

Workload XM 2.2.1

Workload XM Overview

Date published: 2020-12-04

Date modified: 2021-09-21

CLOUDERA

<https://docs.cloudera.com/>

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

- Understanding Workload XM.....4**
- Collecting Workload XM Diagnostic Metrics.....4**
 - Metric Sources Sent to Workload XM..... 4
 - Diagnostic Metrics Collection Details..... 5
 - Redaction Capabilities for Diagnostic Data.....5
- Key Features of Workload XM..... 6**
 - Default Time Range..... 6
 - Understanding the Clusters Page.....6
 - File Size Reporting.....6
 - Understanding the Baseline Metric..... 7
 - Spark RDD Health Check..... 7
 - Analyze Workloads from Auto-Generated Workload Views..... 7
 - Workload Classification for Deep-dive Analysis..... 8
 - Comparing a Job with the Previous Run..... 9
 - Repair Table Statistic Issues..... 9
 - Suppress Sensitive Data..... 10
 - Proxy Server Support for Telemetry Publisher.....10

Understanding Workload XM

Workload XM is a Cloudera service that helps you understand the workloads that you process, the clusters and services that manage those workloads, and the data that is processed. Workload XM provides information that helps you troubleshoot failed jobs and optimize slow jobs that run on your Workload clusters. When a job completes, information about the job and the cluster that processed the job is sent to Workload XM with Telemetry Publisher, a role in the Cloudera Manager Management Service.

Workload XM also displays metrics about a job's performance and compares the current job run with previous runs by creating baselines. This information helps you identify and address abnormal or degraded performance.

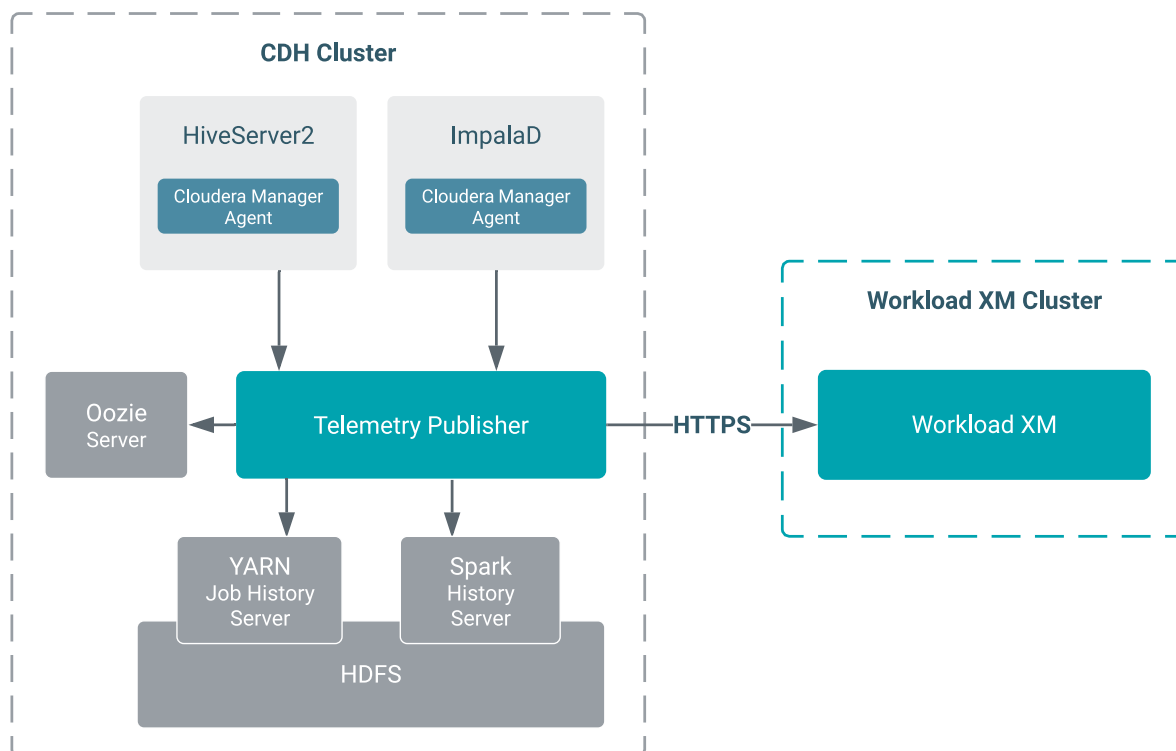
Collecting Workload XM Diagnostic Metrics

When you enable Workload XM, the Cloudera Management Service starts the Telemetry Publisher role. 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 Workload XM. Telemetry Publisher collects metrics for all the clusters that use Workload XM-enabled environments.

