

Troubleshooting Cloudera Data Science Workbench

Date published: 2020-02-28

Date modified: 2021-06-24



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

Troubleshooting Cloudera Data Science Workbench.....	4
Understanding Installation Warnings.....	4
GPU Issues.....	7
Error Encountered Trying to Load Images when Initializing Cloudera Data Science Workbench.....	8
404 Not Found Error.....	8
Troubleshooting Kerberos Errors.....	8
Troubleshooting TLS/SSL Errors.....	9
Troubleshooting Issues with Workloads.....	10
Troubleshooting Issues with Models and Experiments.....	11

Troubleshooting Cloudera Data Science Workbench

Depending on your deployment, use one of the following courses of action to start debugging issues with Cloudera Data Science Workbench.

- CSD Deployments
 - Check the current status of the application and run validation checks. You can use the Status and Validate commands in Cloudera Manager to do so.
 - Make sure your Cloudera Data Science Workbench configuration is correct. Log into Cloudera Manager and review configuration for the CDSW service.
- RPM Deployments
 - Check the status of the application.

```
cdsw status
```

- SSH to your master host and run the following host validation command to check that the key services are running:

```
cdsw validate
```

- Make sure your Cloudera Data Science Workbench configuration is correct.

```
cat /etc/cdsw/config/cdsw.conf
```

Understanding Installation Warnings

This section describes solutions to some warnings you might encounter during the installation process.

Preexisting iptables rules not supported

WARNING: Cloudera Data Science Workbench requires iptables, but does not support preexisting iptables rules.

Kubernetes makes extensive use of iptables. However, it's hard to know how pre-existing iptables rules will interact with the rules inserted by Kubernetes. Therefore, Cloudera recommends you run the following commands to clear all pre-existing rules before you proceed with the installation.

```
sudo iptables -F INPUT ACCEPT
sudo iptables -F FORWARD ACCEPT
sudo iptables -F OUTPUT ACCEPT
sudo iptables -t nat -F
sudo iptables -t mangle -F
sudo iptables -F
sudo iptables -X
```

The warning can be ignored after you clear the pre-existing rules or are sure that there are no pre-existing iptables rules.

Remove the entry corresponding to /dev/xvdc from /etc/fstab

Cloudera Data Science Workbench installs a custom filesystem on its Application and Docker block devices. These filesystems will be used to store user project files and Docker engine images respectively. Therefore, Cloudera Data

Science Workbench requires complete access to the block devices. To avoid losing any existing data, make sure the block devices allocated to Cloudera Data Science Workbench are reserved only for the workbench.

Linux sysctl kernel configuration errors

Kubernetes and Docker require non-standard kernel configuration. Depending on the existing state of your kernel, this might result in sysctl errors such as:

```
sysctl net.bridge.bridge-nf-call-iptables must be set to 1
```

This is because the settings in `/etc/sysctl.conf` conflict with the settings required by Cloudera Data Science Workbench. Cloudera cannot make a blanket recommendation on how to resolve such errors because they are specific to your deployment. Cluster administrators may choose to either remove or modify the conflicting value directly in `/etc/sysctl.conf`, remove the value from the conflicting configuration file, or even delete the module that is causing the conflict.

To start diagnosing the issue, run the following command to see the list of configuration files that are overwriting values in `/etc/sysctl.conf`.

```
SYSTEMD_LOG_LEVEL=debug /usr/lib/systemd/systemd-sysctl
```

You will see output similar to:

```
Parsing /usr/lib/sysctl.d/00-system.conf
Parsing /usr/lib/sysctl.d/50-default.conf
Parsing /etc/sysctl.d/99-sysctl.conf
Overwriting earlier assignment of net/bridge/bridge-nf-call-ip6tables in file
'/etc/sysctl.d/99-sysctl.conf'.
Overwriting earlier assignment of net/bridge/bridge-nf-call-ip6tables in file
'/etc/sysctl.d/99-sysctl.conf'.
Overwriting earlier assignment of net/bridge/bridge-nf-call-ip6tables in file
'/etc/sysctl.d/99-sysctl.conf'.
Parsing /etc/sysctl.d/k8s.conf
Overwriting earlier assignment of net/bridge/bridge-nf-call-iptables in file
'/etc/sysctl.d/k8s.conf'.
Parsing /etc/sysctl.conf
Overwriting earlier assignment of net/bridge/bridge-nf-call-ip6tables in file
'/etc/sysctl.conf'.
Overwriting earlier assignment of net/bridge/bridge-nf-call-ip6tables in file
'/etc/sysctl.conf'.
Setting 'net/ipv4/conf/all/promote_secondaries' to '1'
Setting 'net/ipv4/conf/default/promote_secondaries' to '1'
...
```

`/etc/sysctl.d/k8s.conf` is the configuration added by Cloudera Data Science Workbench. Administrators must make sure that no other file is overwriting values set by `/etc/sysctl.d/k8s.conf`.

