

Cloudera Observability Overview

Date published: 2023-04-31

Date modified: 2023-09-14

CLOUDERA

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

What is Cloudera Observability and how is it useful.....	4
About the Cloudera Observability user interface hierarchy.....	4
About the Cloudera Observability common web user interface features.....	7
Using the Cloudera Observability user interface.....	7
Collecting Cloudera Observability diagnostic metrics for Public Cloud.....	12
Metric sources sent to Cloudera Observability.....	12
Diagnostic metrics collection details.....	13
Enabling the redaction of sensitive data.....	14
Collecting Cloudera Observability diagnostic metrics for Private Cloud.....	15
Metric sources sent to Cloudera Observability.....	15
Diagnostic metrics collection details.....	16
Redaction capabilities for diagnostic data.....	16
Considerations related to telemetry tasks for your environment's deployment.....	17
What is Expedited Support.....	19
Disabling Expedited Support.....	20
Reenabling Expedited Support.....	21

What is Cloudera Observability and how is it useful

Cloudera Observability is a Cloudera service that helps you interactively understand your environment, data services, workloads, clusters, and resources. Its wide range of metrics and health tests help you identify and troubleshoot existing and potential problems. This service also provides prescriptive guidance and recommendations that help you quickly address those problems and optimize solutions. When a workload completes, diagnostic information about the job or query and the cluster that processed them is collected by Telemetry Publisher, a role in the Cloudera Manager Management Service, and sent to Cloudera Observability.

Cloudera Observability enables you to interactively understand your environment, data services, workloads, clusters, and resources, and optimize your systems through:

- A wide range of metrics and health tests that help you identify and troubleshoot both existing and potential issues.
- Prescriptive guidance and recommendations that help you quickly address those problems and optimize solutions.
- Performance baselines and historical analysis that help you identify and address performance problems.

In addition, Cloudera Observability also enables you to:

- Visually display your workload cluster's current and historical costs that help you plan and forecast budgets, future workload environments, and justify current user groups and resources.
- Trigger actions in real-time across jobs and queries that help you take steps to alleviate potential problems.
- Enable the daily delivery of your cluster statistics to your email address that help you to track, compare, and monitor without having to log in to the cluster.
- Break down your workload metrics into more meaningful views for your business requirements that help you analyze specific workload criteria. For example, you can analyze how queries that access a particular database or that use a specific resource pool are performing against your SLAs. Or you can examine how all the queries are performing on your cluster that are sent by a specific user.

About the Cloudera Observability user interface hierarchy

Provides a brief introduction to the web interface, its hierarchical components, and the frequently used interface elements of Cloudera Observability. Cloudera recommends that you take a moment to familiarize yourself with the user interface, its components, and elements.



Note: The features and views are dependent on your Cloudera Observability subscription entitlement.

The Cloudera Observability web UI hierarchically displays the health, performance, and status of your environment, services, workload clusters, engines, and resources, including the costs associated with your data infrastructure, from the top-down. Its dashboard components include statistics, performance, health, and prescriptive guidance visually displayed in chart widgets or tabular views.

After selecting the Cloudera Observability tile from the CDP Public Cloud Enterprise Data web interface landing page, the Cloudera Observability landing page opens. This page is also called the Cloudera Observability HOME page.



Tip: Clicking the Cloudera Observability icon brings you back to this page.

About the main navigation panel

The **Main** navigation panel enables access to the following Cloudera Observability features:

- Financial Governance, which opens the **Chargeback** page that displays the total costs and the hourly CPU and memory usage for all of your cost centers, including the unutilized resource usage costs from the Uncategorised section.

From its Actions list you can:

- Configure your cost centers criteria based on CPU and Memory costs and resource usage.
- Create cost centers, which separate costs across user or pool usage and track their workload resource consumption costs.
- Analytics, which depending on the tier level within the hierarchy displays:
 - The **Environments** page that lists your environments, including:
 - The environment's platform type.
 - The platform's version number.
 - The number and type of service hosted on the platform.
 - The date and time that telemetry data was last collected.

To filter and display only those environment platforms or services of interest, either enter the environment's name in the Search field or from the Environments list, select the environment's Type.

The Cloudera Observability Environment Types are:

- Classic Cluster
- Manually Uploaded
- Private Cloud Base
- Data Lake
- Data Hub
- Database Catalog
- Virtual Warehouse
- Data Engineering
- Virtual Cluster

Where, a Cloudera Observability Environment hierarchically represents the association of your Public or Private Cloud account and its data services, resources, clusters and their workloads (Jobs and Queries).

- User login name, which enables you to access your Cloudera Management Console user profile and to log out. The profile page displays information about you including your cloud region and cloud resource name. You to can also create the Cloudera Observability Telemetry Publisher access key credentials from the Actions list.

About the environment navigation panel

The **Environment** navigation panel hierarchically displays the environment from its parent tier (environment name) to the lower tier levels (services, service components, clusters, engines, jobs, and queries).

The following table describes the environment categories, which are displayed in the Environment panel dependent on the selected Cloudera Observability environment type:

Table 1: Environment categories and summaries

Category	Description
DATA HUB CLUSTERS	This service when expanded in the Environment panel, hierarchically displays the environment's clusters and workload engines.
DATA WAREHOUSE	This service when expanded in the Environment panel, hierarchically displays the data service's Database Catalogs, their Virtual Warehouses and workload Hive and Impala engines.
DATA ENGINEERING	This service when expanded in the Environment panel, hierarchically displays the data service's Virtual Clusters and workload Spark engines.

Category	Description
ENGINES	Lists the Hive, Impala, Spark, Oozie, and MapReduce workload engines for Classic Cluster, Private Cloud Base, and Data Hub environments.
HIVE METASTORE	Lists the Hive and Impala engine metastores. When a metastore is selected the List of Tables View displays details about each table available in your system that were processed in Hive and Impala engines, regardless of whether they have been queried or not.
Unclassified Jobs	Displays the diagnostic data collected from Data Hub environments before April 30th 2023 or collected from CDE and CDW environments using versions older than CDE 1.19 and CDW 1.6.3.