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

Metric Sources Sent to Workload XM

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 Workload XM, 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 Workload XM.

Diagnostic Metrics Collection Details

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

Telemetry Publisher collects and sends the following diagnostic metrics to Workload XM:

- 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 Workload XM. By default, this log collection is turned off.
- Spark Applications — Telemetry Publisher polls the Spark History Server for recently completed Spark applications. For each of these applications, Telemetry Publisher collects their event log from HDFS. Telemetry Publisher can be configured to collect the executor logs of Spark applications from HDFS and send them to Workload XM. By default, this data collection is turned off.



Important: Telemetry Publisher only collects Spark application data from Apache Spark 2.x versions. At this time, Apache Spark 3.x versions are not supported.

As CDH version 5.x is packaged with Spark version 1.6, you cannot configure Telemetry Publisher data collection for CDH 5.x clusters unless you are using CDS 2.2 Powered by Apache Spark or later versions with those clusters.

- Oozie Workflows — Telemetry Publisher polls Oozie servers for recently completed Oozie workflows and sends the details to Workload XM.
- 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.

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 Spark2 on YARN service on CDH clusters has the `spark.redaction.regex` configuration property that can be used to redact sensitive data from event and executor logs. When this configuration property is enabled, Telemetry Publisher sends only redaction data to Workload XM. By default, this configuration property is enabled, but it can be overridden by using safety valves in Cloudera Manager or in the Spark application itself.

Key Features of Workload XM

Lists the key features of Workload XM.

Default Time Range

Workload XM enables you to choose a time period in which your workload results are displayed for analysis and troubleshooting.

By default, Workload XM displays data for the last 24 hours. If there is no data available during that time, Workload XM displays the nearest range that is available.

Understanding the Clusters Page

Describes the features in the Clusters page, which lists your clusters, the services that use them, the last time they were updated, and whether you enabled the daily cluster report.

By default, the Clusters page displays your clusters by the date that they were last updated. You can sort your clusters in ascending or descending order from the Cluster Name column.

The image below shows an example of the Clusters page:

Clusters

| <input type="text" value="Search cluster"/> | | | | | |
|---|-----------------------------------|------------------------|--------------|------------------------------|--|
| Cluster Name | Service | Last Updated | Email Report | | |
| Brontosaurus | Data Engineering Data Warehouse | 05/16/2019 12:00 P... | ✓ | Actions ▾ | |
| Sauropoda | Data Engineering Data Warehouse | 05/16/2019 11:00 A... | | Actions ▾ | |
| Dilophosaurus | Data Warehouse | 05/14/2019 5:00 PM ... | | Delete | |
| Troodon | Data Engineering | 05/14/2019 10:00 A... | | Enable Cluster Report Emails | |
| | | | | Actions ▾ | |

File Size Reporting

The file size reporting feature in Workload XM enables you to identify databases and tables where the data is stored inefficiently, such as in small files or partitions, that can cause performance issues.



Important: At this time the Workload XM File Size Report feature is only supported on CDH Workload clusters, version 6.3 to version 7.0, with Cloudera Navigator enabled. CDP Workload clusters are not supported.

