Cloudera Observability Cluster Optimization

Date published: 2023-04-31
Date modified: 2024-03-27

https://docs.cloudera.com/

Legal Notice

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.

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 your environment

This section describes how to plan, budget, and forecast costs, identify, troubleshoot, and optimize existing and potential problems, create alerts and enable reports, and manage your users with roles and your data with Workload Views.

Note: To display the most current diagnostic metrics and health statistics collected by Telemetry Publisher in the Cloudera Observability web UI, you must upgrade to the latest version of Cloudera Manager and restart Telemetry Publisher.

Logging in to Cloudera Observability

Learn how to access the Cloudera Observability web user interface to start viewing your diagnostic data for analysis.

About this task

Describes how to access Cloudera Observability and begin working with the Main and the Environment navigation panels.

Note: There can be a delay from job completion to when the job is available in Cloudera Observability, where large jobs can take up to 10 minutes to display.

Before you begin

Do the following:

- For CDP Private Cloud:
  - Verify that Telemetry Publisher is enabled for Cloudera Observability on your Workload clusters and that they are associated with Telemetry Publisher.
  - If applicable, verify that your environment's data services are using Cloudera Data Engineering (CDE) version 1.19 or above and/or Cloudera Data Warehouse (CDW) version 1.6.3 or above. Starting with these versions the collected diagnostic data is categorized and displayed within their Data Service category in the Cloudera Observability web UI.
- For CDP Public Cloud:

  Verify that Enable Workload Analytics was turned on when you registered your Data Hub environment and when you installed your Data Engineering service.

Important: When you are not working in Cloudera Observability, Cloudera recommends that you explicitly log out by selecting your user name in the main navigation panel and clicking Log Out.

Procedure

1. In a supported web browser log in to the Cloudera Observability web UI by doing the following:

   a) In a supported browser, log into the Cloudera Data Platform.

      The CDP Public Cloud web interface landing page opens.

   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

   The Cloudera Observability web UI Landing page opens to the main navigation panel.

   Note: For a list of supported web browsers, click the Related Information link below.

2. From the main navigation panel, select **Financial Governance**, which opens the **Chargeback** page. When configured, by you, this page 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. For more information about the Financial Governance feature and how to configure your cost centers and assign them to your resources, click the Related Information link below.

3. From the main navigation panel, select Analytics.

   The Cloudera Observability **Environments** page opens.
4. Select an environment required for analysis.

   Tip: To filter and display only those environment platforms or services of interest, from the Environments list, select the environment's Type.

   The **Environment** navigation panel opens, which hierarchically lists the environment and its services hosted on the selected environment.
5. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required for analysis.

   To display the **Cluster Summary** page for a Data Lake, Database Catalog, Data Engineering, and Data Hub environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, the Data Engineering's Virtual Cluster, or the Data Warehouse's Virtual Warehouse that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, the Data Engineering's Virtual Cluster, or the Data Warehouse's Virtual Warehouse that is required for analysis.

   The **Cluster Summary** page, which is displayed as the title in your browser tab, displays performance trends and 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.
6. From the cluster's ENGINES, the Data Engineering's Virtual Cluster, or the Data Warehouse's Virtual Warehouse, select a workload engine of interest.

   When an engine is selected, the name of the engine is displayed in the browser tab and the page's chart widgets display information about the workload jobs 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.
7. In the workload engine's page, review its chart widgets and then select a chart widget, such as Suboptimal. Select a link or bar and drill down further to view more information, such as health checks, execution details, baselines, and trends.

   Tip: Breadcrumbs are displayed at the top of each page, which displays the name of your current location and its preceding page levels. You can move between these levels by clicking on a breadcrumb location.

Related Information

Supported browsers

Analyzing your environment costs with Cloudera Observability

About the Cloudera Observability user interface hierarchy

Supported browsers

Cloudera validates and tests against the latest version and supports recent versions of the following browsers:

- Google Chrome
- Mozilla Firefox

Note:

- Mozilla Firefox is not supported by Data Engineering.
- Certain accessibility features in DataFlow do not work in Mozilla Firefox.

- Safari
- Microsoft Edge

Managing your workloads and users

Learn the Cloudera Observability administration features that enable you to define workload views for analyzing specific items of interest and assign resource access roles for managing and restricting user access.

Classifying workloads for analysis with Workload Views

The Workload View feature enables you to analyze workloads with much finer granularity. 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 that are sent by a specific user are performing on your cluster.

Note: Workload Views are available for Classic Cluster, Private Cloud Base, Data Hub, Virtual Cluster, and Virtual Warehouse Cloudera Observability environments only.

Automatically generate workload views

If you have not defined workload views you have an option to generate default views by selecting a set of criteria.

About this task

Describes how to generate the Cloudera Observability default views.

Note: Workload Views are available for Classic Cluster, Private Cloud Base, Data Hub, Virtual Cluster, and Virtual Warehouse Cloudera Observability environments only.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.
   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.
   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.
   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

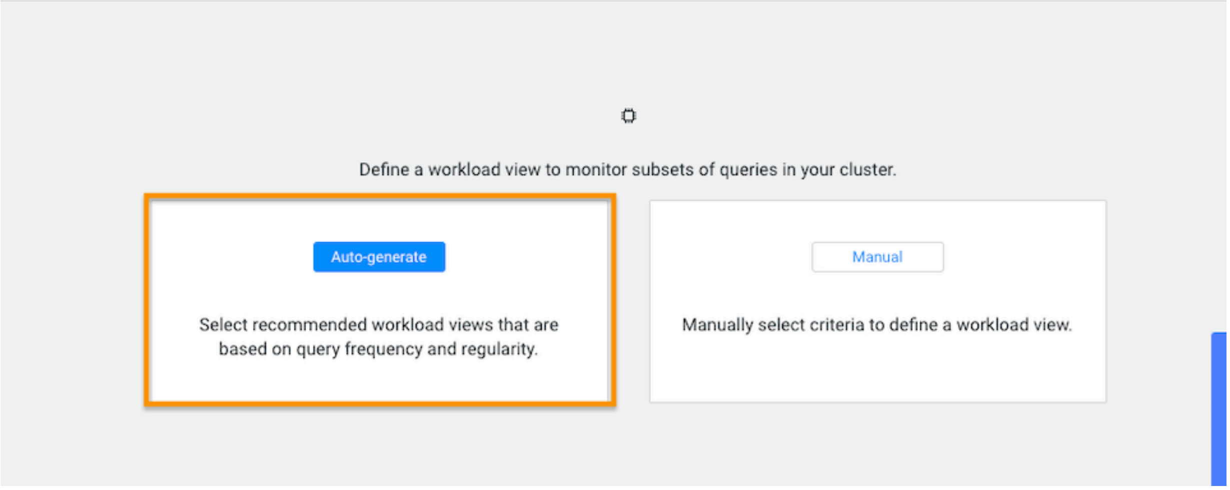
      The Environment navigation panel opens.
2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.

   To display the **Cluster Summary** page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.

   Tip: The page's title is displayed in the browser tab.

3. Select the Workloads tab.
4. In the Workloads page, click Auto-generate:

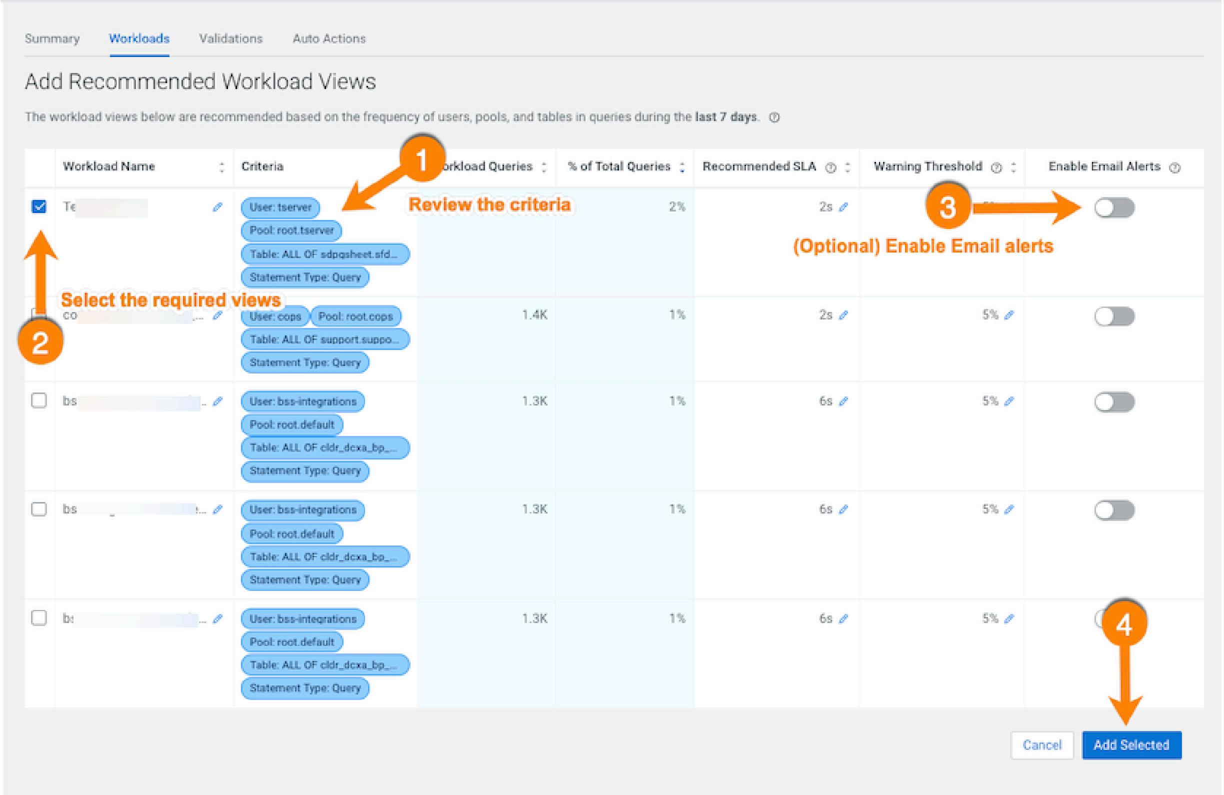


Define a workload view to monitor subsets of queries in your cluster.

Select recommended workload views that are based on query frequency and regularity.

Manually select criteria to define a workload view.

5. From the Criteria column, examine the criteria that is used for each workload view, select the required workload view or views, and then click Add Selected:



Add Recommended Workload Views

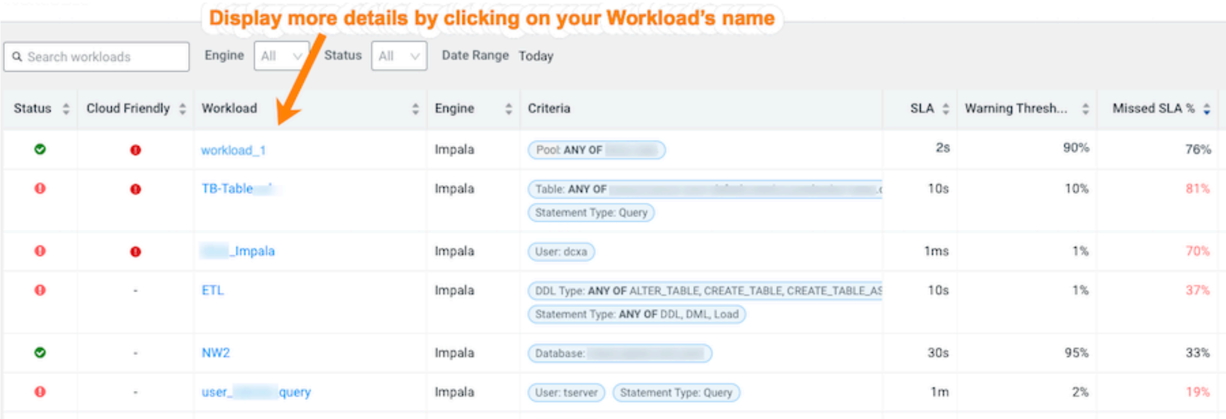
The workload views below are recommended based on the frequency of users, pools, and tables in queries during the **last 7 days**.

	Workload Name	Criteria	orkload Queries	% of Total Queries	Recommended SLA	Warning Threshold	Enable Email Alerts
☑	Te[redacted]	User: tserver Pool: root.tserver Table: ALL OF sdpqsheet.sfd... Statement Type: Query		2%	2s		○
□	co[redacted]...	User: cops Pool: root.cops Table: ALL OF support.suppo... Statement Type: Query	1.4K	1%	2s	5%	○
□	bs[redacted]...	User: bss-integrations Pool: root.default Table: ALL OF cldr_dcxa_bp_... Statement Type: Query	1.3K	1%	6s	5%	○
□	bs[redacted]...	User: bss-integrations Pool: root.default Table: ALL OF cldr_dcxa_bp_... Statement Type: Query	1.3K	1%	6s	5%	○
□	b[redacted]...	User: bss-integrations Pool: root.default Table: ALL OF cldr_dcxa_bp_... Statement Type: Query	1.3K	1%	6s	5%	○

The workload views you selected are saved and displayed on the Workloads page.

Note: For users with the Cloudera Observability Premium license tier, you can enable an email alert notification when a threshold or the number of failed jobs or queries is exceeded, by toggling the Enable Email Alerts switch to ON.

6. To verify your workload views, on the Workloads page, locate the workload view you added. When verified, click the workload to view its details:



Status	Cloud Friendly	Workload	Engine	Criteria	SLA	Warning Thresh...	Missed SLA %
✓	!	workload_1	Impala	Pool: ANY OF [redacted]	2s	90%	76%
!	!	TB-Table[redacted]	Impala	Table: ANY OF [redacted] Statement Type: Query	10s	10%	81%
!	!	[redacted]_Impala	Impala	User: dcxa	1ms	1%	70%
!	-	ETL	Impala	DDL Type: ANY OF ALTER_TABLE, CREATE_TABLE, CREATE_TABLE_AS Statement Type: ANY OF DDL, DML, Load	10s	1%	37%
✓	-	NW2	Impala	Database: [redacted]	30s	95%	33%
!	-	user_[redacted]query	Impala	User: tserver Statement Type: Query	1m	2%	19%

7. To view more information about the workload, open its Summary page by clicking the name of the workload view in the Workload column, which displays the view's details as chart widgets that you can use to further analyze the results.

8. To create a new view do the following:

   a) Verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.

      To display the **Cluster Summary** page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

      - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.
      - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.

      Tip: The page's title is displayed in the browser tab.

   b) Select the Workloads tab.

   c) From the Define New menu in the Workloads page, select one of the following:

      - To create a new manual view, select Manual Definition, in the Criteria Definition widget define a set of criteria for the view, and then click Save.
      - To automatically generate a new view, select Auto-generate Definition.

      The Workloads page reopens and your workload view appears in the Workload column.

9. Workload Views cannot be edited directly. If you require changes to an existing Workload View do the following:

   a) In the Workloads page, locate the Workload View that requires changes.

   b) From its Action list, select Clone.

   c) In the Criteria Definition widget make the changes you require, and then click Save.

      The Workloads page reopens and your workload view appears in the Workload column.

   d) Locate the Workload View that required changes and from its Action list, select Delete and then in the Confirm message, confirm its deletion by clicking OK.

Defining workload views manually

Steps for manually defining your workload views.

About this task

This task describes how to manually define your Workload Views.

Note: Workload Views are available for Classic Cluster, Private Cloud Base, Data Hub, Virtual Cluster, and Virtual Warehouse Cloudera Observability environments only.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.
   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.

   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.

   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

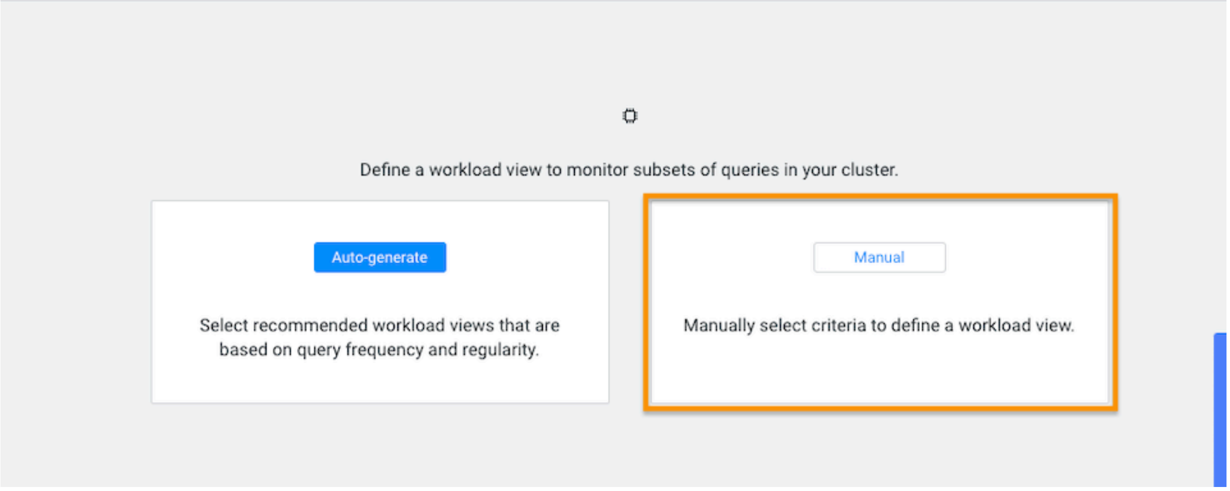
      The Environment navigation panel opens.
2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.

   To display the Cluster Summary page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.

   Tip: The page's title is displayed in the browser tab.
3. Select the Workloads tab.

4. In the Workloads page, click Manual:



Define a workload view to monitor subsets of queries in your cluster.

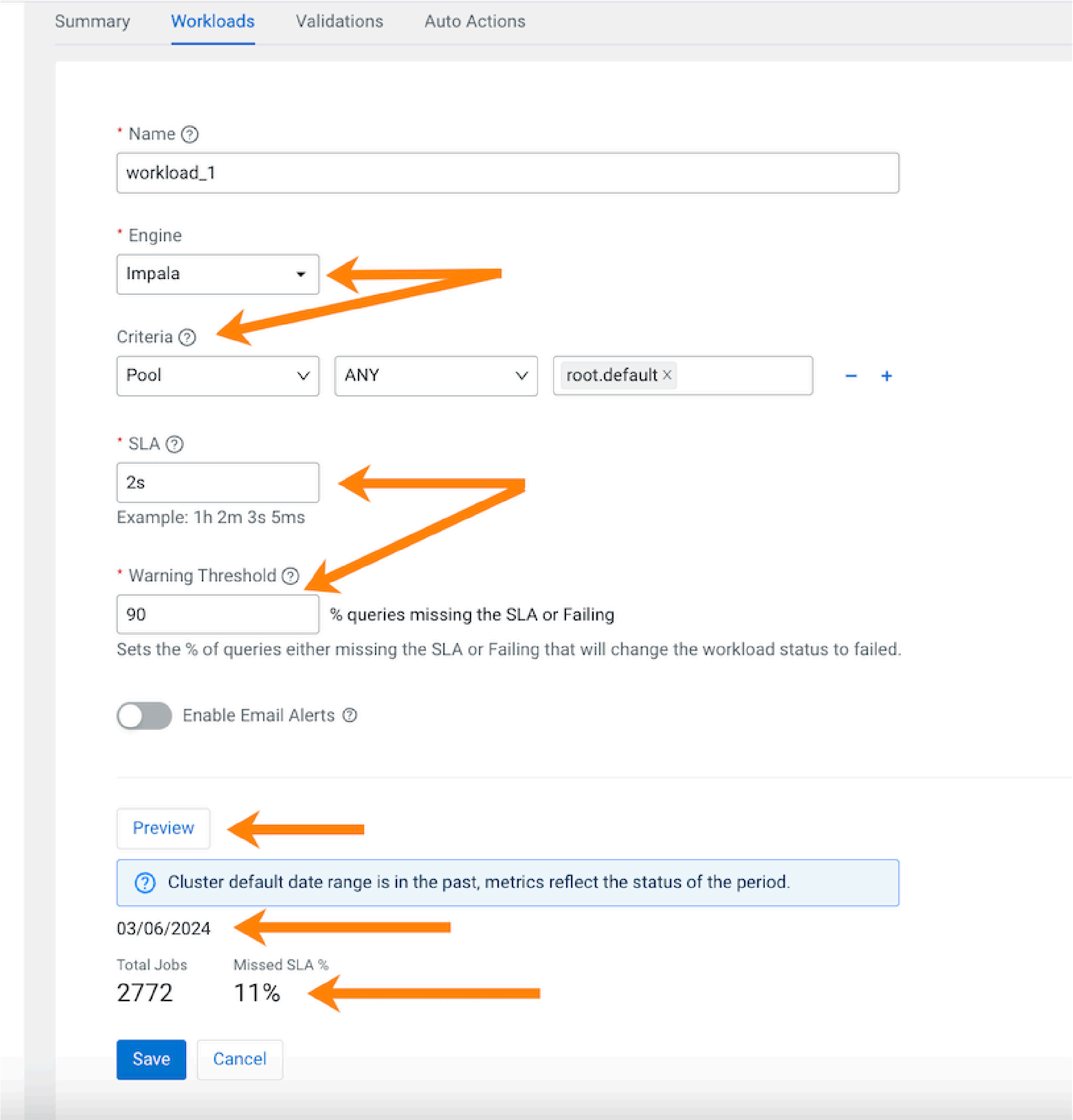
Select recommended workload views that are based on query frequency and regularity.

Manually select criteria to define a workload view.

The Criteria Definition widget opens, where you define a set of criteria that enables you to analyze a specific set of queries.

For example, as shown in the image below, you can list the total amount of failed queries, as a percentage, from a specific engine that are subject to a two second SLA.

Where, as defined by the criteria condition, Cloudera Observability will monitor all query jobs from the Impala engine. When the total query execution time exceeds 2 seconds, as defined by the SLA condition, for 90 percent of these queries, as defined by the Warning Threshold, the workload is flagged with a failed state:



* Name ⓘ

workload_1

* Engine

Impala

Criteria ⓘ

Pool | ANY | root.default ×

* SLA ⓘ

2s

Example: 1h 2m 3s 5ms

* Warning Threshold ⓘ

90 % queries missing the SLA or Failing

Sets the % of queries either missing the SLA or Failing that will change the workload status to failed.

Enable Email Alerts ⓘ

ⓘ Cluster default date range is in the past, metrics reflect the status of the period.

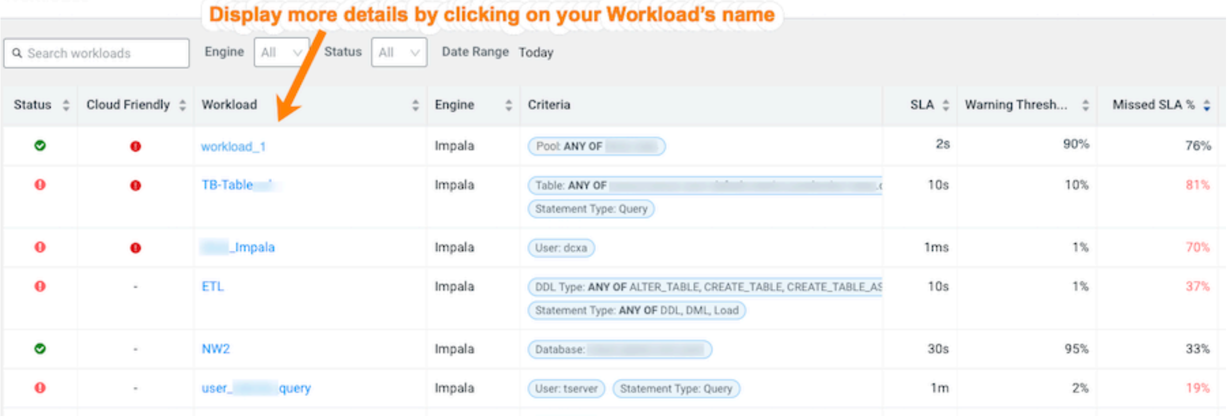
03/06/2024

Total Jobs	Missed SLA %
2772	11%

5. To display a summary of the queries matching your criteria, click Preview. The date range, the number of queries that match the criteria, and the number of queries that missed the SLA condition are displayed.

6. **6.** Click Save.

   The Workloads page opens and your workload view appears in the Workload column.



Display more details by clicking on your Workload's name

Status	Cloud Friendly	Workload	Engine	Criteria	SLA	Warning Thresh...	Missed SLA %
✓	!	workload_1	Impala	Pool: ANY OF [redacted]	2s	90%	76%
!	!	TB-Table[redacted]	Impala	Table: ANY OF [redacted] Statement Type: Query	10s	10%	81%
!	!	[redacted]_Impala	Impala	User: dcxa	1ms	1%	70%
!	-	ETL	Impala	DDL Type: ANY OF ALTER_TABLE, CREATE_TABLE, CREATE_TABLE_AS Statement Type: ANY OF DDL, DML, Load	10s	1%	37%
✓	-	NW2	Impala	Database: [redacted]	30s	95%	33%
!	-	user_[redacted]query	Impala	User: tserver Statement Type: Query	1m	2%	19%

Tip: To locate your new workload view from a long list, sort the Workload column alphabetically in either the ascending or descending order by clicking the Workload column's up and down arrows.

7. **7.** To view more information about the workloads using the view's formula, open its Summary page by clicking the name of the workload view in the Workload column, which displays the view's details as chart widgets that you can use to further analyze the results.
8. **8.** To create a new view do the following:

   a) Verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.

   To display the **Cluster Summary** page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.

   Tip: The page's title is displayed in the browser tab.

   b) Select the Workloads tab.

   c) From the Define New menu in the Workloads page, select one of the following:

   - To create a new manual view, select Manual Definition, in the Criteria Definition widget define a set of criteria for the view, and then click Save.
   - To automatically generate a new view, select Auto-generate Definition.

   The Workloads page reopens and your workload view appears in the Workload column.
9. **9.** Workload Views cannot be edited directly. If you require changes to an existing Workload View do the following:

   a) In the Workloads page, Locate the Workload View that requires changes.

   b) From its Action list, select Clone.

   c) In the Criteria Definition widget make the changes you require, and then click Save.

   The Workloads page reopens and your workload view appears in the Workload column.

   d) Locate the Workload View that required changes and from its Action list, select Delete and then in the Confirm message, confirm its deletion by clicking OK.

Triggering email alerts for your workload views

You can trigger daily email alerts for your Workload Views, based on your defined service-level agreement (SLA) performance threshold and/or your workload job or query failures. When a Workload View's SLA reaches or exceeds the defined threshold, or the workload jobs or queries reach or exceed the failure percentage, an email alert is triggered for you to take action upon its receipt.

About this task

Steps on how to enable email alerts for your Workload Views.

Note: The Cloudera Observability Premium license tier is required for the Workload View Alert feature. If you do not have Cloudera Observability Premium the Workload View's alert enablement and menu items are hidden. For more information about the Cloudera Observability Premium license tier and to request a demo, click the Related Information link below.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.

   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.

   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.

   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

      The Environment navigation panel opens.

2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.

   To display the Cluster Summary page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.

   Tip: The page's title is displayed in the browser tab.

3. Select the Workloads tab.

4. Do one of the following:

   a. If no other Workload Views exist, in the Workloads page, click Manual.

   b. If other Workload Views exist, from the Define New list, select Manual Definition.

   The Criteria Definition widget opens, where you define the criteria for the Workload View that will alert you when the SLA or specific workload jobs or queries reach or exceed the defined threshold or failure percentage.

5. In the Name field, enter a unique name that is easily identifiable.

   Note: The name must be alphanumeric, must start with an alphabetical character, and must not contain spaces. Underscores and minus characters are accepted.

6. From the Engine list, select the engine in which the job or query is run. For example, Impala.

   Note: All jobs or queries that are run on the selected engine will be monitored by Cloudera Observability.

7. Specify the Criteria condition by doing the following:

   a. From the Criteria list, select a criteria filter item from the available options.

      You can set multiple conditions for the selected filter item. For example,

      - If you selected User, you can include ANY or NONE of the available users in the Select user list.
      - If you selected Pool, you can include ANY or NONE of the available pools in the Select pool list.
      - If you selected Query Start Time, you can include IN RANGE or NOT IN RANGE of your selected time period.

      Tip: You can define multiple Criteria filters by clicking the plus sign.

8. In the SLA field, enter the threshold unit for the completion of a job or query, using the following abbreviations as the time units:

   - h = hours
   - m = minutes
   - s = seconds
   - ms = milliseconds

   The time units must be in chronological order, where hours come before minutes, minutes come before seconds, and seconds come before milliseconds, and cannot have a space between the threshold number value and the time unit. For example, 2h 20m 3s. The threshold time value can also be entered as a whole time unit, where instead of entering 1h 10m you can enter 70m.

9. In the Warning Threshold field, enter a percentage value that when exceeded by either the number of jobs or queries failing the SLA value or failing execution completion, triggers a Warning status and if applicable triggers an email alert notification.

   Note: Rounding rules are applied to the Warning Threshold value.