The following table describes the Environment, Cluster, and Engine Summary pages:

Table 2: Environment summary pages

Environment Summary	<p>Displays the Data Lake Data Services page, which enables the exploration of each of your services and their components.</p> <p>The table has the following columns:</p> <ul style="list-style-type: none"> • Data Service, which lists the name of each service. • Type, which lists the Cloudera Observability environment type, such as Data Engineering Service, Database Catalog, or Data Hub Cluster. • Content, which displays the number of workspaces, virtual warehouses, or virtual clusters. • Version, which displays the version number of the service. • Last Data Collection, which displays the last time data was collected for the service.
Cluster Summary	<p>Displays the Summary page and the Cloudera Observability features available for the environment's cluster and subscription entitlement:</p> <ul style="list-style-type: none"> • Summary tab, which displays performance trends and metrics about the processed jobs and queries as chart widgets and enables you to view historical trends for analysis when you select a predefined or custom time period from the Time-Range filter list. • Workloads tab, which enables the creation and management of your Workload views for the environment's cluster. • Validations tab, which displays all the open validation alerts for your environment's cluster. • Auto Actions tab, which enables the creation and management of your Auto Actions events for the environment's cluster.
Engine Summary	Displays information about the workload jobs or queries run by the selected engine, such as which jobs or queries have failed or are slow, their processing time, missed SLAs (thresholds), user and pool metrics, and outlier issues.

About the Cloudera Observability chart widgets

The Cloudera Observability chart widgets enable you to quickly observe real-time and historical patterns, trends, and outliers of your workload data. They provide quick insights into the health and performance of your workloads, clusters, and resources. Where:

- Hovering over an element with your mouse pointer, such as over a time-line or a data point, displays more information about the element underneath.
- Clicking a link within a chart widget or a bar within bar chart widget, such as the Suboptimal bar chart widget, opens the engine's Jobs or Queries page that contains more information in a tabular view for you to investigate further.

About the Jobs and Queries pages

Depending on the engine chosen, the engine's Jobs or Queries page provides information about each of the jobs or queries that are serviced by the engine. You can filter further by selecting a filter option from one of the filter categories; Pool, User, Status, Health Check, or Duration.

For more information about a specific job's or query's health, execution details, baseline, and trends, open their respective page by clicking the link in the Job or Query column and selecting the tab of interest. To investigate further, these pages also provide prescriptive guidance and recommendations that enable you to address problems and optimize solutions.

About the Cloudera Observability common web user interface features

Learn about the Cloudera Observability web UI navigation features, charts, and actions.

As you explore your environment's statistics and health the following UI elements are available to help you navigate and explore your diagnostic data:

Navigation

- Navigation drawer panels (side-bars) that toggle between open and close, which enable you to view more or less real estate space and provide access from the parent tier (environment) to the lower tier levels (services, clusters, engines, jobs, and queries).
- Drawer panels that toggle between open and close on the right-side of the page to provide more detailed information about a component.
- Breadcrumbs that are displayed at the top of the page, which displays the name of your current location and its preceding pages. You can move between these pages by clicking on a breadcrumb location.

Charts and statistic banners

- Statistic banners, which display dynamic and interactive pie charts, rose charts, bar charts, and statistic cards about your jobs, queries, tables, and engines.
- Chart widgets, which enable you to dynamically observe real-time and historical patterns, trends, and outliers of your workload data, jobs, and queries.

Action tasks

- Filters, which enable you to refine your selection and display only the components of interest.
- Action menus, which list the actions that you can perform on an environment or component, such as on a data service or engine.
- Time-range list (time-picker), which displays the current or historical data for the selected time-period.
- Search fields and lists, which enable you to locate a specific component, such as an environment, data service, or cluster ID.

Using the Cloudera Observability user interface

Describes a few frequently used interface elements of Cloudera Observability that help you identify and troubleshoot your workload issues.

The following examples describe a few interface elements of Cloudera Observability that enable you to quickly identify workload problems, health issues, resource contentions, and abnormal or degraded performance problems.

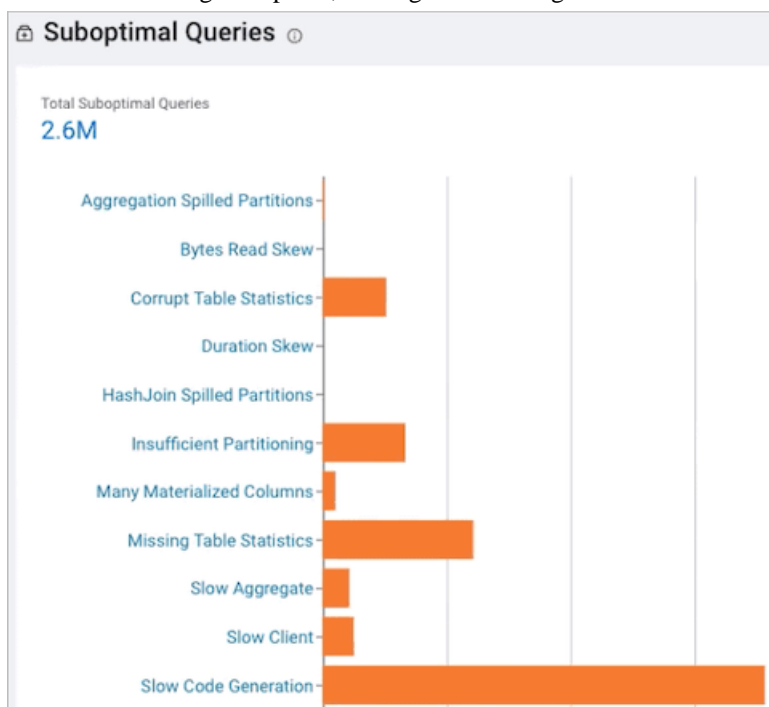
These examples assume that you have logged in to the Cloudera Data Platform and accessed the Cloudera Observability web UI from the Your Enterprise Data Cloud landing page.

Identify workload problems and health issues

You can easily locate what engines are running on your clusters from the environment's Cluster Summary page and what jobs and queries are failing the health tests with the Suboptimal chart widget.

The Suboptimal chart widget, displays the distribution of jobs and the jobs and queries that failed. This chart widget enables you to visually see at a glance what issues are currently impacting your jobs or queries and how they are executing on your cluster.

Location: The Suboptimal chart widget is found on the engine's Summary page, which, depending on the environment selected, is accessed by selecting the environment for analysis in the **Environments** page and then in the Environment navigation panel, drilling down through the environment's categories and selecting an engine of interest.

