

Installing Cloudera Data Science Workbench on CDH

Date published: 2020-02-28

Date modified:

The Cloudera logo is displayed in a bold, orange, sans-serif font. The word "CLOUDERA" is written in all caps, with a stylized 'E' that has a horizontal bar extending to the right.

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

CDSW 1.10.4 Download Information.....	4
Installing Cloudera Data Science Workbench on CDP Private Base.....	5
Reverse Proxy Overview.....	5
Installing the Reverse Proxy Server.....	5
Configuring the Reverse Proxy Server.....	6
Creating a Certificate Signing Request (CSR) and Key/Certificate Pair.....	7
Installing Cloudera Data Science Workbench on CDP Private Base.....	9
Installing Cloudera Data Science Workbench 1.10.4.....	9
Multiple Cloudera Data Science Workbench Deployments.....	9
Airgapped Installations.....	9
Required Pre-Installation Steps.....	9
Set Up a Wildcard DNS Subdomain.....	10
Disable Untrusted SSH Access.....	10
Configure Block Devices.....	10
Installing Cloudera Data Science Workbench 1.10.4 Using Cloudera	
 Manager.....	11
Prerequisites.....	11
Configure Apache Spark 2.....	11
Configure Apache Spark 2 on CDH 6 or CDP Data Center 7.....	12
Configure JAVA_HOME.....	12
Download and Install the Cloudera Data Science Workbench CSD.....	13
Install the Cloudera Data Science Workbench Parcel.....	14
Add the Cloudera Data Science Workbench Service.....	14
Create the Administrator Account.....	17
Next Steps.....	18
Installing Cloudera Data Science Workbench 1.10.4 Using Packages.....	18
Prerequisites.....	18
Configure Gateway Hosts Using Cloudera Manager.....	18
Install Cloudera Data Science Workbench on the Master Host.....	20
(Optional) Install Cloudera Data Science Workbench on Worker Hosts.....	23
Create the Administrator Account.....	25
Next Steps.....	25

CDSW 1.10.4 Download Information

You can use the links provided in this topic to download CDSW 1.10.4.

Parcels

Parcel Type	Location (baseurl)
manifest.json	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/manifest.json
EL5	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-el5.parcel
EL6	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-el6.parcel
EL7	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-el7.parcel
EL8	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-el8.parcel
SLES 11	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-sles11.parcel
SLES 12	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-sles12.parcel
Ubuntu 14	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-trusty.parcel
Ubuntu 16	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-xenial.parcel
Debian 7	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-wheezy.parcel
Debian 8	https://archive.cloudera.com/p/cdsw1/1.10.4/parcels/CDSW-1.10.4.p1.42607632-jessie.parcel

Custom Service Descriptors (CSD)

CSD Type	Location (baseurl)
CDP Private Cloud Base	https://archive.cloudera.com/p/cdsw1/1.10.4/csd/CLOUDERA_DATA_SCIENCE_WORKBENCH-CDPDC-1.10.4.jar
Cloudera Manager CSD (CDH 6)	https://archive.cloudera.com/p/cdsw1/1.10.4/csd/CLOUDERA_DATA_SCIENCE_WORKBENCH-CDH6-1.10.4.jar

RPM

Repository Type	Location (baseurl)
RHEL 7 Compatible	https://archive.cloudera.com/p/cdsw1/1.10.4/redhat7/yum/RPMS/x86_64/cloudera-data-science-workbench-1.10.4.42607632-1.el7.x86_64.rpm

Installing Cloudera Data Science Workbench on CDP Private Base

This topic walks you through the installation paths available for Cloudera Data Science Workbench 1.10.4.

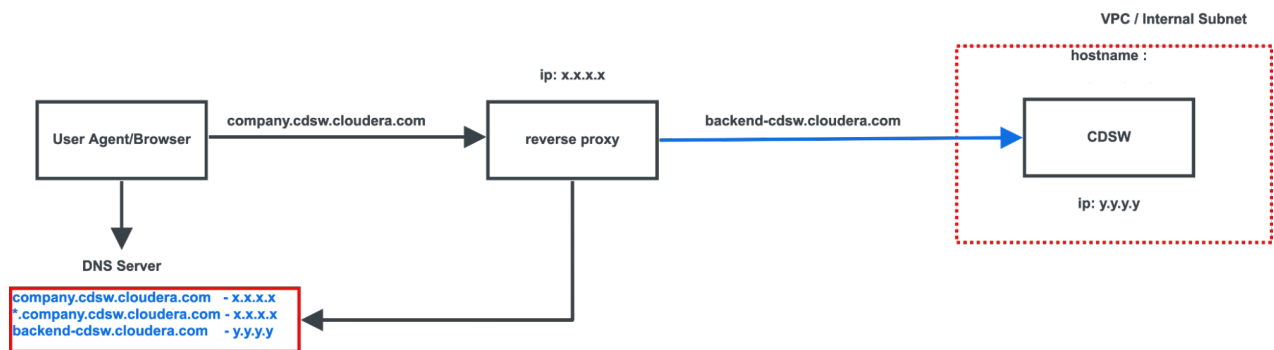
It also describes the steps needed to configure your cluster gateway hosts and block devices before you can begin installing the Cloudera Data Science Workbench parcel/package.

