Workload XM 2.1.3

Workload XM Overview

Date published: 2020-12-04
Date modified: 2021-07-22

CLOUDERA

Legal Notice

Contents

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. You need to know the sources of information sent to Workload XM and how that data is redacted.

Metric Sources Sent to Workload XM

Describes the resources from which you can configure Telemetry Publisher to collect diagnostic metrics.
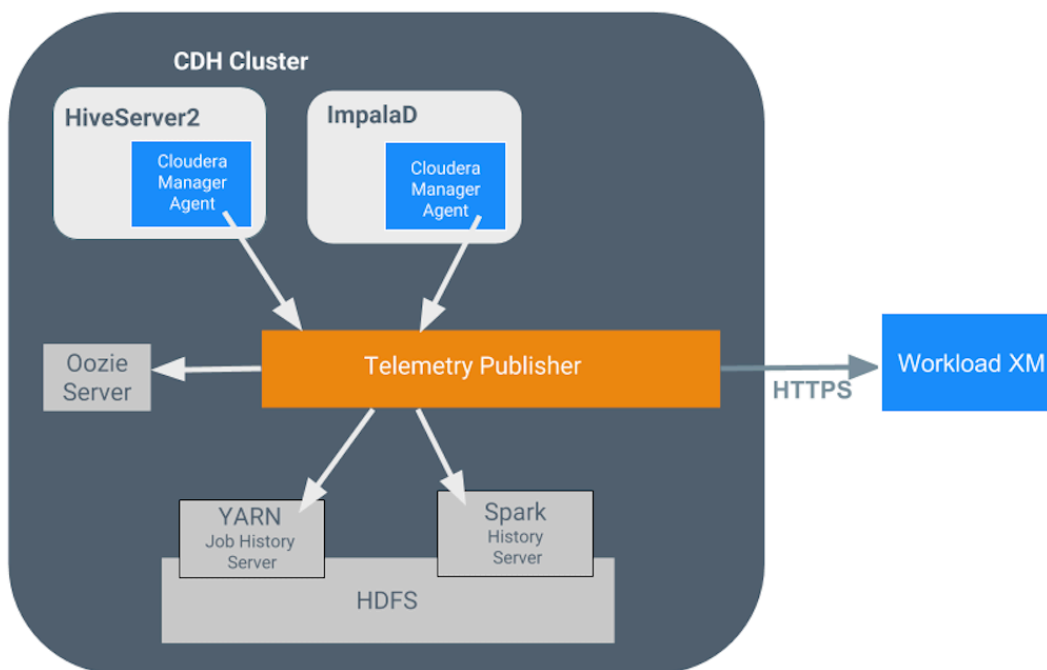
These metrics are collected as follows:

- **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 details to Workload XM.
- **Hive Queries** — The Cloudera Manager agent periodically searches for query detail files that are generated by HiveServer 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.re daction.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 Experience Manager

Lists the key features of Workload Experience Manager.

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:

Understanding the Data Warehouse Summary Page

Describes the features of the Data Warehouse Summary page, which displays the current state of the jobs and queries that were run on your cluster.

The page consists of the following widget charts:

• Trend
• Queries
• Workloads
• Usage Analysis
• Tables
• Statement Types

Note: Previously, the information contained in the Queries, Workloads, and Usage Analysis widgets were displayed as separate tabs within the Outliers widget.

