Installing Cloudera Data Science Workbench Using Packages

Date published: 2020-02-28
Date modified:

https://docs.cloudera.com/

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

Installing Cloudera Data Science Workbench 1.10.1 Using Packages

Use the following steps to install the latest Cloudera Data Science Workbench 1.10.1 using RPM packages.

RPM installations are deprecated and only applicable for HDP. Please choose a different installation method for CDH 6 and CDP.

Prerequisites

Before you begin installing Cloudera Data Science Workbench, make sure you have completed the steps to secure your hosts, set up DNS subdomains, and configure block devices.

Note: You should configure your system journal so that the message file does not grow unchecked. For example, the following sets the maximum size of /var/log/message to 500MB and older log files will be kept until the sum of all message files' sizes exceed 4GB.

```
cat /etc/systemd/journald.conf |grep System
SystemMaxUse=4G
SystemMaxFileSize=500M
```

For information on performing these prerequisite tasks, see: Required Pre-Installation Steps

Note: 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 when required for demonstration purposes. For production deployments, you must have a reserved, dedicated master host and separate worker host(s).

Configure Gateway Hosts Using Cloudera Manager

Cloudera Data Science Workbench hosts must be added to your CDH cluster as gateway hosts, with gateway roles properly configured.

About this task

Note: RPM installations are deprecated and not recommended for CDH 6.

To configure gateway hosts:

Procedure

1. If you have not already done so and plan to use PySpark, install either the Anaconda parcel or Python (versions 2.7.11 and 3.6.1) on your CDH cluster.
2. Configure Apache Spark on your gateway hosts.
   a) (Required CDH 6) To be able to use Spark 2, each user must have their own /home directory in HDFS. If you sign in to Hue first, these directories will automatically be created for you. Alternatively, you can have cluster administrators create these directories.

   ```
   hdfs dfs -mkdir /user/<username>
   hdfs dfs -chown <username>:<username> /user/<username>
   ```

   If you are using CDS 2.3 release 2 (or higher), review the associated known issues here: CDS Powered By Apache Spark.

3. Use Cloudera Manager to create add gateway hosts to your CDH cluster.

a) Create a new host template that includes gateway roles for HDFS, YARN, and Spark 2.

(Required for CDH 6) If you want to run workloads on dataframe-based tables, such as tables from PySpark, sparklyr, SparkSQL, or Scala, you must also add the Hive gateway role to the template.

b) Use the instructions at Adding a Host to the Cluster to add gateway hosts to the cluster. Apply the template created in the previous step to these gateway hosts. If your cluster is kerberized, confirm that the krb5.conf file on your gateway hosts is correct.

4. Test Spark 2 integration on the gateway hosts.

a) SSH to a gateway host.

b) If your cluster is kerberized, run kinit to authenticate to the CDH cluster's Kerberos Key Distribution Center. The Kerberos ticket you create is not visible to Cloudera Data Science Workbench users.

c) Submit a test job to Spark by executing the following command:

CDH 6

```
spark-submit --class org.apache.spark.examples.SparkPi --master yarn \
--deploy-mode client SPARK_HOME/lib/spark-examples*.jar 100
```

To view a sample command, click

```
spark-submit --class org.apache.spark.examples.SparkPi --master yarn \
--deploy-mode client /opt/cloudera/parcels/CDH/lib/spark/examples/jars/
spark-examples*.jar 100
```

d) View the status of the job in the CLI output or in the Spark web UI to confirm that the host you want to use for the Cloudera Data Science Workbench master functions properly as a Spark gateway.

To view sample CLI output, click

```
19/02/15 09:37:39 INFO spark.SparkContext: Running Spark version 2.4.0-c
dh6.1.0
19/02/15 09:37:39 INFO spark.SparkContext: Submitted application: Spark
Pi
...
19/02/15 09:37:40 INFO util.Utils: Successfully started service 'spar
kDriver' on port 37050.
...
19/02/15 09:38:06 INFO scheduler.DAGScheduler: Job 0 finished: reduce at
 SparkPi.scala:38, took 18.659033 s
```

Install Cloudera Data Science Workbench on the Master Host

Use the following steps to install Cloudera Data Science Workbench on the master host.

Before you begin

Note: The airgapped clusters and non-airgapped clusters use different files for installation.

Note: To download Cloudera Data Science Workbench, you must have an active subscription agreement along with the required authentication credentials (namely, the username and password). The authentication credentials are provided in an email sent to the customer account from Cloudera when a new license is issued.

If you do not have the authentication credentials, contact your account representative to receive the same.

You can download the CDSW version 1.10.0 using the URLs listed in *Download and Install the Cloudera Data Science Workbench*.