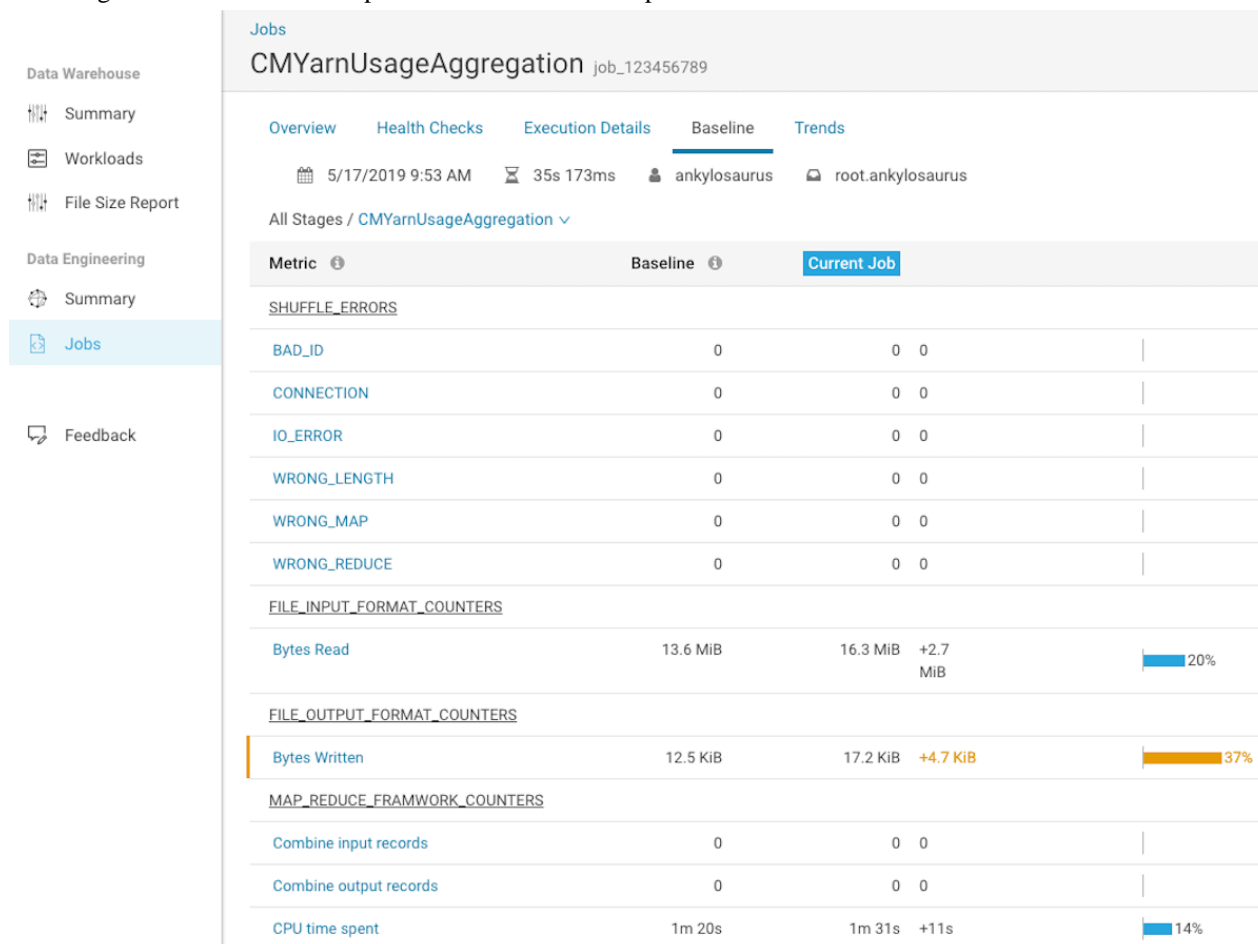
Understanding the Baseline Metric

Describes the Baseline metric feature of the engine's Jobs page.

The Jobs page lists the jobs that were run on your cluster and provides detailed reports about those jobs.

The Baseline information matches the style of the Job Comparison page. Where, the metrics are sorted by their header, and are in alphabetical order.

The image below shows an example of a Baseline metric report:



Spark RDD Health Check

The Spark RDD health check informs you when you have redundant Resilient Distributed Datasets (RDD) cache. Workload XM displays the location of the cache so that you can remove it and save executor memory.

