

Cloudera Observability On-Premises Cluster Optimization

Date published: 2024-01-31

Date modified: 2024-02-14

CLOUDERA

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

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Understanding your environment.....	5
Supported browsers.....	5
Managing your workloads and users.....	5
Classifying workloads for analysis with Workload Views.....	5
Automatically generate workload views.....	5
Defining workload views manually.....	8
Assigning access roles in Cloudera Observability On-Premises.....	10
Understanding the Cloudera Observability On-Premises access roles.....	11
Understanding a Cloudera Observability On-Premises cluster policy.....	13
Configuring a default systems administrator for Cloudera Observability On-Premises.....	13
Assigning Cloudera Observability On-Premises access roles.....	14
Managing Your Cloudera Observability On-Premises access roles.....	18
Purging HDFS data.....	18
Understanding the purge date used by the purge event.....	19
Cloudera Observability On-Premises purge event parameters.....	19
Configuring the Cloudera Observability On-Premises purge event.....	21
Manually executing a Cloudera Observability On-Premises purge event.....	22
Managing your Cloudera Observability On-Premises purge event.....	22
Working with alerts, costs, and reports.....	24
Analyzing your environment costs with Cloudera Observability On-Premises.....	24
Configuring the Cloudera Observability On-Premises cost center criteria.....	25
Creating a Cloudera Observability On-Premises cost center.....	26
Assigning uncategorized resources to a cost center.....	29
Displaying your costs associated with a cost center.....	30
Downloading your chargeback costs.....	32
Triggering action alerts across jobs and queries.....	33
Creating an auto action event.....	33
Events and management details of auto actions.....	36
Managing your auto actions.....	38
Auto action email notification examples.....	39
Understanding, identifying, and addressing problems with Cloudera	
Observability On-Premises.....	40
Specifying a time range.....	40
Exporting a report about your workload jobs and queries.....	42
Analyzing your tables.....	42
Understanding the Cloudera Observability On-Premises metastore analytics UI elements.....	43
Understanding the Hive Metastore category.....	46
Displaying the Metastore Analytics.....	49
Analyzing your Hive queries for debugging and optimization.....	50
Identifying inefficient phases of your Hive queries.....	51
About the Cloudera Observability On-Premises Hive cluster service metrics.....	53

Troubleshooting an abnormal job duration.....	55
Troubleshooting failed jobs.....	59
Determining the cause of slow and failed queries.....	62
Troubleshooting with the Job Comparison Feature.....	65

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, purge your data, 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 On-Premises web UI, you must upgrade to the latest version of Cloudera Manager and restart Telemetry Publisher.

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 On-Premises administration features that enable you to define workload views for analyzing specific items of interest, purge your data, 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.

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 On-Premises default views.

Procedure

1. Verify that you are logged in to the Cloudera Observability On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.

- d) From the **Environment Name** column in the Environment's table, select the environment required for analysis.

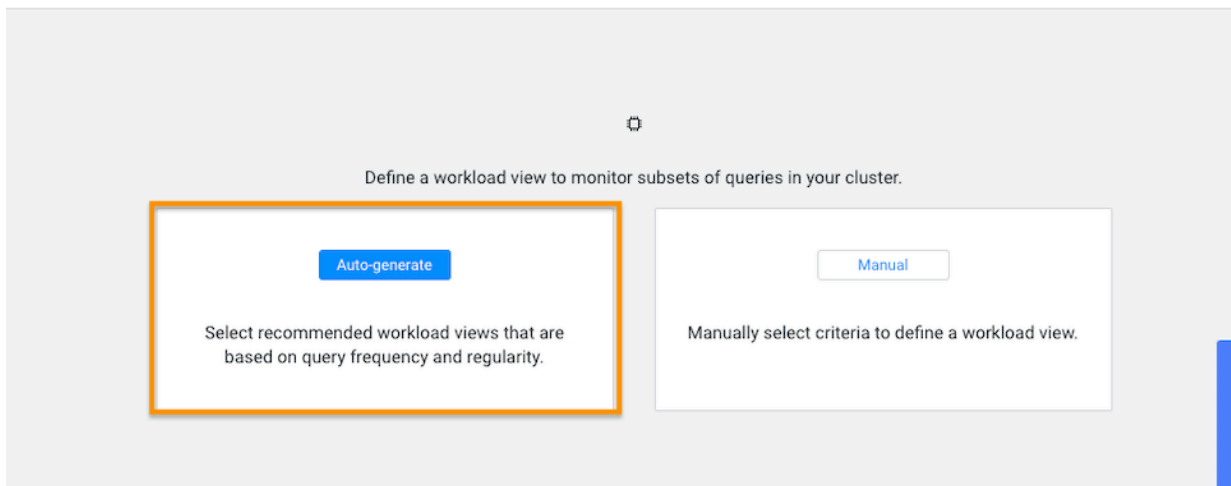
The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.

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



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

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



- 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:

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	Workload Queries	% of Total Queries	Recommended SLA	Warning Threshold
<input checked="" type="checkbox"/> tables_19_58678	User: [criteria] Pool: [criteria] Table: [criteria] Statement Type: Query		16%	5s	5%
<input checked="" type="checkbox"/> tables_19_57837	User: [criteria] Pool: [criteria] Table: [criteria] Statement Type: Query	57.8K	15%	5s	5%
<input checked="" type="checkbox"/> tables_20_57335	User: [criteria] Pool: [criteria] Table: [criteria] Statement Type: Query	57.3K	15%	6s	5%
<input checked="" type="checkbox"/> tables_44_18959	User: [criteria] Pool: [criteria] Table: [criteria] Statement Type: Query	19K	5%	3s	5%
<input checked="" type="checkbox"/> tables_25_17121	User: [criteria] Pool: [criteria] Table: [criteria] Statement Type: Query	17.1K	5%	927ms	5%

Buttons: Cancel, Add Selected

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

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

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 [criteria]	2s	90%	76%
●	●	TB-Table	Impala	Table: ANY OF [criteria] Statement Type: Query	10s	10%	81%
●	●	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: [criteria]	30s	95%	33%
●	-	user_query	Impala	User: tserver Statement Type: Query	1m	2%	19%

- 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.
- To create a new view do the following:
 - Verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.



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

- Select the Workloads tab.
- 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.

Procedure

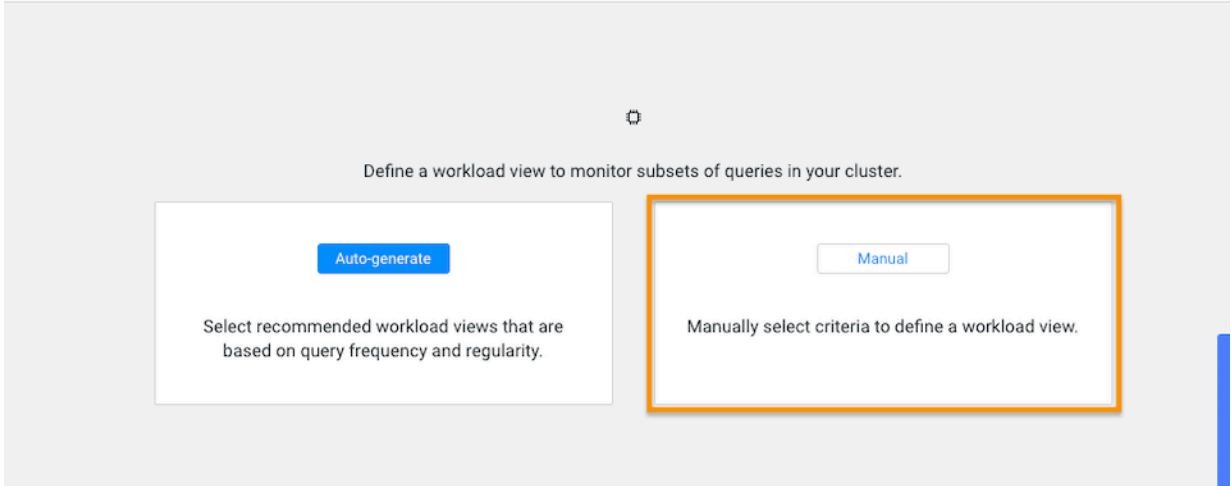
1. Verify that you are logged in to the Cloudera Observability On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.
The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.
 - d) From the **Environment Name** column in the Environment's table, select the environment required for analysis.
The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.
2. Depending on the environment selected, verify that the **Cluster Summary** page is displayed for the environment's cluster required as a workload view.



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

3. Select the Workloads tab.

- In the Workloads page, click Manual:



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 On-Premises 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 ⊙

* Engine

* Criteria ⊙

* SLA ⊙

 Example: 1h 2m 3s 5ms

* Warning Threshold ⊙
 % queries missed SLA
 Sets the percentage of queries missing the SLA that is to be reached before the workload is flagged with a failed status.

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

01/24/2021 - 07/23/2021

Total Jobs	Missed SLA %
30063	76%

- 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. 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	2s	90%	76%
●	●	TB-Table	Impala	Table: ANY OF Statement Type: Query	10s	10%	81%
●	●	_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:	30s	95%	33%
●	-	user_ 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. 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. 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.



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.

Assigning access roles in Cloudera Observability On-Premises

Cloudera Observability On-Premises supports cluster privilege role types that define who is entitled to access jobs and queries that are created by the user, who is entitled to create and administer cost centers and view cluster costs, and who is entitled to access and administer jobs and queries within either a specific cluster or across all clusters within the Cloudera Observability On-Premises environment.

Limiting the trust boundary for jobs, queries, cluster costs, and administrative management at the cluster level, enables more control over the security and access management of your Cloudera Observability On-Premises environment.