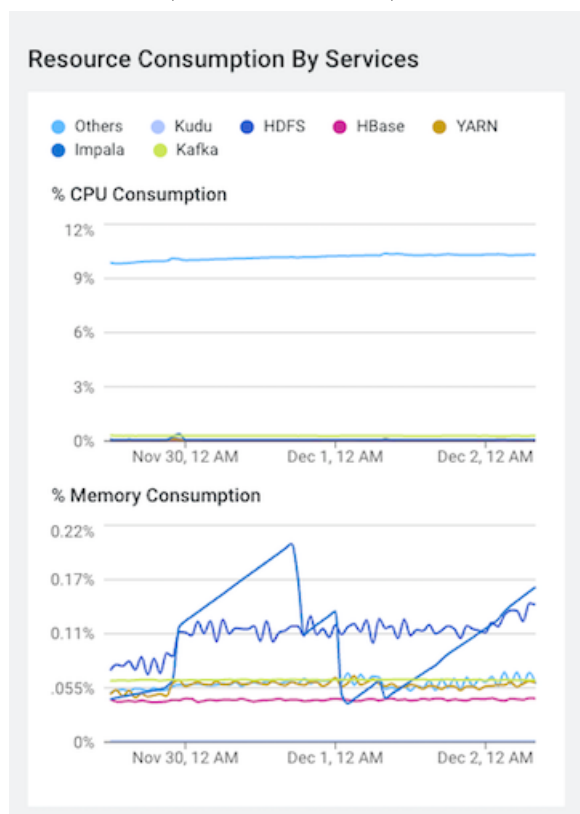


Identify and address resource contentions

Cloudera Observability provides the following chart widgets that help you analyze and identify resource consumption and contention problems:

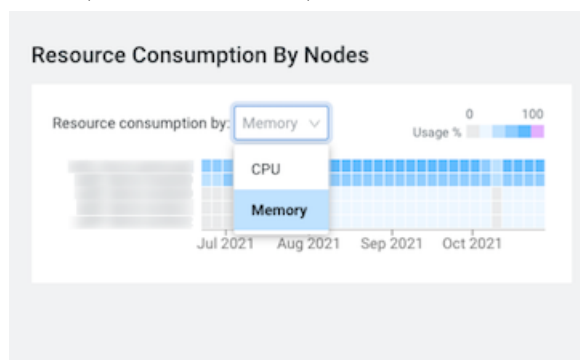
- The Resource Consumption By Services chart widget displays the CPU and memory consumption for each service across the time range you selected. Hover your mouse over the timeline, to display the percentage of CPU or memory consumed by each of the cluster's services.

Location: The Resource Consumption By Services chart widget is found on the **Cluster Summary** page of a Classic Cluster, Private Cloud Base, and a Data Hub cluster.



- The Resource Consumption By Nodes chart widget displays the CPU and memory consumption for each node in the cluster. Hover your mouse over the time line, to display the amount of CPU or memory, as a percentage, that is consumed by each node and its services.

Location: The Resource Consumption By Nodes chart widget is found on the **Cluster Summary** page of a Classic Cluster, Private Cloud Base, and a Data Hub cluster.

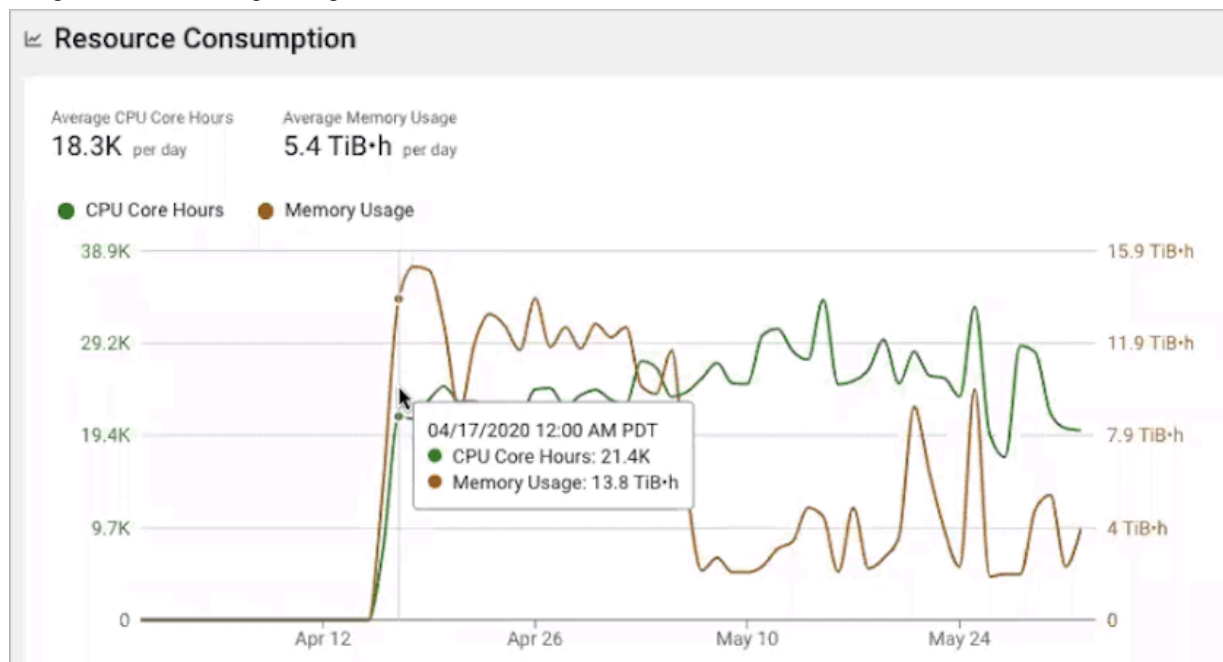


- The Memory Utilization chart widget, displays the aggregated maximum amount of memory that is used by the queries on any node performing the processing. It helps you identify inefficient queries that are consuming the most amount of memory and decide if you need to allocate more memory to continue running your query jobs.

Location: The Memory Utilization chart widget is found on the Impala engine's Summary page, which, depending on the environment selected, is accessed by selecting the environment for analysis in the **Environments** page and then in the Environment navigation panel, drilling down through the environment's categories and selecting an engine of interest.

- The Resource Consumption chart widget, displays the concurrent use of CPU and memory consumption for a workload across the timeline you selected.

Location: The Resource Consumption chart widget is found on the Impala engine's Summary page, which, depending on the environment selected, is accessed by selecting the environment for analysis in the **Environments** page and then in the Environment navigation panel, drilling down through the environment's categories and selecting an engine of interest.



Identify and address abnormal or degraded performance problems

Cloudera Observability enables you to identify and address abnormal or degraded performance problems by establishing baselines from health issues that also enable a performance comparison of your workloads. The Cloudera Observability baseline metrics measure the current performance of a job against the average performance of previous runs. They use performance data from 30 of the most recent runs of a job and require a minimum of three runs. The baseline comparisons start with the fourth run of a job.