The image below shows an example of the Data Warehouse Summary page:
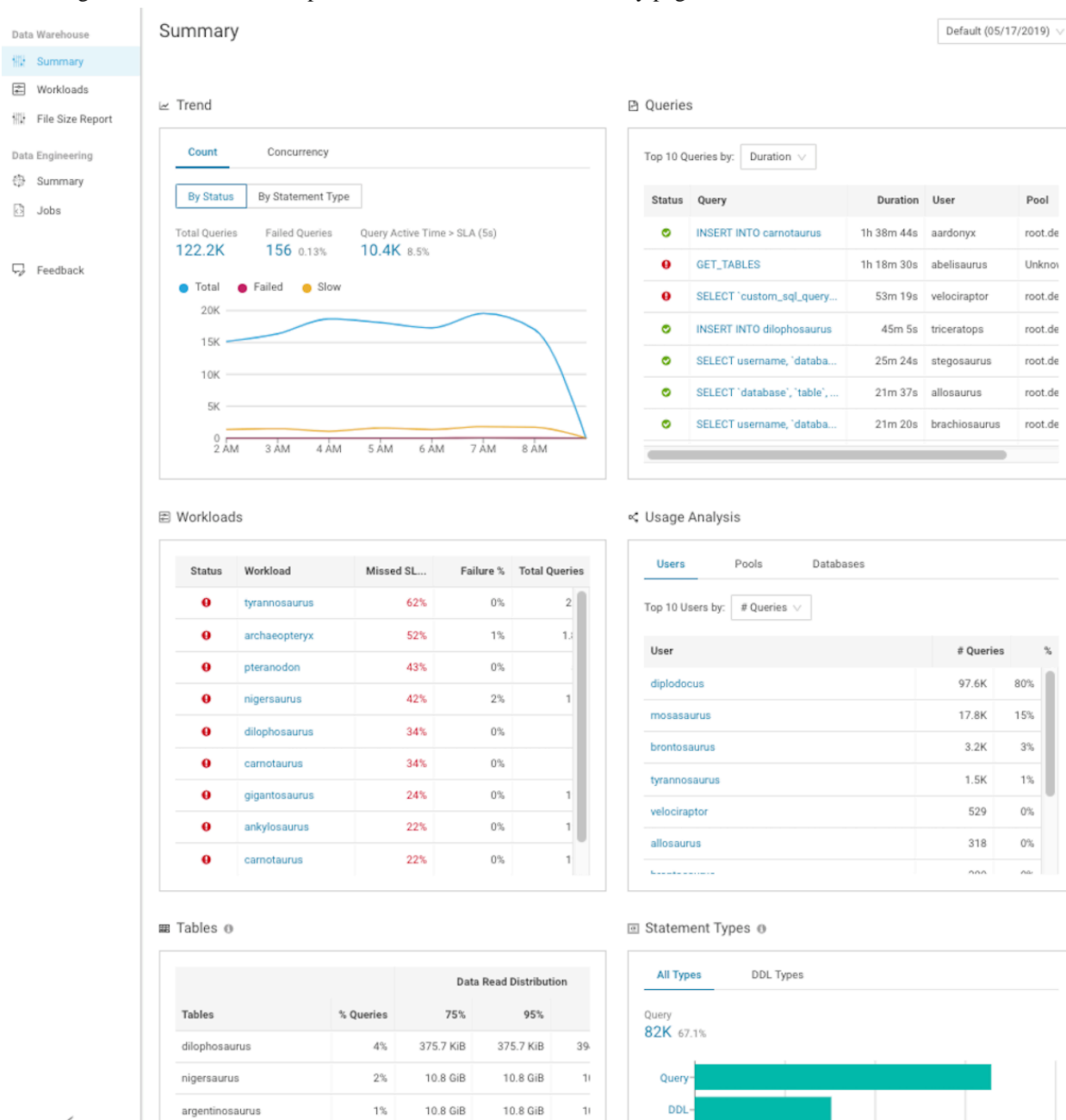
Understanding the Tables Widget

Describes the **Tables** chart, which lists the data table names in your data warehouse and how often they are queried.

The **Tables** widget chart in the Data Warehouse Summary page provides a quick overview of the tables that are predominantly queried in your SQL statements.

The **Data Read Distribution** column, displays the distribution of data, as a percentage, that was read by the queries. When two or more tables are displayed in a single row, it denotes a **join**.

In the following example, the **Data Read Distribution** statistic represents the total amount of data that was read across all the tables in a join clause. Where, the **parrot.employees** table was queried in 1% of the total statements, 75% of the query statements read 209.4 MiB or less data, 95% of the query statements read 395.2 MiB or less data, and the query statement that read the most data accessed 861.5 MiB of data.

⊞ Tables ⓘ

Data Warehouse		Tables	% Queries	Data Read Distribution 75%	95%	Max
▥ Summary						
⊟ Workloads		parrot.employees	1%	209.4 MiB	395.2 MiB	861.5 MiB
▥ File Size Report		penguin.metrics_staging hummingbird.case_history albatross.user_history woodpecker.user_history	0%	1.9 GiB	1.9 GiB	1.9 GiB
Data Engineering						
⊕ Summary		pigeon.user_history	0%	13 MiB	13 MiB	169.3 MiB
⟨⟩ Jobs		owl.employees finch.employees	0%	1.9 GiB	1.9 GiB	1.9 GiB
� Feedback		parrot.user_history	0%	47.2 MiB	374.2 MiB	374.3 MiB

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 Data Engineering Jobs page.