Understanding the Cloudera Observability On-Premises access roles

Describes the Cloudera Observability On-Premises access roles.



Important: Customers are responsible for managing and reviewing access credentials for their Cloudera Observability On-Premises accounts and activities. All user privileges and access rights should periodically be reviewed and monitored, including who should access Cloudera Observability On-Premises, its services, and components. For example, access rights should be reviewed when a user moves to another business unit.

Cloudera Observability On-Premises supports cluster privilege roles that define Cloudera Observability On-Premises users as a:

- System Admin
- Cluster Admin
- Cluster User

The following tables describe these cluster privilege roles, also known as access roles:

System Admin access role

An authentic Cloudera Observability On-Premises user who is assigned the System Admin access role has full access rights and system administrator privileges across all clusters within the Cloudera Observability On-Premises environment. Where they can view, edit, and create cost centers, view, edit, and create auto actions, and view all the jobs and queries in all the Workload clusters. These users have the least restrictive access permissions.

Table 1: System Admin

Resource	Actions
Access Management page	View and manage all the Cloudera Observability On-Premises cluster policies and user access from the Access Management page
Cluster	<ul style="list-style-type: none"> • View all the workload clusters on the Clusters page • Rename a workload cluster • Delete a workload cluster
Workloads	<ul style="list-style-type: none"> • Create workloads • View all the workloads in a cluster • Update all the workloads in a cluster • Delete all the workloads in a cluster
Queries	View all the queries in all the clusters of the Cloudera Observability On-Premises environment
Jobs	View all the jobs in all the clusters of the Cloudera Observability On-Premises environment
Chargeback	<ul style="list-style-type: none"> • Create cost centers • Update cost centers • List cost centers • Delete cost centers • View all the Chargeback related dashboards
Auto Actions	<ul style="list-style-type: none"> • Create auto actions • View auto actions • Update auto actions • Disable auto actions • Delete auto actions • Enable an auto action email

Cluster Admin access role

An authentic Cloudera Observability On-Premises user who is assigned the Cluster Admin access role has full access rights and cluster administrator privileges across an assigned cluster within the Cloudera Observability On-Premises environment. Where they can view all the jobs and queries in the assigned Workload cluster.

Table 2: Cluster Admin

Resource	Actions
Cluster	<ul style="list-style-type: none"> View the assigned Workload cluster on the Clusters page Rename the Workload cluster Delete the Workload cluster
Workloads	<ul style="list-style-type: none"> Create workloads View all workloads in the assigned cluster Update all workloads in the assigned cluster Delete all workloads in the assigned cluster
Queries	View all the queries in the assigned cluster
Jobs	View all the jobs in the assigned cluster

Cluster User access role

An authentic Cloudera Observability On-Premises user who is assigned the Cluster User access role has limited access rights across an assigned cluster within the Cloudera Observability On-Premises environment. Where they can view only those jobs and queries they created and executed in the assigned Workload cluster.

Table 3: Cluster User

Resource	Actions
Cluster	View their assigned cluster on the Clusters page.
Workloads	View their assigned workloads on the Workloads page
Queries	View their queries in the assigned cluster
Jobs	View their jobs in the assigned cluster

The Cluster User access role type has the most restricted access permissions, where the user may only view their own jobs and queries.

This access role further restricts the Cluster User to one cluster per policy. For users who are responsible for jobs and queries in more than one cluster they must also be assigned access rights to those clusters. You can either add them to the Cluster Policy for that cluster or include the pool that contains those workloads in the Cluster Policy in which they are assigned.

Also, for users who require access to jobs and queries executed by other users, you can create a Custom Policy as part of the Cluster Policy. This policy includes the user names of the users who execute those jobs and queries and/or the pool names in which they are executed.

For example, though user A and user B have been granted the same Cluster User role type their access to jobs and queries is different. This is due to the conditions of the Cluster Policy in which they are assigned. Where:

- The cluster policy that defines user A's Cluster User role type does not permit the user to view workloads within a pool or view other user workloads. In this case, user A is restricted to only view their own jobs and queries within their policy's assigned cluster.
- The cluster policy that defines user B's Cluster User role type contains a Custom Policy that permits the user to view workloads within a pool and view other user workloads. In this case, user B can view the jobs and queries executed by other users and the jobs and queries executed in the pool.

Understanding a Cloudera Observability On-Premises cluster policy

Describes the Cloudera Observability On-Premises Cluster Policy criteria that is used to assign Cloudera Observability On-Premises access roles to your users.

Access to your Workload jobs and queries is determined by a Cloudera Observability On-Premises Cluster Policy, which comprises two or more of the following conditions:

- One or more LDAP Group identifier account names.
- One or more user names. By default, Cloudera Observability On-Premises authenticates user access by checking that the user is a member of an LDAP group.
- A Cloudera Observability On-Premises access role type. The access role is assigned to the users that you provide in the Users field and/or the users who are part of the groups you provide in the Groups field and is defined by the conditions in the Cluster Policy.
- (Cluster User and Cluster Admin only) The cluster associated with the access role.
- (Cluster User only) A custom policy whose criteria is defined from the provided user names and/or the provided pools. A custom policy enables the user or users defined in the Cluster Policy to view the jobs and queries executed by other users and/or the jobs and queries executed in a pool.

Cloudera Observability On-Premises Cluster Policies are created, managed, and maintained from the Access Management page. Only users who have been granted the System Admin access role type can view and manage your Cloudera Observability On-Premises cluster policies.



Note: At this time, to access the Access Management page a manual edit in the URL is required. This is a temporary workaround for this known issue.

Configuring a default systems administrator for Cloudera Observability On-Premises

Pre-tasks that are required before you can start enabling role based access in Cloudera Observability On-Premises.

About this task

Describes how to enable role based access in Cloudera Observability On-Premises and configure a Cloudera Observability On-Premises default systems administrator.

Before you can assign access roles in Cloudera Observability On-Premises you must first enable role based access and configure a default systems administrator. Both tasks are completed in Cloudera Manager. Once configured, the default administrator (also known as a superuser) can log into the Cloudera Observability On-Premises UI and assign the System Admin access policy role to one or more users.

Procedure

1. In a supported web browser on the Cloudera Observability On-Premises on-premises cluster, log in to Cloudera Manager.
2. In Cloudera Manager, select Clusters, OBSERVABILITY, and then click the Configuration tab.
3. In the Configuration page, search for the Role Based Access enabled property and then select its OBSERVABILITY (Service-Wide) check box.

4. According to your requirements, do one of the following:
 - a. In the OBSERVABILITY (Service Wide) field of the OBSERVABILITY Default Super Users property, enter either the user name or the account name of a system administrator who is to be granted access to perform administration tasks in Cloudera Observability On-Premises. By default, admin.



Tip: If the OBSERVABILITY (Service Wide) field is not displayed, click the plus sign circle icon.

- b. In the OBSERVABILITY (Service Wide) field of the OBSERVABILITY Default Super Groups property, enter the group account name of your LDAP admin group. For example, admin_grp.

The following image shows the configuration properties:

5. Click Save Changes.
6. Navigate to the top of the Cloudera Observability On-Premises service page and from the Actions menu, restart the Cloudera Observability On-Premises service, by selecting Restart.

Assigning Cloudera Observability On-Premises access roles

Role based access to your Workload jobs and queries requires a Cloudera Observability On-Premises Cluster Policy that defines the conditions for the role based access type and assigns it to your users. You can have multiple Cluster Policies that define the access criteria for all of your workloads.



Note: At this time, to access the Access Management page a manual edit in the URL is required. This is a temporary workaround for this known issue.

Assigning a Cloudera Observability On-Premises system admin access role

Steps for assigning a System Admin access role to your Cloudera Observability On-Premises users.

About this task

Describes how to assign a Cloudera Observability On-Premises Role Based Access (RBAC) role for a system administrator. This access role has full access rights and system administrator privileges across all clusters within the Cloudera Observability On-Premises environment and can create your Cloudera Observability On-Premises Cluster Policies that define your access roles.



Note: Generally, only a user assigned the System Admin access role can create a Cloudera Observability On-Premises Cluster Policy. But until the first System Admin access role is assigned, a Cluster Policy can only be created by a default systems administrator, also known as a default super user.

Before you begin

This task assumes that you have:

- Enabled role based access in Cloudera Manager. For more information, click the Related Information link below.
- Created a default systems administrator, also known as a default super user, in Cloudera Manager.

Procedure

1. In a supported web browser log in to Cloudera Observability On-Premises as the user with default systems administrator privileges.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises landing page opens.

2. Verify that you have enabled role based access in Cloudera Manager. For more information, click the Related Information link below.

3. Open the Access Management feature, by entering access at the end of the Cloudera Observability On-Premises UI's URL. For example, `http://obs_console_server_hostname:port_number/observability/index.html#/access` Where,

- *obs_console_server_hostname* is the host name of the machine on which the Console server is installed in your Cloudera Observability On-Premises cluster.
- *port_number* is the port number for the Console Server.



Note: At this time, after the role based access property is enabled in Cloudera Manager the Access Management link is not appearing on the Cloudera Observability On-Premises UI's main navigation panel. This step's manual edit of the URL is a temporary workaround.

4. In the Access Management page, click New Cluster Policy.

The Create Cluster Policy page opens.

5. Do one or more of the following:

- a. In the Groups field, enter the name of the LDAP administration group account whose users will be assigned this cluster policy's access role.
- b. In the Users field, enter the user name or user names who will be assigned this cluster policy's access role.

6. From the Assign Roles list, select System Admin.

7. Click Create.



Note: Cloudera Observability On-Premises will take at least 60 minutes to assign the access role to the user, users, and/or groups provided in the Cluster Policy.

Results

The Successfully created access policy message appears when the Cluster Policy is created and the policy is displayed in the Access Management's home page.