10. For users with the Cloudera Observability Premium license tier, you can enable the email alert notification when the threshold or the number of failed jobs or queries is exceeded, by toggling the Enable Email Alerts switch to ON. When enabled, a maximum of one email for each calendar day is sent to the current user's email address notifying them of the exceeded threshold.

    Note: To disable email alert notifications, in the Workloads page, select the Workload View and from its Actions list, select Disable Email Alerts.

11. To display a summary of the jobs or queries matching your criteria, click Preview. The date range, the number of jobs or queries that match the criteria, and the number of jobs or queries that missed the SLA condition or failed completion are displayed.

12. Click Save.

    The Workloads page opens and your workload view appears in the Workload column.

    Tip: To locate your new workload view from a long list, sort the Workload column alphabetically in either the ascending or descending order by clicking the Workload column's up and down arrows.

13. To view more information about the workloads using the view's formula, open its Summary page by clicking the name of the Workload View in the Workload column, which displays the view's details as chart widgets that you can use to further analyze the results.

14. To delete an existing Workload View, do the following:

    a) In the Workloads page, locate and select the Workload View that requires deletion.
    b) From the Actions list, select Delete.
    c) In the confirmation message, click OK to confirm. The view is permanently removed.

15. To create a new view do the following:

a) Verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.

To display the **Cluster Summary** page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

- From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.
- In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse that is required for analysis.

Tip: The page's title is displayed in the browser tab.

b) Select the Workloads tab.

c) From the Define New menu in the Workloads page, select one of the following:

- To create a new manual view, select Manual Definition, in the Criteria Definition widget define a set of criteria for the view, and then click Save.
- To automatically generate a new view, select Auto-generate Definition.

The Workloads page reopens and your workload view appears in the Workload column.

16. Workload Views cannot be edited directly. If you require changes to an existing Workload View do the following:

a) In the Workloads page, Locate the Workload View that requires changes.

b) From its Action list, select Clone.

c) In the Criteria Definition widget make the changes you require, and then click Save.

The Workloads page reopens and your workload view appears in the Workload column.

d) Locate the Workload View that required changes and from its Action list, select Delete and then in the Confirm message, confirm its deletion by clicking OK.

Related Information

Cloudera Observability product website

About the Cloudera Observability Workloads page

Describes the fields in the Cloudera Observability Workloads page.

The Workloads page displays the defined settings and state of your workload views.

It contains the following entry fields:

- Status, which displays the current state of the action, as follows:
  - Green, denotes that all the jobs/queries in a Workload View are UNDER the specified threshold for both Missed SLA and Failure Rate.
  - Red, denotes that the percentage of jobs/queries in a Workload View have met or exceeded the specified threshold for EITHER the Missed SLA or the Failure rate.
- Workload, which displays the name of the Workload View. When clicked the Workload View’s Summary page opens.
- Engine, which displays, from the Workload View’s definition settings, the selected engine in which the jobs or queries are run.
- Criteria, which displays the alert's Criteria filters. These are attributes with static values that remain the same during the execution of a job or query.

- SLA, which displays the service level agreement performance measurement, set as the completion duration threshold of a job or query, using the following abbreviations as the time units:
  - h = hours
  - m = minutes
  - s = seconds
  - ms = milliseconds

  Note: Depending on what was entered for the SLA duration threshold, the time unit value may be displayed as a whole time unit, where 70m is displayed for 1h 10m.
- Warning Threshold, which displays the warning threshold value, which is set as a percentage of jobs or queries that either miss the SLA condition or fail completion.
- Missed SLA %, which displays the percentage of jobs or queries that missed the SLA threshold.
- Failure %, which displays the percentage of jobs or queries that failed completion.
- Total Jobs/Queries, which displays the total number of jobs or queries executed, regardless of completion (including those not in a terminal state), during the selected time period that is displayed in the Date Range field in the filter row.
- Action, which when selected lists the Workload View's available actions:
  - Rename
  - Clone
  - Delete
  - Manage Access
  - Enable Email Alerts or Disable Email Alerts

    The action displayed is dependent on whether Email Alerts have been enabled or disabled.

Managing user access to workloads

Describes how to create and manage Cloudera Observability resource access roles.

You can assign user roles in Cloudera Observability that restrict access to your workload clusters, jobs, and queries. These roles provide varying levels of access that prevent users from accessing information that they do not explicitly require.

When a user is assigned to a cluster with one of the Cloudera Observability access roles, the cluster will appear on the Environment navigation page. All other workload clusters are not visible to the user.

Note: Only Account administrator users can assign Cloudera Observability resource access roles.

Cloudera Observability access roles

Describes the Cloudera Observability access roles and the actions that a user can perform in the Cloudera Observability web UI for each access role type.

The following tables list the Cloudera Observability access roles:

Account admin access role type

A user assigned the ObservabilityAccountAdmin access role type has complete access to the Cloudera Observability workload clusters, where they can view, edit, and create cost centers, view, edit, and create auto actions, and view, edit, and create workloads in all the Workload clusters. These users have the least restrictive access permissions.

Note: The Cloudera Observability Account Admin user cannot grant or revoke access for other users.

Table 1: Actions that can be performed by the Account Admin

Resource	Actions
Cluster	- View all workload clusters on the Clusters page - Rename a workload cluster - Delete a workload cluster
Workloads	- Create workloads - View all workloads in a cluster - Update all workloads in a cluster - Delete all workloads in a cluster
Queries	View all queries in a cluster
Jobs	View all jobs in a cluster
Chargeback	- Create cost centers - Update cost centers - List cost centers - Delete cost centers - View all Chargeback related dashboards
Auto Actions	- Create auto actions - View auto actions - Update auto actions - Disable auto actions - Delete auto actions - Enable an auto action email
Cluster Report Emails	Enable cluster report emails
Cloudera Support Access	N/A

Cluster Admin access role type

The ObservabilityClusterAdmin access role type has full access to Cloudera Observability and can view, edit, and create workloads in the Workload cluster.

Note: The Cloudera Observability Cluster Admin user cannot grant or revoke access for other users.

Table 2: Actions that can be performed by the Cluster Admin

Resource	Actions
Cluster	- View the workload cluster on the Clusters page - Rename the workload cluster - Delete the workload cluster
Workloads	- Create workloads - View all workloads in the cluster - Update all workloads in the cluster - Delete all workloads in the cluster
Queries	View all queries in the cluster
Jobs	View all jobs in the cluster
Chargeback	N/A
Auto Actions	N/A
Cluster Report Emails	Enable cluster report emails

Resource	Actions
Cloudera Support Access	Enable and Disable Cloudera Support access to the cluster's diagnostic data

Cluster access role type

A user assigned the ObservabilityClusterUser access role type can view items within Cloudera Observability, including all workloads, but cannot edit workloads.

Table 3: Actions that can be performed by the Cluster User

Resource	Actions
Cluster	View the cluster on the Clusters page
Workloads	View all workloads in the cluster
Queries	View all queries in the cluster
Jobs	View all jobs in the cluster
Chargeback	N/A
Auto Actions	N/A
Cluster Report Emails	Enable cluster report emails
Cloudera Support Access	N/A

Workload access role type

A user assigned the ObservabilityWorkloadUser access role type can only view their assigned workloads in the cluster and the jobs and queries within that workload. These users have the most restrictive access permissions.

Table 4: Actions that can be performed by the Workload user

Resource	Actions
Cluster	Disabled For a Workload user to be able to view the cluster that contains their assigned workloads they must also be assigned the Limited Environment access type. Cloudera recommends assigning the Limited Environment access type to Workload Users because this enables the user to access and view the environment that contains their workloads.
Workloads	View their assigned workloads on the Workloads page
Queries	View all queries within an assigned workload
Jobs	View their jobs within the assigned workload
Chargeback	N/A
Auto Actions	N/A
Cluster Report Emails	Disabled
Cloudera Support Access	N/A

Limited environment access role type

A user assigned the ObservabilityLimitedClusterUser access role type can only view the cluster that contains their workloads. Without this access privilege, Workload users are unable to access or view the environment that contains their workloads.

Note: Cloudera recommends assigning the Limited Environment access type to Workload users.

Table 5: Actions that can be performed by the Limited Environment user

Resource	Action
Cluster	View the cluster on the Clusters page
Workloads	N/A
Queries	N/A
Jobs	N/A
Chargeback	N/A
Auto Actions	N/A
Cluster Report Emails	N/A
Cloudera Support Access	N/A

Assigning access roles in Cloudera Observability

Steps for assigning resource access roles in Cloudera Observability that restrict access to your workload clusters, jobs, and queries.

About this task

Describes how to assign resource access roles to a Cloudera Observability user. The Cloudera Observability Manage Access feature enables you to assign a user to a Cloudera Observability access role that is associated with one or multiple workload clusters, jobs, and queries.

Note: Only Account administrator users can assign Cloudera Observability resource access roles.

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).

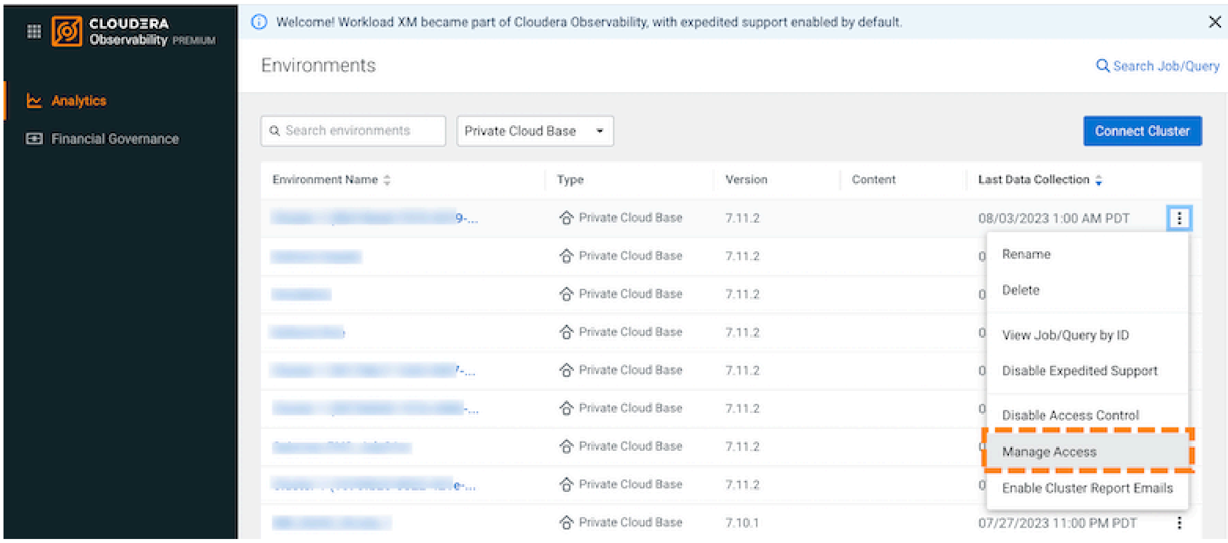
       The CDP Cloud web interface landing page opens.
    b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

       The Cloudera Observability landing page opens.
2. From the Cloudera Observability **Environments** page, locate the environment that contains the workload to which you will assign a Cloudera Observability user resource access role.

   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, select Manage Access.

   The **Manage Access** page opens.


CLOUDERA Observability PREMIUM
Analytics
Financial Governance
Welcome! Workload XM became part of Cloudera Observability, with expedited support enabled by default.
Environments
Search Job/Query
Search environments
Private Cloud Base
Connect Cluster
Rename
Delete
View Job/Query by ID
Disable Expedited Support
Disable Access Control
Manage Access
Enable Cluster Report Emails

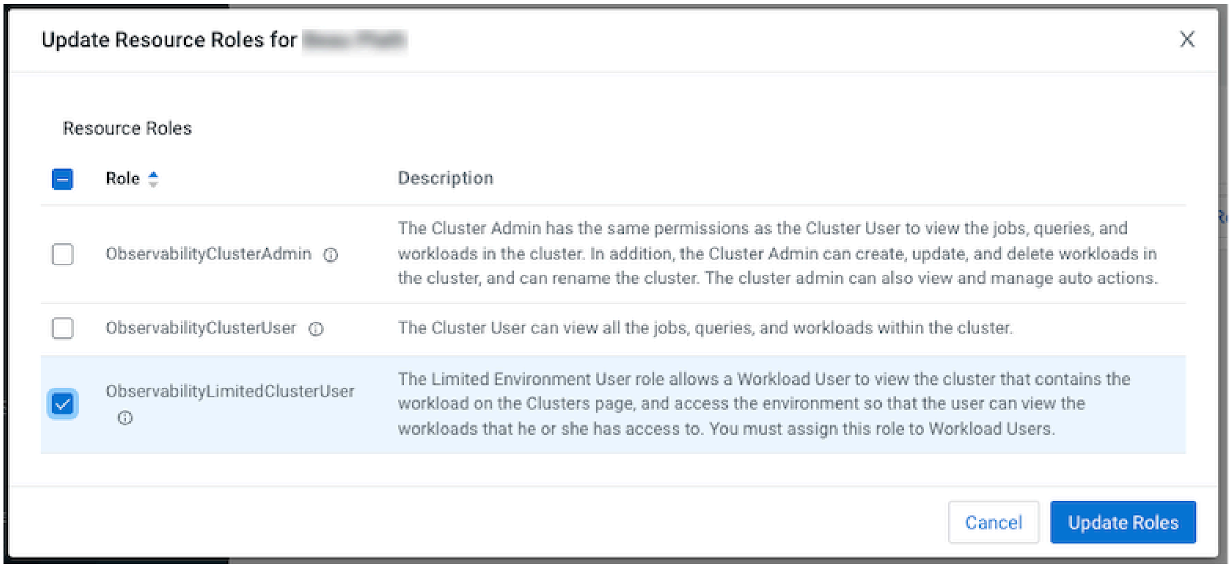
Environment Name	Type	Version	Content	Last Data Collection
[redacted]	Private Cloud Base	7.11.2		08/03/2023 1:00 AM PDT
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.11.2		0
[redacted]	Private Cloud Base	7.10.1		07/27/2023 11:00 PM PDT

4. In the search field, enter and then select the name of the user to which you will assign a Cloudera Observability user resource access role.

   The Update Resource Roles for *nameofuser* dialog box opens, which lists the user resource access role options that you can assign to the user for Cloudera Observability.

5. Select the check box next to the resource role you require for the user.

   In this example, the ObservabilityLimitedClusterUser role check box is selected, which gives the user limited access to the environment, but provides access and visibility to their workloads.



Update Resource Roles for [redacted]

Resource Roles

	Role	Description
□	ObservabilityClusterAdmin	The Cluster Admin has the same permissions as the Cluster User to view the jobs, queries, and workloads in the cluster. In addition, the Cluster Admin can create, update, and delete workloads in the cluster, and can rename the cluster. The cluster admin can also view and manage auto actions.
□	ObservabilityClusterUser	The Cluster User can view all the jobs, queries, and workloads within the cluster.
☑	ObservabilityLimitedClusterUser	The Limited Environment User role allows a Workload User to view the cluster that contains the workload on the Clusters page, and access the environment so that the user can view the workloads that he or she has access to. You must assign this role to Workload Users.

6. Click Update Roles.

   A Success message appears confirming that the resource roles for the user are updated and the name of the user is populated in the Name column of the Manage Access table.

7. In the breadcrumb row, click the name of the environment.

   The Environment Summary page opens.

   Tip: You can navigate between pages in the Cloudera Observability web UI using the breadcrumb row.

8. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster or Virtual Warehouse.

   To display the Cluster Summary page for a Data Hub, Virtual Cluster, and Virtual Warehouse environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, Virtual Cluster, or Virtual Warehouse.

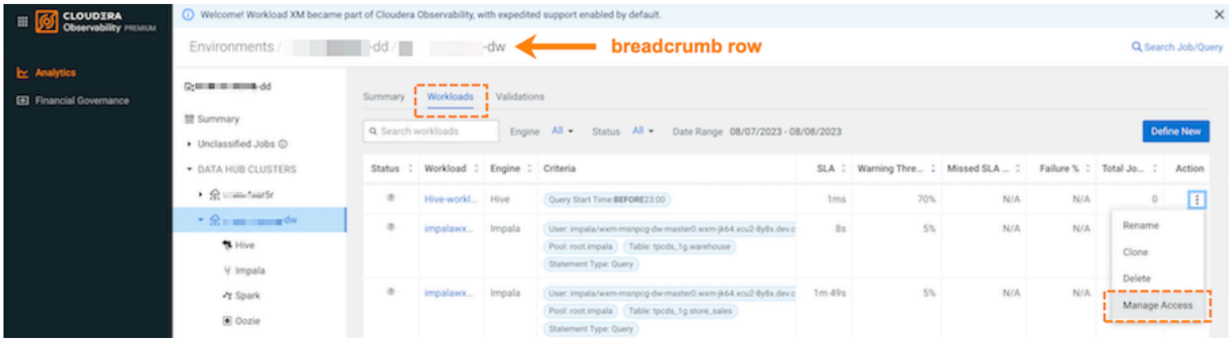
   Tip: The page's title is displayed in the browser tab.

9. In the Cluster Summary page, select the Workloads tab.

   The Workloads page opens.

10. In the **Workloads** page, locate the workload that is to be assigned to the user of the user resource access role, in this case the ObservabilityLimitedClusterUser, and then from its Actions list, select Manage Access.

    The Manager Access page opens.



Welcome! Workload XM became part of Cloudera Observability, with expedited support enabled by default.

Environments / [redacted]-dd / [redacted]-dw

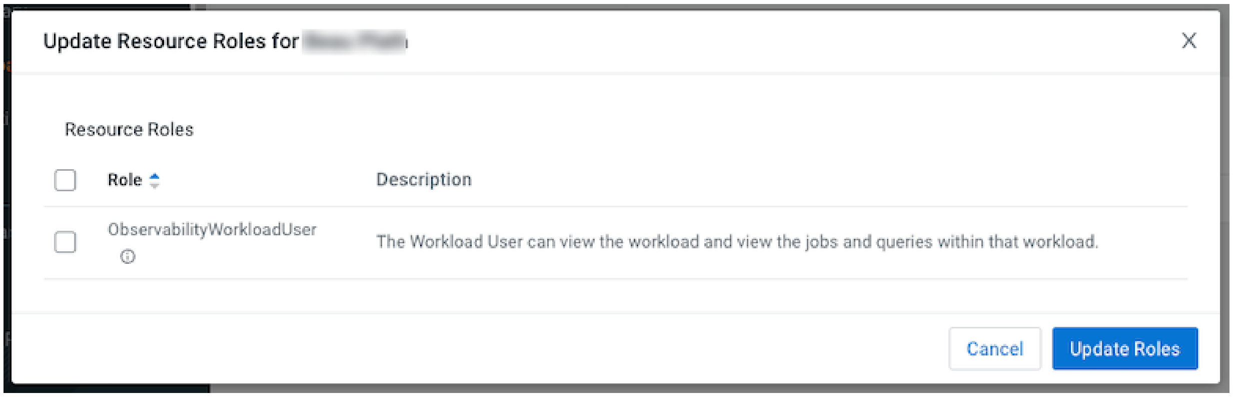
Date Range 08/07/2023 - 08/08/2023

Status	Workload	Engine	Criteria	SLA	Warning Thre...	Missed SLA ...	Failure %	Total Jo...	Action
	Hive-workl...	Hive	Query Start Time BEFORE 23:00	1ms	70%	N/A	N/A	0	⋮
	impalawx...	Impala	User: impala/wxm-msnpcg-dw-master0.wxm-jk64.xcu2-8y8x.dev.c Pool: root.impala Table: tpcds_1g.warehouse Statement Type: Query	8s	5%	N/A	N/A		
	impalawx...	Impala	User: impala/wxm-msnpcg-dw-master0.wxm-jk64.xcu2-8y8x.dev.c Pool: root.impala Table: tpcds_1g.store_sales Statement Type: Query	1m 49s	5%	N/A	N/A		

11. In the search field, enter and then select the name of the user with the assigned user resource access role.

    The Update Resource Roles for *nameofuser* dialog box opens, which displays the workload role option that is associated with the user resource access role.

12. Select the check box next to the resource role, in this case the ObservabilityWorkloadUser role, which gives the user limited access to the workload, but provides access and visibility to their workloads.



Update Resource Roles for [redacted]

Resource Roles

□	Role	Description
□	ObservabilityWorkloadUser ⓘ	The Workload User can view the workload and view the jobs and queries within that workload.

13. Click Update Roles.

    A Success message appears confirming that the resource roles for the user are updated.

    The user is now limited to viewing only those workload jobs and queries associated with the workload cluster that they were assigned.

14. To verify which Cloudera Observability user resource and workload roles are assigned to a user, do the following:

    a) In the Manage Access page, locate and click the name of the user whose roles you require for verification.

       To access the Manage Access page, do the following:

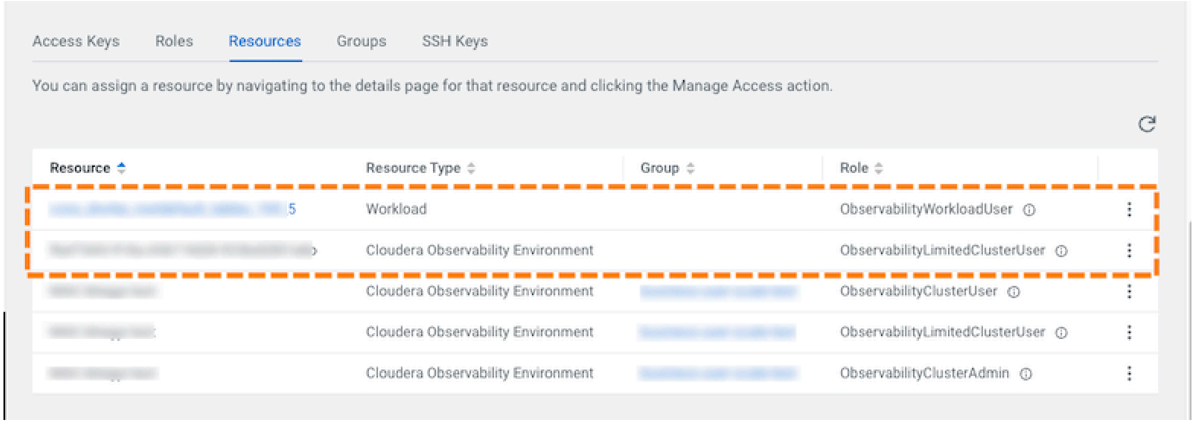
       1. From the cluster's or Virtual Warehouse's Workloads page, locate the workload that is assigned to the user whose user resource and workload roles require verification.
       2. From its Actions list, select Manage Access.

          The Manage Access page opens.

       The user's profile page opens.

    b) Scroll down and select the Resources tab.

       The resources, the resource type, and the roles that are assigned to the user are displayed.



You can assign a resource by navigating to the details page for that resource and clicking the Manage Access action.

Resource	Resource Type	Group	Role
[redacted]	Workload		ObservabilityWorkloadUser
[redacted]	Cloudera Observability Environment		ObservabilityLimitedClusterUser
[redacted]	Cloudera Observability Environment	[redacted]	ObservabilityClusterUser
[redacted]	Cloudera Observability Environment	[redacted]	ObservabilityLimitedClusterUser
[redacted]	Cloudera Observability Environment	[redacted]	ObservabilityClusterAdmin

Working with alerts, costs, and reports

Certain Cloudera Observability features enable you to define cost centers for planning, budgeting, and justifying resources, create alert actions that monitor your workloads and trigger a corrective action when applicable, and enable reports that send you daily statistics about your environment.

Analyzing your environment costs with Cloudera Observability

The Cloudera Observability Financial Governance Chargeback feature collects CPU, memory, and resource usage data from your environment, allocates those charges to your custom cost centers, and visually displays the results. It provides an in-depth visibility into the workload resource costs of your environment's infrastructure that can be used for planning, budgeting, and forecasting.

Note: Only users with the ObservabilityAccountAdmin access role type can define the Chargeback settings, list, create, update, or delete cost centers, and view all the Chargeback related dashboards. For more information about the Cloudera Observability access roles, click the Related Information link below.

About the Cloudera Observability Chargeback feature

The Cloudera Observability Financial Governance Chargeback feature measures and records the costs of your workload resources and allocates them to the users who consume them. For resources that are shared, such as multi-tenant clusters that are shared between different organizations and departments, it also enables you to measure and record those shared costs and charge those users based on their actual consumption. This feature helps you plan and forecast budgets, it helps you ensure that costs are in line with business requirements and the Chargeback cost center reports can be used to raise cost awareness and set limits to control your overall costs.

About cost centers and their criteria

The Cloudera Observability Financial Governance Chargeback feature calculates cost based on the following criteria:

- User or Pool usage, which enables you to separate user and resource pool costs.
- CPU and Memory hourly unit costs, which are based on actual CPU and memory usage using your internal pricing or cost model.

Using the Chargeback criteria that you have set, charges for CPU and memory consumption are calculated and assigned by Cloudera Observability to a cost center that is created by you. Cost centers separate costs across users or pools and track their workload resource consumption costs. They can be divided and/or grouped into members associated with an organization or group for helping you assign the charges to a user’s department.

When you create a Cloudera Observability cost center, detailed summary reports of the costs and resource usage for the environment are generated. After a job has run, the tracked resource costs that is associated with the cost center’s environment, service, or cluster are visually displayed. You can drill down for more detailed reports, such as viewing the costs incurred by a specific user or pool or viewing the top 5 users or pools whose jobs created the highest costs or the top 25 jobs or queries that created the highest costs.

Overtime, as more jobs and queries are run you can view and compare historical trends by selecting specific time periods from the time-range list. By default, data is retained for 6 months.

The Cloudera Observability Chargeback feature uses usage-based metrics for CPU utilization and memory consumption that have an hourly aggregation.

The Chargeback costs are calculated based on an hourly cost per resource unit, where:

- The CPU costs are expressed as the amount of time a job process uses CPU within an hour.
- The Memory costs are expressed as an hourly allocation cost per gigabyte.

Considerations and limitations

The following describes consideration and limitations you must know when using the Cloudera Observability Chargeback feature:

- Cost centers aggregate the charges, where a cost center can be for one individual user, multiple users, or pools. To avoid cost duplication, users and pools must only be assigned one cost center.
- When viewing the Chargeback reports, the costs are adjusted to a user's local timezone. Therefore, total costs, such as daily charges, may differ across timezones.

  Note: The time-range list converts universal time to the user's local timezone.

Assumptions and prerequisites

The Cloudera Observability Chargeback feature assumes the following:

- Your organization has an internal pricing or cost model.
- You have created Cloudera Observability users or resource pools and assigned them to your workloads.

Related Information

Managing user access to workloads

Configuring the Cloudera Observability cost center criteria

The Cloudera Observability Financial Governance Chargeback feature defines cost centers based on certain criteria. Configure your Cloudera Observability Chargeback settings by designating your cost center resource usage across users and pools and defining the unit resource consumption costs.

About this task

The Cloudera Observability Chargeback calculates user and pool costs based on CPU and memory consumption. You decide the CPU and Memory unit rates using your internal pricing or cost model.

Note: Only users with the ObservabilityAccountAdmin access role type can define the Chargeback settings, list, create, update, or delete cost centers, and view all the Chargeback related dashboards.

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 Main navigation panel, select Financial Governance.
3. To configure your Chargeback criteria, do the following:
   a. From the Actions list, select Chargeback Settings.

      The Setup page opens displaying the Chargeback Criteria settings.
   b. From the Select your Chargeback criteria section, select the required user or pool usage criterion option for your cost centers.

      Where, the Users option defines your cost centers based on users, and the Pools option defines your cost centers based on your resource pools.
   c. From the Unit cost section, do the following:
      1. In the CPU ($/CPU core hours) field, enter the amount for each CPU core hour.
      2. In the Memory ($/GB hours) field, enter the amount for each gigabyte hour.

      Note: By default, the decimal currency symbol uses the $ dollar sign.
   d. Click Complete Setup.

   Now that you have configured your Chargeback criteria settings you can start creating cost centers.
4. To change your Chargeback criteria, do the following:

   Important: Cost centers are associated with a specific usage criterion (Users or Pools). Changing the Chargeback usage criteria setting that your cost centers are associated with, such as from Users to Pools, will hide your current cost centers that are associated with the previous usage criterion.
   a. From the Actions list on the Environments page, select Chargeback Settings.

      The Chargeback Criteria Setup page opens.
   b. If required, change the usage criterion option by selecting the option now required.

      A warning message appears explaining that all cost centers associated with your previous usage criterion will be hidden.
   c. If required, from the Unit cost section, make your changes to the CPU and Memory unit costs.
   d. Click Update.

   If you changed the usage criterion, for example from Users to Pools, your cost centers from the previous usage criterion (Users) are hidden. To display the unit consumption costs of your resources based on the new usage criterion value requires creating new cost centers.

   Note: The Cloudera Observability Chargeback feature enables you to have cost centers associated with each usage criterion for tracking the workload consumption costs for both Users and Pools.