The Data Engineering 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:
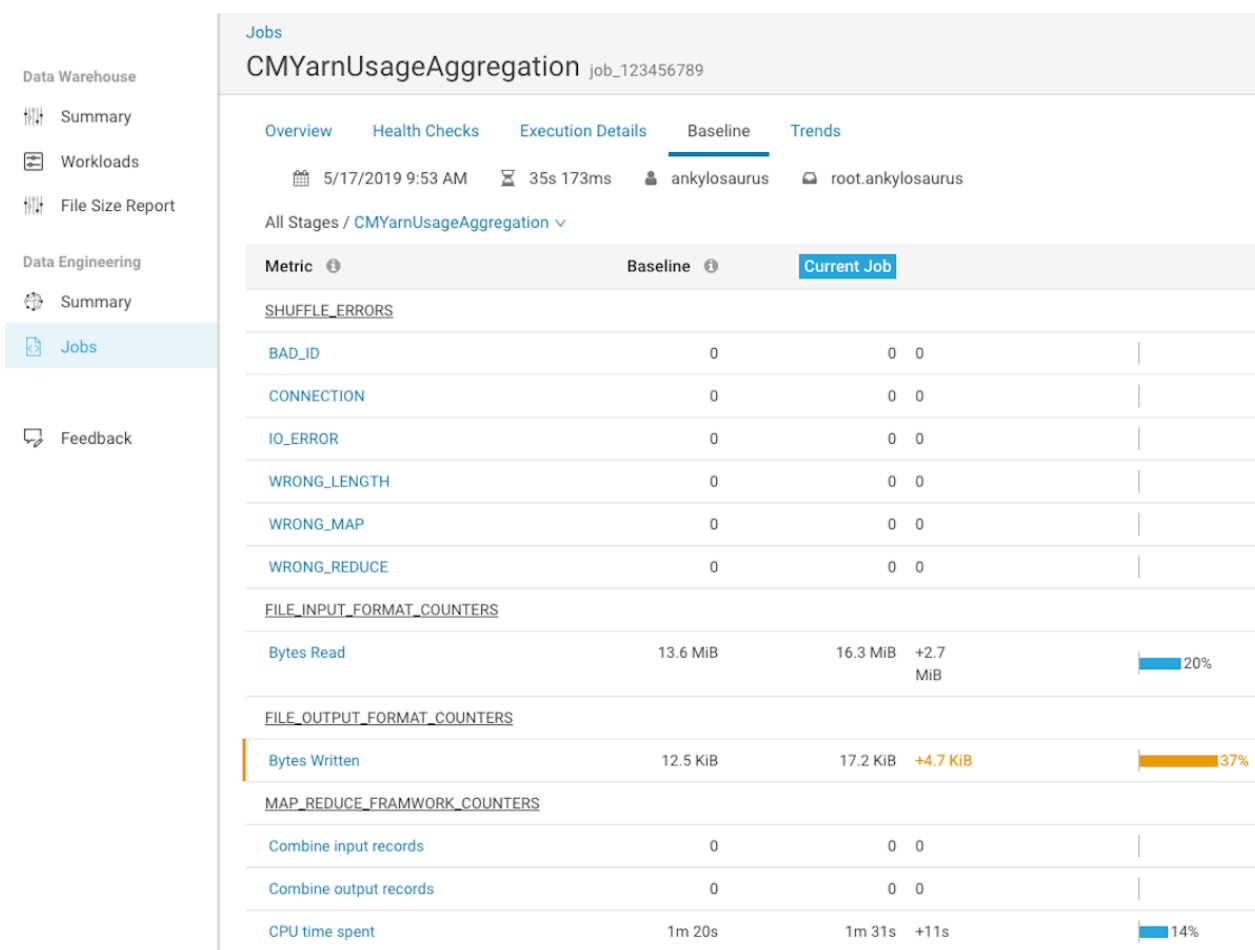
Spark RDD Health Check