Procedure

1. Non-airgapped Installation - Download the Cloudera Data Science Workbench repo file (cloudera-cdsw.repo) from the following location:

   ```
   https://username:password@archive.cloudera.com/p/cdsw1/1.10.0/redhat7/yum/
   cloudera-cdsw.repo
   ```

   Airgapped installation - For airgapped installations, download the Cloudera Data Science Workbench RPM file from the following location:

   ```
   https://username:password@archive.cloudera.com/p/cdsw1/1.10.0/redhat7/yum/
   RPMS/x86_64/
   ```

   Note: Make sure all Cloudera Data Science Workbench hosts (master and worker) are running the same version of Cloudera Data Science Workbench.

2. Skip this step for airgapped installations. Add the Cloudera Public GPG repository key. This key verifies that you are downloading genuine packages.

   ```
   sudo rpm --import https://username:password@archive.cloudera.com/p/cdsw1
   /1.10.0/redhat7/yum/RPM-GPG-KEY-cloudera
   ```

3. Non-airgapped Installation - Install the latest RPM with the following command:

   ```
   sudo yum install cloudera-data-science-workbench
   ```

   Airgapped Installation - Copy the RPM downloaded in the previous step to the appropriate gateway host. Then, use the complete filename to install the package. For example:

   ```
   sudo yum install cloudera-data-science-workbench-1.10.0.12345.rpm
   ```

   For guidance on any warnings displayed during the installation process, see Understanding Installation Warnings.

4. Edit the configuration file at /etc/cdsw/config/cdsw.conf. The following table lists the configuration properties that can be configured in cdsw.conf.

Table 1: cdsw.conf Properties

Properties	Description
Required Configuration	
DOMAIN	Wildcard DNS domain configured to point to the master host. If the wildcard DNS entries are configured as cdsw.<*your_domain*>.com and *.cdsw.<*your_domain*>.com, then DOMAIN should be set to cdsw.<your_domain>.com. Users' browsers should then contact the Cloudera Data Science Workbench web application at http://cdsw.<*your_domain*>.com. This domain for DNS and is unrelated to Kerberos or LDAP domains.
MASTER_IP	IPv4 address for the master host that is reachable from the worker hosts. Within an AWS VPC, MASTER_IP should be set to the internal IP address of the master host; for instance, if your hostname is ip-10-251-50-12.ec2.internal, set MASTER_IP to the corresponding IP address, 10.251.50.12.
DISTRO	The Hadoop distribution installed on the cluster. Set this property to HDP.
DOCKER_BLOCK_DEVICES	Block device(s) for Docker images (space separated if there are multiple). Use the full path to specify the image(s), for instance, /dev/xvde.

Properties	Description
JAVA_HOME	Path where Java is installed on the Cloudera Data Science Workbench hosts. This path must match the JAVA_HOME environment variable that is configured for your HDP cluster. You can find the value in hadoop-env.sh on any node in the HDP cluster. Note that Spark 2.3 requires JDK 1.8. For more details on the specific versions of Oracle JDK recommended for HDP clusters, see the Hortonworks Support Matrix - https://supportmatrix.cloudera.com/.
Optional Configuration	
APPLICATION_BLOCK_DEVICE	(Master Host Only) Configure a block device for application state. If this property is left blank, the filesystem mounted at /var/lib/cdsw on the master host will be used to store all user data. For production deployments, Cloudera strongly recommends you use this option with a dedicated SSD block device for the /var/lib/cdsw mount. (Not recommended) If set, Cloudera Data Science Workbench will format the provided block device as ext4, mount it to /var/lib/cdsw, and store all user data on it. This option has only been provided for proof-of-concept setups, and Cloudera is not responsible for any data loss. Use the full path to specify the mount point, for instance, /dev/xvdf.
RESERVE_MASTER	Set this property to true to reserve the master host for Cloudera Data Science Workbench's internal components and services, such as Livelog, the PostgreSQL database, and so on. User workloads will now run exclusively on worker hosts, while the master is reserved for internal application services. 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.
DISTRO_DIR	Path where the Hadoop distribution is installed on the Cloudera Data Science Workbench hosts. For HDP clusters, the default location of the packages is /usr/hdp. Specify this property only if you are using a non-default location.
ANACONDA_DIR	Path where Anaconda is installed. Set this property only if you are using Anaconda for package management. By default, the Anaconda package is installed at: /home/*<your-username>*/anaconda*<2 or 3>* . Refer to the Anaconda FAQs for more details. If you choose to start using Anaconda anytime post-installation, you must set this property and then restart Cloudera Data Science Workbench to have this change take effect.
TLS_ENABLE	Enable and enforce HTTPS (TLS/SSL) for web access. Set to true to enable and enforce HTTPS access to the web application. You can also set this property to true to enable external TLS termination. For more details on TLS termination, see Enabling TLS/SSL for Cloudera Data Science Workbench.
TLS_CERT TLS_KEY	Certificate and private key for internal TLS termination. Setting TLS_CERT and TLS_KEY will enable internal TLS termination. You must also set TLS_ENABLE to true above to enable and enforce internal termination. Set these only if you are not terminating TLS externally. Make sure you specify the full path to the certificate and key files, which must be in PEM format. For details on certificate requirements and enabling TLS termination, see Enabling TLS/SSL for Cloudera Data Science Workbench.