Installing on CDH uses the same steps as installing on CDP.

Reverse Proxy Overview

Reverse Proxy is a server that sits in front of a web server and forwards client or web browser requests to those web servers.

You can use a reverse proxy server for load balancing, web acceleration, and security and anonymity. CDSW uses the Apache HTTP server as its reverse proxy server.



Installing the Reverse Proxy Server

To install the reverse proxy server, you need to manipulate the DNS configurations to support the reverse proxy.

Before you begin

- *** What should I use as the OS? The POC uses Ubuntu. Is this the logical OS to use in the examples?
- The following examples use Ubuntu.

Procedure

1. Install the Apache HTTP server:

```
sudo apt-get update
sudo apt-get install apache2
```

2. Validate the configuration in the `/etc/apache2/apache2.conf` file:

```
sudo apache2ctl configtest
```

3. If you receive the following warning:

```
AH00558: apache2: Could not reliably determine the server's fully qualified domain name, using 127.0.1.1. Set the 'ServerName' directive globally to suppress this message
```

```
Syntax OK
```

you can fix it by adding the following configuration to `/etc/apache2/apache2.conf`:

```
ServerName server_domain_or_IP
```

4. If you have a firewall installed, ensure the firewall is configured to allow traffic to Apache.
5. Validate the Apache HTTP server installation:

```
http://your_server_IP_address
```

You should see the Apache2 Ubuntu default page.

Configuring the Reverse Proxy Server

To install the reverse proxy server, you need to manipulate the DNS configurations to support the reverse proxy.

Before you begin

- *** What should I use as the OS? The POC uses Ubuntu. Is this the logical OS to use in the examples?

Procedure

1. Enable the following modules to enable Apache2 to act as a reverse proxy for CDSW installations:

```
sudo a2enmod proxy
sudo a2enmod proxy_http
sudo a2enmod proxy_balancer
sudo a2enmod lbmethod_byrequests
sudo a2enmod proxy_wstunnel
sudo a2enmod ssl
sudo a2enmod proxy_ajp
sudo a2enmod rewrite
sudo a2enmod deflate
sudo a2enmod headers
sudo a2enmod proxy_connect
sudo a2enmod proxy_html
sudo a2enmod proxy_http2
```

2. Create the configuration file `/etc/apache2/sites-enabled/000-default.conf` and add the following to enable reverse proxy for CDSW:

```
<VirtualHost *:443>
  ServerName company.cdsw.cloudera.com
  ServerAlias *.company.cdsw.cloudera.com

  SSLEngine on
  SSLProxyEngine on

  RewriteEngine On
  RewriteCond %{HTTP:Connection} Upgrade [NC]
  RewriteCond %{HTTP:Upgrade} websocket [NC]
  RewriteRule /(.*) wss://backend-company.cdsw.cloudera.com/$1 [P,L]

  ProxyPass / https://backend-company.cdsw.cloudera.com/ nocanon
  ProxyPassReverse / https://backend-company.cdsw.cloudera.com/ nocanon

  SSLCertificateFile /root/ca/cdsw/company.cdsw.cloudera.com.crt
```

```

SSLCertificateKeyFile /root/ca/cdsw/private.key
SSLCertificateChainFile /root/ca/intermediate/certs/ca-chain.cert.pem

Header always set Strict-Transport-Security "max-age=31536000; includeS
ubDomains; preload"
Header always set Access-Control-Allow-Origin "*"
Header always set Access-Control-Max-Age "1000"
Header always set Access-Control-Allow-Headers "x-requested-with, Con
tent-Type, origin, authorization, accept, client-security-token"

ProxyPreserveHost On
ProxyRequests Off

RewriteEngine On
RewriteRule ^(.*) - [E=CLIENT_IP:%{REMOTE_ADDR},L]

RequestHeader set X-Forwarded-For %{CLIENT_IP}e
RequestHeader set X-Forwarded-Proto https
RequestHeader set X-Forwarded-SSL on
RequestHeader set X-Forwarded-Port 443

AllowEncodedSlashes On
</VirtualHost>