Related Information

[Configuring a default systems administrator for Cloudera Observability On-Premises](#)

Assigning a Cloudera Observability On-Premises cluster admin access role

Steps for assigning a Cluster Admin access role to your Cloudera Observability On-Premises users.

About this task

Describes how to assign a Cloudera Observability On-Premises Role Based Access (RBAC) role for a cluster administrator.



Note: Only a user assigned the System Admin access role can create a Cloudera Observability On-Premises Cluster Policy.

Procedure

1. In a supported web browser log in to Cloudera Observability On-Premises as a user that has been granted the System Admin access role.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises landing page opens.

2. Verify that you have enabled role based access in Cloudera Manager. For more information, click the Related Information link below.
3. Open the Access Management feature, by entering access at the end of the Cloudera Observability On-Premises UI's URL. For example, `http://obs_console_server_hostname:port_number/observability/index.html#/access` Where,

- *obs_console_server_hostname* is the host name of the machine on which the Console server is installed in your Cloudera Observability On-Premises cluster.
- *port_number* is the port number for the Console Server.



Note: At this time, after the role based access property is enabled in Cloudera Manager the Access Management link is not appearing on the Cloudera Observability On-Premises UI's main navigation panel. This step's manual edit of the URL is a temporary workaround.

4. In the Access Management page, click New Cluster Policy.
The Create Cluster Policy page opens.
5. Do one or more of the following:
 - a. In the Groups field, enter the name of the LDAP group account whose users will be assigned this cluster policy's access role.
 - b. In the Users field, enter the user name or user names who will be assigned this cluster policy's access role.
6. From the Assign Roles list, select Cluster Admin.
7. From the Cluster list, select the name of the cluster that will be assigned to this policy's access role.
8. Click Create.



Note: Cloudera Observability On-Premises will take at least 60 minutes to assign the access role to the user, users, and/or groups provided in the Cluster Policy.

Results

The Successfully created access policy message appears when the Cluster Policy is created and the policy is displayed in the Access Management's home page.

Related Information

[Configuring a default systems administrator for Cloudera Observability On-Premises](#)

Assigning a Cloudera Observability On-Premises cluster user access role

Steps for assigning a Cluster User access role to your Cloudera Observability On-Premises users.

About this task

Describes how to assign a Cloudera Observability On-Premises Role Based Access (RBAC) role for a cluster user.



Note: Only a user assigned the System Admin access role can create a Cloudera Observability On-Premises Cluster Policy.

Procedure

1. In a supported web browser log in to Cloudera Observability On-Premises as a user that has been granted the System Admin access role.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.
The Cloudera Observability On-Premises landing page opens.
2. Verify that you have enabled role based access in Cloudera Manager. For more information, click the Related Information link below.
3. Open the Access Management feature, by entering access at the end of the Cloudera Observability On-Premises UI's URL. For example, `http://obs_console_server_hostname:port_number/observability/index.html#/access`
Where,
 - *obs_console_server_hostname* is the host name of the machine on which the Console server is installed in your Cloudera Observability On-Premises cluster.
 - *port_number* is the port number for the Console Server.



Note: At this time, after the role based access property is enabled in Cloudera Manager the Access Management link is not appearing on the Cloudera Observability On-Premises UI's main navigation panel. This step's manual edit of the URL is a temporary workaround.

4. In the Access Management page, click New Cluster Policy.
The Create Cluster Policy page opens.
5. Do one or more of the following:
 - a. In the Groups field, enter the name of the LDAP group account whose users will be assigned this cluster policy's access role.
 - b. In the Users field, enter the user name or user names who will be assigned this cluster policy's access role.
6. From the Assign Roles list, select Cluster User.
7. From the Cluster list, select the name of the cluster that will be assigned to this policy's access role.
8. Enable the user or users defined in this cluster policy to view executed workloads from other users or executed workloads from a pool by doing the following:
 - a. In the Users field, enter the user name or user names whose jobs and queries can be viewed by the user or users defined in this cluster policy.
 - b. In the Pools field, enter the pool name or pool names whose jobs and queries can be viewed by the user or users defined in this cluster policy.
9. Click Create.



Note: Cloudera Observability On-Premises will take at least 60 minutes to assign the access role to the user, users, and/or groups provided in the Cluster Policy.

Results

The Successfully created access policy message appears when the Cluster Policy is created and the policy is displayed in the Access Management's home page.

Related Information

[Configuring a default systems administrator for Cloudera Observability On-Premises](#)

Managing Your Cloudera Observability On-Premises access roles

Describes how to manage your Cloudera Observability On-Premises cluster policies and access roles.



Note: At this time, to access the Access Management page, enter access at the end of the Cloudera Observability On-Premises UI's URL. For example, `http://obs_console_server_hostname:port_number/observability/index.html#/access` This is a temporary workaround for this known issue.

Information about your Cloudera Observability On-Premises Cluster Policies are displayed on the Access Management page, which are viewed and managed by the user with the System Admin access role.

Each row displays a Cluster Policy and its conditions, where:

- The Status column displays the state of the policy, as either Enabled or Disabled.
- The Clusters column displays the name of the cluster assigned to the Cloudera Observability On-Premises access role.
- The Role column displays the Cloudera Observability On-Premises access role type.
- The Groups column displays the LDAP group users who are assigned the Cluster Policy's access role.
- The Users column displays the user names who are assigned the Cluster Policy's access role.
- The Custom Policy column displays the user and pool filter conditions.
- The Last Updated column displays the date when the policy was last updated.
- The Actions column's vertical ellipses, when selected, lists the management tasks that can be performed.

The following management tasks are performed from the Access Management home page by a user with the System Admin access role, which is accessed by selecting Access Management from the Cloudera Observability On-Premises Navigation side-bar.

Updating a cluster policy

In the Access Management page, click the cluster policy's vertical ellipsis in the Actions column, and select Edit. In the Cluster Policy, make your changes and then click Update.

Deleting a cluster policy

In the Access Management page, click the cluster policy's vertical ellipsis in the Actions column, and select Delete. In the confirmation message, click OK to confirm the action. The policy is permanently removed.



Tip: Cloudera recommends disabling rather than deleting a Cluster Policy.

Disabling a cluster policy

In the Access Management page, click the cluster policy's vertical ellipsis in the Actions column, and select Disable. In the confirmation message, click OK to confirm the action. The Status column displays the state of the policy as Disabled.

Purging HDFS data

Reduce bottlenecks between Telemetry Publisher and Cloudera Observability On-Premises, free up storage space, and increase job and query runtime efficiency by removing obsolete HDFS data that exceeds the maximum retention limit.



Note: Cloudera recommends performing regular purge events for HDFS files that are no longer required.

Understanding the purge date used by the purge event

Describes the Cloudera Observability On-Premises purge event's criteria that is based on the file's data group and the data group's retention limit and how the purge date is calculated.

The purge event's criteria is based on the maximum data retention policy, described in days, for the following HDFS data groups:

- Temporary data, when the retention period exceeds 8 days
- Staging data, when the retention period exceeds 31 days
- Detailed data, when the retention period exceeds 181 days
- Summarized data, when the retention period exceeds 731 days

The purge date is calculated by subtracting the retention days, specified by the maximum data retention period policy, from the current date and comparing the resultant date with the data's timestamp date. If the data's timestamp date is less than or equal to the resultant date the data is removed.

The data's timestamp date is determined by where the data resides:

- If the data resides in the cloudera-dbus root directory, the timestamp date is extracted from the subdirectory name. For example, if the directory name is /cloudera-dbus/HiveAudit/2021030623. The timestamp date extracted by the purge event is 2021/03/06, using the YYYY/MM/DD date format.



Important: The purge event deletes files from the cloudera-dbus directory as follows:

- If the date is successfully extracted and is less than or equal to the resultant date, all the files in the directory are removed and are counted as one file by the maximum deletion limit.
- If the date is successfully extracted, is less than or equal to the resultant date, and a file or files are set in the blobstore.purger.paths.to.keep parameter, all the files except the file or files set in the blobstore.purger.paths.to.keep parameter are removed and each file that is removed is counted by the maximum deletion limit.
- If the data resides in a cloudera-sigma-olap-impala, cloudera-sigma-partial-pse, cloudera-sigma-pse-extended, or cloudera-sigma-sdx-payloads root directory, the timestamp date is extracted from the file's last modified time.


Obsolete data can be purged from the following HDFS root directories:


- cloudera-dbus
- cloudera-sigma-olap-impala
- cloudera-sigma-partial-pse
- cloudera-sigma-pse-extended
- cloudera-sigma-sdx-payloads

Cloudera Observability On-Premises purge event parameters

Lists the Cloudera Observability On-Premises purge event parameter settings that enable you to set the event's execution time, frequency, and maximum purge duration. You can also exclude files and directories from being purged with the blobstore.purger.paths.to.keep parameter setting.