Creating a Cloudera Observability cost center

This topic describes the steps for creating a Cloudera Observability cost center. Cost centers separate costs across user or pool usage and track their workload resource consumption costs. They can be divided and/or grouped into members associated with a specific organization or group for helping you assign actual consumption charges to a user's department.

About this task

Describes how to create a Cloudera Observability cost center.

Important: To avoid cost duplication, resources must only be assigned one cost center.

Note: Only users with the ObservabilityAccountAdmin access role type can define the Chargeback settings, list, create, update, or delete cost centers, and view all the Chargeback related dashboards.

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 Main navigation panel, select Financial Governance.

3. To create a new cost center, do the following:

   a. From the Actions list, select Create a Cost Center.

      The Create a Cost Center page opens displaying the Cost Center details settings.
   b. In the Name field, enter a unique name for your cost center.
   c. In the Description field, enter a meaningful description for the cost center.
   d. From the Environment Selection section, click inside its text field to display a hierarchical list of your environments and their clusters.
   e. From the hierarchical list, locate the cluster in which your jobs or queries run and select its check box.

      The cluster is and its parent service are also highlighted and the cluster's name is populated in the Environment Selection text field.

      Note: For multiple clusters within an environment or within an environment's service, such as a Data Warehouse, you must select the check box for each of the clusters you require.
   f. Continue to locate and add more clusters and their environments.
   g. Depending on the Chargeback usage criterion option you selected when you configured your Chargeback settings, do one of the following:

      - If you selected Pools, click inside the Add Pools field and then select either one or multiple resource pools, or select All (denoted as a star *), which highlights all the resource pools associated with the selected cluster.

        The Add Pools field is populated with the selected pools.
      - If you selected Users, click inside the Add Users field and then select either one or multiple users, or select All (denoted as a star *), which highlights all the users associated with the selected cluster.

        The Add Users field is populated with the selected users.

      Note: Only those pools or users that are associated with the environment's cluster that you previously selected in the Environment Selection's field are listed.
   h. Click Create.

      The CDP Chargeback page opens displaying a Success message, which denotes that the cost center was successfully created, and your new cost center is listed under the Cost Centers column.

      Note: Until data is available from a job run, within the clusters you selected, the costs and resource usage will not display. Zero cost and resource usage values for the cost center denote that no charges have been incurred. If this continues after a job has run, check that the correct time-period is displayed.

   Now that you have created a cost center you can now view the costs and resource usage associated with your cost center.
4. To edit an existing cost center, do the following:

   a. In the Chargeback page, locate and select the cost center that requires changes.
   b. From the Actions list, select Edit Cost Center.

      The Cost Center details page opens displaying the Cost Center details settings.
   c. Make your changes.
   d. Click Update.
5. To delete an existing cost center, do the following:

   a. In the Chargeback page, locate and select the cost center that requires deletion.
   b. From the Actions list, select Delete Cost Center.

      A confirmation message appears confirming the deletion.
   c. Click OK.

   The cost center is deleted and removed from the environment's cost center list and all the user and pool costs associated with the cost center are moved into the Uncategorized section.

Assigning uncategorized resources to a cost center

Unassigned resource costs are included in the total cost of all your cost centers. They represent user and pool costs that have not been assigned to a cost center. Learn how to move unassigned resource costs into an existing or a new Cloudera Observability cost center.

About this task

Steps on how to locate and move unassigned resource costs into an existing or a new Cloudera Observability cost center.

Important: To avoid cost duplication, resources must only be assigned one cost center.

Note: Only users with the ObservabilityAccountAdmin access role type can define the Chargeback settings, list, create, update, or delete cost centers, and view all the Chargeback related dashboards.

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 Main navigation panel, select Financial Governance.

   The Chargeback page opens.
3. Scroll down and select Uncategorized.

   The Uncategorized page opens.
4. Select the uncategorised usage criteria tab that is associated with your cost center settings.
5. Depending on the Chargeback usage criterion option you selected when you configured your Chargeback settings, from the Pools or Users page, select the check boxes of the uncategorized resources you require for your cost center.

   Note: Users and Pools may contain multiple instances of the same name, for example an *admin* user. In this case, select the name that is associated with the environment you require for your cost center.

   The Assign Cost Center dialog box opens.

6. Do one of the following:

   - To add the unassigned resource costs in a new cost center, do the following:

     a. From the Select Cost Center list, select New Cost Center.

        The Create a Cost Center page opens displaying the Cost Center details settings.
     b. In the Name field, enter a unique name for the cost center.
     c. In the Description field, enter a meaningful description for the cost center.
     d. Click Create.

     The CDP Chargeback page opens displaying a Success message, which denotes that the cost center was successfully created. Your new cost center is listed under the Cost Centers column and the Uncategorized page no longer displays the unassigned resource costs.
   - To add the unassigned resource costs in an existing cost center, do the following:

     a. From the Select Cost Center list, select the existing cost center that you require.
     b. Click Assign to Cost Center.

     The CDP Chargeback page opens displaying a Success message, which denotes that the unassigned costs were moved into the selected cost center, and the Uncategorized page no longer displays the unassigned resource costs.
7. Repeat steps 4 through 6 until all your uncategorized resources are placed in your Cloudera Observability cost centers.

Displaying your costs associated with a cost center

When a Cloudera Observability cost center is created, detailed summary Chargeback reports of the costs and resource usage for the environment are also generated that enable you to analyze the costs and the cost break-down associated with the cost center. You can view the current and historical costs and the resource usage associated with your cost centers.

About this task

Steps on how to view the detailed summary reports associated with a cost center.

The Cloudera Observability Chargeback reports visually display the tracked resource consumption and usage costs associated with the cost center for a specific time period that you select from the time-range list.

Within each report you can to drill down further:

- To view the users, resource pools, jobs, and queries with the highest costs.
- To view the health of a job or query of interest.
- To optimize costs by using the Cloudera Observability prescriptive guidance and recommendations that enable you to improve performance and resource usage.

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 Main navigation panel, select Financial Governance.

   The Cloudera Observability Chargeback page opens, which displays:

   - The total cost and CPU and memory hourly usage for all of your cost centers including those uncategorised resource usage costs not yet utilized.
   - Lists your existing cost centers that use the current criteria settings and displays the total costs and CPU and Memory hourly usage associated with each cost center.
   - The total cost and CPU and memory hourly usage for users and pools that are not yet assigned to a cost center in the uncategorized section.
3. To display a cost center's detailed report that includes costs for each chosen environment, do the following:

   a. From the time-range list, select a time-period that meets your requirements.
   b. From the Chargeback page, click inside the cost center row that requires analysis.

      The cost center's report page opens, which displays the following:

      - The total costs and CPU and Memory hourly usage for the cost center.
      - Lists the environments that are associated with the cost center and displays their total costs and CPU and Memory hourly usage.
   c. To view more details, such as which clusters created the highest costs within a specific environment, expand the environment by clicking its plus sign (+).
4. To display the users, pools, jobs, and queries that created the highest costs on a specific cluster of interest, do the following:

   a. Click inside the cluster row that requires more analysis.

      The cluster report Overview page opens, which displays the following:

      - The top 5 users whose jobs created the highest costs.
      - The top 5 pools whose jobs created the highest costs.
      - The top 25 jobs or queries that created the highest costs.
   b. To gain more insights on the health of a job or query, click the name of the job or query listed in the Top Jobs panel that requires more investigation.

      The job or query's summary page opens.
   c. Select the Health Checks and Execution Details tabs for more insights and if available read the optimization recommendations.
5. To view a full list of users, their job costs, and their CPU and Memory hourly usage, from the Overview page, select the Users tab.

   The Users report opens, which displays the following:

   - The name of the user.
   - The total cost that the user incurred.
   - The number of jobs that the user ran.
   - The CPU and Memory hourly usage.
   - The total job costs that the Cloudera Observability engines incurred; Impala, Hive, Spark, MapReduce, and Oozie.
6. To view a full list of pools, their job costs, and their CPU and Memory hourly usage, select the Pools tab.

   The Pools report opens, which displays the following:

   - The name of the pool.
   - The environment that the pool is associated with.
   - The total cost that the pool incurred.
   - The number of jobs that the pool ran.
   - The CPU and Memory hourly usage.
   - The total job costs that the Cloudera Observability incurred; Impala, Hive, Spark, MapReduce, and Oozie.

7. To view historical costs, change the time period currently displayed in the time-range field. For information on how to change the time period, click the Related Information link below.

Related Information

Specifying a time range

Downloading your chargeback costs

You can save the CPU, memory, and resource usage costs displayed in the Chargeback or Uncategorized pages of the Cloudera Observability UI as a spreadsheet report on your system, which can be used at a later time for further analysis using other tools or for printing and sharing with others. Learn how to download reports for both assigned and unassigned resource costs in Cloudera Observability.

About this task

Steps on how to locate and download a report for assigned and unassigned resource costs.

Note: The downloaded file contains raw data and as such may not display exactly as the format in the Cloudera Observability UI.

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 Main navigation panel, select Financial Governance.

   The Cloudera Observability Chargeback page opens.
3. To download a report of your assigned chargeback costs, do the following:
   a) From the time-range list, select a time-period that meets your requirements.
   b) From the Actions list, click Download Report.

      A Download Report message appears stating that the report is generating.

      Important: Navigating to another page or browser tab during the generation process will automatically stop the generation and download process.

      When completed, your assigned chargeback costs for the time-period selected are downloaded as a Microsoft Excel file to your Downloads folder.
4. To download a report of your unassigned chargeback costs, do the following:
   a) From the time-range list, select a time-period that meets your requirements.
   b) Scroll down and select Uncategorized.

      The Uncategorized page opens.
   c) From its topmost menu bar, click Download Report.

      A Download Report message appears stating that the report is generating.

      Important: Navigating to another page or browser tab during the generation process will automatically stop the generation and download process.

      When completed, your unassigned chargeback costs for the time-period selected are downloaded as a Microsoft Excel file to your Downloads folder.

Triggering action alerts across jobs and queries

You can trigger action alerts, that are defined by you, across your workload applications, jobs, and queries whilst they are running with the Cloudera Observability Auto Actions feature. When a workload application, job, or query matches the action's defined threshold value, the auto action event is triggered. For example, you may have a scenario where too much memory is being allocated to specific jobs and you would like to take an action before a problem occurs, such as avoiding memory exhaustion. In this case, you can create an auto action that triggers a notification alert when a job is identified as having an over-allocation of memory. You can then either manually take steps to alleviate the problem or include the Kill action option that stops the job in question.

Important: Before you can use the Auto Actions feature you must set the required auto actions configuration properties in Telemetry Publisher. For information on how to enable the Telemetry Publisher Auto Actions property settings, click the Related Information link below.

Note: At this time the Auto Actions feature is only available for Classic Cluster and CDP Private Cloud Base using Cloudera Manager version 7.6.2, or above, environments. Users using CDP Data Hub clusters require Cloudera Runtime version 7.2.18 running Cloudera Manager version 7.12.0, or above.

Considerations and Limitations

The following describes consideration decisions and limitations when using Auto Actions:

- Terminating a workload application, job, or query could impact other workloads. Especially when another workload is dependent on the results of the terminated workload application, job, or query. Before triggering a Kill type action, Cloudera recommends using the Notification only action alert until you have verified that no issues will arise.
- By default, the Cloudera Observability UI limits the amount of displayed audit events to 500 and sorts them in ascending order (newest time stamp first). To display older audit events, change the date range duration and/or the time range duration from the time-range list on the Auto Actions Events page.
- Too Fast To Collect: The minimum polling interval is one minute. If you have jobs or queries that overlap or start before the minimum polling interval is completed there may not be enough time for Cloudera Observability to evaluate your auto action’s definition.

  For example, if Cloudera Observability starts polling at 1:00:00 PM and polling finishes by 1:00:10 PM (10 seconds) and then a job starts at 1:00:12 PM and finishes before 1:01:00 PM, there is not enough of a time lapse for Cloudera Observability to evaluate and trigger your action alert.
- Too Fast to Kill: Under normal conditions the evaluation and invocation phases of an auto action is within the span of one minute. If you have jobs or queries whose run time is less than one minute, Cloudera Observability may complete the evaluation phase but not have time to complete the invocation phase, such as terminating the job. Depending on the context of your auto action, this may or may not be an issue. But if, for example, you have a workload cluster that is dedicated for specific jobs and a rogue job is run before the action is triggered, then this could be an issue

Related Information

Enabling the Auto Actions feature in Telemetry Publisher

Creating an auto action event

The steps to create a Cloudera Observability auto action definition, which is triggered when a workload application, job, or query matches the auto action’s definition threshold. For example, when a job is taking too long to run it may delay other jobs waiting in the queue. With Auto Actions, you can create an auto action alert that informs you through an email when a job is exceeding its usual runtime so that you can decide whether to manually take steps to alleviate the problem or have an auto action that will terminate the job or query process for you.

About this task

Describes how to create a Cloudera Observability Auto Action definition.

Note: These instructions assume that you have set the required auto actions configuration properties in Telemetry Publisher. For information about the properties and how to enable the Telemetry Publisher Auto Actions property settings, click the Related Information links below.

Note: At this time the Auto Actions feature is only available for Classic Cluster and CDP Private Cloud Base using Cloudera Manager version 7.6.2, or above, environments. Users using CDP Data Hub clusters require Cloudera Runtime version 7.2.18 running Cloudera Manager version 7.12.0, or above.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.
   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.
   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.
   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

      The Environment navigation panel opens.
2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required for analysis.

   Tip: The page's title is displayed in the browser tab.
3. Select the Auto Actions tab.
4. Do one of the following:

   - If no other auto actions exist, select the Management tab and then click Auto Actions Setup.
   - If other auto actions exist, click Create Auto Action.

   The Auto Actions Create page opens.
5. In the Auto Action Name field, enter a unique name that is easily identifiable.
6. From the Scope list, select the workload component service that is to be monitored by the action.

   For example, if you want your action to only evaluate Spark related applications, select Spark Application.

7. Define the conditions for the auto action by doing at least one of the following:
   - Specify the Criteria:
     a. From the Criteria list, select a criteria item.
     b. From the Operator list, select the required operator.

        Important: Cloudera Observability does not validate regular expressions when using the matches regex operator for string criteria types, such as User, Pool, or Query Name. Neither does it display help for poor syntax. Cloudera recommends validating your code and syntax before entering your regular expression in the Value field.
     c. In the Value field, enter the value for this filter.

        Tip: You can define multiple AND filters for the Criteria by clicking the plus sign.

        Note: An Auto Action does not require the Criteria filter as long as a Trigger condition is defined:
        - When included, only those workloads that are run on the selected workload component service and meet the criteria conditions are tested by the Trigger.
        - When not included, all workloads that are run on the selected workload component service are tested by the Trigger.
   - Specify the trigger for the auto action by doing the following:
     a. From the Metric list, select a metric item.
     b. From the Operator list, select the required operator.

        Note: The in between operator is inclusive.
     c. In the Value field, enter the value for this trigger condition.

        Tip: You can define multiple OR conditions for the trigger by clicking the plus sign.

        Note: An Auto Action does not require the Trigger filter as long as a Criteria condition is defined:
        - When included, workloads that are run on the selected workload component service and meet the criteria conditions are tested by the Trigger.
        - When not included, only those workloads that are run on the selected workload component service and meet the criteria conditions are evaluated.
8. From the Select Action options, select the action that is to be performed when the condition is met.

   Warning: Terminating a workload application, job, or query could impact other workloads. Especially when another workload is dependent on the results of the terminated workload application, job, or query. Before triggering a Kill type action, Cloudera recommends using the Notification only action until you have verified that no issues will arise if the workload application, job, or query is terminated.
9. From the Notification section do the following:
   a. In the Emails field, enter the email address that you use to log into Cloudera Observability.
   b. In the Subject field, enter the subject for the email that distinguishes the subject matter from other auto action emails.
10. Click Create, which creates the action and its audit log, adds it on the Auto Actions Events and Management pages, and displays its status as Enabled and its most recent event type as Create.

Results

When a workload application, job, or query meets the auto action's criteria and trigger conditions, the action event is triggered.

Related Information

Enabling the Collection of Auto Action Data by Telemetry Publisher

Telemetry Publisher configuration settings for Auto Actions

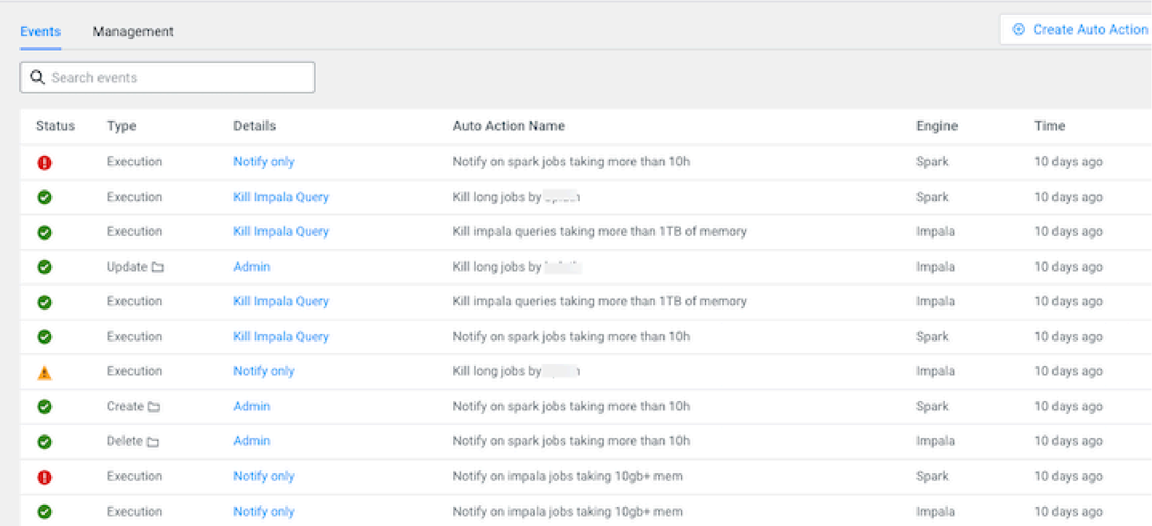
Events and management details of auto actions

Describes the fields in the Auto Actions Events and Management pages of Cloudera Observability.

The Events and Management pages help you monitor, manage, and troubleshoot your Auto Actions.

Events

The **Events** page displays information about your auto action audit events:



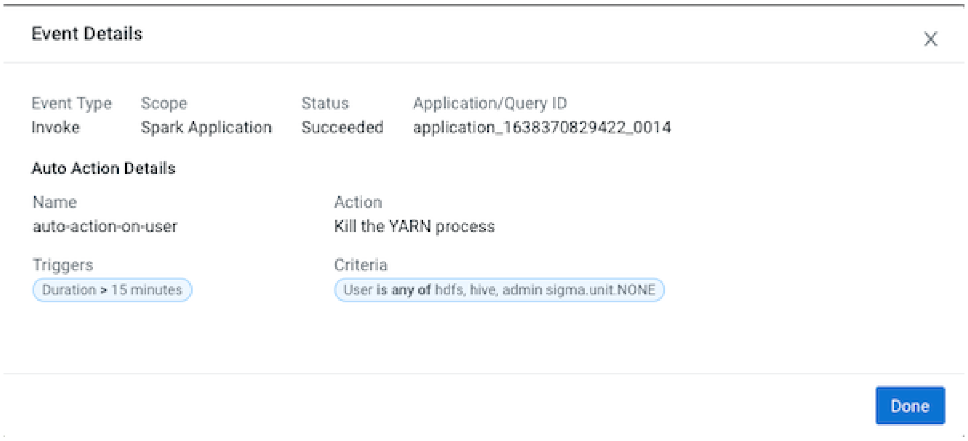
Status	Type	Details	Auto Action Name	Engine	Time
!	Execution	Notify only	Notify on spark jobs taking more than 10h	Spark	10 days ago
✓	Execution	Kill Impala Query	Kill long jobs by [redacted]	Spark	10 days ago
✓	Execution	Kill Impala Query	Kill impala queries taking more than 1TB of memory	Impala	10 days ago
✓	Update	Admin	Kill long jobs by [redacted]	Impala	10 days ago
✓	Execution	Kill Impala Query	Kill impala queries taking more than 1TB of memory	Impala	10 days ago
✓	Execution	Kill Impala Query	Notify on spark jobs taking more than 10h	Spark	10 days ago
⚠	Execution	Notify only	Kill long jobs by [redacted]	Impala	10 days ago
✓	Create	Admin	Notify on spark jobs taking more than 10h	Spark	10 days ago
✓	Delete	Admin	Notify on spark jobs taking more than 10h	Impala	10 days ago
!	Execution	Notify only	Notify on impala jobs taking 10gb+ mem	Spark	10 days ago
✓	Execution	Notify only	Notify on impala jobs taking 10gb+ mem	Impala	10 days ago

It contains the following audit entry fields:

- Status, which displays the results of the auto action event as an icon. Where, green indicates that the action was successful (SUCCEEDED). All the other icons indicate that the action was unsuccessful (FAILED).
- Type, which displays the auto action's audit event category type, such as Create or Update.

- Details, which displays the type of action. When clicked the auto action’s Event Details audit log opens, as shown in the following Invoke and Update event type example images:

Figure 1: Invoke Event Type Example



Event Details

Event Type	Scope	Status	Application/Query ID
Invoke	Spark Application	Succeeded	application_1638370829422_0014

Auto Action Details

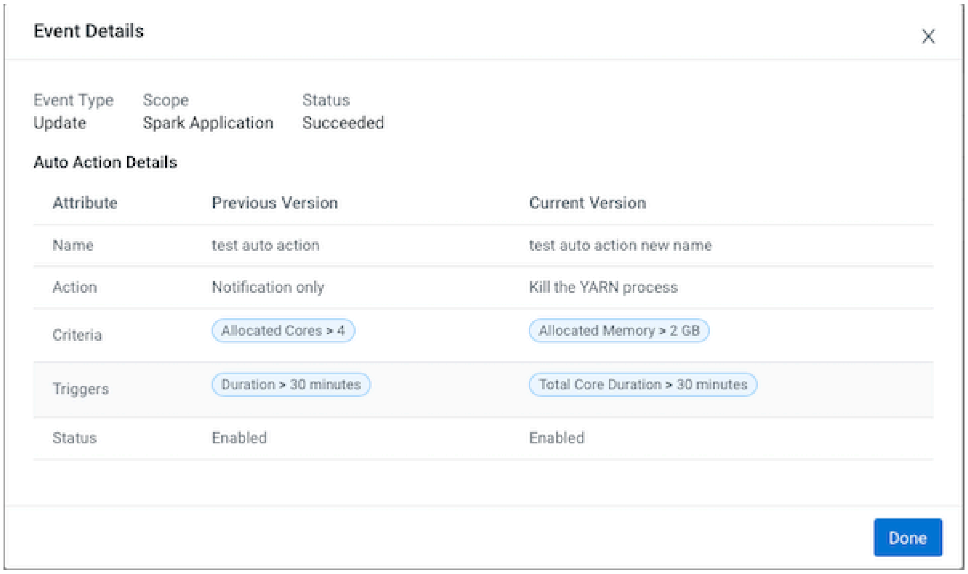
Name: auto-action-on-user

Action: Kill the YARN process

Triggers: Duration > 15 minutes

Criteria: User is any of hdfs, hive, admin sigma.unit.NONE

Figure 2: Update Event Type Example



Event Details

Event Type	Scope	Status
Update	Spark Application	Succeeded

Auto Action Details

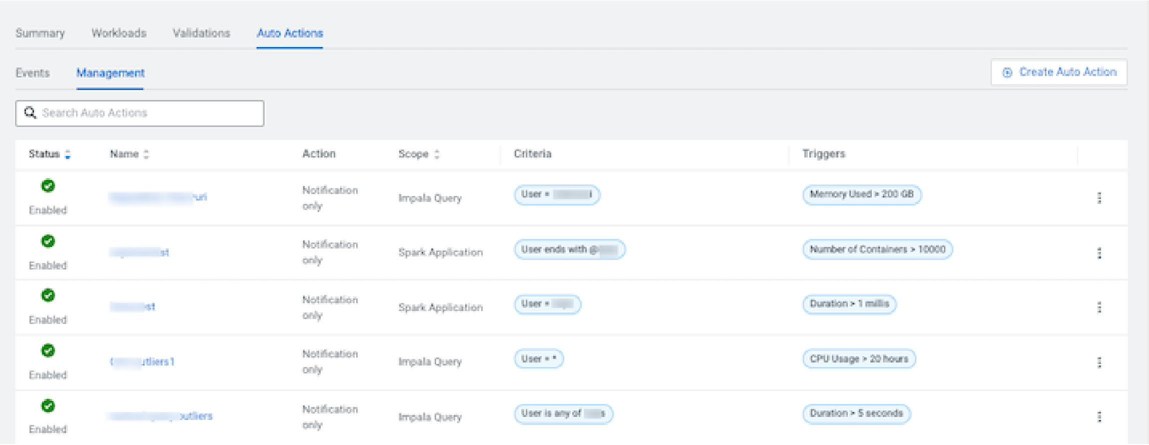
Attribute	Previous Version	Current Version
Name	test auto action	test auto action new name
Action	Notification only	Kill the YARN process
Criteria	Allocated Cores > 4	Allocated Memory > 2 GB
Triggers	Duration > 30 minutes	Total Core Duration > 30 minutes
Status	Enabled	Enabled

- Auto Action Name, which displays the unique name you entered for the auto action.
- Scope, which displays the workload component service that is monitored by the action.
- Time, which displays the time stamp of when the auto action's audit event occurred.

Important: By default, the Cloudera Observability UI limits the number of displayed audit events to 500 and sorts them in ascending order (newest time stamp first). To display older audit events, from the time-range list on the Auto Actions Events page, change the date range duration and/or the time range duration.

Management

The **Management** page displays your auto action's defined settings and state:



Status	Name	Action	Scope	Criteria	Triggers
Enabled	[redacted]uri	Notification only	Impala Query	User = [redacted]	Memory Used > 200 GB
Enabled	[redacted]st	Notification only	Spark Application	User ends with @[redacted]	Number of Containers > 10000
Enabled	[redacted]st	Notification only	Spark Application	User = [redacted]	Duration > 1 millis
Enabled	[redacted]utliers1	Notification only	Impala Query	User = *	CPU Usage > 20 hours
Enabled	[redacted]outliers	Notification only	Impala Query	User is any of [redacted]s	Duration > 5 seconds

It contains the following entry fields:

- Status, which displays the current state of the action, as either Enabled or Disabled.
- Name, which contains the name of the auto action. When clicked the auto action's definition settings page opens.
- Action, which displays the name of the action that is invoked when the auto action is triggered, such as Notify Only.
- Scope, which displays the workload component service that is monitored by the action.
- Criteria, which displays the action's Criteria filters. These are attributes with static values that remain the same during the execution of a job or query.
- Triggers, which displays the action's Trigger conditions. These are attributes with dynamic values that change during the execution of a job or query.

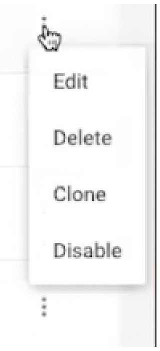
Managing your auto actions

Describes how to update, delete, duplicate, and disable an auto action event.

The following Auto Actions management tasks are performed in the **Management** page, which is accessed by selecting the Auto Actions tab in the Cluster Summary page and then selecting the **Management** tab.

Updating your auto action

In the Management page, click the action’s vertical ellipsis, as shown in the following image, and select Edit. Make your changes and then click Update.



Deleting an auto action

In the **Management** page, click the action's vertical ellipsis, and select Delete. In the confirmation message, click OK to confirm. The action is permanently removed.

Note: Unless the action is no longer required, Cloudera recommends disabling the action, as you may require the action at another time.

Duplicating an auto action

In the **Management** page, click the action's vertical ellipsis, and select Clone. Replace the existing name with a new unique name for the cloned auto action, make any other changes, and then click Create. A new auto action is created and is displayed on the **Management** page.

Note: You must change the name of the cloned auto action before a new one can be created.

Disabling an auto action

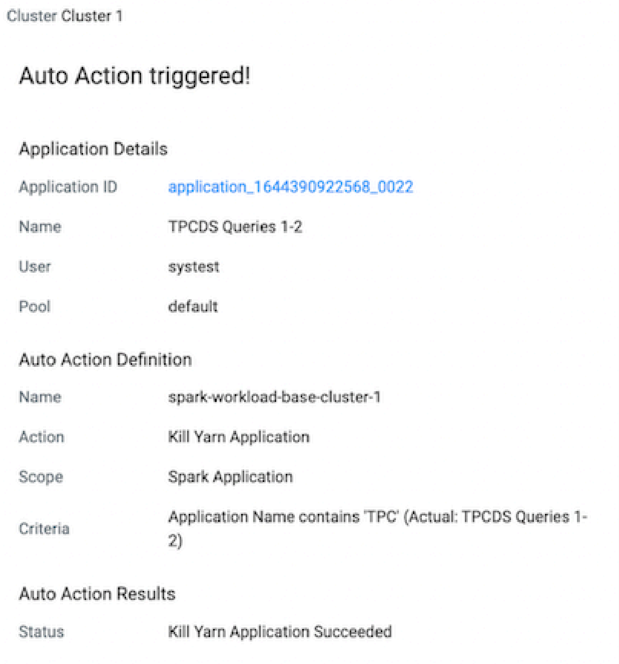
In the **Management** page, click the action's vertical ellipsis, and select Disable. In the confirmation message, click OK to confirm. The action is no longer active and the Disabled state is displayed in the action’s Status column on the **Management** page.