```

3. Ensure that port 443 is associated with the host `company.cds.cds.cloudera.com` and is listed first in the default SSL configuration, `/etc/apache2/sites-enabled/000-default-ssl.conf`.

```

root@ip-10-80-161-148:/etc/apache2/sites-enabled# apachectl -S
VirtualHost configuration: *:443 company.cds.cds.cloudera.com (/etc/apache2/
sites-enabled/000-default.conf:1)
ServerRoot: "/etc/apache2"
Main DocumentRoot: "/var/www/html"
Main ErrorLog: "/var/log/apache2/error.log"
Mutex proxy-balancer-shm: using_defaults
Mutex rewrite-map: using_defaults
Mutex ssl-stapling-refresh: using_defaults
Mutex ssl-stapling: using_defaults
Mutex proxy: using_defaults
Mutex ssl-cache: using_defaults
Mutex default: dir="/var/run/apache2/" mechanism=default
Mutex watchdog-callback: using_defaults
PidFile: "/var/run/apache2/apache2.pid"
Define: DUMP_VHOSTS
Define: DUMP_RUN_CFG
User: name="www-data" id=33
Group: name="www-data" id=33

```

You need to ensure this configuration because there is a default SSL configuration that listens to port 443 and the Apache HTTP server serves the request in the order it is defined in the configuration.

Creating a Certificate Signing Request (CSR) and Key/Certificate Pair

Use the following steps to create a Certificate Signing Request (CSR) to submit to your CA. Then, create a private key/certificate pair that can be used to authenticate incoming communication requests to Cloudera Data Science Workbench.

About this task



Important: Make sure you use `openssl`, and not `keytool`, to perform these steps. `Keytool` does not support a wildcard Subject Alternative Name (SAN) and cannot create flat files.

Procedure

1. Create a `cdsw.cnf` file and populate it with the required configuration parameters including the SAN field values.

```
vi cdsw.cnf
```

2. Copy and paste the default `openssl.cnf` from: <http://web.mit.edu/crypto/openssl.cnf>.
3. Modify the following sections and save the `cdsw.cnf` file:

```
[ CA_default ]
default_md = sha256

[ req ]
default_bits = 2048
distinguished_name = req_distinguished_name
req_extensions = req_ext

[ req_distinguished_name ]
countryName = Country Name (2 letter code)
stateOrProvinceName = State or Province Name (full name)
localityName = Locality Name (eg, city)
organizationName = Organization Name (eg, company)
commonName = Common Name (e.g. server FQDN or YOUR name)

[ req_ext ]
subjectAltName = @alt_names
[alt_names]
DNS.1 = *.cdsw.company.com
DNS.2 = cdsw.company.com
```

Key points to note:

- The domains set in the `DNS.1` and `DNS.2` entries above must match the `DOMAIN` set in `cdsw.conf`.
 - The `default_md` parameter must be set to `sha256` at a minimum. Older hash functions such as `SHA1` are deprecated and will be rejected by browsers, either currently or in the very near future.
 - The `commonName` (CN) parameter will be ignored by browsers. You must use Subject Alternative Names.
4. Run the following command to generate the CSR.

```
openssl req -out cert.csr -newkey rsa:2048 -nodes -keyout private.key -c
onfig cdsw.cnf
```

This command generates the private key and the CSR in one step. The `-nodes` switch disables encryption of the private key (which is not supported by Cloudera Data Science Workbench at this time).

5. Use the CSR and private key generated in the previous step to request a certificate from the CA. If you have access to your organization's internal CA or PKI, use the following command to request the certificate. If you do not have access, or are using a third-party/commercial CA, use your organization's respective internal process to submit the request.

```
openssl x509 -req -days 365 -in cert.csr -CA ca.crt -CAkey ca.key -CAcre
ateserial -out <your_tls_cert>.crt -sha256 -extfile cdsw.cnf -extensions r
eq_ext
```

6. Run the following command to verify that the certificate issued by the CA lists both the required domains, `cdsw.company.com` and `*.cdsw.company.com`, under X509v3 Subject Alternative Name.

```
openssl x509 -in <your_tls_cert>.crt -noout -text
```

You should also verify that a valid hash function is being used to create the certificate. For `SHA-256`, the value under Signature Algorithm will be `sha256WithRSAEncryption`.