Location: The baseline for a job or a query is found on the Baseline page, which is accessed by selecting the environment, followed by the cluster and then the engine for analysis, then depending on the engine, clicking the Total Jobs or Total Queries in the Job Trend chart widget, selecting the job or query of interest, and then selecting the Baseline tab.



Note: Cloudera Observability requires at least four job runs in order to create the job or query's baseline.

The following images show a few of the baseline and comparison metrics that are provided:

- This image shows the comparison between the baseline performance metrics and the current job run:

Metric	Baseline	Current Job	
HIVE_EXEC_PERFORMANCE			
TezBuildDag	< 1s	< 1s	-26ms -15%
TezCompiler	< 1s	< 1s	+4ms 6%
TezCreateVertex.Map 1	< 1s	< 1s	-1ms -2%
TezCreateVertex.Map 3	< 1s	< 1s	-6ms -11%
TezCreateVertex.Map 4	< 1s	< 1s	-12ms -19%
TezGetSession	< 1s	< 1s	+1ms 7%
TezRunDag	34m 26s	47m 30s	+13m 4s 38%
TezRunVertex.Map 1	34m 17s	47m 22s	+13m 5s 38%
TezRunVertex.Map 3	16s	6s	-10s -65%
TezRunVertex.Map 4	14s	12s	-2s -16%
TezSubmitDag	13s	2s	-11s -87%
TezSubmitToRunningDag	< 1s	< 1s	-86ms -65%

- To display only those metrics with performance issues, select Show only abnormal metrics:

Q Search	Show only abnormal metrics		
Metric	Baseline	Current Job	
HIVE_EXEC_PERFORMANCE			
acquireReadWriteLocks	< 1s	0s	-8ms -100%
PreHook.org.apache.hadoop.hive.ql.hooks.HiveProtoLoggingHook	< 1s	< 1s	-4ms -44%
runTasks	34m 34s	47m 32s	+12m 58s 37%
TezRunDag	34m 26s	47m 30s	+13m 4s 38%
TezRunVertex.Map 1	34m 17s	47m 22s	+13m 5s 38%
TezRunVertex.Map 3	16s	6s	-10s -65%
TezSubmitDag	13s	2s	-11s -87%
TezSubmitToRunningDag	< 1s	< 1s	-86ms -65%
Ungrouped			
Duration	34m 34s	47m 32s	+12m 58s 37%

Identify performance trends

You can identify trends as well as baselines by analyzing your engine's or cluster's performance trends from the Trends chart widget and the Trend tab. Where:

- The engine's job or query Trends time-series chart widget, displays more detailed metrics about the processed jobs and queries and enables you to view historical trends for analysis when you select a predefined or custom time period from the Time-Range filter list.

Location: This chart widget is found on the Cluster Summary and the engine Summary pages, which are accessed by selecting the environment and then the cluster for analysis or by selecting the environment, followed by the cluster, and then the engine for analysis.

- The Trends tab, displays the job or query's instances executed during the selected time period. Depending on the engine, the Trends page displays a job's historical trend from Duration, Data Input, and Data Output histogram charts or lists the runs of the query to show how its performance changes overtime.

Location: The Trends tab is found on the Jobs or Query's page, which is accessed by selecting the environment, followed by the cluster and then the engine for analysis, then depending on the engine, clicking the Total Jobs or Total Queries in the Job Trend chart widget, selecting the job or query of interest, and then selecting the Trends tab.

Collecting Cloudera Observability diagnostic metrics for Public Cloud

When you enable Workload Analytics for Cloudera Observability, the Cloudera Management Service starts the Telemetry Publisher and the Databus WXM Client role. Telemetry Publisher and Databus WXM Client collect and transmit metrics, as well as configuration and log files, to Cloudera Observability from Impala, Oozie, Hive, YARN, and Spark services for jobs running on your clusters. Telemetry Publisher and Databus WXM Client collect metrics for all the clusters that use Cloudera Observability-enabled environments.

Understanding the sources of information sent to Cloudera Observability and how that data is redacted is described in the following topics.



Note: The collected diagnostic data is managed by Cloudera and stored in S3 and DynamoDB with a typical retention of 180 days. Where, data retention is dependent on how long an environment exists and its data is accessible. By default, all data stored in S3 and DynamoDB is encrypted.

Metric sources sent to Cloudera Observability

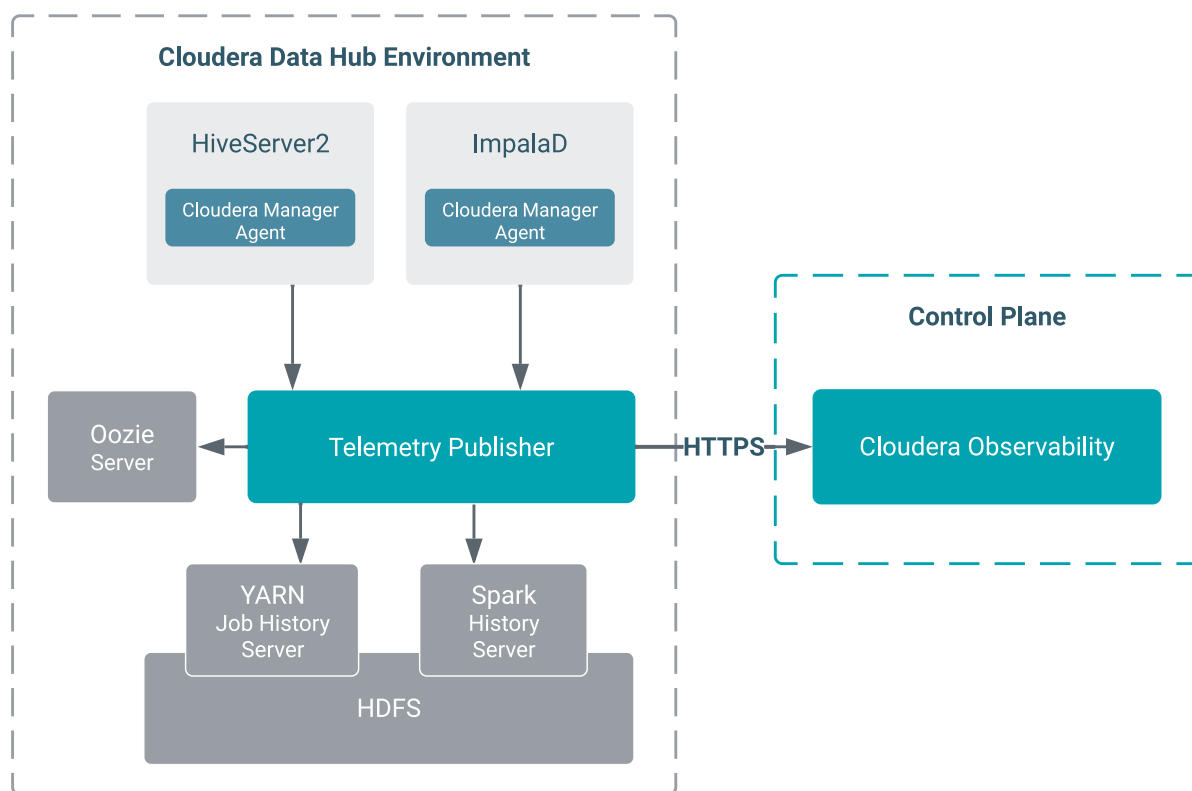
Describes the resources from which Telemetry Publisher and Databus WXM Client collects diagnostic metrics.