Auto action email notification examples

Examples of a Cloudera Observability Auto Actions alert notification email.

The following sample email notifications are sent when the listed application meets the action's criteria and the trigger conditions, which are also included in the email notification. The sample email notifications are split into three sections:

- In the Application Details section, the Application ID contains a link to the workload application, job, or query.
- In the Auto Action Definition section, the Trigger and the Criteria definition display both the value and file size type that you defined and in brackets the Actual value, in megabytes, captured by the engine.
- In the Auto Action Results section, the results of the invoked auto action is displayed.



Cluster Cluster 1

Auto Action triggered!

Application Details

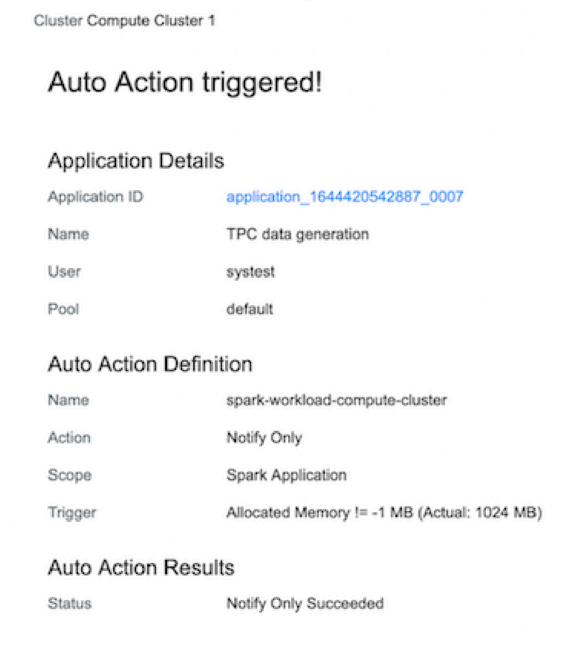
Application ID	application_1644390922568_0022
Name	TPCDS Queries 1-2
User	systest
Pool	default

Auto Action Definition

Name	spark-workload-base-cluster-1
Action	Kill Yarn Application
Scope	Spark Application
Criteria	Application Name contains 'TPC' (Actual: TPCDS Queries 1-2)

Auto Action Results

Status	Kill Yarn Application Succeeded



Cluster Compute Cluster 1

Auto Action triggered!

Application Details

Application ID	application_1644420542887_0007
Name	TPC data generation
User	systest
Pool	default

Auto Action Definition

Name	spark-workload-compute-cluster
Action	Notify Only
Scope	Spark Application
Trigger	Allocated Memory != -1 MB (Actual: 1024 MB)

Auto Action Results

Status	Notify Only Succeeded

Working with cluster reports

When enabled, Cloudera Observability mails you daily statistics about your cluster, such as your cluster's performance.

You can use this information to keep track of the number of queries and jobs that are running on your cluster, identify users that are running large numbers of queries, and be alerted to spikes in the number of failed jobs or queries. This saves you from logging in to your cluster and analyzing these numbers yourself.

Cloudera Observability sends the Cluster Report daily at 1:00 AM PDT with the previous day's statistics. It is split into a Data Warehouse section and a Data Engineering section, where:

- The Data Warehouse section shows the total number of queries and the number of failed queries for the day, as well as how those statistics compare to the same day of the previous week. It also lists the users that ran the most queries and the number of queries they ran in the 24 hour period.
- The Data Engineering section shows the total number of jobs, the number of failed jobs, and how those numbers compare to the same day of the previous week.

Note: Cluster reports are available for Classic Cluster, Private Cloud Base, Data Hub, Virtual Warehouse, and Virtual Cluster Cloudera Observability environments only and cannot be managed for other users.

Enabling and disabling cluster reports

The steps that enable you to receive daily reports that help you track and identify changes that could be or become potential problems with your jobs and queries. When enabled reports that compare your jobs and queries with historical data collected from the previous week are sent to the email address that you use to log into Cloudera Observability.

About this task

Describes how to enable and disable daily cluster reports.

Note: Cluster reports are available for Classic Cluster, Private Cloud Base, Data Hub, Virtual Warehouse, and Virtual Cluster Cloudera Observability environments only and cannot be managed for other users.

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.

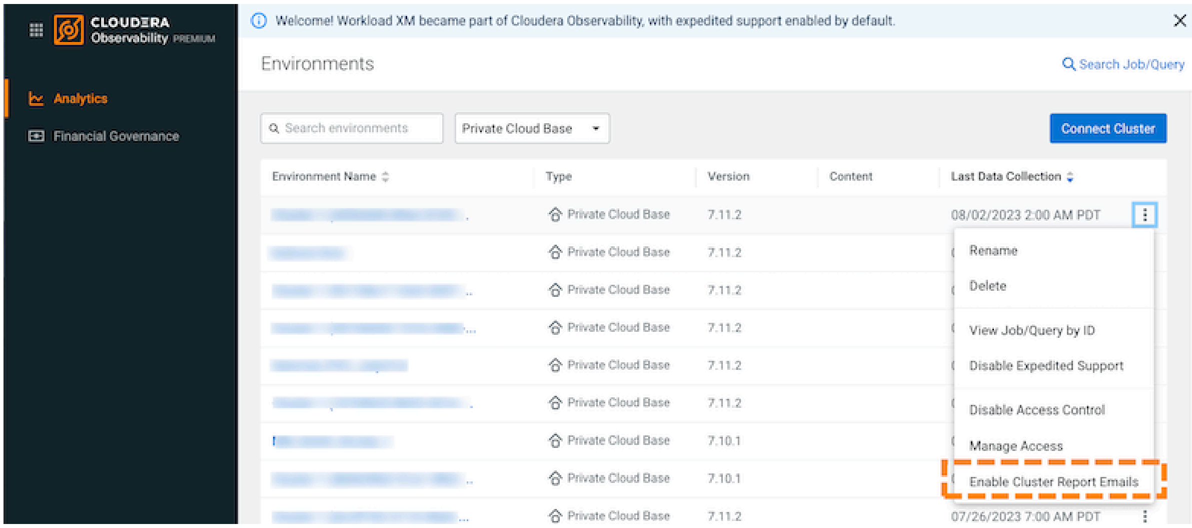
      The Cloudera Observability landing page opens.
2. From the Cloudera Observability **Environments** page, locate the environment from which you require a daily report.

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

3. Do one of the following:

    - To enable Cluster reports:

      From the environment's Actions list, select Enable Cluster Report Emails.

      Welcome! Workload XM became part of Cloudera Observability, with expedited support enabled by default.

      Environments

Environment Name	Type	Version	Content	Last Data Collection
[redacted]	Private Cloud Base	7.11.2		08/02/2023 2:00 AM PDT
[redacted]	Private Cloud Base	7.11.2		
[redacted]	Private Cloud Base	7.11.2		
[redacted]	Private Cloud Base	7.11.2		
[redacted]	Private Cloud Base	7.11.2		
[redacted]	Private Cloud Base	7.11.2		
[redacted]	Private Cloud Base	7.10.1		
[redacted]	Private Cloud Base	7.10.1		
[redacted]	Private Cloud Base	7.11.2		07/26/2023 7:00 AM PDT

      An Email Report message appears confirming enablement.
    - To disable Cluster reports:

      When you no longer require a daily report for your environment, from the environment's Actions list, select Disable Cluster Report Emails.

      An Email Report message appears confirming disablement.

Understanding, identifying, and addressing problems with Cloudera Observability

Learn the tasks that help you analyze, identify and troubleshoot job and query abnormalities and failures, optimize workloads, and improve job performance with Cloudera Observability.

Specifying a time range

Enable a more in-depth analysis about your costs and workloads by displaying current or historical data for a specific time period.

About this task

Describes how to change the currently displayed time period from the time-range list, which appears on the **Cluster Summary**, **Engine Summary**, and **Workload Summary** pages.

By default, Cloudera Observability displays workload data for the last 24 hours. If there is no data available during that time, Cloudera Observability displays the nearest date range that is available.

Note: The time-range list converts universal time to the user's local timezone.

The following steps describe, with examples, how to change the time period from the **Cluster Summary** page.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.
   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.

   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.

   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

      The Environment navigation panel opens.

2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required for analysis.

   To display the **Cluster Summary** page for a Data Lake, Database Catalog, Data Engineering, and Data Hub environment type, do one of the following:

   - From the Environment panel, expand the service's category and depending on the service, locate and select the Data Hub's cluster, the Data Engineering's Virtual Cluster, or the Data Warehouse's Virtual Warehouse that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster, the Data Engineering's Virtual Cluster, or the Data Warehouse's Virtual Warehouse that is required for analysis.

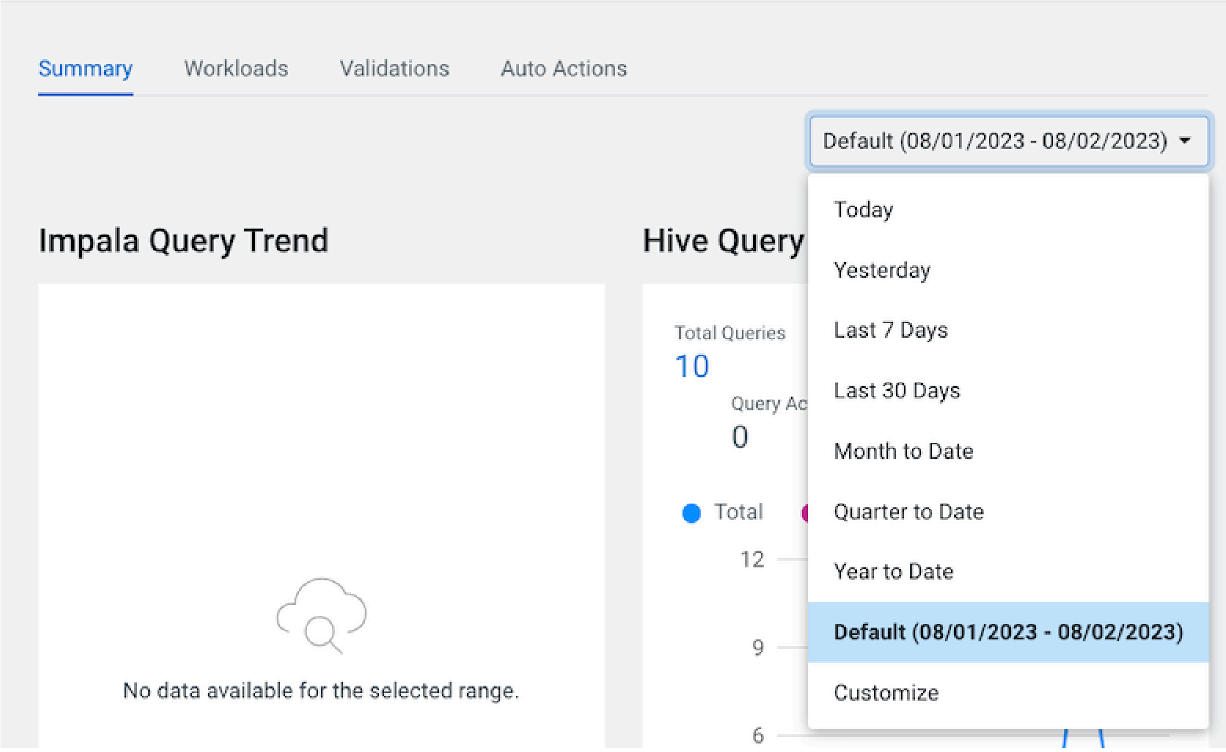
   Tip: The page's title is displayed in the browser tab.

3. From the time-range list, do one of the following:

   - For a predefined period, select one of the default periods of time that meets your requirements.
   - For an exact date and time range, select Customize and then either, enter the date and time range using the YYYY/MM/DD HH:MM:SS format for the beginning and the ending time period, or in the calendar element, select the beginning and ending time period.

   The following image shows an example of the time-range list on a Cluster Summary or Engine Summary page:

   Figure 3: Analytics time-range list



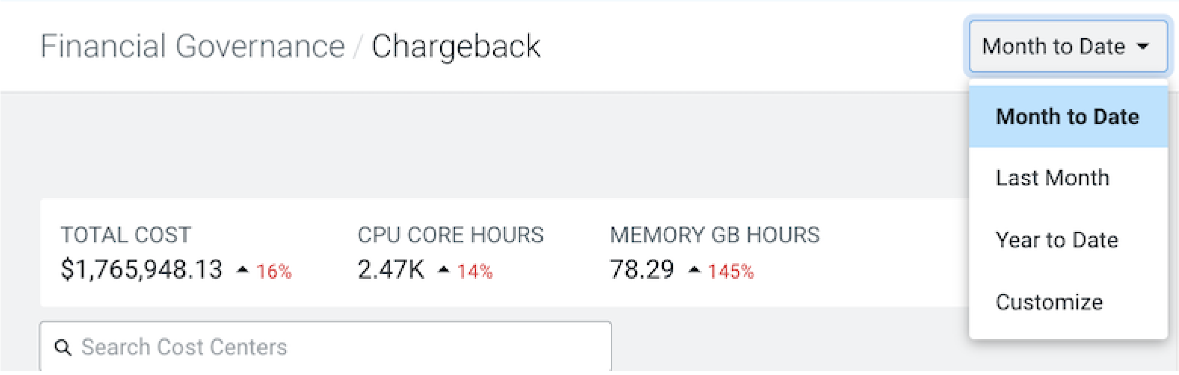
4. Click Ok, which clears any existing workload data from the chart and table components and updates your workload data for the chosen time period.

Results

All charts and tables in Cloudera Observability are updated to reflect the workload data for the chosen time period.

Note: The time-range list is also available on the Financial Governance Chargeback pages for historical analysis, as shown in the following example:

Figure 4: Financial Governance time-range list



Analyzing your tables

Minimize costs and maximize query performance by gaining more insights into your tables, including which tables are frequently or infrequently accessed, with the Cloudera Observability Metastore Analytics feature. By understanding your tables and their metadata, such as a table's data volume or how often a query accesses a table, helps you troubleshoot, make informed decisions about your data, and ensures that your table data is in accordance with your Storage Policy.

The Cloudera Observability Metastore Analytics feature, collects and filters the Hive Metastore (HMS) metadata into meaningful views of your tables, including which tables are hot (high frequency of access) and which tables are cold (little or no frequency of access).

Note: At this time, the Cloudera Observability Metastore Analytics feature is only available for CDP Private Cloud Base using Cloudera Manager version 7.10.1, or above, and CDP Data Hub clusters using Cloudera Manager version 7.11.0 environments. Also, to view the table metadata values in the UI, the Hive Metastore (HMS) must be deployed in the Workload cluster.

The Cloudera Observability Metastore Analytics feature displays information that enables you to:

- Identify and track sudden table changes, such as a table's data size, the number of partitions, or the number of rows that may impact the processing of your queries. The HMS Extract, which is updated daily, lists the details about each table available in your system regardless of whether they have been queried or not. It includes the table's configurations and if enabled the table's statistics, as well as size related information, such as the table's volume, the number of partitions, and the number of rows.
- View and analyze the most frequently accessed tables from the Data Temperature's Hot Tables chart widget on the environment's Cluster Summary page. With this information you can decide which tables should be moved to performance-efficient storage, such as an SSD that can improve a query's performance due to its fast processing speed, especially queries that access large amounts of data.
- View and analyze the least frequently accessed tables from the Data Temperature's Cold Tables chart widget on the environment's Cluster Summary page. With this information you can decide which tables should be purged or moved to cost-efficient storage, which will save platform costs.
- Analyze and troubleshoot inefficiencies within your tables, such as the wrong table type or storage format. The HMS Tables view and the Table Details panel display details about each table within your system, such as the table's location, database, column names, and properties.
- Identify tables that contain huge amounts of data. With this information you can decide if partitioning is required or if more partitioning is required, which improves query performance and costs by reducing the amount of data that has to be retrieved, manipulated, and outputted, as well as making your tables easier to manage.

Understanding the Cloudera Observability metastore analytics UI elements

Learn about the Cloudera Observability Metastore Analytics UI elements that display the Hive Metastore (HMS) metadata information about your tables.

Note: At this time, the Cloudera Observability Metastore Analytics feature is only available for CDP Private Cloud Base using Cloudera Manager version 7.10.1, or above, and CDP Data Hub clusters using Cloudera Manager version 7.11.0 environments. Also, to view the table metadata values in the UI, the Hive Metastore (HMS) must be deployed in the Workload cluster.

About the Cloudera Observability data temperatures

In Cloudera Observability Hot and Cold represents the number of times a query accesses the table. Where, the color and the depth of color represents the number of times a query accesses the table in relation to all the other tables in your system:

- Hot tables (red) - are tables that were frequently accessed during the selected time-period.

  Note: Cloudera recommends moving Hot tables into performance-efficient storage, like an SSD, due to its fast processing speed.

- Cold tables (blue) - are tables that were infrequently accessed during the selected time-period. This includes tables where no queries (zero) accessed their data during the selected time-period and by definition are considered the coldest tables.

  Note: Cloudera recommends purging tables that are no longer required or moving infrequently accessed tables into cost-efficient storage. This saves platform costs and improves the performance of jobs that access the table data more frequently by creating more storage capacity.

About the data temperature charts

The Cloudera Observability Metastore Analytics feature has several UI elements that describe your table data.

The following charts display the data temperature information:

- Located on the environment's **Cluster Summary** page, the Data Temperature chart automatically displays the top 25 most frequently queried and the bottom 25 least frequently queried tables from both the Hive and Impala engines in the Hot Tables and the Cold Tables chart widgets respectively.
- Located on the Hive and Impala engine summary pages, the Data Temperature chart automatically displays the tables that were most frequently queried by their engine in the Hot Tables chart widget respectively.

Note: The Hot and Cold table chart widgets do not reflect tables queried by the Spark application.

Hovering over a Hot or Cold table with your mouse pointer, displays general information, such as, the number of queries that accessed the table, the total table size in gibibytes, the number of partitions that comprise the table, the number of files that make up the table, and whether statistics were enabled on the table's rows.

Clicking the table's name of interest in either the Hot Tables or Cold Tables chart widget or in the HMS Tables view in the HIVE METASTORE category of your environment's cluster, opens the table's **Overview Details** side drawer panel, which displays more information about the table.

About the Overview Details side drawer panel

The **Overview Details** side drawer panel describes more information about the table. Based on the table’s HMS metadata, such as the table's schema, database location, partitions, structure, and relationships, the information displayed may vary. It also describes the table's columns, such as the column names and their data types, and the table's metadata properties that include user-defined and predefined key-value pairs.

It is accessed by clicking on the table's name of interest in either the Hot or Cold Tables chart widget or from the HMS Tables view, which is found by selecting the Tables tab in the HIVE METASTORE category of your environment's cluster.

The information collected from your table's HMS metadata is divided into sub categories and displayed in the following tabs:

- Details
- Columns
- Properties

Where, each tab displays the following general table values:

- Volume, which displays the total table size in Kilobytes.
- Rows, which displays the number of records in the table.
- Partitions, which, if applicable, displays the number of segments that comprise the table.
- Total Files, which displays the number of files that make up the table.

If your table contains partitions, the Distribution Across Partitions section is also displayed, which contains the following distribution cards:

- DATA SIZE, which displays the total data size of the table selected and the distribution across its partitions.
- NUMBER OF FILES, which displays the total number of files within the table selected and the distribution across its partitions.

- NUMBER OF ROWS, which displays the total number of rows within the table selected and the distribution across its partitions.

Note: The distribution cards display the minimum and maximum values, the median value, and the median Q1 and Q3 quantiles (25th and 75th percentiles) that summarize a specific set of metrics and how they are distributed across the table's partitions. These cards enable you to analyze and gain insights into the lowest and highest values, the spread of these values and where the majority of the values reside within the spread, and where outliers reside.

The HMS metadata that is displayed in each tab is dependent on the table’s underlying data on which it is built. The following tables describe the most common parameters displayed in the Details, Columns, and Properties tabs:

Table 6: Details

Parameter	Description
Historical Trend chart	Displays the historical values for the Rows, Data Volume, and Partitions.
Database name	The database in which the table resides.
Compressed	Displays a True or False value depending on whether data compression been applied.
Location	The table’s location in HDFS.
Partition Keys	The name/s of the partition keys that are responsible for data distribution across the nodes.
Raw Data Size	The raw data size of the table, in the nearest byte unit.
Storage Format	The table’s storage format, such as but not limited to: • JDBC • LazySimple • Orc • Parquet
Stats Enabled	Displays a True or False value depending on whether statistics were enabled.
Table Type	The table’s type, such as but not limited to: • EXTERNAL_TABLE, which defines a table whose data is stored in the location specified during table creation. • MANAGED_TABLE, which defines a table whose data is stored in the warehouse directory. • VIRTUAL_TABLE, which defines a table that is the result of a query which has not materialized and whose data is not stored.
Transactional	Displays a True or False value depending on whether the table contains one or more ACID semantic properties.
Created	The date when the table was created, using the MM- DD-YYYY date format. For example, 06-25-2023.

Table 7: Columns

Parameter	Description
Column Name	Lists the Column field names.

Parameter	Description
Type	The Hive data type, as one of the following: • bigint • binary • boolean • chara • date • decimal • double • float • int • smallint • string • timestamp • tinyint • varchar
Comment	An informative note about the column that was added during table creation.

Table 8: Properties

Parameter Sections	Description
Table Properties	Predefined and user-defined metadata key-value pair properties.
SerDe Properties	Serialization and deserialization properties.
Storage Descriptor Properties	Metadata that describes the physical storage properties of the data residing in the table.

Understanding the Hive Metastore category

Learn about the Cloudera Observability Hive Metastore category that lists the details about each table available in your system and visually displays the current state and activity of your tables in the selected environment.

For users with a Hive Metastore deployment, Cloudera Observability captures the Hive Metastore (HMS) metadata about your tables and displays it into meaningful cards and views. These can be found in the HMS Summary and HMS Tables views, which display the current state and activity of all your tables and list details about each table available in your system, regardless of whether they have been queried or not.

The metric results displayed are dependent on your table’s schema and their HMS properties and parameters.

Note: Rounding rules are applied.

About the HMS Summary view

The Hive Metastore (HMS) Summary view visually displays information about the current state and activity of all your tables in the selected Environment.

It contains three sections:

- Overview
- Table Insights
- Table Statistics

Overview

The Overview section displays general information about your tables and the number of databases in which they reside.

It displays the following cards:

Table 9: Overview cards

Card	Description
DATABASES	The number of databases in which your tables reside.
TABLES	The number of tables and the percentage of tables that are External and Managed.
VIEWS	The number of views and the percentage of views that are Materialized and Virtual.
PARTITIONS	The number of partitions.

Tables Insights

This section displays the physical structure of your tables using the base table metrics. These cards enable you to identify how well your tables are structured for increased performance.

Note: The gathering of metric data for this section does not require statistics enablement.

It displays the following cards:

Note: The distribution cards display the minimum and maximum values, the median value, and the median Q1 and Q3 quantiles (25th and 75th percentiles) that summarize a specific set of metrics and how they are distributed across the table's partitions. These cards enable you to analyze and gain insights into the lowest and highest values, the spread of these values and where the majority of the values reside within the spread, and where outliers reside.

Table 10: Tables insights cards

Card	Description
NUMBER OF PARTITIONS	The partition distribution across all tables.
PARTITION KEY SIZE	The partition key size distribution across all tables.
COLUMN SIZE	The column size distribution across all tables.
NUMBER OF BUCKET COLUMNS SIZE	The bucket column size distribution across all tables.
BUCKETED TABLES	The number of bucketed tables and the percentage across all tables.
COMPRESSED TABLES	The number of compressed tables and the percentage across all tables.
NON PARTITIONED TABLES	The number of non partitioned tables and the percentage across all tables.
PARTITIONED TABLES	The number of partitioned tables and the percentage across all tables.
TABLES WITH ARRAY COLUMNS	The number of tables with array columns and the percentage across all tables.
TABLES WITH BINARY COLUMNS	The number of tables with binary columns and the percentage across all tables.
TABLES WITH MAP COLUMNS	The number of tables with map columns and the percentage across all tables.
TABLES WITH STRUCT COLUMNS	The number of tables with struct data type columns and the percentage across all tables.
TEMPORARY TABLES	The number of temporary tables and the percentage across all tables.

Table Statistics

This section displays the physical characteristic metrics of those tables that have statistics enabled, such as the volume of data, the number of rows and files, and how these values are distributed. Table statistics improve the optimization of queries by the engine for increased performance. Understanding the size and volume of a table helps the engine organize the workload appropriately, such as for a join or insert operation.

Note: To display this section's metrics, statistics must be enabled on your most important tables and materialized views. To verify that table statistics are available for a table, click the Related Information link below.

It displays the following cards:

Table 11: Table statistics cards

Card	Description
STATISTICS ENABLED	The number and percentage across all tables with statistics enabled.
DATA VOLUME	The total size of tables with statistics enabled and the distribution across all tables.
NUMBER OF FILES displayed by distribution	The total number of files with statistics enabled and the distribution across all tables.
NUMBER OF ROWS	The total number of rows with statistics enabled and the distribution across all tables.
TOTAL DATA VOLUME	The total size of your tables with statistics enabled and the size of each storage format.
NUMBER OF FILES displayed by type	The total number of files that form the tables with statistics enabled and their storage formats displayed as a percentage as a whole.

About the HMS Tables view

The HMS Extract, which is updated daily, is displayed in the Hive Metastore (HMS) Tables view. It lists the details about each table available in your system, regardless of whether they have been queried or not.

The Tables view contains the following columns:

Column Name	Description
Table	The name of the table.
Database	The database in which the table resides.
Partitions	The number of partitions.
Volume	The total table size in bytes.
Rows	The number of records in the table.
Files	The number of files that make up the table.
Frequency of Access	The number of times queries have accessed the table. **Note:** This value does not reflect tables used by the Spark application.
Storage Format	The table's storage format, such as but not limited to: • JDBC • LazySimple • Orc • Parquet

Column Name	Description
Table Type	The table’s type, such as but not limited to: • EXTERNAL_TABLE, which defines a table whose data is stored in the location specified during table creation. • MANAGED_TABLE, which defines a table whose data is stored in the warehouse directory. • VIRTUAL_TABLE, which defines a table that is the result of a query which has not materialized and whose data is not stored.

Related Information

Statistics generation and viewing commands in Data Hub

Displaying the Metastore Analytics

Learn how to analyze, identify, and troubleshoot table changes and inefficiencies, including which tables are hot and which tables are cold.

About this task

Steps for troubleshooting your tables and their data with the Cloudera Observability Metastore Analytics feature.

Note: At this time, the Cloudera Observability Metastore Analytics feature is only available for CDP Private Cloud Base using Cloudera Manager version 7.10.1, or above, and CDP Data Hub clusters using Cloudera Manager version 7.11.0 environments. Also, to view the table metadata values in the UI, the Hive Metastore (HMS) must be deployed in the Workload cluster.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.
   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.
   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.
   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

      The Environment navigation panel opens.
2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required for analysis.

   To display the **Cluster Summary** page for a Data Hub environment type, do one of the following:

   - From the Environment's panel, expand the service's category and locate and select the Data Hub's cluster that is required for analysis.
   - In the Data Services table, drill-down through the service links to locate and select the Data Hub's cluster that is required for analysis.

   Tip: The page's title is displayed in the browser tab.

3. To display the top 25 hot tables and the bottom 25 cold tables, do the following:
   a) Locate the Data Temperature chart.
   b) In the Hot Tables chart widget, hover over each table to view information about how often the table was accessed, its volume, and the number of partitions and files it contains.
   c) View more details about a table of interest, such as the hottest table, by clicking on the table's component element.

      The **Overview Details** side drawer panel opens, which enables you to view more information about the table, such as historical trends, column names, data types, and key-value pair properties. This information can be useful before you process or make changes to a query.
   d) Review the table's metadata from the Details, Columns, and Property tabs.
   e) Close the **Overview Details** side drawer panel and do the same steps in the Cold Tables chart widget.
4. To display the top 25 hot tables that were most frequently queried by either the Hive or Impala engine, do the following:
   a) From the cluster's ENGINES, select the Hive or Impala engine of interest.
   b) In the workload engine's Summary page, locate the Data Temperature chart.
   c) In the Hot Tables chart widget, hover over each table to view information about how often the table was accessed, its volume, and the number of partitions and files it contains.
   d) View more details about a table of interest, such as the hottest table, by clicking on the table's component element.

      The **Overview Details** side drawer panel opens.
   e) Review the table's metadata from the Details, Columns, and Property tabs.
5. As your tables and data increases it becomes more difficult for you keep track of your tables and their data, the HMS Tables view lists your tables and provides details about each table available in your system, regardless of whether they have been queried or not.

   To list the details about each table available in your system, do the following:
   a) Expand the HIVE METASTORE category for the cluster of interest.