Installing Cloudera Data Science Workbench on CDP Private Base

This topic walks you through the installation paths available for Cloudera Data Science Workbench 1.10.4.

It also describes the steps needed to configure your cluster gateway hosts and block devices before you can begin installing the Cloudera Data Science Workbench parcel/package.

Installing on CDH uses the same steps as installing on CDP.

Installing Cloudera Data Science Workbench 1.10.4

There are two ways to install Cloudera Data Science Workbench 1.10.4.

- Using a Custom Service Descriptor (CSD) and Parcel - Starting with version 1.2.x, Cloudera Data Science Workbench is available as an add-on service for Cloudera Manager. Two files are required for this type of installation: 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.

Note that this installation mode does not apply to CDSW-on-HDP deployments.

or

- Using a Package (RPM) - You can install the Cloudera Data Science Workbench package directly on your cluster's gateway or edge hosts. In this case, you will not be able to manage the Cloudera Data Science Workbench service from a cluster manager such as Cloudera Manager or Ambari.

Multiple Cloudera Data Science Workbench Deployments

Starting with version 1.6, you can add more than one Cloudera Data Science Workbench CSD deployment to a single instance of Cloudera Manager. The deployments must install the same version of CDSW.

To add a second Cloudera Data Science Workbench to Cloudera Manager, complete the [Required Pre-Installation Steps](#) for a second set of gateway hosts. Then, install the parcel and add the service as described in [the CSD Installation](#).

Airgapped Installations

Sometimes organizations choose to restrict parts of their network from the Internet for security reasons. Isolating segments of a network can provide assurance that valuable data is not being compromised by individuals out of maliciousness or for personal gain. However, in such cases isolated hosts are unable to access Cloudera repositories for new installations or upgrades. Effective version 1.1.1, Cloudera Data Science Workbench supports installation on CDP clusters that are not connected to the Internet.

For CSD-based installs in an airgapped environment, put the Cloudera Data Science Workbench parcel into a [new hosted or local parcel repository](#), and then configure the Cloudera Manager Server to target this newly-created repository.

Required Pre-Installation Steps

You will need to complete steps and ensure you meet requirements prior to installing Cloudera Data Science Workbench.

Review the complete list of [Cloudera Data Science Workbench 1.10.4 Requirements and Supported Platforms](#) before you proceed with the installation.



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).

Set Up a Wildcard DNS Subdomain

Cloudera Data Science Workbench uses DNS to route HTTP requests to specific engines and services. Wildcard subdomains (such as *.cdsw.<your_domain>.com) are required in order to provide isolation for user-generated content.

Wildcard subdomains help:

- Securely expose interactive session services, such as visualizations, the terminal, and web UIs such as TensorBoard, Shiny, Plotly, and so on.
- Securely isolate user-generated content from the application.

To set up subdomains for Cloudera Data Science Workbench, configure your DNS server with an A record for a wildcard DNS name such as *.cdsw.<your_domain>.com for the master host, and a second A record for the root entry of cdsw.<your_domain>.com.

For example, if your master IP address is 172.46.47.48, you'd configure two A records as follows:

```
cdsw.<your_domain>.com.    IN A 172.46.47.48
*.cdsw.<your_domain>.com.  IN A 172.46.47.48
```

You can also use a wildcard CNAME record if it is supported by your DNS provider.

Starting with version 1.5, the wildcard DNS hostname configured for Cloudera Data Science Workbench must now be resolvable from both, the CDSW cluster, and your Computer from where you are accessing the CDSW UI.

Disable Untrusted SSH Access

Cloudera Data Science Workbench assumes that users only access the gateway hosts through the web application. Untrusted users with SSH access to a Cloudera Data Science Workbench host can gain full access to the cluster, including access to other users' workloads. Therefore, untrusted (non-sudo) SSH access to Cloudera Data Science Workbench hosts must be disabled to ensure a secure deployment.

For more information on the security capabilities of Cloudera Data Science Workbench, see the [Cloudera Data Science Workbench Security Guide](#).

Configure Block Devices

Cloudera Data Science Workbench works with two block devices: Docker block device and application block device or mount point. You need to configure each of these devices to prepare for installation.