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 under Data Warehouse, select **Workloads**. In the Data Warehouse Workloads page, click **Define New** and then select **Select recommended views**, as shown in the following image:
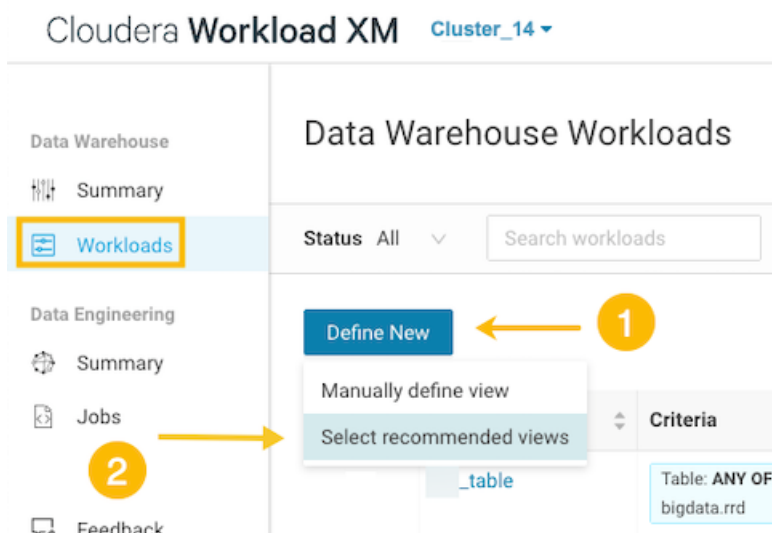
## Workload Classification for Deep-dive Analysis

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 under Data Warehouse, select **Workloads** and then click **Define Workload**, as shown in the following image:
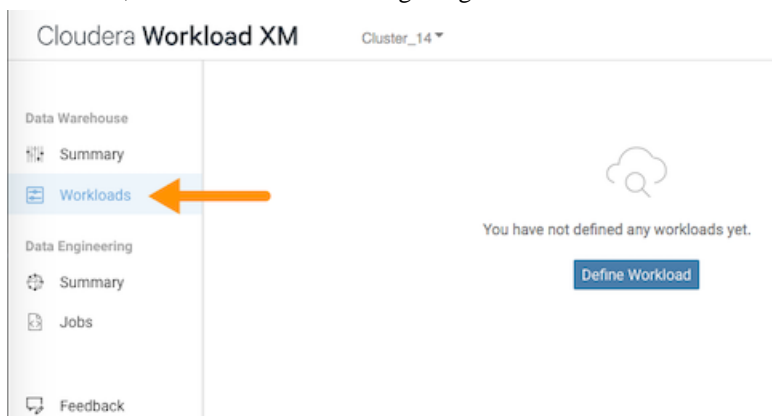

Comparing a Job with the Previous Run

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:
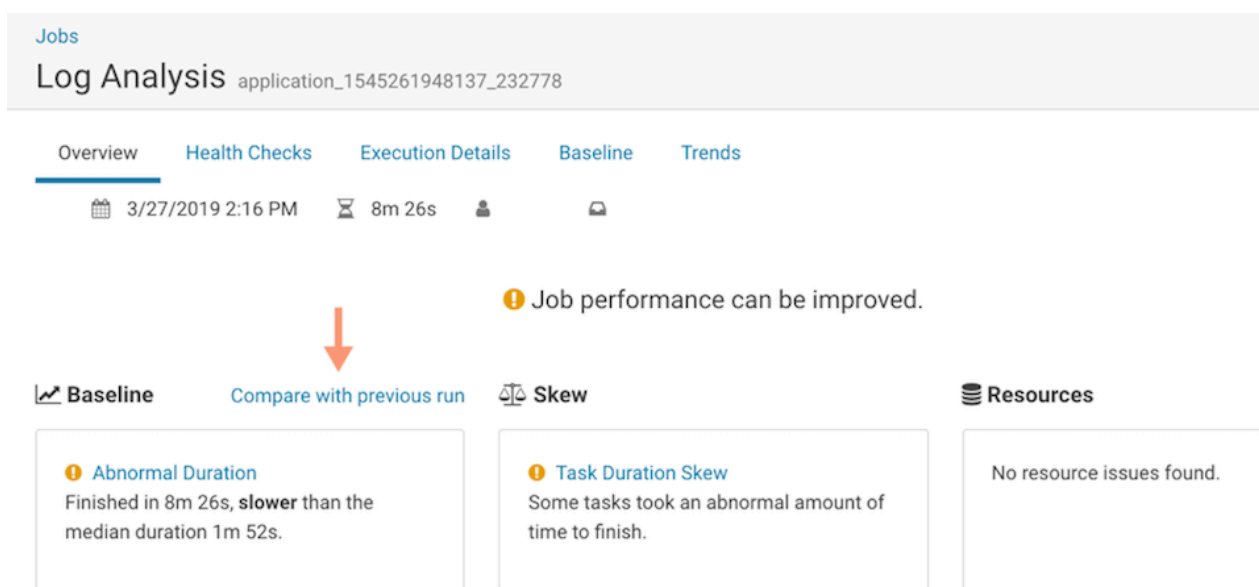
Repair Table Statistic Issues

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 under **Data Warehouse**, select **Queries**.
2. On the Queries page, select the time period you want to investigate for the **Range** column.
3. 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.
4. Select the query to view its details.
5. List the health checks that were triggered for this query by clicking the **Health Check Violations** tab in the **Performance Issues** region.