      One or multiple metastores are displayed.
   b) Select the metastore of interest.

      The metastore's HMS Summary page opens displaying information about the current state and activity of all your tables in the selected Environment.
   c) To open the HMS Tables view, click the Tables tab.
   d) Locate specific tables of interest with the filter and sort functions. For example:

      - Sort the tables by their name or by a table's column value, such as the highest number of Partitions.
      - Reduce and locate tables by a specific value, such as filtering by their Table Type.
      - Locate the tables with a specific number of rows by selecting the Rows filter, entering the minimum and maximum row values that you require, and clicking Apply.
   e) Analyze the details of those tables of interest and look for inconsistencies or issues that may interfere with optimal query performance.
   f) View more details about a specific table by doing the following:

      1. Click the table's name.

         The **Overview Details** side drawer panel opens.
      2. Review the table's metadata from the Details, Columns, and Property tabs.
      3. Close the **Overview Details** side drawer panel and analyze another table.

Understanding Validations in Cloudera Observability

Learn about Cloudera Support’s Validations predictive alerting tool whose alerts are displayed in the Cloudera Observability web UI. The Validations feature identifies problems within your environment arising from your support bundles and automatically displays an alert when an issue appears and/or when conditions are not met, including details about the problem and recommended solutions that enable you to take corrective action before a more serious issue arises.

Important: The Validations feature is only available for CDP Private Cloud Base, CDP Data Hub, and Classic Cluster environments and requires that a diagnostic bundle has been sent to Cloudera Support, either as part of a Support case, through the scheduled support bundle delivery of diagnostic data, or through a phone home python script that is available on Cloudera.com.

Cloudera Observability provides access to the Cloudera Support's Validations predictive alerts through its Validations feature. The Validations predictive alerting tool uses predictive checks, known as validators, that are automatically run against every diagnostic bundle that is received from a supported CDH or CDP Private Cloud Base environment.

The validators analyze and identify known problem signatures contained in the diagnostic data of your bundles, such as the state and configuration settings of your cluster. There are over 400 validator problem signatures relating to misconfigurations, security vulnerabilities, performance degradation, and deviations from Cloudera's known best practices, as well as many other types of commonly experienced issues that can affect the stability, performance, security, and health of your environment.

The Cluster validation signatures also ensure that Cloudera Data Hub (CDH) clusters are optimized for a successful upgrade to Cloudera Data Platform (CDP). Where, configuration best practices that are required for a smooth upgrade are reviewed and an alert is raised accordingly on each cluster asset if it's out of alignment.

The following support bundles are validated:

- Diagnostic bundle, which is created by Cloudera Manager and contains all the configurations, logs, and details about your cluster and its services.
- Application bundle, which is created from your applications, such as the Spark application, a specific Hive query, or a specific item, such as a workload job execution.

Understanding the severity values

Each validation error includes associated severity levels, based on the impact to the cluster:

- Critical, which indicates a serious problem that must be resolved immediately.
- Error, which indicates incorrect settings and configurations that require attention.
- Warn, which indicates a potential problem that eventually must be resolved but does not have to be completed at this time.
- Info, which is displayed for informational purposes only, such as performance improvements. For example, to handle all services, configure your log and query redactions in Cloudera Manager rather than in HDFS.
- Curiosity, which flags unusual deployments and settings. For example, Cloudera is curious as to why this type of file system is used.
- Pass, which indicates a normal result and within the acceptable range.
- Insufficient_information, which indicates that there is not enough information at this time.

Based on the Validator's threshold severity values, the most serious alert issues are displayed in Cloudera Observability, including details about the identified error, its cause, severity level, affected hosts, the corrective actions you should consider to resolve the problem, and links to applicable documentation.

Note: Customer Support also has access to your validation alerts through Expedited Support. If you are having problems in resolving an alert, create a Support ticket for Cloudera Support's help. You can also review and manage all your validation alerts in **MyCloudera** under **Assets**.

Considerations and limitations

The following describes considerations and limitations that you must know when using the Validations feature:

- At this time your validations critical alerts are only available for CDP Data Hub, CDP Private Cloud Base, and Classic Cluster environments.

- The Validations feature requires that a support bundle has been sent to Cloudera Support through one of the following methods:
    - Directly, as a scheduled diagnostic delivery of diagnostic data from CDP Private Cloud Base with Cloudera Manager.
    - Attachment, as a support bundle attachment in a Support case.
    - Manually, through a phone home python script that is available on Cloudera.com.

  For more information, click the Related Information links below.
- New and updated validator signatures are continuously added to the Validations feature by Cloudera Support. To have the latest signatures run on your bundles, you must regularly send your support bundles to Cloudera.

Related Information

Sending Usage and Diagnostic Data to Cloudera

phone home python script

Validations – Cloudera Support’s Predictive Alerting Program

Working with Validations

Learn how to use the Validations predictive alerting tool feature in Cloudera Observability that displays details about the issue, it's severity level, the affected hosts, and the corrective actions you should consider to resolve the problem.

About this task

Describes where to view the Validations alerts in Cloudera Observability that help you fix problems within your environment's cluster and engines.

Important: The Validations feature is only available for CDP Private Cloud Base, CDP Data Hub, and Classic Cluster environments and requires that a diagnostic bundle has been sent to Cloudera Support, either as part of a Support case, through the scheduled support bundle delivery of diagnostic data, or through a phone home python script that is available on Cloudera.com.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.
    a) In a supported browser, log into the Cloudera Data Platform (CDP).

       The CDP web interface landing page opens.

    b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

       The Cloudera Observability landing page opens to the main navigation panel.

    c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

       The Environment navigation panel opens.

2. To view all open validation alerts for your environment's cluster, do the following:
   a) Depending on the environment selected, verify that the Validations page is displayed for the environment's cluster required for analysis.

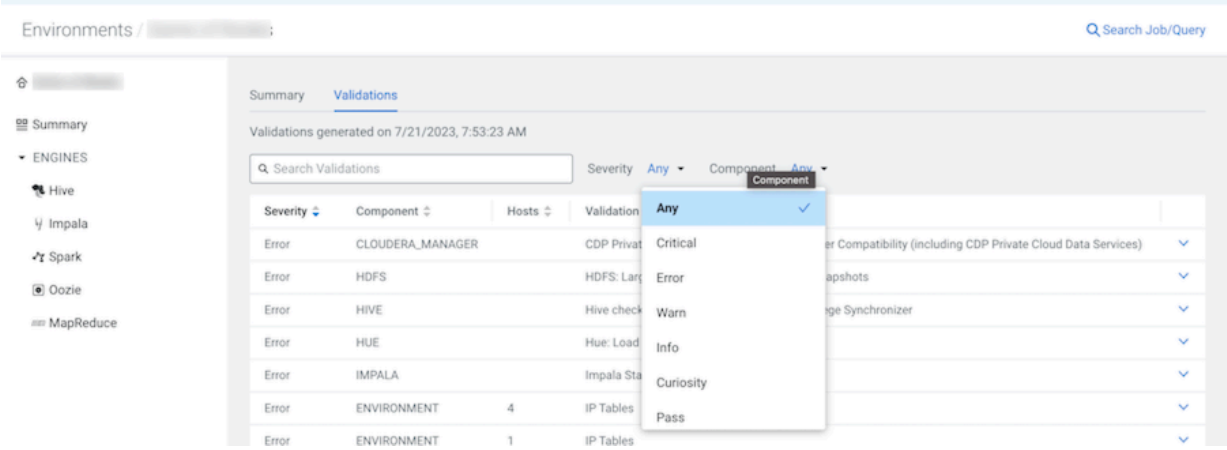
      To display the Validations tab, do one of the following:

      - For a Private Cloud Base and a Classic Cluster environment type, select the Validations tab in the Cluster Summary page.
      - For a Data Hub environment type, expand its Data Lake category in the Environment's panel, locate and select the Data Hub cluster, and then select the Validations tab in the Cluster Summary page.

      The Validations page opens, which displays the current alerts for the cluster that requires attention.
   b) Sort your alerts from any column, such as by the number of hosts affected, the severity level, and/or by the component that is affected.
   c) Filter the alerts displayed by doing one or more of the following:

      - Select a severity alert level from the **Severity** filter list, such as Critical.
      - Select a specific component from the **Component** list, such as HBASE, HDFS, or HIVE.
   d) Locate the alert of interest, expand and display the full details about the alert, and then follow the recommended course of action to fix the issue.

   The following example, shows the Validations page and the Severity list for an environments cluster.



Environments / [redacted]

Validations generated on 7/21/2023, 7:53:23 AM

Severity Any Component Any

Severity	Component	Hosts	Validation	
Error	CLOUDERA_MANAGER		CDP Privat	er Compatibility (including CDP Private Cloud Data Services)
Error	HDFS		HDFS: Larg	apshots
Error	HIVE		Hive check	ge Synchronizer
Error	HUE		Hue: Load	
Error	IMPALA		Impala Sta	
Error	ENVIRONMENT	4	IP Tables	
Error	ENVIRONMENT	1	IP Tables	

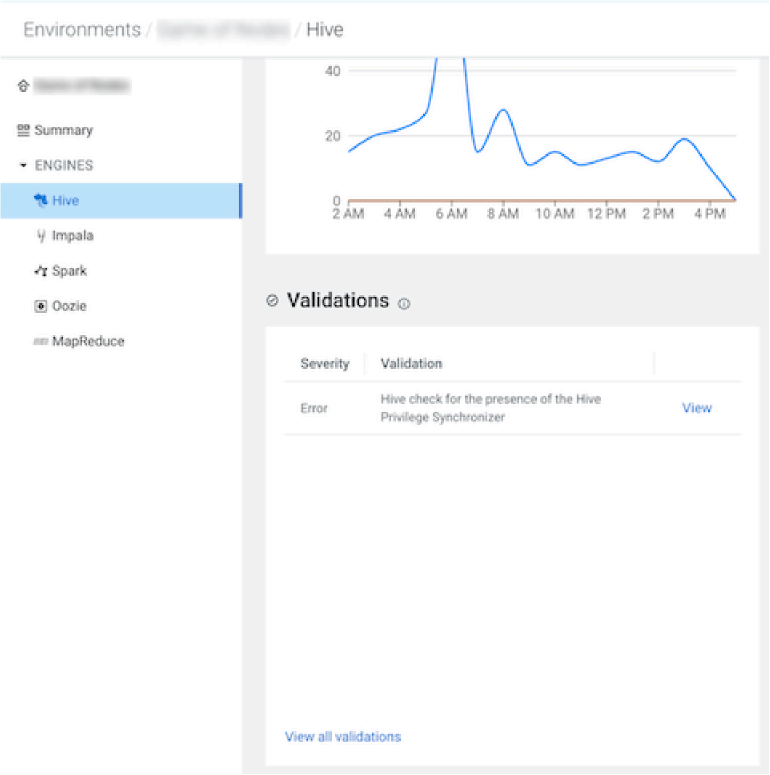
3. To view the open validation alerts for an engine in the environment's cluster, do the following:
   a) Verify that the engine's Summary page is displayed for the environment's cluster.

      To display the engine's Summary page, do one of the following:

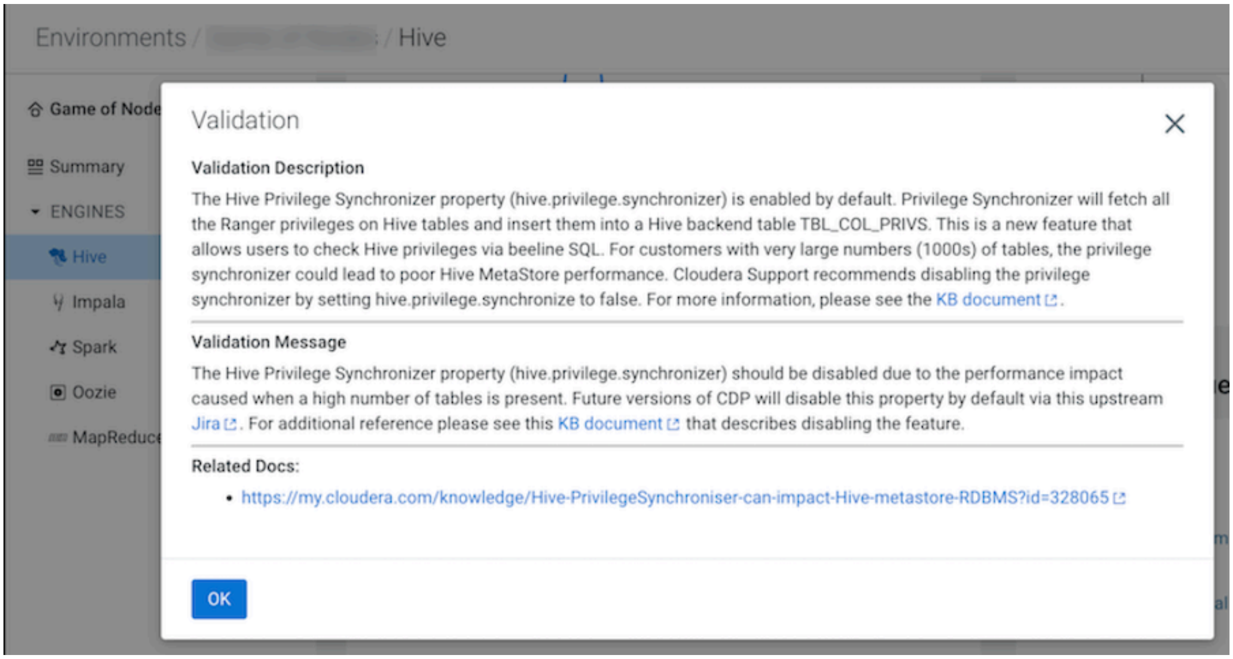
      - For Private Cloud Base and Classic Cluster environment type, locate the cluster's ENGINES category and select a workload engine of interest.
      - For a Data Hub environment type, expand its Data Lake category in the Environment's panel, locate and select the Data Hub cluster, and then from the ENGINES category, select a workload engine of interest.

      The engine's Summary page opens, which displays a series of chart widgets that display metrics about the workload jobs run by the selected engine.
   b) Scroll down and locate the **Validations** chart widget, which displays the current alerts for the engine that requires attention.
   c) Locate the alert of interest and then display the full details about the alert in a new dialog box, by clicking View.
   d) Follow the recommended course of action that will help you fix the issue.

   The following examples show the Validations chart widget and an Error severity Validation alert for a Hive engine, which indicates incorrect settings and configurations that require attention. The Validation message explains why the Hive Privilege Synchronizer property should be disabled and provides a link on how to disable.



Severity	Validation	
Error	Hive check for the presence of the Hive Privilege Synchronizer	View



Validation

Validation Description

The Hive Privilege Synchronizer property (hive.privilege.synchronizer) is enabled by default. Privilege Synchronizer will fetch all the Ranger privileges on Hive tables and insert them into a Hive backend table TBL_COL_PRIVS. This is a new feature that allows users to check Hive privileges via beeline SQL. For customers with very large numbers (1000s) of tables, the privilege synchronizer could lead to poor Hive MetaStore performance. Cloudera Support recommends disabling the privilege synchronizer by setting hive.privilege.synchronize to false. For more information, please see the KB document.

Validation Message

The Hive Privilege Synchronizer property (hive.privilege.synchronizer) should be disabled due to the performance impact caused when a high number of tables is present. Future versions of CDP will disable this property by default via this upstream Jira. For additional reference please see this KB document that describes disabling the feature.

Related Docs:

- https://my.cloudera.com/knowledge/Hive-PrivilegeSynchroniser-can-impact-Hive-metastore-RDBMS?id=328065

OK

4. To view the open validation alerts specifically for a Hive query, do the following:

a) Verify that the environment's Cluster Summary page is displayed for the Cloudera Observability environment.

To display the Cluster Summary page, do one of the following:

- For Private Cloud Base and Classic Cluster environment type, verify that the Cluster Summary title is displayed in the browser tab.
- For a Data Hub environment type, expand its Data Lake category in the Environment's panel and then locate and select the Data Hub cluster of interest.

The Summary page opens, which displays a series of performance trends and metric chart widgets about the processed jobs and queries.

b) From the Hive Query Trend chart widget, click its Total Queries value.

c) From the Job column in the Queries page, locate and click the query of interest.

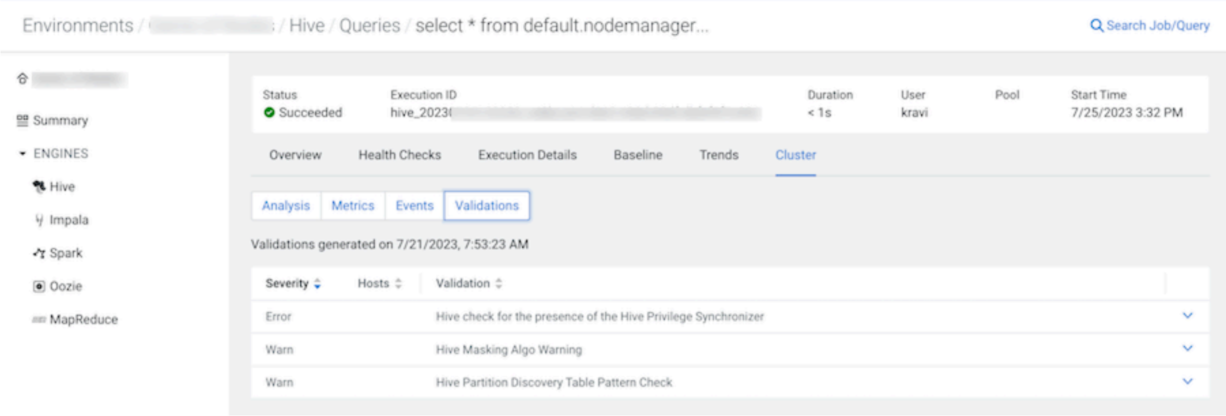
d) From the query's page, select the Cluster tab and then select the Validations tab.

The Validations page opens, which displays the current alerts for the Hive query that require attention.

e) Locate the alert of interest, expand and display the full details about the alert, and then follow the recommended course of action to fix the issue.

The following examples show the **Cluster** tab's Validations page and several error and warning Validation alerts. Where an Error severity indicates incorrect settings and configurations that require attention and a Warn severity indicates a potential problem that eventually must be resolved but does not have to be completed at this time. The

Validation message for the Error alert explains why the Hive Privilege Synchronizer property should be disabled and provides a link on how to disable.



Environments / [redacted] / Hive / Queries / select * from default.nodemanager...

Search Job/Query

[redacted]

Summary

ENGINES

Hive

Impala

Spark

Oozie

MapReduce

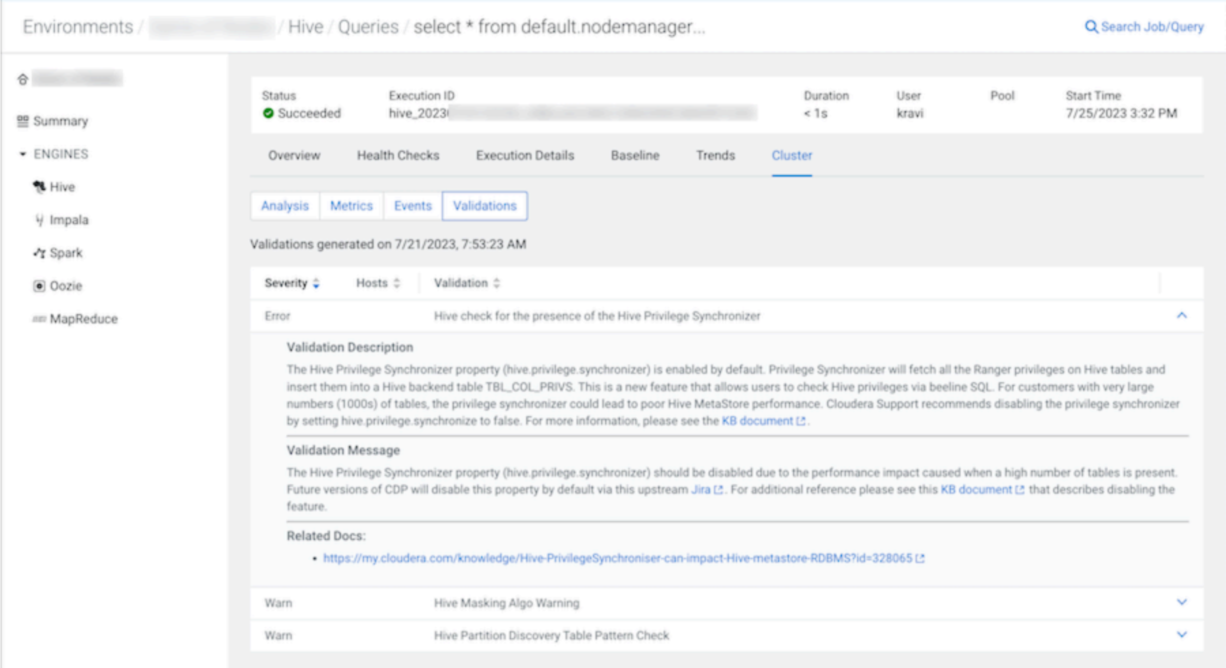
Status	Execution ID	Duration	User	Pool	Start Time
Succeeded	hive_2023[redacted]	< 1s	kravi		7/25/2023 3:32 PM

Overview · Health Checks · Execution Details · Baseline · Trends · Cluster

Analysis · Metrics · Events · Validations

Validations generated on 7/21/2023, 7:53:23 AM

Severity	Hosts	Validation
Error		Hive check for the presence of the Hive Privilege Synchronizer
Warn		Hive Masking Algo Warning
Warn		Hive Partition Discovery Table Pattern Check



Environments / [redacted] / Hive / Queries / select * from default.nodemanager...

Search Job/Query

[redacted]

Summary

ENGINES

Hive

Impala

Spark

Oozie

MapReduce

Status	Execution ID	Duration	User	Pool	Start Time
Succeeded	hive_2023[redacted]	< 1s	kravi		7/25/2023 3:32 PM

Overview · Health Checks · Execution Details · Baseline · Trends · Cluster

Analysis · Metrics · Events · Validations

Validations generated on 7/21/2023, 7:53:23 AM

Severity	Hosts	Validation
Error		Hive check for the presence of the Hive Privilege Synchronizer

Validation Description

The Hive Privilege Synchronizer property (hive.privilege.synchronizer) is enabled by default. Privilege Synchronizer will fetch all the Ranger privileges on Hive tables and insert them into a Hive backend table TBL_COL_PRIVS. This is a new feature that allows users to check Hive privileges via beeline SQL. For customers with very large numbers (1000s) of tables, the privilege synchronizer could lead to poor Hive MetaStore performance. Cloudera Support recommends disabling the privilege synchronizer by setting hive.privilege.synchronize to false. For more information, please see the KB document.

Validation Message

The Hive Privilege Synchronizer property (hive.privilege.synchronizer) should be disabled due to the performance impact caused when a high number of tables is present. Future versions of CDP will disable this property by default via this upstream Jira. For additional reference please see this KB document that describes disabling the feature.

Related Docs:

- https://my.cloudera.com/knowledge/Hive-PrivilegeSynchroniser-can-impact-Hive-metastore-RDBMS?id=328065

Severity	Hosts	Validation
Warn		Hive Masking Algo Warning
Warn		Hive Partition Discovery Table Pattern Check

Analyzing your Hive queries for debugging and optimization

Identify operational, performance, and health issues of your Hive workloads, queries, and cluster. The following topics, guide you through the Cloudera Observability Hive features that enable you to identify and troubleshoot performance and health issues.

Note: If you do not see any of these features and/or metrics, verify that Cloudera Manager has been upgraded to the latest version and that Telemetry Publisher was restarted.

Identifying inefficient phases of your Hive queries

Identify inefficient phases of your Hive queries for query optimization and performance tuning, such as viewing the execution phases, the order in which the operations are executed, comparing two execution plans, and validating the events performed.

About this task

Describes how to locate the Cloudera Observability Hive SQL Query Plan and DAG, the Hive Query Plan Graph, and the Counters and Configuration panels for identifying and troubleshooting inefficient operational phases of your Hive queries.

Note: The Query Plan and the Query Plan Graph are only available for Hive.

Note: If you do not see any of these features and/or metrics, verify that Cloudera Manager has been upgraded to the latest version and that Telemetry Publisher was restarted.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.

   a) In a supported browser, log into the Cloudera Data Platform (CDP).

      The CDP web interface landing page opens.

   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

      The Cloudera Observability landing page opens to the main navigation panel.

   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

      The Environment navigation panel opens.

2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required for analysis.

   To display the Cluster Summary page, do one of the following:

   - For Private Cloud Base and Classic Cluster environment type, verify that the Cluster Summary title is displayed in the browser tab.
   - For Data Hub and Database Catalog environment types, expand its Data Lake category in the Environment's panel and then locate and select the cluster or Virtual Warehouse of interest.

   The Cluster Summary page displays a series of performance trends and metric chart widgets about the processed jobs and queries.

   Tip: The page's title is displayed in the browser tab.

3. From the time-range list in the Cluster Summary page, select a time period that meets your requirements.
4. Locate the Hive query of interest by doing one of the following:

   - In the Cluster Summary page, locate the Hive Query Trend chart widget, click its Total Queries value, and then from the Job column in the Queries page, locate and click the query of interest.
   - If not already expanded, from the Environment navigation panel, expand the ENGINES or Virtual Warehouse category and then select the **Hive** engine. Locate the Slow Queries chart widget and click the query of interest.

5. From the .../Hive/Queries/ *queryname* page, select the Execution Details tab.

   The execution stages appear, displaying the Query Details Summary panel.

6. To review the query's execution instructions and logical steps, do the following:

   a) In the stages column, verify that the query and not the DAG is selected.
   b) From the Query Details Summary panel, click Text.

   The query's Query Plan panel opens.

   The query plan displays the execution statistics in a list of execution stages that include the execution instructions and steps, such as the operations that are performed, the operators that are used, the resources that are allocated, and the stage dependencies. These stages can help you diagnose and improve a query's performance.

   Note: You can save the Query Plan as a JSON file to your computer by clicking the download icon ⤓.

7. To visually display a graphical representation of the Query Plan's DAG, which contains individual components that represent each event and the order they are executed, do the following:

   a) In the stages column, verify that the query and not the DAG is selected.
   b) From the Query Details Summary panel, click Graphical.

   The query's Query Plan Graph page opens in another tab of your browser.

   The Query Plan Graph displays the order of events and the steps and phases of the query. This page enables you to visually inspect where each operation is executed and if the order is the most efficient. For example, an operator that could be draining your CPU and memory because it is joining tables that contain a large number of records before a filtering operation was performed.

   You can also view an operator's location within the Query Plan and the operator's execution details by doing the following:

   - To display where the operator is located in the Query Plan, hover over an event box.
   - To display, in the right-side panel, the operator's execution details from the Query Plan, click on an event box.

8. To identify and validate which tasks completed or are taking too long to run, do the following:

   a) In the stages column, click the dag_*xxx* link. Where, *xxx* is the DAG ID number.
   b) From the Dag Details Summary panel, click Counters.

   The Counters panel opens.

   This panel lists in detail the events performed and the total number of occurrences, which enable you to track, compare, and validate the events that were run for the query. For example, you can verify that the correct number of tasks were run and completed, that the number of records, rows, and the amount of bytes were read and written, and that the correct amount of CPU and memory was consumed for the query.

9. To verify that the query's configuration settings align with your expectations, do one or more of the following:

   - To understand the query's execution configuration setting details:

     a. In the stages column, click the query link.
     b. From the Query Details Summary panel, click Configurations.
   - To understand the query's DAG execution configuration setting details:

     a. In the stages column, click the dag_*xxx* link. Where, *xxx* is the DAG ID number.
     b. From the Dag Details Summary panel, click Configurations.

10. To troubleshoot performance-related issues between two different runs of the same query, do the following:
    a) From the query's page, select the Trends tab.
    b) Scroll down and from the table, select the check boxes adjacent to the query's job runs that you require, such as the latest run with a run from a week ago, and then click Compare.

       The Job Comparison page opens displaying more details about each job.
    c) From the Details section, select the Query Plan tab.

    You can view and analyze the selected query plans Side By Side or as a Unified plan that highlights the differences in color, which enables you to quickly identify what changed between the selected execution runs of the query.

    The Job Comparison page not only enables you to compare the query plans but also the following:

    - The Duration, Data Input, and Data Output of the selected job runs from the Performance section.
    - Their run-times by selecting the Structure tab.
    - Configuration differences by selecting the Configurations tab.
    - Statistical differences by selecting the Metrics tab.

About the Cloudera Observability Hive cluster service metrics

The Cloudera Observability Hive cluster service metrics are displayed as graphical reports that show the state, activity, and performance of your workload Hive service, including recommendations on how to resolve a problem. They help you monitor the health, performance, and workload usage of your Hive service for identifying and troubleshooting existing and potential problems.

Cloudera Observability collects diagnostic data from a series of health checks that are performed on your Hive service. When completed they are compared to their defined thresholds that determine if the service is Good, Concerning, or Bad and the results are displayed on the **Analysis** page, which is accessed from the Hive engine's Queries page.