Analyze Workloads from Auto-Generated Workload Views

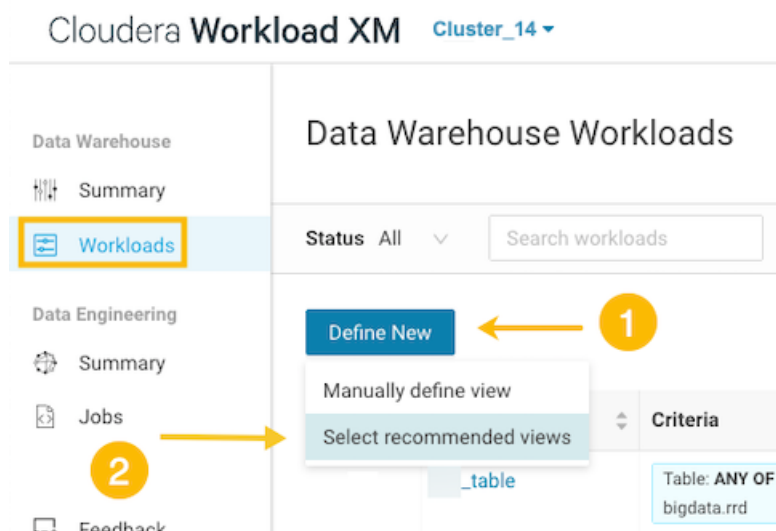
Describes how to access the auto-generated predefined Workload XM workload views, which enable you to immediately start analyzing jobs on your cluster without having to perform an initial analysis that determines which criteria to use when creating a new workload view.

When enabled, recommended predefined workload views are generated for each of your workloads. You can choose to accept or reject a view.

The auto-generated predefined views are based on the following criteria that frequently appear in queries:

- tables
- resource pools
- users who initiate the query

To enable the auto-generated predefined views, from the navigation panel, select Workloads. In the Workloads page, click Auto Generate and then select Select recommended views, as shown in the following image:



Workload Classification for Deep-dive Analysis

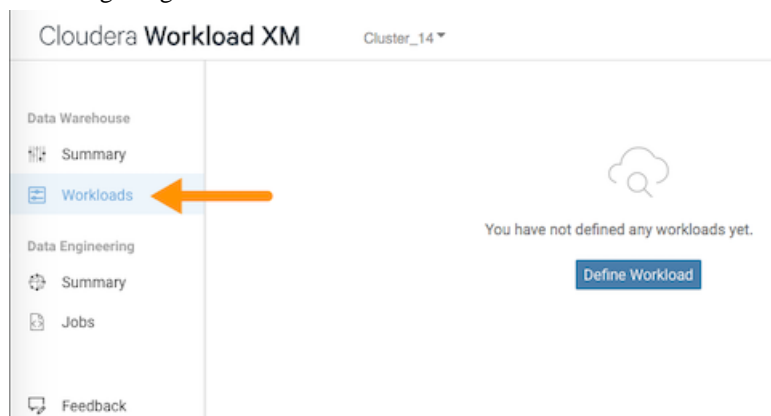
Describes how to access the Workload XM workload view feature where you define your workload views.

The Workload XM Workload Classification feature, enables you to examine your workloads by specific criteria to perform deep-dive analysis on your query statements.

For example you can:

- Determine which users are executing workloads that do not adhere to SLAs.
- Examine how queries are being sent to specific databases.
- Examine which queries are using specific resource pools.
- Examine how your queries are performing against SLAs.

To access this feature, from the navigation panel, select Workloads and then click Define Workload, as shown in the following image:



Comparing a Job with the Previous Run

Describes how to access the Job Comparison feature, which compares two different runs of the same job.

When a job is flagged as slow on the Jobs page, you can click the Compare with previous run link to open the Job Comparison tool and compare the current run with the last run of the job.

The image below shows the location of this link:

Jobs

Log Analysis application_1545261948137_232778

Overview **Health Checks** Execution Details Baseline Trends

📅 3/27/2019 2:16 PM ⏱ 8m 26s 👤 📄

⚠ Job performance can be improved.

📈 **Baseline** [Compare with previous run](#) ⚖ **Skew** 📊 **Resources**

⚠ **Abnormal Duration**

Finished in 8m 26s, **slower** than the median duration 1m 52s.

⚠ **Task Duration Skew**

Some tasks took an abnormal amount of time to finish.

No resource issues found.

Repair Table Statistic Issues

When your queries trigger either the Corrupt Table Statistics or the Missing Table Statistics health checks, Workload XM generates SQL code for you to copy and run on your cluster to address these issues.