Table 4: Purge event parameters

Parameter	Description	Example
blobstore.purger.frequency	<p>The purge event's recurring schedule, based on one of the following values:</p> <ul style="list-style-type: none"> • None. By default, the purge process is set to none. • Daily. When this value is set for the first time, files are automatically deleted the next day at 1am. • Weekly. By default, files are automatically deleted every Saturday at 1am. • Monthly. When this value is set for the first time, files are automatically deleted the last Saturday of the month at 1am. Thereafter, files are deleted every 28th day. The monthly parameter uses the 28 day calendar format 	blobstore.purger.frequency = none
blobstore.purger.start.time	<p>The purge event's start time, based on the 24-hour time format. Where, 01:00 and 0:00 are valid time values, and 24:00, 1:0, and 01:0 are not valid time values</p> <p>By default, Cloudera Observability On-Premises schedules the purge process when it will cause the least amount of disruption to users.</p> <p> Note: Cloudera recommends scheduling a time during non-peak working hours or job execution hours.</p>	blobstore.purger.start.time = 01:00
blobstore.purger.paths.to.keep	<p>Lists the files and directories that are to be excluded from the purge event.</p> <p>Where each file and/or directory is separated by a comma and where:</p> <ul style="list-style-type: none"> • a file value must use its full path, directory name, and file name. • a directory value must use its full path and directory name. 	<pre>blobstore.purger.paths.to.keep= /cloudera-dbus/ImpalaQueryProfile/2021030217/7d2bcefa-8819-4fa1-be0c-4529ee4eb98f, /cloudera-dbus/HiveAudit, /cloudera-sigma-olap-impala/02f54999-b9a4-4dca-8237-d1b047755efb, /cloudera-sigma-sdx-payloads/2bc85719-7a3e-4438-96a4-8fc0f77ff</pre>

Parameter	Description	Example
blobstore.purger.delete.request.limit	<p>The maximum deletion limit.</p> <p>By default, the maximum number of files that can be deleted by the purge process is 500,000. This ensures that a purge cycle is not overloaded, does not introduce bugs, or takes up too much time.</p> <p>When the deletion limit is met, the purge process:</p> <ul style="list-style-type: none"> • Stops processing for a daily scheduled value. • Stops processing and restarts the next day for all other scheduled values. <p> Note: The purge event's maximum deletion limit calculates all the files in a dbus directory as one file. When you exclude a file or files that reside in the dbus directory from the purge process, the purge event's maximum deletion limit condition calculates all the files in the directory minus those files you have excluded.</p>	blobstore.purger.delete.request.limit=500000

Configuring the Cloudera Observability On-Premises purge event

Steps for scheduling and configuring a purge event.

About this task

Describes how to schedule and configure the Cloudera Observability On-Premises purge event.

Procedure

1. In a supported web browser, log in to Cloudera Manager as a user with full system administrator privileges.
2. From the Navigation panel, select Clusters and then OBSERVABILITY.
3. In the Status Summary panel of the OBSERVABILITY page, select Admin API Server.
4. Click the Configuration tab.
5. Search for the Admin API Server Advanced Configuration Snippet (Safety Valve) for the wxm-conf/sigmaadminapi.properites option.
6. In the text field enter your purge event's parameter settings, using the *Purge Event Parameters* table.

For example,

```
blobstore.purger.delete.request.limit=9990000
blobstore.purger.paths.to.keep=/cloudera-dbus/ImpalaQueryProfile/202103021
7/7d2bcefa-8819-4fa1-be0c-4529ee4eb98f,/cloudera-dbus/HiveAudit,/cloudera-
sigma-olap-impala/02f54999-b9a4-4dca-8237-d1b047755efb,/cloudera-sigma-sdx
-payloads/2bc85719-7a3e-4438-96a4-8fc0f77ff79e
blobstore.purger.frequency=daily
blobstore.purger.start.time = 0:00
```

7. Click Save Changes, which sets and schedules the purge process.
8. From the Actions menu, select Restart this Admin API Server.
9. In the Restart this Admin API Server message, confirm your changes by clicking Restart this Admin API Server.
10. When the Restart API Server step window displays Completed, click Close.

Manually executing a Cloudera Observability On-Premises purge event

You can manually run your purge event immediately with a one-time operation, rather than scheduling a purge event.

About this task

Describes how to manually run a Cloudera Observability On-Premises purge event.

A one-time purge event is based on the maximum data retention policy using the Cloudera Observability On-Premises purge event's parameter values, without the frequency value.

Procedure

1. In a supported web browser, log in to Cloudera Manager as a user with full system administrator privileges.
2. From the Navigation panel, select Clusters and then OBSERVABILITY.
3. In the Status Summary panel of the OBSERVABILITY page, select Admin API Server.
4. Click the Configuration tab.
5. Search for the Admin API Server Advanced Configuration Snippet (Safety Valve) for the `wxm-conf/sigmaadminapi.properites` option.
6. In the text field enter your purge event's parameter settings, using the *Purge Event Parameters* table.

For example,

```
blobstore.purger.delete.request.limit=9990000
blobstore.purger.paths.to.keep=/cloudera-dbus/ImpalaQueryProfile/202103021
7/7d2bcefa-8819-4fa1-be0c-4529ee4eb98f,/cloudera-dbus/HiveAudit,/cloudera-
sigma-olap-impala/02f54999-b9a4-4dca-8237-d1b047755efb,/cloudera-sigma-sdx
-payloads/2bc85719-7a3e-4438-96a4-8fc0f77ff79e
blobstore.purger.frequency=none
blobstore.purger.start.time = 0:00
```

7. Click Save Changes.
8. From the Actions menu, select Restart this Admin API Server.
9. In the Restart this Admin API Server message, confirm your changes by clicking Restart this Admin API Server.
10. When the Restart API Server step window displays Completed, click Close.
11. When a manual purge event run is required, do the following:
 - a) Log in to Cloudera Manager.
 - b) From the Navigation panel, select Clusters and then OBSERVABILITY.
 - c) From the Actions menu, select Purge HDFS Bucket Data.
 - d) In the Purge HDFS Bucket Data confirmation message, confirm the purge event by clicking Purge HDFS Bucket Data.
 - e) When the Purge HDFS Bucket Data window displays Completed, click Close.

Managing your Cloudera Observability On-Premises purge event

Steps for updating, stopping, and troubleshooting your Cloudera Observability On-Premises Purge event.

The following management tasks can be performed:

Updating your Cloudera Observability On-Premises purge event

To update your purge event:

1. In a supported web browser, log in to Cloudera Manager as a user with full system administrator privileges.
2. From the Navigation panel, select Clusters and then OBSERVABILITY.
3. From the Status Summary panel, select Admin API Server.
4. Click the Configuration tab.

5. Search for the Admin API Server Advanced Configuration Snippet (Safety Valve) for the wxm-conf/sigmaadminapi.properites option field.
6. In the text field, change the required values.
7. Click Save Changes.
8. From the Actions menu, select Restart this Admin API Server.
9. In the Restart this Admin API Server message, confirm your changes by clicking Restart this Admin API Server.
10. When the Restart API Server step window displays Completed, click Close.

Stopping the Cloudera Observability On-Premises purge event

You can stop a recurring purge event or stop a scheduled purge event whilst still running.

- To stop a recurring purge event:
 1. In a supported web browser, log in to Cloudera Manager as a user with full system administrator privileges.
 2. From the Navigation panel, select Clusters and then OBSERVABILITY.
 3. From the Status Summary panel, select Admin API Server.
 4. Click the Configuration tab.
 5. Search for the Admin API Server Advanced Configuration Snippet (Safety Valve) for the wxm-conf/sigmaadminapi.properites option field.
 6. In the text field, replace the blobstore.purger.frequency value with none.
 7. Click Save Changes.
 8. From the Actions menu, select Restart this Admin API Server.
 9. In the Restart this Admin API Server message, confirm your changes by clicking Restart this Admin API Server.
 10. When the Restart API Server step window displays Completed, click Close.
- To stop a scheduled purge event whilst still running:
 1. In a supported web browser, log in to Cloudera Manager as a user with full system administrator privileges.
 2. From the Navigation panel, select Clusters and then OBSERVABILITY.
 3. From the Status Summary panel, select Admin API Server.
 4. From the Actions menu, select Stop this Admin API Server.
 5. Still in the Admin API Server page, click the Configuration tab.
 6. Search for the Admin API Server Advanced Configuration Snippet (Safety Valve) for the wxm-conf/sigmaadminapi.properites option field.
 7. Replace the blobstore.purger.frequency value with none.
 8. Click Save Changes.
 9. From the Actions menu, select Restart this Admin API Server.
 10. In the Restart this Admin API Server message, confirm your changes by clicking Restart this Admin API Server.
 11. When the Restart API Server step window displays Completed, click Close.

Troubleshooting

The Cloudera Observability On-Premises purge event does not delete directories and files that do not have the full observability owner and file permissions. Files and directories may revert back to the hdfs owner when a restore is created from a snapshot. In this case and before creating an automatic or manual purge event you must verify the owner and file permissions of the required files to be purged.

To reset your HDFS files and directories as the observability owner with full administrative permissions do the following:

1. In a terminal go to the /etc directory and open the hdfs password file by entering:


```
vim passwd
```
2. Search for the kafka parameter.

3. Replace `/sbin/nologin` with `/bin/hash`.
4. Save the file.
5. Grant full observability access permissions to the `hdfs` password file by using the `chown` command.

Tracking your purge event from log entries

You can determine if the purge event was successful or identify potential problems from the Cloudera Manager Admin API Server log files.

The Admin API Server log file entries also list the names of the files and directories that were deleted and provide details about how many files and directories were deleted, the sum total size of the files and directories that were deleted, and the time they were deleted.

Working with alerts, costs, and reports

Certain Cloudera Observability On-Premises features enable you to define cost centers for planning, budgeting, and justifying resources and create alert actions that monitor your workloads and trigger a corrective action when applicable.

Analyzing your environment costs with Cloudera Observability On-Premises

The Cloudera Observability On-Premises 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 System Admin 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 On-Premises access roles, click the Related Information link below.

About the Cloudera Observability On-Premises Chargeback feature

The Cloudera Observability On-Premises 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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises Chargeback feature assumes the following:

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

Related Information

[Assigning access roles in Cloudera Observability On-Premises](#)

Configuring the Cloudera Observability On-Premises cost center criteria

The Cloudera Observability On-Premises Financial Governance Chargeback feature defines cost centers based on certain criteria. Configure your Cloudera Observability On-Premises 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 On-Premises 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 System Admin 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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises landing page opens.