Telemetry Publisher and Databus WXM Client collect metrics, as well as configuration and log files, from Impala, Oozie, Hive, YARN, and Spark services for jobs running on your clusters and transmits this information to Cloudera Observability.

The following example, describes how metrics are collected from a Data Hub Public Cloud environment:

- Pull — Telemetry Publisher pulls diagnostic metrics from Oozie, YARN, and Spark periodically (by default, once per minute).
- Push — An Agent pushes diagnostic data from Hive and Impala to Telemetry Publisher within 5 seconds after a job finishes.

The following diagram, shows an example of a Data Hub Public Cloud environment:



After the diagnostic data reaches Telemetry Publisher or Databus WXM Client, it is stored temporarily in its data directory and periodically (once per minute) exported to Cloudera Observability.

Diagnostic metrics collection details

Describes the type of data provided by the Cloudera data services and collected by Telemetry Publisher and Databus WXM Client.

Telemetry Publisher and Databus WXM Client collect and send the following diagnostic metrics and data to Cloudera Observability:

- **Cloudera Manager Metrics** — The Telemetry Publisher agent pulls a subset of Cloudera Manager metrics from the Cloudera Manager API endpoint in Data Hub clusters and sends it to Cloudera Observability. For more information, click the Related Information link below.
- **Cloudera Manager Events** — The Telemetry Publisher agent pulls the Cloudera Manager Events from the Cloudera Manager API endpoint in Data Hub clusters and sends its diagnostic data to Cloudera Observability. For more information, click the Related Information link below.
- **Hive MetaStore (HMS) data source** — The Telemetry Publisher or Databus Producer agent polls Hive and Impala for HMS metadata about your tables and their database and sends the details to Cloudera Observability. This data includes the table's schema, database location, partitions, structure and relationships, columns, column names and their data types, and the table's metadata properties that include user-defined and predefined key-value pairs.
- **Hive Queries** — An agent periodically searches for query detail files that are generated by HiveServer2 after a query completes and then sends the details from those files to Telemetry Publisher or Databus Producer.



Important: Hive query audits must be enabled.

- **Impala Queries** — An agent periodically looks for query profiles of recently completed queries and sends them to Telemetry Publisher and Databus Producer.

- **MapReduce Jobs** — Telemetry Publisher and Databus Producer poll the YARN Job History Server for recently completed MapReduce jobs. For each of these jobs, Telemetry Publisher and Databus Producer collects the configuration and jhist file, which is the job history file that contains job and task counters, from HDFS. Telemetry Publisher and Databus Producer can be configured to collect MapReduce task logs from HDFS and send them to Cloudera Observability. By default, this log collection is turned off.
- **Oozie Workflows** — Telemetry Publisher and Databus Producer polls Oozie servers for recently completed Oozie workflows and sends the details to Cloudera Observability.
- **Spark Applications** — Telemetry Publisher and Databus Producer poll the Spark History Server for recently completed Spark applications. For each of these applications, Telemetry Publisher and Databus Producer collect their event log from HDFS. You can configure Telemetry Publisher to collect the executor logs of Spark applications from HDFS and send them to Cloudera Observability. By default, this data collection is turned off.

Related Information

[Cloudera Manager Metrics](#)

[Cloudera Manager Events](#)

Enabling the redaction of sensitive data

Telemetry Publisher and Databus Producer collect diagnostic data from logs, job configurations, and SQL queries, and then send this data to Cloudera Observability. As this diagnostic information may contain sensitive information it is important to mask this data before it is sent to Cloudera Observability.

About this task

Describes how to redact your sensitive data in the Cloudera Management Console.

By default, the Cloudera Management Console provides the following default anonymization rules that mask the following using regular expression patterns and replacement strings:

- email addresses
- credit card numbers
- Social Security numbers



Important: Cloudera recommends enabling redaction even if you are not sending diagnostic data to Telemetry Publisher or Databus Producer.

Procedure

1. Verify that you are logged in to the Cloudera Data Platform.
2. From the Your Enterprise Data Cloud landing page, select the Management Console tile.
3. From the Navigation panel, select Global Settings and then the Telemetry tab.
4. Verify that the Workload Analytics and the Deployment Cluster Logs Collection options are enabled.
5. In the Anonymization rules section, click New rule and add your regex pattern and its replacement string.
6. Test your rule in the Test rules section, by doing the following:
 - a. In the Input test text box, enter an example string that contains sensitive content.
 - b. Click Test all rules.

The Anonymized result text box is populated with your example and its sensitive data is masked by the replacement string that you defined in your anonymization rule.

7. Add more rules.
8. Click Save Changes.

Collecting Cloudera Observability diagnostic metrics for Private Cloud

When you enable the Telemetry Publisher service for Cloudera Observability, the Cloudera Management Service starts the Telemetry Publisher role. Cloudera Telemetry Publisher collects and transmits metrics, as well as configuration and log files, from Impala, Oozie, Hive, YARN, and Spark services for jobs running on your clusters to Cloudera Observability. Telemetry Publisher collects metrics for all the clusters that use Cloudera Observability-enabled environments.