CDH parcels not found at `/opt/cloudera/parcels`

There are two possible reasons for this warning:

- If you are using a custom parcel directory, you can ignore the warning and proceed with the installation. Once the Cloudera Data Science Workbench is running, set the path to the CDH parcel in the admin dashboard.
- This warning can be an indication that you have not added gateway roles to the Cloudera Data Science Workbench hosts. In this case, do not ignore the warning. Exit the installer and go to Cloudera Manager to add gateway roles to the cluster.

CDSW docker daemons fail to start

CDSW docker daemons fail to start with the following error:

```
Error starting daemon: error initializing graphdriver: devmapper: Unable to take ownership of thin-pool (docker-thinpool) that already has used data blocks.
```

This issue occurs when the block devices you specified for the Docker Block Device field already have data on them. This is a safeguard to prevent block devices from being wiped inadvertently. Note that resolving this issue involves deleting data from the block devices.

To resolve this issue, perform the following steps:

1. Verify that it is okay to delete the data on the block device.
2. SSH to the Cloudera Data Science Workbench master host.
3. Run the following script:

```
/opt/cloudera/parcels/CDSW/scripts/teardown-docker.sh
```

4. In the Cloudera Manager Admin Console, select the Cloudera Data Science Workbench service.
5. On the Instances tab, select the Docker Daemons.
6. Click **Actions for Selected (n) Prepare Node**.
7. Start the Cloudera Data Science Workbench service by clicking **Actions Start**.

User Process Limit

During host validation, you may encounter the following warning message:

```
{WARN} Cloudera Data Science Workbench recommends that all users have a max-user-processes limit of at least 65536.
```

This message appears if the user process limit is under 65536. You can increase the user process limit by adding the following line to the `/etc/security/limits.conf` file:

```
ulimit -u 65536
```

Set this configuration on every Cloudera Data Science Workbench host. You can also edit `/etc/security/limits.conf` to configure the user process limit.

Open Files Limit

During host validation, you may encounter the following warning message:

```
{WARN} Cloudera Data Science Workbench recommends that all users have a max-open-files limit set to 1048576.
```

This message appears if the open files limit is under 1048576. Note that on HDP clusters, the open file limit recommendation is 10000 at a minimum. Cloudera recommends a higher limit for clusters with Cloudera Data Science Workbench.

You can configure the file limit with the following command:

```
ulimit -n 1048576
```

Set this configuration on every Cloudera Data Science Workbench host. You can also edit `/etc/security/limits.conf` to configure the open files limit.

Disable SE Linux

During installation, you may encounter the following message:

```
Please disable SELinux by setting SELINUX=disabled|permissive in /etc/selinux/config, then reboot or using setenforce 0 command"
```

SELinux enforces additional control policies for what a user, process, or daemon can do. If SELinux is enabled or not in permissive mode, Cloudera Data Science Workbench may not have the proper permissions to run.

To resolve this issue, you must change the SELinux mode on every host by doing one of the following:

- Edit the configuration file for SELinux and set it to disabled or permissive. Note that if you set SELinux to permissive mode, events such as access denials will be logged, but the denial will not be enforced. You can find the SELinux configuration file in the following location: `/etc/selinux/config`.
- Run the following command: `setenforce 0`. This command disables SELinux completely.

DNS is not configured properly

During installation, you might encounter the messages such as:

```
DNS doesn't resolve <CDSW_domain> to <CDSW_Master_IP_address>; DNS is not configured properly
```

or

```
DNS doesn't resolve <CDSW_Master_IP_address> to <CDSW_domain>; DNS is not configured properly"
```

This indicates that the CDSW domain name configured does not resolve to the IP address of the Master host. You must enable DNS forward and reverse lookup for the CDSW domain and IP address to proceed.

GPU Issues

This section describes GPU issues and solutions you might encounter during GPU set up.

Known Issues with Tensorflow 2.4 and CUDA 11.2

During GPU set up, the dynamic library `libcusolver.so.10` is not read. For more information, see <https://github.com/tensorflow/tensorflow/issues/44777>.