Cloudera Observability helps you distinguish between a healthy and an unhealthy state by including a severity level icon adjacent to the health test using the following colors:

Table 12: Severity Colors

Severity Color	Description
Green	Good- The health check result is normal and within the acceptable range.
Yellow	Concerning- The health check result has exceeded the Warning threshold limit and indicates a potential problem, which eventually must be resolved but does not have to be completed at this time. See the Recommendation actions.
Red	Bad- The health check result has exceeded the Critical threshold limit and indicates a serious problem, which must be resolved immediately. See the Recommendation actions.

For descriptions of the Hive cluster health checks performed by Cloudera Observability, the severity conditions and thresholds, and what actions you should consider to resolve a problem, click the Related Information link below.

You can also manually build a Hive service chart in the **Metrics** page, without having to leave Cloudera Observability, using the Cloudera Manager service metrics and Chart Builder.

Note: This feature is intended for Advanced Hive Users and requires knowledge of the Cloudera Manager service metrics and the Cloudera Manager’s Chart Builder.

For more information about the Cloudera Manager health checks performed on the Hive service, click the Related Information link below.

Related Information

Cloudera Observability Hive Cluster Metrics

Cloudera Manager Metrics

Monitoring your Hive service

Identify Hive service problems that may be affecting your Hive workloads, such as queries that are running slow or that are failing, with the Cloudera Observability Hive cluster service metrics.

About this task

Describes where to view the Cloudera Observability Hive cluster service metrics and how to build your own service chart from the Cloudera Manager service metrics and Chart Builder.

Note:

- The Cloudera Observability Hive cluster service metrics are not available for a Virtual Warehouse environment.
- If you do not see any of these features and/or metrics, verify that Cloudera Manager has been upgraded to the latest version and that Telemetry Publisher was restarted.

Procedure

1. Verify that you are logged in to the Cloudera Observability web UI and that you selected an environment from the **Analytics Environments** page.

   a) In a supported browser, log into the Cloudera Data Platform (CDP).

   The CDP web interface landing page opens.

   b) From the Your Enterprise Data Cloud landing page, select the Observability tile.

   The Cloudera Observability landing page opens to the main navigation panel.

   c) From the Cloudera Observability **Environments** page, select the environment required for analysis.

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

   The Environment navigation panel opens.
2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required for analysis.

   To display the Cluster Summary page, do one of the following:

   - For Private Cloud Base and Classic Cluster environment type, verify that the Cluster Summary title is displayed in the browser tab.
   - For a Data Hub environment type, expand its Data Lake category in the Environment's panel and then locate and select the cluster of interest.

   The Cluster Summary page displays a series of performance trends and metric chart widgets about the processed jobs and queries.

   Tip: The page's title is displayed in the browser tab.
3. From the time-range list in the Cluster Summary page, select a time period that meets your requirements.
4. Locate the Hive query of interest by doing one of the following:

   - In the Cluster Summary page, locate the Hive Query Trend chart widget, click its Total Queries value, and then from the Job column in the Queries page, locate and click the query of interest.
   - If not already expanded, from the Environment navigation panel, expand the ENGINES category and then select the **Hive** engine. Locate the Slow Queries chart widget and click the query of interest.
5. From the .../Hive/Queries/ *queryname* page, select the Cluster tab.

   The Analysis page opens, which lists the Hive cluster health check metrics that are performed by Cloudera Observability at the end of a Hive job.

6. Select, either the metric you require for analysis or a metric that displays a red or yellow icon adjacent to the metric, which represents the threshold warning or error state for at least one unhealthy role instance.

   The Analysis summary page opens, which describes the health check metric performed by Cloudera Observability on the Hive service, the severity conditions and thresholds, and the remediation actions you should consider to resolve a problem. It also contains:

   - The Analysis chart, which displays the severity condition of the operation during the job run and displays the state of each role instance 5 minutes before the start of the job and 5 minutes after the job has completed.
   - The Host Status section, which displays the full list of workload role instances and the hosts they are running on, their health check result, and the severity state icon.
7. To build your own chart from a Cloudera Manager health check service metric, click the Metrics option and do the following:

   a. From the Service Name list, select a service that you are running on your workloads cluster.
   b. From the Metric Name list, select the name of the Cloudera Manager health test metric.
   c. Click View.

Troubleshooting an abnormal job duration

You can identify areas of risk from jobs running on your workload cluster that complete within an unusual time period, using Cloudera Observability.

About this task

Describes how to locate and troubleshoot an abnormal job duration.

Steps with examples from a Virtual Cluster's Spark engine are used to explain how to further investigate and troubleshoot the cause of an abnormal job duration.

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 Environment Name column in the Environments page, locate and click the name of the environment whose workload diagnostic information requires analysis and troubleshooting.

   For this example, select **Virtual Cluster** from the Environments list and then select a Virtual Cluster required for analysis.

   The Environment navigation panel opens, which hierarchically lists the environment and its services hosted on the selected environment.
3. Verify that the **Cluster Summary** page is displayed.

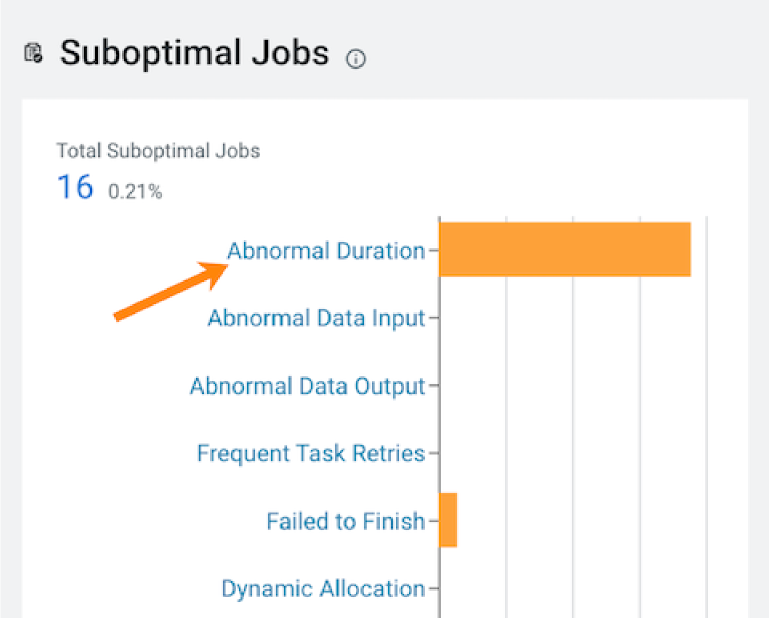
   Tip: The page's title is displayed in the browser tab.

   The **Cluster Summary** page, displays performance trends and metrics about the cluster's processed jobs and queries.
4. From the time-range list, select a time period that meets your requirements.
5. If not already expanded, from the Environment navigation panel expand the Virtual Cluster and then select the **Spark** engine.

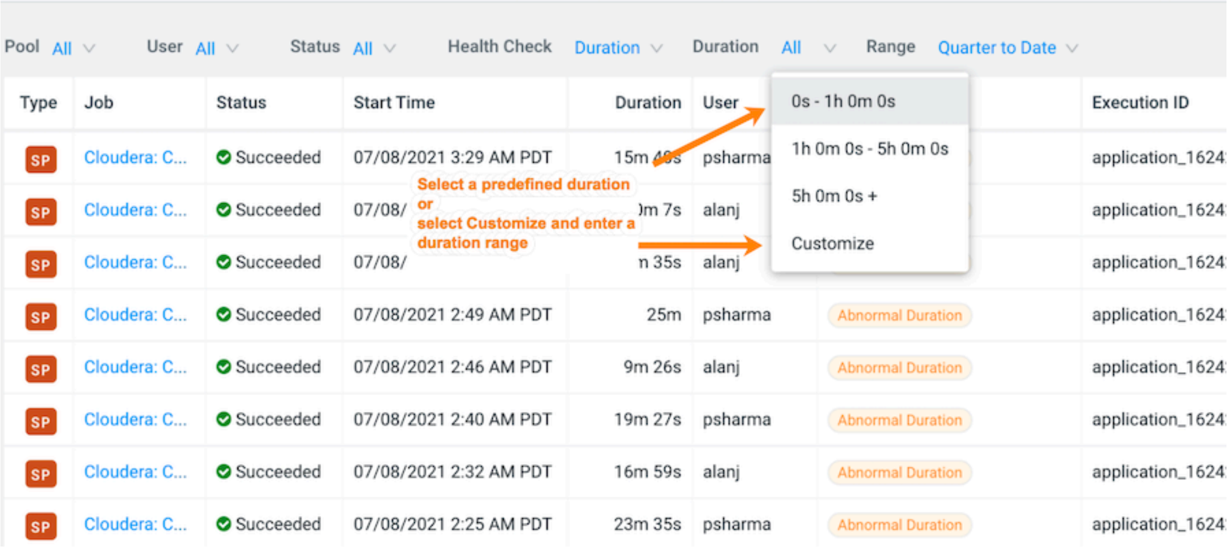
6. Scroll down to the Suboptimal Jobs chart widget and click the Abnormal Duration health check bar.

   The Jobs page opens, listing all the jobs that triggered the Abnormal Duration Health check during the time period, including their health status, the length of time the job took to run, the user who ran the job, the job identification number, and the amount of CPU used to run the job.

   Tip: Any jobs or queries that fall outside of their baseline are counted. You can hover over each bar within the chart to view how many jobs or queries triggered each health check.



7. Specify a specific amount of time in which the job either ran less than or more than the Health check rule by either selecting a predefined time duration or selecting Customize and enter the minimum or maximum time period from the Duration filter.


Pool All User All Status All Health Check Duration Duration All Range Quarter to Date

Type	Job	Status	Start Time	Duration	User		Execution ID
SP	Cloudera: C...	Succeeded	07/08/2021 3:29 AM PDT	15m 40s	psharma		application_1624
SP	Cloudera: C...	Succeeded	07/08/	m 7s	alanj		application_1624
SP	Cloudera: C...	Succeeded	07/08/	n 35s	alanj		application_1624
SP	Cloudera: C...	Succeeded	07/08/2021 2:49 AM PDT	25m	psharma	Abnormal Duration	application_1624
SP	Cloudera: C...	Succeeded	07/08/2021 2:46 AM PDT	9m 26s	alanj	Abnormal Duration	application_1624
SP	Cloudera: C...	Succeeded	07/08/2021 2:40 AM PDT	19m 27s	psharma	Abnormal Duration	application_1624
SP	Cloudera: C...	Succeeded	07/08/2021 2:32 AM PDT	16m 59s	alanj	Abnormal Duration	application_1624
SP	Cloudera: C...	Succeeded	07/08/2021 2:25 AM PDT	23m 35s	psharma	Abnormal Duration	application_1624

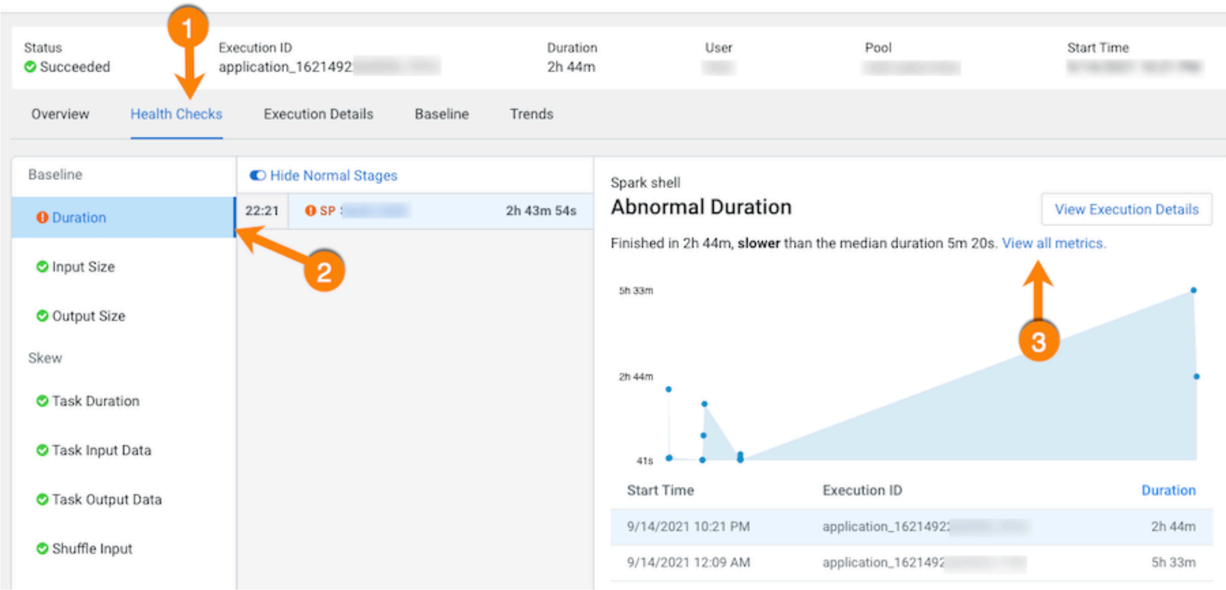
Select a predefined duration or select Customize and enter a duration range

8. View more details about a job by selecting a job's name from the Job column and then clicking the Health Checks tab.

   The Baseline and Skew Health checks are displayed.

9. Display more information about the job's duration by selecting Duration from the Baseline section. As shown in the image below.

In the following example, the job finished much slower than the baseline duration, which is the aggregate calculated over multiple jobs.



Status	Execution ID	Duration	User	Pool	Start Time
Succeeded	application_1621492[redacted]	2h 44m	[redacted]	[redacted]	[redacted]

22:21 SP [redacted] 2h 43m 54s

Spark shell

Abnormal Duration

Finished in 2h 44m, **slower** than the median duration 5m 20s. View all metrics.

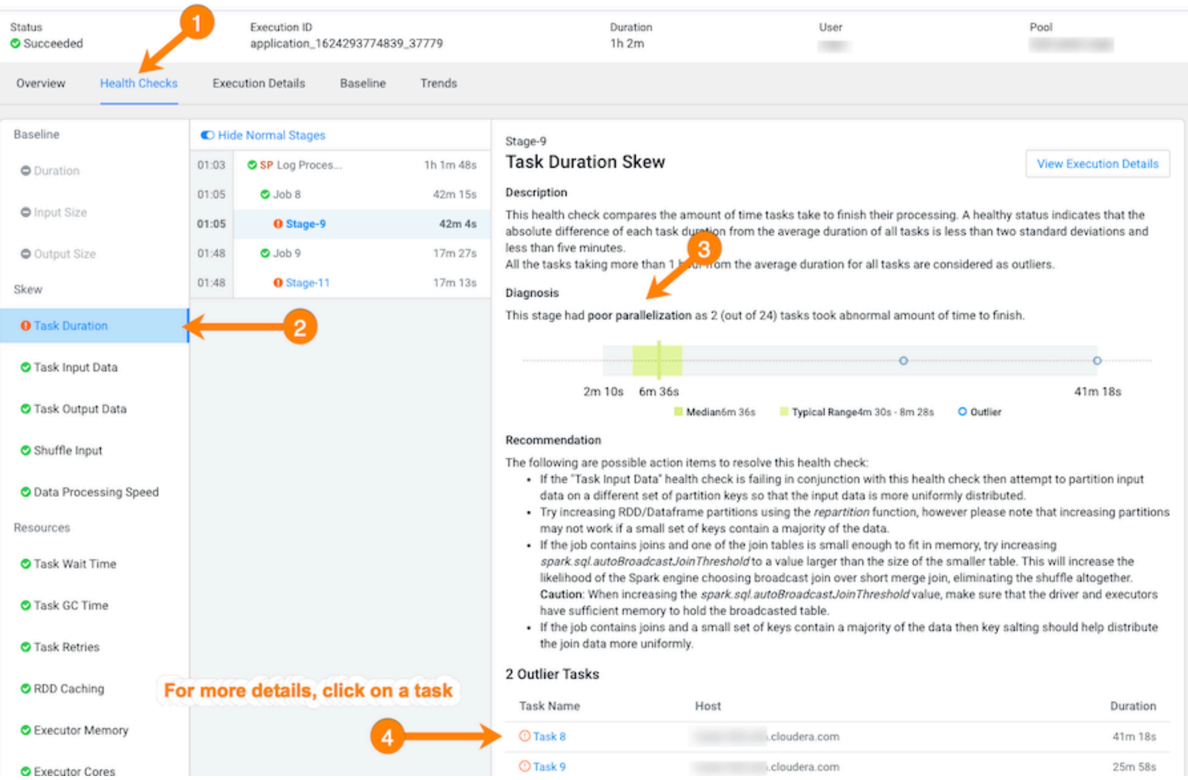
Start Time	Execution ID	Duration
9/14/2021 10:21 PM	application_1621492[redacted]	2h 44m
9/14/2021 12:09 AM	application_1621492[redacted]	5h 33m

10. Compare and analyze this job against other baseline metrics by clicking View all metrics.

11. Continue to analyze and search for probable causes by doing one or more of the following:

- To display more information about the length of time the processing tasks took within a job, select Task Duration, which opens a panel that describes the health check, displays information about the possible causes, and lists recommended solutions.

  In the following example, issues arose during Stage-9 of the job due to poor parallelization. The Recommendation section lists items for you to complete that may resolve the problem and the specific outlier tasks that produced the unusual results:



Status
Succeeded

Execution ID
application_1624293774839_37779

Duration
1h 2m

User
[redacted]

Pool
[redacted]

01:03	SP Log Proces...	1h 1m 48s
01:05	Job 8	42m 15s
01:05	Stage-9	42m 4s
01:48	Job 9	17m 27s
01:48	Stage-11	17m 13s

Stage-9

Task Duration Skew

Description

This health check compares the amount of time tasks take to finish their processing. A healthy status indicates that the absolute difference of each task duration from the average duration of all tasks is less than two standard deviations and less than five minutes.
All the tasks taking more than 1 hour from the average duration for all tasks are considered as outliers.

Diagnosis

This stage had **poor parallelization** as 2 (out of 24) tasks took abnormal amount of time to finish.

Recommendation

The following are possible action items to resolve this health check:

- If the "Task Input Data" health check is failing in conjunction with this health check then attempt to partition input data on a different set of partition keys so that the input data is more uniformly distributed.
- Try increasing RDD/Dataframe partitions using the *repartition* function, however please note that increasing partitions may not work if a small set of keys contain a majority of the data.
- If the job contains joins and one of the join tables is small enough to fit in memory, try increasing *spark.sql.autoBroadcastJoinThreshold* to a value larger than the size of the smaller table. This will increase the likelihood of the Spark engine choosing broadcast join over short merge join, eliminating the shuffle altogether. **Caution**: When increasing the *spark.sql.autoBroadcastJoinThreshold* value, make sure that the driver and executors have sufficient memory to hold the broadcasted table.
- If the job contains joins and a small set of keys contain a majority of the data then key salting should help distribute the join data more uniformly.

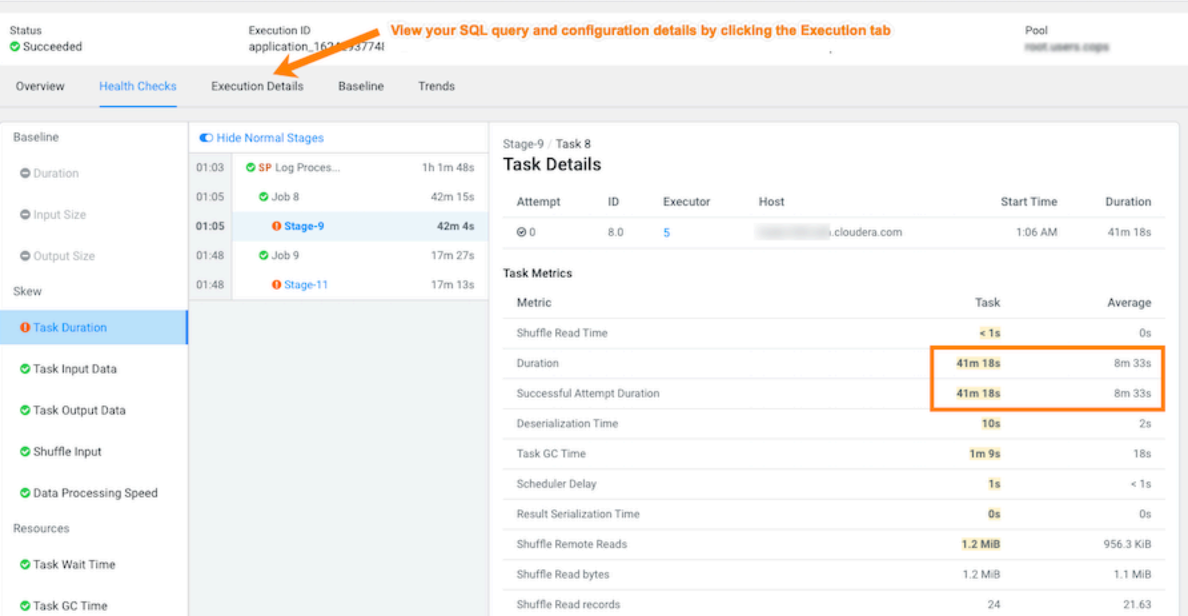
2 Outlier Tasks

Task Name	Host	Duration
Task 8	[redacted].cloudera.com	41m 18s
Task 9	[redacted].cloudera.com	25m 58s

For more details, click on a task

- To display more details about an outlier, click the outlier task, which opens the Task Details panel.

  In the following example, the Task Details show that the outlier task took significantly more time to complete compared to previous runs. In this case, 41 minutes as compared to 8 minutes:



View your SQL query and configuration details by clicking the Execution tab

Status
Succeeded

Execution ID
application_1624293774

Pool
[redacted]

01:03	SP Log Proces...	1h 1m 48s
01:05	Job 8	42m 15s
01:05	Stage-9	42m 4s
01:48	Job 9	17m 27s
01:48	Stage-11	17m 13s

Stage-9 / Task 8

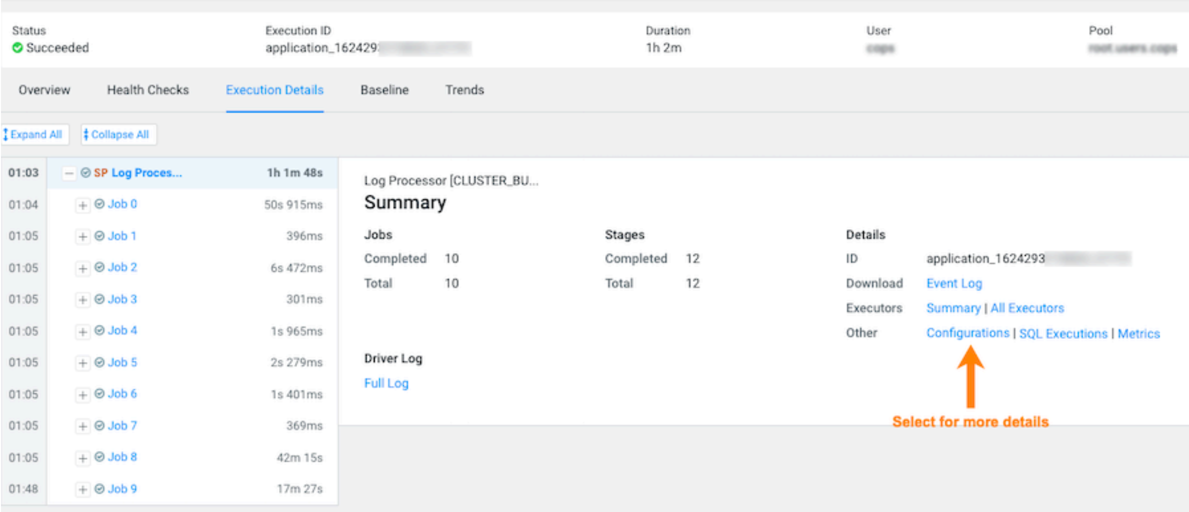
Task Details

Attempt	ID	Executor	Host	Start Time	Duration
0	8.0	5	[redacted].cloudera.com	1:06 AM	41m 18s

Task Metrics

Metric	Task	Average
Shuffle Read Time	< 1s	0s
Duration	41m 18s	8m 33s
Successful Attempt Duration	41m 18s	8m 33s
Deserialization Time	10s	2s
Task GC Time	1m 9s	18s
Scheduler Delay	1s	< 1s
Result Serialization Time	0s	0s
Shuffle Remote Reads	1.2 MiB	956.3 KiB
Shuffle Read bytes	1.2 MiB	1.1 MiB
Shuffle Read records	24	21.63

- To gain more insights about the task's duration, such as checking memory allocation, click the Execution Details tab and then in the Summary panel, click Configurations:



Status: Succeeded | Execution ID: application_162429[redacted] | Duration: 1h 2m | User: [redacted] | Pool: [redacted]

01:03	SP Log Proces...	1h 1m 48s
01:04	Job 0	50s 915ms
01:05	Job 1	396ms
01:05	Job 2	6s 472ms
01:05	Job 3	301ms
01:05	Job 4	1s 965ms
01:05	Job 5	2s 279ms
01:05	Job 6	1s 401ms
01:05	Job 7	369ms
01:05	Job 8	42m 15s
01:48	Job 9	17m 27s

Log Processor [CLUSTER_BU...

Summary

Jobs: Completed 10, Total 10

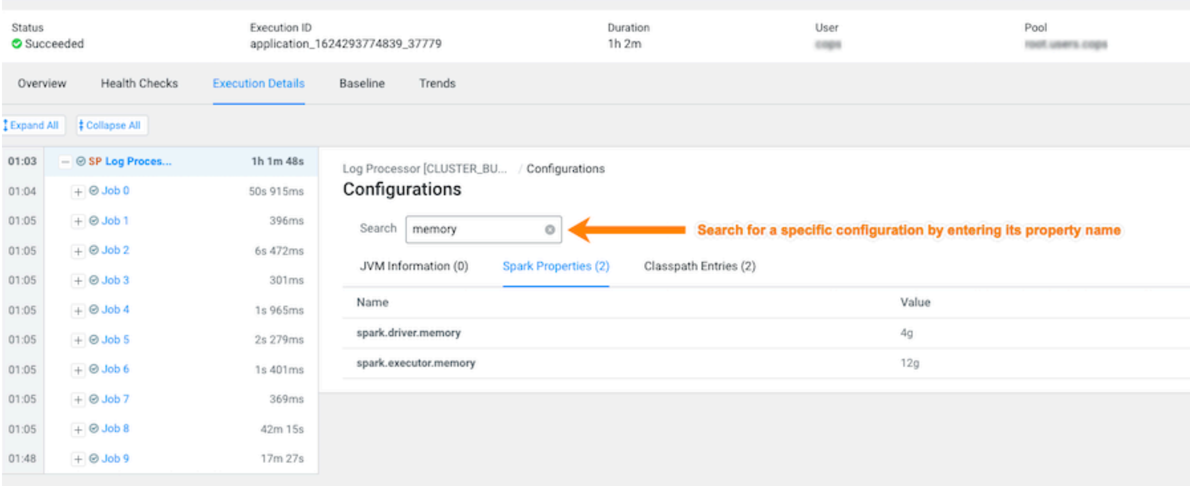
Stages: Completed 12, Total 12

Details: ID application_1624293[redacted]; Download Event Log; Executors Summary | All Executors; Other Configurations | SQL Executions | Metrics

Driver Log: Full Log

Select for more details

- In the Configurations panel, click the Spark Properties tab and search for the memory configuration settings and their values. If memory is less than the recommended value, increasing its value will improve cluster performance:



Status: Succeeded | Execution ID: application_1624293774839_37779 | Duration: 1h 2m | User: [redacted] | Pool: [redacted]

01:03	SP Log Proces...	1h 1m 48s
01:04	Job 0	50s 915ms
01:05	Job 1	396ms
01:05	Job 2	6s 472ms
01:05	Job 3	301ms
01:05	Job 4	1s 965ms
01:05	Job 5	2s 279ms
01:05	Job 6	1s 401ms
01:05	Job 7	369ms
01:05	Job 8	42m 15s
01:48	Job 9	17m 27s

Log Processor [CLUSTER_BU... / Configurations

Configurations

Search: memory

Search for a specific configuration by entering its property name

Name	Value
spark.driver.memory	4g
spark.executor.memory	12g

Troubleshooting failed jobs

You can identify and troubleshoot incomplete jobs on your cluster using Cloudera Observability.

About this task

Describes how to locate and troubleshoot jobs that have failed to complete.

Steps with examples from a Virtual Cluster's Spark engine are used to describe how to further investigate and troubleshoot the root cause of a job that failed to finish.

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 Environment Name column in the Environments page, locate and click the name of the environment whose workload diagnostic information requires analysis and troubleshooting.

   For this example, select **Virtual Cluster** from the Environments list and then select a Virtual Cluster required for analysis.

   The Environment navigation panel opens, which hierarchically lists the environment and its services hosted on the selected environment.

3. Verify that the **Cluster Summary** page is displayed.

   Tip: The page's title is displayed in the browser tab.

   The **Cluster Summary** page, displays performance trends and metrics about the cluster's processed jobs and queries.