2. From the Cloudera Observability On-Premises 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 On-Premises 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 On-Premises cost center

This topic describes the steps for creating a Cloudera Observability On-Premises 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 On-Premises cost center.



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



Note: Only users with the System Admin 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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.
The Cloudera Observability On-Premises landing page opens.
2. From the Cloudera Observability On-Premises 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 highlighted and its name is populated in the Environment Selection text field.



Note: For multiple clusters within an environment 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 On-Premises cost center.

About this task

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



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



Note: Only users with the System Admin 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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.
The Cloudera Observability On-Premises landing page opens.
2. From the Cloudera Observability On-Premises 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 On-Premises cost centers.

Displaying your costs associated with a cost center

When a Cloudera Observability On-Premises 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 breakdown 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 On-Premises 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 On-Premises 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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises landing page opens.

2. From the Cloudera Observability On-Premises Main navigation panel, select Financial Governance.

The Cloudera Observability On-Premises Chargeback page opens, which displays:

- The total cost and CPU and memory hourly usage for all of your cost centers including those uncategorized 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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises.

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 On-Premises UI.

Procedure

1. Verify that you are logged in to the Cloudera Observability On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises landing page opens.

2. From the Cloudera Observability On-Premises Main navigation panel, select Financial Governance.
The Cloudera Observability On-Premises 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 On-Premises 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 available for Classic Cluster and Private Cloud Base Cloudera Observability On-Premises environments only.

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 On-Premises 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 On-Premises to evaluate your auto action's definition.

For example, if Cloudera Observability On-Premises 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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises 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 available for Classic Cluster and Private Cloud Base Cloudera Observability On-Premises environments only.

Procedure

1. Verify that you are logged in to the Cloudera Observability On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.
The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.
 - d) From the **Environment Name** column in the Environment's table, select the environment required for analysis.
The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.
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 On-Premises 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 On-Premises.
- 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 Auto Actions feature in 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 On-Premises.

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 ...	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 ...	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 ...	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

The screenshot shows a dialog titled "Event Details" with a close button (X) in the top right corner. The main content is organized into two sections:

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

Below this is the "Auto Action Details" section:

Name	Action
auto-action-on-user	Kill the YARN process

Triggers and Criteria are listed at the bottom:

- Triggers: Duration > 15 minutes
- Criteria: User is any of hdfs, hive, admin sigma.unit NONE

A "Done" button is located in the bottom right corner of the dialog.

Figure 2: Update Event Type Example

The screenshot shows a dialog titled "Event Details" with a close button (X) in the top right corner. The main content is organized into two sections:

Event Type	Scope	Status
Update	Spark Application	Succeeded

Below this is the "Auto Action Details" section, which is presented as a comparison table:

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

A "Done" button is located in the bottom right corner of the dialog.

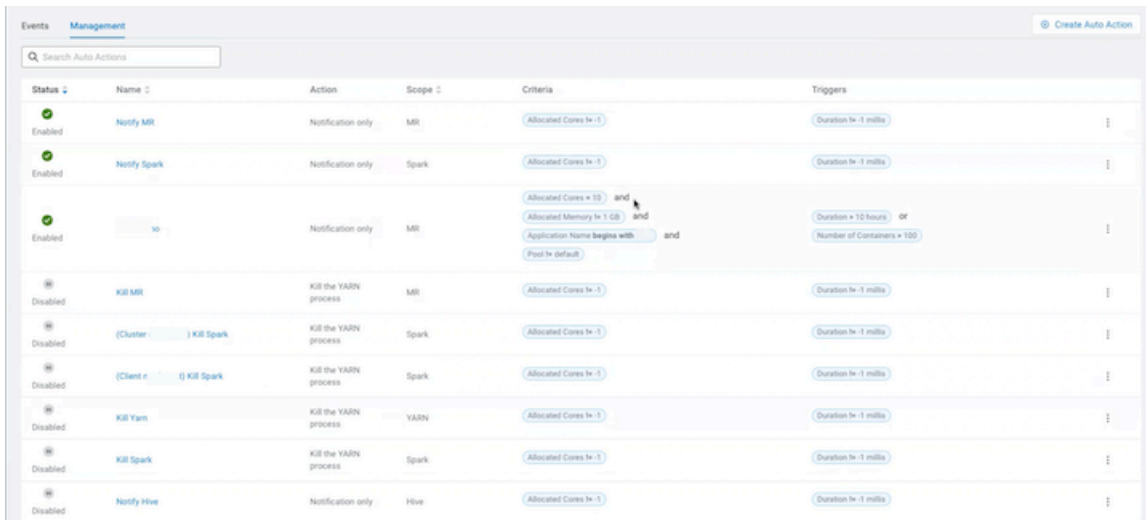
- Auto Action Name, which displays the unique name you entered for the auto action.
- Scope, which displays the workload component service that is monitored by the action.
- Time, which displays the time stamp of when the auto action's audit event occurred.



Important: By default, the Cloudera Observability On-Premises 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	Notify MR	Notification only	MR	Allocated Cores > 1	Duration > 1 min
Enabled	Notify Spark	Notification only	Spark	Allocated Cores > 1	Duration > 1 min
Enabled	to	Notification only	MR	Allocated Cores > 10 and Allocated Memory > 1 GB and Application Name begins with Pool < default	Duration > 10 hours or Number of Containers > 100
Disabled	Kill MR	Kill the YARN process	MR	Allocated Cores > 1	Duration > 1 min
Disabled	[Cluster] Kill Spark	Kill the YARN process	Spark	Allocated Cores > 1	Duration > 1 min
Disabled	[Client] Kill Spark	Kill the YARN process	Spark	Allocated Cores > 1	Duration > 1 min
Disabled	Kill Yarn	Kill the YARN process	YARN	Allocated Cores > 1	Duration > 1 min
Disabled	Kill Spark	Kill the YARN process	Spark	Allocated Cores > 1	Duration > 1 min
Disabled	Notify Hive	Notification only	Hive	Allocated Cores > 1	Duration > 1 min

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.

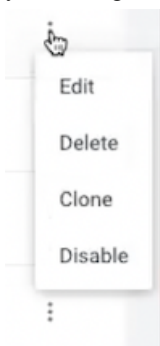
Managing your auto actions

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 On-Premises 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

Understanding, identifying, and addressing problems with Cloudera Observability On-Premises

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 On-Premises.

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 On-Premises displays workload data for the last 24 hours. If there is no data available during that time, Cloudera Observability On-Premises 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 On-Premises web UI.

- a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
- b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
- c) Click Log in.

The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.

- d) From the **Environment Name** column in the Environment's table, select the environment required for analysis.

The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.

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



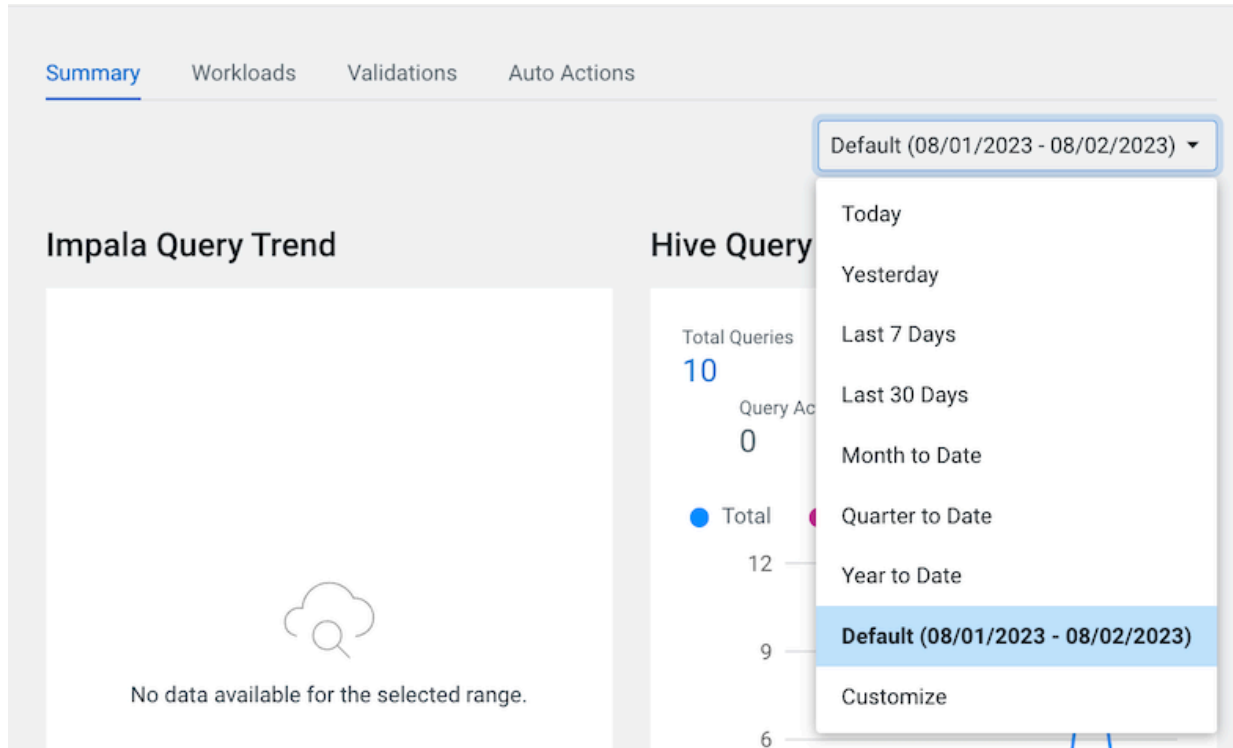
Tip: The page's title is displayed in the browser tab.

