on all Cloudera Data Science Workbench hosts.

The Docker Daemon role must be assigned to every Cloudera Data Science Workbench gateway host.

### Application

Runs the Cloudera Data Science Workbench web application. The Application role must only be assigned to the Cloudera Data Science Workbench master host.

As of version 1.6, the Application role can be restarted independently of the other roles. However, the Master role must not be restarted independently of the Application role.

Similarly, do not attempt to restart the underlying Docker Daemon role while the Master/Worker roles are still running on a host. This will result in the operation hanging indefinitely. To avoid this, always perform a full service restart.

## Accessing Cloudera Data Science Workbench from Cloudera Manager

This topic describes how to access Cloudera Manager from Data Science Workbench.

### Procedure

1. Log into the Cloudera Manager Admin Console.
2. Go to the CDSW service.
3. Click CDSW Web UI to visit the Cloudera Data Science Workbench web application.

## Configuring Cloudera Data Science Workbench Properties

In a CSD-based deployment, Cloudera Manager allows you to configure Cloudera Data Science Workbench properties without having to directly edit any configuration file.

### Procedure

1. Log into the Cloudera Manager Admin Console.
2. Go to the CDSW service.
3. Click the Configuration tab.
4. Use the search bar to look for the property you want to configure. You can use Cloudera Manager to [configure](#), [enable TLS](#), reserve the master host, and [enable GPU support](#) for Cloudera Data Science Workbench.  
If you have recently migrated from an RPM-based deployment to a CSD-based deployment, a list of the properties in `cdsw.conf`, along with their corresponding properties in Cloudera Manager can be found in the upgrade guide [here](#)
5. Click Save Changes.

## Starting, Stopping, and Restarting the Service

You can start, stop, and restart Cloudera Data Science Workbench services.

### About this task

On Cloudera Data Science Workbench 1.4.0 (and lower), do not stop or restart Cloudera Data Science Workbench without using the `cdsw_protect_stop_restart.sh` script. This is to help avoid the data loss issue detailed in [TSB-346](#).

#### Points to Remember

- Make sure to stop the CDSW service when doing any work on the nodes, including upgrading the OS or upgrading Cloudera Manager.
- After a restart, the Cloudera Data Science Workbench service in Cloudera Manager will display Good health even though the Cloudera Data Science Workbench web application might need a few more minutes to get ready to serve requests.
- The CDSW service must be restarted every time client configuration is redeployed to the Cloudera Data Science Workbench hosts.

### Procedure

1. Log into the Cloudera Manager Admin Console.
2. On the Home Status tab, click  to the right of the CDSW service and select the action (Start, Stop, or Restart) you want to perform from the dropdown.
3. Confirm your choice on the next screen. When you see a Finished status, the action is complete.

## Checking the Status of the CDSW Service

You can check the status of the CDSW service.

### About this task

Starting with version 1.6, the CDSW service in Cloudera Manager includes the following commands:

- Status: Checks the current status of Cloudera Data Science Workbench.
- Validate: Runs common diagnostic checks to ensure all internal components are configured and running as expected.

To run these commands on the Cloudera Data Science Workbench service:

### Procedure

1. Log into the Cloudera Manager Admin Console.
2. On the Home Status tab, click  to the right of the CDSW service and select the action (Status or Validate) you want to perform from the dropdown.
3. Confirm your choice on the next screen. When you see a Finished status, the action is complete. If the commands fail, click on the stdout tab to view the complete output from the commands.

## Managing Cloudera Data Science Workbench Worker Hosts

You can add or remove workers from Cloudera Data Science Workbench using Cloudera Manager.

For instructions, see:

- [Adding a Worker Host](#)
- [Removing a Worker Host](#)

## Health Tests

Cloudera Manager runs a few health tests to confirm whether Cloudera Data Science Workbench and its components (Master and Workers) are running, and ready to serve requests.

You can choose to enable or disable individual or summary health tests, and in some cases specify what should be included in the calculation of overall health for the service, role instance, or host. See [Configuring Monitoring Settings](#) for more information.

## Tracking Disk Usage on the Application Block Device

This section demonstrates how to use Cloudera Manager to chart disk usage on the Application block device over time, and to create a trigger to notify cluster administrators when free space on the block device falls below a certain threshold.

The latter is particularly important because once the Application block device runs out of disk space, Cloudera Data Science Workbench will stop launching any new sessions or jobs. Advance notifications will give administrators a chance to expand the block device or cleanup existing data before Cloudera Data Science Workbench users run into any problems.

### Create a Chart to Track Disk Usage on the Application Block Device

The following steps use Cloudera Manager's Chart Builder to track disk usage on the Application Block Device (mounted to /var/lib/cdsw on the CDSW master host) over time.

## Procedure

1. Log into the Cloudera Manager Admin Console.
2. Click Charts Chart Builder .
3. Enter a [tsquery](#) that charts disk usage on the block device. For example, the following tsquery creates a chart to track unallocated disk space on the Application block device.

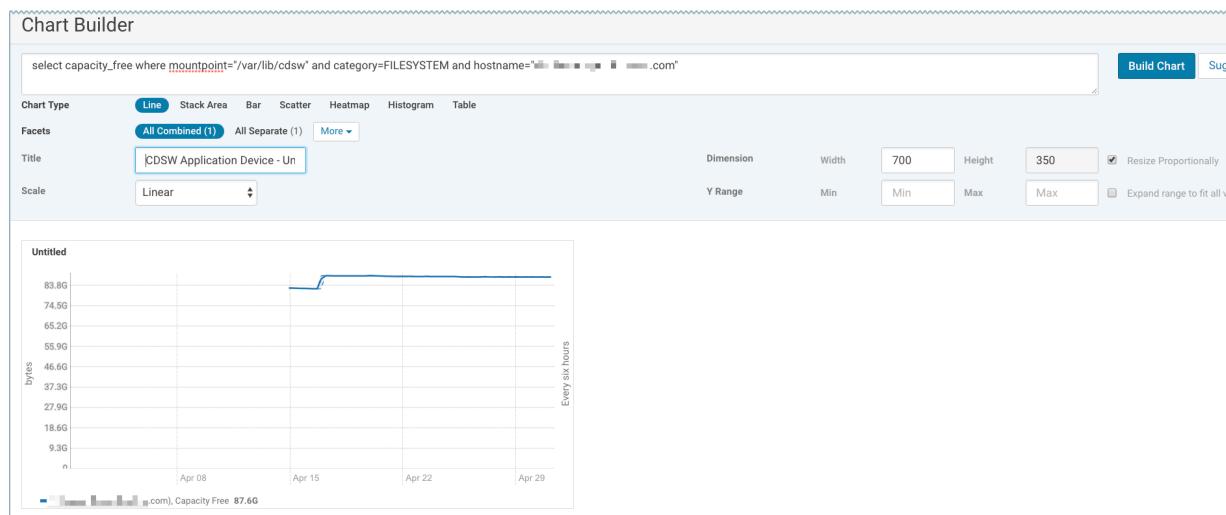
```
select capacity_free where mountpoint="/var/lib/cdsw" and category=FILESYSTEM and hostname=<CDSW_Master_hostname>"
```

Alternatively, you could use the following tsquery to track the disk space already in use on the block device.

```
select capacity, capacity_used where mountpoint="/var/lib/cdsw" and category=FILESYSTEM and hostname=<CDSW_Master_hostname>"
```

Make sure you insert the hostname for your master host as indicated in the queries.

4. Click Build Chart. You should see a preview of the chart below.



5. Click Save.
6. Enter a name for the chart.
7. Select Add chart to another dashboard. From the dropdown list of available System Dashboards, select CDH Cloudera Data Science Workbench Status Page.
8. Click Save Chart. If you navigate back to the CDSW service page, you should now see the new chart on this page.

For more details about Cloudera Manager's Chart Builder, see the following topic in the Cloudera Manager documentation: [Charting Time Series Data](#).

## Create a Trigger to Notify Cluster Administrators when Free Space Runs Low

The following steps create a [trigger](#) to alert Cloudera Manager cluster administrators when free space on the Application Block Device has fallen below a specific threshold.

## Procedure

1. Log in to Cloudera Manager and go to the CDSW service page.
2. Click Create Trigger.
3. Give the trigger a name.

4. Modify the Expression field to include a condition for the trigger to fire. For example, if the trigger should fire when unallocated disk space on the Application Block Device falls below 250GB, the expression should be:

```
IF (select capacity_free where mountpoint="/var/lib/cdsw" and category=FILESYSTEM and hostname="<CDSW_Master_hostname>" and LAST (capacity_free) < 250GB) DO health:concerning
```

On the right hand side of the page, you should see a preview of the query you have entered and a chart that displays the result of the query as in the following sample image. Note that if the query is incorrect or incomplete you will not see the preview on the right.

**Create New Trigger**

**Name**  
CDSW Application Device Memory Low

**Expression**  
IF (select capacity\_free where mountpoint="/var/lib/cdsw" and category=FILESYSTEM and hostname='[REDACTED].com' and LAST (capacity\_free) < 55GB) DO health:concerning

**Preview**  
2 minutes preceding May 1, 12:03 AM PDT  
CDSW Application Device Memory Low  
CDSW Application Device Memory Low trigger is not firing.

**Expression**  
The trigger's name must be unique in the context for which the trigger is defined. That is, there cannot be two triggers for the same service or role with the same name. Different services or different roles can have triggers with the same name.

**Use editor**  
The trigger's expression. A trigger expression is of the form  
IF (CONDITIONS) DO HEALTH\_ACTION

A condition is any valid tsquery. In most cases conditions employ stream filters to filter out streams below or above a certain threshold. For example, the following tsquery can be used to retrieve the streams for DataNodes with more than 500 open file descriptors:

```
SELECT fd_open WHERE roleType=DataNode AND last(fd_open) > 500
```

Conditions can be combined using the logical operators AND and OR. For example, here is a trigger expression with two conditions:

```
IF ((SELECT fd_open WHERE roleType=DataNode AND last(fd_open) > 500) OR (SELECT fd_open WHERE roleType=NameNode AND last(fd_open) > 500)) DO health:bad
```

**Charts**  
capacity\_free  
bytes  
74.5G  
55.9G  
37.3G  
18.6G  
30 12:02 30 12:03  
[REDACTED].com... 87.6G  
Threshold: capacity\_free < 5905... 59.1B

5. Click Create Trigger. If you navigate back to the CDSW service page, you should now see the new trigger in the list of Health Tests.

For more details about Triggers, refer the following topic in the Cloudera Manager documentation: [Triggers](#).

## Creating Diagnostic Bundles

Diagnostic data for Cloudera Data Science Workbench is now available as part of the Cloudera Manager diagnostic bundle. For details on usage and diagnostic data collection in Cloudera Data Science Workbench, see <https://docs.cloudera.com/cdsw/1.10.0/data-collection/topics/cdsw-data-collection.html>