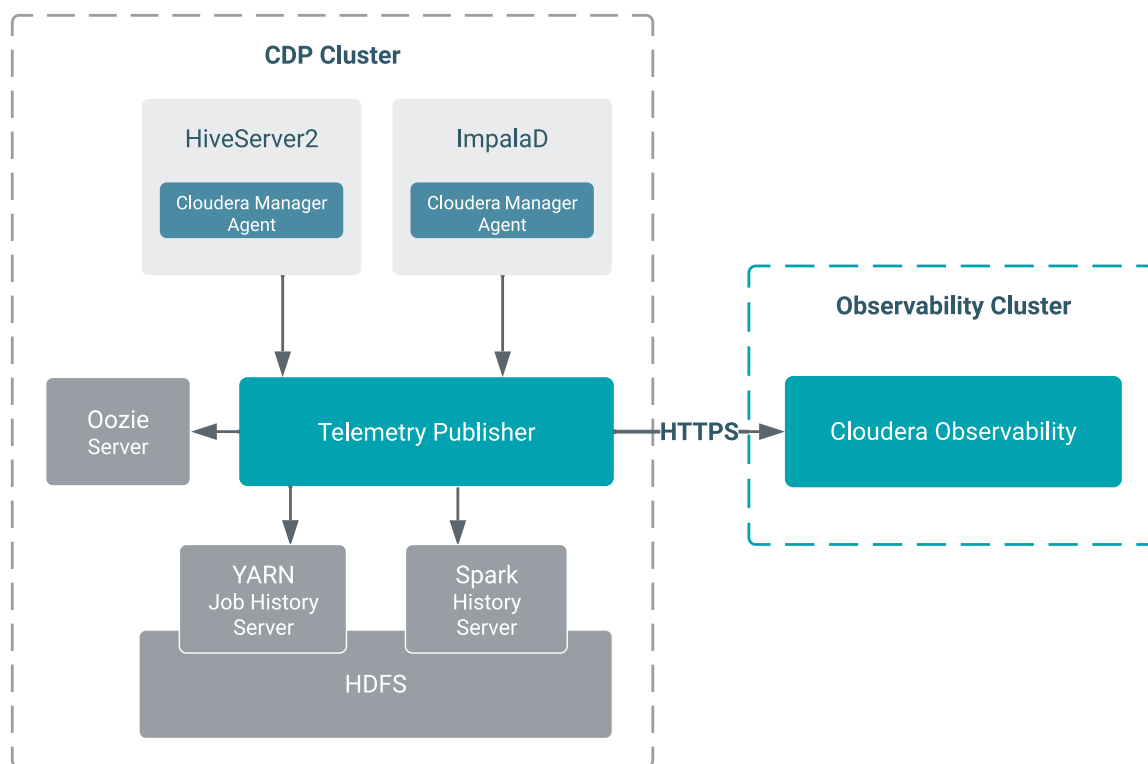
Understanding the sources of information sent to Cloudera Observability and how that data is redacted is described in the following topics.



Note: The collected diagnostic data is managed by Cloudera and stored in S3 and DynamoDB with a typical retention of 180 days. Where, data retention is dependent on how long an environment exists and its data is accessible. By default, all data stored in S3 and DynamoDB is encrypted.

Metric sources sent to Cloudera Observability

Describes the resources from which you can configure Telemetry Publisher to collect diagnostic metrics.



Telemetry Publisher collects and transmits metrics as well as configuration and log files from Impala, Oozie, Hive, YARN, and Spark services for jobs running on your clusters to Cloudera Observability, as shown in the above diagram. The metrics are collected as follows:

- Pull — Telemetry Publisher pulls diagnostic metrics from Oozie, YARN, and Spark periodically (by default, once per minute).
- Push — A Cloudera Manager Agent pushes diagnostic data from Hive and Impala to Telemetry Publisher within 5 seconds after a job finishes.

After the diagnostic data reaches Telemetry Publisher, it is stored temporarily in its data directory and periodically (once per minute) exported to Cloudera Observability.

Diagnostic metrics collection details

Describes the type of data provided by the Cloudera data services and collected by Telemetry Publisher.

Telemetry Publisher collects and sends the following diagnostic metrics and data to Cloudera Observability:

- **Cloudera Manager Metrics** — The Telemetry Publisher agent pulls a subset of Cloudera Manager metrics from the Cloudera Manager API endpoint in Private Cloud Base clusters and sends them to Cloudera Observability. For more information, click the [Related Information](#) link below.
- **Cloudera Manager Events** — The Telemetry Publisher agent pulls the Cloudera Manager Events from the Cloudera Manager API endpoint in Private Cloud Base clusters and sends them to Cloudera Observability. For more information, click the [Related Information](#) link below.
- **Hive MetaStore (HMS) data source** — Telemetry Publisher polls Hive and Impala for HMS metadata about your tables and their database and sends the details to Cloudera Observability. This data includes the table's schema, database location, partitions, structure and relationships, columns, column names and their data types, and the table's metadata properties that include user-defined and predefined key-value pairs.



Important: The Hive Metastore (HMS) must be deployed in the Workload cluster.

- **Hive Queries** — The Cloudera Manager agent periodically searches for query detail files that are generated by HiveServer2 after a query completes, and then sends the details from those files to Telemetry Publisher.



Important: Hive query audits must be enabled.

- **Impala Queries** — A Cloudera Manager agent periodically looks for query profiles of recently completed queries and sends them to Telemetry Publisher.
- **MapReduce Jobs** — Telemetry Publisher polls the YARN Job History Server for recently completed MapReduce jobs. For each of these jobs, Telemetry Publisher collects the configuration and jhist file, which is the job history file that contains job and task counters, from HDFS. Telemetry Publisher can be configured to collect MapReduce task logs from HDFS and send them to Cloudera Observability. By default, this log collection is turned off.
- **Oozie Workflows** — Telemetry Publisher polls Oozie servers for recently completed Oozie workflows and sends the details to Cloudera Observability.
- **Spark Applications** — Telemetry Publisher polls the Spark History Server for recently completed Spark applications. For each of these applications, Telemetry Publisher collects the event log from HDFS. You can configure Telemetry Publisher to collect the executor logs of Spark applications from HDFS and send them to Cloudera Observability. By default, this data collection is turned off.



Important:

To collect Spark application data from Apache Spark 3.x versions, Telemetry Publisher requires Cloudera Manager version 7.11.0 or above.

For CDH 5.x clusters, Telemetry Publisher requires the CDS 2.2 Powered by Apache Spark release 2 or later, which is packaged with Apache Spark version 1.6.

Related Information

[Cloudera Manager Metrics](#)

[Cloudera Manager Event Schema Reference](#)

Redaction capabilities for diagnostic data

Describes the resources that you can configure for redaction. Cloudera recommends enabling redaction even if you are not sending diagnostic data to Telemetry Publisher.