3. 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 Custom 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.

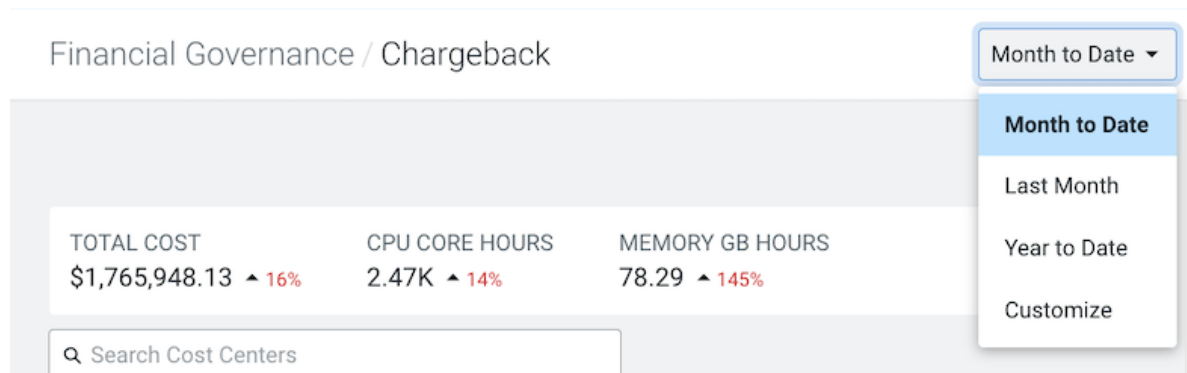
Results

All charts and tables in Cloudera Observability On-Premises 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



Exporting a report about your workload jobs and queries

You can save the job or query information displayed in the Jobs or Queries page of the Cloudera Observability On-Premises UI as a CSV formatted file 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 export reports about your workload Jobs and Queries in Cloudera Observability On-Premises.

About this task

Steps on how to export a report about your Spark, MapReduce, and Oozie Jobs, and Impala and Hive Queries.



Note: The downloaded CSV file contains raw data and as such may not display exactly as the format in the Cloudera Observability On-Premises UI. Also, in compliance with the Microsoft Excel row limit the maximum number of jobs exported is limited to one million. This limit also avoids exporting a Job or Query list that contains a huge number of jobs or queries.

Procedure

1. Verify that you are logged in to the Cloudera Observability On-Premises web UI.

- In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
- When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
- Click Log in.

The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.

- From the **Environment Name** column in the Environment's table, select the environment required for analysis.

The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.

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. From the time-range list in the Cluster Summary page, select a time period that meets your requirements.

4. In the Cluster Summary page, locate the Trend chart widget for the engine that processed your jobs or queries of interest and click its Total value. For example, locate the Hive Query Trend chart widget and click its Total Queries value.

Depending on the engine selected, the engine's Jobs or Queries page opens.

5. From the filters menu bar, click Export.

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, the data displayed in the engine's Jobs or Queries page is downloaded as a CSV formatted file to your Downloads folder.

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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises 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 On-Premises metastore analytics UI elements

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



Note: At this time, the Cloudera Observability On-Premises 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 On-Premises data temperatures

In Cloudera Observability On-Premises 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 On-Premises 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 5: 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 has 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: <ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> • 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 6: Columns

Parameter	Description
Column Name	Lists the Column field names.

Parameter	Description
Type	The Hive data type, as one of the following: <ul style="list-style-type: none"> • 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 7: 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 On-Premises 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 On-Premises 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 8: 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 9: 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 10: 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: <ul style="list-style-type: none"> • JDBC • LazySimple • Orc • Parquet

Column Name	Description
Table Type	<p>The table's type, such as but not limited to:</p> <ul style="list-style-type: none"> 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 On-Premises Metastore Analytics feature.



Note: At this time, the Cloudera Observability On-Premises 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

- Verify that you are logged in to the Cloudera Observability On-Premises web UI.
 - In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - Click Log in.

The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.
 - From the **Environment Name** column in the Environment's table, select the environment required for analysis.

The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.
- 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. 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.

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 On-Premises 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 On-Premises 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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.
 - d) From the **Environment Name** column in the Environment's table, select the environment required for analysis.

The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.
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. 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 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 On-Premises Hive cluster service metrics

The Cloudera Observability On-Premises 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 On-Premises 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 On-Premises 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 11: 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 On-Premises, 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 On-Premises, 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 On-Premises Hive cluster service metrics.

About this task

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



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 On-Premises web UI.

- a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
- b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
- c) Click Log in.

The Cloudera Observability On-Premises web user interface landing page opens, which by default displays the Analytics Environments page that lists your Workload cluster environments.

- d) From the **Environment Name** column in the Environment's table, select the environment required for analysis.

The Environments navigation panel opens, which hierarchically lists the environment's cluster, engines, and if applicable the Hive Metastore category.

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. 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 On-Premises 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-Premises 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 On-Premises.

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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises 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.

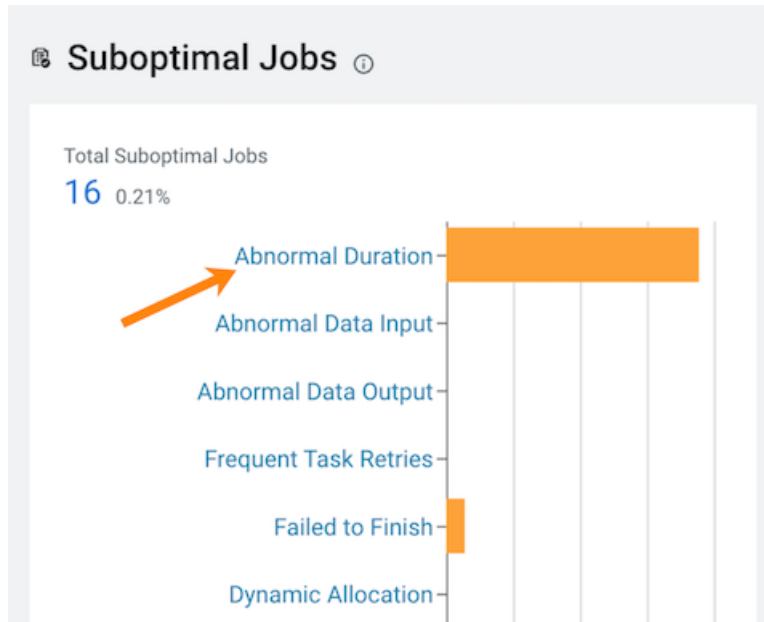
4. From the time-range list, select a time period that meets your requirements.
5. If not already expanded, from the Environment navigation panel expand the Virtual Cluster and then select the **Spark** engine.

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.

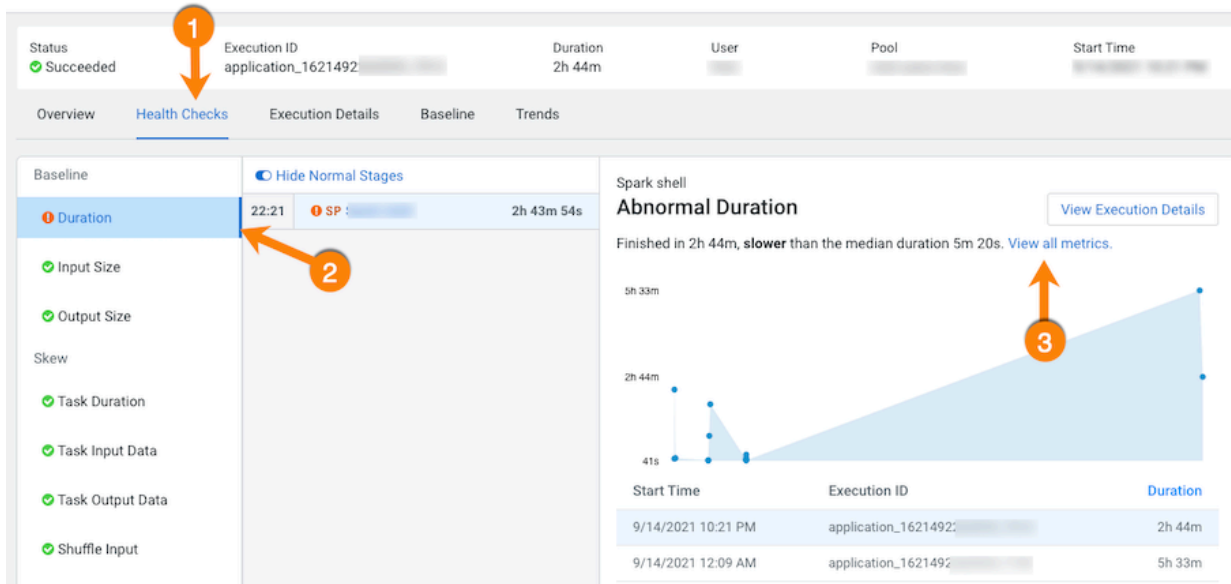
Type	Job	Status	Start Time	Duration	User	Health Check	Execution ID
SP	Cloudera: C...	✔ Succeeded	07/08/2021 3:29 AM PDT	15m 46s	psharma		application_1624
SP	Cloudera: C...	✔ Succeeded	07/08/	1m 7s	alanj		application_1624
SP	Cloudera: C...	✔ Succeeded	07/08/	1m 35s	alanj		application_1624
SP	Cloudera: 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

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.



