

Managing Cloudera Data Science Workbench Hosts

Date published: 2020-02-28

Date modified:



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 Cloudera Data Science Workbench Hosts.....	4
Customize Workload Scheduling.....	4
Labeling Auxiliary Hosts for CSD Deployments.....	4
Labeling Auxiliary Hosts for RPM Deployments.....	4
Reserving the Master Host for Internal CDSW Components.....	5
Adding and Removing Worker Hosts.....	6
Adding a Worker Host Using Cloudera Manager.....	6
Adding a Worker Host Using Packages.....	7
Validating Sessions on New Worker Host.....	7
Removing a Worker Host Using Cloudera Manager.....	8
Removing a Worker Host Using Packages.....	8
Changing the Domain Name Using Cloudera Manager.....	8
Changing the Domain Name Using Packages.....	9

Managing Cloudera Data Science Workbench Hosts

This topic describes how to perform some common tasks related to managing Cloudera Data Science Workbench hosts.

Customize Workload Scheduling

Starting with version 1.6, Cloudera Data Science Workbench allows you to specify a list of CDSW gateway hosts that are labeled as Auxiliary Nodes. These hosts will be deprioritized during workload scheduling. That is, they will be chosen to run workloads that can't be scheduled on any other hosts; for example sessions with very large resource requests, or when the other hosts are fully utilized..

Cloudera Data Science Workbench will use the following order of preference when scheduling non-GPU workloads (session, job, experiment, or model):

Worker Hosts > Master Host > GPU-equipped Hosts | Labeled Auxiliary Hosts

When selecting a host to schedule an engine, Cloudera Data Science Workbench will give first preference to unlabeled Worker hosts. If Workers are unavailable or at capacity, CDSW will then leverage the Master host. And finally, any GPU-equipped hosts OR labeled auxiliary hosts will be leveraged.

Points to Note:

- GPU-equipped Hosts - Hosts equipped with GPUs will be labeled auxiliary by default so as to reserve them for GPU-intensive workloads. They do not need to be explicitly configured to be labeled. A GPU-equipped host and a labeled auxiliary host will be given equal priority when scheduling workloads.
- Master Host - The Master host must not be labeled an auxiliary node. If you want to reserve the Master for running internal Cloudera Data Science Workbench application components, use the Reserve Master Host property.

Labeling Auxiliary Hosts for CSD Deployments

You can label your auxiliary hosts.

Before you begin

Before you proceed, make sure you have reviewed the guidelines on customizing workload scheduling in Cloudera Data Science Workbench.

Procedure

Use the Auxiliary Nodes property in the CDSW service in Cloudera Manager to specify a comma-separated list of auxiliary hosts:

- a) Log into the Cloudera Manager Admin Console.
- b) Go to the CDSW service.
- c) Click the Configuration tab.
- d) Search for the following property: Auxiliary Nodes.
- e) Enter the hostnames that you want to label as auxiliary.
- f) Click Save Changes.
- g) Restart the CDSW service to have this change go into effect.

Labeling Auxiliary Hosts for RPM Deployments

You can label your auxiliary hosts.

Before you begin

Before you proceed, make sure you have reviewed the guidelines on customizing workload scheduling in Cloudera Data Science Workbench.

Procedure

Use the Auxiliary Nodes property in `cdsw.conf` to specify a comma-separated list of auxiliary hosts:

Reserving the Master Host for Internal CDSW Components

Cloudera Data Science Workbench allows you to reserve the master host for running internal application components and services such as Livelog, the PostgreSQL database, and so on, while user workloads run exclusively on worker hosts.



Note: This feature only applies to deployments with more than one Cloudera Data Science Workbench host. Enabling this feature on single-host deployments will leave Cloudera Data Science Workbench incapable of scheduling any workloads.