The diagnostic data collected by Telemetry Publisher may contain sensitive information in job configuration or log files. The following lists the data and resources that you can configure for redacting sensitive data before it is sent to Telemetry Publisher:

- Log and query redaction — You can redact information in logs and queries collected by Telemetry Publisher based on filters created with regular expressions.
- MapReduce job properties redaction — You can redact job configuration properties before they are stored in HDFS. Since Telemetry Publisher reads the job configuration files from HDFS, it only fetches redacted configuration information.
- Spark event and executor log redaction — The `spark.redaction.regex` configuration property is used to redact sensitive data from event and executor logs in your YARN service. When this configuration property is enabled, Telemetry Publisher sends only redaction data to Cloudera Observability. By default, this configuration property is enabled, but it can be overridden by using the advanced configuration snippet in Cloudera Manager or in the Spark application itself.

Related Information

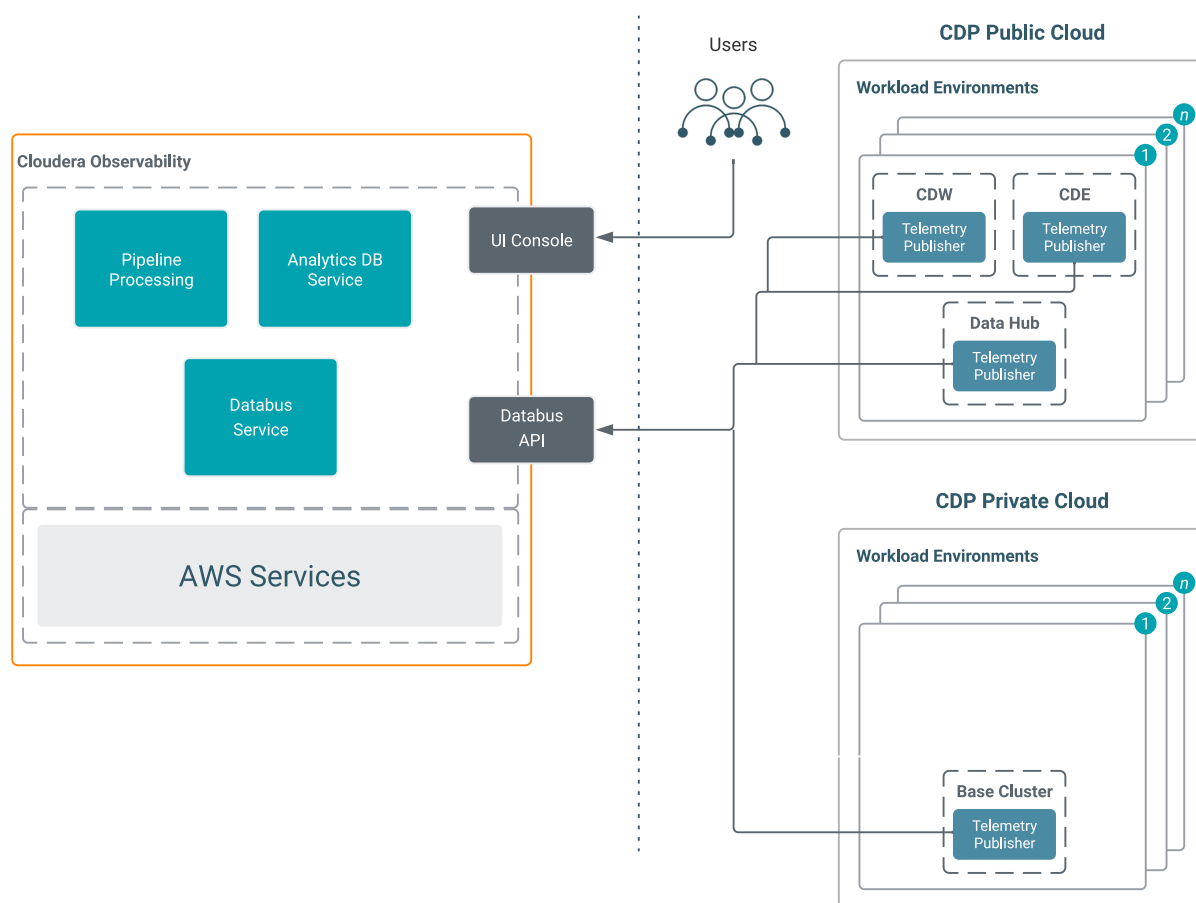
[Redacting data](#)

Considerations related to telemetry tasks for your environment's deployment

Your environment's form-factor (CDP Public Cloud or CDP Private Cloud) and where your workload cluster is hosted determines how you enable telemetry for Cloudera Observability. This topic explains how you decide the telemetry tasks for your environment's form-factor and the steps required for telemetry enablement.

Determining what information to provide and which telemetry configuration steps to choose is dependent on your deployment architecture and where your workload cluster is hosted.

The below diagram shows the communication between Cloudera Observability and your workload clusters in CDP Public Cloud and CDP Private Cloud through Cloudera Manager's Telemetry Publisher. The Cloudera Observability service, including its main component services, runs in the Cloudera Cloud Environment, and your workload jobs run in the Working Environment of your CDP cloud platform that contains the clusters and/or services required to run your workload processes:



CDP Public Cloud Form-factor 1

If your workload cluster is hosted in CDP Public Cloud, you connect to the Cloudera Observability cloud service in the Cloudera cloud environment through Telemetry Publisher as follows:

- If your workload cluster is hosted on one or multiple Data Hub clusters in CDP Public Cloud, you enable telemetry when you register your environment in the Cloudera Management Console by turning on the Enable Workload Analytics option.

For more information on *Enabling Telemetry for Cloudera Observability*, click the Related Information link below.

- If your workload cluster is hosted on one or multiple CDP clusters using Cloudera Data Engineering (CDE) and/or a Cloudera Data Warehouse (CDW) data service in CDP Public Cloud, do the following:
 - For CDW, do nothing. By default telemetry is enabled.
 - For CDE, you enable telemetry when you enable your CDE service by selecting the Enable Workload Analytics check box.

For more information on *Enabling a Cloudera Data Engineering service*, click the Related Information link below.

CDP Private Cloud Form-factor 2

If your workload cluster is hosted in your environment on a Cloudera Data Platform (CDP) Private Cloud Base cluster and you want to connect it to the Cloudera Observability cloud service in the Cloudera cloud environment, you configure the Telemetry Publisher service using the Databus endpoint EC2 service URL for your cloud region:

- For a Us-west-2 cloud region use:

```
https://dbusapi.us-west-1.sigma.altus.cloudera.com
```

- For a EU-central-1 cloud region use:

```
https://dbusapi.eu-1.cdp.cloudera.com
```

- For a AP-southeast-2 cloud region use:

```
https://dbusapi.ap-1.cdp.cloudera.com
```

To enable telemetry for CDP Private Cloud the following tasks must be performed:

1. Ensure your firewall rules will allow communication to the endpoint. For more information on *Configuring Your Firewall*, click the Related Information link below.
2. Generate the Telemetry Publisher access credentials. For more information on *Generating Telemetry Publisher Access Credentials*, click the Related Information link below.
3. Enable the Telemetry Publisher service on each of your workload clusters using the Telemetry Publisher access credentials you generated in task 2. For more information on *Configuring Telemetry Publisher*, click the Related Information link below.