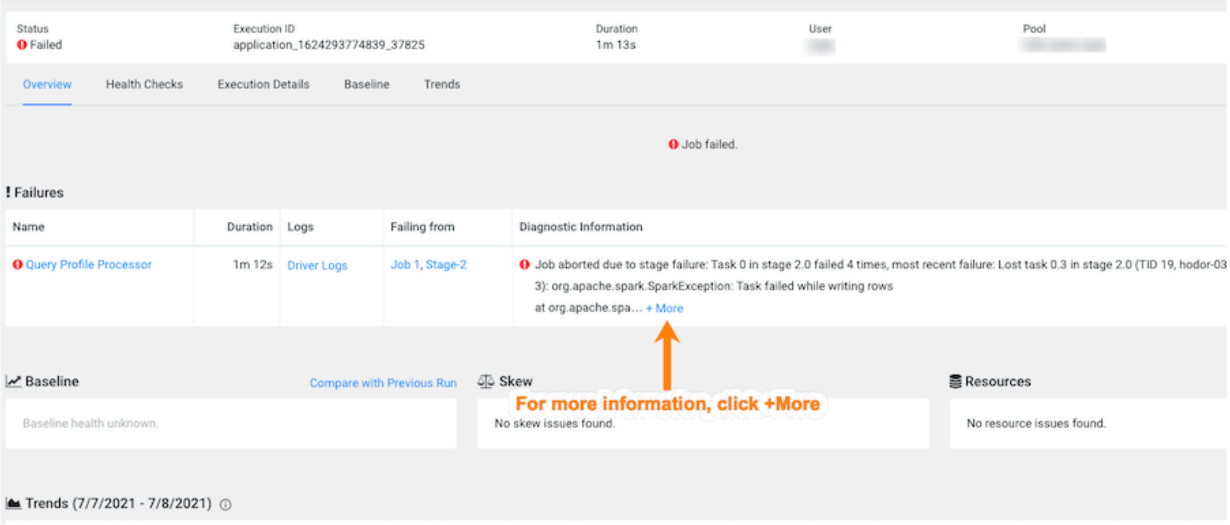
4. From the time-range list, select a time period that meets your requirements.

5. In the **Cluster Summary** page, locate the Spark Jobs Trend chart widget and then click its Failed/Killed Jobs value.

   The engine's Jobs page opens.

6. From the Health Check filter's list, select Failed to Finish, which filters the list to display a list of jobs that did not complete.

7. To view more details about why a job failed to complete, from the Job column select a job's name. The job's page opens displaying information about the job you selected and where the failure happened.



Status	Execution ID	Duration	User	Pool
Failed	application_1624293774839_37825	1m 13s	[redacted]	[redacted]

Job failed.

! Failures

Name	Duration	Logs	Failing from	Diagnostic Information
Query Profile Processor	1m 12s	Driver Logs	Job 1, Stage-2	Job aborted due to stage failure: Task 0 in stage 2.0 failed 4 times, most recent failure: Lost task 0.3 in stage 2.0 (TID 19, hodor-03 3): org.apache.spark.SparkException: Task failed while writing rows at org.apache.spa... + More

Baseline — Compare with Previous Run

Baseline health unknown.

Skew

No skew issues found.

Resources

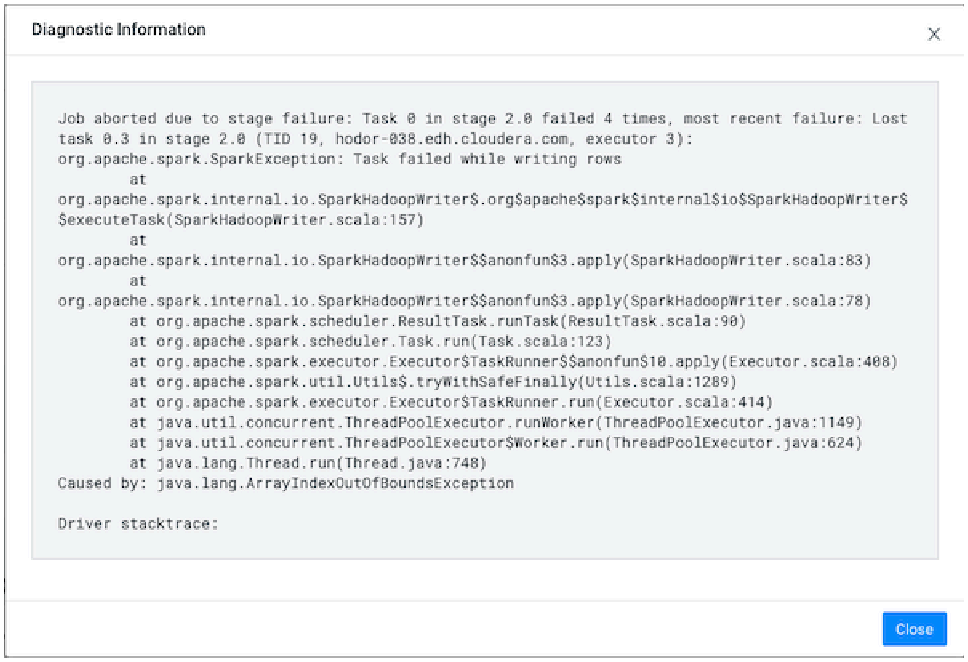
No resource issues found.

For more information, click +More

Trends (7/7/2021 - 7/8/2021)

8. From the Failures section in the Diagnostic Information column, click More.

   The Diagnostic Information dialog box opens, which describes more details about why the job aborted. In the following example, the job was aborted whilst writing rows due to an out of bounds java exception:



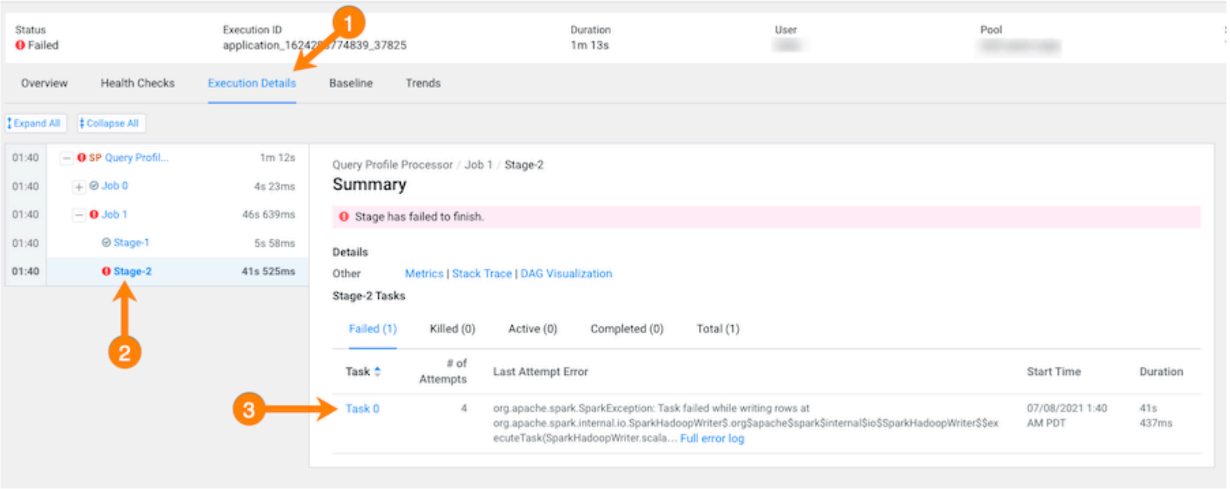
Diagnostic Information

```
Job aborted due to stage failure: Task 0 in stage 2.0 failed 4 times, most recent failure: Lost
task 0.3 in stage 2.0 (TID 19, hodor-038.edh.cloudera.com, executor 3):
org.apache.spark.SparkException: Task failed while writing rows
        at
org.apache.spark.internal.io.SparkHadoopWriter$.org$apache$spark$internal$io$SparkHadoopWriter$
$executeTask(SparkHadoopWriter.scala:157)
        at
org.apache.spark.internal.io.SparkHadoopWriter$$anonfun$3.apply(SparkHadoopWriter.scala:83)
        at
org.apache.spark.internal.io.SparkHadoopWriter$$anonfun$3.apply(SparkHadoopWriter.scala:78)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:123)
        at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.apply(Executor.scala:408)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1289)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:414)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
Caused by: java.lang.ArrayIndexOutOfBoundsException

Driver stacktrace:
```

9. Click Close.
10. To display more information about the stage where the job failed, in this case the Stage-2 process, in the Failing from column, click the stage's link. Or select the Execution Details tab and then click the failed stage's link.

    In the following example's Summary panel, it shows that Task 0 was attempted 4 times:



Status: Failed
Execution ID: application_16242[illegible]774839_37825
Duration: 1m 13s
User: [redacted]
Pool: [redacted]

```
01:40  SP Query Profil...   1m 12s
01:40  Job 0                4s 23ms
01:40  Job 1                46s 639ms
01:40  Stage-1              5s 58ms
01:40  Stage-2              41s 525ms
```

Query Profile Processor / Job 1 / Stage-2

Summary

Stage has failed to finish.

Details

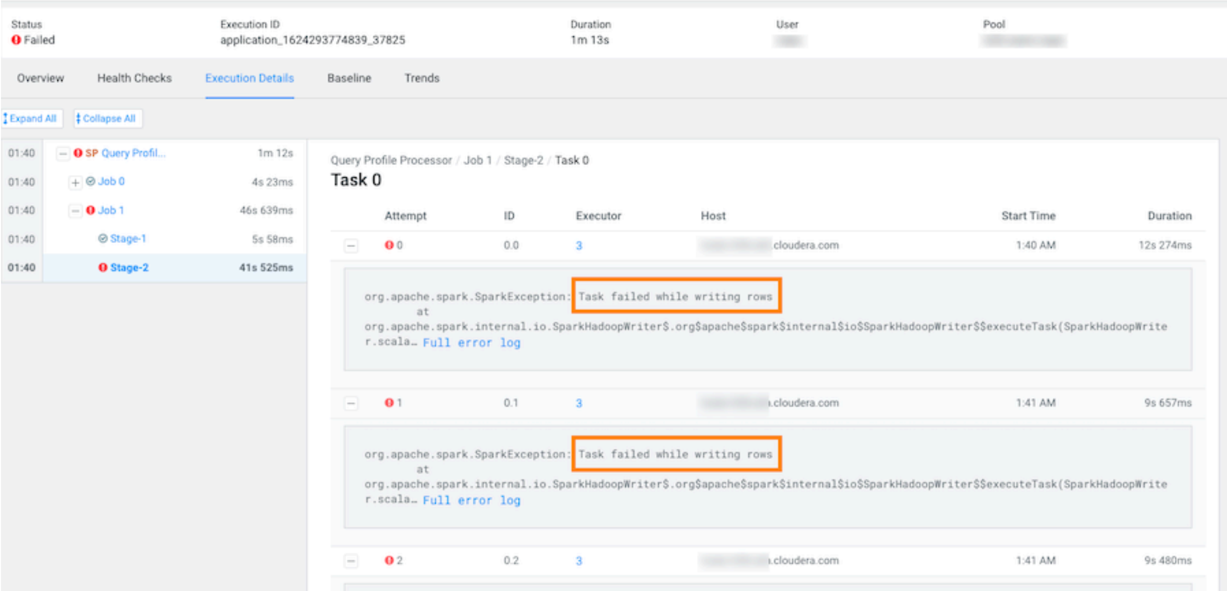
Other: Metrics | Stack Trace | DAG Visualization

Stage-2 Tasks

Task	# of Attempts	Last Attempt Error	Start Time	Duration
Task 0	4	org.apache.spark.SparkException: Task failed while writing rows at org.apache.spark.internal.io.SparkHadoopWriter$.org$apache$spark$internalioSparkHadoopWriter$$executeTask(SparkHadoopWriter.scala... Full error log	07/08/2021 1:40 AM PDT	41s 437ms

11. To display more information about all the failed attempts, in the Summary panel, click the Failed task value.

In the following example, the job aborted when Task 0 was writing rows. To understand more about what triggered the SparkException error message and to further troubleshoot the root cause, you can open the associated log file by clicking Full error log.



Status	Execution ID	Duration	User	Pool
Failed	application_1624293774839_37825	1m 13s	[redacted]	[redacted]

01:40	SP Query Profil...	1m 12s
01:40	Job 0	4s 23ms
01:40	Job 1	46s 639ms
01:40	Stage-1	5s 58ms
01:40	Stage-2	41s 525ms

Query Profile Processor / Job 1 / Stage-2 / Task 0

Task 0

Attempt	ID	Executor	Host	Start Time	Duration
0	0.0	3	[redacted]cloudera.com	1:40 AM	12s 274ms

```
org.apache.spark.SparkException: Task failed while writing rows
        at
org.apache.spark.internal.io.SparkHadoopWriter$.org$apache$spark$internal$io$SparkHadoopWriter$$executeTask(SparkHadoopWrite
r.scala… Full error log
```

Attempt	ID	Executor	Host	Start Time	Duration
1	0.1	3	[redacted].cloudera.com	1:41 AM	9s 657ms

```
org.apache.spark.SparkException: Task failed while writing rows
        at
org.apache.spark.internal.io.SparkHadoopWriter$.org$apache$spark$internal$io$SparkHadoopWriter$$executeTask(SparkHadoopWrite
r.scala… Full error log
```

Attempt	ID	Executor	Host	Start Time	Duration
2	0.2	3	[redacted].cloudera.com	1:41 AM	9s 480ms

Determining the cause of slow and failed queries

You can identify the cause of slow query run times and queries that fail to complete using Cloudera Observability.

About this task

Describes how to determine the cause of slow and failed queries.

Steps with examples from a Virtual Cluster's Spark engine are used to explain how to further investigate and troubleshoot the cause of slow query run times.

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 Environment Name column in the Environments page, locate and click the name of the environment whose workload diagnostic information requires analysis and troubleshooting.

   For this example, select **Virtual Cluster** from the Environments list and then select a Virtual Cluster required for analysis.

   The Environment navigation panel opens, which hierarchically lists the environment and its services hosted on the selected environment.

3. Verify that the **Cluster Summary** page is displayed.

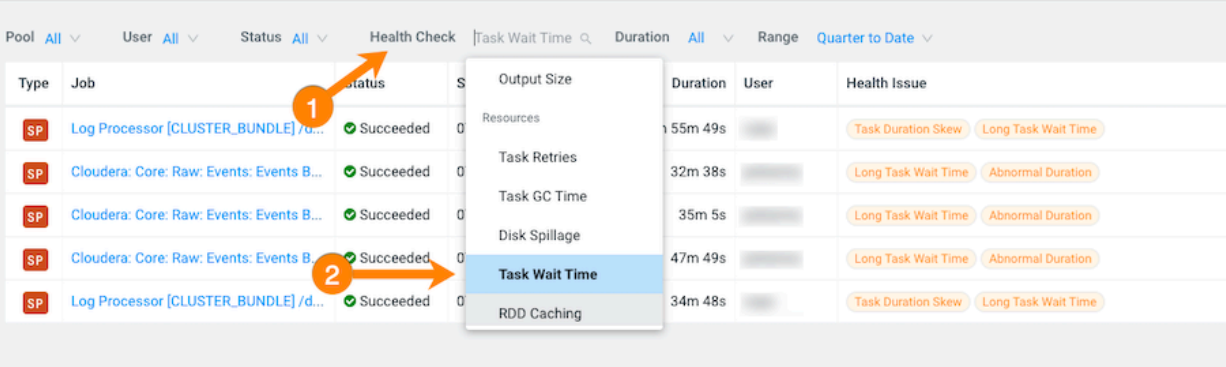
   Tip: The page's title is displayed in the browser tab.

   The **Cluster Summary** page, displays performance trends and metrics about the cluster's processed jobs and queries.

4. From the time-range list, select a time period that meets your requirements.
5. If not already expanded, from the Environment navigation panel expand the Virtual Cluster and then select the **Spark** engine.

   The engine's Summary page opens, in this case the Spark Summary page.
6. From the Job Trend widget, click its Total Jobs value.

   The engine's Jobs page opens.
7. From the Health Check filter's list, select Task Wait Time, which filters and displays a list of jobs with longer than average wait times before the process was executed.



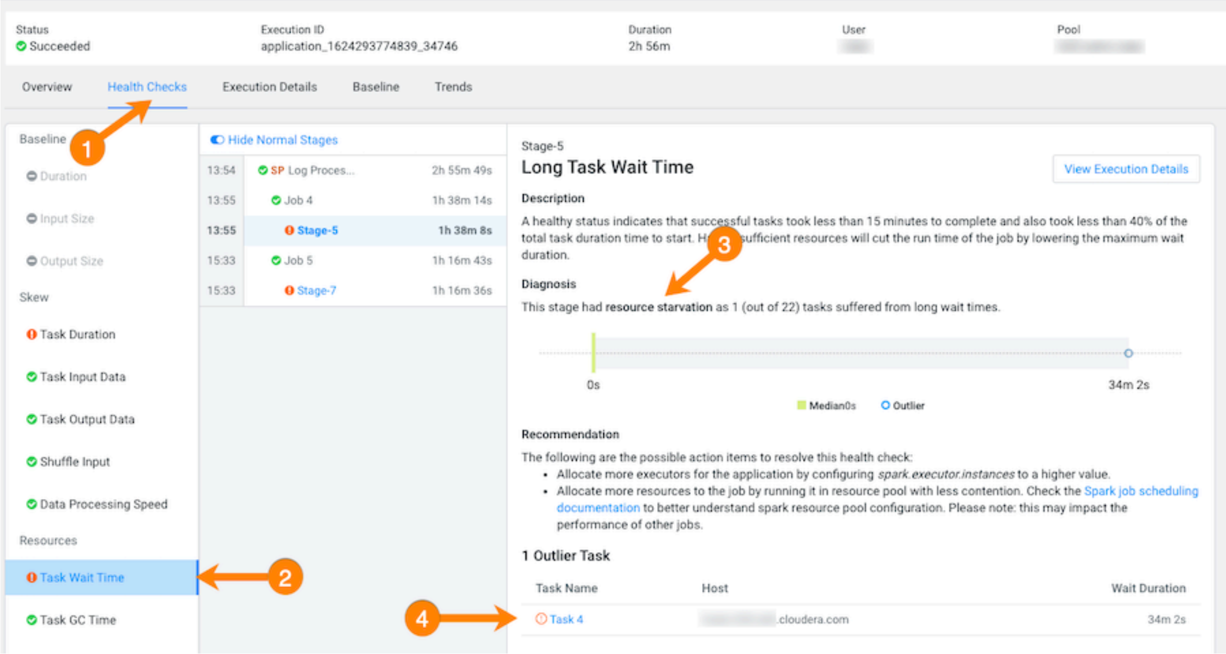
Pool All User All Status All Health Check Task Wait Time Duration All Range Quarter to Date

Type	Job	Status	S	Duration	User	Health Issue
SP	Log Processor [CLUSTER_BUNDLE] /d...	Succeeded	0	55m 49s	[redacted]	Task Duration Skew Long Task Wait Time
SP	Cloudera: Core: Raw: Events: Events B...	Succeeded	0	32m 38s	[redacted]	Long Task Wait Time Abnormal Duration
SP	Cloudera: Core: Raw: Events: Events B...	Succeeded	0	35m 5s	[redacted]	Long Task Wait Time Abnormal Duration
SP	Cloudera: Core: Raw: Events: Events B...	Succeeded	0	47m 49s	[redacted]	Long Task Wait Time Abnormal Duration
SP	Log Processor [CLUSTER_BUNDLE] /d...	Succeeded	0	34m 48s	[redacted]	Task Duration Skew Long Task Wait Time

8. Display more details by selecting a job's name from the Job column and then clicking the Health Checks tab.

   The Baseline Health checks are displayed.
9. From the Health Checks panel on the left, click the Task Wait Time health check, which opens a panel that describes the health check, displays information about the possible causes, and lists recommended solutions.

   In the following example, the long wait time occurred in Stage-5 of the job process due to insufficient resources. The Recommendation section lists items for you to complete that may resolve the problem and the specific outlier tasks that produced the unusual results:



Status: Succeeded

Execution ID: application_1624293774839_34746

Duration: 2h 56m

User: [redacted]

Pool: [redacted]

13:54	SP Log Proces...	2h 55m 49s
13:55	Job 4	1h 38m 14s
13:55	Stage-5	1h 38m 8s
15:33	Job 5	1h 16m 43s
15:33	Stage-7	1h 16m 36s

Stage-5

Long Task Wait Time

Description

A healthy status indicates that successful tasks took less than 15 minutes to complete and also took less than 40% of the total task duration time to start. H sufficient resources will cut the run time of the job by lowering the maximum wait duration.

Diagnosis

This stage had **resource starvation** as 1 (out of 22) tasks suffered from long wait times.

Recommendation

The following are the possible action items to resolve this health check:

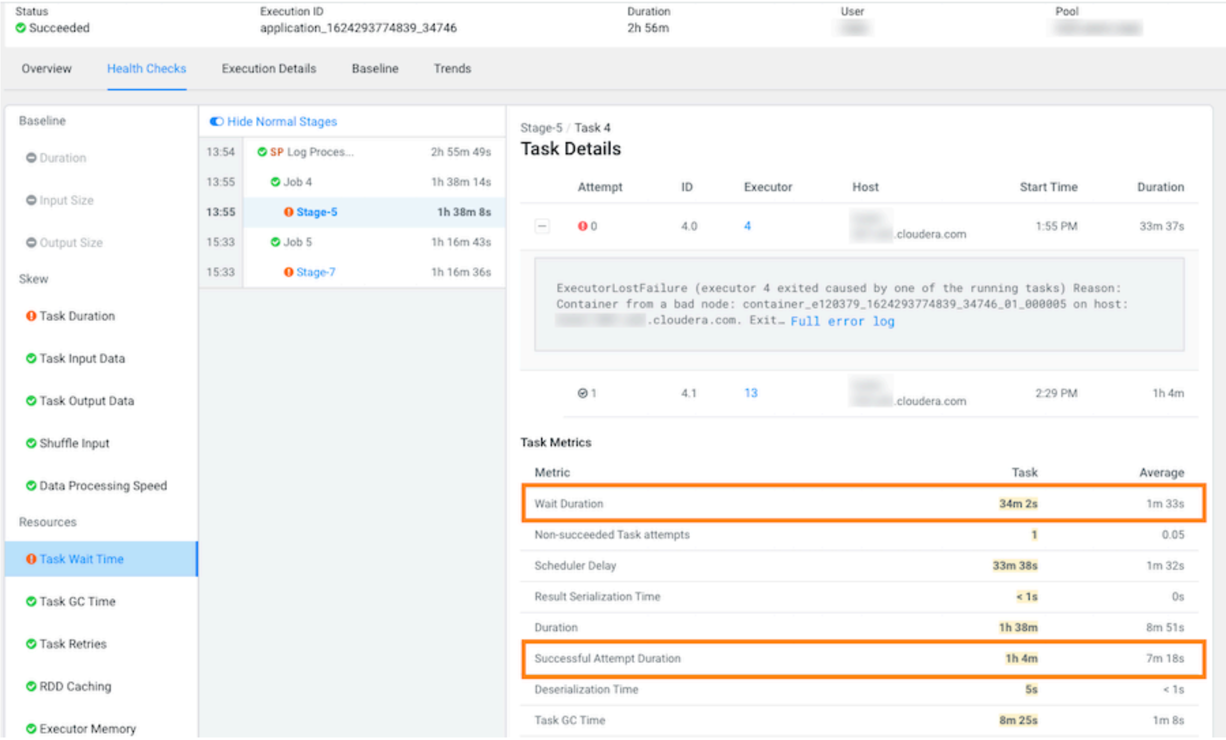
- Allocate more executors for the application by configuring *spark.executor.instances* to a higher value.
- Allocate more resources to the job by running it in resource pool with less contention. Check the Spark job scheduling documentation to better understand spark resource pool configuration. Please note: this may impact the performance of other jobs.

1 Outlier Task

Task Name	Host	Wait Duration
Task 4	[redacted].cloudera.com	34m 2s

10. To display more details about why this job is experiencing longer than average wait times, click one of the tasks listed under Outlier Tasks.

    In the following example, the Task Metrics section shows higher than average criteria measurement results and the Task Details reveal an ExecutorLostFailure error. This indicates a probable memory issue, where the container is exceeding the memory limits. In this case, more details may be found by clicking Full error log and reviewing the log:



Status	Execution ID	Duration	User	Pool
Succeeded	application_1624293774839_34746	2h 56m	[redacted]	[redacted]

13:54	SP Log Proces...	2h 55m 49s
13:55	Job 4	1h 38m 14s
13:55	Stage-5	1h 38m 8s
15:33	Job 5	1h 16m 43s
15:33	Stage-7	1h 16m 36s

Stage-5 / Task 4

Task Details

Attempt	ID	Executor	Host	Start Time	Duration
0	4.0	4	[redacted].cloudera.com	1:55 PM	33m 37s

```
ExecutorLostFailure (executor 4 exited caused by one of the running tasks) Reason:
Container from a bad node: container_e120379_1624293774839_34746_01_000005 on host:
[redacted].cloudera.com. Exit… Full error log
```

Attempt	ID	Executor	Host	Start Time	Duration
1	4.1	13	[redacted].cloudera.com	2:29 PM	1h 4m

Task Metrics

Metric	Task	Average
Wait Duration	34m 2s	1m 33s
Non-succeeded Task attempts	1	0.05
Scheduler Delay	33m 38s	1m 32s
Result Serialization Time	< 1s	0s
Duration	1h 38m	8m 51s
Successful Attempt Duration	1h 4m	7m 18s
Deserialization Time	5s	< 1s
Task GC Time	8m 25s	1m 8s

Troubleshooting with the Job Comparison Feature

You can compare two different runs of the same job, which is especially useful when you notice unexpected changes, such as when you have a job that consistently completes within a specific amount of time and then it starts taking longer. Comparing two runs of the same job enables you to analyze the performance and differences so that you can troubleshoot the cause.

About this task

Describes how to compare any two runs of a job using the Job Comparison tool.

Steps with examples from a Virtual Cluster's Spark engine are used to explain how to use the job comparison feature for further investigation and troubleshooting.

Note: When a job is flagged as slow, you can select the slow job from the Slow Jobs chart widget in the job's engine page and then in the details page, click Compare with Previous Run. The job is compared with its last run and the results are displayed in the **Job Comparison** page for you to analyze.

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 Environment Name column in the Environments page, locate and click the name of the environment whose workload diagnostic information requires analysis and troubleshooting.

   For this example, select **Virtual Cluster** from the Environments list and then select a Virtual Cluster required for analysis.

   The Environment navigation panel opens, which hierarchically lists the environment and its services hosted on the selected environment.
3. Verify that the **Cluster Summary** page is displayed.

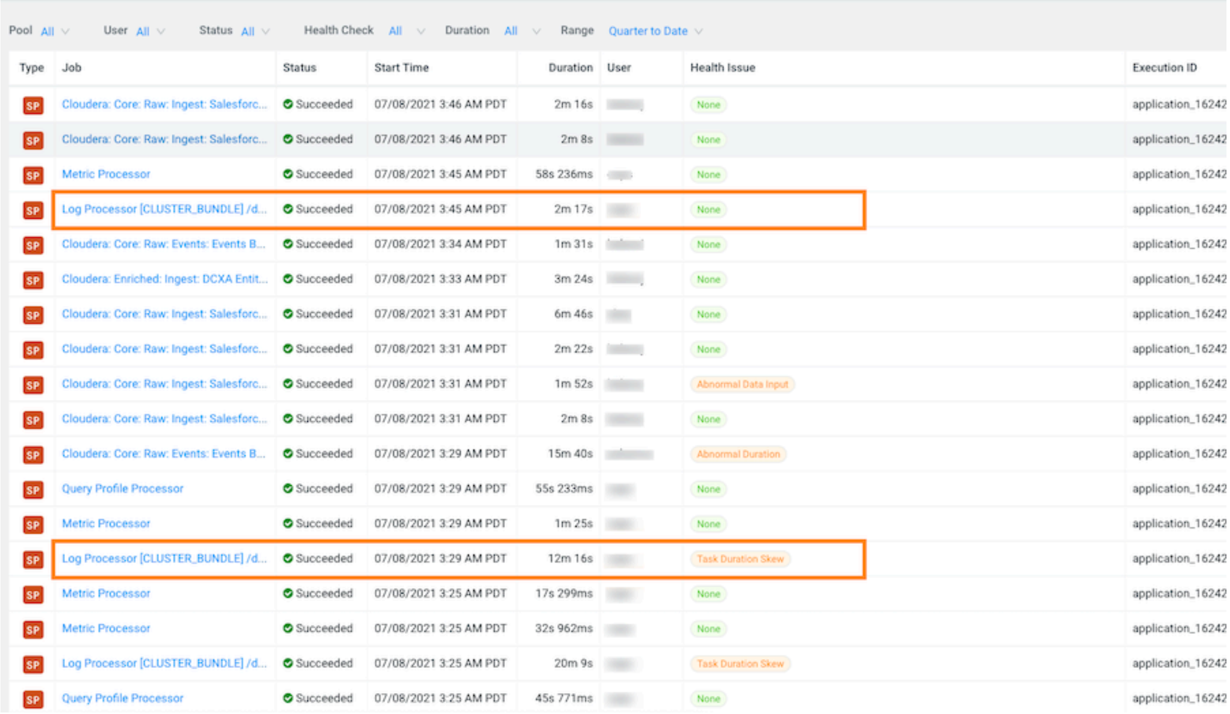
   Tip: The page's title is displayed in the browser tab.