Depending on your deployment type, use one of the following sets of instructions to enable this feature:

By default, the master host runs both, user workloads as well as the application's internal services. However, depending on the size of your CDSW deployment and the number of workloads running at any given time, it's possible that user workloads might dominate resources on the master host. Enabling this feature will ensure that CDSW's application components always have access to the resources they need on the master host and are not adversely affected by user workloads.

Reserving the Master Host for CSD Deployments

On CSD-based deployments, this feature can be enabled in Cloudera Manager. Note that this feature is not yet available as a configuration property in Cloudera Manager. However, you can use an Advanced Configuration Snippet (Safety Valve) to configure this.

Before you begin



Note: This feature only applies to deployments with more than one Cloudera Data Science Workbench host. Enabling this feature on single-host deployments will leave Cloudera Data Science Workbench incapable of scheduling any workloads.

Procedure

1. Log into the Cloudera Manager Admin Console.
2. Go to the CDSW service.
3. Click the Configuration tab.
4. Search for the following property: Reserve Master Host., then select the checkbox to enable it.
5. Click Save Changes.
6. Restart the CDSW service to have this change go into effect.

Reserving the Master Host for RPM Deployments

You can reserve the master host for internal CDSW components on RPM deployments.

Before you begin



Note: This feature only applies to deployments with more than one Cloudera Data Science Workbench host. Enabling this feature on single-host deployments will leave Cloudera Data Science Workbench incapable of scheduling any workloads.

Procedure

To enable this feature on RPM-based deployments, go to the `/etc/cdsw/config/cdsw.conf` file and set the `RESERVE_MASTER` property to `true`.

Adding and Removing Worker Hosts

The following topics describe how to add and remove worker hosts from Cloudera Data Science Workbench.

Adding a Worker Host Using Cloudera Manager

Worker hosts are not required for a fully-functional Cloudera Data Science Workbench deployment. For proof-of-concept deployments, you can deploy a 1-host cluster with just a Master host. The Master host can run user workloads just as a worker host can.

About this task

To add a worker host using Cloudera Manager, complete the following steps:

Procedure

1. Log in to the Cloudera Manager Admin Console.
2. Add a new host to your cluster.

Make sure this is a gateway host and you are not running any services on this host.

3. Assign the HDFS, YARN, and Spark 2 gateway roles to the new host. For instructions, refer the Cloudera Manager documentation at [Adding a Role Instance](#).



Note: If you are using Spark 2.2 (or higher), review [JDK Requirements](#) for any additional steps you might need to perform to configure JAVA_HOME on the new nodes.

4. Go to the Cloudera Data Science Workbench service.
5. Click the Instances tab.
6. Click Add Role Instances.
7. Assign the Worker and Docker Daemon roles to the new host. Click Continue.
8. Review your changes and click Continue. The wizard finishes by performing any actions necessary to add the new role instances.

Do not start the new roles at this point. You must run the Prepare Node command as described in the next steps before the roles are started.

9. The new host must have the following packages installed on it.

```
nfs-utils
libseccomp
lvm2
bridge-utils
libtool-ltdl
iptables
rsync
policycoreutils-python
selinux-policy-base
selinux-policy-targeted
ntp
ebtables
bind-utils
openssl
e2fsprogs
redhat-lsb-core
conntrack-tools
bash
```

```
curl
```

You must either manually install these packages now, or, allow Cloudera Manager to install them in the next step.

If you choose the latter, make sure that Cloudera Manager has the permission needed to install the required packages. To do so, go to the Cloudera Data Science Workbench service and click Configuration. Search for the Install Required Packages property and make sure it is enabled.

10. Click Instances and select the new host. From the list of available actions, select the Prepare Node command to install the required packages on the new node.
11. On the Instances page, select the new role instances and click Actions for Selected Start .



Note: It can take several minutes to initialize the host, load Docker images, and be available after you start it. Cloudera Manager may show the host in a healthy state (or green) while the host is being initialized in the background.