Important:

- To display the most current diagnostic metrics and health statistics collected by Telemetry Publisher, you must upgrade to the latest version of Cloudera Manager and restart Telemetry Publisher.
- For environments using Cloudera Data Engineering (CDE) and Cloudera Data Warehouse (CDW), you must upgrade to CDE 1.19 and/or CDW 1.6.3 or above. Starting with these versions the collected diagnostic data is categorized and displayed within their Data Service category in the Environment navigation panel of the Cloudera Observability web UI.

For environments using older CDE and CDW versions and/or where telemetry was collected from Data Hub environments before April 30th 2023, the diagnostic data is displayed within the **Unclassified Jobs** category in the Environment navigation panel of the Cloudera Observability web UI.

Related Information

[Configuring your firewall](#)

[Generating Telemetry Publisher access credentials](#)

[Configuring Telemetry Publisher](#)

[Enabling telemetry for a Data Hub deployment](#)

[Enabling a Cloudera Data Engineering service](#)

What is Expedited Support

Expedited Support grants read-only access to your workload diagnostic information for the purpose of delivering a direct and expedited troubleshooting experience. Access to only the diagnostic data that is sent to Cloudera Observability from Telemetry Publisher is enabled. The Cloudera Support engineers can not access your workload clusters or any data that is within them.

The Cloudera Observability Expedited Support feature, provides the Cloudera Support team member, who is assisting you with a support case, read-only troubleshooting access to the same details and information that was sent by Telemetry Publisher to Cloudera Observability.

By default, the Expedited Support feature is enabled. When a support case is received, the Support team member assigned to your case, securely logs in to the Cloudera Observability web UI to view your environment's diagnostic data collected by Telemetry Publisher. Cloudera Observability provides essential and historical troubleshooting features and health details about your environment, data services, and workloads, which enables Cloudera Support

to solve your issue without requiring you to collect and send more information, such as diagnostic bundles, logs, and traces.

This results in a direct support response, faster resolution, and minimal downtime.



Note: Cloudera highly recommends retaining enablement for the Expedited Support feature as this can significantly reduce resolution time for any future support tickets. Access to only diagnostic data that is sent to Cloudera Observability is enabled. The Cloudera Support team members are unable to access your workload clusters or any data within them.

Disabling Expedited Support

Describes how to disable the Cloudera Support temporary read-only access to your environment's workload diagnostic information that is collected by Telemetry Publisher for Cloudera Observability.

About this task

By default, the Expedited Support feature is enabled. Learn how to disable read-only access to your workload diagnostic information by Cloudera Support when troubleshooting help is no longer required.

Expedited Support grants read-only troubleshooting access to your Cloudera Support team. It provides access to the same details and information that was sent by Telemetry Publisher to Cloudera Observability. Expedited Support enables the Cloudera Support team member who is assisting you with a support case the ability to review any workload related issues without requiring you to collect and send the workload details. This results in a direct and accelerated troubleshooting experience.



Note: Cloudera highly recommends retaining enablement for the Expedited Support feature as this can significantly reduce resolution time for any future support tickets. Access to only the diagnostic data that is already sent to Cloudera Observability is enabled. The Cloudera Support team members can not access your workload clusters or any data within them.

Before you begin

These steps assume that you have been assigned by your administrator the Cluster Admin access role (ObservabilityClusterAdmin).

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI.
 - a) In a supported browser, log into the Cloudera Data Platform (CDP).
The CDP Cloud web interface landing page opens.
 - b) From the Your Enterprise Data Cloud landing page, select the Observability tile.
The Cloudera Observability landing page opens.
2. From the Cloudera Observability **Environments** page, locate the environment whose diagnostic data no longer requires troubleshooting help from Cloudera Support.



Tip: You can reduce the number of environments by selecting your environment's type from the Environments list.

3. From the environment's Actions list (ellipsis icon), select Disable Expedited Support.

A Cloudera Observability Expedited Support Disabled notification email is sent to the user who initiated the Expedited Support action. The email which confirms the disablement condition, lists the cluster affected by the action, and provides a date and time for when the action was initiated.



Note: Disabling support access for a cluster can significantly increase the resolution time on possible future cluster issues.

Reenabling Expedited Support

Describes how to grant Cloudera Support temporary read-only access to your environment's workload diagnostic information that is collected by Telemetry Publisher for Cloudera Observability.

About this task

By default, the Expedited Support feature is enabled. Learn how to re-enable read-only access to your workload diagnostic information by Cloudera Support when troubleshooting help is required after disabling.

Expedited Support grants read-only troubleshooting access to your Cloudera Support team. It provides access to the same details and information that was sent by Telemetry Publisher to Cloudera Observability. Expedited Support enables the Cloudera Support team member who is assisting you with a support case the ability to review any workload related issues without requiring you to collect and send the workload details with a support case. This results in a direct and accelerated troubleshooting experience.



Note: Cloudera highly recommends retaining enablement for the Expedited Support feature as this can significantly reduce resolution time for any future support tickets. Access to only the diagnostic data that is already sent to Cloudera Observability is enabled. The Cloudera Support team members can not access your workload clusters or any data within them.

Before you begin

These steps assume that you have been assigned the Cluster Admin access role (ObservabilityClusterAdmin) by your administrator.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI.
 - a) In a supported browser, log into the Cloudera Data Platform (CDP).
The CDP Cloud web interface landing page opens.
 - b) From the Your Enterprise Data Cloud landing page, select the Observability tile.
The Cloudera Observability landing page opens.
2. From the Cloudera Observability **Environments** page, locate the environment whose diagnostic data requires troubleshooting help from Cloudera Support



Tip: You can reduce the number of environments by selecting your environment's type from the Environments list.

3. From the environment's Actions list (ellipsis icon), select Enable Expedited Support.
The Enable Expedited Support confirmation message opens.
4. In the Additional Note field, enter a comment that uniquely describes why access is granted. For example, enter the Cloudera Support ticket number with a brief description of the problem.
5. Click Enable.

A Cloudera Observability Expedited Support Enabled notification email is sent to the user who initiated the Expedited Support action. The email confirms the access condition, lists the cluster affected by the action, and provides a date and time for when the action was initiated.