Properties	Description

HTTP_PROXY
HTTPS_PROXY
	If your deployment is behind an HTTP or HTTPS proxy, set the respective HTTP_PROXY or HTTPS_PROXY property in /etc/cdsw/config/cdsw.conf to the hostname of the proxy you are using.
	`HTTP_PROXY="<http://proxy_host>:<proxy-port>"`
	`HTTPS_PROXY="<http://proxy_host>:<proxy_port>"`
	If you are using an intermediate proxy, such as Cntlm, to handle NTLM authentication, add the Cntlm proxy address to the HTTP_PROXY or HTTPS_PROXY fields in cdsw.conf.
	`HTTP_PROXY="http://localhost:3128"`
	`HTTPS_PROXY="http://localhost:3128"`
	If the proxy server uses TLS encryption to handle connection requests, you will need to add the proxy's root CA certificate to your host's store of trusted certificates. This is because proxy servers typically sign their server certificate with their own root certificate. Therefore, any connection attempts will fail until the Cloudera Data Science Workbench host trusts the proxy's root CA certificate. If you do not have access to your proxy's root certificate, contact your Network / IT administrator.
	To enable trust, copy the proxy's root certificate to the trusted CA certificate store (ca-trust) on the Cloudera Data Science Workbench host.
	`cp /tmp/<proxy-root-certificate>.crt /etc/pki/ca-trust/source/anchors/`
	Use the following command to rebuild the trusted certificate store.
	`update-ca-trust extract`

ALL_PROXY	If a SOCKS proxy is in use, set to socks5://<host>:<port>/.

Properties	Description
NO_PROXY	Comma-separated list of hostnames that should be skipped from the proxy. Starting with version 1.4, if you have defined a proxy in the HTTP_PROXY(S) or ALL_PROXY properties, Cloudera Data Science Workbench automatically appends the following list of IP addresses to the NO_PROXY configuration. Note that this is the minimum required configuration for this field. This list includes 127.0.0.1, localhost, and any private Docker registries and HTTP services inside the firewall that Cloudera Data Science Workbench users might want to access from the engines. `"127.0.0.1,localhost,100.66.0.1,100.66.0.2,100.66.0.3,100.66.0.4,100.66.0.5,100.66.0.6,100.66.0.7,100.66.0.8,100.66.0.9,100.66.0.10,100.66.0.11,100.66.0.12,100.66.0.13,100.66.0.14,100.66.0.15,100.66.0.16,100.66.0.17,100.66.0.18,100.66.0.19,100.66.0.20,100.66.0.21,100.66.0.22,100.66.0.23,100.66.0.24,100.66.0.25,100.66.0.26,100.66.0.27,100.66.0.28,100.66.0.29,100.66.0.30,100.66.0.31,100.66.0.32,100.66.0.33,100.66.0.34,100.66.0.35,100.66.0.36,100.66.0.37,100.66.0.38,100.66.0.39,100.66.0.40,100.66.0.41,100.66.0.42,100.66.0.43,100.66.0.44,100.66.0.45,100.66.0.46,100.66.0.47,100.66.0.48,100.66.0.49,100.66.0.50,100.77.0.10,100.77.0.128,100.77.0.129,100.77.0.130,100.77.0.131,100.77.0.132,100.77.0.133,100.77.0.134,100.77.0.135,100.77.0.136,100.77.0.137,100.77.0.138,100.77.0.139"`
NVIDIA_GPU_ENABLE	Set this property to true to enable GPU support for Cloudera Data Science Workbench workloads. When this property is enabled on a host is equipped with GPU hardware, the GPU(s) will be available for use by Cloudera Data Science Workbench hosts. If this property is set to true on a host that does not have GPU support, there will be no effect. By default, this property is set to false. For detailed instructions on how to enable GPU-based workloads on Cloudera Data Science Workbench, see Using NVIDIA GPUs for Cloudera Data Science Workbench Projects.
NVIDIA_LIBRARY_PATH	Complete path to the NVIDIA driver libraries.