Adding a Worker Host Using Packages

Worker hosts are not required for a fully-functional Cloudera Data Science Workbench deployment. For proof-of-concept deployments, you can deploy a 1-host cluster with just a Master host. The Master host can run user workloads just as a worker host can.

About this task

To add a worker using packages on a RPM Deployment, complete the following steps:

Procedure

On an RPM deployment, the procedure to add a worker host to an existing deployment is the same as that required when you first install Cloudera Data Science Workbench on a worker.

For instructions, see [Installing Cloudera Data Science Workbench on a Worker Host](#).

Validating Sessions on New Worker Host

After adding one or more worker hosts or nodes to your cluster, you should validate that the CDSW/CML sessions run properly on the new hosts. If you do not validate that the sessions run properly on the new hosts, you might find intermittent issues that are difficult to debug when a session is scheduled on the worker host.

About this task

You can validate that the session works on the new host by forcing a new session to be scheduled on the host with the following procedure on the master node.

Procedure

1. Enter the following command:

```
kubectl get nodes #
```

2. Validate that the new node is displayed.
3. Cordon all the CDSW nodes to prevent them from being scheduled:

```
kubectl get nodes | awk '{if (NR!=1) {print $1}}' | xargs -I {} kubectl
cordon {}
```

4. Uncordon the new worker node:

```
kubectl uncordon [new cdsw worker node]
```

5. Start a new session in CDSW and verify that this new session is scheduled on the worker node.

```
kubectl get pods -A -o wide #
```

You should see your session in this list.

6. Verify that the CDSW session starts on the new node without any issues.
7. Uncordon all of the CDSW nodes after testing the new worker node:

```
kubectl get nodes | awk '{if (NR!=1) {print $1}}' | xargs -I {} kubectl  
uncordon {}
```

8. If the session does not properly start, you can check the following
 - a) Ensure that the new worker node has the same system level configurations as all other worker nodes, including things such as the `/etc/resolv.conf` file.
 - b) Ensure that any folders in the CDSW -> Admin > Mounts section exist on the new worker host.
 - c) If you see errors starting the session such as `ImgPullBack` error, then validate that any custom Docker images also exist on the new worker node.

Removing a Worker Host Using Cloudera Manager

Perform the following steps to remove a worker host from Cloudera Data Science Workbench using Cloudera Manager.

Procedure

1. Log into the Cloudera Manager Admin Console.
2. Click the Instances tab.
3. Select the Docker Daemon and Worker roles on the host to be removed from Cloudera Data Science Workbench.
4. Select Actions for Selected Stop and click Stop to confirm the action. Click Close when the process is complete.
5. On the Instances page, re-select the Docker Daemon and Worker roles that were stopped in the previous step.
6. Actions for Selected Delete and then click Delete to confirm the action.

Removing a Worker Host Using Packages

Perform the following steps to remove a worker host from Cloudera Data Science Workbench using Packages.

Procedure

1. On the master host, run the following command to delete the worker host:

```
kubectl delete node <worker_host_domain_name>
```

2. Reset the worker host.

```
cdsw stop
```

Changing the Domain Name Using Cloudera Manager

Cloudera Data Science Workbench allows you to change the domain of the web console.

Procedure

1. Log into the Cloudera Manager Admin Console.
2. Go to the Cloudera Data Science Workbench service.
3. Click the Configuration tab.

4. Search for the Cloudera Data Science Workbench Domain property and modify the value to reflect the new domain.
5. Click Save Changes.
6. Restart the Cloudera Data Science Workbench service to have the changes go into effect.

Changing the Domain Name Using Packages

Cloudera Data Science Workbench allows you to change the domain of the web console.

Procedure

1. Open `/etc/cdsw/config/cdsw.conf` and set the DOMAIN variable to the new domain name.

```
DOMAIN="cdsw.<your_new_domain>.com"
```

2. Run the following commands to have the new domain name go into effect.

```
cdsw stop  
cdsw start
```