   The **Cluster Summary** page, displays performance trends and metrics about the cluster's processed jobs and queries.
4. From the time-range list, select a time period that meets your requirements.
5. If not already expanded, from the Environment navigation panel expand the Virtual Cluster and then select the **Spark** engine.

   The engine's Summary page opens, in this case the Spark Summary page.
6. From the Job Trend widget, click its Total Jobs value.

   The engine's Jobs page opens.
7. Examine the list of jobs that have executed during the selected time period and manually compare runs of the same job.

   For example, as shown in the following image, when manually comparing the last two runs of the Log Processor job we can see that there are duration differences. In this example, the older run had a Task duration skew health issue, which appears to be fixed.

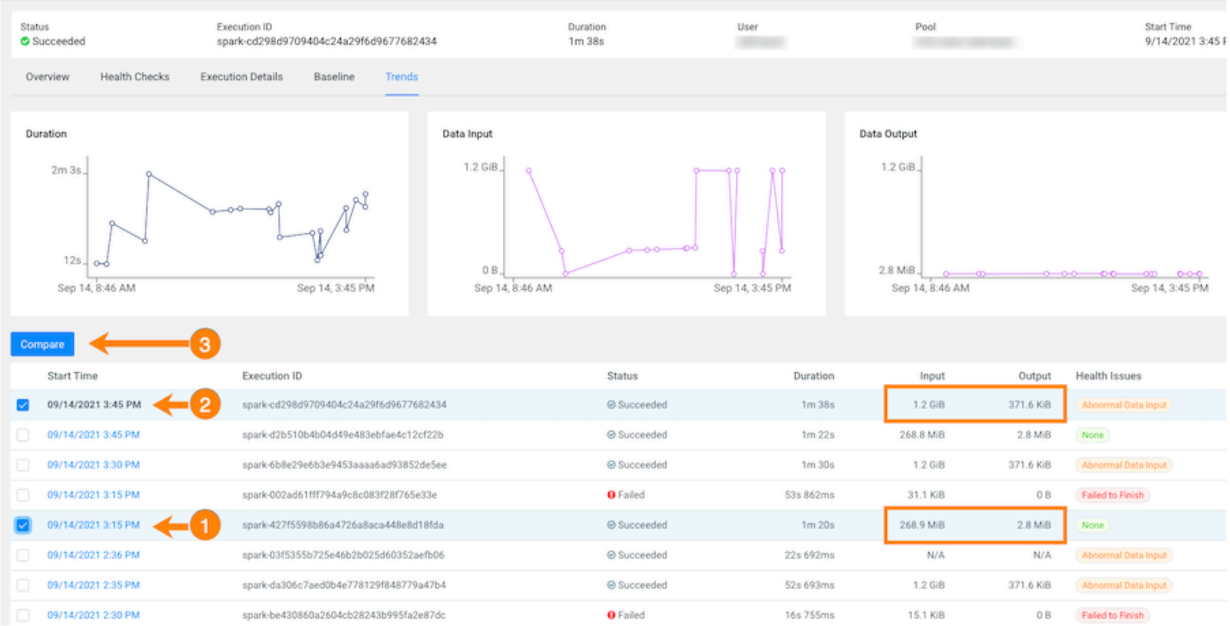


Pool All ∨ User All ∨ Status All ∨ Health Check All ∨ Duration All ∨ Range Quarter to Date ∨

Type	Job	Status	Start Time	Duration	User	Health Issue	Execution ID
SP	Cloudera: Core: Raw: Ingest: Salesforc...	✓ Succeeded	07/08/2021 3:46 AM PDT	2m 16s	[redacted]	None	application_16242
SP	Cloudera: Core: Raw: Ingest: Salesforc...	✓ Succeeded	07/08/2021 3:46 AM PDT	2m 8s	[redacted]	None	application_16242
SP	Metric Processor	✓ Succeeded	07/08/2021 3:45 AM PDT	58s 236ms	[redacted]	None	application_16242
SP	Log Processor [CLUSTER_BUNDLE] /d...	✓ Succeeded	07/08/2021 3:45 AM PDT	2m 17s	[redacted]	None	application_16242
SP	Cloudera: Core: Raw: Events: Events B...	✓ Succeeded	07/08/2021 3:34 AM PDT	1m 31s	[redacted]	None	application_16242
SP	Cloudera: Enriched: Ingest: DCXA Entit...	✓ Succeeded	07/08/2021 3:33 AM PDT	3m 24s	[redacted]	None	application_16242
SP	Cloudera: Core: Raw: Ingest: Salesforc...	✓ Succeeded	07/08/2021 3:31 AM PDT	6m 46s	[redacted]	None	application_16242
SP	Cloudera: Core: Raw: Ingest: Salesforc...	✓ Succeeded	07/08/2021 3:31 AM PDT	2m 22s	[redacted]	None	application_16242
SP	Cloudera: Core: Raw: Ingest: Salesforc...	✓ Succeeded	07/08/2021 3:31 AM PDT	1m 52s	[redacted]	Abnormal Data Input	application_16242
SP	Cloudera: Core: Raw: Ingest: Salesforc...	✓ Succeeded	07/08/2021 3:31 AM PDT	2m 8s	[redacted]	None	application_16242
SP	Cloudera: Core: Raw: Events: Events B...	✓ Succeeded	07/08/2021 3:29 AM PDT	15m 40s	[redacted]	Abnormal Duration	application_16242
SP	Query Profile Processor	✓ Succeeded	07/08/2021 3:29 AM PDT	55s 233ms	[redacted]	None	application_16242
SP	Metric Processor	✓ Succeeded	07/08/2021 3:29 AM PDT	1m 25s	[redacted]	None	application_16242
SP	Log Processor [CLUSTER_BUNDLE] /d...	✓ Succeeded	07/08/2021 3:29 AM PDT	12m 16s	[redacted]	Task Duration Skew	application_16242
SP	Metric Processor	✓ Succeeded	07/08/2021 3:25 AM PDT	17s 299ms	[redacted]	None	application_16242
SP	Metric Processor	✓ Succeeded	07/08/2021 3:25 AM PDT	32s 962ms	[redacted]	None	application_16242
SP	Log Processor [CLUSTER_BUNDLE] /d...	✓ Succeeded	07/08/2021 3:25 AM PDT	20m 9s	[redacted]	Task Duration Skew	application_16242
SP	Query Profile Processor	✓ Succeeded	07/08/2021 3:25 AM PDT	45s 771ms	[redacted]	None	application_16242

8. List and display details of all the runs of a specific job of interest by selecting one of the job runs and then in its jobs details page, click the Trends tab.

In the following example, notice how the amount of Input and Output data changes between runs. The Job Comparison tool enables you to examine more details about two runs to determine why the amount of data changed. In this case you can compare a run with no health issues with the last run of the job:



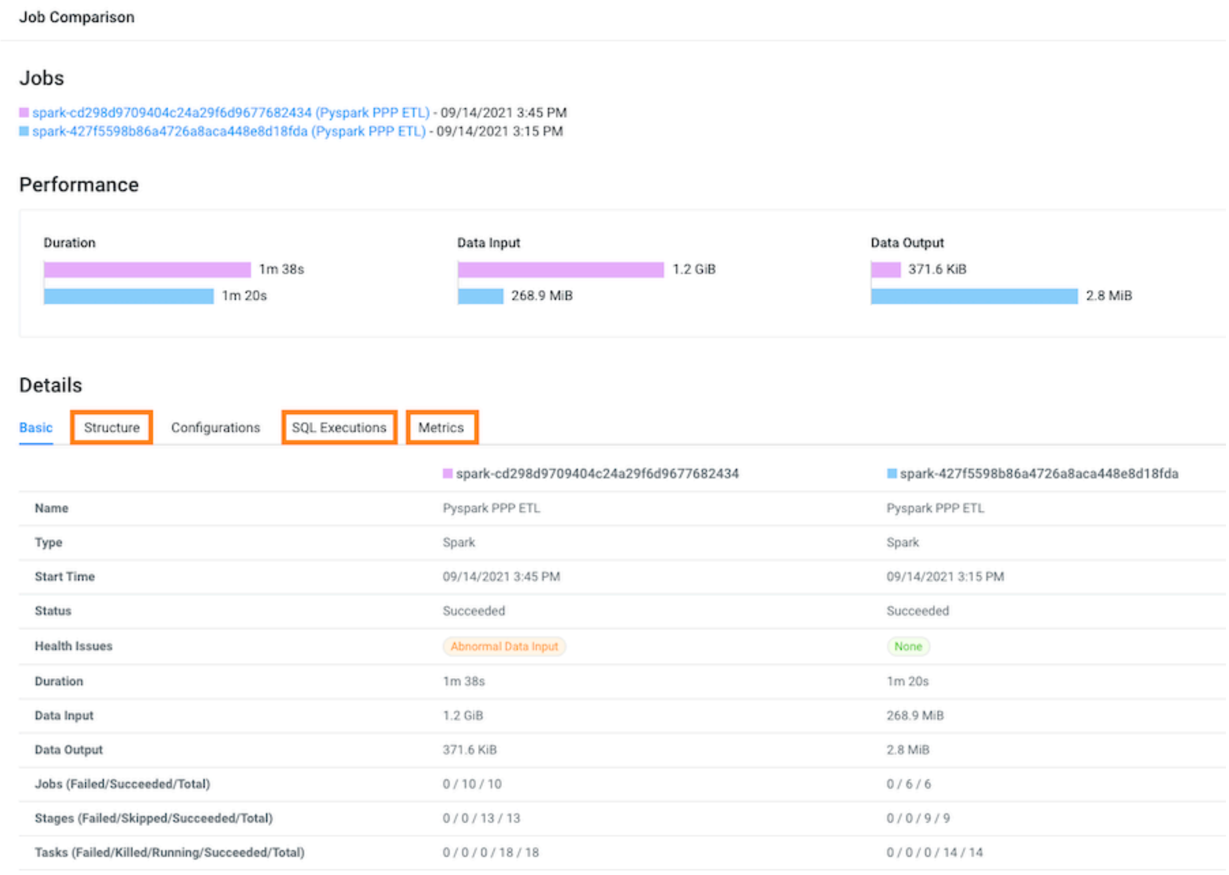
Status	Execution ID	Duration	User	Pool	Start Time
Succeeded	spark-cd298d9709404c24a29f6d9677682434	1m 38s	[redacted]	[redacted]	9/14/2021 3:45

	Start Time	Execution ID	Status	Duration	Input	Output	Health Issues
☑	09/14/2021 3:45 PM	spark-cd298d9709404c24a29f6d9677682434	Succeeded	1m 38s	1.2 GiB	371.6 KiB	Abnormal Data Input
□	09/14/2021 3:45 PM	spark-d2b510b4b04d49e483ebfae4c12cf22b	Succeeded	1m 22s	268.8 MiB	2.8 MiB	None
□	09/14/2021 3:30 PM	spark-6b8e29e6b3e9453aaaa6ad93852de5ee	Succeeded	1m 30s	1.2 GiB	371.6 KiB	Abnormal Data Input
□	09/14/2021 3:15 PM	spark-002ad61fff794a9c8c083f28f765e33e	Failed	53s 862ms	31.1 KiB	0 B	Failed to Finish
☑	09/14/2021 3:15 PM	spark-427f5598b86a4726a8aca448e8d18fda	Succeeded	1m 20s	268.9 MiB	2.8 MiB	None
□	09/14/2021 2:36 PM	spark-03f5355b725e46b2b025d60352aefb06	Succeeded	22s 692ms	N/A	N/A	Abnormal Data Input
□	09/14/2021 2:35 PM	spark-da306c7aed0b4e778129f848779a47b4	Succeeded	52s 693ms	1.2 GiB	371.6 KiB	Abnormal Data Input
□	09/14/2021 2:30 PM	spark-be430860a2604cb28243b995fa2e87dc	Failed	16s 755ms	15.1 KiB	0 B	Failed to Finish

9. To compare two job runs, select the check boxes adjacent to the job runs you require and then click Compare.

   The Job Comparison page opens, displaying more details about each job.

   For this example's comparison, the tabs that contain more information about the job runs are the Structure, SQL Executions, and the Metrics tabs:



Job Comparison

Jobs

- spark-cd298d9709404c24a29f6d9677682434 (Pyspark PPP ETL) - 09/14/2021 3:45 PM
- spark-427f5598b86a4726a8aca448e8d18fda (Pyspark PPP ETL) - 09/14/2021 3:15 PM

Performance

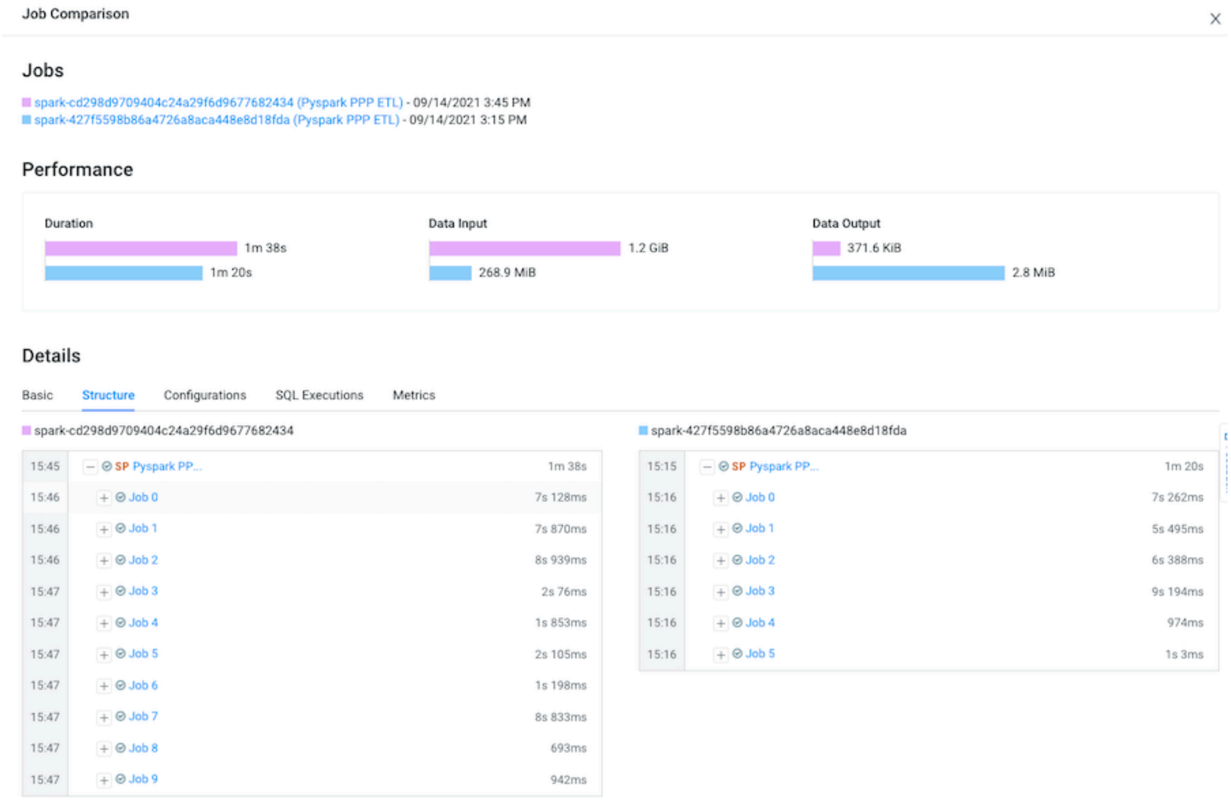
Details

	spark-cd298d9709404c24a29f6d9677682434	spark-427f5598b86a4726a8aca448e8d18fda
Name	Pyspark PPP ETL	Pyspark PPP ETL
Type	Spark	Spark
Start Time	09/14/2021 3:45 PM	09/14/2021 3:15 PM
Status	Succeeded	Succeeded
Health Issues	Abnormal Data Input	None
Duration	1m 38s	1m 20s
Data Input	1.2 GiB	268.9 MiB
Data Output	371.6 KiB	2.8 MiB
Jobs (Failed/Succeeded/Total)	0 / 10 / 10	0 / 6 / 6
Stages (Failed/Skipped/Succeeded/Total)	0 / 0 / 13 / 13	0 / 0 / 9 / 9
Tasks (Failed/Killed/Running/Succeeded/Total)	0 / 0 / 0 / 18 / 18	0 / 0 / 0 / 14 / 14

Note: The SQL Executions tab is only available for Spark jobs.

10. Display and compare the sub-jobs executed for both of your selected job runs by selecting the Structure tab.

    For example, as shown in the following image, the last run of the job (with health issues) completed in 1 minute and 38 seconds and executed 9 sub-jobs and the run that had no health issues took 1 minute and 20 seconds but only executed 5 sub-jobs. Clicking any of the listed sub-jobs displays more details.



Jobs

- spark-cd298d9709404c24a29f6d9677682434 (Pyspark PPP ETL) - 09/14/2021 3:45 PM
- spark-427f5598b86a4726a8aca448e8d18fda (Pyspark PPP ETL) - 09/14/2021 3:15 PM

Performance

	Duration	Data Input	Data Output
spark-cd298d9709404c24a29f6d9677682434	1m 38s	1.2 GiB	371.6 KiB
spark-427f5598b86a4726a8aca448e8d18fda	1m 20s	268.9 MiB	2.8 MiB

Details

spark-cd298d9709404c24a29f6d9677682434

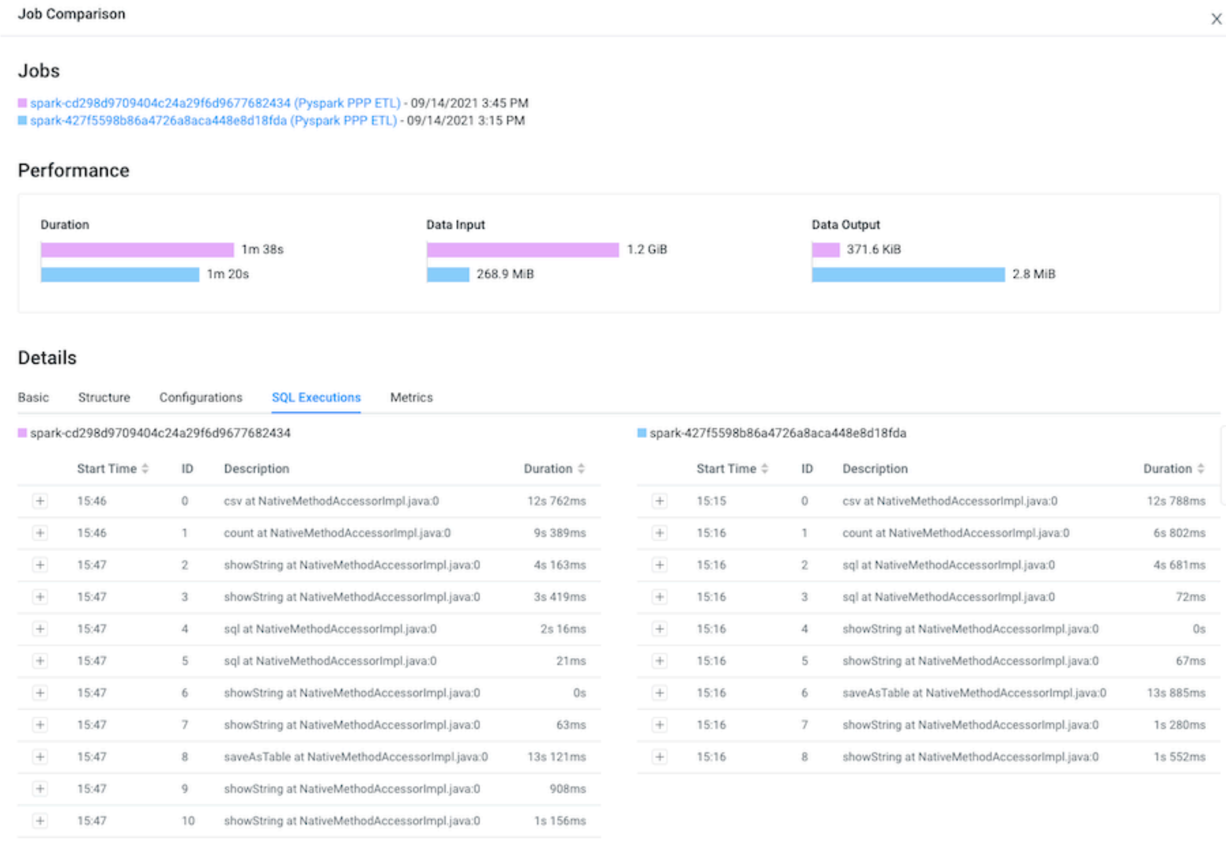
15:45	− SP Pyspark PP...	1m 38s
15:46	+ Job 0	7s 128ms
15:46	+ Job 1	7s 870ms
15:46	+ Job 2	8s 939ms
15:47	+ Job 3	2s 76ms
15:47	+ Job 4	1s 853ms
15:47	+ Job 5	2s 105ms
15:47	+ Job 6	1s 198ms
15:47	+ Job 7	8s 833ms
15:47	+ Job 8	693ms
15:47	+ Job 9	942ms

spark-427f5598b86a4726a8aca448e8d18fda

15:15	− SP Pyspark PP...	1m 20s
15:16	+ Job 0	7s 262ms
15:16	+ Job 1	5s 495ms
15:16	+ Job 2	6s 388ms
15:16	+ Job 3	9s 194ms
15:16	+ Job 4	974ms
15:16	+ Job 5	1s 3ms

11. Display and compare Spark SQL queries that were run and how long they ran for both of your selected job runs by selecting the SQL Executions tab.

    For example, as shown in the following image, more Spark SQL queries were run on the data in the last job run.



Job Comparison

Jobs

🟪 spark-cd298d9709404c24a29f6d9677682434 (Pyspark PPP ETL) - 09/14/2021 3:45 PM
🟦 spark-427f5598b86a4726a8aca448e8d18fda (Pyspark PPP ETL) - 09/14/2021 3:15 PM

Performance

Duration	Data Input	Data Output
🟪 1m 38s	🟪 1.2 GiB	🟪 371.6 KiB
🟦 1m 20s	🟦 268.9 MiB	🟦 2.8 MiB

Details

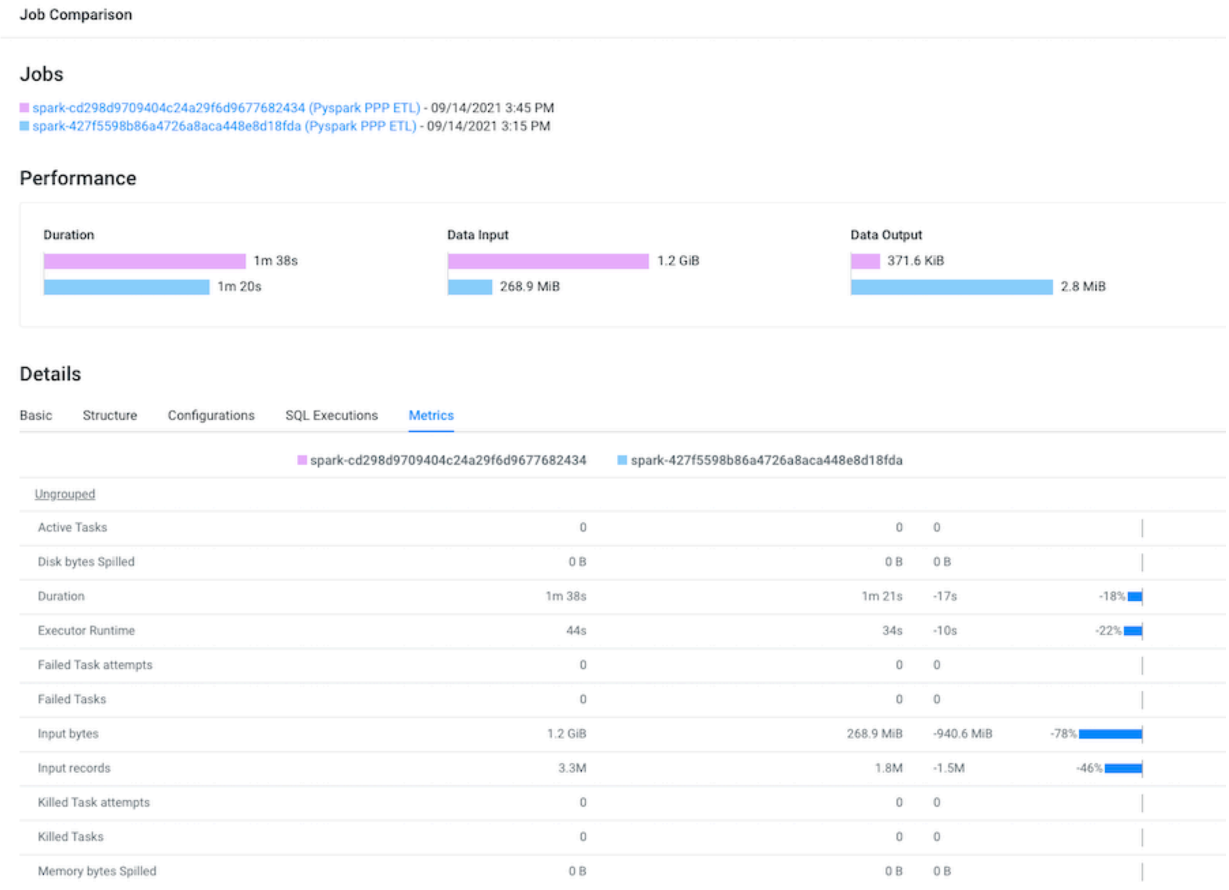
🟪 spark-cd298d9709404c24a29f6d9677682434

	Start Time	ID	Description	Duration
+	15:46	0	csv at NativeMethodAccessorImpl.java:0	12s 762ms
+	15:46	1	count at NativeMethodAccessorImpl.java:0	9s 389ms
+	15:47	2	showString at NativeMethodAccessorImpl.java:0	4s 163ms
+	15:47	3	showString at NativeMethodAccessorImpl.java:0	3s 419ms
+	15:47	4	sql at NativeMethodAccessorImpl.java:0	2s 16ms
+	15:47	5	sql at NativeMethodAccessorImpl.java:0	21ms
+	15:47	6	showString at NativeMethodAccessorImpl.java:0	0s
+	15:47	7	showString at NativeMethodAccessorImpl.java:0	63ms
+	15:47	8	saveAsTable at NativeMethodAccessorImpl.java:0	13s 121ms
+	15:47	9	showString at NativeMethodAccessorImpl.java:0	908ms
+	15:47	10	showString at NativeMethodAccessorImpl.java:0	1s 156ms

🟦 spark-427f5598b86a4726a8aca448e8d18fda

	Start Time	ID	Description	Duration
+	15:15	0	csv at NativeMethodAccessorImpl.java:0	12s 788ms
+	15:16	1	count at NativeMethodAccessorImpl.java:0	6s 802ms
+	15:16	2	sql at NativeMethodAccessorImpl.java:0	4s 681ms
+	15:16	3	sql at NativeMethodAccessorImpl.java:0	72ms
+	15:16	4	showString at NativeMethodAccessorImpl.java:0	0s
+	15:16	5	showString at NativeMethodAccessorImpl.java:0	67ms
+	15:16	6	saveAsTable at NativeMethodAccessorImpl.java:0	13s 885ms
+	15:16	7	showString at NativeMethodAccessorImpl.java:0	1s 280ms
+	15:16	8	showString at NativeMethodAccessorImpl.java:0	1s 552ms

12. Display and compare what metrics were performed on both of your selected job runs by selecting the Metrics tab. For example, as shown in the following image, more input records were digested in the last job run.



Job Comparison

Jobs

- spark-cd298d9709404c24a29f6d9677682434 (Pyspark PPP ETL) - 09/14/2021 3:45 PM
- spark-427f5598b86a4726a8aca448e8d18fda (Pyspark PPP ETL) - 09/14/2021 3:15 PM

Performance

Details

	spark-cd298d9709404c24a29f6d9677682434	spark-427f5598b86a4726a8aca448e8d18fda		
Ungrouped				
Active Tasks	0	0	0	
Disk bytes Spilled	0 B	0 B	0 B	
Duration	1m 38s	1m 21s	-17s	-18%
Executor Runtime	44s	34s	-10s	-22%
Failed Task attempts	0	0	0	
Failed Tasks	0	0	0	
Input bytes	1.2 GiB	268.9 MiB	-940.6 MiB	-78%
Input records	3.3M	1.8M	-1.5M	-46%
Killed Task attempts	0	0	0	
Killed Tasks	0	0	0	
Memory bytes Spilled	0 B	0 B	0 B	