To repair table statistics:

1. From the navigation panel, select Impala.
2. From the Trend widget, click the number under Total Queries.
3. On the Queries page, select the time period you want to investigate for the Range column.
4. Filter out queries that do not trigger these health checks, by selecting either Corrupt Table Statistics or Missing Table Statistics in the Health Check column.
5. Select the query to view its details.
6. List the health checks that were triggered for this query by clicking the Health Check Violations tab in the Performance Issues region.

- Copy and run the commands to repair the table statistics issues.

Performance Issues

Potential SQL Issues **Health Check Violations** 3

Optimal Configuration


Slow Client 3m 0s (96.7%)
[+ Details](#)

Metadata/Statistics

Corrupt Table Statistics 1 table >

Run the following commands to address this issue:

```
drop stats sfdc.case_history;  
compute stats sfdc.case_history;
```

[Copy to clipboard](#) 

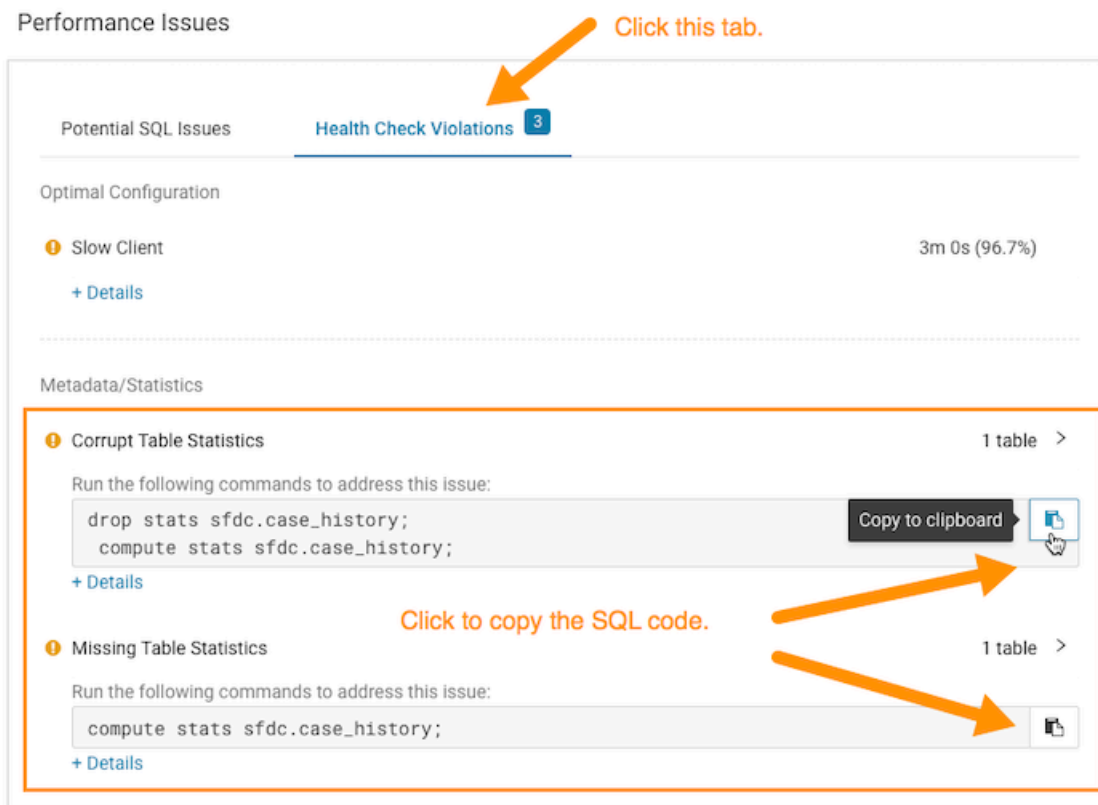
[+ Details](#)

Missing Table Statistics 1 table >

Run the following commands to address this issue:

```
compute stats sfdc.case_history;
```

[+ Details](#)



Suppress Sensitive Data

The suppression of sensitive data is controlled by setting the log and query redaction configuration in Cloudera Manager for the Telemetry Publisher service. By default, this configuration is enabled.

Proxy Server Support for Telemetry Publisher

You can configure the Telemetry Publisher service to send metrics as well as configuration and log files to Workload XM by way of a proxy server for database and Altus metrics uploads. This intermediary gateway also adds extra security when sending your workload data to Workload XM.