

Data Collection in Cloudera Data Science Workbench

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

Data Collection in Cloudera Data Science Workbench.....	4
Usage Tracking.....	4
Disable Usage Tracking.....	4
Diagnostic Bundles.....	4
Using Cloudera Manager.....	4
Disabling Usage Metrics.....	5
Using the Command Line.....	5
Information Collected in Diagnostic Bundles.....	5

Data Collection in Cloudera Data Science Workbench

Cloudera Data Science Workbench collects usage and diagnostic data in two ways, both of which can be controlled by system administrators.

Usage Tracking

Cloudera Data Science Workbench collects aggregate usage data by sending limited tracking events to Google Analytics and Cloudera servers. No customer data or personal information is sent as part of these bundles.

Disable Usage Tracking

You can disable Cloudera Data Science Workbench usage tracking.

Procedure

1. Log in to Cloudera Data Science Workbench as a site administrator.
2. Click Admin.
3. Click the Settings tab.
4. Uncheck Send usage data to Cloudera.

In addition to this, Cloudera Manager also collects and sends anonymous usage information using Google Analytics to Cloudera. If you are running a CSD-based deployment and want to disable data collection by Cloudera Manager, see [Managing Anonymous Usage Data Collection](#).

Diagnostic Bundles

Diagnostic bundles are used to aid in debugging issues filed with Cloudera Support. They can be created using either Cloudera Manager (only for CSD-based deployments), or the command line. This section also provides details on the information collected by Cloudera Data Science workbench as part of these bundles.

Using Cloudera Manager

If you are working on a CSD-based deployment, Cloudera Data Science Workbench logs and diagnostic data are available as part of the diagnostic bundles created by Cloudera Manager.

By default, Cloudera Manager is configured to collect diagnostic data weekly and to send it to Cloudera automatically. You can schedule the frequency of data collection on a daily, weekly, or monthly schedule, or even disable the scheduled collection of data. To learn how to configure the frequency of data collection or disable it entirely, see [Diagnostic Data Collection by Cloudera Manager](#).

You can also manually trigger a collection and transfer of diagnostic data to Cloudera at any time. For instructions, see [Manually Collecting and Sending Diagnostic Data to Cloudera](#).



Note: If Cloudera Data Science Workbench data is missing from Cloudera Manager diagnostic bundles, it might be due to [this known issue](#).

You can configure what directory Cloudera Manager uses as a staging directory for diagnostic bundles. Change the Log Staging Directory property for your Cloudera Data Science Workbench instance in Cloudera Manager to set a different directory.

Disabling Usage Metrics

Starting with version 1.7, CDSW also gathers highly redacted information on which feature is being used. When you create a diagnostic bundle, this information is packed alongside the diagnostic information.

About this task

If you want to disable collection of usage metrics, configure the Feature flag overrides property in Cloudera Manager as described here:

Procedure

1. Log in to Cloudera Manager.
2. Go to the CDSW service.
3. Click Configuration.
4. Search for the Feature flag overrides property. Add the following JSON code to disable usage me.

```
{"usage-reporting": false}
```

5. Click Save Changes.
6. Restart the CDSW service.

Using the Command Line

Diagnostic bundles can be created by system administrators using the `cdsw logs` command.

By default, sensitive information will be redacted from the log files in the bundle. This is the bundle that you should attach to any case opened with Cloudera Support. The filename of this generated bundle will be of the form, `cdsw-logs-$hostname-$date-$time.redacted.tar.gz`.

If you want to turn off redaction of log files, you can use the `-x|--skip-redaction` option as demonstrated below.

```
cdsw logs --skip-redaction
```

The diagnostic bundle is only meant for internal use. It should be retained at least for the duration of the support case, in case any critical information was redacted. However, it can be shared with Cloudera at your discretion. The filename of this bundle will be of the form, `cdsw-logs-$hostname-$date-$time.tar.gz`.

The contents of both archives are stored in text and can easily be inspected by system administrators. Both forms are designed to be easily diff-able.

Usage Metrics

Starting with version 1.7, CDSW also gathers highly redacted information on which feature is being used. When you create a diagnostic bundle, this information is packed alongside the diagnostic information. If you want to turn off collection of information on feature usage, use the `-u|--skip-usage-events` flag when you generate the diagnostic bundle. For example:

```
cdsw logs --skip-usage-events
```

Information Collected in Diagnostic Bundles

Cloudera Data Science Workbench diagnostic bundles collect the following information.

- System information such as hostnames, operating system, kernel modules and settings, available hardware, and system logs.
- Cloudera Data Science Workbench version, status information, configuration, and the results of install-time validation checks.
- Details about file systems, devices, and mounts in use.
- CDH cluster configuration, including information about Java, Kerberos, installed parcels, and CDH services such as Spark 2.

- Network configuration and status, including interfaces, routing configuration, and reachability.
- Status information for system services such as Docker, Kubernetes, NFS, and NTP.
- Listings for processes, open files, and network sockets.
- Reachability, configuration, and logs for Cloudera Data Science Workbench application components.
- Hashed Cloudera Data Science Workbench user names.
- Information about Cloudera Data Science Workbench workloads (sessions, jobs, experiments, models, applications), including editor, kernel type, engine, ownership, termination status, and performance data.