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:

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	cloudera.com	41m 18s
Task 9	cloudera.com	25m 58s

- 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:

Task Details

Attempt	ID	Executor	Host	Start Time	Duration
0	8.0	5	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:

The screenshot shows the Cloudera Observability On-Premises interface. At the top, the status is 'Succeeded', the execution ID is 'application_162429...', the duration is '1h 2m', and the user is 'cmap'. Below this, there are tabs for 'Overview', 'Health Checks', 'Execution Details', 'Baseline', and 'Trends'. The 'Execution Details' tab is active. On the left, there is a list of jobs with their durations. The 'Summary' panel is displayed, showing 'Jobs Completed: 10', 'Stages Completed: 12', and 'Total: 10'. In the 'Other' section, there are links for 'Event Log', 'Summary | All Executors', and 'Configurations | SQL Executions | Metrics'. An orange arrow points to the 'Configurations' link with the text '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:

The screenshot shows the Cloudera Observability On-Premises interface. At the top, the status is 'Succeeded', the execution ID is 'application_1624293774839_37779', the duration is '1h 2m', and the user is 'cmap'. Below this, there are tabs for 'Overview', 'Health Checks', 'Execution Details', 'Baseline', and 'Trends'. The 'Execution Details' tab is active. On the left, there is a list of jobs with their durations. The 'Configurations' panel is displayed, showing 'JVM Information (0)', 'Spark Properties (2)', and 'Classpath Entries (2)'. The 'Spark Properties' tab is selected. A search box contains the word 'memory'. An orange arrow points to the search box with the text 'Search for a specific configuration by entering its property name'. Below the search box, there is a table with the following data:

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 On-Premises.

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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises 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.

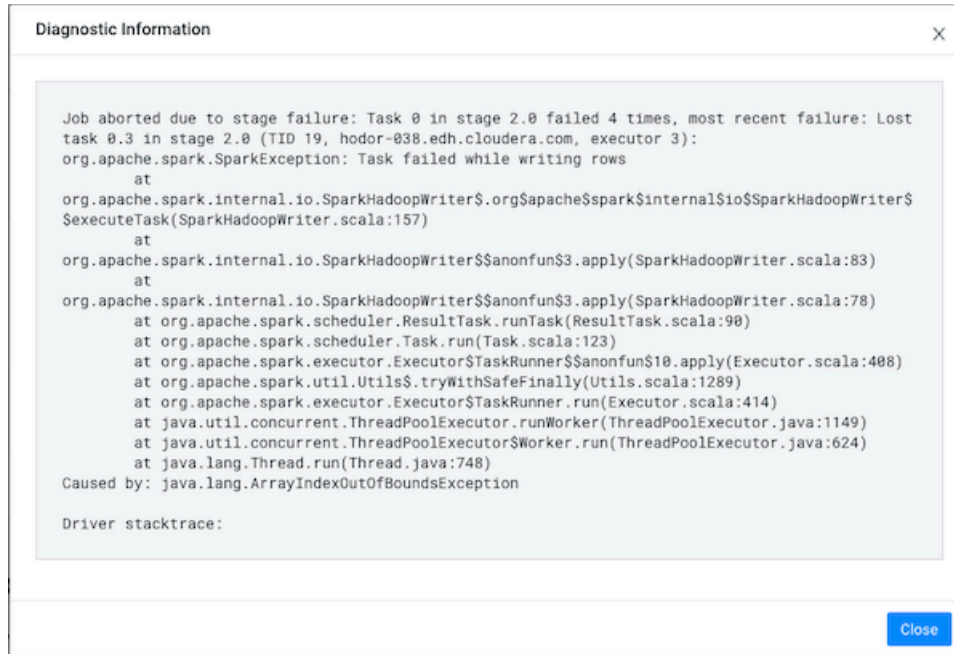
6. From the Health Check filter's list, select Failed to Finish, which filters the list to display a list of jobs that did not complete.
7. To view more details about why a job failed to complete, from the Job column select a job's name. The job's page opens displaying information about the job you selected and where the failure happened.

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): org.apache.spark.SparkException: Task failed while writing rows at org.apache.spa... + More

For more information, click +More

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

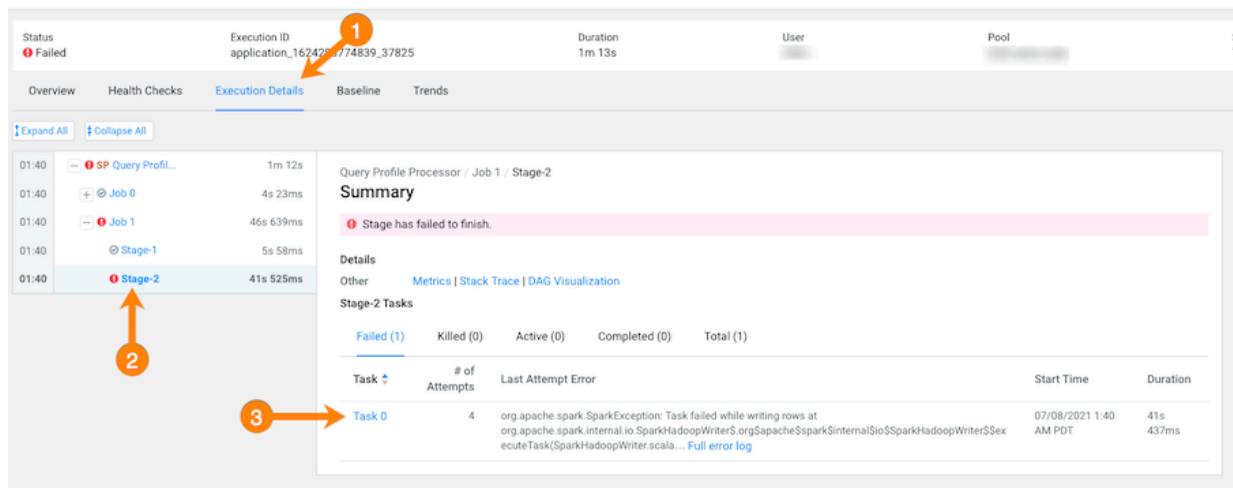
The Diagnostic Information dialog box opens, which describes more details about why the job aborted. In the following example, the job was aborted whilst writing rows due to an out of bounds java exception:



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:



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.

The screenshot displays the Cloudera Observability On-Premises interface. At the top, the status is 'Failed', the execution ID is 'application_1624293774839_37825', and the duration is '1m 13s'. The 'Execution Details' tab is active, showing a tree view of the execution plan. Under 'Stage-2', 'Task 0' is highlighted. A table below shows three failed attempts of Task 0, each with a duration of approximately 12-13 seconds. The error message for each attempt is 'org.apache.spark.SparkException: Task failed while writing rows'. The full error log is expanded for each attempt, showing the exception details and a link to the full error log.

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 On-Premises.

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 On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cloudera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cloudera Observability On-Premises Log in page opens, enter your Cloudera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cloudera Observability On-Premises 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.

- 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.

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

- From the Job Trend widget, click its Total Jobs value.

The engine's Jobs page opens.

- 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.

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

- 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:

The screenshot displays the Cloudera Observability On-Premises interface. At the top, the status is 'Succeeded', the execution ID is 'application_1624293774839_34746', and the duration is '2h 56m'. The 'Health Checks' tab is active, showing a list of checks on the left. The 'Task Wait Time' check is selected, opening a detailed view for 'Stage-5 Long Task Wait Time'. The description states that a healthy status indicates successful tasks took less than 15 minutes to complete and less than 40% of the total task duration time to start. The diagnosis notes 'resource starvation' for 1 out of 22 tasks. A chart shows a median wait time of 0s and an outlier at 34m 2s. The recommendation section provides action items: 'Allocate more executors for the application by configuring spark.executor.instances to a higher value.' and '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.' Below, a table lists '1 Outlier Task' with columns for Task Name, Host, and Wait Duration. The outlier task is 'Task 4' on host 'cloudera.com' with a wait duration of 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:

The screenshot displays the Cludera Observability On-Premises interface for a job execution. The top bar shows the status as 'Succeeded', Execution ID as 'application_1624293774839_34746', Duration as '2h 56m', and the User and Pool. The main content area is divided into several sections:

- Baseline:** A list of metrics including Duration (2h 55m 49s), Input Size (1h 38m 14s), Output Size (1h 16m 43s), Skew (1h 16m 36s), Task Duration (1h 38m 8s), Task Input Data, Task Output Data, Shuffle Input, Data Processing Speed, Task Wait Time (1h 38m 8s), Task GC Time, Task Retries, RDD Caching, and Executor Memory.
- Task Details:** A table showing task attempts. The first attempt (Attempt 0) failed with a duration of 33m 37s. The error message is: '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: .cludera.com. Exit- Full error log'. A second attempt (Attempt 1) succeeded with a duration of 1h 4m.
- Task Metrics:** A table summarizing various metrics for the task. Two rows are highlighted with orange boxes: 'Wait Duration' (34m 2s, Average 1m 33s) and 'Successful Attempt Duration' (1h 4m, Average 7m 18s).

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 Cludera Observability On-Premises web UI.
 - a) In the URL field of a supported web browser, enter the Cludera Observability On-Premises URL that you were given by your system administrator and press Enter.
 - b) When the Cludera Observability On-Premises Log in page opens, enter your Cludera Observability On-Premises user name and password access credentials.
 - c) Click Log in.

The Cludera Observability On-Premises landing page opens.

- 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.

- 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.

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

- From the Job Trend widget, click its Total Jobs value.

The engine's Jobs page opens.