Docker Block Device

The Cloudera Data Science Workbench installer will format and mount Docker on each gateway host.

Make sure there is no important data stored on these devices. Do not mount or format the Docker Block Devices before installing Cloudera Data Science Workbench. Ensure that these devices are formatted as disks and are not LVMs. The "lsblk" output should simply say "disk" under the Type column.

Every Cloudera Data Science Workbench gateway host must have one or more block devices with at least 1 TB dedicated to storage of Docker images. The Docker block devices store the Cloudera Data Science Workbench Docker images including the Python, R, and Scala engines. Each engine image can occupy 15GB.

Application Block Device or Mount Point

The master host on Cloudera Data Science Workbench requires at least 1 TB for database and project storage. This recommended capacity is contingent on the expected number of users and projects on the cluster.

While large data files should be stored on HDFS, it is not uncommon to find gigabytes of data or libraries in individual projects. Running out of storage will cause the application to fail. Cloudera recommends allocating at least 5 GB per project and at least 1 TB of storage in total. Make sure you continue to carefully monitor disk space usage and I/O using Cloudera Manager.

Cloudera Data Science Workbench stores all application data at `/var/lib/cdsw`. On a CSD-based deployment, this location is not configurable. The Application Block Device should be formatted before installing Cloudera Data Science Workbench. Cloudera Data Science Workbench will assume the system administrator has formatted and mounted one or more block devices to `/var/lib/cdsw` on the master host. Note that Application Block Device mounts are not required on worker hosts.

Regardless of the application data storage configuration you choose, `/var/lib/cdsw` must be stored on a separate block device. Given typical database and user access patterns, an SSD is strongly recommended.

The `/var/lib/cdsw` directory contains persistent information such as database, configurations, image details and so on. By default, data in `/var/lib/cdsw` is not backed up or replicated to HDFS or other hosts. Reliable storage and backup strategy is critical for production installations. For more information, see [Backup and Disaster Recovery for Cloudera Data Science Workbench](#). To migrate the user-related information to a new host, you can transfer the `/var/lib/cdsw` directory to the new host.

UUID cannot be used in place of the disk name to identify a block device in CDSW. You can determine whether a given path is a block device or not by using the following expression:

```
if [ -b $FILE ]
```

If this expression returns True or 1, then the file exists and is a block special file. This means that the parameter is a path to the block device and cannot be a plain UUID.

Installing Cloudera Data Science Workbench 1.10.4 Using Cloudera Manager

Use the following steps to install Cloudera Data Science Workbench using Cloudera Manager.

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 Apache Spark 2

Depending on the platform you are using, use one of the following sections to configure Apache Spark 2.

Configure Apache Spark 2 on CDH 6 or CDP Data Center 7

Provides steps for configuring Apache Spark 2 on CDH 6 or CDP Data Center 7.

Procedure

1. 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>
```

2. Use Cloudera Manager to add gateway hosts to your CDH cluster.
 - a) Create a new [host template](#) that includes gateway roles for HDFS, YARN, and Spark 2.

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.
3. 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:

```
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.

```
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
```

Configure JAVA_HOME

On CSD-based deployments, Cloudera Manager automatically detects the path and version of Java installed on Cloudera Data Science Workbench gateway hosts.

You do not need to explicitly set the value for JAVA_HOME unless you want to use a custom location, use JRE, or (in the case of Spark 2) force Cloudera Manager to use JDK 1.8 as explained below.

Setting a value for JAVA_HOME - The value for JAVA_HOME depends on whether you are using JDK or JRE. For example, if you're using JDK 1.8_162, set JAVA_HOME to /usr/java/jdk1.8.0_162. If you are only using JRE, set it to /usr/java/jdk1.8.0_162/jre.

Issues with Spark 2.2 and higher - Spark 2.2 (and higher) requires JDK 1.8. However, if a host has both JDK 1.7 and JDK 1.8 installed, Cloudera Manager might choose to use JDK 1.7 over JDK 1.8. If you are using Spark 2.2 (or higher), this will create a problem during the first run of the service because Spark will not work with JDK 1.7. To work around this, explicitly configure Cloudera Manager to use JDK 1.8 on the gateway hosts that are running Cloudera Data Science Workbench.

For instructions on how to set JAVA_HOME, see [Configuring a Custom Java Home Location in Cloudera Manager](#).

To upgrade the whole CDH cluster to JDK 1.8, see [Upgrading to JDK 1.8](#).