Workaround: Enter the following command when starting a session:

```
!ln -s /usr/local/cuda-11.1/targets/x86_64-linux/lib/libcusolver.so.11.0.1.105 /usr/local/cuda-11.1/targets/x86_64-linux/lib/libcusolver.so.10
!ln -s /usr/lib/x86_64-linux-gnu/libcuda.so.460.73.01 /usr/lib/x86_64-linux-gnu/libcuda.so.1
```

Error Encountered Trying to Load Images when Initializing Cloudera Data Science Workbench

Here are some sample error messages you might see when initializing Cloudera Data Science Workbench.

```
Error encountered while trying to load images.: 1
```

```
Unable to load images from [/etc/cdsd/images/cdsd_<version>.tar.gz].: 1
```

```
Error processing tar file(exit status 1): write ../../tar: no space left on device
```

These errors are an indication that the root volume is running out of space when trying to initialize Cloudera Data Science Workbench. During the initialization process, the Cloudera Data Science Workbench installer temporarily decompresses the engine image file located in `/etc/cdsd/images` to the `/var/lib/cdsd/docker-tmp/` directory.

If you have previously partitioned the root volume (which should be at least 200 GB), make sure you allocate at least 200 GB to `/var/lib/cdsd/docker-tmp/` so that the installer can proceed without running out of space.

404 Not Found Error

The 404 Not Found error might appear in the browser when you try to reach the Cloudera Data Science Workbench web application.

This error is an indication that your installation of Cloudera Data Science Workbench was successful, but there was a mismatch in the domain configured in `cdsd.conf` and the domain referenced in the browser. To fix the error, go to `/etc/cdsd/config/cdsd.conf` and check that the URL you supplied for the `DOMAIN` property matches the one you are trying to use to reach the web application. This is the wildcard domain dedicated to Cloudera Data Science Workbench, not the hostname of the master host.

If this requires a change to `cdsd.conf`, after saving the changes run `cdsd stop` followed by `cdsd start`.

Troubleshooting Kerberos Errors

Here are troubleshooting guidelines for Kerberos errors.

HDFS commands fail with Kerberos errors even though Kerberos authentication is successful in the web application

If Kerberos authentication is successful in the web application, and the output of `klist` in the engine reveals a valid-looking TGT, but commands such as `hdfs dfs -ls /` still fail with a Kerberos error, it is possible that your cluster is missing the [Java Cryptography Extension \(JCE\) Unlimited Strength Jurisdiction Policy File](#). The JCE policy file is required when Red Hat uses AES-256 encryption. This library should be installed on each cluster host and will live under `$JAVA_HOME`. For more information, see [Using AES-256 Encryption](#).

Cannot find renewable Kerberos TGT

Cloudera Data Science Workbench runs its own Kerberos TGT renewer which produces non-renewable TGT. However, this confuses Hadoop's renewer which looks for renewable TGTs. If the Spark 2 logging level is set to WARN or lower, you may see exceptions such as:

```
16/12/24 16:38:40 WARN security.UserGroupInformation: Exception encountered while running the renewal command. Aborting renew thread. ExitCodeException exitCode=1: kinit: Resource temporarily unavailable while renewing credentials
```



```
16/12/24 16:41:23 WARN security.UserGroupInformation: PrivilegedActionException as:user@CLLOUDERA.LOCAL (auth:KERBEROS) cause:javax.security.sasl.SaslException: GSS initiate failed [Caused by GSSException: No valid credentials provided (Mechanism level: Failed to find any Kerberos tgt)]
```

This is not a bug. Spark 2 workloads will not be affected by this. Access to Kerberized resources should also work as expected.

Troubleshooting TLS/SSL Errors

This section describes some common issues with TLS configuration on Cloudera Data Science Workbench. Common errors include.

- Cloudera Data Science Workbench initialisation fails with an error such as:

```
Error preparing server: tls: failed to parse private key
```

- Your browser reports that the Cloudera Data Science Workbench web application is not secure even though you have enabled TLS settings as per [Enabling TLS/SSL for Cloudera Data Science Workbench](#).

Possible Causes and Solutions

- Certificate does not include the wildcard domain - Confirm that the TLS certificate issued by your CA lists both, the Cloudera Data Science Workbench domain, as well as a wildcard for all first-level subdomains. For example, if your Cloudera Data Science Workbench domain is `cdsw.company.com`, then the TLS certificate must include both `cdsw.company.com` and `*.cdsw.company.com`.
- Path to the private key and/or certificate is incorrect - Confirm that the path to the private key file is correct by comparing the path and file name to the values for `TLS_KEY` and/or `TLS_CERT` in `cdsw.conf` or Cloudera Manager. For example:

```
TLS_CERT="/path/to/cert.pem"
TLS_KEY="/path/to/private.key"
```

- Private key file does not have the right permissions - The private key file must have read-only permissions. Set it as follows:

```
chmod 444 private.key
```

- Private key is encrypted - Cloudera Data Science Workbench does not support encrypted private keys. Check to see if your private key is encrypted:

```
cat private.key
```

```
-----BEGIN RSA PRIVATE KEY-----
Proc-Type: 4, ENCRYPTED
DEK-Info: DES-EDE3-CBC, 11556F53E4A2824A
```