5. Initialize and start Cloudera Data Science Workbench.

```
cdsw start
```

The application will take a few minutes to bootstrap. You can watch the status of application installation and startup with watch cdsw status.

(Optional) Install Cloudera Data Science Workbench on Worker Hosts

Cloudera Data Science Workbench supports adding and removing additional worker hosts at any time. Worker hosts allow you to transparently scale the number of concurrent workloads users can run.

About this task

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.

Use the following steps to add worker hosts to Cloudera Data Science Workbench. Note that airgapped clusters and non-airgapped clusters use different files for installation.

Note: To download Cloudera Data Science Workbench, you must have an active subscription agreement along with the required authentication credentials (namely, the username and password). The authentication credentials are provided in an email sent to the customer account from Cloudera when a new license is issued.

If you do not have the authentication credentials, contact your account representative to receive the same.

You can download the CDSW version 1.8 or later using one of the following methods:

- Log in to the Cloudera Downloads web page and download the required files
- Use the URLs listed in this section and provide the login information as provided with your Cloudera subscription.

Procedure

1. Non-airgapped Installation - Download the Cloudera Data Science Workbench repo file (cloudera-cdsw.repo) from the following location:

```
https://archive.cloudera.com/p/cdsw1/1.8.0/redhat7/yum/cloudera-cdsw.repo
```

   Airgapped installation - For airgapped installations, download the Cloudera Data Science Workbench RPM file from the following location:

```
https://archive.cloudera.com/p/cdsw1/1.8.0/redhat7/yum/RPMS/x86_64/
```

   Note: Make sure all Cloudera Data Science Workbench hosts (master and worker) are running the same version of Cloudera Data Science Workbench.

2. Skip this step for airgapped installations. Add the Cloudera Public GPG repository key. This key verifies that you are downloading genuine packages.

```
sudo rpm --import https://archive.cloudera.com/p/cdsw1/1.8.0/redhat7/yum/
RPM-GPG-KEY-cloudera
```

3. Non-airgapped Installation - Install the latest RPM with the following command:

```
sudo yum install cloudera-data-science-workbench
```

   Airgapped Installation - Copy the RPM downloaded in the previous step to the appropriate gateway host. Then, use the complete filename to install the package. For example:

```
sudo yum install cloudera-data-science-workbench-1.8.0.12345.rpm
```

   For guidance on any warnings displayed during the installation process, see Understanding Installation Warnings.

4. Copy cdsw.conf file from the master host:

   ```
   scp root@<cdsw-master-hostname.your_domain.com>:/etc/cdsw/config/cdsw.co
   nf /etc/cdsw/config/cdsw.conf
   ```

   After initialization, the cdsw.conf file includes a generated bootstrap token that allows worker hosts to securely join the cluster. You can get this token by copying the configuration file from master and ensuring it has 600 permissions.

   If your hosts have heterogeneous block device configurations, modify the Docker block device settings in the worker host configuration file after you copy it. Worker hosts do not need application block devices, which store the project files and database state, and this configuration option is ignored.
5. Create /var/lib/cdsw on the worker host. This directory must exist on all worker hosts. Without it, the next step that registers the worker host with the master will fail.

   Unlike the master host, the /var/lib/cdsw directory on worker hosts does not need to be mounted to an Application Block Device. It is only used to store client configuration for HDP services on workers.
6. On the worker host, run the following command to add the host to the cluster:

   ```
   cdsw join
   ```

   This causes the worker hosts to register themselves with the Cloudera Data Science Workbench master host and increase the available pool of resources for workloads.
7. Return to the master host and verify the host is registered with this command:

   ```
   cdsw status
   ```

Create the Administrator Account

After your installation is complete, set up the initial administrator account.

Go to the Cloudera Data Science Workbench web application at http://cdsw.<*your_domain*>.com.

You must access Cloudera Data Science Workbench from the Cloudera Data Science Workbench Domain configured when setting up the service, and not the hostname of the master host. Visiting the hostname of the master host will result in a 404 error.

The first account that you create becomes the site administrator. You may now use this account to create a new project and start using the workbench to run data science workloads. For a brief example, see Getting Started with the Cloudera Data Science Workbench.

Next Steps

As a site administrator, you can invite new users, monitor resource utilization, secure the deployment, and upload a license key for the product.

Depending on the size of your deployment, you might also want to customize how Cloudera Data Science Workbench schedules workloads on your gateway hosts. For more details on these tasks, see:

- Site Administration
- Customize Workload Scheduling
- Security

You can also start using the product by configuring your personal account and creating a new project. For a quickstart that walks you through creating and running a simple template project, see Getting Started with Cloudera Data Science Workbench. For more details on collaborating with teams, working on projects, and sharing results, see the Managing Cloudera Data Science Workbench Users.