Download and Install the Cloudera Data Science Workbench CSD

Provides instructions to download and install the Cloudera Data Science Workbench CSD.

Before you begin

- 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 1.10.4 using the URLs listed in [CDSW Download Information](#).

- Because Cloudera Manager distributes parcels to every host in the cluster, ensure that all nodes in your cluster have enough space (50 GB) in their parcel directory (default opt/cloudera/parcels/) for the CDSW parcel.

Procedure

1. Download the Cloudera Data Science Workbench CSD. Make sure you download the CSD that corresponds to the version of CDH or Cloudera Runtime you are using.

See [CDSW Download Information](#) for the download urls.

2. Log on to the Cloudera Manager Server host, and place the CSD file under /opt/cloudera/csd, which is the default location for CSD files. To configure a custom location for CSD files, refer to the Cloudera Manager documentation at [Configuring the Location of Custom Service Descriptor Files](#).

3. Set the file ownership to cloudera-scm:cloudera-scm with permission 644.

Set the file ownership.

CDH

```
chown cloudera-scm:cloudera-scm CLOUDERA_DATA_SCIENCE_WORKBENCH-CDH<X>-1.10.4<Y>.jar
```

CDP Data Center

```
chown cloudera-scm:cloudera-scm CLOUDERA_DATA_SCIENCE_WORKBENCH-CDPDC-1.10.4<Y>.jar
```

4. Set the file permissions.

CDH

```
chmod 644 CLOUDERA_DATA_SCIENCE_WORKBENCH-CDH<X>-1.10.4<Y>.jar
```

CDP Data Center

```
chmod 644 CLOUDERA_DATA_SCIENCE_WORKBENCH-CDPDC-1.10.4<Y>.jar
```

- Restart the Cloudera Manager Server:

```
service cloudera-scm-server restart
```

- Log into the Cloudera Manager Admin Console and restart the Cloudera Management Service:
 - Select Clusters Cloudera Management Service .
 - Select Actions Restart .

Install the Cloudera Data Science Workbench Parcel

Provides steps to install the Cloudera Data Science Workbench parcel.

About this task

For installation and upgrades, you must manually add the Remote Parcel Repository URL for your CDSW version to Cloudera Manager.



Note: Starting with CDSW 1.10.x, you must have at least 50 Gb of space in the parcel directory on all hosts registered in Cloudera Manager to unzip the CDSW parcel. This is typically the `/opt/` folder.

Procedure

- Log into the Cloudera Manager Admin Console.
- Click **Hosts Parcels** in the main navigation bar.
- Add the remote parcel repository URL to Cloudera Manager.

For detailed steps, click

- On the **Parcels** page, click **Configuration**.
- In the **Remote Parcel Repository URLs** list, click the addition symbol to open an additional row.
- Enter the path to the repository. Cloudera Data Science Workbench publishes placeholder parcels for other operating systems as well. However, note that these do not work and have only been included to support mixed-OS clusters.


Version	Remote Parcel Repository URL
Cloudera Data Science Workbench 1.10.0.0	<code>https://archive.cloudera.com/p/cdsw1/1.10.0.0/parcels/</code>

- Click **Save Changes**.
 - Go to the **Hosts Parcels** page. The external parcel should now appear in the set of parcels available for download.
- Click **Download**. Once the download is complete, click **Distribute** to distribute the parcel to all the CDH hosts in your cluster. Then click **Activate**. For more detailed information on each of these tasks, see [Managing Parcels](#).
For airgapped installations, [create your own local repository](#), put the Cloudera Data Science Workbench parcel there, and then configure the Cloudera Manager Server to target this newly-created repository.

Add the Cloudera Data Science Workbench Service

Perform the following steps to add the Cloudera Data Science Workbench service to your cluster.

Procedure

1. Log into the Cloudera Manager Admin Console.
2. On the Home Status tab, click  to the right of the cluster name and select Add a Service to launch the wizard. A list of services will be displayed.
3. Select the Cloudera Data Science Workbench service and click Continue.
4. Select the services which the new CDSW service should depend on. At a minimum, the HDFS, Spark 2, and YARN services are required for the CDSW service to run successfully. Click Continue.
(Required for CDH 6) If you want to run SparkSQL workloads, you must also add the Hive service as a dependency.
5. Assign the CDSW roles, HDFS, Spark 2, and YARN, to gateway hosts:

Master

Assign the Master role to a gateway host that is the designated Master host. This is the host that should have the Application Block Device mounted to it.