If the private key is encrypted as shown above, use the following steps to decrypt it:

1. Make a backup of the private key file.

```
mv private.key private.key.encrypted
```

2. Decrypt the backup private key and save the file to `private.key`. You will be asked to enter the private key password.

```
openssl rsa -in private.key.encrypted -out private.key
```

- Private key and certificate are not related - Check to see if the private key matches the public key in the certificate.
 1. Print a hash of the private key modulus.

```
openssl rsa -in private.key -noout -modulus | openssl md5
```

```
(stdin)= 7a8d72ed61bb4be3c1f59e4f0161c023
```

2. Print a hash of the public key modulus.

```
openssl x509 -in cert.pem -noout -modulus | openssl md5
```

```
(stdin)= 7a8d72ed61bb4be3c1f59e4f0161c023
```

If the md5 hash output of both keys is different, they are not related to each other, and will not work. You must revoke the old certificate, regenerate a new private key and Certificate Signing Request (CSR), and then apply for a new certificate.

Troubleshooting Issues with Workloads

This section describes some potential issues data scientists might encounter once the application is running workloads.

404 error in Workbench after starting an engine

This is typically caused because a wildcard DNS subdomain was not set up before installation. While the application will largely work, the engine consoles are served on subdomains and will not be routed correctly unless a wildcard DNS entry pointing to the master host is properly configured. You might need to wait 30-60 minutes until the DNS entries propagate. For instructions, see [Set Up a Wildcard DNS Subdomain](#).

.

Engines cannot be scheduled due to lack of CPU or memory

A symptom of this is the following error message in the Workbench: "Unschedulable: No node in the cluster currently has enough CPU or memory to run the engine."

Either shut down some running sessions or jobs or provision more hosts for Cloudera Data Science Workbench.

Workbench prompt flashes red and does not take input

The Workbench prompt flashing red indicates that the session is not currently ready to take input.

Cloudera Data Science Workbench does not currently support non-REPL interaction. One workaround is to skip the prompt using appropriate command-line arguments. Otherwise, consider using the terminal to answer interactive prompts.

PySpark jobs fail due to HDFS permission errors

```
: org.apache.hadoop.security.AccessControlException: Permission denied: user
=alice, access=WRITE, inode="/user":hdfs:supergroup:drwxr-xr-x
```

(Required for CDH 5 and 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>
```

PySpark jobs fail due to Python version mismatch

Exception: Python in worker has different version 2.6 than that in driver 2.7, PySpark cannot run with different minor versions

One solution is to install the matching Python 2.7 version on all the cluster hosts. Another, more recommended solution is to install the Anaconda parcel on all CDH cluster hosts. Cloudera Data Science Workbench Python engines will use the version of Python included in the Anaconda parcel which ensures Python versions between driver and workers will always match. Any library paths in workloads sent from drivers to workers will also match because Anaconda is present in the same location across all hosts. Once the parcel has been installed, set the PYSARK_PYTHON environment variable in the Cloudera Data Science Workbench Admin dashboard. Alternatively, you can use Cloudera Manager to set the path.

Jobs fail due to incorrect JAVA_HOME on HDP

Commands, such as hdfs commands, and jobs fail with an error similar to the following message:

```
ERROR: JAVA_HOME /usr/lib/jvm/java does not exist.
```

The JAVA_HOME path you configure for Cloudera Data Science Workbench in cdsw.conf must match the JAVA_HOME configured by hadoop-env.sh for the HDP cluster. After you update JAVA_HOME in cdsw.conf, you must restart Cloudera Data Science Workbench. For more information, see [Changes to cdsw.conf](#).

The user_events table is growing in size and affecting performance

The user_events table is used to monitor and audit user events. It can grown in size in long running deployments and can decrease performance. To clean the table manually:

1. SSH to the Cloudera Data Science Workbench Master host and log in as root.

```
ssh root@<cdsw_master_host_domain_name>
```

2. Get the name of the database pod:

```
kubectl get pods -l role=db
```

The command returns information similar to the following example:

NAME	READY	STATUS	RESTARTS	AGE
db-6d56584f76-phn2f	1/1	Running	0	4h46m

3. Enter the following command to log into the database as the sense user::

```
kubectl exec -it <database pod> -- psql -U sense
```

4. Run the following query to get the number of rows from the user_events table:

```
select count(id) from user_events;
```

5. Delete the records older than 30 days by running the following query:

```
delete from user_events where created_at < NOW() - INTERVAL '30 days';
```

Troubleshooting Issues with Models and Experiments

This section provides troubleshooting recommendations for models and experiments.

See the following topics:

- [Debugging Issues with Experiments](#)
- [Debugging Issues with Models](#)