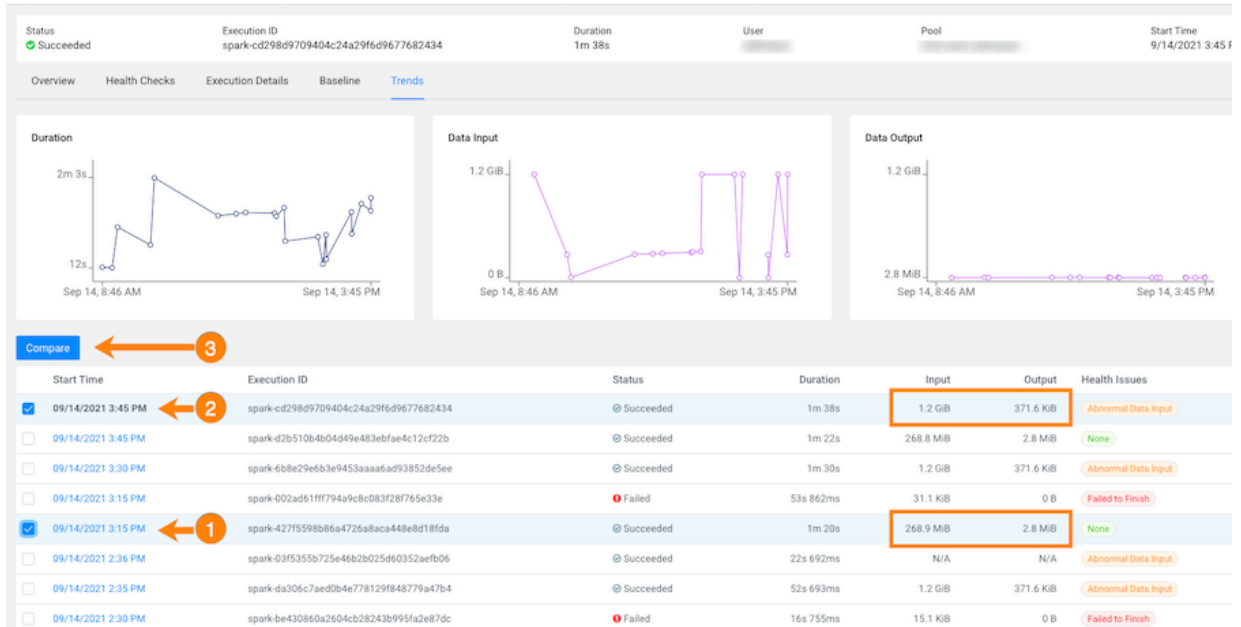
- Examine the list of jobs that have executed during the selected time period and manually compare runs of the same job.

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.

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

- 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:



- 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

- Basic
- Structure
- Configurations
- SQL Executions
- Metrics

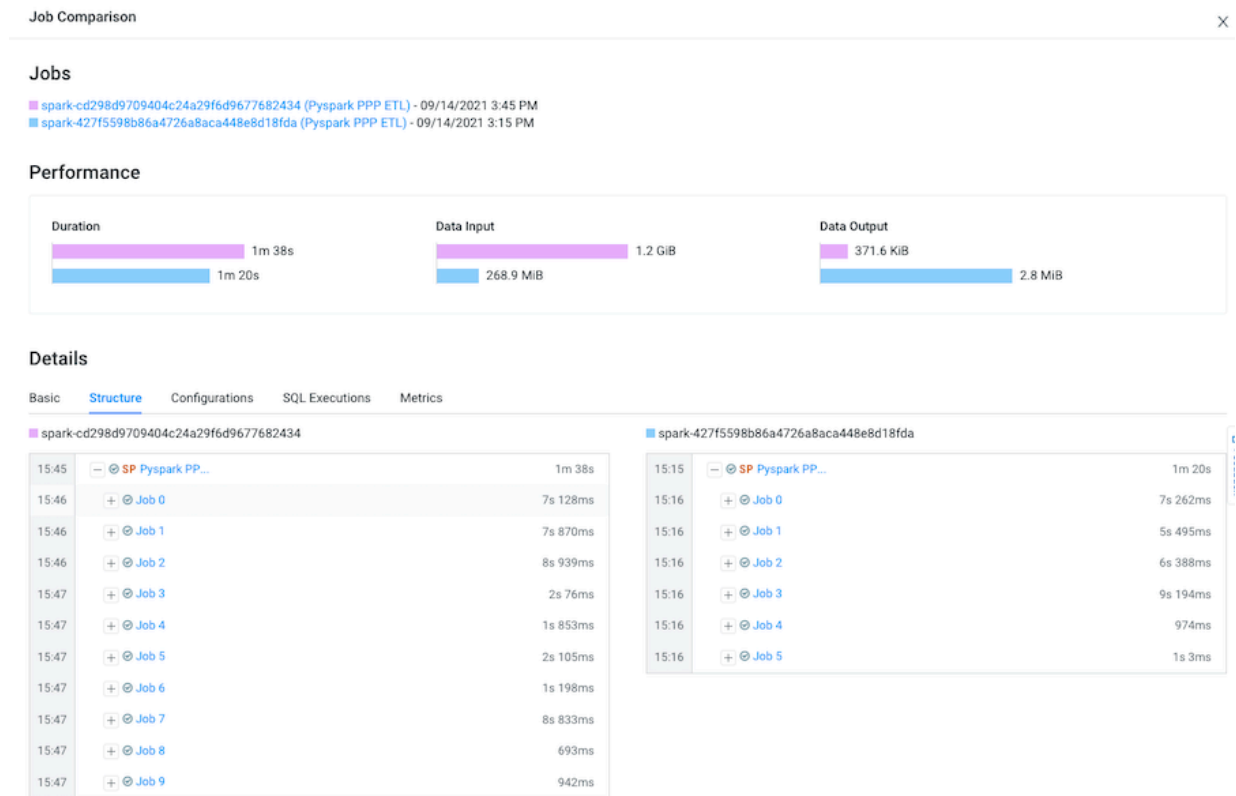
	spark-cd298d9709404c24a29f6d9677682434	spark-427f5598b86a4726a8aca448e8d18fda
Name	Pyspark PPP ETL	Pyspark PPP ETL
Type	Spark	Spark
Start Time	09/14/2021 3:45 PM	09/14/2021 3:15 PM
Status	Succeeded	Succeeded
Health Issues	Abnormal Data Input	None
Duration	1m 38s	1m 20s
Data Input	1.2 GiB	268.9 MiB
Data 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.

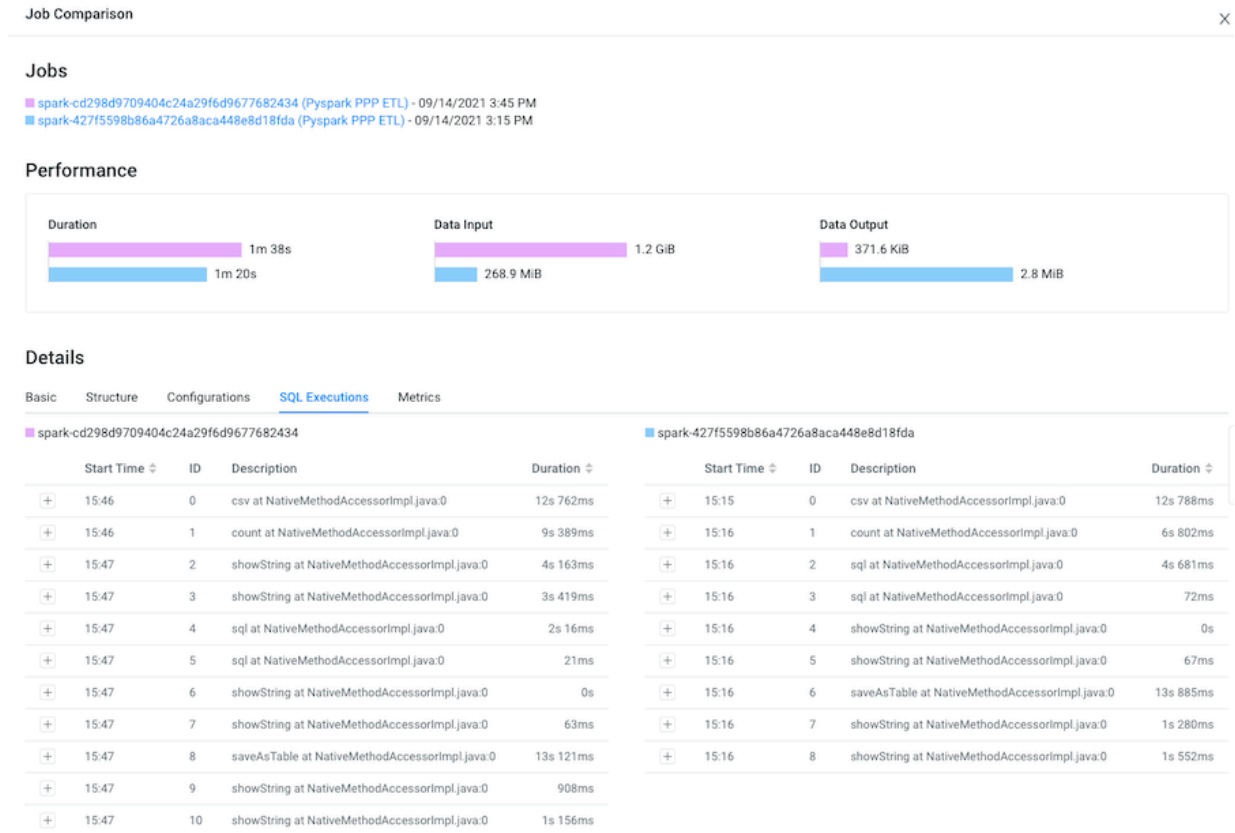
10. Display and compare the sub-jobs executed for both of your selected job runs by selecting the Structure tab.

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.



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.



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

- Basic
- Structure
- Configurations
- SQL Executions
- Metrics**

	spark-cd298d9709404c24a29f6d9677682434	spark-427f5598b86a4726a8aca448e8d18fda	
<u>Ungrouped</u>			
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