Worker

Assign the Worker role to any other gateway hosts that will be used for Cloudera Data Science Workbench. Note that 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.

Even if you are setting up a multi-host deployment, do not assign the Master and Worker roles to the same host. By default, the Master host doubles up to perform both functions: those of the Master and those of a worker.

Docker Daemon

This role 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.

On First Run, Cloudera Manager will automatically assign this role to each Cloudera Data Science Workbench gateway host. However, if any more hosts are added or reassigned to Cloudera Data Science Workbench, you must explicitly assign the Docker Daemon role to them.

Application

This role runs the Cloudera Data Science Workbench application. This role runs only on the CDSW Master host.

On First Run, Cloudera Manager will assign the Application role to the host running the Cloudera Data Science Workbench Master role. The Application role is always assigned to the same host as the Master. Consequently, this role must never be assigned to a Worker host.


The following image shows the role assignments for a Cloudera Data Science Workbench Master host and Worker host:

Hosts	Count	Roles
...	1	B, G, NN, NF..., SNN, AP, ES, HM, NAS, NMS, RM, SM, G, HS, JHS, RM
...	1	A, DD, M, G
...	1	DD, W, G
...	2	DN, G, NM

This table is grouped by hosts having the same roles assigned to them.

6. Configure the following parameters and click Continue.

Properties	Description
Cloudera Data Science Workbench Domain	<p>DNS domain configured to point to the master host.</p> <p>If the previously configured DNS subdomain entries are <code>cdsw.<your_domain>.com</code> and <code>*.cdsw.<your_domain>.com</code>, then this parameter should be set to <code>cdsw.<your_domain>.com</code>.</p> <p>Users' browsers contact the Cloudera Data Science Workbench web application at <code>http://cdsw.<your_domain>.com</code>.</p> <p>This domain for DNS only and is unrelated to Kerberos or LDAP domains.</p>
Master Node IPv4 Address	<p>IPv4 address for the master host that is reachable from the worker host. By default, this field is left blank and Cloudera Manager uses the IPv4 address of the Master host.</p> <p>Within an AWS VPC, set this parameter to the internal IP address of the master host; for instance, if your hostname is <code>ip-10-251-50-12.ec2.internal</code>, set this property to the corresponding IP address, <code>10.251.50.12</code>.</p>

Properties	Description
Install Required Packages	<p>When this parameter is enabled, the Prepare Node command will install all the required package dependencies on First Run. If you choose to disable this property, you must manually install the following packages on all gateway hosts running Cloudera Data Science Workbench roles:</p> <pre>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 bash curl contrack-tools</pre> <p> Note: Be sure to reboot the host machines after the Prepare Node command installs all the required package dependencies.</p>
Docker Block Device	<p>Block device(s) for Docker images. Use the full path to specify the image(s), for instance, /dev/xvde.</p> <p>For the First Run of the Cloudera Data Science workbench service, you only need to enter the list of block devices for the Master node.</p> <p>To add block devices for other (worker) nodes use Role Groups in Cloudera Manager.</p> <p>The Cloudera Data Science Workbench installer will format and mount Docker on each gateway host that is assigned the Docker Daemon role. Do not mount these block devices prior to installation.</p>
RESERVE_MASTER Node	<p>Cloudera recommends setting the RESERVE_MASTER node option to TRUE for clusters with 100+ users and/or 10+ nodes.</p>

- The wizard will now begin a First Run of the Cloudera Data Science Workbench service. This includes deploying client configuration for HDFS, YARN and Spark 2, installing the package dependencies on all hosts, and formatting the Docker block device. The wizard will also assign the Application role to the host running Master and the Docker Daemon role to all the gateway hosts running Cloudera Data Science Workbench.



Note: Ensure you have 200GB of space devoted to DOCKER_TMPDIR (default to /var/lib/cdsd/docker-tmp) on the master node. This is needed to unzip all of the new docker images.

- Once the First Run command has completed successfully, click Finish to go back to the Cloudera Manager home page.

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](#).

Installing Cloudera Data Science Workbench 1.10.4 Using Packages

Use the following steps to install the latest Cloudera Data Science Workbench 1.10.4 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 1.10.4 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/cdsw1/1.10.4/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/cdsw1/1.10.4/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
/1.10.4/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-1.10.4.12345.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	<p>Wildcard DNS domain configured to point to the master host.</p> <p>If the wildcard DNS entries are configured as <code>cdsw.<your_domain>.com</code> and <code>*.cdsw.<your_domain>.com</code>, then DOMAIN should be set to <code>cdsw.<your_domain>.com</code>. Users' browsers should then contact the Cloudera Data Science Workbench web application at <code>http://cdsw.<your_domain>.com</code>.</p> <p>This domain for DNS and is unrelated to Kerberos or LDAP domains.</p>
MASTER_IP	<p>IPv4 address for the master host that is reachable from the worker hosts.</p> <p>Within an AWS VPC, MASTER_IP should be set to the internal IP address of the master host; for instance, if your hostname is <code>ip-10-251-50-12.ec2.internal</code>, set MASTER_IP to the corresponding IP address, <code>10.251.50.12</code>.</p>
DISTRO	The Hadoop distribution installed on the cluster. Set this property to HDP.
DOCKER_BLOCK_DEVICES	<p>Block device(s) for Docker images (space separated if there are multiple).</p> <p>Use the full path to specify the image(s), for instance, <code>/dev/xvde</code>.</p>
JAVA_HOME	<p>Path where Java is installed on the Cloudera Data Science Workbench hosts.</p> <p>This path must match the JAVA_HOME environment variable that is configured for your HDP cluster. You can find the value in <code>hadoop-env.sh</code> on any node in the HDP cluster.</p> <p>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/.</p>
Optional Configuration	
APPLICATION_BLOCK_DEVICE	<p>(Master Host Only) Configure a block device for application state.</p> <p>If this property is left blank, the filesystem mounted at <code>/var/lib/cdsw</code> 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 <code>/var/lib/cdsw</code> mount.</p> <p>(Not recommended) If set, Cloudera Data Science Workbench will format the provided block device as <code>ext4</code>, mount it to <code>/var/lib/cdsw</code>, 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.</p> <p>Use the full path to specify the mount point, for instance, <code>/dev/xvdf</code>.</p>
RESERVE_MASTER	<p>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.</p> <p>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.</p>
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 <code>/usr/hdp</code> . Specify this property only if you are using a non-default location.
ANACONDA_DIR	<p>Path where Anaconda is installed. Set this property only if you are using Anaconda for package management.</p> <p>By default, the Anaconda package is installed at: <code>/home/<your-username>/anaconda<2 or 3></code>. Refer to the Anaconda FAQs for more details.</p> <p>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.</p>

Properties	Description
TLS_ENABLE	<p>Enable and enforce HTTPS (TLS/SSL) for web access.</p> <p>Set to true to enable and enforce HTTPS access to the web application.</p> <p>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.</p>
TLS_CERT TLS_KEY	<p>Certificate and private key for internal TLS termination.</p> <p>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.</p> <p>Make sure you specify the full path to the certificate and key files, which must be in PEM format.</p> <p>For details on certificate requirements and enabling TLS termination, see Enabling TLS/SSL for Cloudera Data Science Workbench.</p>
HTTP_PROXY HTTPS_PROXY	<p>If your deployment is behind an HTTP or HTTPS proxy, set the respective HTTP_PROXY or HTTPS_PROXY property in <code>/etc/cdsw/config/cdsw.conf</code> to the hostname of the proxy you are using.</p> <pre>HTTP_PROXY="<http://proxy_host>:<proxy-port>" HTTPS_PROXY="<http://proxy_host>:<proxy-port>"</pre> <p>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 <code>cdsw.conf</code>.</p> <pre>HTTP_PROXY="http://localhost:3128" HTTPS_PROXY="http://localhost:3128"</pre> <p>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.</p> <p>To enable trust, copy the proxy's root certificate to the trusted CA certificate store (<code>ca-trust</code>) on the Cloudera Data Science Workbench host.</p> <pre>cp /tmp/<proxy-root-certificate>.cert /etc/pki/ca-trust/source/anchors/</pre> <p>Use the following command to rebuild the trusted certificate store.</p> <pre>update-ca-trust extract</pre>
ALL_PROXY	<p>If a SOCKS proxy is in use, set to <code>socks5://<host>:<port>/</code>.</p>

Properties	Description
NO_PROXY	<p>Comma-separated list of hostnames that should be skipped from the proxy.</p> <p>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.</p> <p>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.</p> <pre>"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"</pre>
NVIDIA_GPU_ENABLE	<p>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.</p> <p>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.</p> <p>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.</p>
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-cds.w.repo) from the following location:

```
https://archive.cloudera.com/p/cdsw1/1.8.0/redhat7/yum/cloudera-cds.w.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.conf /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](#).