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

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.

Usability Improvements

Lists the Workload XM usability improvements.

The following usability upgrades for an improved user experience are as follows:

• Support for parsing Spark 2.3 application history logs.

- Job history files and Spark event logs are now downloadable from the **Execution Details** tab in the Jobs detail page:

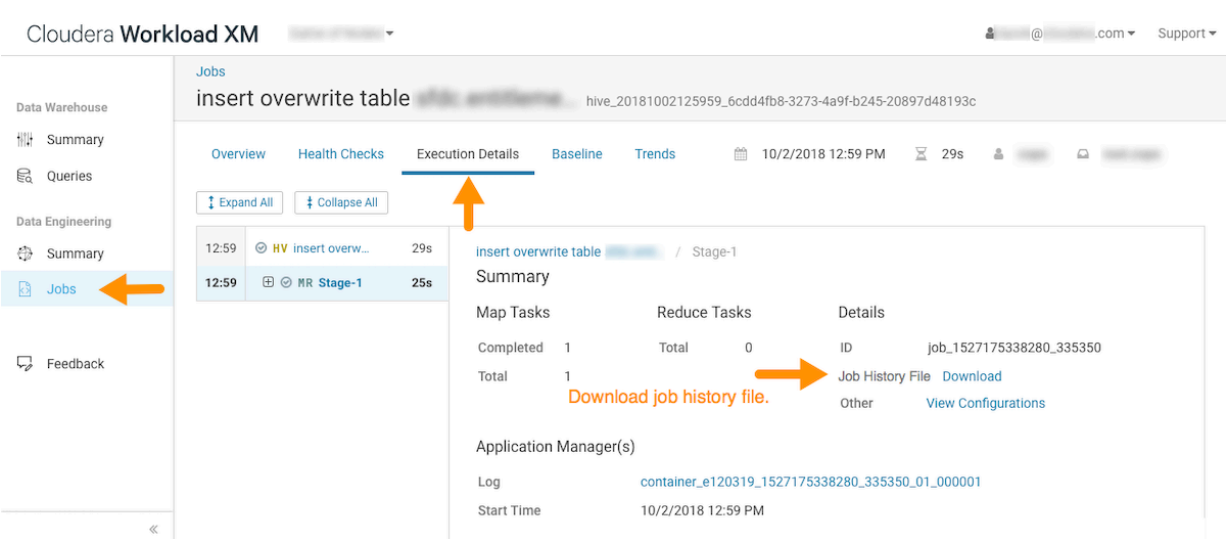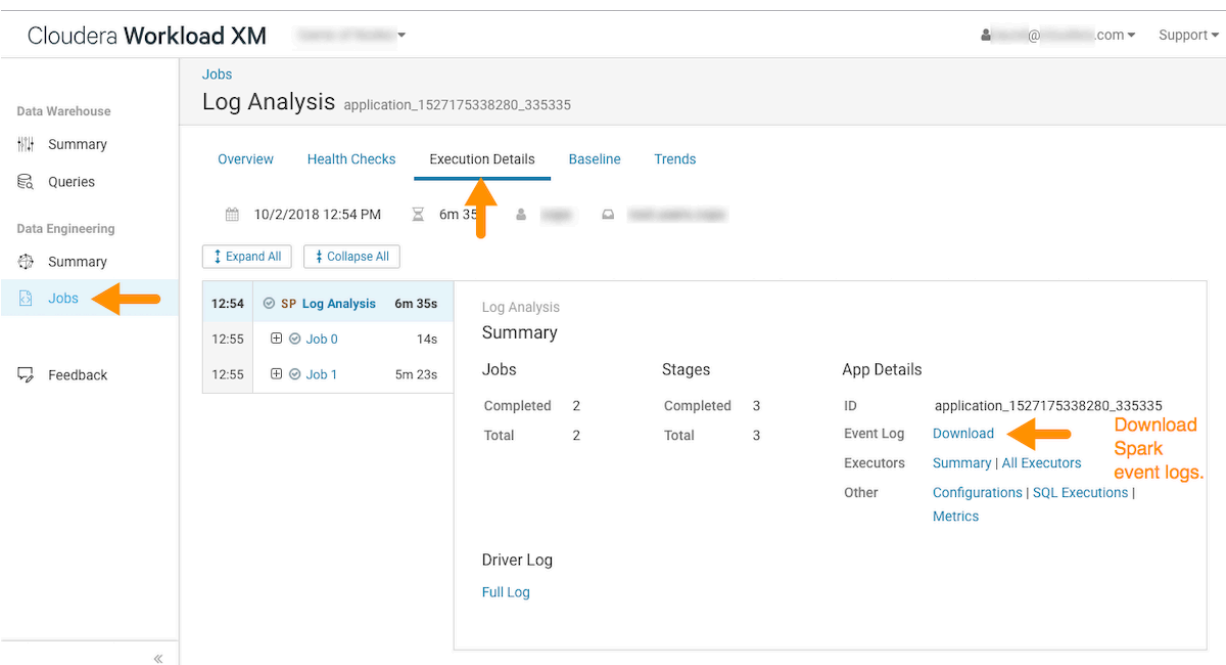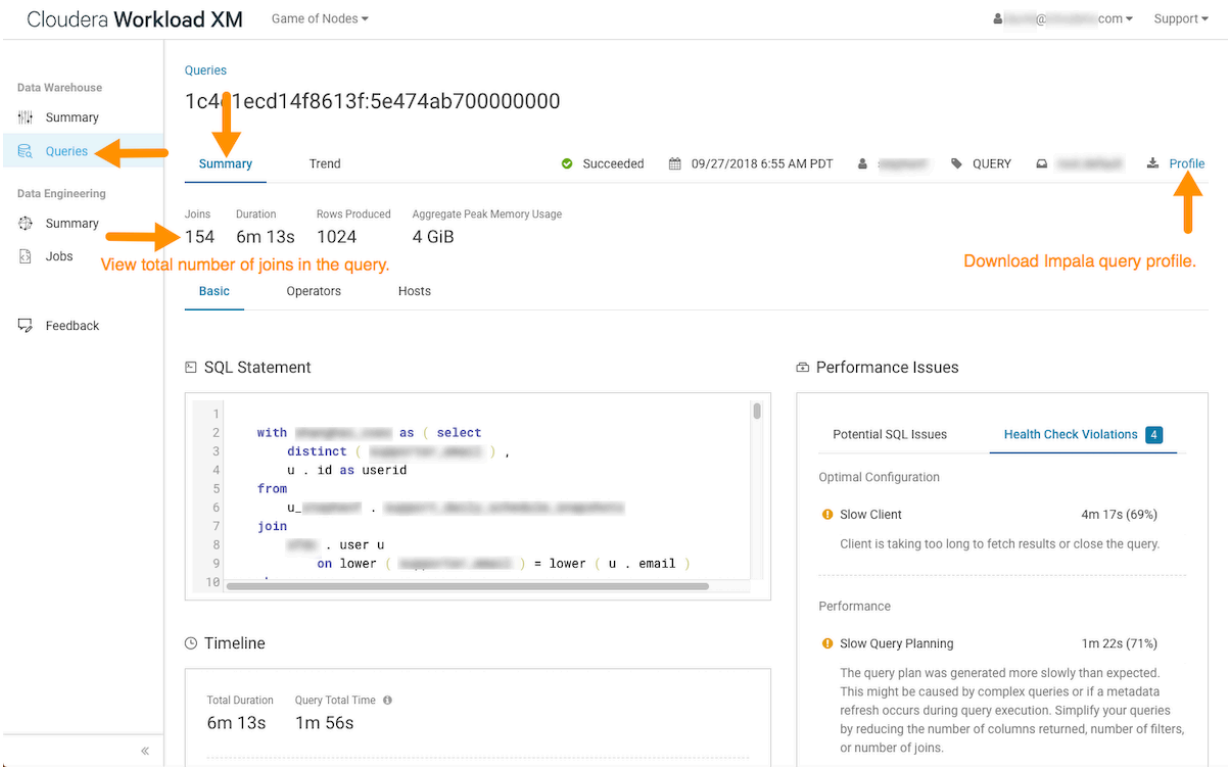**Figure 1: Downloading Job History Files**

Figure 2: Downloading Spark Event Logs

- The **Queries** detail page, enables you to download a query profile for Impala queries and view the total number of JOINS performed for a specific query:

- The **Concurrency** chart in the Data Warehouse Summary page, displays the query concurrency in the cluster during a selected time range. You can use this chart to gain insight, such as identifying a potential resource contention in the cluster or identifying the busiest time of day on your cluster: