Cloudera Manager 7.9.5

Monitoring and Diagnostics

Date published: 2020-11-30
Date modified: 2023-01-25

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

Accessing the Cloudera Manager Admin Console 9

Monitoring and Diagnostics 9

Time Line 9
- Selecting a Point In Time or a Time Range 10

Health Tests 11
- Viewing Health Test Results 12
- Suppressing Health Test Results 12
- Suppressing a Health Test 12
- Configuring Suppression of Health Tests Before Tests Run 13
- Viewing a List of Suppressed Health Tests 13
- Unsuppressing Health Tests 14

Viewing Charts for Cluster, Service, Role, and Host Instances 14
- Exporting Data from Charts 15
- Adding and Removing Charts from a Dashboard 15
- Creating Triggers from Charts 16

Configuring Monitoring Settings 16
- Configuring Health Monitoring 17
- Configuring Service Monitoring 17
- Configuring Host Monitoring 17
- Configuring Directory Monitoring 18
- Configuring YARN Application Monitoring 18
- Configuring Impala Query Monitoring 18
- Configuring Impala Query Data Store Maximum Size 19
- Enabling Configuration Change Alerts 19
- Filtering Metrics 20
- Configuring Log Events 21
  - Configuring Logs 22
  - Configuring Logging Thresholds 22
  - Configuring Log Directories 22
  - Enabling and Disabling Log Event Capture 23
  - Configuring Which Log Messages Become Events 23
  - Configuring Log Alerts 24

Monitoring Clusters 24

Cluster Utilization Report overview 26
- Enable the Cluster Utilization Report 27

Configure the Cluster Utilization Report....28
Use the Cluster Utilization Report to manage resources....28
Overview Tab....29
YARN Tab....30
Impala Tab....32
Download the Cluster Utilization Report....33
Creating a Custom Cluster Utilization Report....34
Metrics and queries....34
Impala query counter metrics....42
Calculations for reports....43
Retrieving metric data....43
Querying metric data....45

Inspecting Network Performance....45
Running the Network Performance Inspector From the Cloudera Manager Admin Console....46
Running the Network Performance Inspector From the Cloudera Manager API....48

Monitoring Services....48
Monitoring Service Status....48
Viewing the URLs of the Client Configuration Files....49
Viewing the Status of a Service Instance....49
Viewing the Health and Status of a Role Instance....49
Viewing the Maintenance Mode Status of a Cluster....49
Viewing Service Status....50
Viewing Past Status....50
Status Summary....51
Service Summary....52
Health Tests and Health History....52
Viewing Service Instance Details....53
Role Instance Reference....53
Viewing Role Instance Status....54
The Actions Menu....55
Viewing Past Status....55
Summary....55
Health Tests and Health History....55
Status Summary....56
Charts....56
The Processes Tab....56
Running Diagnostic Commands for Roles....56
Periodic Stacks Collection....57
Configuring Periodic Stacks Collection....57
Viewing and Downloading Stacks Logs....58
Viewing Running and Recent Commands....58
Viewing Running and Recent Commands For a Cluster....58
Viewing Running and Recent Commands for a Service or Role....59
Command Details....59

Monitoring Hosts....60
Viewing All Hosts....60
Role Assignments....61
Viewing the Disks Overview....61
Viewing the Hosts in a Cluster....62
Viewing Individual Hosts....62

Host Details....62
Viewing Host Details....62
Status....63
Processes....64
Resources....64
Commands....64
Configuration....65
Components....65
Audits....65
Charts Library....65
Host Inspector....65
Running the Host Inspector....66
Viewing Past Host Inspector Results....66

Monitoring Activities....66
Selecting Columns to Show in the Activities List....67
Sorting the Activities List....67
Filtering the Activities List....67
Activity Charts....68
Viewing the Jobs in a Pig, Oozie, or Hive Activity....68
Task Attempts....69
Viewing a Job's Task Attempts....69
Selecting Columns to Show in the Tasks List....70
Sorting the Tasks List....70
Filtering the Tasks List....70
Viewing Activity Details in a Report Format....71
Comparing Similar Activities....71
Viewing the Distribution of Task Attempts....72
The Task Distribution Chart....72
TaskTracker Hosts....72
Monitoring Impala Queries....73
Viewing Queries....73
Configuring Impala Query Monitoring....73
Impala Best Practices....74
Results Tab....74
Filtering Queries....75
Filter Expressions....75
Filter Attributes....76
Choosing and Running a Filter....80
Query Details....82
Monitoring YARN Applications....83
Viewing Jobs....83
Configuring YARN Application Monitoring....83
Results Tab....84
Filtering Jobs....85
Filter Expressions....85
Choosing and Running a Filter....85
Filter Attributes....87
Sending Diagnostic Data to Cloudera for YARN Applications....94
Monitoring Spark Applications....95
Viewing and Debugging Spark Applications Using Logs....95
Managing Spark Driver Logs....95
Visualizing Spark Applications Using the Web Application UI....96
Accessing the Web UI of a Running Spark Application....96
Accessing the Web UI of a Completed Spark Application....96

Events **103**
Viewing Events 104
Filtering Events 104
Adding an Event Filter 105
Removing an Event Filter 105

Alerts **105**
Managing Alerts 105
Configuring Alert Email Delivery 106
Configuring Alert SNMP Delivery 107
Configuring Custom Alert Scripts 108
Sample Custom Alert Script 114
Alert email routing 118
Configuring user profiles for email alerts 119
Configuring host groups for email alerts 119
Configuring service groups for email alerts 120
Configuring role groups for email alerts 120
Configuring alert rules for email alerts 121
Enabling Configuration Change Alerts 122
Enabling HBase Alerts 122
Enabling Health Alerts 123
Modifying the Health Threshold 123
Configuring Alerts Transitioning Out of Alerting Health Threshold 124
Configuring Log Alerts 124
Configuring Alert Delivery 124

Triggers **124**
Creating a Trigger Using the Expression Editor 126
Editing, Deleting, Suppressing, or Deleting a Trigger 128
Cloudera Manager Trigger Use Cases 128
Creating a Trigger for Memory Capacity 128
Creating a Trigger for CPU Capacity 129

Lifecycle and Security Auditing **130**
Viewing Audit Events 131
Filtering Audit Events 131
Adding a Filter 132
Removing a Filter 132
Downloading Audit Events 132

Charting Time-Series Data **133**
Terminology 133
Building a Chart with Time-Series Data 133
Configuring Time-Series Query Results 134
Using Context-Sensitive Variables in Charts 135
Chart Properties 135
Changing the Chart Type 136
Grouping (Faceting) Time Series 138
Displaying Chart Details 138
Editing a Chart 141

Saving a Chart.... 141
Obtaining Time-Series Data Using the API.... 142
Dashboards.... 142
Dashboard Types.... 142
Creating a Dashboard.... 143
Managing Dashboards.... 143
Configuring Dashboards.... 144
Saving Charts to Dashboards.... 144
Saving Charts to a New Dashboard.... 144
Saving Charts to an Existing Dashboard.... 144
Adding a New Chart to the Custom Dashboard.... 145
Removing a Chart from a Custom Dashboard.... 145
Moving and Resizing Charts.... 145
tsquery Language.... 146
tsquery Syntax.... 146
Metric Expressions.... 147
Metric Expression Functions.... 147
Predicates.... 149
Discovering possible predicates.... 150
Filtering by Day of Week or Hour of Day.... 151
Time Series Attributes.... 152
Time Series Entities and their Attributes.... 153
FAQ.... 154
Metric Aggregation.... 156
Presentation of Aggregate Data.... 158
Accessing Aggregate Statistics Through tsquery.... 159
Filtering Metrics.... 160

Logs.... 161
Viewing Logs.... 162
Logs List.... 162
Filtering Logs.... 162
Log Details.... 163
Viewing the Cloudera Manager Server Log.... 163
Viewing Cloudera Manager Server Logs in the Logs Page.... 163
Viewing the Cloudera Manager Server Log.... 164
Viewing the Cloudera Manager Agent Logs.... 164
Viewing Cloudera Manager Agent Logs in the Logs Page.... 164
Viewing the Cloudera Manager Agent Log.... 164
Managing Disk Space for Log Files.... 165
Disk Space Requirements.... 165
Managing Log Files.... 165

Reports.... 165
Directory Usage Report.... 166
Accessing the Directory Usage Report.... 166
Using the Directory Usage Report.... 166
Disk Usage Reports.... 168
Viewing Current Disk Usage by User, Group, or Directory.... 169
Viewing Historical Disk Usage by User, Group, or Directory.... 169
Downloading Reports as CSV and XLS Files.... 170
Activity, Application, and Query Reports.... 170
The File Browser.... 170
Searching Within the File System.... 171

Enabling Snapshots....171
Setting Quotas....171
Designating Directories to Include in Disk Usage Reports....171
Downloading HDFS Directory Access Permission Reports....171

Cluster Support Tokens using Cloudera Manager....172

Sending Usage and Diagnostic Data to Cloudera....173
Configuring a Proxy Server....173
Managing Anonymous Usage Data Collection....174
Diagnostic Data Collection....174
Log support in Cloudera Manager for ECS cluster....175
Configuring the Frequency of Diagnostic Data Collection....176
Configuring collection of Cloudera Manager table data....176
Specifying the Diagnostic Data Directory....177
Redaction of Sensitive Information from Diagnostic Bundles....177
Disabling the Automatic Sending of Diagnostic Data from a Manually Triggered Collection....178
Manually Triggering Collection and Transfer of Diagnostic Data to Cloudera....178

Troubleshooting Cluster Configuration and Operation....179
Solutions to Common Problems....179
Logs and Events....181

Accessing the Cloudera Manager Admin Console

How to open and log in to the Cloudera Manager Admin Console.

About this task

To access the Cloudera Manager Admin Console:

Procedure

1. Open the Cloudera Management Console in a Web browser using the following URL:

```
http://<Cloudera Manager Server URL>:7180
```

2. Enter your user name.
3. Enter your password.
4. Click the Sign In button.

Results

The Cloudera Manager Admin Console opens.

Monitoring and Diagnostics

This section is for system administrators who want to use Cloudera Manager to monitor and diagnose their cluster. You can use the Cloudera Manager Admin Console to monitor cluster health, metrics, and usage, view processing activities, and view events, logs, and reports to troubleshoot problems and monitor compliance.

Time Line

The Time Line allows you to view status and health information for a specific point in time or across a range of time. The Time Line appears on many pages in Cloudera Manager.

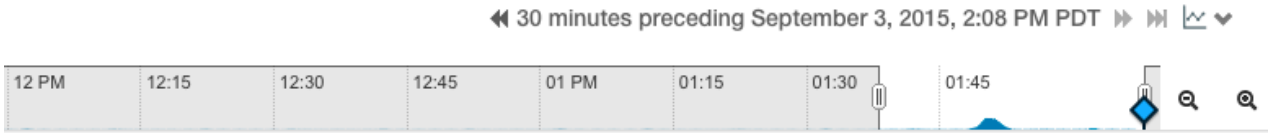
When you view the top level service and Hosts tabs, the Time Line shows status and health only for a specific point in time. When you are viewing the Logs and Events tabs, and when you are viewing the Status, Commands, Audits, Jobs, Applications, and Queries pages of individual services, roles, and hosts, the Time Line appears as a Time Range Selector, which lets you highlight a range of time over which to view historical data.

Click the (📈) icon at the far right to turn on and turn off the display of the Time Line.

Cloudera Manager displays timestamped data using the time zone of the host where Cloudera Manager server is running. The time zone information can be found under the Support About menu.

The background chart in the Time Line shows the percentage of CPU utilization on all hosts in the cluster, updated at approximately one-minute intervals, depending on the total visible time range. You can use this graph to identify periods of activity that may be of interest.

In the pages that support a time range selection, the area between the handles shows the selected time range.



There are a variety of ways to change the time range in this mode.

The Reports screen (Clusters Reports) does not support the Time Range Selector: the historical reports accessed from the Reports screen have their own time range selection mechanism.

Use the Zoom In and Zoom Out buttons (and) to zoom the time line graph in or out.

- Zoom In shows a shorter time period with more detailed interval segments. Zooming does not change your selected time range. However, the ability to zoom the Time Line can make it easier to use the selector to highlight a time range.
- Zoom Out lets you show a longer time period on the time range graph (with correspondingly less granular segmentation).

Selecting a Point In Time or a Time Range

Depending on what page the Time Line appears, you can select a point in time or a time range.

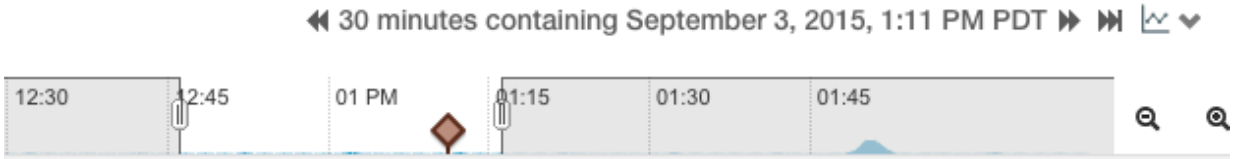
There are two ways to look at information about your cluster—its current status and health, or its status and health at some point (or during some interval) in the past. When you are looking at a point in the past, some functions may not be available. For example, on a Service Status page, the Actions menu (where you can take actions like stopping, starting, or restarting services or roles) is accessible only when you are looking at Current status.

Selecting a Point in Time

Status information on pages such as the service Status pages, reflects the state at a single point in time (a snapshot of the health and status). When displayed data is from a single point in time (a snapshot), the panel or column displays a small version of the Time Marker icon ( ) in the panel. This indicates that the data corresponds to the time at the location of the Time Marker on the Time Line.

By default, the status is shown at the current time. If you specify an earlier point on the time range graph, you see the status as it was at the selected point in the past.

- When the Time Marker is set to the current time, it is blue ( ).
- When the Time Marker is set to a time in the past, it is orange ( ).



You can select the point in time in one of the following ways:

- By moving the Time Marker ( )
- When the Time Marker is set to a past time, you can quickly switch back to view the current time using the Now button ( ).
- By clicking the date, choosing the date and time, and clicking Apply.

Selecting a Time Range

Pages such as the Logs, Events, and Activities show data over a time range rather than at a single point. These default to showing the past 30 minutes of data (ending at the current time). The charts that appear on the individual Service Status and Host Status pages also show data over a time range. For this type of display, there are several ways to select a time range of interest:

- Drag one (or both) edges of the time range handles to expand or contract the range.

- Choose a duration by clicking a duration link 30m 1h 2h 6h 12h 1d 7d 30d and then do one of the following:
  - Click the next ⏭ or ⏮ previous buttons to select the next or previous duration.
  - Click somewhere in the dark portion of the time range to choose the selected duration.
- Click the date range

  ⏮ 30 minutes containing September 3, 2015, 1:11 PM PDT ⏭ ⏭ 📈 ⌄

  to open the time selection widget. Enter a start and end time and click Apply to put your choice into effect.
- When you are under the Clusters tab with an individual activity selected, a Zoom to Duration button is available. This lets you zoom the time selection to include just the time range that corresponds to the duration of your selected activity.

Health Tests

Cloudera Manager monitors the health of the services, roles, and hosts that are running in your clusters using *health tests*.

The Cloudera Management Service also provides health tests for its roles. Role-based health tests are enabled by default. For example, a simple health test is whether there's enough disk space in every NameNode data directory. A more complicated health test may evaluate when the last checkpoint for HDFS was compared to a threshold or whether a DataNode is connected to a NameNode. Some of these health tests also aggregate other health tests: in a distributed system like HDFS, it's normal to have a few DataNodes down (assuming you've got dozens of hosts), so we allow for setting thresholds on what percentage of hosts should color the entire service down.

Health tests can return one of three values: Good, Concerning, and Bad. A test returns Concerning health if the test falls below a warning threshold. A test returns Bad if the test falls below a critical threshold. The overall health of a service or role instance is a roll-up of its health tests. If any health test is Concerning (but none are Bad) the role's or service's health is Concerning; if any health test is Bad, the service's or role's health is Bad.

In the Cloudera Manager Admin Console, health tests results are indicated with colors: Good ✅, Concerning 🟠, and Bad ❗.

There are two types of health tests:

- Pass-fail tests - there are two types:
  - Compare a property to a yes-no value. For example, whether a service or role started as expected, a DataNode is connected to its NameNode, or a TaskTracker is (or is not) blocked.
  - Exercise a service lightly to confirm it is working and responsive. HDFS (NameNode role), HBase, and ZooKeeper services perform these tests, which are referred to as "canary" tests.

  Both types of pass-fail tests result in the health reported as being either Good or Bad.
- Metric tests - compare a property to a numeric value. For example, the number of file descriptors in use, the amount of disk space used or free, how much time spent in garbage collection, or how many pages were swapped to disk in the previous 15 minutes. In these tests the property is compared to a threshold that determine whether everything is Good, (for example, plenty of disk space available), whether it is Concerning (disk space getting low), or is Bad (a critically low amount of disk space).

By default most health tests are enabled and (if appropriate) configured with reasonable thresholds. You can modify threshold values by editing the monitoring properties under the entity's Configuration tab. You can also enable or disable individual or summary health tests, and in some cases specify what should be included in the calculation of overall health for the service, role instance, or host.

Related Information

Configuring Monitoring Settings
Health Test Reference

Viewing Health Test Results

You can view health test results in multiple locations.

Health test results are available in the following locations:

- Home Status tab where various health results determine an overall health assessment of the service or role. The overall health of a role or service is a roll-up of its health tests; if any health test is Bad, the service's or role's health will be Bad. If any health test is Concerning (but none are Bad) the role's or service's health will be Concerning.
- Hosts tab, which shows summary result for the hosts.
- Status tab - which shows metrics for services, role instances, and hosts. These are reflected in the results shown in the Health Tests panel when you have selected a service, role instance, or host.
- The All Health Issues tab of the Home page displays all health issues. You can sort the display by entity or by Health Test.

For some health test results, you can chart the associated metrics over a time range. See the related information for details.

Related Information

Viewing Service Status
Viewing Role Instance Status
Host Details

Suppressing Health Test Results

Cloudera Manager displays warnings when health tests indicate a problem in the cluster. Sometimes these warnings are expected or do not indicate a real problem in your deployment. You can suppress display of these warnings in Cloudera Manager.

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

You can suppress health test warnings as they appear or before any tests run. Suppressed health tests are hidden in Cloudera Manager and their status does not affect the roll-up of health tests that display for a service, host, or role instance. Suppressed health test warnings remain available in Cloudera Manager, and the tests continue to run but the results are hidden. You can unsuppress a suppressed health test at any time.

On pages where you have suppressed validations, you will see a link that says Show # Suppressed Test. On this screen, you can:

- Click the Show # Suppressed Test link to view all suppressed health tests for the page.
- Click the Unsuppress... link to unsuppress the health test.
- Click Hide Suppressed Tests to re-hide the suppressed tests.

Note: Suppressing a health test is different than disabling a health test. A disabled health test never runs, whereas a suppressed health test runs but its results are hidden.

Suppressing a Health Test

You can suppress a health test to hide its results.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Go to the health test you want to suppress.
2. Click the Suppress... link to the right of the health test description.
   A dialog box opens where you can enter a comment about the suppression action.
3. Click Confirm.
   The display changes to Suppressing... while the change is propagated.

Related Information

Viewing Health Test Results

Configuring Suppression of Health Tests Before Tests Run

You can configure a health test to suppress its results before the health test runs.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Go to the service or host with the health test that you want to suppress.
2. Click the Configuration tab.
3. In the filters on the left, select Category Suppressions .
   A list of suppression properties displays. The names of the properties begin with Suppress Health Test.
4. Select a health test suppression property to suppress the test.
5. Enter a Reason for change, and then click Save Changes to commit the changes.

Viewing a List of Suppressed Health Tests

You can view a list of suppressed health tests from the Configuration menu.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. From the Home page or the Status page of a cluster, select Configuration Suppressed Health and Configuration Issues .
2. Select Status Non-default .
   A list of suppressed health tests and configuration issues displays.
3. To limit the list to health tests, enter “health test” in the Search box.

Unsuppressing Health Tests

You can unsuppress a health test from where it displays, or unsuppress one or more health tests from the configuration page of the service or host.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. To unsuppress a single health test where it displays, click the Unsuppress... link next to a suppressed test. (You may need to click the Show # Suppressed Test link first.)
2. To unsuppress one or more health tests from the configuration screen:
   a) Go to the service or host with the health test you want to unsuppress.
   b) Select Status Non-default .
      A list of suppressed health tests and configuration issues displays.
   c) Optionally, type the name of the health test in the Search box to locate it.
   d) Clear the suppression property for the health test.
   e) Enter a Reason for change, and then click Save Changes to commit the changes.

Viewing Charts for Cluster, Service, Role, and Host Instances

For cluster, service, role, and host instances, you can see dashboards of charts of various metrics relevant to the entity you are viewing. While the metrics displayed are different for each entity, the basic functionality works in the same way.

The HomeStatus tab for clusters and the Status tab for a service, role, or host display dashboards containing a limited set of charts.

The Status page Charts Library tab displays a dashboard containing a much larger set of charts, organized by categories such as process charts, host charts, CPU charts, and so on, depending on the entity (service, role, or host) that you are viewing.

A custom dashboard is displayed by default when you view the Status tab for an entity. You can switch between custom and default dashboards by using the edit button to the upper right of the chart.

Displaying Information from Charts

There are various ways to display information from charts.

- Click the ⚙ icon at the top right to see a menu for opening the chart in the Chart Builder or exporting its data.
- Change the size of a chart on a dashboard by dragging the lower-right corner of the chart.
- Hovering with the mouse over a stream on a chart (for example, a line on a line chart) opens a small pop-up window that displays information about that stream. Move the mouse horizontally to see the data values change in the small pop-up window, based on the time represented at the mouse's position along the chart's horizontal axis. Click any stream within the chart to display a larger pop-up window that includes additional information for the stream at the point in time where the mouse was clicked. At the bottom of the large pop-up window is a button for viewing the Cloudera Manager page for the entity (service, host, role, query, or application) associated with the chart, if applicable (View Service, View Host, and so on). Click the button View Entity Chart to display a chart

for the stream on its own page. If the chart displays more than one stream, the new chart displays only the stream that was selected when the button was clicked.

- The chart page includes an editable text field containing a default title based on the select statement that was used to create the chart. This title will be used if you save the chart as a dashboard. Type a new title for the chart into this field, if desired.

Related Information

Dashboards
Charting Time-Series Data
Dashboard Types

Exporting Data from Charts

You can export data from charts in either JSON or CSV format.

Procedure

1. On the HomeStatus tab, hover over the chart that you want to export data from.
2. Click the ⚙ icon in the upper right corner of the chart.
   - Click Export JSON to display the chart data in JSON format in a new browser window.
   - Click Export CSV to open a Save dialog box enabling you to save the data as a CSV file. Choose a program to open the CSV, or open the file with your system's default program for editing and displaying CSV files.

   Note: Time values that appear in Cloudera Manager charts reflect the time zone setting on the Cloudera Manager client machine, but time values returned by the Cloudera Manager API (including those that appear in JSON and CSV files exported from charts) reflect Coordinated Universal Time (UTC). For more information on the timestamp format, see the Cloudera Manager API documentation, for example, ApiT imeSeriesData.java.

Adding and Removing Charts from a Dashboard

You can add a chart to a dashboard or remove a chart from a custom dashboard. When you add a chart to a dashboard, you can add it to an existing dashboard or a new custom dashboard, which creates a new dashboard at the same time.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. On the HomeStatus tab, hover over the chart that you want to export data from.
2. To add a chart to a custom dashboard, click the ⚙ icon in the upper right corner of the chart and then click Add to Dashboard.
   - To add the chart to an existing dashboard, select Add chart to an existing custom or system dashboard and then the dashboard name.
   - To add the chart to a new dashboard, select Add chart to a new custom dashboard and enter a name for the dashboard in the Dashboard Name field.

   To remove a chart from a custom dashboard, click the ⚙ icon in the upper right corner of the chart and then click Remove. The Remove button does not appear in the menu when the default dashboard is used because the default dashboard does not allow removing the original charts. Use the edit button to the upper right of the chart

to switch between custom and default dashboards. The Remove button is only available to users with the required roles.

Creating Triggers from Charts

You can create a trigger for most charts. Triggers allow you to define actions to be taken when a specified condition is met.

About this task

Minimum Required Role: Full Administrator. This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Procedure

1. On the HomeStatus tab, hover over the chart that you want to create a trigger for.
2. Click the icon in the upper right corner of the chart and select Create Trigger.

   For information on creating triggers, see *Triggers*.

Related Information

Triggers

Configuring Monitoring Settings

You can configure settings for various types of monitoring in Cloudera Manager, such as health tests, activities, and more.

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

There are several types of monitoring settings you can configure in Cloudera Manager:

- Health tests - For a service or role for which monitoring is provided, you can enable and disable selected health tests and events, configure how those health tests factor into the overall health of the service, and modify thresholds for the status of certain health tests. For hosts you can disable or enable selected health tests, modify thresholds, and enable or disable health alerts.
- Free space - For hosts, you can set threshold-based monitoring of free space in the various directories on the hosts Cloudera Manager monitors.
- Activities - For MapReduce, YARN, and Impala services, you can configure aspects of how Cloudera Manager monitors activities, applications, and queries.
- Alerts - For all roles you can configure health alerts and configuration change alerts. You can also configure some service specific alerts and how alerts are delivered.
- Log events - For all roles you can configure logging thresholds, log directories, log event capture, when log messages become events, and when to generate log alerts.
- Monitoring roles - For the Cloudera Management Service you can configure monitoring settings for the monitoring roles themselves—enable and disable health tests on the monitoring processes as well as configuring some general settings related to events and alerts (specifically with the Event Server and Alert Publisher). Each of the Cloudera Management Service roles has its own parameters that can be modified to specify how much data is retained by that service. For some monitoring functions, the amount of retained data can grow very large, so it may become necessary to adjust the limits.

Related Information

Modifying Configuration Properties Using Cloudera Manager

Configuring Health Monitoring

Depending on the service or role you select, and the configuration category, you can enable or disable health tests, determine when health tests cause alerts, or determine whether specific health tests are used in computing the overall health of a role or service. In most cases you can disable these "roll-up" health tests separately from the individual health tests.

The initial health monitoring configuration is handled during the installation and configuration of your cluster, and most monitoring parameters have default settings. However, you can set or modify these at any time.

As a rule, a health test whose result is considered "Concerning" or "Bad" is forwarded as an event to the Event Server. That includes health tests whose results are based on configured Warning or Critical thresholds, as well pass-fail type health tests. An event is also published when the health test result returns to normal.

You can control when an individual health test is forwarded as an event or as an alert by modifying the threshold values for the relevant health test.

Configuring Service Monitoring

You can configure different aspects of service monitoring. After you configure monitoring for a service, restart the cluster.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Select Clusterscluster_nameservice_name.
2. Click the Configuration tab.
3. Select Scope*service_name* (Service-Wide).
4. Select CategoryMonitoring.
5. Locate the property to change or search for it by typing its name in the Search box.
6. Configure the property.
7. Enter a Reason for change, and then click Save Changes to commit the changes.
8. Click the Cloudera Manager logo to return to the Home page.
9. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Host Monitoring

You can configure different aspects of host monitoring. After you configure monitoring for a service, restart the cluster.

About this task

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Procedure

1. Click the Hosts tab.
2. Select a host.
3. Click the Configuration tab.

4. Select ScopeAll.
5. Click the Monitoring category.
6. Configure the property.
7. Enter a Reason for change, and then click Save Changes to commit the changes.
8. Click the Cloudera Manager logo to return to the Home page.
9. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Directory Monitoring

Cloudera Manager can perform threshold-based monitoring of free space in the various directories on the hosts it monitor, such as log directories or checkpoint directories (for the Secondary NameNode).

These thresholds can be set in one of two ways—as absolute thresholds (in terms of MiB and GiB, and so on) or as percentages of space. As with other threshold properties, you can set values that trigger events at both the Warning and Critical levels.

If you set both thresholds, the Absolute Threshold setting is used.

Configuring YARN Application Monitoring

You can configure the visibility of the YARN application monitoring results.

About this task

To configure whether admin and non-admin users can view all applications, only that user's applications, or no applications:

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Go to the YARN service.
2. Click the Configuration tab.
3. Select Scope *YARN service_name* (Service-Wide) .
4. Click the Monitoring category.
5. Set the Applications List Visibility Settings properties for admin and non-admin users.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Impala Query Monitoring

You can configure the visibility of the Impala query results and the size of the storage allocated to Impala query results.

About this task

To configure whether admin or non-admin users can view all queries, only that user's queries, or no queries:

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Go to the Impala service.
2. Click the Configuration tab.
3. Select Scope*Impala service_name* (Service-Wide).
4. Click the Monitoring category.
5. Set the Visibility Settings properties for admin and non-admin users.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Impala Query Data Store Maximum Size

You can configure the Impala query data store size. The query store stores enough information to make the query searchable through the filter language.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Do one of the following:
   - Select ClustersCloudera Management Service.
   - On the HomeStatustab, in the Cloudera Management Service table, click the Cloudera Management Service link.
2. Click the Configuration tab.
3. Select ScopeService Monitor .
4. Click the Main category.
5. In the Impala Storage section, set the firehose_impala_storage_bytes property. The default is 1 GiB.

   The firehose_impala_storage_bytes property determines the approximate amount of disk space dedicated to storing Impala query data. Once the store reaches its maximum size, older data is deleted to make room for newer queries. The disk usage is approximate because data deletion begins only when the limit has been reached.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Enabling Configuration Change Alerts

You can set configuration change alerts to be service-wide, or on specific roles for the service.

About this task

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Click a service, role, or host.
2. Click the Configuration tab.

3. Click the Monitoring category.
4. Check the Enable Configuration Change Alerts checkbox.
5. Enter a Reason for change, and then click Save Changes to commit the changes.
6. Click the Cloudera Manager logo to return to the Home page.
7. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Filtering Metrics

Metric Filters limit the amount of metric data reported to the Cloudera Manager server to avoid gaps in the reported data.

About this task

Metrics Filters allow you to limit the amount of metric data sent to the Cloudera Manager Service Monitor. In large clusters, some services, such as Kudu, send a high volume of non-essential metrics data to the Service Monitor, which can overload it, causing gaps in the data reported from these metrics in charts/dashboards and metrics queries, and potentially limiting the ability for Cloudera Manager to effectively monitor cluster health . To mitigate this problem, you can configure Metric Filters that limit the amount of data sent to the Service Monitor and Host Monitor. You can configure Metric Filters for any service deployed in a cluster.

You can configure the filter in several ways:

- Limit the collected metrics to only include those required for Health Tests.
- Limit the collected metrics to only include metrics used in the charts on the main page (‘Dashboard’) of the service.
- Include specific metrics - only the specified metrics will be collected for this service
- Exclude specific metrics - the specified metrics will not be collected for this service

Procedure

To configure a Metric Filter:

1. Log in to the Cloudera Manager Admin Console.
2. Navigate to the cluster where you want to add the filter.
3. Click ConfigurationMetric Filters.

   The Configuration page displays the Metric Filter parameter for all roles in the cluster.

   You can now choose whether to edit all the Metric Filters using the same values, or edit the filter for a specific role.
4. Do one of the following:
   - To configure the filter for a specific role, click the Edit Individual Values link and locate the parameter for the service. Follow the steps below.
   - To configure the same filter for all roles, edit the Default Group.
5. Select one of the following:
   - Include Only Health Test Metric Set
   - Include Only Default Dashboard Metric Set

6. To filter specific metrics:
   a) Select either Include or Exclude from the 'Include/Exclude Custom Metrics' drop-down list.
   If you selected either of the Metric sets, and select Include, the custom metrics will be added to the selected Metrics sets. If you select Exclude, they will be excluded from the selected sets.
   b) Click the "Plus" icon to add a metric.
   c) Enter the metric name.
   d) Click the "Plus" icon to add additional metrics.

   Note: You can find the names of specific metrics you would like to include or exclude either from the Cloudera Manager Metrics Reference (see link below) or from a chart that is displaying the required metrics in the Cloudera Manager Admin Console, as follows:

   a. Go to the Status page for a cluster or service.
   b. In a chart, click the gear icon and select Open in Chart Builder.

   The query displays at the top of the page. Metric names are part of the query. For example, the following query:

```
select dfs_capacity, dfs_capacity_used, dfs_capacity_used_non_hdfs
 where entityName=$SERVICENAME
```

   contains the following metric names:

   - dfs_capacity
   - Dfs_capacity_used
   - dfs_capacity_used_non_hdfs
7. Click Save Changes.
8. Refresh the cluster:
   a) Go the cluster's Status page.
   b) Click ActionsRefresh Cluster. This action does not restart the cluster.
   The cluster configuration is refreshed.

   Note: You can configure Cloudera Manager to automatically refresh the cluster after you update Metric configurations by enabling the Enable auto refresh for metric configurations parameter. This parameter is set at the role level, and, when set to true, Cloudera Manager refreshes only the role affected by a change to the metric filter or collection enable settings, not the entire cluster.

Results

Note: If you do not select any options, all metrics for the service will be sent to the Service Monitor and Host Monitor.

Related Information

Cloudera Manager Metrics Reference

Configuring Log Events

You can enable or disable the forwarding of selected log events to the Event Server.

This functionality is enabled by default, and is a service-wide setting (Enable Log Event Capture) for each service for which monitoring is provided. You can enable and disable event capture for Cloudera Runtime services or for the Cloudera Management Service.

Important: Cloudera does not recommend logging to a network-mounted file system. If a role is writing its logs across the network, a network failure or the failure of a remote file system can cause that role to freeze up until the network recovers.

Configuring Logs

You can configure log properties.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Go to a service.
2. Click the Configuration tab.
3. Select *role_name* (Service-Wide) Logs .
4. Edit a log property.
5. Enter a Reason for change, and then click Save Changes to commit the changes.
6. Click the Cloudera Manager logo to return to the Home page.
7. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Logging Thresholds

A logging threshold determines what level of log message is reported.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

The available levels are:

- TRACE - Informational events finer-grained than DEBUG.
- DEBUG - Informational events useful to debug an application.
- INFO - Informational events that highlight progress at coarse-grained level.
- WARN - Events that indicate a potential problem which is handled by the application.
- ERROR - Error events that allows the application to continue running.
- FATAL - Very severe error events that typically lead the application to abort.

The number of messages is greater and severity is least for TRACE. The default setting is INFO.

Procedure

1. Go to a service.
2. Click the Configuration tab.
3. Enter Logging Threshold in the Search text field.
4. For the desired role group, select a logging threshold level.
5. Enter a Reason for change, and then click Save Changes to commit the changes.
6. Click the Cloudera Manager logo to return to the Home page.
7. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Related Information

Levels

Configuring Log Directories

You can configure log directories for a cluster or a service.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Do one of the following:
   - Cluster:
     a) On the Home Status tab, click a cluster name.
     b) Select Configuration Log Directories .
     c) Edit a *role_name* Log Directory property.
   - Service:
     a) Go to a service.
     b) Click the Configuration tab.
     c) Select *role_name* (Service-Wide) Logs .
     d) Edit the Log Directory property.
2. Enter a Reason for change, and then click Save Changes to commit the changes.
3. Click the Cloudera Manager logo to return to the Home page.
4. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Enabling and Disabling Log Event Capture

You can enable and disable log event capture for a service.

About this task

You can also modify the rules that determine how log messages are turned into events. Editing these rules is not recommended.

For each role, there are rules that govern how its log messages are turned into events by the custom log4j appender for the role. These are defined in the Rules to Extract Events from Log Files property.

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Select Clusters*cluster_nameservice_name*.
2. Click the Configuration tab.
3. Select Scope *service_name* (Service-Wide) .
4. Click the Monitoring category.
5. Modify the Enable Log Event Capture setting.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Which Log Messages Become Events

You can configure rules to determine which log messages become events.

About this task

Cloudera defines a number of rules by default. For example:

- The line {"rate": 10, "threshold":"FATAL"}, means log entries with severity FATAL should be forwarded as events, up to 10 a minute.
- The line {"rate": 0, "exceptiontype": "java.io.EOFException"}, means log entries with the exception `java.io.EOFException` should always be forwarded as an event.

The syntax for these rules is defined in the Description field for this property: the syntax lets you create rules that identify log messages based on log4j severity, message content matching, or the exception type. These rules must result in valid JSON.

Note: Editing these rules is not recommended. Cloudera Manager provides a default set of rules that should be sufficient for most users.

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Select Clusters*cluster_nameservice_name*.
2. Click the Configuration tab.
3. Enter Rules to Extract Events from Log Files in the Search text field.
4. Click the Monitoring category.
5. Select the role group for the role for which you want to configure log events, or search for "Rules to Extract Events from Log Files."

   Note that for some roles there may be more than one role group, and you may need to modify all of them. The easiest way to ensure that you have found all occurrences of the property you need to modify is to search for the property by name. Cloudera Manager shows all copies of the property that matches the search filter.
6. In the Content field, edit the rules as needed. Rules can be written as regular expressions.
7. Enter a Reason for change, and then click Save Changes to commit the changes.
8. Click the Cloudera Manager logo to return to the Home page.
9. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Log Alerts

You specify that a log event should generate an alert (by setting "alert" :true in the rule).

If you specify a content match, the entire content must match — if you want to match on a partial string, you must provide wildcards as appropriate to allow matching the entire string.

Monitoring Clusters

There are several locations in Cloudera Manager where you can monitor clusters.

The **Clusters** tab in the top navigation bar displays each cluster's services in its own section, with the Cloudera Management Service separately below. You can select the following cluster-specific pages: hosts, reports, activities, and resource management.

The HomeStatus tab displays the clusters being displayed by Cloudera Manager.

To display a cluster Status page, click the cluster name on the HomeStatus tab Status tab. The cluster Status page displays a table containing links to the Hosts page and the status pages of the services running in the cluster.

Each service row in the table has a menu of actions that you select by clicking and can contain one or more of the following indicators:

Indicator	Meaning	Description
2	Health issue	Indicates that the service has at least one health issue. The indicator shows the number of health issues at the highest severity level. If there are Bad health test results, the indicator is red. If there are no Bad health test results, but Concerning test results exist, then the indicator is yellow. No indicator is shown if there are no Bad or Concerning health test results. **Important:** If there is one Bad health test result and two Concerning health results, there will be three health issues, but the number will be one. Click the indicator to display the Health Issues pop-up dialog box. By default only Bad health test results are shown in the dialog box. To display Concerning health test results, click the Also show *n* concerning issue(s) link.Click the link to display the Status page containing with details about the health test result.
4	Configuration issue	Indicates that the service has at least one configuration issue. The indicator shows the number of configuration issues at the highest severity level. If there are configuration errors, the indicator is red. If there are no errors but configuration warnings exist, then the indicator is yellow. No indicator is shown if there are no configuration notifications. **Important:** If there is one configuration error and two configuration warnings, there will be three configuration issues, but the number will be one. Click the indicator to display the Configuration Issues pop-up dialog box. By default only notifications at the Error severity level are listed, grouped by service name are shown in the dialog box. To display Warning notifications, click the Also show *n* warning(s) link.Click the message associated with an error or warning to be taken to the configuration property for which the notification has been issued where you can address the issue. For more information see the topic *Managing Services*.
Restart Needed Refresh Needed	Configuration modified	Indicates that at least one of a service's roles is running with a configuration that does not match the current configuration settings in Cloudera Manager. Click the indicator to display the Stale Configurations page.To bring the cluster up-to-date, click the Refresh or Restart button on the Stale Configurations page to refresh the cluster, restart the cluster, or restart the stale service.
	Client configuration redeployment required	Indicates that the client configuration for a service should be redeployed. Click the indicator to display the Stale Configurations page. To bring the cluster up-to-date, click the Deploy Client Configuration button on the Stale Configurations page or manually redeploy the Client Configuration.

The right side of the status page displays charts that summarize resource utilization (IO, CPU usage) and processing metrics.

Note: If you delete a cluster, the deleted cluster still displays in some charts. This is because the charts also show historical data. Over time, data from the deleted cluster will drop off as older data is replaced by more current data. You can work around this by:

- Waiting for the data from the deleted cluster to drop off.
- Editing the where clause of query for the chart to include only the cluster(s) you are interested in. (For example: clusterDisplayName=Cluster_1). You can revert to the original query at a later date, after the data for the deleted cluster has dropped off. See the topic Charting Time-Series Data.
- Deleting all data in the Host Monitor and Service Monitor storage directories and starting from scratch. You will, however, loose all historical data from both current and deleted clusters. See Configuring Service Monitor Data Storage and Configuring Host Monitor Data Storage to learn where the storage directories are located.

Related Information

Monitoring Activities
Reports
Managing Cloudera Runtime Services
Dashboards
Charting Time-Series Data

Client Configuration Files
Manually Redeploying Client Configuration Files
Stale Configurations
Viewing Role Instance Status

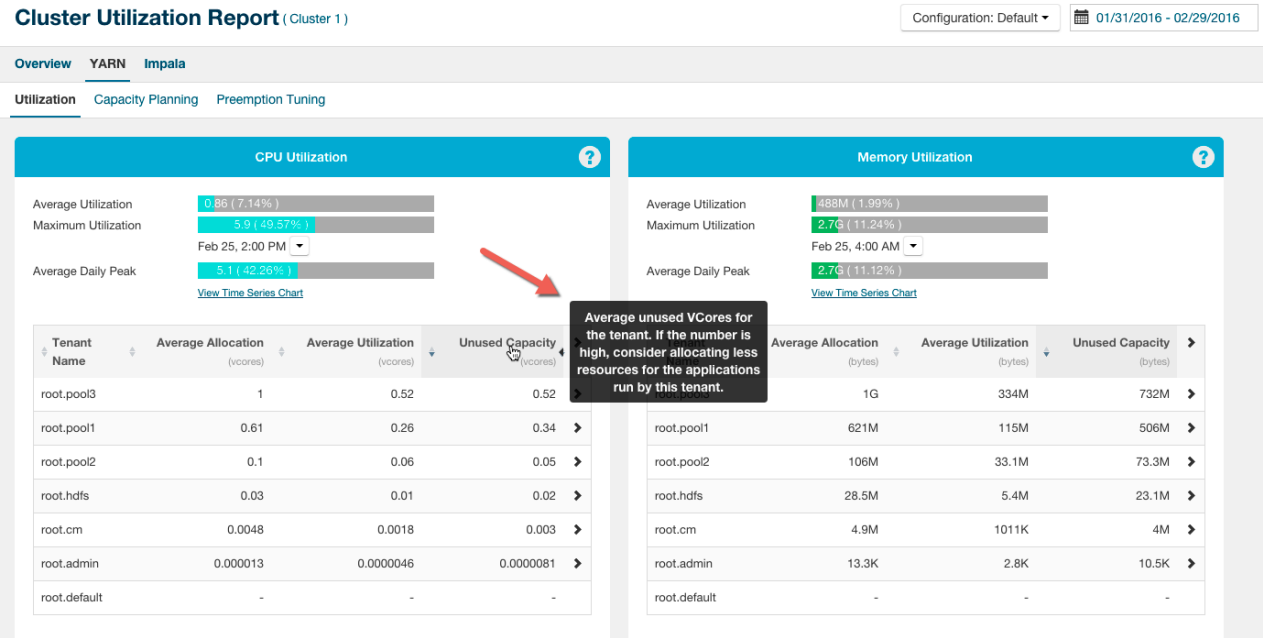
Cluster Utilization Report overview

The Cluster Utilization Report screens in Cloudera Manager display aggregated utilization information for YARN and Impala jobs.

The reports display CPU utilization, memory utilization, resource allocations made due to the YARN capacity scheduler, and Impala queries. The report displays aggregated utilization for the entire cluster and also breaks out utilization by tenant, which is either a user or a resource pool. You can configure the report to display utilization for a range of dates, specific days of the week, and time ranges.

The report displays the current utilization of CPU and memory resources and the resources that were allocated using the Cloudera Manager resource management features.

Using the information displayed in the Cluster Utilization Report, a Cloudera Runtime cluster administrator can verify that sufficient resources are available for the number and types of jobs running in the cluster. An administrator can use the reports to tune resource allocations so that resources are used efficiently and meet business requirements. Tool tips in the report pages provide suggestions about how to improve performance based on the information displayed in the report. Hover over a label to see these suggestions and other information. For example:



You can tune the following:

- CPU and memory allocations
- Weights for each pool
- Scheduling rules
- Preemption thresholds
- Maximum number of running and queued Impala queries
- Maximum timeout for the queue of Impala queries
- Placement rules
- Number of hosts in a cluster

- Memory capacity of hosts
- Impala Admission Control pool and queue configurations

Enable the Cluster Utilization Report

You must configure several parameters to enable the Cluster Utilization Report.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

By default, the Cluster Utilization Report displays aggregated CPU and memory utilization for an entire cluster and for YARN and Impala utilization. You can also view this utilization by tenants, which include Linux users and Dynamic Resource Pools. To see utilization for a tenant, you must configure the tenant and define resource limits for it.

Procedure

Enable YARN utilization metrics collection:

1. In Cloudera Manager, select the YARN service.
2. Click the Configuration tab.
3. Use the Search function to locate the configuration properties mentioned below.
4. In the Container Usage MapReduce Job User property, enter a username for the MapReduce job that collects the metrics.

   The username you enter must be a Linux user on all the cluster hosts. If you are using an Active Directory KDC, the username must also exist in Active Directory. For secure clusters, the user must not be banned or below the minimum user ID. You can view the list of banned users (banned.users) and the minimum user ID (min.user.id) by under YARN Configuration in Cloudera Manager.

   The user that is configured with the Container Usage MapReduce Job User property in the YARN service requires permissions to read the subdirectories of the HDFS directory specified with the Cloudera Manager Container Usage Metrics Directory property. The default umask of 022 allows any user to read from that directory. However, if a more strict umask (for example, 027) is used, then those directories are not readable by any user. In that case the user specified with the Container Usage MapReduce Job User property should be added to the same group that owns the subdirectories.

   For example, if the /tmp/cmYarnContainerMetrics/20161010 subdirectory is owned by user and group yarn:had oop, the user specified in Container Usage MapReduce Job User should be added to the hadoop group.

   The directories you specify with the Cloudera Manager Container Usage Metrics Directory and Container Usage Output Directory properties should not be located in encryption zones.
5. Optionally, enter the resource pool in which the container usage collection MapReduce job runs in the Container Usage MapReduce Job Pool property.

   Cloudera recommends that you dedicate a resource pool for running this MapReduce job.

   If you specify a custom resource pool, ensure that the placement rules for the cluster allow for it. The first rule must be for resource pools to be specified at run time with the Create pool if it does not exist option selected. Alternatively, ensure that the pool you specify already exists. If the placement rule is not properly configured or the resource pool does not already exist, the job may run in a different pool.
6. Click Save Changes.
7. Click the Actions button.
8. Select Create CM Container Usage Metrics Dir.
9. Restart the YARN service.

Enable Impala utilization collection:

10. In Cloudera Manager, select the Impala service.
11. Click the Configuration tab.
12. Search for admission control.
13. Find the Enable Impala Admission Control and the Enable Dynamic Resource Pools properties and enable both of them.
14. Click Save Changes.
15. Restart the Impala service.

Configure the Cluster Utilization Report

You can configure the Cluster Utilization Report using Cloudera Manager.

About this task

To access the Cluster Utilization Report, go to Clusters and then select Utilization Report for the cluster. The Overview tab displays when you first open the report.

The upper-right part of the page has two controls that you use to configure the Cluster Utilization Report:

Procedure

1. Click the Configuration drop-down menu.
2. Select one of the configured options, or create a new configuration.

If you want to create a new configuration, do the following:

3. Click Create New Configuration.
4. Enter a Configuration Name.
5. Select the Tenant Type: Pool or User.
6. Select the days of the week for which you want to report utilization.
7. Select All Day, or use the drop-down menus to specify a utilization time range for the report.
8. Click Create.
   The configuration you created is now available from the Configuration drop-down menu.

Select a date range for the report:

9. Click the date range button.
10. Select one of the range options (Today, Yesterday, Last 7 Days, Last 30 Days, or This Month) or click Custom Range and select the beginning and ending dates for the date range.

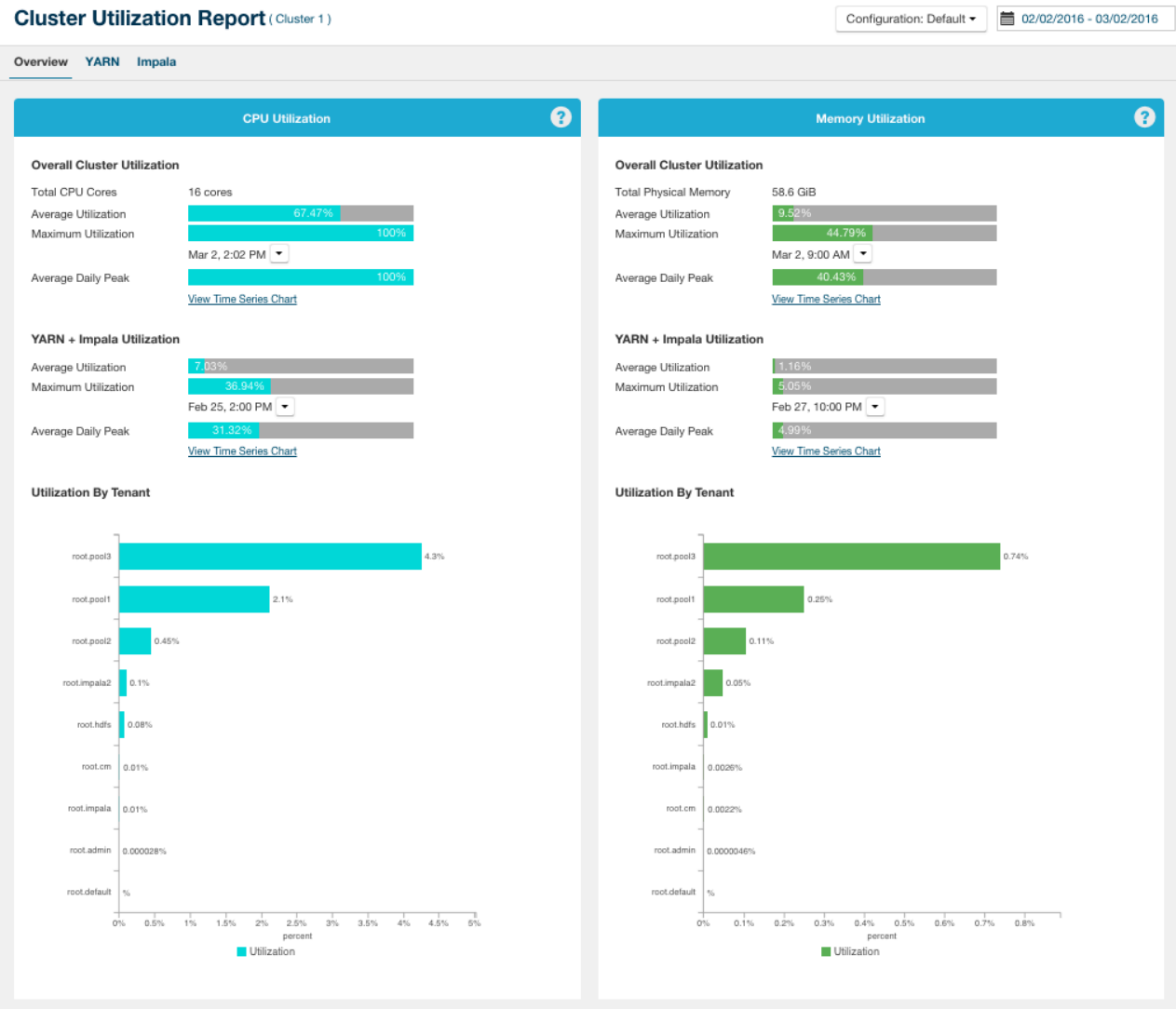
Use the Cluster Utilization Report to manage resources

The Cluster Utilization Report provides information about CPU and memory utilization in three tabs: Overview, YARN and Impala tabs.

To access the Cluster Utilization Report, go to Clusters and then select Utilization Report for the cluster. The Overview tab of the report displays.

Note: The report updates utilization information every hour. The utilization information for Impala and YARN queries does not display in the Cluster Utilization Report until captured by the hourly update.

Figure 1: Cluster Utilization Report Overview Tab



The Cluster Utilization Report is divided into the following tabs:

- Overview Tab
- YARN Tab
- Impala Tab

Overview Tab

The overview tab of the cluster utilization report provides a summary of CPU and memory utilization.

The Overview tab provides a summary of CPU and memory utilization for the entire cluster and also for only YARN applications and Impala queries. Two sections, CPU Utilization and Memory Utilization, display the following information:

CPU Utilization	Memory Utilization
Overall Cluster Utilization • Total CPU Cores – Average number of CPU cores available during the reporting window. • Average Utilization – Average CPU utilization for the entire cluster, including resources consumed by user applications and Cloudera Runtime services. • Maximum Utilization – Maximum CPU utilization for the entire cluster during the reporting window, including resources consumed by user applications and Cloudera Runtime services. If this value is high, consider adding more hosts to the cluster. Click the drop-down menu next to the date and select one of the following to view details about jobs running when maximum utilization occurred: • View YARN Applications Running at the Time • View Impala Queries Running at the Time • Average Daily Peak – Average daily peak CPU consumption for the entire cluster during the reporting window. This includes resources consumed by user applications and Cloudera Runtime services. The number is computed by averaging the maximum resource consumption for each day of the reporting period. Click View Time Series Chart to view a chart of peak utilization.	Overall Cluster Utilization • Total Physical Memory – Average physical memory available in the cluster during the reporting window. • Average Utilization – Average memory consumption for the entire cluster, including resources consumed by user applications and Cloudera Runtime services. • Maximum Utilization – Maximum memory consumption for the entire cluster during the reporting window, including resources consumed by user applications and Cloudera Runtime services. If this value is high, consider adding more hosts to the cluster. Click the drop-down menu next to the date and select one of the following to view details about jobs running when maximum utilization occurred: • View YARN Applications Running at the Time • View Impala Queries Running at the Time • Average Daily Peak – Average daily peak memory consumption for the entire cluster during the reporting window, including resources consumed by user applications and Cloudera Runtime services. The number is computed by averaging the maximum memory utilization for each day of the reporting period. Click View Time Series Chart to view a chart of peak utilization.
YARN + Impala Utilization • Average Utilization – Average resource consumption by YARN applications and Impala queries that ran on the cluster. • Maximum Utilization – Maximum resource consumption by YARN applications and Impala queries that ran on the cluster. Click the drop-down menu next to the date and select one of the following to view details about jobs running when maximum utilization occurred: • View YARN Applications Running at the Time • View Impala Queries Running at the Time • Average Daily Peak – Average daily peak resource consumption by YARN applications and Impala queries during the reporting window. The number is computed by finding the maximum resource consumption per day and calculating the mean. Click View Time Series Chart to view a chart of peak utilization.	YARN + Impala Utilization • Average Utilization – Average memory consumption by YARN applications and Impala queries that ran on the cluster. • Maximum Utilization – Maximum memory consumption for the entire cluster during the reporting window, including resources consumed by user applications and Cloudera Runtime services. If this is high, consider adding more hosts to the cluster. Click the drop-down menu next to the date and select one of the following to view details about jobs running when maximum utilization occurred: • View YARN Applications Running at the Time • View Impala Queries Running at the Time • Average Daily Peak – Average daily peak memory consumption by YARN applications and Impala queries during the reporting window. The number is computed by finding the maximum resource consumption per day and then calculating the mean. Click View Time Series Chart to view a chart of peak utilization.
Utilization by Tenant Displays overall utilization for each tenant. Tenants can be either pools or users.	Utilization by Tenant Displays overall utilization for each tenant. Tenants can be either pools or users.

YARN Tab

The YARN tab displays CPU and memory utilization for YARN application on two tabs: Utilization and Capacity Planning tab.

Utilization Tab

Table 1: Utilization Tab

CPU Utilization	Memory Utilization
Summary section: • Average Utilization – Average number of vcores used by YARN applications. The percentage reported is of the total number of vcores configured for YARN. • Maximum Utilization – Maximum number of vcores used by YARN applications. The percentage reported is of the total number of vcores configured for YARN. Click the drop-down menu next to the date and select View YARN Applications Running at the Time to view details about jobs running when maximum utilization occurred. • Average Daily Peak – Average daily peak vcores used by YARN applications. The number is computed by finding the maximum resource consumption per day and calculating the mean. The percentage reported is of the total number of vcores configured for YARN. Click View Time Series Chart to view a chart of peak utilization.	Summary section: • Average Utilization – Average memory used by YARN applications. The percentage reported is of the total container memory configured for YARN. • Maximum Utilization – Maximum memory used by YARN applications. The percentage reported is of the total container memory configured for YARN. Click the drop-down menu next to the date and select View YARN Applications Running at the Time to view details about jobs running when maximum utilization occurred. • Average Daily Peak – Average daily peak memory used by YARN applications. The number is computed by finding the maximum resource consumption per day and calculating the mean. The percentage reported is of the total container memory configured for YARN. Click View Time Series Chart to view a chart of peak utilization.
Utilization by Tenant Displays overall utilization for each tenant. The tenants can be either pools (queues) or users. Utilization by tenant is displayed in a table with the following columns: • Tenant Name • Average Allocation – Average number of vcores allocated to YARN applications of the tenant. The percentage reported is of the total number of vcores configured for YARN. • Average Utilization – Average memory used by YARN applications. The percentage reported is of the total container memory configured for YARN. • Unused Capacity – Average unused vcores for the tenant. If this number is high, consider allocating less resources for the applications run by this tenant. Click the right arrow icon in the header row of the table to view utilization charts for all tenants. Click right arrow in a row to view CPU utilization for a single tenant.	Utilization by Tenant Displays overall utilization for each tenant. The tenants can be either pools (queues) or users. Utilization by tenant is displayed in a table with the following columns: • Tenant Name • Average Allocation – Average memory allocated to YARN applications of the tenant. The percentage reported is of the total container memory configured for YARN. • Average Utilization – Average memory used by YARN applications. The percentage reported is of the total container memory configured for YARN. • Unused Capacity – Average unused memory for the tenant. If this number is high, consider allocating less resources for the applications run by this tenant. Click the right arrow icon in the header row of the table to view utilization charts for all tenants. Click right arrow in a row to view CPU utilization for a single tenant.

Adjusting YARN Resources

To adjust YARN resources. Go to the YARN service and select Configuration Category Resource Management and configure the following properties:

- vcores: Container Virtual CPU Cores
- Memory: Container Memory

Capacity Planning Tab

The Capacity Planning tab displays a table showing how the capacity assigned to YARN Queues affects CPU and memory allocations. The table displays the following columns:

- Tenant Name
- Guaranteed Vcores – Displays the guaranteed number of virtual CPU vcores allocated to each tenant based on the configured capacity of the queue.
- Guaranteed MB – Displays the guaranteed memory in bytes allocated to each tenant based on the configured capacity of the queue.
- Wait Ratio During Contention – The wait ratio is the average percentage of containers in the YARN pool that were pending when there was at least one pending container in the queue. If a queue running critical applications

has a high wait ratio, consider increasing the capacity of that pool. If several pools in the cluster have a high wait ratio, consider adding more hosts to the cluster.

Impala Tab

The Impala tab displays CPU and memory utilization for Impala queries using three tabs: Queries, Peak Memory Usage and Spilled Memory tab.

Queries Tab

The Overview tab displays information about Impala queries.

The top part of the page displays summary information about Impala queries for the entire cluster. The table in the lower part displays the same information by tenant. Both sections display the following:

- Total – Total number of queries.

  Click the link with the total to view details and charts about the queries.
- Avg Wait Time in Queue – Average time, in milliseconds, spent by a query in an Impala pool while waiting for resources. If this number is high, consider increasing the resources allocated to the pool. If this number is high for several pools, consider increasing the number of hosts in the cluster.
- Successful – The number and percentage of queries that finished successfully.

  Click the link with the total to view details and charts about the queries.
- Memory Limit Exceeded – Number and percentage of queries that failed due to insufficient memory. If there are such queries, consider increasing the memory allocated to the pool. If there are several pools with such queries, consider increasing the number of hosts in the cluster.
- Timed Out in Queue – Number of queries that timed out while waiting for resources in a pool. If there are such queries, consider increasing the maximum number of running queries allowed for the pool. If there are several pools with such queries, consider increasing the number of hosts in the cluster.
- Rejected – Number of queries that were rejected by Impala because the pool was full. If this number is high, consider increasing the maximum number of queued queries allowed for the pool.

Click the column header to sort the table by that column.

Peak Memory Usage Tab

This report shows how Impala consumes memory at peak utilization. If utilization is high for a pool, consider adding resources to the pool. If utilization is high for several pools, consider adding more hosts to the cluster.

The Summary section of this page displays aggregated peak memory usage information for the entire cluster and the Utilization by Tenant section displays peak memory usage by tenant. Both sections display the following:

- Max Allocated
  - Peak Allocation Time – The time when Impala reserved the maximum amount of memory for queries.

    Click the drop-down list next to the date and time and select View Impala Queries Running at the Time to see details about the queries.
  - Max Allocated – The maximum memory that was reserved by Impala for executing queries. If the percentage is high, consider increasing the number of hosts in the cluster.
  - Utilized at the Time – The amount of memory used by Impala for running queries at the time when maximum memory was reserved.

    Click View Time Series Chart to view a chart of peak memory allocations.
  - Histogram of Allocated Memory at Peak Allocation Time – Distribution of memory reserved per Impala daemon for executing queries at the time Impala reserved the maximum memory. If some Impala daemons have reserved memory close to the configured limit, consider adding more physical memory to the hosts.

    Note: This histogram is generated from the minute-level metrics for Impala daemons. If the minute-level metrics for the timestamp at which peak allocation happened are no longer present in the Cloudera Service Monitor Time-Series Storage, the histogram shows no data. To maintain a longer history for the minute-level metrics, increase the value of the Time-Series Storage property for the Cloudera Service Monitor. (Go to the Cloudera Management Service Configuration and search for Time-Series Storage.)
- Max Utilized
  - Peak Usage Time – The time when Impala used the maximum amount of memory for queries.

    Click the drop-down list next to the date and time and select View Impala Queries Running at the Time to see details about the queries.
  - Max Utilized – The maximum memory that was used by Impala for executing queries. If the percentage is high, consider increasing the number of hosts in the cluster.
  - Reserved at the Time – The amount of memory reserved by Impala at the time when it was using the maximum memory for executing queries.

    Click View Time Series Chart to view a chart of peak memory utilization.
  - Histogram of Utilized Memory at Peak Usage Time – Distribution of memory used per Impala daemon for executing queries at the time Impala used the maximum memory. If some Impala daemons are using memory close to the configured limit, consider adding more physical memory to the hosts.

    Note: This histogram is generated from the minute-level metrics for Impala daemons. If the minute-level metrics for the timestamp at which peak allocation happened are no longer present in the Cloudera Service Monitor Time-Series Storage, the histogram shows no data. To maintain a longer history for the minute-level metrics, increase the value of the Time-Series Storage property for the Cloudera Service Monitor. (Go to the Cloudera Management Service Configuration and search for Time-Series Storage.)

Spilled Memory Tab

The Spilled Memory tab displays information about Impala spilled memory. These disk spills can deteriorate the performance of Impala queries significantly. This report shows the amount of disk spills for Impala queries by tenant. If disk spill is high for a pool, consider adding resources to the pool. If disk spill is high for several pools, consider adding more hosts to the cluster.

For each tenant, the following are displayed:

- Average Spill – Average spill per query
- Maximum Spill – Maximum memory spilled per hour

Download the Cluster Utilization Report

You can download the Cluster Utilization Reports as a JSON file using the Cloudera Manager API.

See the Cloudera Manager REST API documentation for the following API endpoints:

- Cluster Utilization:path__clusters_-clusterName-_utilization.html
- Impala Utilization: path__clusters_-clusterName-_impalaUtilization.html
- YARN Utilization: path__clusters_-clusterName-_yarnUtilization.html

Creating a Custom Cluster Utilization Report

You can create a custom Cluster Utilization Report using different metrics and queries.

Cloudera Manager provides a Cluster Utilization Report that displays aggregated utilization information for YARN and Impala jobs. If you wish to export the data from this report, you can build custom reports based on the same metrics data using the Cloudera Manager Admin console or the Cloudera Manager API. These reports all use the tsquery Language to chart time-series data.

Metrics and queries

You can use metrics and queries to create a custom Cluster Utilization Report in Cloudera Manager.

Many of the metrics described below use a data granularity of hourly. This is not required, but is recommended because some of the YARN utilization metrics are only available hourly and using the hourly granularity allows for consistent reporting.

Cluster-Level CPU and Memory Metrics

Total cluster CPU usage

Data Granularity: hourly

Units: percentage

tsquery:

```
SELECT
   cpu_percent_across_hosts
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Total CPU Cores in the cluster

Data Granularity: hourly

Units: CPU cores

tsquery:

```
SELECT
   total_cores_across_hosts
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Total cluster memory usage

Data Granularity: hourly

Units: percentage

tsquery:

```
SELECT
   100 * total_physical_memory_used_across_hosts/total_physical_m
emory_total_across_hosts
```

```
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Total cluster memory usage

Time series of total cluster memory usage.

Data Granularity:hourly

Units: Byte seconds

tsquery:

```
SELECT
   total_physical_memory_total_across_hosts
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

CPU used by Impala

Time series of total Impala CPU usage in milliseconds.

Data Granularity: hourly

Units: milliseconds

tsquery:

```
SELECT
   counter_delta(impala_query_thread_cpu_time_rate)
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Memory used by Impala

Time series of Impala memory usage

Data Granularity: hourly

Units: byte seconds

tsquery:

```
SELECT
   counter_delta(impala_query_memory_accrual_rate)
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

CPU used by YARN

The yarn_reports_containers_used_cpu metric used in this tsquery is generated per hour, therefore the data granularity used for this query is the raw metric value.

Data Granularity: Raw

Units: percent seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_cpu FROM REPORTS
WHERE
   category=SERVICE
```

```
   AND clusterName=Cluster_Name
```

Memory used by YARN

Yarn memory usage. The yarn_reports_containers_used_memory metric used in this tsquery is generated per hour, therefore the data granularity used for this query is the raw metric value.

Data Granularity: raw metric value

Units: megabyte seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_memory
FROM
   REPORTS
WHERE
   category=SERVICE
   AND clusterName=Cluster_Name
```

Pool-Level CPU and Memory Metrics

CPU used by Impala pool

CPU usage for an Impala pool.

Data Granularity: hourly

Units: milliseconds

tsquery:

```
SELECT
   counter_delta(impala_query_thread_cpu_time_rate)
WHERE
   category=IMPALA_POOL
   AND poolName=Pool_Name
```

Memory used by Impala pool

Data Granularity: hourly

Units: byte seconds

tsquery:

```
SELECT
   counter_delta(impala_query_memory_accrual_rate)
WHERE
   category=IMPALA_POOL
   AND poolName=Pool_Name
```

CPU used by YARN pool

Provides CPU metrics per YARN pool and user. You can aggregate a pool-level metric from this query.

Data Granularity: Raw

Units: percent seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_cpu FROM REPORTS
```

```
WHERE
   category=YARN_POOL_USER
```

Memory used by YARN pool

Provides memory metrics per YARN pool and user. You can aggregate a pool-level metric from this query.

Data Granularity: hourly

Units: megabyte seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_memory
FROM
   REPORTS
WHERE
   category=YARN_POOL_USER
```

YARN Metrics

YARN VCore usage

Data Granularity: Raw

Units: VCore seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_vcores
FROM
   REPORTS
WHERE
   category=SERVICE
   AND clusterName=Cluster_Name
```

Total VCores available to YARN

Data Granularity: hourly

Units: Number of VCores (Note that this value is not multiplied by the time unit.)

tsquery:

```
SELECT
   total_allocated_vcores_across_yarn_pools + total_available_vco
res_across_yarn_pools
WHERE
   category=SERVICE
   AND clusterName=Cluster_Name
```

YARN Memory usage

Data Granularity: Raw

Units: MB seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_memory FROM REPORTS
WHERE
   category=SERVICE
```

```
   AND clusterName=Cluster_Name
```

Total memory available to YARN

Data Granularity: hourly

Units: MB (Note that this value is not multiplied by the time unit.)

tsquery:

```
SELECT
   total_available_memory_mb_across_yarn_pools + total_allocated_
memory_mb_across_yarn_pools
WHERE
   category=SERVICE
   AND clusterName=Cluster_Name
```

Pool-level VCore usage

The results of this query return the usage for each user in each pool. To see the total usage for a pool, sum all users of the pool.

Data Granularity: Raw

Units: VCore seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_vcores FROM REPORTS
WHERE
   category=YARN_POOL_USER
```

To view metrics for a specific pool, add poolName=*Pool Name* to the tsquery statement.

Pool-level memory usage

The results of this query return the usage for each user in each pool. To see the total usage for a pool, sum all users of the pool.

Data Granularity: Raw

Units: MB seconds

tsquery:

```
SELECT
   yarn_reports_containers_used_memory FROM REPORTS
WHERE
   category=YARN_POOL_USER
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool-level allocated VCores

The results of this query return the usage for each user in each pool. To see the total usage for a pool, sum all users of the pool.

Data Granularity: raw metric value

Units: VCore seconds

tsquery:

```
SELECT
   yarn_reports_containers_allocated_vcores FROM REPORTS
WHERE
```

```
   category=YARN_POOL_USER
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool-level allocated memory

The results of this query return the usage for each user in each pool. To see the total usage for a pool, sum all users of the pool.

Data Granularity: raw metric value

Units: megabyte seconds

tsquery:

```
SELECT
   yarn_reports_containers_allocated_memory
FROM
   REPORTS
WHERE
   category=YARN_POOL_USER
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool-level steady fair share VCore

Data Granularity: hourly

Units: VCores

tsquery:

```
SELECT
   steady_fair_share_vcores
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool-level fair share VCore

Data Granularity: hourly

Units: VCores

tsquery:

```
SELECT
   fair_share_vcores
WHERE
   category=YARN_POOL
```

Pool-level steady fair share memory

Data Granularity: hourly

Units: MB

tsquery:

```
SELECT
   steady_fair_share_mb
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool-level fair share memory

Data Granularity: hourly

Units: MB

tsquery:

```
SELECT
   fair_share_mb
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Metric indicating contention

Data Granularity: hourly

Units: percentage

tsquery:

```
SELECT
   container_wait_ratio
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

YARN Contention-Related Metrics

Use the following metrics to monitor resource contention.

Pool-level allocated VCores when contention occurs

Data Granularity: hourly

Units: VCores

tsquery:

```
SELECT
   allocated_vcores_with_pending_containers
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool level steady fair share VCores when contention occurs

Data Granularity: hourly

Units: VCores

tsquery:

```
SELECT
   steady_fair_share_vcores_with_pending_containers
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool level fair share VCores when contention occurs

Data Granularity: hourly

Units: VCores

tsquery:

```
SELECT
   fair_share_vcores_with_pending_containers
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool level allocated memory when contention occurs

Data Granularity: hourly

Units: MB

tsquery:

```
SELECT
   allocated_memory_mb_with_pending_containers
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool level steady fair share memory when contention occurs

Data Granularity: hourly

Units: MB

tsquery:

```
SELECT
   steady_fair_share_mb_with_pending_containers
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Pool level fair share memory when contention occurs

Data Granularity: hourly

Units: MB

tsquery:

```
SELECT
   fair_share_mb_with_pending_containers
WHERE
   category=YARN_POOL
```

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Impala-Specific Metrics

To view metrics for a specific pool, add poolName=Pool Name to the tsquery statement.

Total reserved memory

Data Granularity: hourly

Units: MB seconds

tsquery:

```
SELECT
   total_impala_admission_controller_local_backend_mem_reserved_a
cross_impala_daemon_pools
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Total used memory

Data Granularity: hourly

Units: MB seconds

tsquery:

```
SELECT
   total_impala_admission_controller_local_backend_mem_usage_acro
ss_impala_daemon_pools
 WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Total available memory

Data Granularity: hourly

Units: MB seconds

tsquery:

```
SELECT
   total_mem_tracker_process_limit_across_impalads
WHERE
   category=CLUSTER
   AND clusterName=Cluster_Name
```

Note: To query for pool-level metrics, change the category to IMPALA-POOL in the above tsquery statements.

Impala query counter metrics

Use Impala query counter metrics to get information about the rate of Impala queries.

Include the following in the SELECT statement of the tsquery:

- counter_delta(queries_ingested_rate)
- counter_delta(queries_successful_rate)
- counter_delta(queries_rejected_rate)
- counter_delta(queries_oom_rate)
- counter_delta(queries_timed_out_rate)
- counter_delta(impala_query_admission_wait_rate)
- counter_delta(impala_query_memory_spilled_rate)

For example:

```
SELECT
   counter_delta(queries_ingested_rate)
   WHERE
   category=IMPALA_POOL
   AND clusterName=Cluster_Name
```

```
    AND serviceName=Service_Name
```

Calculations for reports

Learn about how to correctly perform calculation using metrics values.

All the metrics return a time series of metric values. Depending on the collection frequency of the metric itself and the data granularity you use when issuing tsquery statements, the results return metric values in different frequencies and therefore there are different ways to handle the metric values.

Note the following about how to correctly perform calculations using metric values:

- YARN container metrics are generated once per hour resulting in one raw metric value every hour. Therefore, the most detailed results possible for YARN CPU and memory usage are hourly reports.
- Hourly aggregates are summarized from raw metric values. These aggregates include a set of statistics that include the sum, maximum, minimum, count and other statistics that summarize the raw metric values. When you use the hourly granularity, you lose the single values of the raw metric values. However, you can still get peak usage data for such metrics.
- For some of the YARN metrics described in this topic, the tsquery statement aggregates from the pool and user level to pool level in the Cloudera Manager Cluster Utilization reports. For these queries, because the maximum and minimum for different pool and user combinations are not likely to happen at the same time, there is no way to get the peak usage across pool and user combinations, or at the pool level. The only meaningful results possible are average and sum.
- When calculating CPU/Memory usage percentage, pay attention to the units for each metric. For example, if the cluster consistently has 8 VCores, the total VCore seconds for each hour would be 8 * 3600 VCore seconds. You can then use this adjusted number to compare with the VCore seconds used by YARN or YARN pools.

Retrieving metric data

You can vire the Cloudera Manager Service Monitor data storage granularities in Cloudera Manager.

There is a Time series endpoint exposed by the Cloudera Manager REST API. The API accepts tsquery statements as input for which metrics need to be retrieved during the specified time window. The API provides functionality to specify the desired data granularity (for example, raw metric values, TEN_MINUTES, HOURLY etc.). Each granularity level of data is maintained in a leveldb table. This data is aggregated from raw metric values such as minimum, maximum, etc. within the corresponding data window.

For example, if you do not need the metric data at a specific timestamp but care more about the hourly usage, HOURLY data should be good enough. In general, the longer the granular window it is, the less storage it is taking, and thus the longer period of time you are able to keep that level of data without being purged when the storage raches the configured limit. In the case of Cloudera Manager Cluster Utilization Reports, Cloudera Manager generates the reports based on an hourly window.

To view the Cloudera Manager Service Monitor data storage granularities, go to ClustersCloudera Management ServiceService MonitorCharts LibraryService Monitor Storage and scroll down to see the Data Duration Covered table to see the earliest available data points for each level of granularity. The value in the last(duration_covered) column indicates the age of the oldest data in the table.

Data Duration Covered

September 14, 2016 3:40 PM

Search

Entity	last(duration_covered)
impala-query-monitoring - profiles (RAW)	9.4h
impala-query-monitoring - profiles_end_time (RAW)	9.4h
impala-query-monitoring - queries (RAW)	9.4h
service-monitoring - reports_stream (RAW)	9.4h
service-monitoring - reports_type (RAW)	9.4h
service-monitoring - stream (RAW)	9.4h
service-monitoring - subject_ts (RAW)	9.4h
service-monitoring - ts_stream_rollup_PT21600S (SIX_HOURLY)	9.4h
service-monitoring - ts_stream_rollup_PT3600S (HOURLY)	9.4h
service-monitoring - ts_stream_rollup_PT600S (TEN_MINUTELY)	9.4h
service-monitoring - ts_stream_rollup_PT604800S (WEEKLY)	9.4h
service-monitoring - ts_stream_rollup_PT86400S (DAILY)	9.4h
service-monitoring - ts_subject (RAW)	9.4h
service-monitoring - ts_type_rollup_PT21600S (SIX_HOURLY)	9.4h
service-monitoring - ts_type_rollup_PT3600S (HOURLY)	9.4h
service-monitoring - ts_type_rollup_PT600S (TEN_MINUTELY)	9.4h
service-monitoring - ts_type_rollup_PT604800S (WEEKLY)	9.4h
service-monitoring - ts_type_rollup_PT86400S (DAILY)	9.4h
service-monitoring - type (RAW)	9.4h
yarn-application-monitoring - application_details (RAW)	9.4h
yarn-application-monitoring - applications (RAW)	9.4h
yarn-application-monitoring - applications_end_time (RAW)	9.4h

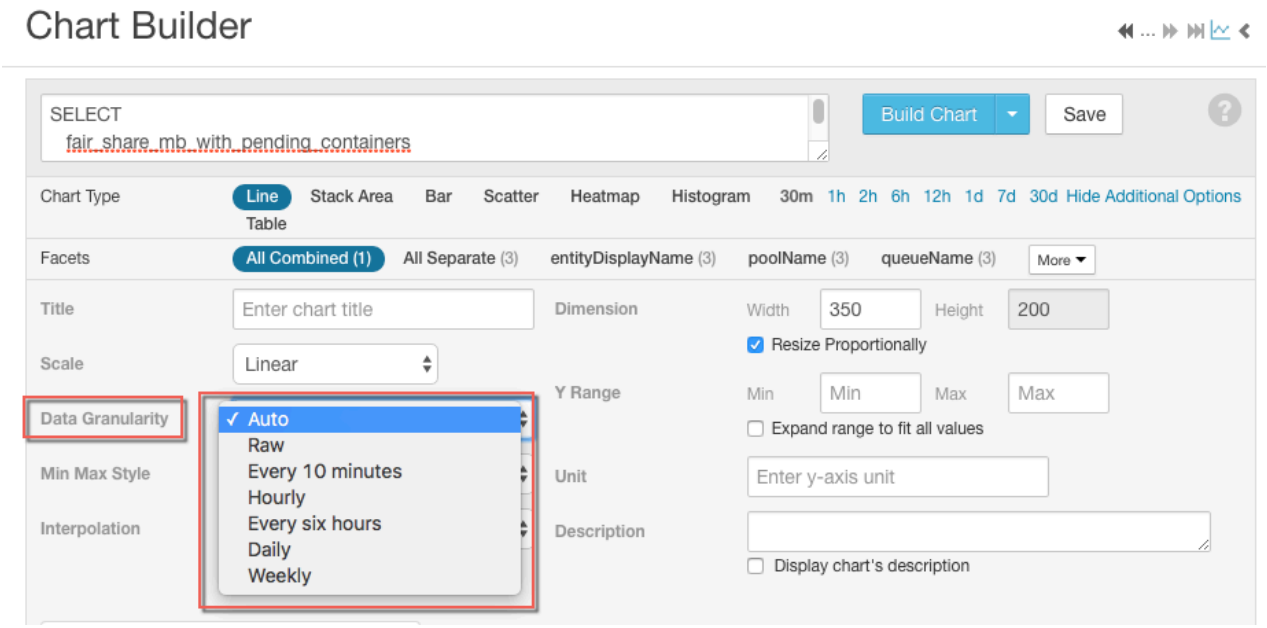
To configure the Time series storage used by the Service Monitor, go to ClustersCloudera Management ServiceConfiguration and search for "Time-Series Storage".

Querying metric data

You can build charts that query time series data using the Cloudera Manager Admin console.

Go to Charts Chart Builder . When building charts, it may be useful to choose the data granularity by clicking the Show additional options link on the chart builder page and then selecting the Data Granularity drop-down list.

Selecting data granularity in chart builder:



Inspecting Network Performance

The Network Performance Inspector allows you to examine the network performance among hosts and clusters managed by an instance of Cloudera Manager. You can use this tool to diagnose networking issues that can significantly affect the performance of workloads such as MapReduce jobs, Spark jobs, Hive queries, and Impala queries, particularly when using remote storage. You can run the inspector on-demand, and it also available when adding a new cluster. You can also run the inspector using the Cloudera Manager API.

The Network Performance Inspector runs the following types of inspections:

- Latency test

  The inspector runs ping commands from each host to all other hosts, and reports the average ping time and packet loss percentage.

  You can run the latency test among the hosts in a single cluster and between hosts in different clusters.
- Bandwidth Test

  The bandwidth test measures the speed of data transfer between the hosts. You can use this information to identify problematic hosts or networking infrastructure issues so that you take corrective action.

  You can run the bandwidth test only between clusters managed by an instance of Cloudera Manager.

Important:

The Network Performance Inspector uses the `ping` command on each host in the cluster to check the latency between the hosts. In some operating system configurations, only the root user can run the `ping` command. Otherwise, it has insufficient privileges to open a raw socket in order to send the required ICMP messages.

If the `ping` command has insufficient privileges, then the Network Performance Inspector displays the following error message:invocation of script- /opt/cloudera/cm-agent/service/perf/host_perf_diag.py could not execute due to - "ping: socket: Operation not permitted"

To avoid this error, you must ensure that the `ping` command have the setuid permissions (not recommended) or CAP_NET_RAW capability enabled to open a raw socket.

Include the specific capability for the `ping` command on each host in the cluster by running the following command::

```
# setcap cap_net_admin,cap_net_raw=+p $(which ping)
```

Check the current capabilities by running the following command:

```
# getcap $(which ping)
```

Running the Network Performance Inspector From the Cloudera Manager Admin Console

Minimum Required Role: Key Administrator or Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

To run the Network Performance Inspector:

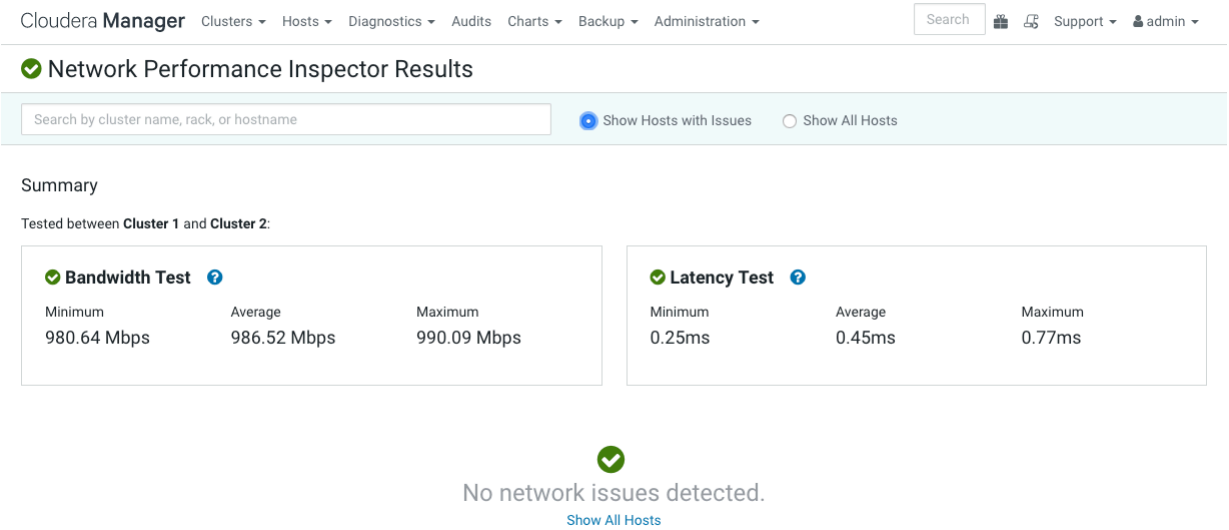
1. Open the Network Performance Inspector from one of the following pages in the Cloudera Manager Admin Console:
   - From the All Hosts page:
     a. Select HostsAll Hosts.
     b. Click the Inspect Network Performance button to launch the inspector.
   - From the Cluster Status page:
     a. Select Clusters*Cluster Name*
     b. In the Status section, click the Hosts link at the top of the list of Cluster Services to open the Hosts page.
     c. Click the Inspect Cluster Network Performance button to launch the inspector.

   The Inspect Network Performance dialog box opens.
2. Enter the following information:
   - Source Cluster – Select a cluster to inspect from the drop-down list.

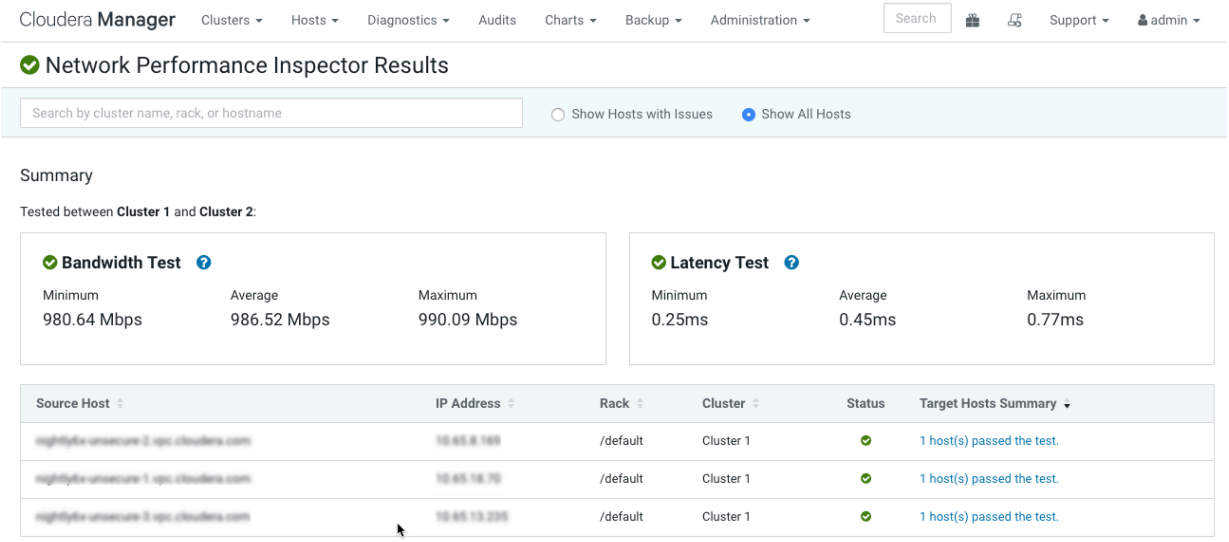
     Select Run against another cluster to select additional clusters. An additional drop-down list displays from which you can select an additional cluster to inspect. Select either the Latency Test or the Latency and Bandwidth Test option.
   - Ping Timeout – Amount of time, in seconds, after which the inspector reports a failure.
   - Ping Count – Number of times the inspector pings each host.
   - Ping Packet Size – Size, in bytes, of the test packet sent when pinging the hosts.

3. Click Run.

The Cluster Performance Inspector command window opens. When the Inspector finishes, click the Show Inspector Results button to open the Network Performance Inspector Results page.



Select Show Hosts with Issues to display any problems found by the inspector, or select Show All Hosts. The Network Diagnostic Result page displays a table of results. Each row represents a single host. The Target Hosts Summary column summarizes the performance of the host. Click the summary text to view detailed performance statistics from this host to each of the other hosts.



The inspector summarizes the latency and bandwidth tests for the hosts using three icons:

Table 2:

Status	Bandwidth Test	Latency Test
Green Good network performance.	More than 500 MBits/Second	No packet loss.
Orange Concerning network performance	250 - 500 MBits/Second	Any host with a ping latency in the range of 1 to 4 milliseconds.

Status	Bandwidth Test	Latency Test
Red Bad network performance.	Less than 250 MBits/Second	Any target host that is unreachable by hostname (has a 100% packet loss), any host with a ping latency greater than 4 milliseconds, or any target host with a packet loss of 1% or greater.

Poor performance can result from firewall settings, router configurations, network topology, and other factors. You may need to work with your network administrator to mitigate these issues.

Running the Network Performance Inspector From the Cloudera Manager API

You can invoke the Network Performance Inspector using the Cloudera Manager API and the following endpoints:

- /cm/commmands/hostsPerfInspector

  Invokes the inspector across an arbitrary set of hosts, (including hosts that are not part of the cluster).
- /cm/commands/clusterPerfInspector

  Invokes the inspector across the hosts in two clusters.
- /cm/clusters/*cluster Name*/commands/perfInspector

  Invokes the inspector across the hosts of a specified cluster.

For more information, see the Cloudera Manager API REST documentation.

Monitoring Services

Cloudera Manager's Service Monitoring feature monitors dozens of service health and performance metrics about the services and role instances running on your cluster.

Service Monitoring includes the following functions:

- Presents health and performance data in a variety of formats, including interactive charts.
- Monitors metrics against configurable thresholds.
- Generates events related to system and service health and critical log entries and makes them available for searching and alerting.
- Maintains a complete record of service-related actions and configuration changes.

The following topics describe how to monitor the services and role instances installed on your cluster:

Monitoring Service Status

From a service page, you can monitor the status of services, manage services and roles, and more.

From a service page, you can:

- Monitor the status of the services running on your clusters.
- Manage the services and roles in your clusters.
- Add new services.
- Access the client configuration files generated by Cloudera Manager that enable Hadoop client users to work with the HDFS, MapReduce, HBase, and YARN services you added. (These configuration files are normally deployed automatically when you install a cluster or add a service).
- View the maintenance mode status of a cluster.

You can also pull down a menu from an individual service name to go directly to one of the tabs for that service to its Status, Instances, Commands, Configuration, Audits, or Charts Library tabs.

Viewing the URLs of the Client Configuration Files

To allow Hadoop client users to work with the services you created, Cloudera Manager generates client configuration files that contain the relevant configuration files with the settings from your services.

About this task

These files are deployed automatically by Cloudera Manager based on the services you have installed, when you add a service, or when you add a Gateway role on a host. You can manually download and distribute these client configuration files to the users of a service, if necessary.

Procedure

1. Go to the Status page for a cluster. (Clusters*cluster name*).
2. Click ActionsView Client Configuration URLs.

   The Actions button is not enabled if you are viewing status for a point of time in the past.

   A pop-up window displays links to the client configuration zip files.
3. Click a link to download a zip file.

Related Information

Client Configuration Files

Viewing the Status of a Service Instance

You can view the status of a service instance from the **Status** page or the **Clusters** menu.

Procedure

1. Click the Cloudera Manager logo to go to the Home page.
2. Access the service.
   - In the **Status** tab, select *ClusterName* *ServiceName* .
   - Select Clusters *ClusterName* *ServiceName* .

Results

The **Status** page opens. On the **Status** page you can view a variety of information about a service and its performance.

Related Information

Viewing Service Status

Viewing the Health and Status of a Role Instance

You can view the health and status of a role instance by clicking the role instance under the Role Counts column.

If there is just one instance of this role, this opens the **Status** tab for the role instance.

If there are multiple instances of a role, clicking the role link under Role Counts will open the **Instances** tab for the service, showing instances of the role type you have selected.

If you are viewing a point in time in the past, the Role Count links will be greyed out, but still functional. Their behavior will depend on whether historical data is available for the role instance.

Related Information

Viewing Role Instance Status

Viewing the Maintenance Mode Status of a Cluster

For any cluster, you can view the components (service, roles, or hosts) that are in maintenance mode.

About this task

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Procedure

1. From the Cloudera Manager Home page, select the cluster that you want to view the maintenance mode status for.
2. Click Actions View Maintenance Mode Status... .
   This pops up a dialog box that shows the components in your cluster that are in maintenance mode, and indicates which are in effective maintenance mode as well as those that have been explicitly placed into maintenance mode.

   From this dialog box you can select any of the components shown there and remove them from maintenance mode.

   If individual services are in maintenance mode, you will see the maintenance mode icon next to the Actions button for that service.

   Note: The Actions button is not enabled if you are viewing status for a point of time in the past.

Viewing Service Status

You can view a summary of the status for each service.

Procedure

1. Click the Cloudera Manager logo to go to the Home page.
2. Access the status summary.
   - In the Home Status tab, if the cluster is displayed in full form, click *ServiceName* in a *ClusterName* table.
   - In the Home Status tab, click *ClusterName* and then click *ServiceName*.
   - Select Clusters *ClusterName* *ServiceName* .

   For all service types there is a Status Summary that shows, for each configured role, the overall status and health of the role instance(s).

   Note: Not all service types provide complete monitoring and health information. Hive, Hue, Oozie, Solr, and YARN only provide the basic Status Summary.

   Each service that supports monitoring provides a set of monitoring properties where you can enable or disable health tests and events, and set thresholds for tests and modify thresholds for the status of certain health tests.

   The HDFS, MapReduce, HBase, ZooKeeper, and Flume services also provide additional information: a snapshot of service-specific metrics, health test results, health history, and a set of charts that provide a historical view of metrics of interest.

Related Information

Status Summary

Configuring Monitoring Settings

Viewing Past Status

You can expand the Time Range Selector to view historical health, status, and chart data.

The health and status information on the **Status** page represents the state of the service or role instance at a given point in time. The charts (and the Logs and Events under Diagnostics) represent the time range selected on the Time Range Selector, which defaults to the past 30 minutes. You can view health, status, and chart historical data by expanding the Time Range Selector. Click the mini line chart under "admin" and move the time marker ( ) to a point in the past.

When you move the time marker to a point in the past (for services and roles that support health history), the entire **Status** page updates to the time selected. A Now button (⏭) allows you to quickly return to the current state of the service. The Actions menu is disabled while viewing a past status to ensure that you cannot accidentally act on outdated status information.

Related Information

Time Line

Status Summary

The Status Summary shows the status of each service instance being managed by Cloudera Manager.

Even services such as Hue, Oozie, or YARN (which are not monitored by Cloudera Manager) show a status summary. The overall status for a service is a roll-up of the health test results for the service and all its role instances. The Status can be:

Table 3: Status

Indicator	Status	Description
	Started with outdated configuration	For a service, this indicates the service is running, but at least one of its roles is running with a configuration that does not match the current configuration settings in Cloudera Manager. For a role, this indicates a configuration change has been made that requires a restart, and that restart has not yet occurred.
	Starting	The entity is starting up but is not yet running.
	Stopping	The entity is stopping but has not stopped yet.
	Stopped	The entity is stopped, as expected.
	Down	The entity is not running, but it is expected to be running.
	History not available	Cloudera Manager is in historical mode, and the entity does not have historical monitoring support. This is the case for services other than HDFS, MapReduce and HBase such as ZooKeeper, Oozie, and Hue.
	None	The entity does not have a status. For example, it is not something that can be running and it cannot have health. Examples are the HDFS Balancer (which runs from the HDFS Rebalance action) or Gateway roles. The Start and Stop commands are not applicable to these instances.
	Good health	The entity is running with good health. For a specific health test, the returned result is normal or within the acceptable range. For a role or service, this means all health tests for that role or service are Good.
	Concerning health	The entity is running with concerning health. For a specific health test, the returned result indicates a potential problem. Typically this means the test result has gone above (or below) a configured Warning threshold. For a role or service, this means that at least one health test is Concerning.
	Bad health	The entity is running with bad health. For a specific health test, the test failed, or the returned result indicates a serious problem. Typically this means the test result has gone above (or below) a configured Critical threshold. For a role or service, this means that at least one health test is Bad.
	Disabled health	The entity is running, but all of its health tests are disabled.
	Unknown health	The status of a service or role instance is unknown. This can occur for a number of reasons, such as the Service Monitor is not running, or connectivity to the Agent doing the health monitoring has been lost.

To see the status of one or more role instances, click the role type link under Status Summary. If there is a single instance of the role type, the link directs you to the Status page of the role instance.

If there are multiple role instances (such as for DataNodes, TaskTrackers, and RegionServers), the role type link directs you to the Role Instances page for that role type. Click on each instance, under Role Type, to be taken to the corresponding Status page.

To display the results for each health test that applies to this role type, expand the Health Tests filter on the left and expand Good Health, Warnings, Bad Health, or Disabled Health. Health test results that have been filtered out by your role type selection appear as unavailable.

Related Information

Stale Configurations

Viewing Role Instance Status

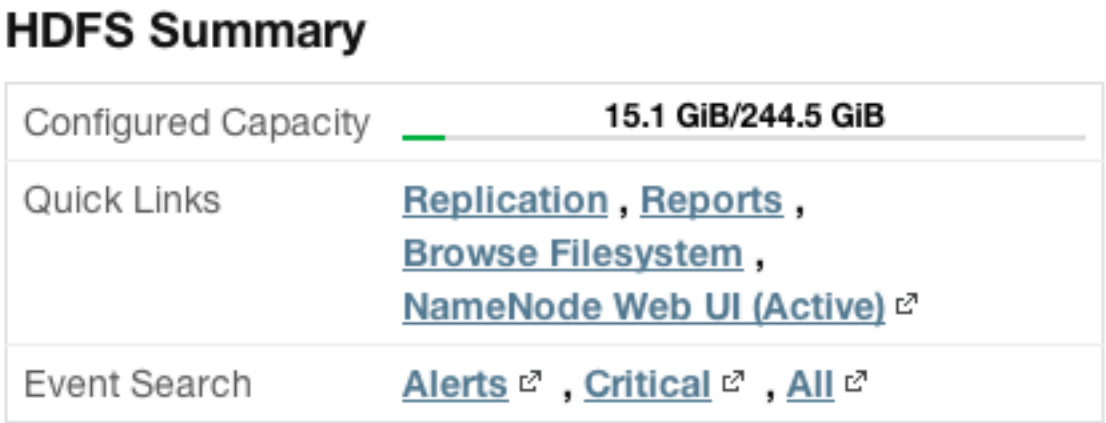
Service Summary

Some services (specifically HDFS, MapReduce, HBase, Flume, and ZooKeeper) provide additional statistics about their operation and performance. These are shown in a Summary panel at the left side of the page.

The contents of this panel depend on the service:

- The HDFS Summary shows disk space usage.
- The MapReduce Summary shows statistics on slot usage, jobs and so on.
- The Flume Summary provides a link to a page of Flume metric details.
- The ZooKeeper Summary provides links to the ZooKeeper role instances (nodes) as well as Zxid information if you have a ZooKeeper Quorum (multiple ZooKeeper servers).

For example:



Other services such as Hue, Oozie, Impala, and Cloudera Manager itself, do not provide a Service Summary.

Health Tests and Health History

The **Health Tests** panel shows health test results in an expandable and collapsible list, typically with the specific metrics that the test returned. You can Expand All or Collapse All from the links at the upper right of the Health Tests panel.

The **Health Tests** and **Health History** panels appear for HDFS, MapReduce, HBase, Flume, Impala, ZooKeeper, and the Cloudera Manager Service. Other services such as Hue, Oozie, and YARN do not provide a Health Test panel.

- The color of the text (and the background color of the field) for a Health Test result indicates the status of the results. The tests are sorted by their health status – Good, Concerning, Bad, or Disabled. The entries are collapsed by default. Click the arrow to the left of an entry to expand the entry and display further information.
- Clicking the Details link for a health test displays further information about the test, such as the meaning of the test and its possible results, suggestions for actions you can take or how to make configuration changes related to the test. The help text may include a link to the relevant monitoring configuration section for the service.

- In the **Health Tests** panel:
  - Clicking ❯ displays the lists of health tests that contributed to the health test.
  - Clicking the Details link displays further information about the health test.
- In the **Health History** panel:
  - Clicking ❯ displays the lists of health tests that contributed to the health history.
  - Clicking the Show link moves the time range to the historical time period.

Related Information

Configuring Monitoring Settings

Viewing Service Instance Details

You can view service instance details such as the name of the role instance, the host on which it is running, the rack assignment, and more.

Procedure

1. Do one of the following:
   - In the HomeStatus tab, if the cluster is displayed in full form, click *ServiceName* in a *ClusterName* table.
   - In the HomeStatus tab, click *ClusterName* and then click *ServiceName*.
   - Select Clusters*ClusterNameServiceName*.
2. Click the Instances tab on the service's navigation bar. This shows all instances of all role types configured for the selected service.

Results

The Instances page displays the results of the configuration validation checks it performs for all the role instances for this service.

Note: The information on this page is always the Current information for the selected service and roles. This page does not support a historical view: thus, the Time Range Selector is not available.

The information on this page shows:

- The name of the role instance. Click the name to view the role status for that role.
- The host on which it is running. Click the hostname to view the host status details for the host.
- The rack assignment.
- The status. A single value summarizing the state and health of the role instance.
- Whether the role is currently in maintenance mode. If the role has been set into maintenance mode explicitly, you will see the following icon ( ). If it is in effective maintenance mode due to the service or its host having been set into maintenance mode, the icon will be this ( ).
- Whether the role is currently decommissioned.

What to do next

You can sort or filter the Instances list by criteria in any of the displayed columns. To filter, type a property value in the Search box or select the value from the facets at the left of the page.

Related Information

Viewing Role Instance Status

Host Details

Role Instance Reference

The following tables contain reference information on the status, role state, and health columns for role instances.

Table 4: Status

Indicator	Status	Description
	Started with outdated configuration	For a service, this indicates the service is running, but at least one of its roles is running with a configuration that does not match the current configuration settings in Cloudera Manager. For a role, this indicates a configuration change has been made that requires a restart, and that restart has not yet occurred.
	Starting	The entity is starting up but is not yet running.
	Stopping	The entity is stopping but has not stopped yet.
	Stopped	The entity is stopped, as expected.
	Down	The entity is not running, but it is expected to be running.
	History not available	Cloudera Manager is in historical mode, and the entity does not have historical monitoring support. This is the case for services other than HDFS, MapReduce and HBase such as ZooKeeper, Oozie, and Hue.
	None	The entity does not have a status. For example, it is not something that can be running and it cannot have health. Examples are the HDFS Balancer (which runs from the HDFS Rebalance action) or Gateway roles. The Start and Stop commands are not applicable to these instances.
	Good health	The entity is running with good health. For a specific health test, the returned result is normal or within the acceptable range. For a role or service, this means all health tests for that role or service are Good.
	Concerning health	The entity is running with concerning health. For a specific health test, the returned result indicates a potential problem. Typically this means the test result has gone above (or below) a configured Warning threshold. For a role or service, this means that at least one health test is Concerning.
	Bad health	The entity is running with bad health. For a specific health test, the test failed, or the returned result indicates a serious problem. Typically this means the test result has gone above (or below) a configured Critical threshold. For a role or service, this means that at least one health test is Bad.
	Disabled health	The entity is running, but all of its health tests are disabled.
	Unknown health	The status of a service or role instance is unknown. This can occur for a number of reasons, such as the Service Monitor is not running, or connectivity to the Agent doing the health monitoring has been lost.

Related Information

Stale Configurations

Viewing Role Instance Status

You can view the status for a role instance.

Procedure

1. Select a service instance to display the Status page for that service.
2. Click the Instances tab.
3. From the list of roles, select one to display that role instance's Status page.

The Actions Menu

The Actions menu provides a list of commands relevant to the role type you are viewing. These commands typically include Stopping, Starting, or Restarting the role instance, accessing the Web UI for the role, and may include many other commands, depending on the role you are viewing.

Minimum Required Role: Operator (also provided by Configurator, Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

The Actions menu is available from the Role Status page only when you are viewing Current time status. The menu is disabled if you are viewing a point of time in the past.

Viewing Past Status

You can expand the Time Range Selector to view historical health, status, and chart data.

The status and health information shown on this page represents the state of the service or role instance at a given point in time. The exceptions are the charts tabs, which show information for the time range currently selected on the Time Range Selector, which defaults to the past 30 minutes. By default, the information shown on this page is for the current time. You can view status for a past point in time by moving the time marker (◆) to a point in the past.

When you move the time marker to a point in the past (for Services/Roles that support health history), the Health Status clearly indicates that it is referring to a past time. A Now button (⏭) enables you to quickly switch to view the current state of the service. In addition, the Actions menu is disabled while you are viewing status in the past – to ensure that you cannot accidentally take an action based on outdated status information.

You can also view past status by clicking the Show link in the **Health Tests and Health History** panel.

Related Information

Time Line

Health Tests and Health History

Summary

The **Summary** panel provides basic information about the role instance, where it resides, and the health of its host.

All role types provide the **Summary** panel. Some role instances related to HDFS, MapReduce, and HBase also provide a **Health Tests** panel and associated charts.

Health Tests and Health History

The **Health Tests** panel shows health test results in an expandable and collapsible list, typically with the specific metrics that the test returned. You can Expand All or Collapse All from the links at the upper right of the Health Tests panel.

The **Health Tests** and **Health History** panels appear for HDFS, MapReduce, HBase, Flume, Impala, ZooKeeper, and the Cloudera Manager Service. Other services such as Hue, Oozie, and YARN do not provide a Health Test panel.

- The color of the text (and the background color of the field) for a Health Test result indicates the status of the results. The tests are sorted by their health status – Good, Concerning, Bad, or Disabled. The entries are collapsed by default. Click the arrow to the left of an entry to expand the entry and display further information.
- Clicking the Details link for a health test displays further information about the test, such as the meaning of the test and its possible results, suggestions for actions you can take or how to make configuration changes related to the test. The help text may include a link to the relevant monitoring configuration section for the service.
- In the **Health Tests** panel:
  - Clicking ❯ displays the lists of health tests that contributed to the health test.
  - Clicking the Details link displays further information about the health test.
- In the **Health History** panel:
  - Clicking ❯ displays the lists of health tests that contributed to the health history.
  - Clicking the Show link moves the time range to the historical time period.

Related Information

Configuring Monitoring Settings

Status Summary

The **Status Summary** panel reports a roll-up of the status of all the roles.

Related Information

Role Instance Reference

Charts

Charts are shown for roles that are related to HDFS, MapReduce, HBase, ZooKeeper, Flume, and Cloudera Management Service. Roles related to other services such as Hue, Hive, Oozie, and YARN, do not provide charts.

See the topic *Viewing Charts for Cluster, Service, Role, and Host Instances* for detailed information on the charts that are presented, and the ability to search and display metrics of your choice.

Related Information

Viewing Charts for Cluster, Service, Role, and Host Instances

The Processes Tab

The Processes page shows the processes that run as part of this service role, with a variety of metrics about those processes.

Procedure

1. To view the processes running for a role instance, select a service instance to display the Status page for that service.
2. Click the Instances tab.
3. From the list of roles, select one to display that role instance's Status page.
4. Click the Processes tab.

What to do next

- To see the location of a process' configuration files, and to view the Environment variable settings, click the Show link under Configuration Files/Environment.
- If the process provides a Web UI (as is the case for the NameNode, for example) click the link to open the Web UI for that process.
- To see the most recent log entries, click the Show Recent Logs link.
- To see the full log, stderr, or stdout log files, click the appropriate links.

Running Diagnostic Commands for Roles

You can run diagnostic utility tools such as "collect stack traces" and "heap dump" against most role processes.

About this task

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Cloudera Manager allows administrators to run the following diagnostic utility tools against most Java-based role processes:

- List Open Files (lsof) - Lists the open files of the process.
- Collect Stack Traces (jstack) - Captures Java thread stack traces for the process.
- Heap Dump (jmap) - Captures a heap dump for the process.
- Heap Histogram (jmap -histo) - Produces a histogram of the heap for the process.

These commands are found on the Actions menu of the Cloudera Manager page for the instance of the role. For example, to run diagnostics commands for the HDFS active NameNode, perform these steps:

Procedure

1. Click the HDFS service on the HomeStatus tab or select it on the Clusters menu.
2. Click InstancesNameNode (Active).
3. Click the Actions menu.
4. Choose one of the diagnostics commands listed in the lower section of the menu.
5. Click the button in the confirmation dialog box to confirm your choice.
6. When the command is executed, click Download Result Data and save the file to view the command output.

Periodic Stacks Collection

Periodic stacks collection allows you to enable and configure the periodic collection of thread stack traces in Cloudera Manager.

When stacks collection is enabled for a role, call stacks are output to a log file at regular intervals. The logs can help with diagnosis of performance issues such as deadlock, slow processing, or excessive numbers of threads.

Stacks collection may impact performance for the processes being collected as well as other processes on the host, and is turned off by default. For troubleshooting performance issues, you may be asked by Cloudera Support to enable stacks collection and send the resulting logs to Cloudera for analysis.

Stacks collection is available for the majority of roles in Cloudera Manager. For the HDFS service, for example, you can enable stacks collection for the DataNode, NameNode, Failover Controller, HttpFS, JournalNode, and NFS Gateway. If the Stacks Collection category does not appear in the role's configuration settings, the feature is not available for that role.

Configuring Periodic Stacks Collection

You can enable and configure periodic stacks collection.

About this task

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Open the Cloudera Manager page for a specific service or role.
2. Access the configuration settings in one of the following ways:
   - **a.** From the service page in Cloudera Manager, click the Configuration tab.
     b. Select Scope NameNode .
     c. Select Category Stacks Collection .
   - **a.** From the service page in Cloudera Manager, click the Instances tab.
     b. Click the Configuration tab.
     c. Select Scope *role type* .
     d. Select Category Stacks Collection .

The configuration settings are as follows:

- Stacks Collection Enabled - Whether or not periodic stacks collection is enabled.
- Stacks Collection Directory - The directory in which stack logs will be placed. If not set, stacks will be logged into a stacks subdirectory of the role's log directory.
- Stacks Collection Frequency - The frequency with which stacks will be collected.

- Stacks Collection Data Retention - The amount of stacks data that will be retained. When the retention limit is reached, the oldest data will be deleted.
- Stacks Collection Method - The method that will be used to collect stacks. The jstack option involves periodically running the jstack command against the role's daemon process. The servlet method is available for those roles with an HTTP server endpoint that exposes the current stacks traces of all threads. When the servlet method is selected, that HTTP endpoint is periodically scraped.

Example

As an example, to configure stacks collection for an HDFS NameNode, complete the following steps:

1. Go to the HDFS service page.
2. Click the Configuration tab.
3. Select Scope NameNode .
4. Select Category Stacks Collection .
5. Locate the property or search for it by typing its name in the Search box.
6. Modify the configuration settings if desired.
7. Click Save Changes.

Stacks collection configuration settings are stored in a per-role configuration file called cloudera-stacks-monitor.properties. Cloudera Manager reads the configuration file and coordinates stack collection. Changes to the configuration settings take effect after a short delay. It is not necessary to restart the role.

Viewing and Downloading Stacks Logs

Stacks are collected and logged to a compressed, rotated log file. You can view and download stacks logs.

About this task

A certain amount of the log data is in an uncompressed file. When that file reaches a limit, the file is rotated and bzip2 compressed. Once the total number of files exceeds the configured retention limit, the oldest files are deleted. Collected stacks data is available for download through the Cloudera Manager UI and API. To view or download stacks logs through the UI, complete the following steps:

Procedure

1. On the service page, click the Instances tab.
2. Click the role in the Role Type column.
3. Click the Stacks Logs tab.
4. Click Stacks Log File to view the most recent stacks file. Click Download Stacks Logs to download a zipped bundle of the stacks logs.

Viewing Running and Recent Commands

You can view and run recent commands for a cluster, service, or role. You can also view the details of recent or running commands.

Viewing Running and Recent Commands For a Cluster

You can use the command indicator to view the running and recent commands for a cluster.

Procedure

1. Click the Recent Commands in the left navigation panel.
   The indicator displays the number of commands currently running for all services or roles.

2. To display all commands that have run and finished recently, do one of the following:
   - Click the All Recent Commands button in the window that pops up when you click the indicator. This command displays information on all running and recent commands in the same form, as described below.
   - Click the Cloudera Manager logo in the main navigation bar and click the All Recent Commands tab.

   The command indicator shows the number of commands running on all clusters you are managing. Likewise, All Recent Commands shows all commands that were run and finished within the search time range you specified, across all your managed clusters.
3. Select a value from the pager

   to control how many commands are listed, or click the arrows to view pages.

Viewing Running and Recent Commands for a Service or Role

You can view running and recent commands for a service or role instance in the **Commands** tab.

About this task

For a selected service or role instance, the **Commands** tab shows which commands are running or have been run for that instance, and what the status, progress, and results are. For example, if you go to the HDFS service shortly after you have installed your cluster and look at the **Commands** tab, you will see recent commands that created the directories, started the HDFS role instances (the NameNode, Secondary NameNode, and DataNode instances), and the command that initially formatted HDFS on the NameNode. This information is useful if a service or role seems to be taking a long time to start up or shut down, or if services or roles are not running or do not appear to have been started correctly. You can view both the status and progress of currently running commands, as well as the status and results of commands run in the past.

Procedure

1. Click the **Clusters** tab on the top navigation bar.
2. Click the service name to go to the Status tab for that service.
3. For a role instance, click the Instances tab and select the role instance name to go to its **Status** tab.
4. Click the **Commands** tab.

Command Details

You can view command details. The details available for a command depend on whether the command is running or recently completed.

Running Commands

The Running Commands area shows commands that are in progress.

While the status of the command is In Progress, an Abort button displays so that you can abort the command if necessary.

The Commands status information is updated automatically while the command is running.

After the command has finished running (all its subcommands have finished), the status is updated, the Abort buttons disappear, and the information for Recent Commands appears as described below.

Recent Commands

The Recent Commands area shows commands that were run and finished within the search time range you specified.

If no commands were run during the selected time range, you can double the time range selection by clicking the Try expanding the time range selection link. If you are in the "current time" mode, the beginning time will move; if you

are looking at a time range in the past, both the beginning and ending times of the range are changed. You can also change the time range using the options described in the topic *Time Line*.

Select a value from the pager

to control how many commands are listed, or click the arrows to view pages.

Commands are shown with the most recent ones at the top.

The icon associated with the status (which typically includes the time that the command finished) plus the result message tells you whether the command succeeded ✔ or failed ⊗. If the command failed, it indicates if it was one of the subcommands that actually failed. In many cases, multiple subcommands result from the top level command.

The First Run command runs during the initial startup of your cluster. Click this link to view the command history of the cluster startup.

Command Details

In the Running Commands dialog box or Recent Commands page, click a command in the Command column to display its details and any subcommands. The page title is the name of the command.

The **Summary** section at the top shows information about the command:

- The current status
- The context, which can be a cluster, service, host, or role
- The time the command started
- The duration of the command
- A message about the command completion
- If the context is a role, links to role instance logs

The **Details** section shows how many steps, if any, the selected command has and lists any subcommands.

Expand a command to view subcommands. In the Running Commands dialog box, each subcommand also has an Abort button that is present as long as the subcommand is in progress.

You can perform the following actions:

- Select the option to display all the subcommands or only failed or running commands.
- Click the link in the Context column to go to the **Status** page for the component (host, service, or role instance) to which this command is related.
- Click a Role Log tab to display the log for that role, and stdout and stderr if available for the role.

Related Information

Time Line

Monitoring Hosts

Cloudera Manager's Host Monitoring features let you manage and monitor the status of the hosts in your clusters.

Viewing All Hosts

You can view summary information about all the hosts managed by Cloudera Manager on the **All Hosts** page.

Procedure

1. To display summary information about all the hosts managed by Cloudera Manager, click Hosts in the main navigation bar.

   The list of hosts shows the overall status of the Cloudera Manager-managed hosts in your cluster.
2. Optionally, to change the columns, click the Columns: *n* Selected drop-down and select the checkboxes next to the columns to display.

   The information provided varies depending on which columns are selected.
3. Optionally, click › to the left of the number of roles to list all the role instances running on that host.
4. Optionally, filter the hosts list by entering search terms (hostname, IP address, or role) in the search box separated by commas or spaces. Use quotes for exact matches (for example, strings that contain spaces, such as a role name) and brackets to search for ranges. Hosts that match any of the search terms are displayed. For example:

   ```
   hostname[1-3], hostname8 hostname9, "hostname.example.com"
   hostname.example.com “HDFS DataNode”
   ```

   You can also search for hosts by selecting a value from the facets in the Filters section at the left of the page.

   - If the Agent Heartbeat and Health Status options are configured as follows:
     - Send Agent heartbeat every *x*
     - Set health status to Concerning if the Agent heartbeats fail *y*
     - Set health status to Bad if the Agent heartbeats fail *z*

     The value *v* for a host's Last Heartbeat facet is computed as follows:

     - $v < x * y$ = Good
     - $v >= x * y$ and $<= x * z$ = Concerning
     - $v >= x * z$ = Bad

Related Information

Configuring Agent Heartbeat and Health Status Options

Role Assignments

You can view the assignment of roles to hosts from the **Roles** tab.

Procedure

1. Click the Roles tab.
2. Click a cluster name or All Clusters.

Viewing the Disks Overview

You can view an overview of the status of all disks in the deployment. The statistics displayed match or build on those in iostat, and are shown in a series of histograms that by default cover every physical disk in the system.

Procedure

1. Click HostsDisks Overview.
2. Optionally, adjust the endpoints of the time line to see the statistics for different time periods.
3. Optionally, specify a filter in the box to limit the displayed data. For example, to see the disks for a single rack rack1, set the filter to: logicalPartition = false and rackId = "rack1" and click Filter.
4. Optionally, click a histogram to drill down and identify outliers. Mouse over the graph and click ⤢ to display additional information about the chart.

Viewing the Hosts in a Cluster

You can view the hosts in a cluster. The **All Hosts** page displays a list of the hosts filtered by the cluster name.

Procedure

Do one of the following:

- Select Clusters*Cluster name*Hosts.
- In the **Home** screen, click Hosts in a full-form cluster table.

Viewing Individual Hosts

You can view detailed information about an individual host—resources (CPU/memory/storage) used and available, which processes it is running, details about the host agent, and much more—by clicking a host link on the **All Hosts** page.

Related Information

Host Details

Host Details

You can view details about each host from the status page for each host.

The host details include:

- Name, IP address, rack ID
- Health status of the host and last time the Cloudera Manager Agent sent a heartbeat to the Cloudera Manager Server
- Number of cores
- System load averages for the past 1, 5, and 15 minutes
- Memory usage
- File system disks, their mount points, and usage

  Important: If you have multiple mount points under the same device, then the available free space on that device is counted multiple times and adds to the total available disk space.

- Health test results for the host
- Charts showing a variety of metrics and health test results over time.
- Role instances running on the host and their health
- CPU, memory, and disk resources used for each role instance

Viewing Host Details

You can view detailed information about each host, such as name, IP address, and rack ID, and more from the **All Hosts** page.

Procedure

1. To view detailed host information, click HostsAll Hosts.
2. Click the name of one of the hosts. The Status page is displayed for the host you selected.
3. Click tabs to access specific categories of information. Each tab provides various categories of information about the host, its services, components, and configuration.

Status

The **Status** page is displayed when a host is initially selected. It provides summary information about the status of the selected host. Use the **Status** page to gain an understanding of work being done by the system, the configuration, and health status.

If this host has been decommissioned or is in maintenance mode, you will see the following icon(s) (,) in the top bar of the page next to the status message.

Details

This panel provides basic system configuration such as the host's IP address, rack, health status summary, and disk and CPU resources. This information summarizes much of the detailed information provided in other panes on this tab. To view details about the Host agent, click the Host Agent link in the Details section.

Health Tests

Cloudera Manager monitors a variety of metrics that are used to indicate whether a host is functioning as expected. The Health Tests panel shows health test results in an expandable/collapsible list, typically with the specific metrics that the test returned. (You can Expand All or Collapse All from the links at the upper right of the Health Tests panel).

- The color of the text (and the background color of the field) for a health test result indicates the status of the results. The tests are sorted by their health status – Good, Concerning, Bad, or Disabled. The list of entries for good and disabled health tests are collapsed by default; however, Bad or Concerning results are shown expanded.
- The text of a health test also acts as a link to further information about the test. Clicking the text will pop up a window with further information, such as the meaning of the test and its possible results, suggestions for actions you can take or how to make configuration changes related to the test. The help text for a health test also provides a link to the relevant monitoring configuration section for the service.

Health History

The Health History provides a record of state transitions of the health tests for the host.

- Click the arrow symbol at the left to view the description of the health test state change.
- Click the View link to open a new page that shows the state of the host at the time of the transition. In this view some of the status settings are greyed out, as they reflect a time in the past, not the current status.

File Systems

The File systems panel provides information about disks, their mount points and usage. Use this information to determine if additional disk space is required.

Roles

Use the Roles panel to see the role instances running on the selected host, as well as each instance's status and health. Hosts are configured with one or more role instances, each of which corresponds to a service. The role indicates which daemon runs on the host. Some examples of roles include the NameNode, Secondary NameNode, Balancer, JobTrackers, DataNodes, RegionServers and so on. Typically a host will run multiple roles in support of the various services running in the cluster.

Clicking the role name takes you to the role instance's status page.

You can delete a role from the host from the Instances tab of the Service page for the parent service of the role. You can add a role to a host in the same way.

Charts

Charts are shown for each host instance in your cluster.

See the topic *Viewing Charts for Cluster, Service, Role, and Host Instances* for detailed information on the charts that are presented, and the ability to search and display metrics of your choice.

Related Information

Configuring Monitoring Settings
Role Instances
Viewing Charts for Cluster, Service, Role, and Host Instances

Processes

The **Processes** page provides information about each of the processes that are currently running on this host. Use this page to access management web UIs, check process status, and access log information.

Note: The **Processes** page may display exited startup processes. Such processes are cleaned up within a day.

The **Processes** page includes a variety of categories of information.

- Service - The name of the service. Clicking the service name takes you to the service status page. Using the triangle to the right of the service name, you can directly access the tabs on the role page (such as the Instances, Commands, Configuration, Audits, or Charts Library tabs).
- Instance - The role instance on this host that is associated with the service. Clicking the role name takes you to the role instance's status page. Using the triangle to the right of the role name, you can directly access the tabs on the role page (such as the Processes, Commands, Configuration, Audits, or Charts Library tabs) as well as the status page for the parent service of the role.
- Name - The process name.
- Links - Links to management interfaces for this role instance on this system. These is not available in all cases.
- Status - The current status for the process. Statuses include stopped, starting, running, and paused.
- PID - The unique process identifier.
- Uptime - The length of time this process has been running.
- Full log file - A link to the full log (a file external to Cloudera Manager) for this host log entries for this host.
- Stderr - A link to the stderr log (a file external to Cloudera Manager) for this host.
- Stdout - A link to the stdout log (a file external to Cloudera Manager) for this host.

Resources

The **Resources** page provides information about the resources (Memory, disk, and ports) used by every service and role instance running on the selected host.

Each entry on this page lists:

- The service name
- The name of the particular instance of this service
- A brief description of the resource
- The amount of the resource being consumed or the settings for the resource

The resource information provided depends on the type of resource:

- Memory - The number of bytes consumed
- Disk - The disk location where this service stores information
- Ports - The port number being used by the service to establish network connections

Commands

The **Commands** page shows you running or recent commands for the host you are viewing.

Related Information

Viewing Running and Recent Commands

Configuration

The **Configuration** page for a host lets you set properties for the selected host.

You can set properties in the following categories:

- Advanced - Advanced configuration properties. These include the Java Home Directory, which explicitly sets the value of JAVA_HOME for all processes. This overrides the auto-detection logic that is normally used.
- Monitoring - Monitoring properties for this host. The monitoring settings you make on this page will override the global host monitoring settings you make on the Configuration tab of the Hosts page. You can configure monitoring properties for:
  - Health check thresholds
  - The amount of free space on the filesystem containing the Cloudera Manager Agent's log and process directories
  - A variety of conditions related to memory usage and other properties
  - Alerts for health check events

  For some monitoring properties, you can set thresholds as either a percentage or an absolute value (in bytes).
- Other - Other configuration properties
- Parcels - Configuration properties related to parcels. Includes the Parcel Director property, the directory that parcels will be installed into on this host. If the parcel_dir variable is set in the Agent's config.ini file, it will override this value.
- Resource Management - Enables resource management using control groups (cgroups)

Related Information

Modifying Configuration Properties Using Cloudera Manager

Components

The **Components** page lists every component installed on this host. This may include components that have been installed but have not been added as a service (such as YARN, Flume, or Impala).

This includes the following information:

- Component - The name of the component.
- Version - The version of Cloudera Runtime from which each component came.
- Component Version - The detailed version number for each component.

Audits

The **Audits** page lets you filter for audit events related to this host.

Charts Library

The **Charts** Library page for a host instance provides charts for all metrics kept for that host instance, organized by category. Each category is collapsible/expandable.

Related Information

Viewing Charts for Cluster, Service, Role, and Host Instances

Host Inspector

You can use the host inspector to gather information about hosts that Cloudera Manager is currently managing.

You can review this information to better understand system status and troubleshoot any existing issues. For example, you might use this information to investigate potential DNS misconfiguration.

The inspector runs tests to gather information for functional areas including:

- Networking
- System time

- User and group configuration
- HDFS settings
- Component versions

Common cases in which this information is useful include:

- Installing components
- Upgrading components
- Adding hosts to a cluster
- Removing hosts from a cluster

Running the Host Inspector

You can run the host inspector to inspect all hosts and display a list of validations and their results.

Procedure

1. Click HostsAll Hosts.
2. Click the Inspect All Hosts button. Cloudera Manager begins several tasks to inspect the managed hosts.
3. After the inspection completes, click Download Result Data or Show Inspector Results to review the results. The results of the inspection displays a list of all the validations and their results, and a summary of all the components installed on your managed hosts.

   If the validation process finds problems, the **Validations** section will indicate the problem. In some cases the message may indicate actions you can take to resolve the problem. If an issue exists on multiple hosts, you may be able to view the list of occurrences by clicking a small triangle that appears at the end of the message.

   The **Version Summary** section shows all the components that are available from Cloudera, their versions (if known) and the Cloudera Runtime distribution to which they belong.

Viewing Past Host Inspector Results

You can view the results of a past host inspection by looking for the Host Inspector command using the Recent Commands feature.

Procedure

1. Click the Running Commands indicator ( ) to the left of the Search box at the right side of the navigation bar.
2. Click the Recent Commands button.
3. If the command is too far in the past, you can use the Time Range Selector to move the time range back to cover the time period you want.
4. When you find the Host Inspector command, click its name to display its subcommands.
5. Click the **Show Inspector Results** button to view the report.

Related Information

Viewing Running and Recent Commands

Monitoring Activities

Cloudera Manager's activity monitoring capability monitors the Hive, Oozie, and streaming jobs, Impala queries, and YARN applications running or that have run on your cluster.

When the individual jobs are part of larger workflows (using Oozie, or Hive), these jobs are aggregated into MapReduce jobs that can be monitored as a whole, as well as by the component jobs.

If you are running multiple clusters, there will be a separate link in the **Clusters** tab for each cluster's MapReduce activities, Impala queries, and YARN applications.

The following sections describe how to view and monitor activities that run on your cluster.

Selecting Columns to Show in the Activities List

In the Activities list, you can display or hide any of the statistics that Cloudera Manager collects. By default only a subset of the possible statistics are displayed.

Procedure

1. Click the Select Columns to Display icon ( ). A pop-up panel lets you turn on or off a variety of metrics that may be of interest.
2. Check or uncheck the columns you want to include or remove from the display. As you check or uncheck an item, its column immediately appears or disappears from the display.
3. Click the × in the upper right corner to close the panel.

Note: You cannot hide the shortcut menu or chart icon columns. Also, column selections are retained only for the current session.

Sorting the Activities List

You can sort the Activities list by the contents of any column.

Procedure

1. Click the column header to initiate a sort. The small arrow that appears next to the column header indicates the sort direction.
2. Click the column header to reverse the sort direction.

Filtering the Activities List

You can filter the list of activities based on values of any of the metrics that are available. You can also easily filter for certain common queries from the drop-down menu next to the Search button at the top of the Activities list. By default, it is set to show All Activities.

Procedure

1. To use one of the predefined filters, click the to the right of the Search button and select the filter you want to run.

   There are predefined filters to search by job type (for example Pig activities, MapReduce jobs, and so on) or for running, failed, or long-running activities.
2. To create a filter, click the to the right of the Search button and select Custom.
3. Select a metric from the drop-down list in the first field; you can create a filter based on any of the available metrics.
4. Once you select a metric, fill in the rest of the fields; your choices depend on the type of metric you have selected. Use the percent character % as a wildcard in a string; for example, Id matches job%0001 will look for any MapReduce job ID with suffix 0001.
5. To create a compound filter, click the plus icon at the end of the filter row to add another row. If you combine filter criteria, all criteria must be true for an activity to match.

6. To remove a filter criteria from a compound filter, click the minus icon at the end of the filter row. Removing the last row removes the filter.
7. To include any children of a Pig, Hive, or Oozie activity in your search results, check the Include Child Activities checkbox. Otherwise, only the top-level activity will be included, even if one or more child activities matched the filter criteria.
8. Click the Search button (which appears when you start creating the filter) to run the filter.

Note: Filters are remembered across user sessions — that is, if you log out the filter will be preserved and will still be active when you log back in. Newly-submitted activities will appear in the Activity List only if they match the filter criteria.

Activity Charts

By default the charts show aggregated statistics about the performance of the cluster: Tasks Running, CPU Usage, and Memory Usage. There are additional charts you can enable from a pop-up panel. You can also superimpose individual job statistics on any of the displayed charts.

Most charts display multiple metrics within the same chart. For example, the Tasks Running chart shows two metrics: Cluster, Running Maps and Cluster, Running Reduces in the same chart. Each metric appears in a different color.

- To see the exact values at a given point in time, move the cursor over the chart – a movable vertical line pinpoints a specific time, and a tooltip shows you the values at that point.
- You can use the time range selector at the top of the page to zoom in – the chart display will follow. In order to zoom out, you can use the Time Range Selector at the top of the page or click the link below the chart.

To select additional charts:

1. Click at the top right of the chart panel to open the Customize dialog box.
2. Check or uncheck the boxes next to the charts you want to show or hide.

To show or hide cluster-wide statistics:

- Check or uncheck the Cluster checkbox at the top of the Charts panel.

To chart statistics for an individual job:

- Click the chart icon ( ) in the row next to the job you want to show on the charts. The job ID will appear in the top bar next to the Cluster checkbox, and the statistics will appear on the appropriate chart.
- To remove a job's statistics from the chart, click the ✖ next to the job ID in the top bar of the chart.

Note: Chart selections are retained only for the current session.

To expand, contract, or hide the charts:

- Move the cursor over the divider between the Activities list and the charts, grab it and drag to expand or contract the chart area compared to the Activities list.
- Drag the divider all the way to the right to hide the charts, or all the way to the left to hide the Activities list.

Viewing the Jobs in a Pig, Oozie, or Hive Activity

The Activity Children tab shows the same information as does the Activities tab, except that it shows only jobs that are children of a selected Pig, Hive or Oozie activity. In addition, from this tab you can view the details of the Pig, Hive or Oozie activity as a whole, and compare it to similar activities.

Procedure

1. Click the Activities tab.
2. Click the Pig, Hive or Oozie activity you want to inspect. This presents a list of the jobs that make up the Pig, Hive or Oozie activity.

   The functions under the Children tab are the same as those seen under the Activities tab. You can filter the job list, show and hide columns in the job list, show and hide charts and plot job statistics on those charts.

   Click an individual job to view Task information and other information for that child. See the topic *Viewing and Filtering MapReduce Activities* for details of how the functions on this page work.

   In addition, viewing a Pig, Hive or Oozie activity provides the following tabs:

   - The Details tab shows Activity details in a report form.
   - The Compare tab compares this activity to other similar activity. The main difference between this and a comparison for a single MapReduce activity is that the comparison is done looking at other activities of the same type (Pig, Hive or Oozie) but does include the child jobs of the activity. See the topic *Comparing Similar Activities* for an explanation of that tab.

Related Information

Viewing Activity Details in a Report Format
Comparing Similar Activities

Task Attempts

The **Tasks** tab contains a list of the Map and Reduce task attempts that make up a job.

Viewing a Job's Task Attempts

You can view a job's task attempts to inspect statistics about the performance of and resources used by the task attempts spawned by a selected job.

Procedure

1. From the Clusters tab, in the section marked Other, select the activity you want to inspect.

   - If the activity is a MapReduce job, the Tasks tab opens.
   - If the activity is a Pig, Hive, or Oozie activity, select the job you want to inspect from the activity's Children tab to open the Tasks tab.

   The columns shown under the Tasks tab display statistics about the performance of and resources used by the task attempts spawned by the selected job. By default only a subset of the possible metrics are displayed — you can modify the columns that are displayed to add or remove the columns in the display.

   The status of an attempt is shown in the Attempt Status column:

	The attempt is running.
	The attempt has succeeded.
	The attempt has failed.
	The attempt has been unassigned.
	The attempt has been killed.
	The attempt's final state is unknown.

2. Click the task ID to view details of the individual task.

3. Optionally, use the Zoom to Duration button to zoom the Time Range Selector to the exact time range spanned by the activity whose tasks you are viewing.

Selecting Columns to Show in the Tasks List

In the Tasks list, you can display or hide any of the metrics the Cloudera Manager collects for task attempts. By default a subset of the possible metrics are displayed.

Procedure

1. Click the Select Columns to Display icon ( ). A pop-up panel lets you turn on or off a variety of metrics that may be of interest.
2. Check or uncheck the columns you want to include or remove from the display. As you check or uncheck an item, its column immediately appears or disappears from the display.
3. Click the × in the upper right corner to close the panel.

Sorting the Tasks List

You can sort the tasks list by any of the information displayed in the list.

Procedure

1. Click the column header to initiate a sort. The small arrow that appears next to the column header indicates the sort direction.
2. Click the column header to reverse the sort direction.

Filtering the Tasks List

You can filter the list of tasks based on values of any of the metrics that are available.

Procedure

1. To use one of the predefined filters, click the to the right of the Search button and select the filter you want to run. There are predefined filters to search by job type (for example Pig activities, MapReduce jobs, and so on) or for running, failed, or long-running activities.
2. To create a filter, click the to the right of the Search button and select Custom.
3. Select a metric from the drop-down list in the first field; you can create a filter based on any of the available metrics.
4. Once you select a metric, fill in the rest of the fields; your choices depend on the type of metric you have selected. Use the percent character % as a wildcard in a string; for example, Id matches job%0001 will look for any MapReduce job ID with suffix 0001.
5. To create a compound filter, click the plus icon at the end of the filter row to add another row. If you combine filter criteria, all criteria must be true for an activity to match.
6. To remove a filter criteria from a compound filter, click the minus icon at the end of the filter row. Removing the last row removes the filter.
7. To include any children of a Pig, Hive, or Oozie activity in your search results, check the Include Child Activities checkbox. Otherwise, only the top-level activity will be included, even if one or more child activities matched the filter criteria.
8. Click the Search button (which appears when you start creating the filter) to run the filter.

Note: The filter persists only for this user session — when you log out, tasks list filter is removed.

Viewing Activity Details in a Report Format

The **Details** tab for an activity shows the job or activity statistics in a report format.

To view activity details for an individual MapReduce job:

1. Select a MapReduce job from the Clusters tab or Select a Pig, Hive or Oozie activity, then select a MapReduce job from the Children tab.
2. Select the Details tab after the job page is displayed.

This displays information about the individual MapReduce job in a report format.

From this page you can also access the Job Details and Job Configuration pages on the JobTracker web UI.

- Click the Job Details link at the top of the report to be taken to the job details web page on the JobTracker host.
- Click the Job Configuration link to be taken to the job configuration web page on the JobTracker host.

To view activity details for a Pig, Hive, or Oozie activity:

1. Select a Pig, Hive or Oozie activity.
2. Select the Details tab after the list of child jobs is displayed.

This displays information about the Pig, Oozie, or Hive job as a whole.

Note that this the same data you would see for the activity if you displayed all possible columns in the Activities list.

Comparing Similar Activities

It can be useful to compare the performance of similar activities if, for example, you suspect that a job is performing differently than other similar jobs that have run in the past. The Compare tab shows you the performance of the selected job compared with the performance of other similar jobs.

Cloudera Manager identifies jobs that are similar to each other (jobs that are basically running the same code – the same Map and Reduce classes, for example).

To compare an activity to other similar activities:

1. Select the job or activity from the Activities list.
2. Click the Compare tab.

The activity comparison feature compares performance and resource statistics of the selected job to the mean value of those statistics across a set of the most recent similar jobs. The table provides visual indicators of how the selected job deviates from the mean calculated for the sample set of jobs, as well as providing the actual statistics for the selected job and the set of the similar jobs used to calculate the mean.

- The first row in the comparison table displays a set of visual indicators of how the selected job deviates from the mean of all the similar jobs (the combined Average values). This is displayed for each statistic for which a comparison makes sense. The diagram in the ID column shows the elements of the indicator, as follows:
  - The line at the midpoint of the bar represents the mean value of all similar jobs. The colored portion of the bar indicates the degree of deviation of your selected job from the mean. The top and bottom of the bar represent two standard deviations (plus or minus) from the mean.
  - For a given metric, if the value for your selected job is within two standard deviations of the mean, the colored portion of the bar is blue.
  - If a metric for your selected job is more than two standard deviations from the mean, the colored portion of the bar is red.
- The following rows show the actual values for other similar jobs. These are the sets of values that were used to calculate the mean values shown in the Combined Averages row. The most recent ten similar jobs are used to calculate the average job statistics, and these are the jobs that are shown in the table.

Viewing the Distribution of Task Attempts

The **Task Distribution** tab provides a graphical view of the performance of the Map and Reduce tasks that make up a job.

Procedure

1. To display the task distribution metrics for a job, do one of the following:
   - Select a MapReduce job from the Activities list.
   - Select a job from the Children tab of a Pig, Hive, or Oozie activity.
2. Click the Task Distribution tab.

   The chart that appears initially shows the distribution of Map Input Records by Duration; you can change the Y-axis to chart a number of different metrics.
3. Optionally, you can use the Zoom to Duration button to zoom the Time Range Selector to the exact time range spanned by the activity whose tasks you are viewing.

The Task Distribution Chart

The Task Distribution chart shows the distribution of attempts according to their duration on the X-axis and a number of different metrics on the Y-axis.

Each cell of the chart represents the number of tasks whose performance statistics fall within the parameters of the cell. The Task Distribution chart is useful for detecting tasks that are outliers in your job, either because of skew, or because of faulty TaskTrackers. The chart can clearly show if some tasks deviate significantly from the majority of task attempts.

Normally, the distribution of tasks will be fairly concentrated. If, for example, some Reducers receive much more data than others, that will be represented by having two discrete sections of density on the graph. That suggests that there may be a problem with the user code, or that there's skew in the underlying data. Alternately, if the input sizes of various Map or Reduce tasks are the same, but the time it takes to process them varies widely, it might mean that certain TaskTrackers are performing more poorly than others.

You can click in a cell and see a list of the TaskTrackers that correspond to the tasks whose performance falls within the cell.

The X-axis show the task duration is seconds. From the drop-down you can chose different metrics for the Y-axis: Input or Output records or bytes for Map tasks, or the number of CPU seconds for the user who ran the job:

- Map Input Records vs. Duration
- Map Output Records vs. Duration
- Map Input Bytes vs. Duration
- Map Output Bytes vs. Duration
- Map Total User CPU seconds vs. Duration
- Reduce Input Records vs. Duration
- Reduce Output Records vs. Duration
- Reduce Total User CPU seconds vs. Duration

TaskTracker Hosts

To the right of the chart is a table that shows the TaskTracker hosts that processed the tasks in the selected cell, along with the number of task attempts each host executed.

You can select a cell in the table to view the TaskTracker hosts that correspond to the tasks in the cell.

- The area above the TaskTracker table shows the type of task and range of data volume (or User CPUs) and duration times for the task attempts that fall within the cell.
- The table itself shows the TaskTracker hosts that executed the tasks that are represented within the cell, and the number of task attempts run on that host.

Clicking a TaskTracker hostname takes you to the **Role Status** page for that TaskTracker instance.

Monitoring Impala Queries

The Impala Queries page displays information about Impala queries that are running and have run in your cluster. You can filter the queries by time period and by specifying simple filtering expressions.

Note: The Impala query monitoring feature requires Impala 1.0.1 and higher.

Related Information

Filtering Queries

Viewing Queries

You can view queries, filter queries, and more from the Impala service **Queries** tab.

Do one of the following:

- Select Clusters *Cluster name* *Impala service name* Queries .
- On the Home Status tab, select *Impala service name* and click the Queries tab.

The Impala queries run during the selected time range display in the Results tab.

You can also perform the following actions on this page:

Table 5: Viewing Queries Actions

Action	Description
Filter the displayed queries	Create filter expressions manually, select preconfigured filters, or use the Workload Summary section to build a query interactively.
Select additional attributes for display.	Click Select Attributes. Selected attributes also display as available filters in the Workload Summary section. To display information about attributes, hover over a field label. . Only attributes that support filtering appear in the Workload Summary section.
View a histogram of the attribute values.	Click the icon to the right of each attribute displayed in the Workload Summary section.
Display charts based on the filter expression and selected attributes.	Cick the Charts tab.
View charts that help identify whether Impala best practices are being followed.	Click the Best Practices link.
Export a JSON file with the query results that you can use for further analysis.	Click Export.

Related Information

Filtering Queries

Filter Attributes

Configuring Impala Query Monitoring

You can configure the visibility of the Impala query results and the size of the storage allocated to Impala query results.

About this task

To configure whether admin or non-admin users can view all queries, only that user's queries, or no queries:

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

1. Go to the Impala service.
2. Click the Configuration tab.
3. Select Scope*Impala service_name* (Service-Wide).
4. Click the Monitoring category.
5. Set the Visibility Settings properties for admin and non-admin users.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Impala Best Practices

The **Impala Best Practices** page contains charts that include description of each best practice and how to determine if it is being followed.

To open the Impala Best Practices page, click the Best Practices tab on the Impala service page. See the Impala documentation for more detail on each best practice and for additional best practices.

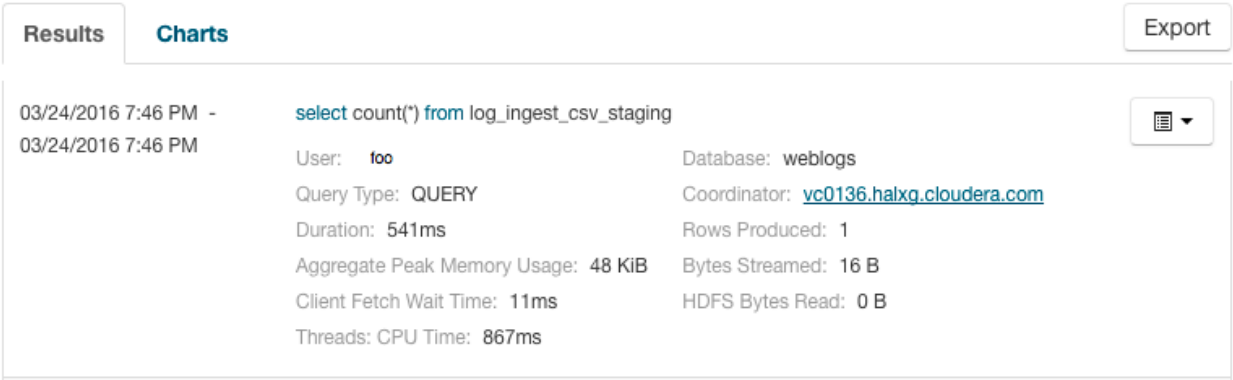
Adjust the time range to see data on queries run at different times. Click the charts to get more detail on individual queries. Use the filter box at the top right of the Best Practices page to adjust which data is shown on the page. For example, to see just the queries that took more than ten seconds, make the filter query_duration > 10s.

Create a trigger based on any best practice by choosing Create Trigger from the individual chart drop-down menu.

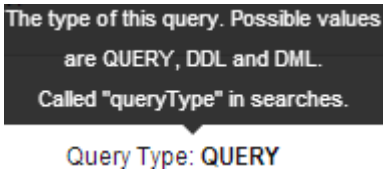
Results Tab

Queries appear on the Results tab, with the most recent at the top. Each query has summary and detail information.

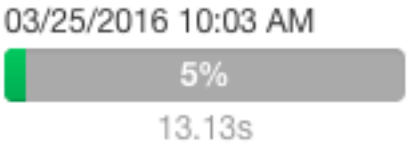
A query summary includes the following default attributes: start and end timestamps, statement, duration, rows produced, user, coordinator, database, and query type. For example:

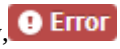


You can add additional attributes to the summary by clicking the Attribute Selector. In each query summary, the query statement is truncated if it is too long to display. To display the entire statement, click ❯. The query entry expands to display the entire query string. To collapse the query display, click ❮. To display information about query attributes and possible values, hover over a field in a query. For example:



A running job displays a progress bar under the starting timestamp:



If an error occurred while processing the query,  displays under the complete timestamp.

Use the Actions drop-down menu to the right of each query listing to do the following. (Not all options display, depending on the type of job.)

- Query Details – Opens a details page for the job.
- User's Impala Queries – Displays a list of queries run by the user for the current job.
- Cancel (running queries only) – Cancel a running query (administrators only). Canceling a running query creates an audit event. When you cancel a query, replaces the progress bar.
- Queries in the same YARN pool – Displays queries that use the same resource pool.

Related Information

Filter Attributes
Attribute Selector
Query Details

Filtering Queries

You filter queries by selecting a time range and specifying a filter expression in the search box.

You filter queries by selecting a time range and specifying a filter expression in the search box.

You can use the Time Range Selector or a duration link (

) to set the time range.

Related Information

Time Line

Filter Expressions

Filter expressions specify which entries should display when you run the filter.

The simplest expression consists of three components:

- Attribute - Query language name of the attribute.
- Operator - Type of comparison between the attribute and the attribute value. Cloudera Manager supports the standard comparator operators =, !=, >, <, >=, <=, and RLIKE. (RLIKE performs regular expression matching as specified in the Java Pattern class documentation.) Numeric values can be compared with all operators. String values can be compared with =, !=, and RLIKE. Boolean values can be compared with = and !=.
- Value - The value of the attribute. The value depends on the type of the attribute. For a Boolean value, specify either true or false. When specifying a string value, enclose the value in double quotes.

You create compound filter expressions using the AND and OR operators. When more than one operator is used in an expression, AND is evaluated first, then OR. To change the order of evaluation, enclose subexpressions in parentheses.

Compound Expressions

To find all the queries issued by the root user that produced over 100 rows, use the expression:

```
user = "root" AND rowsProduced > 100
```

To find all the executing queries issued by users Jack or Jill, use the expression:

```
executing = true AND (user = "Jack" OR user = "Jill")
```

Related Information

Java Pattern

Filter Attributes

The following table includes available filter attributes and their names in Cloudera Manager, types, and descriptions.

Note: Only attributes for which the Supports Filtering? column value is TRUE appear in the Workload Summary section.

Table 6: Attributes

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Admission Result (admission_result)	STRING	TRUE	The result of admission, whether immediately, queued, rejected, or timed out. Called 'admission_result' in searches.
Admission Wait Time (admission_wait)	MILLISECONDS	TRUE	The time from submission for admission to completion of admission. Called 'admission_wait' in searches.
Aggregate Peak Memory Usage (memory_aggregate_peak)	BYTES	TRUE	The highest amount of memory allocated by this query at a particular time across all nodes. Called 'memory_aggregate_peak' in searches.
Bytes Streamed (bytes_streamed)	BYTES	TRUE	The total number of bytes sent between Impala Daemons while processing this query. Called 'bytes_streamed' in searches.
Client Fetch Wait Time (client_fetch_wait_time)	MILLISECONDS	TRUE	The total amount of time the query spent waiting for the client to fetch row data. Called 'client_fetch_wait_time' in searches.
Client Fetch Wait Time Percentage (client_fetch_wait_time_percentage)	NUMBER	TRUE	The total amount of time the query spent waiting for the client to fetch row data divided by the query duration. Called 'client_fetch_wait_time_percentage' in searches.
Connected User (connected_user)	STRING	TRUE	The user who created the Impala session that issued this query. This is distinct from 'user' only if delegation is in use. Called 'connected_user' in searches.
Coordinator (coordinator_host_id)	STRING	TRUE	The host coordinating this query. Called 'coordinator_host_id' in searches.
Database (database)	STRING	TRUE	The database on which the query was run. Called 'database' in searches.
DDL Type (ddl_type)	STRING	TRUE	The type of DDL query. Called 'ddl_type' in searches.
Delegated User (delegated_user)	STRING	TRUE	The effective user for the query. This is set only if delegation is in use. Called 'delegated_user' in searches.
Duration (query_duration)	MILLISECONDS	TRUE	The duration of the query in milliseconds. Called 'query_duration' in searches.
Estimated per Node Peak Memory (estimated_per_node_peak_memory)	BYTES	TRUE	The planning process's estimate of per-node peak memory usage for the query. Called 'estimated_per_node_peak_memory' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Executing (executing)	BOOLEAN	FALSE	Whether the query is currently executing. Called 'executing' in searches.
File Formats (file_formats)	STRING	FALSE	An alphabetically sorted list of all the file formats used in the query. Called 'file_formats' in searches.
HBase Bytes Read (hbase_bytes_read)	BYTES	TRUE	The total number of bytes read from HBase by this query. Called 'hbase_bytes_read' in searches.
HBase Scanner Average Read Throughput (hbase_scanner_average_bytes_read_per_second)	BYTES_PER_SECOND	TRUE	The average HBase scanner read throughput for this query. This is computed by dividing the total bytes read from HBase by the total time spent reading by all HBase scanners. Called 'hbase_scanner_average_bytes_read_per_second' in searches.
HDFS Average Scan Range (hdfs_average_scan_range)	BYTES	TRUE	The average HDFS scan range size for this query. HDFS scan nodes that contained only a single scan range are not included in this computation. Low numbers for a query might indicate reading many small files which negatively impacts performance. Called 'hdfs_average_scan_range' in searches.
HDFS Bytes Read (hdfs_bytes_read)	BYTES	TRUE	The total number of bytes read from HDFS by this query. Called 'hdfs_bytes_read' in searches.
HDFS Bytes Read From Cache (hdfs_bytes_read_from_cache)	BYTES	TRUE	The total number of bytes read from HDFS that were read from the HDFS cache. This is only for completed queries. Called 'hdfs_bytes_read_from_cache' in searches.
HDFS Bytes Read From Cache Percentage (hdfs_bytes_read_from_cache_percentage)	NUMBER	TRUE	The percentage of all bytes read by this query that were read from the HDFS cache. This is only for completed queries. Called 'hdfs_bytes_read_from_cache_percentage' in searches.
HDFS Bytes Skipped (hdfs_bytes_skipped)	BYTES	TRUE	The total number of bytes that had to be skipped by this query while reading from HDFS. Any number above zero may indicate a problem. Called 'hdfs_bytes_skipped' in searches.
HDFS Bytes Written (hdfs_bytes_written)	BYTES	TRUE	The total number of bytes written to HDFS by this query. Called 'hdfs_bytes_written' in searches.
HDFS Local Bytes Read (hdfs_bytes_read_local)	BYTES	TRUE	The total number of local bytes read from HDFS by this query. This is only for completed queries. Called 'hdfs_bytes_read_local' in searches.
HDFS Local Bytes Read Percentage (hdfs_bytes_read_local_percentage)	NUMBER	TRUE	The percentage of all bytes read from HDFS by this query that were local. This is only for completed queries. Called 'hdfs_bytes_read_local_percentage' in searches.
HDFS Remote Bytes Read (hdfs_bytes_read_remote)	BYTES	TRUE	The total number of remote bytes read from HDFS by this query. This is only for completed queries. Called 'hdfs_bytes_read_remote' in searches.
HDFS Remote Bytes Read Percentage (hdfs_bytes_read_remote_percentage)	NUMBER	TRUE	The percentage of all bytes read from HDFS by this query that were remote. This is only for completed queries. Called 'hdfs_bytes_read_remote_percentage' in searches.
HDFS Scanner Average Read Throughput (hdfs_scanner_average_bytes_read_per_second)	BYTES_PER_SECOND	TRUE	The average HDFS scanner read throughput for this query. This is computed by dividing the total bytes read from HDFS by the total time spent reading by all HDFS scanners. Called 'hdfs_scanner_average_bytes_read_per_second' in searches.
HDFS Short Circuit Bytes Read (hdfs_bytes_read_short_circuit)	BYTES	TRUE	The total number of bytes read from HDFS by this query that used short-circuit reads. This is only for completed queries. Called 'hdfs_bytes_read_short_circuit' in searches.
HDFS Short Circuit Bytes Read Percentage (hdfs_bytes_read_short_circuit_percentage)	NUMBER	TRUE	The percentage of all bytes read from HDFS by this query that used short-circuit reads. This is only for completed queries. Called 'hdfs_bytes_read_short_circuit_percentage' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Impala Version (impala_version)	STRING	TRUE	The version of the Impala Daemon coordinating this query. Called 'impala_version' in searches.
Memory Accrual (memory_accrual)	BYTE_SECONDS	TRUE	The total accrued memory usage by the query. This is computed by multiplying the average aggregate memory usage of the query by the query's duration. Called 'memory_accrual' in searches.
Memory Spilled (memory_spilled)	BYTES	TRUE	Amount of memory spilled to disk. Called 'memory_spilled' in searches.
Network Address (network_address)	STRING	TRUE	The network address that issued this query. Called 'network_address' in searches.
Node with Peak Memory Usage (memory_per_node_peak_node)	STRING	TRUE	The node with the highest peak memory usage for this query. See Per Node Peak Memory Usage for the actual peak value. Called 'memory_per_node_peak_node' in searches.
Out of Memory (oom)	BOOLEAN	TRUE	Whether the query ran out of memory. Called 'oom' in searches.
Per Node Peak Memory Usage (memory_per_node_peak)	BYTES	TRUE	The highest amount of memory allocated by any single node that participated in this query. See Node with Peak Memory Usage for the name of the peak node. Called 'memory_per_node_peak' in searches.
Planning Wait Time (planning_wait_time)	MILLISECONDS	TRUE	The total amount of time the query spent waiting for planning to complete. Called 'planning_wait_time' in searches.
Planning Wait Time Percentage (planning_wait_time_percentage)	NUMBER	TRUE	The total amount of time the query spent waiting for planning to complete divided by the query duration. Called 'planning_wait_time_percentage' in searches.
Pool (pool)	STRING	TRUE	The name of the resource pool in which this query executed. Called 'pool' in searches. If YARN is in use, this corresponds to a YARN pool. Within YARN, a pool is referred to as a queue.
Query ID (query_id)	STRING	FALSE	The id of this query. Called 'query_id' in searches.
Query State (query_state)	STRING	TRUE	The current state of the query (running, finished, and so on). Called 'query_state' in searches.
Query Status (query_status)	STRING	TRUE	The status of the query. If the query hasn't failed the status will be 'OK', otherwise it will provide more information on the cause of the failure. Called 'query_status' in searches.
Query Type (query_type)	STRING	TRUE	The type of the query's SQL statement (DML, DDL, Query). Called 'query_type' in searches.
Resource Reservation Wait Time (resources_reserved_wait_time)	MILLISECONDS	TRUE	The total amount of time the query spent waiting for pool resources to become available . Called 'resources_reserved_wait_time' in searches.
Resource Reservation Wait Time Percentage (resources_reserved_wait_time_percentage)	NUMBER	TRUE	The total amount of time the query spent waiting for pool resources to become available divided by the query duration. Called 'resources_reserved_wait_time_percentage' in searches.
Rows Inserted (rows_inserted)	NUMBER	TRUE	The number of rows inserted by the query. Called 'rows_inserted' in searches.
Rows Produced (rows_produced)	NUMBER	TRUE	The number of rows produced by the query. Called 'rows_produced' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Service Name (service_name)	STRING	FALSE	The name of the Impala service. Called 'service_name' in searches.
Session ID (session_id)	STRING	TRUE	The ID of the session that issued this query. Called 'session_id' in searches.
Session Type (session_type)	STRING	TRUE	The type of the session that issued this query. Called 'session_type' in searches.
Statement (statement)	STRING	FALSE	The query's SQL statement. Called 'statement' in searches.
Statistics Missing (stats_missing)	BOOLEAN	TRUE	Whether the query was flagged with missing table or column statistics warning during the planning process. Called 'stats_missing' in searches.
Threads: CPU Time (thread_cpu_time)	MILLISECONDS	TRUE	The sum of the CPU time used by all threads of the query. Called 'thread_cpu_time' in searches.
Threads: CPU Time Percentage (thread_cpu_time_percentage)	NUMBER	TRUE	The sum of the CPU time used by all threads of the query divided by the total thread time. Called 'thread_cpu_time_percentage' in searches.
Threads: Network Receive Wait Time (thread_network_receive_wait_time)	MILLISECONDS	TRUE	The sum of the time spent waiting to receive data over the network by all threads of the query. A query will almost always have some threads waiting to receive data from other nodes in the query's execution tree. Unlike other wait times, network receive wait time does not usually indicate an opportunity for improving a query's performance. Called 'thread_network_receive_wait_time' in searches.
Threads: Network Receive Wait Time Percentage (thread_network_receive_wait_time_percentage)	NUMBER	TRUE	The sum of the time spent waiting to receive data over the network by all threads of the query divided by the total thread time. A query will almost always have some threads waiting to receive data from other nodes in the query's execution tree. Unlike other wait times, network receive wait time does not usually indicate an opportunity for improving a query's performance. Called 'thread_network_receive_wait_time_percentage' in searches.
Threads: Network Send Wait Time (thread_network_send_wait_time)	MILLISECONDS	TRUE	The sum of the time spent waiting to send data over the network by all threads of the query. Called 'thread_network_send_wait_time' in searches.
Threads: Network Send Wait Time Percentage (thread_network_send_wait_time_percentage)	NUMBER	TRUE	The sum of the time spent waiting to send data over the network by all threads of the query divided by the total thread time. Called 'thread_network_send_wait_time_percentage' in searches.
Threads: Storage Wait Time (thread_storage_wait_time)	MILLISECONDS	TRUE	The sum of the time spent waiting for storage by all threads of the query. Called 'thread_storage_wait_time' in searches.
Threads: Storage Wait Time Percentage (thread_storage_wait_time_percentage)	NUMBER	TRUE	The sum of the time spent waiting for storage by all threads of the query divided by the total thread time. Called 'thread_storage_wait_time_percentage' in searches.
Threads: Total Time (thread_total_time)	MILLISECONDS	TRUE	The sum of thread CPU, storage wait and network wait times used by all threads of the query. Called 'thread_total_time' in searches.
User (user)	STRING	TRUE	The effective user for the query. This is the delegated user if delegation is in use. Otherwise, this is the connected user. Called 'user' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Work CPU Time (cm_cpu_milliseconds)	MILLISECONDS	TRUE	Attribute measuring the sum of CPU time used by all threads of the query, in milliseconds. Called 'work_cpu_time' in searches. For Impala queries, CPU time is calculated based on the 'TotalCpuTime' metric. For YARN MapReduce applications, this is calculated from the 'cpu_milliseconds' metric.

Examples

Consider the following filter expressions: user = "root", rowsProduced > 0, fileFormats RLIKE ".TEXT.*", and executing = true. In the examples:

- The filter attributes are user, rowsProduced, fileFormats, and executing.
- The operators are =, >, and RLIKE.
- The filter values are root, 0, .TEXT.*, and true.

Choosing and Running a Filter

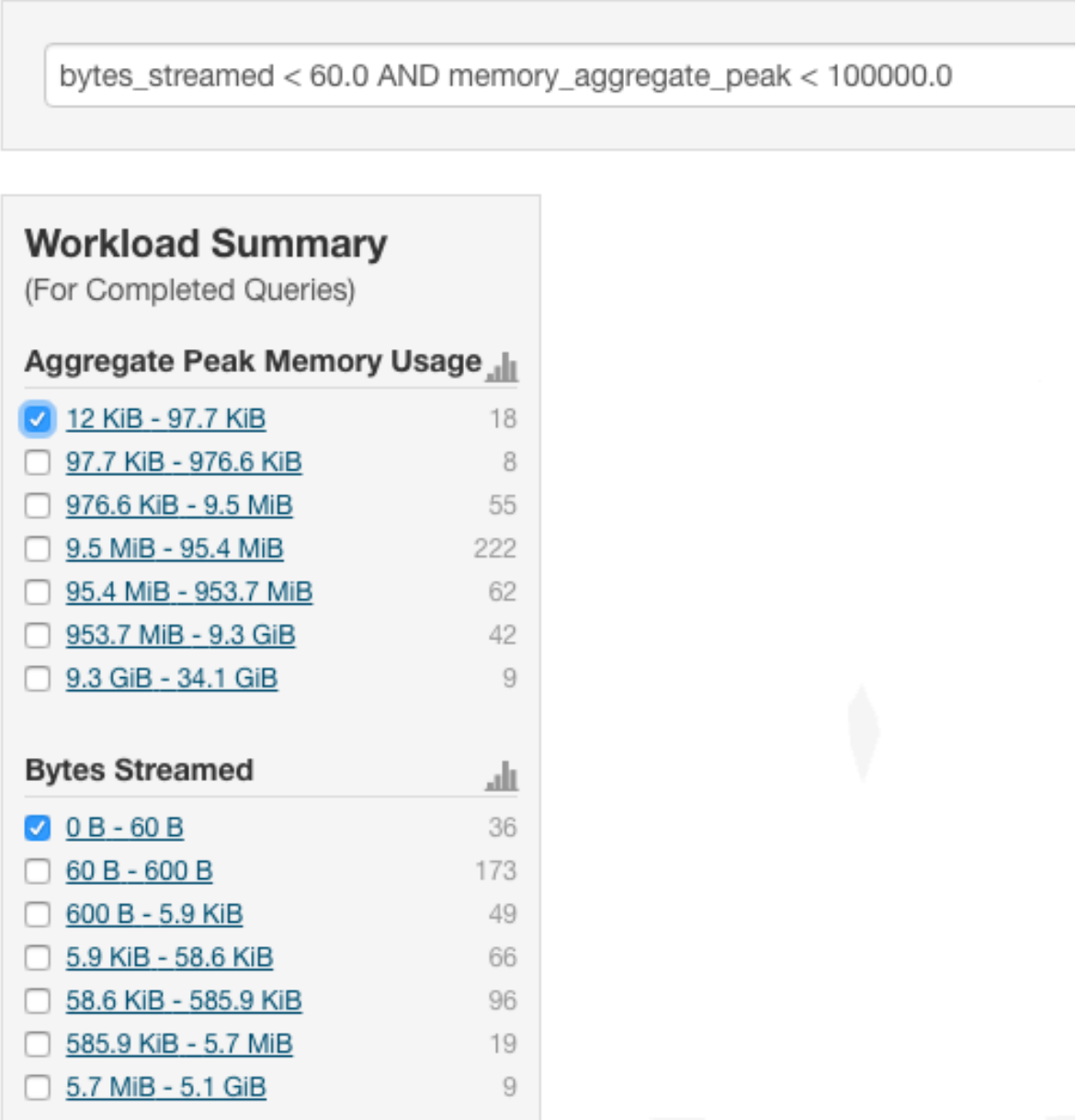
You can construct a filter, type a filter, or select a suggested or recently run filter.

Procedure

1. Do one of the following:

   - Select a Suggested or Recently Run Filter: Click the to the right of the Search button to display a list of sample and recently run filters, and select a filter. The filter text displays in the text box.
   - Construct a Filter from the Workload Summary Attributes: Optionally, click Select Attributes to display a dialog box where you can chose which attributes to display in the Workload Summary section. Select the checkbox next to one or more attributes, and click Close.

     The attributes display in the Workload Summary section along with values or ranges of values that you can filter on. The values and ranges display as links with checkboxes. Select one or more checkboxes to add the range or value to the query. Click a link to run a query on that value or range. For example:

   - Type a Filter: Start typing or press Spacebar in the text box.

     As you type, filter attributes matching the typed letter display. If you press Spacebar, standard filter attributes display. These suggestions are part of typeahead, which helps build valid queries. For information about the attribute name and supported values for each field, hover over the field in an existing query.

   a. Select an attribute and press Enter.
   b. Press Spacebar to display a drop-down list of operators.
   c. Select an operator and press Enter.
   d. Specify an attribute value. For attribute values that support typeahead, press the spacebar to display a drop-down list of values and press Enter. Alternatively, you can type a value.
2. Click in the text box and press Enter or click Search. The list displays the results that match the specified filter. The Workload Summary section refreshes to show only the values for the selected filter. The filter is added to the Recently Run list.

Query Details

The Query Details page contains the low-level details of how a SQL query is processed through Impala.

The initial information on the page can help you tune the performance of some kinds of queries, primarily those involving joins. The more detailed information on the page is primarily for troubleshooting with the assistance of Cloudera Support; you might be asked to attach the contents of the page to a trouble ticket. The Query Details page displays the following information that is also available in the Query Profile.

To download the contents of the query profile details, select one of the following:

- Download Profile... or Download Profile...Download Text Profile... - to download a text version of the query detail.
- Download Profile...Download Thrift Encoded Profile... - to download a binary version of the query detail.

Query Plan

The Query Plan section can help you diagnose and tune performance issues with queries. This information is especially useful to understand performance issues with join queries, such as inefficient order of tables in the SQL statement, lack of table and column statistics, and the need for query hints to specify a more efficient join mechanism. You can also learn valuable information about how queries are processed for partitioned tables.

The information in this section corresponds to the output of the EXPLAIN statement for the Impala query. Each fragment shown in the query plan corresponds to a processing step that is performed by the central coordinator host or distributed across the hosts in the cluster.

Query Timeline

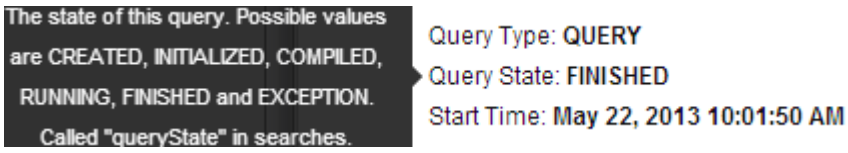
The Query Timeline section reports statistics about the execution time for phases of the query.

Planner Timeline

The Planner Timeline reports statistics about the execution time for phases of the query planner.

Query Info

The Query Info section reports the attributes of the query, start and end time, duration, and statistics about HDFS access. You can hover over an attribute for information about the attribute name and supported values (for enumerated values). For example:



Query Fragments

The Query Fragments section reports detailed low-level statistics for each query plan fragment, involving physical aspects such as CPU utilization, disk I/O, and network traffic. This is the primary information that Cloudera Support might use to help troubleshoot performance issues and diagnose bugs. The details for each fragment display on separate tabs.

Related Information

Understanding Performance using Query Profile

Monitoring YARN Applications

The YARN Applications page displays information about the YARN jobs that are running and have run in your cluster. You can filter the jobs by time period and by specifying simple filtering expressions.

Viewing Jobs

You can view YARN jobs, filter YARN jobs, and more from the YARN service **Applications** tab.

Do one of the following:

- Select Clusters *Cluster name* *YARN service name* Applications .
- On the Home Status tab, select *YARN service name* and click the Applications tab.

The YARN jobs run during the selected time range display in the Results tab. The results displayed can be filtered by creating filter expressions.

You can also perform the following actions on this page:

Table 7: Viewing Jobs Actions

Action	Description
Filter jobs that display.	Create filter expressions manually, select preconfigured filters, or use the Workload Summary section to build a query interactively.
Select additional attributes for display.	Click Select Attributes. Selected attributes also display as available filters in the Workload Summary section. To display information about attributes, hover over a field label. Only attributes that support filtering appear in the Workload Summary section.
View a histogram of the attribute values.	Click the icon to the right of each attribute displayed in the Workload Summary section.
Display charts based on the filter expression and selected attributes.	Click the Charts tab.
Send a YARN application diagnostic bundle to Cloudera support.	Click Collect Diagnostics Data.
Export a JSON file with the query results that you can use for further analysis.	Click Export.

Related Information

Results Tab

Filtering Jobs

Filter Attributes

Sending Diagnostic Data to Cloudera for YARN Applications

Configuring YARN Application Monitoring

You can configure the visibility of the YARN application monitoring results.

About this task

To configure whether admin and non-admin users can view all applications, only that user's applications, or no applications:

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

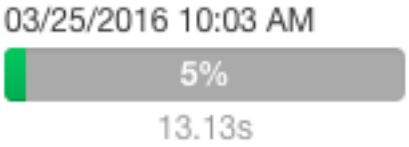
1. Go to the YARN service.
2. Click the Configuration tab.
3. Select Scope *YARN service_name* (Service-Wide) .
4. Click the Monitoring category.
5. Set the Applications List Visibility Settings properties for admin and non-admin users.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Results Tab

Jobs appear on the Results tab, with the most recent at the top. Each job has summary and detail information.

Jobs are ordered with the most recent at the top. Each job has summary and detail information. A job summary includes start and end timestamps, query (if the job is part of a Hive query) name, pool, job type, job ID, and user. For example:

A running job displays a progress bar under the start timestamp:



Use the Actions drop-down menu to the right of each job listing to do the following. (Not all options display, depending on the type of job.)

- Application Details – Open a details page for the job.
- Collect Diagnostic Data – Send a YARN application diagnostic bundle to Cloudera support.
- Similar MR2 Jobs – Display a list of similar MapReduce 2 jobs.
- User's YARN Applications – Display a list of all jobs run by the user of the current job.
- View on JobHistory Server – View the application in the YARN JobHistory Server.
- Kill (running jobs only) – Stop a job (administrators only). Stopping a job creates an audit event. When you stop a job,

replaces the progress bar.

- Applications in Hive Query (Hive jobs only)
- Applications in Oozie Workflow (Oozie jobs only)
- Applications in Pig Script (Pig jobs only)

Filtering Jobs

You filter jobs by selecting a time range and specifying a filter expression in the search box.

You can use the Time Range Selector or a duration link (

) to set the time range.

Related Information

Time Line

Filter Expressions

Filter expressions specify which entries should display when you run the filter.

Filter Expressions

The simplest expression consists of three components:

- Attribute - Query language name of the attribute.
- Operator - Type of comparison between the attribute and the attribute value. Cloudera Manager supports the standard comparator operators =, !=, >, <, >=, <=, and RLIKE. (RLIKE performs regular expression matching as specified in the Java Pattern class documentation.) Numeric values can be compared with all operators. String values can be compared with =, !=, and RLIKE. Boolean values can be compared with = and !=.
- Value - The value of the attribute. The value depends on the type of the attribute. For a Boolean value, specify either true or false. When specifying a string value, enclose the value in double quotes.

You create compound filter expressions using the AND and OR operators. When more than one operator is used in an expression, AND is evaluated first, then OR. To change the order of evaluation, enclose subexpressions in parentheses.

Compound Expressions

To find all the jobs issued by the root user that ran for longer than ten seconds, use the expression:

```
user = "root" AND application_duration >= 100000.0
```

To find all the jobs that had more than 200 maps issued by users Jack or Jill, use the expression:

```
maps_completed >= 200.0 AND (user = "Jack" OR user = "Jill")
```

Related Information

Java Pattern

Choosing and Running a Filter

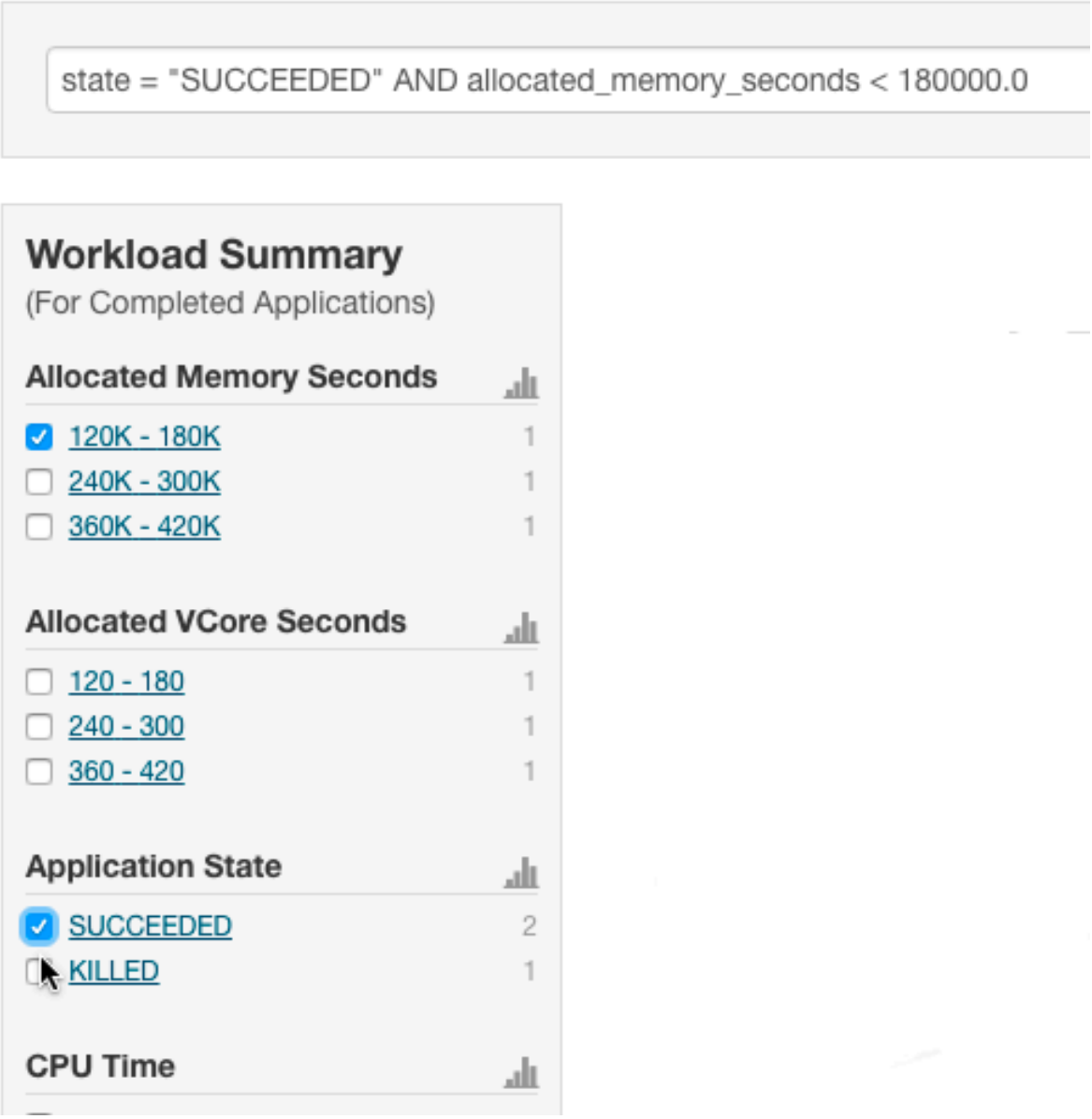
You can construct a filter, type a filter, or select a suggested or recently run filter.

Procedure

1. Do one of the following:
   - Select a Suggested or Recently Run Filter: Click the to the right of the Search button to display a list of sample and recently run filters, and select a filter. The filter text displays in the text box.
   - Construct a Filter from the Workload Summary Attributes: Optionally, click Select Attributes to display a dialog box where you can chose attributes to display in the Workload Summary section. Select the checkbox

next to one or more attributes and click Close. Only attributes that support filtering appear in the Workload Summary section. See the Attributes table.

The attributes display in the Workload Summary section along with values or ranges of values that you can filter on. The values and ranges display as links with checkboxes. Select one or more checkboxes to add the range or value to the query. Click a link to run a query on that value or range. For example:



- Type a Filter: Start typing or press Spacebarin the text box.

  As you type, filter attributes matching the typed letter display. If you press Spacebar, standard filter attributes display. These suggestions are part of typeahead, which helps build valid queries. For information about the attribute name and supported values for each field, hover over the field in an existing query.

  a. Select an attribute and press Enter.
  b. Press Spacebar to display a drop-down list of operators.
  c. Select an operator and press Enter.
  d. Specify an attribute value. For attribute values that support typeahead, press the spacebar to display a drop-down list of values and press Enter. Alternatively, you can type a value.

2. Click in the text box and press Enter or click Search. The list displays the results that match the specified filter. If the histograms are showing, they are redrawn to show only the values for the selected filter. The filter is added to the Recently Run list.

Related Information

Filter Attributes

Filter Attributes

The table below describes filter attributes, their names as they are displayed in Cloudera Manager, their types, and descriptions.

Note: Only attributes where the Supports Filtering? column value is TRUE appear in the Workload Summary section.

Table 8: Attributes

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Allocated Memory (allocated_mb)	NUMBER	FALSE	The sum of memory in MB allocated to the application's running containers. Called 'allocated_mb' in searches.
Allocated Memory Seconds (allocated_memory_seconds)	NUMBER	TRUE	The amount of memory the application has allocated (megabyte-seconds). Called 'allocated_memory_seconds' in searches.
Allocated VCores (allocated_vcores)	NUMBER	FALSE	The sum of virtual cores allocated to the application's running containers. Called 'allocated_vcores' in searches.
Allocated VCore Seconds (allocated_vcore_seconds)	NUMBER	TRUE	The amount of CPU resources the application has allocated (virtual core-seconds). Called 'allocated_vcore_seconds' in searches.
Application ID (application_id)	STRING	FALSE	The ID of the YARN application. Called 'application_id' in searches.
Application State (state)	STRING	TRUE	The state of this YARN application. This reflects the ResourceManager state while the application is running and the JobHistory Server state after the application has completed. Called 'state' in searches.
Application Tags (application_tags)	STRING	FALSE	A list of tags for the application. Called 'application_tags' in searches.
Application Type (application_type)	STRING	TRUE	The type of the YARN application. Called 'application_type' in searches.
Bytes Read (bytes_read)	BYTES	TRUE	Bytes read. Called 'bytes_read' in searches.
Bytes Written (bytes_written)	BYTES	TRUE	Bytes written. Called 'bytes_written' in searches.
Combine Input Records (combine_input_records)	NUMBER	TRUE	Combine input records. Called 'combine_input_records' in searches.
Combine Output Records (combine_output_records)	NUMBER	TRUE	Combine output records. Called 'combine_output_records' in searches.
Committed Heap (committed_heap_bytes)	BYTES	TRUE	Total committed heap usage. Called 'committed_heap_bytes' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Completed Maps and Reduces (tasks_completed)	NUMBER	TRUE	The number of completed map and reduce tasks in this MapReduce job. Called 'tasks_completed' in searches. Available only for running jobs.
CPU Allocation (vcores_millis)	NUMBER	TRUE	CPU allocation. This is the sum of 'vcores_millis_maps' and 'vcores_millis_reduces'. Called 'vcores_millis' in searches.
CPU Time (cpu_milliseconds)	MILLISECONDS	TRUE	CPU time. Called 'cpu_milliseconds' in searches.
Data Local Maps (data_local_maps)	NUMBER	TRUE	Data local maps. Called 'data_local_maps' in searches.
Data Local Maps Percentage (data_local_maps_percentage)	NUMBER	TRUE	The number of data local maps as a percentage of the total number of maps. Called 'data_local_maps_percentage' in searches.
Diagnostics (diagnostics)	STRING	FALSE	Diagnostic information on the YARN application. If the diagnostic information is long, this may only contain the beginning of the information. Called 'diagnostics' in searches.
Duration (application_duration)	MILLISECONDS	TRUE	How long YARN took to run this application. Called 'application_duration' in searches.
Executing (executing)	BOOLEAN	FALSE	Whether the YARN application is currently running. Called 'executing' in searches.
Failed Map and Reduce Attempts (failed_tasks_attempts)	NUMBER	TRUE	The number of failed map and reduce attempts for this MapReduce job. Called 'failed_tasks_attempts' in searches. Available only for failed jobs.
Failed Map Attempts (failed_map_attempts)	NUMBER	TRUE	The number of failed map attempts for this MapReduce job. Called 'failed_map_attempts' in searches. Available only for running jobs.
Failed Maps (num_failed_maps)	NUMBER	TRUE	Failed maps. Called 'num_failed_maps' in searches.
Failed Reduce Attempts (failed_reduce_attempts)	NUMBER	TRUE	The number of failed reduce attempts for this MapReduce job. Called 'failed_reduce_attempts' in searches. Available only for running jobs.
Failed Reduces (num_failed_reduces)	NUMBER	TRUE	Failed reduces. Called 'num_failed_reduces' in searches.
Failed Shuffles (failed_shuffle)	NUMBER	TRUE	Failed shuffles. Called 'failed_shuffle' in searches.
Failed Tasks (num_failed_tasks)	NUMBER	TRUE	The total number of failed tasks. This is the sum of 'num_failed_maps' and 'num_failed_reduces'. Called 'num_failed_tasks' in searches.
Fallow Map Slots Time (fallow_slots_millis_maps)	MILLISECONDS	TRUE	Fallow map slots time. Called 'fallow_slots_millis_maps' in searches.
Fallow Reduce Slots Time (fallow_slots_millis_reduces)	MILLISECONDS	TRUE	Fallow reduce slots time. Called 'fallow_slots_millis_reduces' in searches.
Fallow Slots Time (fallow_slots_millis)	MILLISECONDS	TRUE	Total fallow slots time. This is the sum of 'fallow_slots_millis_maps' and 'fallow_slots_millis_reduces'. Called 'fallow_slots_millis' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
File Bytes Read (file_bytes_read)	BYTES	TRUE	File bytes read. Called 'file_bytes_read' in searches.
File Bytes Written (file_bytes_written)	BYTES	TRUE	File bytes written. Called 'file_bytes_written' in searches.
File Large Read Operations (file_large_read_ops)	NUMBER	TRUE	File large read operations. Called 'file_large_read_ops' in searches.
File Read Operations (file_read_ops)	NUMBER	TRUE	File read operations. Called 'file_read_ops' in searches.
File Write Operations (file_write_ops)	NUMBER	TRUE	File write operations. Called 'file_large_write_ops' in searches.
Garbage Collection Time (gc_time_millis)	MILLISECONDS	TRUE	Garbage collection time. Called 'gc_time_millis' in searches.
HDFS Bytes Read (hdfs_bytes_read)	BYTES	TRUE	HDFS bytes read. Called 'hdfs_bytes_read' in searches.
HDFS Bytes Written (hdfs_bytes_written)	BYTES	TRUE	HDFS bytes written. Called 'hdfs_bytes_written' in searches.
HDFS Large Read Operations (hdfs_large_read_ops)	NUMBER	TRUE	HDFS large read operations. Called 'hdfs_large_read_ops' in searches.
HDFS Read Operations (hdfs_read_ops)	NUMBER	TRUE	HDFS read operations. Called 'hdfs_read_ops' in searches.
HDFS Write Operations (hdfs_write_ops)	NUMBER	TRUE	HDFS write operations. Called 'hdfs_write_ops' in searches.
Hive Query ID (hive_query_id)	STRING	FALSE	If this MapReduce job ran as a part of a Hive query, this field contains the ID of the Hive query. Called 'hive_query_id' in searches.
Hive Query String (hive_query_string)	STRING	TRUE	If this MapReduce job ran as a part of a Hive query, this field contains the string of the query. Called 'hive_query_string' in searches.
Hive Sentry Subject Name (hive_sentry_subject_name)	STRING	TRUE	If this MapReduce job ran as a part of a Hive query on a secured cluster using impersonation, this field contains the name of the user that initiated the query. Called 'hive_sentry_subject_name' in searches.
Input Directory (input_dir)	STRING	TRUE	The input directory for this MapReduce job. Called 'input_dir' in searches.
Input Split Bytes (split_raw_bytes)	BYTES	TRUE	Input split bytes. Called 'split_raw_bytes' in searches.
Killed Map and Reduce Attempts (killed_tasks_attempts)	NUMBER	TRUE	The number of map and reduce attempts that were killed by user(s) for this MapReduce job. Called 'killed_tasks_attempts' in searches. Available only for killed jobs.
Killed Map Attempts (killed_map_attempts)	NUMBER	TRUE	The number of map attempts killed by user(s) for this MapReduce job. Called 'killed_map_attempts' in searches. Available only for running jobs.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Killed Reduce Attempts (killed_reduce_attempts)	NUMBER	TRUE	The number of reduce attempts killed by user(s) for this MapReduce job. Called 'killed_reduce_attempts' in searches. Available only for running jobs.
Launched Map Tasks (total_launched_maps)	NUMBER	TRUE	Launched map tasks. Called 'total_launched_maps' in searches.
Launched Reduce Tasks (total_launched_reduces)	NUMBER	TRUE	Launched reduce tasks. Called 'total_launched_reduces' in searches.
Map and Reduce Attempts in NEW State (new_tasks_attempts)	NUMBER	TRUE	The number of map and reduce attempts in NEW state for this MapReduce job. Called 'new_tasks_attempts' in searches. Available only for running jobs.
Map Attempts in NEW State (new_map_attempts)	NUMBER	TRUE	The number of map attempts in NEW state for this MapReduce job. Called 'new_map_attempts' in searches. Available only for running jobs.
Map Class (mapper_class)	STRING	TRUE	The class used by the map tasks in this MapReduce job. Called 'mapper_class' in searches. You can search for the mapper class using the class name alone, for example 'QuasiMonteCarlo $QmcMapper', or the fully qualified classname, for example, 'org.apache.hadoop.examples.QuasiMonteCarlo$QmcMapper'.
Map CPU Allocation (vcores_millis_maps)	NUMBER	TRUE	Map CPU allocation. Called 'vcores_millis_maps' in searches.
Map Input Records (map_input_records)	NUMBER	TRUE	Map input records. Called 'map_input_records' in searches.
Map Memory Allocation (mb_millis_maps)	NUMBER	TRUE	Map memory allocation. Called 'mb_millis_maps' in searches.
Map Output Bytes (map_output_bytes)	BYTES	TRUE	Map output bytes. Called 'map_output_bytes' in searches.
Map Output Materialized Bytes (map_output_materialized_bytes)	BYTES	TRUE	Map output materialized bytes. Called 'map_output_materialized_bytes' in searches.
Map Output Records (map_output_records)	NUMBER	TRUE	Map output records. Called 'map_output_records' in searches.
Map Progress (map_progress)	NUMBER	TRUE	The percentage of maps completed for this MapReduce job. Called 'map_progress' in searches. Available only for running jobs.
Maps Completed (maps_completed)	NUMBER	TRUE	The number of map tasks completed as a part of this MapReduce job. Called 'maps_completed' in searches.
Map Slots Time (slots_millis_maps)	MILLISECONDS	TRUE	Total time spent by all maps in occupied slots. Called 'slots_millis_maps' in searches.
Maps Pending (maps_pending)	NUMBER	TRUE	The number of maps waiting to be run for this MapReduce job. Called 'maps_pending' in searches. Available only for running jobs.
Maps Running (maps_running)	NUMBER	TRUE	The number of maps currently running for this MapReduce job. Called 'maps_running' in searches. Available only for running jobs.
Maps Total (maps_total)	NUMBER	TRUE	The number of Map tasks in this MapReduce job. Called 'maps_total' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Memory Allocation (mb_millis)	NUMBER	TRUE	Total memory allocation. This is the sum of 'mb_millis_maps' and 'mb_millis_reduces'. Called 'mb_millis' in searches.
Merged Map Outputs (merged_map_outputs)	NUMBER	TRUE	Merged map outputs. Called 'merged_map_outputs' in searches.
Name (name)	STRING	TRUE	Name of the YARN application. Called 'name' in searches.
Oozie Workflow ID (oozie_id)	STRING	FALSE	If this MapReduce job ran as a part of an Oozie workflow, this field contains the ID of the Oozie workflow. Called 'oozie_id' in searches.
Other Local Maps (other_local_maps)	NUMBER	TRUE	Other local maps. Called 'other_local_maps' in searches.
Other Local Maps Percentage (other_local_maps_percentage)	NUMBER	TRUE	The number of other local maps as a percentage of the total number of maps. Called 'other_local_maps_percentage' in searches.
Output Directory (output_dir)	STRING	TRUE	The output directory for this MapReduce job. Called 'output_dir' in searches.
Pending Maps and Reduces (tasks_pending)	NUMBER	TRUE	The number of maps and reduces waiting to be run for this MapReduce job. Called 'tasks_pending' in searches. Available only for running jobs.
Physical Memory (physical_memory_bytes)	BYTES	TRUE	Physical memory. Called 'physical_memory_bytes' in searches.
Pig Script ID (pig_id)	STRING	FALSE	If this MapReduce job ran as a part of a Pig script, this field contains the ID of the Pig script. Called 'pig_id' in searches.
Pool (pool)	STRING	TRUE	The name of the resource pool in which this application ran. Called 'pool' in searches. Within YARN, a pool is referred to as a queue.
Progress (progress)	NUMBER	TRUE	The progress reported by the application. Called 'progress' in searches.
Rack Local Maps (rack_local_maps)	NUMBER	TRUE	Rack local maps. Called 'rack_local_maps' in searches.
Rack Local Maps Percentage (rack_local_maps_percentage)	NUMBER	TRUE	The number of rack local maps as a percentage of the total number of maps. Called 'rack_local_maps_percentage' in searches.
Reduce Attempts in NEW State (new_reduce_attempts)	NUMBER	TRUE	The number of reduce attempts in NEW state for this MapReduce job. Called 'new_reduce_attempts' in searches. Available only for running jobs.
Reduce Class (reducer_class)	STRING	TRUE	The class used by the reduce tasks in this MapReduce job. Called 'reducer_class' in searches. You can search for the reducer class using the class name alone, for example 'QuasiMonteCarlo$QmcReducer', or fully qualified classname, for example, 'org.apache.hadoop.examples.QuasiMonteCarlo $QmcReducer'.
Reduce CPU Allocation (vcores_millis_reduces)	NUMBER	TRUE	Reduce CPU allocation. Called 'vcores_millis_reduces' in searches.
Reduce Input Groups (reduce_input_groups)	NUMBER	TRUE	Reduce input groups. Called 'reduce_input_groups' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Reduce Input Records (reduce_input_records)	NUMBER	TRUE	Reduce input records. Called 'reduce_input_records' in searches.
Reduce Memory Allocation (mb_millis_reduces)	NUMBER	TRUE	Reduce memory allocation. Called 'mb_millis_reduces' in searches.
Reduce Output Records (reduce_output_records)	NUMBER	TRUE	Reduce output records. Called 'reduce_output_records' in searches.
Reduce Progress (reduce_progress)	NUMBER	TRUE	The percentage of reduces completed for this MapReduce job. Called 'reduce_progress' in searches. Available only for running jobs.
Reduces Completed (reduces_completed)	NUMBER	TRUE	The number of reduce tasks completed as a part of this MapReduce job. Called 'reduces_completed' in searches.
Reduce Shuffle Bytes (reduce_shuffle_bytes)	BYTES	TRUE	Reduce shuffle bytes. Called 'reduce_shuffle_bytes' in searches.
Reduce Slots Time (slots_millis_reduces)	MILLISECONDS	TRUE	Total time spent by all reduces in occupied slots. Called 'slots_millis_reduces' in searches.
Reduces Pending (reduces_pending)	NUMBER	TRUE	The number of reduces waiting to be run for this MapReduce job. Called 'reduces_pending' in searches. Available only for running jobs.
Reduces Running (reduces_running)	NUMBER	TRUE	The number of reduces currently running for this MapReduce job. Called 'reduces_running' in searches. Available only for running jobs.
Reduces Total (reduces_total)	NUMBER	TRUE	The number of reduce tasks in this MapReduce job. Called 'reduces_total' in searches.
Running Containers (running_containers)	NUMBER	FALSE	The number of containers currently running for the application. Called 'running_containers' in searches.
Running Map and Reduce Attempts (running_tasks_attempts)	NUMBER	TRUE	The number of map and reduce attempts currently running for this MapReduce job. Called 'running_tasks_attempts' in searches. Available only for running jobs.
Running Map Attempts (running_map_attempts)	NUMBER	TRUE	The number of running map attempts for this MapReduce job. Called 'running_map_attempts' in searches. Available only for running jobs.
Running MapReduce Application Information Retrieval Duration. (running_application_info_retrieval_time)	NUMBER	TRUE	How long it took, in seconds, to retrieve information about the MapReduce application.
Running Maps and Reduces (tasks_running)	NUMBER	TRUE	The number of maps and reduces currently running for this MapReduce job. Called 'tasks_running' in searches. Available only for running jobs.
Running Reduce Attempts (running_reduce_attempts)	NUMBER	TRUE	The number of running reduce attempts for this MapReduce job. Called 'running_reduce_attempts' in searches. Available only for running jobs.
Service Name (service_name)	STRING	FALSE	The name of the YARN service. Called 'service_name' in searches.
Shuffle Bad ID Errors (shuffle_errors_bad_id)	NUMBER	TRUE	Shuffle bad ID errors. Called 'shuffle_errors_bad_id' in searches.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
Shuffle Connection Errors (shuffle_errors_connection)	NUMBER	TRUE	Shuffle connection errors. Called 'shuffle_errors_connection' in searches.
Shuffled Maps (shuffled_maps)	NUMBER	TRUE	Shuffled maps. Called 'shuffled_maps' in searches.
Shuffle IO Errors (shuffle_errors_io)	NUMBER	TRUE	Shuffle IO errors. Called 'shuffle_errors_io' in searches.
Shuffle Wrong Length Errors (shuffle_errors_wrong_length)	NUMBER	TRUE	Shuffle wrong length errors. Called 'shuffle_errors_wrong_length' in searches.
Shuffle Wrong Map Errors (shuffle_errors_wrong_map)	NUMBER	TRUE	Shuffle wrong map errors. Called 'shuffle_errors_wrong_map' in searches.
Shuffle Wrong Reduce Errors (shuffle_errors_wrong_reduce)	NUMBER	TRUE	Shuffle wrong reduce errors. Called 'shuffle_errors_wrong_reduce' in searches.
Slots Time (slots_millis)	MILLISECONDS	TRUE	Total slots time. This is the sum of 'slots_millis_maps' and 'slots_millis_reduces'. Called 'slots_millis' in searches.
Spilled Records (spilled_records)	NUMBER	TRUE	Spilled Records. Called 'spilled_records' in searches.
Successful Map and Reduce Attempts (successful_tasks_attempts)	NUMBER	TRUE	The number of successful map and reduce attempts for this MapReduce job. Called 'successful_tasks_attempts' in searches. Available only for successful jobs.
Successful Map Attempts (successful_map_attempts)	NUMBER	TRUE	The number of successful map attempts for this MapReduce job. Called 'successful_map_attempts' in searches. Available only for running jobs.
Successful Reduce Attempts (successful_reduce_attempts)	NUMBER	TRUE	The number of successful reduce attempts for this MapReduce job. Called 'successful_reduce_attempts' in searches. Available only for running jobs.
Total Maps and Reduces Number (total_task_num)	NUMBER	TRUE	The number of map and reduce tasks in this MapReduce job. Called 'tasks_total' in searches. Available only for running jobs.
Total Tasks (total_launched_tasks)	NUMBER	TRUE	The total number of tasks. This is the sum of 'total_launched_maps' and 'total_launched_reduces'. Called 'total_launched_tasks' in searches.
Tracking Url (tracking_url)	STRING	FALSE	The MapReduce application tracking URL.
Uberized Job (uberized)	BOOLEAN	FALSE	Whether this MapReduce job is uberized - running completely in the ApplicationMaster. Called 'uberized' in searches. Available only for running jobs.
Unused Memory Seconds (unused_memory_seconds)	NUMBER	TRUE	The amount of memory the application has allocated but not used (megabyte-seconds). This metric is calculated hourly if container usage metric aggregation is enabled. Called 'unused_memory_seconds' in searches.
Unused VCore Seconds (unused_vcore_seconds)	NUMBER	TRUE	The amount of CPU resources the application has allocated but not used (virtual core-seconds). This metric is calculated hourly if container usage metric aggregation is enabled. Called 'unused_vcore_seconds' in searches.
Used Memory Max (used_memory_max)	NUMBER	TRUE	The maximum container memory usage for a YARN application. This metric is calculated hourly if container usage metric aggregation is enabled and a Cloudera Manager Container Usage Metrics Directory is specified.

Display Name (Attribute Name)	Type	Supports Filtering?	Description
User (user)	STRING	TRUE	The user who ran the YARN application. Called 'user' in searches.
Virtual Memory (virtual_memory_bytes)	BYTES	TRUE	Virtual memory. Called 'virtual_memory_bytes' in searches.
Work CPU Time (cm_cpu_milliseconds)	MILLISECONDS	TRUE	Attribute measuring the sum of CPU time used by all threads of the query, in milliseconds. Called 'work_cpu_time' in searches. For Impala queries, CPU time is calculated based on the 'TotalCpuTime' metric. For YARN MapReduce applications, this is calculated from the 'cpu_milliseconds' metric.

Examples

Consider the following filter expressions: user = "root", rowsProduced > 0, fileFormats RLIKE ".TEXT.*", and executing = true. In the examples:

- The filter attributes are user, rowsProduced, fileFormats, and executing.
- The operators are =, >, and RLIKE.
- The filter values are root, 0, .TEXT.*, and true.

Sending Diagnostic Data to Cloudera for YARN Applications

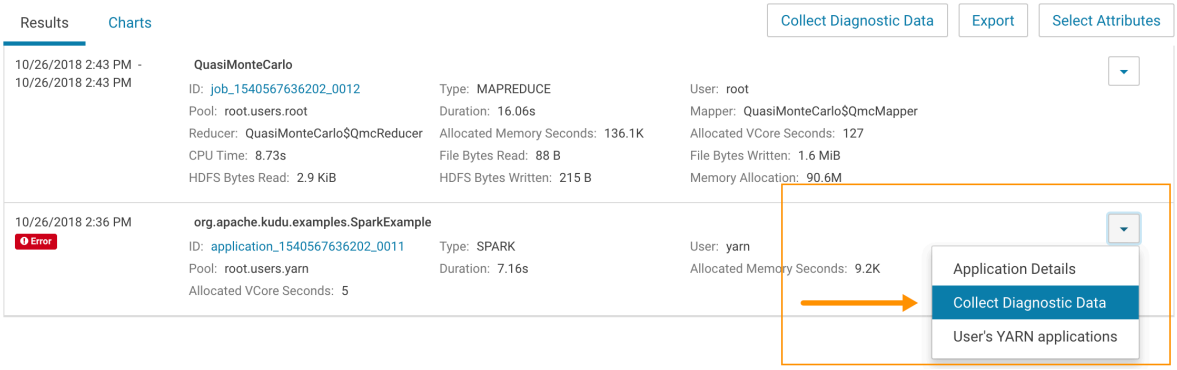
You can send diagnostic data collected from YARN applications, including metadata, configurations, and log data, to Cloudera Support for analysis. Include a support ticket number if one exists to enable Cloudera Support to address the issue more quickly and efficiently.

About this task

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Procedure

1. From the YARN page in Cloudera Manager, click the Applications menu.
2. Collect diagnostics data. There are two ways to do this:
   - To collect data from all applications that are visible in the list, click the top Collect Diagnostics Data button on the upper right, above the list of YARN applications.
   - To collect data from only one specific application, click the down arrow on the right-hand end of the row of the application and select Collect Diagnostics Data.



3. In the Send YARN Applications Diagnostic Data dialog box, provide the following information:
   - If applicable, the Cloudera Support ticket number of the issue being experienced on the cluster.
   - Optionally, add a comment to help the support team understand the issue.
4. Click the checkbox Send Diagnostic Data to Cloudera.
5. Click the button Collect and Send Diagnostic Data.

Note: Passwords from configuration will not be retrieved.

Monitoring Spark Applications

To obtain information about Spark application behavior you can consult cluster manager logs and the Spark web application UI. These two methods provide complementary information.

Logs enable you to see fine grained events in the lifecycle of an application. The web UI provides both a broad overview of the various aspects of Spark application behavior and fine grained metrics. This section provides an overview of both methods.

For further information on Spark monitoring, see the Spark documentation on monitoring and instrumentation.

Related Information

Monitoring and Instrumentation

Viewing and Debugging Spark Applications Using Logs

You can view overview information about all running Spark applications.

Procedure

1. Go to the **YARN Applications** page in the Cloudera Manager Admin Console.
2. To debug Spark applications running on YARN, view the logs for the NodeManager role. Open the log event viewer.
3. Filter the event stream.
4. For any event, click View Log File to view the entire log file.

Related Information

Monitoring YARN Applications

Viewing Logs

Filtering Logs

Managing Spark Driver Logs

You can change the default directory for Spark driver logs, or disable the log collection.

About this task

The Spark service collects Spark driver logs when Spark applications are run in YARN-client mode or with the Spark Shell. This feature is enabled by default, and the logs are persisted to an HDFS directory and included in YARN Diagnostic Bundles. The default directory for the logs is /user/spark/driverLogs.

Procedure

1. In the Cloudera Manager Admin Console, go to the Spark service.
2. Select the Configuration page.

3. Configure the default log directory, or disable the log collection:
   - To configure the default log directory, search for the Spark Driver Log Location configuration and change the directory specified in the field.
   - To disable the log collection, search for the Persist Driver Logs to Dfs configuration.
4. Save the changes.
5. Deploy the updated client configuration.

Visualizing Spark Applications Using the Web Application UI

Every Spark application launches a web application UI that displays useful information about the application.

The web application UI includes the following information:

- An event timeline that displays the relative ordering and interleaving of application events. The timeline view is available on three levels: across all jobs, within one job, and within one stage. The timeline also shows executor allocation and deallocation.
- A list of stages and tasks.
- The execution directed acyclic graph (DAG) for each job.
- A summary of RDD sizes and memory usage.
- Environment - runtime information, property settings, library paths.
- Information about Spark SQL jobs.

The web UI is available in different ways depending on whether the application is running or has completed.

Accessing the Web UI of a Running Spark Application

You can access the web UI of a running Spark application from a web browser.

To access the web application UI of a running Spark application, open http://*spark_driver_host*:4040 in a web browser. If multiple applications are running on the same host, the web application binds to successive ports beginning with 4040 (4041, 4042, and so on). The web application is available only for the duration of the application.

Accessing the Web UI of a Completed Spark Application

You can access the web application UI of a completed Spark application through the Spark history server.

Procedure

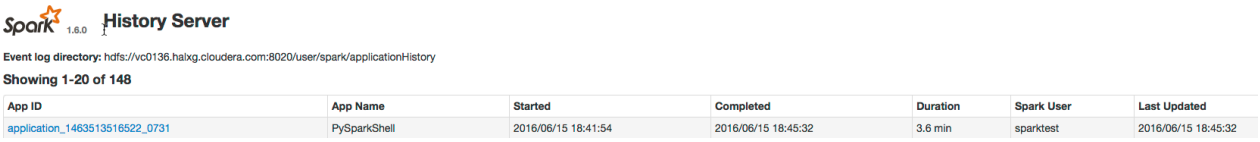
1. Open the Spark History Server UI in one of the following ways:
   - Open the URL http://*spark_history_server_host*:18088.
   - Open the UI in the Cloudera Manager Admin Console:
     a. Go to the Spark service.
     b. Click the History Server Web UI link.

   The History Server displays a list of completed applications.
2. In the list of applications, click an App ID link. The application UI displays.

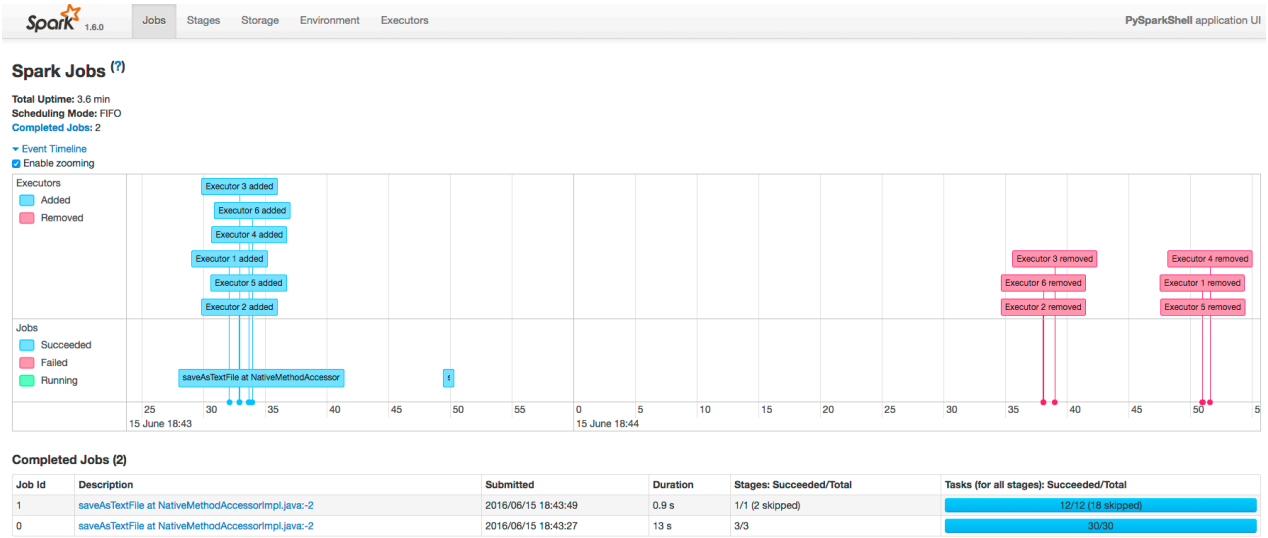
   Note: In CDH 5.10 and higher, and in CDK 2.x Powered By Apache Spark, the Storage tab of the Spark History Server is always blank. To see storage information while an application is running, use the web UI of the application as described in the previous section. After the application is finished, storage information is not available.

Example Spark Application Web Application

Consider a job consisting of a set of transformation to join data from an accounts dataset with a weblogs dataset in order to determine the total number of web results for every account and then an action write the result to HDFS. In this example, the write is performed twice, resulting in two jobs. To view the application UI, in the History Server click the link in the App ID column:



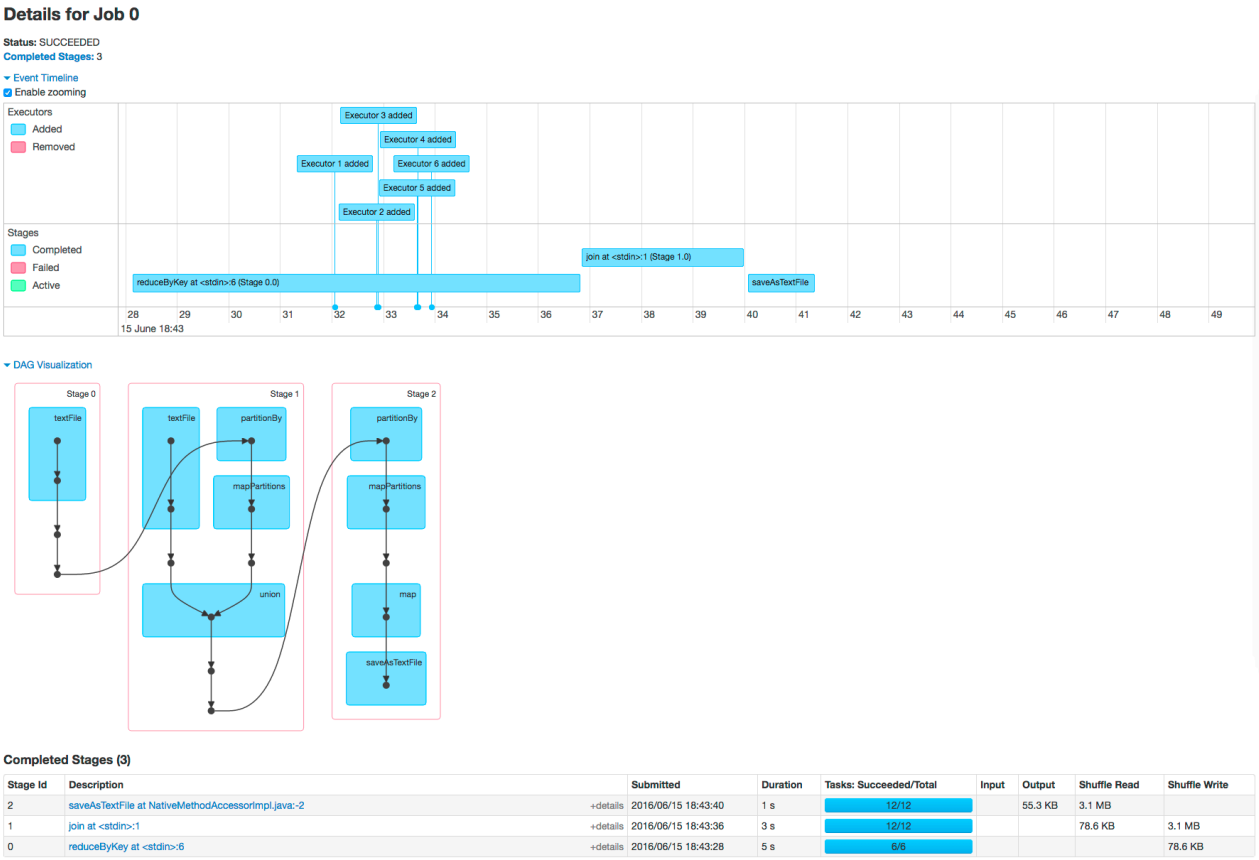
The following screenshot shows the timeline of the events in the application including the jobs that were run and the allocation and deallocation of executors. Each job shows the last action, saveAsTextFile, run for the job. The timeline shows that the application acquires executors over the course of running the first job. After the second job finishes, the executors become idle and are returned to the cluster.



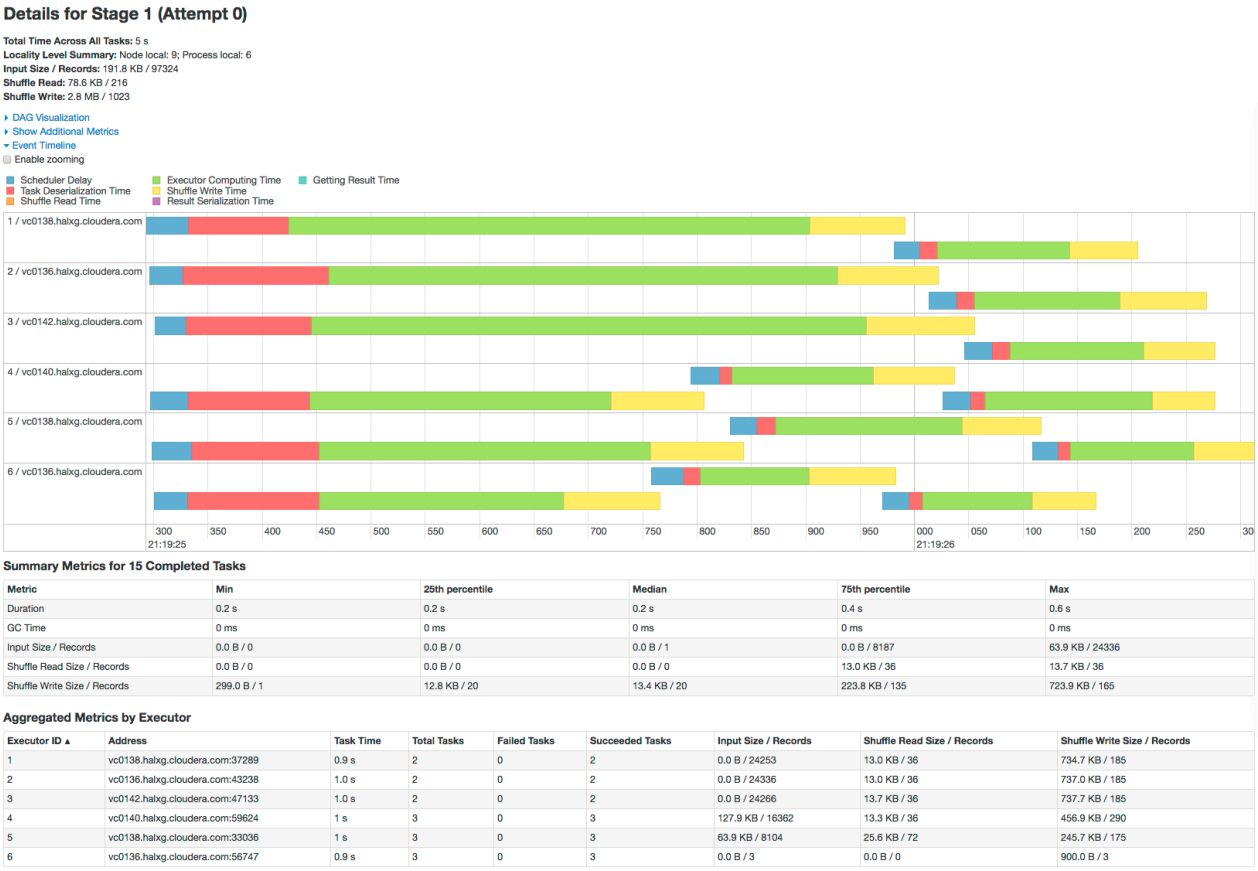
You can manipulate the timeline as follows:

- Pan - Press and hold the left mouse button and swipe left and right.
- Zoom - Select the Enable zooming checkbox and scroll the mouse up and down.

To view the details for Job 0, click the link in the Description column. The following screenshot shows details of each stage in Job 0 and the DAG visualization. Zooming in shows finer detail for the segment from 28 to 42 seconds:



Clicking a stage shows further details and metrics:



Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.2 s	0.2 s	0.2 s	0.4 s	0.6 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Input Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 1	0.0 B / 8187	63.9 KB / 24336
Shuffle Read Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 0	13.0 KB / 36	13.7 KB / 36
Shuffle Write Size / Records	299.0 B / 1	12.8 KB / 20	13.4 KB / 20	223.8 KB / 135	723.9 KB / 165

Executor ID ▲	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Read Size / Records	Shuffle Write Size / Records
1	vc0138.halxg.cloudera.com:37289	0.9 s	2	0	2	0.0 B / 24253	13.0 KB / 36	734.7 KB / 185
2	vc0136.halxg.cloudera.com:43238	1.0 s	2	0	2	0.0 B / 24336	13.0 KB / 36	737.0 KB / 185
3	vc0142.halxg.cloudera.com:47133	1.0 s	2	0	2	0.0 B / 24266	13.7 KB / 36	737.7 KB / 185
4	vc0140.halxg.cloudera.com:59624	1 s	3	0	3	127.9 KB / 16362	13.3 KB / 36	456.9 KB / 290
5	vc0138.halxg.cloudera.com:33036	1 s	3	0	3	63.9 KB / 8104	25.6 KB / 72	245.7 KB / 175
6	vc0136.halxg.cloudera.com:56747	0.9 s	3	0	3	0.0 B / 3	0.0 B / 0	900.0 B / 3

The web page for Job 1 shows how preceding stages are skipped because Spark retains the results from those stages:

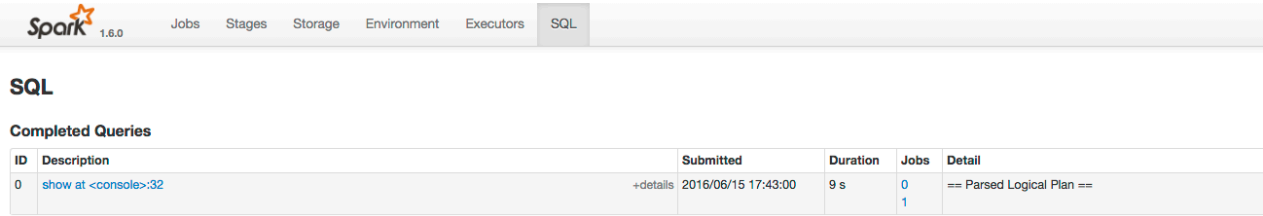
Stage Id	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
5	saveAsTextFile at NativeMethodAccessorImpl.java:-2	+details	2016/06/15 18:43:49	0.8 s	12/12		55.3 KB	3.1 MB	

Stage Id	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
4	join at <stdin>:1	+details	Unknown	Unknown	0/12				
3	reduceByKey at <stdin>:6	+details	Unknown	Unknown	0/6				

Example Spark SQL Web Application

In addition to the screens described above, the web application UI of an application that uses the Spark SQL API also has an SQL tab. Consider an application that loads the contents of two tables into a pair of DataFrames, joins

the tables, and then shows the result. After you click the application ID, the SQL tab displays the final action in the query:

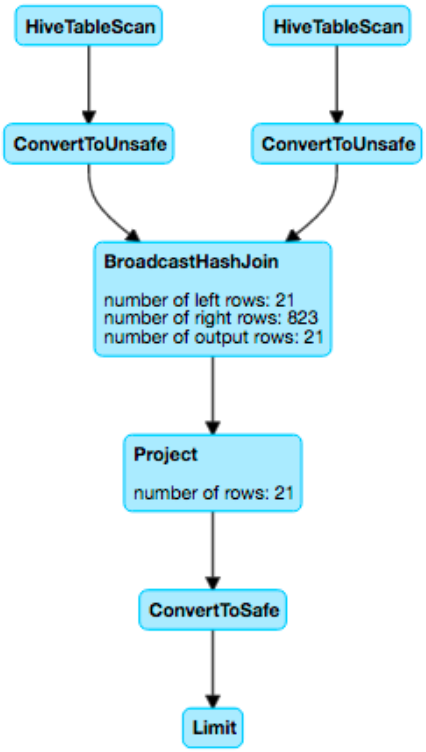


If you click the show link you see the DAG of the job. Clicking the Details link on this page displays the logical query plan:

Spark 1.6.0 Jobs Stages Storage Environment Executors SQL

Details for Query 0

Submitted Time: 2016/06/15 17:43:00
Duration: 9 s
Succeeded Jobs: 0 1



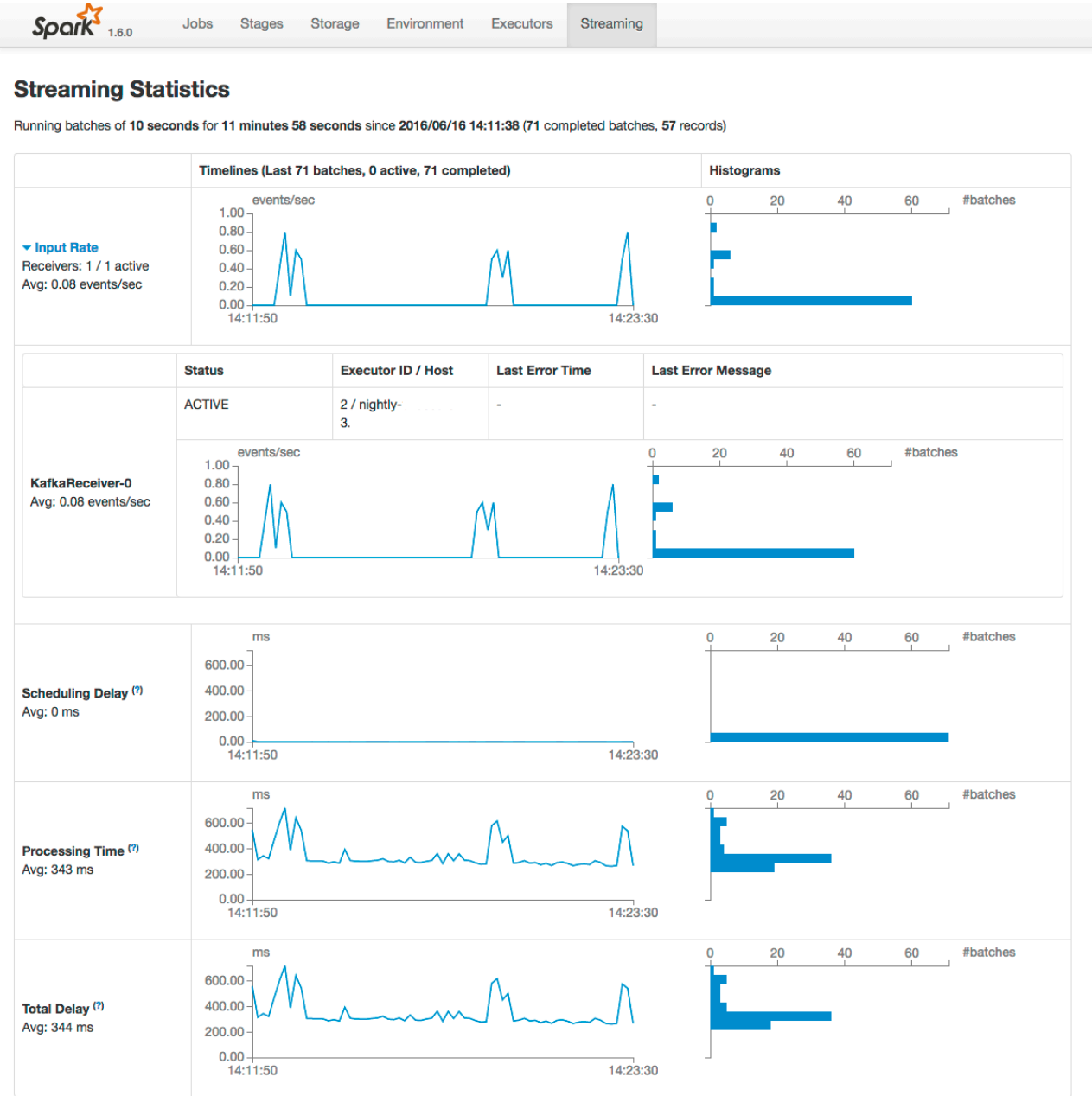
▾ Details

```
== Parsed Logical Plan ==
Limit 21
+- Project [code#0,description#1]
   +- Join Inner, Some((code#0 = code#4))
      :- Project [code#0,description#1,total_emp#2,salary#3]
      :  +- MetastoreRelation default, sample_07, None
      +- Project [code#4,description#5,total_emp#6,salary#7]
         +- MetastoreRelation default, sample_08, None

== Analyzed Logical Plan ==
code: string, description: string
Limit 21
+- Project [code#0,description#1]
   +- Join Inner, Some((code#0 = code#4))
      :- Project [code#0,description#1,total_emp#2,salary#3]
      :  +- MetastoreRelation default, sample_07, None
      +- Project [code#4,description#5,total_emp#6,salary#7]
         +- MetastoreRelation default, sample_08, None

== Optimized Logical Plan ==
Limit 21
+- Project [code#0,description#1]
   +- Join Inner, Some((code#0 = code#4))
      :- Project [code#0,description#1]
      :  +- MetastoreRelation default, sample_07, None
      +- Project [code#4]
         +- MetastoreRelation default, sample_08, None

== Physical Plan ==
```

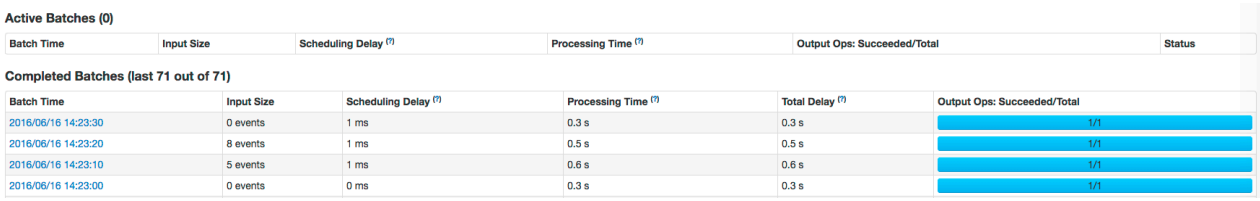
Example Spark Streaming Web Application

Note: The following example demonstrates the Spark driver web UI. Streaming information is not captured in the Spark History Server.

The Spark driver web application UI also supports displaying the behavior of streaming applications in the Streaming tab. If you run the example described in the Spark streaming example, and provide three bursts of data, the top of the tab displays a series of visualizations of the statistics summarizing the overall behavior of the streaming application:



The application has one receiver that processed 3 bursts of event batches, which can be observed in the events, processing time, and delay graphs. Further down the page you can view details of individual batches:



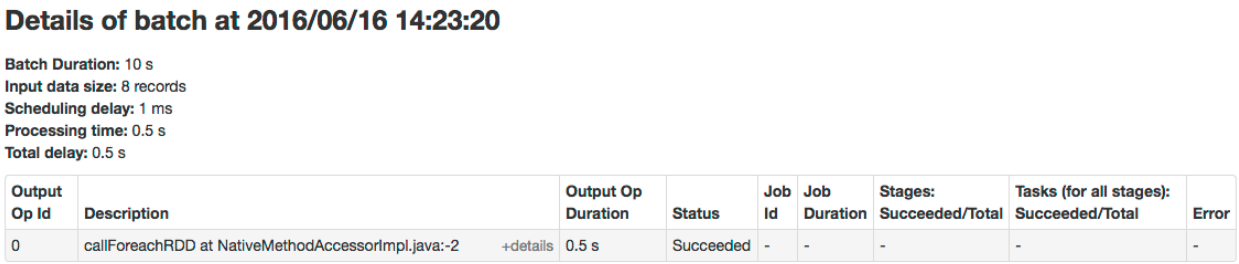
Active Batches (0)

Batch Time	Input Size	Scheduling Delay (?)	Processing Time (?)	Output Ops: Succeeded/Total	Status

Completed Batches (last 71 out of 71)

Batch Time	Input Size	Scheduling Delay (?)	Processing Time (?)	Total Delay (?)	Output Ops: Succeeded/Total
2016/06/16 14:23:30	0 events	1 ms	0.3 s	0.3 s	1/1
2016/06/16 14:23:20	8 events	1 ms	0.5 s	0.5 s	1/1
2016/06/16 14:23:10	5 events	1 ms	0.6 s	0.6 s	1/1
2016/06/16 14:23:00	0 events	0 ms	0.3 s	0.3 s	1/1

To view the details of a specific batch, click a link in the Batch Time column. Clicking the 2016/06/16 14:23:20 link with 8 events in the batch, provides the following details:



Details of batch at 2016/06/16 14:23:20

Batch Duration: 10 s
Input data size: 8 records
Scheduling delay: 1 ms
Processing time: 0.5 s
Total delay: 0.5 s

Output Op Id	Description	Output Op Duration	Status	Job Id	Job Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total	Error
0	callForeachRDD at NativeMethodAccessorImpl.java:-2 +details	0.5 s	Succeeded	-	-	-	-	-

Events

An *event* is a record that something of interest has occurred – a service's health has changed state, a log message (of the appropriate severity) has been logged, and so on. Many events are enabled and configured by default.

From the **Events** page you can filter for events for services or role instances, hosts, users, commands, and much more. You can also search against the content information returned by the event.

The Event Server aggregates relevant events and makes them available for alerting and for searching. This way, you have a view into the history of all relevant events that occur cluster-wide.

Cloudera Manager supports the following categories of events:

Category	Description
ACTIVITY_EVENT	Generated by the Activity Monitor; specifically, for jobs that fail, or that run slowly (as determined by comparison with duration limits). In order to monitor your workload for slow-running jobs, you must specify activity duration rules.
AUDIT_EVENT	Generated by actions performed • In Cloudera Manager, such as creating, configuring, starting, stopping, and deleting services or roles • By services that are being audited by Cloudera Navigator.
HBASE	Generated by HBase with the exception of log messages, which have the LOG_MESSAGE category.
HEALTH_CHECK	Indicate that certain health test activities have occurred, or that health test results have met specific conditions (thresholds). Thresholds for various health tests can be set under the Configuration tabs for HBase, HDFS, Impala, and MapReduce service instances, at both the service and role level.
LOG_MESSAGE	Generated for certain types of log messages from HDFS, MapReduce, and HBase services and roles. Log events are created when a log entry matches a set of rules for identifying messages of interest. The default set of rules is based on Cloudera experience supporting Hadoop clusters. You can configure additional log event rules if necessary.
SYSTEM	Generated by system events such as parcel availability.

For detailed information on each supported event, see the Cloudera Manager Events reference documentation.

Related Information

Configuring Health Monitoring

Configuring Log Events
Cloudera Manager Events Reference

Viewing Events

About this task

The Events page lets you display events and alerts that have occurred within a time range you select anywhere in your clusters. From the Events page you can filter for events for services or role instances, hosts, users, commands, and much more. You can also search against the content information returned by the event.

Procedure

- To view events, click the Diagnostics tab on the top navigation bar, then select Events.

  Event entries are ordered (within the time range you've selected) with the most recent at the top. If the event generated an alert, that is indicated by a red alert icon (Alert) in the entry.

  This page supports infinite scrolling: you can scroll to the end of the displayed results and the page will fetch more results and add them to the end of the list automatically.
- To display event details, click › Expand at the right side of the event entry.

  Clicking the View link at the far right of the entry has different results depending on the category of the entry:

  - ACTIVITY_EVENT - Displays the activity **Details** page.
  - AUDIT_EVENT - If the event was a restart, displays the service's Commands page. If the event was a configuration change, the Revision Details dialog box displays.
  - HBASE - Displays a health report or log details.
  - HEALTH_CHECK - Displays the status page of the role instance.
  - LOG_MESSAGE - Displays the event's log entry. You can also click › Expand to display details of the entry, then click the URL link. When you perform one of these actions the time range in the Time Line is shifted to the time the event occurred.
  - SYSTEM - Displays the Parcels page.

Related Information

Alerts
Viewing Activity Details in a Report Format
Viewing Running and Recent Commands
Viewing Role Instance Status

Filtering Events

You filter events by selecting a time range and adding filters.

You can use the Time Range Selector or a duration link (30m 1h 2h 6h 12h 1d 7d 30d) to set the time range. The time it takes to perform a search will typically increase for a longer time range, as the number of events to be searched will be larger.

To re-run a recently performed search, click to the right of the Search button and select a search.

Related Information

Time Line

Adding an Event Filter

Procedure

1. Choose a property in the drop-down list. You can search by properties such as Username, Service, Command, or Role. The properties vary depending on the service or role.
2. If the property allows it, choose an operator in the operator drop-down list.
3. Type a property value in the value text field. For some properties you can include multiple values in the value field. For example, you can create a filter like Category = HEALTH_CHECK LOG_MESSAGE. To drop individual values, click the × to the right of the value. For properties where the list of values is finite and known, you can start typing and then select from a drop-down list of potential matches.
4. Click Search. The log displays all events that match the filter criteria.
5. Click ✚ to add more filters and repeat steps 1 through 4.

Removing an Event Filter

You can remove an event filter.

Procedure

1. Click the ➖ at the right of the filter. The filter is removed.
2. Click Search. The log displays all events that match the filter criteria.

Alerts

An *alert* is an event that is considered especially noteworthy and is triggered by a selected event. Alerts are shown with an Alert badge when they appear in a list of events. You can configure the Alert Publisher to send alert notifications by email or by SNMP trap to a trap receiver.

Service instances of type HDFS, MapReduce, and HBase (and their associated roles) can generate alerts if so configured. Alerts can also be configured for the monitoring roles that are a part of the Cloudera Management Service.

The settings to enable or disable specific alerts are found under the Configuration tab for the services to which they pertain.

Related Information

Events

Managing Alerts

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

The AdministrationAlerts page provides a summary of the settings for alerts in your clusters.

Alert Type The left column lets you select by alert type (Health, Log, or Activity) and within that by service instance. In the case of Health alerts, you can look at alerts for Hosts as well. You can select an individual service to see just the alert settings for that service.

Health/Log/Activity Alert Settings Depending on your selection in the left column, the right hand column show you the list of alerts that are enabled or disabled for the selected service type.

To change the alert settings for a service, click Edit next to the service name. This will take you to the Monitoring section of the Configuration tab for the service. From here you can enable or disable alerts and configure thresholds as needed.

Recipients You can also view the list of recipients configured for the enabled alerts.

Configuring Alert Delivery

When you install Cloudera Manager you can configure the mail server you will use with the Alert Publisher. However, if you need to change these settings, you can do so under the Alert Publisher section of the Management Services configuration tab. Under the Alert Publisher role of the Cloudera Manager Management Service, you can configure email or SNMP delivery of alert notifications and you can also configure a custom script that runs in response to an alert.

Viewing Which Alerts are Enabled and Disabled

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Select AdministrationAlerts.

Configuring Alert Email Delivery

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Sending A Test Alert E-mail

Select the AdministrationAlerts tab and click the Send Test Alert link.

Configuring the List Of Alert Recipient Email Addresses

1. Click ClustersCloudera Management Service.
2. Select the Configuration tab.
3. Select ScopeAlert Publisher.
4. Select CategoryMain.
5. Locate the Alerts: Mail Message Recipients property or search for it by typing its name in the Search box.
6. Configure the Alerts: Mail Message Recipients property. Use the + icon to add new recipients.
7. Click the Save Changes button at the top of the page to save your settings.

   To apply this configuration property to other role groups as needed, edit the value for the appropriate role group. See the topic *Modifying Configuration Properties Using Cloudera Manager*.
8. Restart the Alert Publisher role.

Configuring Alert Email Properties

1. Display the Cloudera Management Service status page.
2. Click the Configuration tab.
3. Select ScopeAlert Publisher.

4. Select CategoryMain to see the list of properties. To receive email alerts, you must set (or verify) the following settings:
   - Enable email alerts
   - Email protocol to use.
   - Your mail server hostname and port.
   - The username and password of the email user that will be logged into the mail server as the "sender" of the alert emails.

     Important: If a password contains special characters such as @, #, $, %, ^, &, and *, then enclose the password inside the RAW() method. For example, "RAW(pass@*word)".
   - A comma-separated list of email addresses that will be the recipients of alert emails.
   - The format of the email alert message. Select json if you need the message to be parsed by a script or program.

   To apply this configuration property to other role groups as needed, edit the value for the appropriate role group. See the topic *Modifying Configuration Properties Using Cloudera Manager*.
5. Click the Save Changes button at the top of the page to save your settings.
6. Restart the Alert Publisher role.

Related Information

Modifying Configuration Properties Using Cloudera Manager

Configuring Alert SNMP Delivery

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Enabling, Configuring, and Disabling SNMP Traps

1. Before you enable SNMP traps, configure the trap receiver (Network Management System or SNMP server) with the Cloudera MIB.
2. Do one of the following:
   - Select Clusters Cloudera Management Service .
   - On the HomeStatus tab, in Cloudera Management Service table, click the Cloudera Management Service link.
3. Click the Configuration tab.
4. Select ScopeAlert Publisher SNMP.
5. Select CategorySNMP
   - Enter the DNS name or IP address of the Network Management System (SNMP server) acting as the trap receiver in the SNMP NMS Hostname property.
   - In the SNMP Security Level property, select the version of SNMP you are using: SNMPv2, SNMPv3 without authentication and without privacy (noAuthNoPriv), or SNMPv3 with authentication and without privacy (auth NoPriv) and specify the required properties:
     - SNMPv2 - SNMPv2 Community String.
     - SNMPv3 without authentication (noAuthNoPriv) - SNMP Server Engine Id and SNMP Security UserName.
     - SNMPv3 with authentication (authNoPriv) - SNMP Server Engine Id, SNMP Security UserName, SNMP Authentication Protocol, and SNMP Authentication Protocol Pass Phrase.
   - You can also change other settings such as the port, retry, or timeout values.

   To apply this configuration property to other role groups as needed, edit the value for the appropriate role group. See the topic *Modifying Configuration Properties Using Cloudera Manager*.
6. Click Save Changes when you are done.
7. Restart the Alert Publisher role.

To disable SNMP traps, remove the hostname from the SNMP NMS Hostname property (alert.snmp.server.hostname).

Viewing the Cloudera MIB

1. Do one of the following:
   - Select Clusters Cloudera Management Service .
   - On the HomeStatus tab, in Cloudera Management Service table, click the Cloudera Management Service link.
2. Click the Configuration tab.
3. Select ScopeAlert Publisher SNMP.
4. Select CategorySNMP.
5. Locate the SNMP NMS Hostname property and click the ? icon to display the property description.
6. Click the SMNP Mib link.

Related Information

Modifying Configuration Properties Using Cloudera Manager

Configuring Custom Alert Scripts

You can configure the Alert Publisher to call a user-provided script in response to an alert.

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

You can configure the Alert Publisher to run a user-written script in response to an alert. The Alert Publisher passes a single argument to the script that is a UTF-8 JSON file containing a list of alerts. The script runs on the host where the Alert Publisher service is running and must have read and execute permissions for the cloudera-scm user. Only one instance of a script runs at a time. The standard out and standard error messages from the script are logged to the Alert Publisher log file.

You use the Alert Publisher: Maximum Batch Size and Alert Publisher: Maximum Batch interval to configure when the Alert Publisher delivers alerts.

To configure the Alert Publisher to deliver alerts using a script:

1. Save the script on the host where the Alert Publisher role is running.
2. Change the owner of the file to cloudera-scm and set its permissions to read and execute:

```
sudo chown cloudera-scm:cloudera-scm path_to_script
sudo chmod u+rx path_to_script
```

3. Open the Cloudera Manager Admin console and select ClustersCloudera Management Service.
4. Click the Configuration tab.
5. Select ScopeAlert Publisher .
6. Enter the path to the script in the Custom Alert Script property.
7. Enter a Reason for change, and then click Save Changes to commit the changes.

A sample Python script is available that parses the alert file and sends alerts to different email addresses based on the content of the alert. See Sample Custom Alert Script on page 114.

Sample JSON Alert File

When a custom script runs, it passes a JSON file that contains the alerts. For example:

```
[ {
  "body" : {
    "alert" : {
```

```
      "content" : "The health test result for MAPREDUCE_HA_JOB_TRACKER_HEAL
TH has become bad: JobTracker summary: myCluster.com (Availability: Active,
Health: Bad). This health test reflects the health of the active JobTracker.
",
      "timestamp" : {
        "iso8601" : "2015-06-11T03:52:56Z",
        "epochMs" : 1433994776083
      },
      "source" : "http://myCluster.com:7180/cmf/eventRedirect/89521139-0859-
4bef-bf65-eb141e63dbba",
      "attributes" : {
        "__persist_timestamp" : [ "1433994776172" ],
        "ALERT_SUPPRESSED" : [ "false" ],
        "HEALTH_TEST_NAME" : [ "MAPREDUCE_HA_JOB_TRACKER_HEALTH" ],
        "SEVERITY" : [ "CRITICAL" ],
        "HEALTH_TEST_RESULTS" : [ {
          "content" : "The health test result for MAPREDUCE_HA_JOB_TRACKER_H
EALTH has become bad: JobTracker summary: myCluster.com (Availability: Activ
e, Health: Bad). This health test reflects the health of the active JobTrack
er.",
          "testName" : "MAPREDUCE_HA_JOB_TRACKER_HEALTH",
          "eventCode" : "EV_SERVICE_HEALTH_CHECK_BAD",
          "severity" : "CRITICAL"
        } ],
        "CLUSTER_DISPLAY_NAME" : [ "Cluster 1" ],
        "ALERT" : [ "true" ],
        "CATEGORY" : [ "HEALTH_CHECK" ],
        "BAD_TEST_RESULTS" : [ "1" ],
        "SERVICE_TYPE" : [ "MAPREDUCE" ],
        "EVENTCODE" : [ "EV_SERVICE_HEALTH_CHECK_BAD", "EV_SERVICE_HEALTH
_CHECK_GOOD" ],
        "ALERT_SUMMARY" : [ "The health of service MAPREDUCE-1 has become ba
d." ],
        "CLUSTER_ID" : [ "1" ],
        "SERVICE" : [ "MAPREDUCE-1" ],
        "__uuid" : [ "89521139-0859-4bef-bf65-eb141e63dbba" ],
        "CLUSTER" : [ "Cluster 1" ],
        "CURRENT_COMPLETE_HEALTH_TEST_RESULTS" : [ "{\"content\":\"The hea
lth test result for MAPREDUCE_HA_JOB_TRACKER_HEALTH has become bad: JobTrack
er summary: myCluster.com (Availability: Active, Health: Bad). This health t
est reflects the health of the active JobTracker.\",\"testName\":\"MAPREDUCE
_HA_JOB_TRACKER_HEALTH\",\"eventCode\":\"EV_SERVICE_HEALTH_CHECK_BAD\",\"sev
erity\":\"CRITICAL\"}", "{\"content\":\"The health test result for MAPREDUCE
_TASK_TRACKERS_HEALTHY has become good: Healthy TaskTracker: 3. Concerning T
askTracker: 0. Total TaskTracker: 3. Percent healthy: 100.00%. Percent healt
hy or concerning: 100.00%.\",\"testName\":\"MAPREDUCE_TASK_TRACKERS_HEALTHY\
",\"eventCode\":\"EV_SERVICE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONA
L\"}" ],
        "PREVIOUS_HEALTH_SUMMARY" : [ "GREEN" ],
        "CURRENT_HEALTH_SUMMARY" : [ "RED" ],
        "MONITOR_STARTUP" : [ "false" ],
        "PREVIOUS_COMPLETE_HEALTH_TEST_RESULTS" : [ "{\"content\":\"The hea
lth test result for MAPREDUCE_HA_JOB_TRACKER_HEALTH has become good: JobTrac
ker summary: myCluster.com (Availability: Active, Health: Good)\",\"testName
\":\"MAPREDUCE_HA_JOB_TRACKER_HEALTH\",\"eventCode\":\"EV_SERVICE_HEALTH_CHE
CK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health test
result for MAPREDUCE_TASK_TRACKERS_HEALTHY has become good: Healthy TaskTrac
ker: 3. Concerning TaskTracker: 0. Total TaskTracker: 3. Percent healthy: 10
0.00%. Percent healthy or concerning: 100.00%.\",\"testName\":\"MAPREDUCE_TA
SK_TRACKERS_HEALTHY\",\"eventCode\":\"EV_SERVICE_HEALTH_CHECK_GOOD\",\"sever
ity\":\"INFORMATIONAL\"}" ],
        "SERVICE_DISPLAY_NAME" : [ "MAPREDUCE-1" ]
      }
    }
```

```
  },
  "header" : {
    "type" : "alert",
    "version" : 2
  }
}, {
  "body" : {
    "alert" : {
      "content" : "The health test result for JOB_TRACKER_SCM_HEALTH has bec
ome bad: This role's process exited. This role is supposed to be started.",
      "timestamp" : {
        "iso8601" : "2015-06-11T03:52:56Z",
        "epochMs" : 1433994776083
      },
      "source" : "http://myCluster.com:7180/cmf/eventRedirect/67b4d1c4-791b
-428e-a9ea-8a09d4885f5d",
      "attributes" : {
        "__persist_timestamp" : [ "1433994776173" ],
        "ALERT_SUPPRESSED" : [ "false" ],
        "HEALTH_TEST_NAME" : [ "JOB_TRACKER_SCM_HEALTH" ],
        "SEVERITY" : [ "CRITICAL" ],
        "ROLE" : [ "MAPREDUCE-1-JOBTRACKER-10624c438dee9f17211d3f33fa899957"
 ],
        "HEALTH_TEST_RESULTS" : [ {
          "content" : "The health test result for JOB_TRACKER_SCM_HEALTH has
 become bad: This role's process exited. This role is supposed to be started
.",
          "testName" : "JOB_TRACKER_SCM_HEALTH",
          "eventCode" : "EV_ROLE_HEALTH_CHECK_BAD",
          "severity" : "CRITICAL"
        } ],
        "CLUSTER_DISPLAY_NAME" : [ "Cluster 1" ],
        "HOST_IDS" : [ "75e763c2-8d22-47a1-8c80-501751ae0db7" ],
        "ALERT" : [ "true" ],
        "ROLE_TYPE" : [ "JOBTRACKER" ],
        "CATEGORY" : [ "HEALTH_CHECK" ],
        "BAD_TEST_RESULTS" : [ "1" ],
        "SERVICE_TYPE" : [ "MAPREDUCE" ],
        "EVENTCODE" : [ "EV_ROLE_HEALTH_CHECK_BAD", "EV_ROLE_HEALTH_CHECK_
GOOD", "EV_ROLE_HEALTH_CHECK_DISABLED" ],
        "ALERT_SUMMARY" : [ "The health of role jobtracker (nightly-1) has
 become bad." ],
        "CLUSTER_ID" : [ "1" ],
        "SERVICE" : [ "MAPREDUCE-1" ],
        "__uuid" : [ "67b4d1c4-791b-428e-a9ea-8a09d4885f5d" ],
        "CLUSTER" : [ "Cluster 1" ],
        "CURRENT_COMPLETE_HEALTH_TEST_RESULTS" : [ "{\"content\":\"The healt
h test result for JOB_TRACKER_SCM_HEALTH has become bad: This role's process
 exited. This role is supposed to be started.\",\"testName\":\"JOB_TRACKER_S
CM_HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_BAD\",\"severity\":\"CRITIC
AL\"}", "{\"content\":\"The health test result for JOB_TRACKER_UNEXPECTED_EX
ITS has become good: This role encountered 0 unexpected exit(s) in the previ
ous 5 minute(s).\",\"testName\":\"JOB_TRACKER_UNEXPECTED_EXITS\",\"eventCode
\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"conte
nt\":\"The health test result for JOB_TRACKER_FILE_DESCRIPTOR has become goo
d: Open file descriptors: 244. File descriptor limit: 32,768. Percentage in
use: 0.74%.\",\"testName\":\"JOB_TRACKER_FILE_DESCRIPTOR\",\"eventCode\":\"E
V_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\
"The health test result for JOB_TRACKER_SWAP_MEMORY_USAGE has become good: 0
 B of swap memory is being used by this role's process.\",\"testName\":\"JOB
_TRACKER_SWAP_MEMORY_USAGE\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"s
everity\":\"INFORMATIONAL\"}", "{\"content\":\"The health test result for JO
B_TRACKER_LOG_DIRECTORY_FREE_SPACE has become good: This role's Log Director
y (/var/log/hadoop-0.20-mapreduce) is on a filesystem with more than 20.00%
```

```
of its space free.\",\"testName\":\"JOB_TRACKER_LOG_DIRECTORY_FREE_SPACE\",\
"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}",
 "{\"content\":\"The health test result for JOB_TRACKER_HOST_HEALTH has beco
me good: The health of this role's host is good.\",\"testName\":\"JOB_TRACKE
R_HOST_HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"I
NFORMATIONAL\"}", "{\"content\":\"The health test result for JOB_TRACKER_WEB
_METRIC_COLLECTION has become good: The web server of this role is respondin
g with metrics. The most recent collection took 49 millisecond(s).\",\"testN
ame\":\"JOB_TRACKER_WEB_METRIC_COLLECTION\",\"eventCode\":\"EV_ROLE_HEALTH_C
HECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health tes
t result for JOB_TRACKER_GC_DURATION has become good: Average time spent in
garbage collection was 0 second(s) (0.00%) per minute over the previous 5 mi
nute(s).\",\"testName\":\"JOB_TRACKER_GC_DURATION\",\"eventCode\":\"EV_ROLE_
HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The he
alth test result for JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE has become d
isabled: Test disabled because role is not configured to dump heap when out
of memory. Test of whether this role's heap dump directory has enough free s
pace.\",\"testName\":\"JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE\",\"eventC
ode\":\"EV_ROLE_HEALTH_CHECK_DISABLED\",\"severity\":\"INFORMATIONAL\"}" ],
        "CURRENT_HEALTH_SUMMARY" : [ "RED" ],
        "PREVIOUS_HEALTH_SUMMARY" : [ "GREEN" ],
        "MONITOR_STARTUP" : [ "false" ],
        "ROLE_DISPLAY_NAME" : [ "jobtracker (nightly-1)" ],
        "PREVIOUS_COMPLETE_HEALTH_TEST_RESULTS" : [ "{\"content\":\"The heal
th test result for JOB_TRACKER_SCM_HEALTH has become good: This role's statu
s is as expected. The role is started.\",\"testName\":\"JOB_TRACKER_SCM_HEAL
TH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONA
L\"}", "{\"content\":\"The health test result for JOB_TRACKER_UNEXPECTED_EXI
TS has become good: This role encountered 0 unexpected exit(s) in the previo
us 5 minute(s).\",\"testName\":\"JOB_TRACKER_UNEXPECTED_EXITS\",\"eventCode\
":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"conten
t\":\"The health test result for JOB_TRACKER_FILE_DESCRIPTOR has become good
: Open file descriptors: 244. File descriptor limit: 32,768. Percentage in u
se: 0.74%.\",\"testName\":\"JOB_TRACKER_FILE_DESCRIPTOR\",\"eventCode\":\"EV
_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"
The health test result for JOB_TRACKER_SWAP_MEMORY_USAGE has become good: 0
B of swap memory is being used by this role's process.\",\"testName\":\"JOB_
TRACKER_SWAP_MEMORY_USAGE\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"se
verity\":\"INFORMATIONAL\"}", "{\"content\":\"The health test result for JOB
_TRACKER_LOG_DIRECTORY_FREE_SPACE has become good: This role's Log Directory
 (/var/log/hadoop-0.20-mapreduce) is on a filesystem with more than 20.00% o
f its space free.\",\"testName\":\"JOB_TRACKER_LOG_DIRECTORY_FREE_SPACE\",\"
eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}",
"{\"content\":\"The health test result for JOB_TRACKER_HOST_HEALTH has becom
e good: The health of this role's host is good.\",\"testName\":\"JOB_TRACKER
_HOST_HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"IN
FORMATIONAL\"}", "{\"content\":\"The health test result for JOB_TRACKER_WEB_
METRIC_COLLECTION has become good: The web server of this role is responding
 with metrics. The most recent collection took 49 millisecond(s).\",\"testNa
me\":\"JOB_TRACKER_WEB_METRIC_COLLECTION\",\"eventCode\":\"EV_ROLE_HEALTH_CH
ECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health test
 result for JOB_TRACKER_GC_DURATION has become good: Average time spent in g
arbage collection was 0 second(s) (0.00%) per minute over the previous 5 min
ute(s).\",\"testName\":\"JOB_TRACKER_GC_DURATION\",\"eventCode\":\"EV_ROLE_H
EALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The hea
lth test result for JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE has become di
sabled: Test disabled because role is not configured to dump heap when out o
f memory. Test of whether this role's heap dump directory has enough free sp
ace.\",\"testName\":\"JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE\",\"eventCo
de\":\"EV_ROLE_HEALTH_CHECK_DISABLED\",\"severity\":\"INFORMATIONAL\"}" ],
        "SERVICE_DISPLAY_NAME" : [ "MAPREDUCE-1" ],
        "HOSTS" : [ "myCluster.com" ]
      }
    }
```

```
  },
  "header" : {
    "type" : "alert",
    "version" : 2
  }
}, {
  "body" : {
    "alert" : {
      "content" : "The health test result for JOB_TRACKER_UNEXPECTED_EXITS
has become bad: This role encountered 1 unexpected exit(s) in the previous 5
 minute(s).This included 1 exit(s) due to OutOfMemory errors. Critical thres
hold: any.",
      "timestamp" : {
        "iso8601" : "2015-06-11T03:53:41Z",
        "epochMs" : 1433994821940
      },
      "source" : "http://myCluster.com:7180/cmf/eventRedirect/b8c4468d-08c
2-4b5b-9bda-2bef892ba3f5",
      "attributes" : {
        "__persist_timestamp" : [ "1433994822027" ],
        "ALERT_SUPPRESSED" : [ "false" ],
        "HEALTH_TEST_NAME" : [ "JOB_TRACKER_UNEXPECTED_EXITS" ],
        "SEVERITY" : [ "CRITICAL" ],
        "ROLE" : [ "MAPREDUCE-1-JOBTRACKER-10624c438dee9f17211d3f33fa8999
57" ],
        "HEALTH_TEST_RESULTS" : [ {
          "content" : "The health test result for JOB_TRACKER_UNEXPECTED_
EXITS has become bad: This role encountered 1 unexpected exit(s) in the prev
ious 5 minute(s).This included 1 exit(s) due to OutOfMemory errors. Critical
 threshold: any.",
          "testName" : "JOB_TRACKER_UNEXPECTED_EXITS",
          "eventCode" : "EV_ROLE_HEALTH_CHECK_BAD",
          "severity" : "CRITICAL"
        } ],
        "CLUSTER_DISPLAY_NAME" : [ "Cluster 1" ],
        "HOST_IDS" : [ "75e763c2-8d22-47a1-8c80-501751ae0db7" ],
        "ALERT" : [ "true" ],
        "ROLE_TYPE" : [ "JOBTRACKER" ],
        "CATEGORY" : [ "HEALTH_CHECK" ],
        "BAD_TEST_RESULTS" : [ "1" ],
        "SERVICE_TYPE" : [ "MAPREDUCE" ],
        "EVENTCODE" : [ "EV_ROLE_HEALTH_CHECK_BAD", "EV_ROLE_HEALTH_CHECK
_GOOD", "EV_ROLE_HEALTH_CHECK_DISABLED" ],
        "ALERT_SUMMARY" : [ "The health of role jobtracker (nightly-1) has
 become bad." ],
        "CLUSTER_ID" : [ "1" ],
        "SERVICE" : [ "MAPREDUCE-1" ],
        "__uuid" : [ "b8c4468d-08c2-4b5b-9bda-2bef892ba3f5" ],
        "CLUSTER" : [ "Cluster 1" ],
        "CURRENT_COMPLETE_HEALTH_TEST_RESULTS" : [ "{\"content\":\"The heal
th test result for JOB_TRACKER_SCM_HEALTH has become bad: This role's proces
s exited. This role is supposed to be started.\",\"testName\":\"JOB_TRACKER_
SCM_HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_BAD\",\"severity\":\"CRITI
CAL\"}", "{\"content\":\"The health test result for JOB_TRACKER_UNEXPECTED_E
XITS has become bad: This role encountered 1 unexpected exit(s) in the previ
ous 5 minute(s).This included 1 exit(s) due to OutOfMemory errors. Critical
threshold: any.\",\"testName\":\"JOB_TRACKER_UNEXPECTED_EXITS\",\"eventCode\
":\"EV_ROLE_HEALTH_CHECK_BAD\",\"severity\":\"CRITICAL\"}", "{\"content\":\"
The health test result for JOB_TRACKER_FILE_DESCRIPTOR has become good: Open
 file descriptors: 244. File descriptor limit: 32,768. Percentage in use: 0.
74%.\",\"testName\":\"JOB_TRACKER_FILE_DESCRIPTOR\",\"eventCode\":\"EV_ROLE_
HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The he
alth test result for JOB_TRACKER_SWAP_MEMORY_USAGE has become good: 0 B of s
wap memory is being used by this role's process.\",\"testName\":\"JOB_TRACKE
```

```
R_SWAP_MEMORY_USAGE\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity
\":\"INFORMATIONAL\"}", "{\"content\":\"The health test result for JOB_TRACK
ER_LOG_DIRECTORY_FREE_SPACE has become good: This role's Log Directory (/var
/log/hadoop-0.20-mapreduce) is on a filesystem with more than 20.00% of its
space free.\",\"testName\":\"JOB_TRACKER_LOG_DIRECTORY_FREE_SPACE\",\"eventC
ode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"co
ntent\":\"The health test result for JOB_TRACKER_HOST_HEALTH has become good
: The health of this role's host is good.\",\"testName\":\"JOB_TRACKER_HOST_
HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMAT
IONAL\"}", "{\"content\":\"The health test result for JOB_TRACKER_WEB_METRIC
_COLLECTION has become good: The web server of this role is responding with
metrics. The most recent collection took 49 millisecond(s).\",\"testName\":\
"JOB_TRACKER_WEB_METRIC_COLLECTION\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GO
OD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health test resul
t for JOB_TRACKER_GC_DURATION has become good: Average time spent in garbage
 collection was 0 second(s) (0.00%) per minute over the previous 5 minute(s)
.\",\"testName\":\"JOB_TRACKER_GC_DURATION\",\"eventCode\":\"EV_ROLE_HEALTH_
CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health te
st result for JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE has become disabled
: Test disabled because role is not configured to dump heap when out of memo
ry. Test of whether this role's heap dump directory has enough free space.\"
,\"testName\":\"JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE\",\"eventCode\":\
"EV_ROLE_HEALTH_CHECK_DISABLED\",\"severity\":\"INFORMATIONAL\"}" ],
        "CURRENT_HEALTH_SUMMARY" : [ "RED" ],
        "PREVIOUS_HEALTH_SUMMARY" : [ "RED" ],
        "MONITOR_STARTUP" : [ "false" ],
        "ROLE_DISPLAY_NAME" : [ "jobtracker (nightly-1)" ],
        "PREVIOUS_COMPLETE_HEALTH_TEST_RESULTS" : [ "{\"content\":\"The h
ealth test result for JOB_TRACKER_SCM_HEALTH has become bad: This role's pro
cess exited. This role is supposed to be started.\",\"testName\":\"JOB_TRACK
ER_SCM_HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_BAD\",\"severity\":\"CR
ITICAL\"}", "{\"content\":\"The health test result for JOB_TRACKER_UNEXPECTE
D_EXITS has become good: This role encountered 0 unexpected exit(s) in the p
revious 5 minute(s).\",\"testName\":\"JOB_TRACKER_UNEXPECTED_EXITS\",\"event
Code\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"c
ontent\":\"The health test result for JOB_TRACKER_FILE_DESCRIPTOR has become
 good: Open file descriptors: 244. File descriptor limit: 32,768. Percentage
 in use: 0.74%.\",\"testName\":\"JOB_TRACKER_FILE_DESCRIPTOR\",\"eventCode\"
:\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content
\":\"The health test result for JOB_TRACKER_SWAP_MEMORY_USAGE has become goo
d: 0 B of swap memory is being used by this role's process.\",\"testName\":\
"JOB_TRACKER_SWAP_MEMORY_USAGE\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\"
,\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health test result fo
r JOB_TRACKER_LOG_DIRECTORY_FREE_SPACE has become good: This role's Log Dire
ctory (/var/log/hadoop-0.20-mapreduce) is on a filesystem with more than 20.
00% of its space free.\",\"testName\":\"JOB_TRACKER_LOG_DIRECTORY_FREE_SPACE
\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\
"}", "{\"content\":\"The health test result for JOB_TRACKER_HOST_HEALTH has
become good: The health of this role's host is good.\",\"testName\":\"JOB_TR
ACKER_HOST_HEALTH\",\"eventCode\":\"EV_ROLE_HEALTH_CHECK_GOOD\",\"severity\"
:\"INFORMATIONAL\"}", "{\"content\":\"The health test result for JOB_TRACKER
_WEB_METRIC_COLLECTION has become good: The web server of this role is respo
nding with metrics. The most recent collection took 49 millisecond(s).\",\"t
estName\":\"JOB_TRACKER_WEB_METRIC_COLLECTION\",\"eventCode\":\"EV_ROLE_HEAL
TH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"The health
 test result for JOB_TRACKER_GC_DURATION has become good: Average time spent
 in garbage collection was 0 second(s) (0.00%) per minute over the previous
5 minute(s).\",\"testName\":\"JOB_TRACKER_GC_DURATION\",\"eventCode\":\"EV_R
OLE_HEALTH_CHECK_GOOD\",\"severity\":\"INFORMATIONAL\"}", "{\"content\":\"Th
e health test result for JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE has beco
me disabled: Test disabled because role is not configured to dump heap when
out of memory. Test of whether this role's heap dump directory has enough fr
ee space.\",\"testName\":\"JOB_TRACKER_HEAP_DUMP_DIRECTORY_FREE_SPACE\",\"ev
```

```
entCode\":\"EV_ROLE_HEALTH_CHECK_DISABLED\",\"severity\":\"INFORMATIONAL\"}"
 ],
        "SERVICE_DISPLAY_NAME" : [ "MAPREDUCE-1" ],
        "HOSTS" : [ "myCluster.com" ]
      }
    }
  },
  "header" : {
    "type" : "alert",
    "version" : 2
  }
} ]
```

Related Information

Managing Alerts

Configuring Alert Email Delivery

Configuring Alert SNMP Delivery

Sample Custom Alert Script

You can use this sample script to respond to an alert by sending emails based on the content of the alert.

The Sample Custom Alert Script provided below provides an example of how custom alert scripts can be implemented. This example uses a Python script, which includes multiple routing rules called Alert Templates. Once the script is called, it parses the alert JSON file using a path provided by the Alert Publisher. Based on the Alert Templates, it sends out alerts to different email addresses based on the content of the alert.

The script requires an SMTP server (defined by its host and port) and an email address to send out the emails. You need to set these parameters in the script.

There are multiple Alert Templates in the sample script as examples. Pick the ones you need and delete the rest, or create new templates.

Python should be always available on your hosts because hosts managed by Cloudera Manager require Python for the Cloudera Manager Agent to run. Because the Alert Publisher is not able to execute python scripts, a small Bash script is needed to execute the main Python script.

Bash Script

This Bash script is required to execute the Python script.

```
# © 2022 by Cloudera, Inc. All rights reserved.
# Scripts and sample code are licensed under the Apache License,
# Version 2.0
#!/bin/bash

python /bin/script_test/main.py $1
```

Python script

Setting up the Python script

To use the Python script, you will need to provide the values described below.

- SMTP hostname:

  ```
  host = "mailserver.mycompany.com"
  ```

- SMTP port:

  ```
  port = 25
  ```

- Sender email address:

```
sender_address = "noreply@mycompany.com"
```

- Values for the AlertTemplate(email, attribute_key, attribute_value, remove_after) method:
  - Email addresses:

    ```
    AlertTemplate("admin@mycompany.com", "SEVERITY", "CRITICAL", False)
    ```

    Specify two or more recipients using this format:

    ```
    AlertTemplate("admin@mycompany.com, admin2@mycompany.com", "SEVERITY",
     "CRITICAL", False)
    ```

  - An attribute_key and attribute_value containing the rules that your alert should match for the email to be sent:

    ```
     AlertTemplate("admin@mycompany.com", "SEVERITY", "CRITICAL", False),
    ```

    To use Regular Expression matching for your alerts, use “REGEX” as the attribute_key:

    ```
    AlertTemplate("admin@mycompany.com", "REGEX", '.*content": "This is a
     test alert that was generated.*', False)
    ```

    A None attribute_key and attribute_value pair indicates a default template, which should be applied to all alerts. You can use this when there's no matching AlertTemplate, but you still do not want to lose the Alert.
  - remove_after:

    Set this argument to True if you want the alert to not be sent, even if further Alert Templates match the alert:

    ```
     AlertTemplate("admin@mycompany.com", "SEVERITY", "CRITICAL", True),
    ```

Sample Python script

You can copy this script and modify it for your deployment.

```
# -*- coding: utf-8 -*-
# © 2022 by Cloudera, Inc. All rights reserved.
# Scripts and sample code are licensed under the Apache License,
# Version 2.0

#!/usr/bin/env python2
from __future__ import print_function

import sys
import smtplib
import json
import re

from email.mime.text import MIMEText
from email.mime.multipart import MIMEMultipart

class AlertReader(object):
    def __init__(self, file_path):
        self.alerts = []
        self.file_path = file_path

    def read_alert(self):
        alert_file = open(self.file_path, "r")
        data = json.load(alert_file)
        for alert in data:
            self.alerts.append(alert["body"]["alert"])
        print("Done reading " + self.file_path)
```

```
    def print_alerts(self):
        print(self.alerts)

def alert_as_string(alert):
    return json.dumps(alert)

class AlertTemplate(object):
    def __init__(self, email, attribute_key, attribute_value, remove_after):
        self.email = email
        self.attribute_key = attribute_key
        self.attribute_value = attribute_value
        self.remove_after = remove_after
        """
        A None attribute_key and attribute_value pair
        indicates a default template, which should be
        applied to all alerts. You can use this when
        there's no matching AlertTemplate, but you still
        don't want to lose the Alert.

        Put these AlertTemplates to the end of your
        AlertTemplate list to catch all remaining alerts.
        But if only one of the attribute_key or attribute_value is None,
        that is an invalid state.
        """
        if (self.attribute_key is None) != (self.attribute_value is None):
            raise ValueError("AlertTemplate is in invalid state, "
                             "one of attribute_key or attribute_value is N
one.")

    def does_apply(self, alert):
        if self.attribute_key == "REGEX":
            alert_string = alert_as_string(alert)
            search_result = re.search(self.attribute_value, alert_string)
            return search_result is not None
        if self.attribute_key == "CONTENT":
            if self.attribute_value in alert["content"]:
                return True
            return False
        if self.attribute_key is None and self.attribute_value is None:
            return True
        try:
            if self.attribute_value in alert["attributes"][self.attribute
_key]:
                return True
        except KeyError:
            print("There's no match for attribute_key: " + self.attribute_ke
y)
        return False
    def to_string(self):
        return "AlertTemplate[" + \
               str(self.attribute_key) + ": " + \
               str(self.attribute_value) + ", " + \
               str(self.email) + "]"


class EmailSender(object):
    def __init__(self, smtp_host, smtp_port, sender):
        self.server = smtplib.SMTP(smtp_host, smtp_port)
        self.sender = sender

    def send_email_to_recipients(self, alerts, alert_templates):
        print("Processing alerts to send emails.")
        for alert in alerts:
            for alert_template in alert_templates:
```

```
                if alert_template.does_apply(alert):
                    """
                    With a template with no email address, you usually want
to exclude
                    a type of alert from the processed alert list. Use None
 to email address
                    with remove_after = True if you want an alert to be remo
ved from the
                    list of alerts sent.
                    These templates should be in the beginning of the temp
late list.
                    """
                    if alert_template.email is not None:
                        self.send_email(alert_template, alert)
                        print(create_message_text(alert_template, alert))
                    print("Sending alert based on " + alert_template.to_str
ing())
                    if alert_template.remove_after:
                        break
        print("All alerts are processed.")
    def send_email(self, alert_template, alert):
        message = MIMEMultipart("alternative")
        message["Subject"] = "Cloudera Alert"
        message["From"] = self.sender
        message["To"] = alert_template.email
        message.attach(MIMEText(create_message_text(alert_template, alert),
 "plain"))
        self.server.sendmail(self.sender, alert_template.email.split(","), m
essage.as_string())

def create_message_text(alert_template, alert):
    out = alert_template.to_string()
    out += " firing for: \n"
    out += alert_as_string(alert)
    return out


if __name__ == '__main__':
    """
    SMTP server parameters:
      - email where the emails sent from
      - file name, where the alerts are coming from. This is provided by
the script interface.
    Alert templates:
      - email address to send email to when the alert contains the specif
ied key with the specified value.
      - Set remove_after to True if you don't want the alert sent to other
 recipients,
        even if the template is matching. Because of this, it is required t
hat the recipients
        list is in a specific order.
    """
    host = "mailserver.mycompany.com"  # SMTP server host name
    port = 25  # SMTP Server port
    sender_address = "noreply@"MyCompanyMailServer"  # Email address to send
 emails from
    # The file's name which contains the alerts is provided by the script
framework
    if len(sys.argv) < 2:
        print("Please provide a file path for alerts. Usually this comes fro
m the script framework.")
        exit(2)
    file_name = sys.argv[1]
    alert_reader = AlertReader(file_name)
```

```
    recipients = []
    try:
        recipients = [AlertTemplate(None, None, None, True),  # This templ
ate matches every alert
                      # This template matches to all alerts which matches to
 the given regex pattern
                      AlertTemplate("admin@mycompany.com", "REGEX", '.*cont
ent": "This is a test alert that was generated.*', False),
                      # This template matches only for issues with critical
 severity
                      AlertTemplate("admin@mycompany.com", "SEVERITY", "CRIT
ICAL", False),
                      # This template matches only for Kafka related issues
                      AlertTemplate("kafka_admin@mycompany.com", "SERVICE
_TYPE", "KAFKA", True),
                      # This template matches only for Kafka Broker related
issues
                      AlertTemplate("user1@gmail.com", "ROLE_TYPE", "KAFK
A_BROKER", True),
                      # This template matches only for a specific host's a
lerts
                      AlertTemplate("admin2@mycompany.com", "HOSTS", "myho
st", True),
                      # This template matches for a specific cluster's alert
s
                      AlertTemplate("admin3@mycompany.com", "CLUSTER", "Cl
uster 1", True),
                      # This template matches for a specific health test me
ssage
                      AlertTemplate("admin@mycompany.com", "CONTENT", "This
 is a test alert that was generated by user request from "
                                                          "Cloudera Manager", Tru
e)]
    except ValueError as error:
        print(error)
        exit(2)

    alert_reader.read_alert()
    alert_reader.print_alerts()
    email_sender = EmailSender(host, port, sender_address)
    email_sender.send_email_to_recipients(alert_reader.alerts, recipients)
```

Alert email routing

In multi-tenant or large-cluster environments, routing specific alerts to the right person can ensure a healthy deployment. With Alert email routing, you can define the alert rules in Cloudera Manager to send email alerts to selected email addresses and avoid sending all alert notifications to one or more email addresses.

Alert rules are defined based on user profiles, clusters, hosts, services, roles, host groups, service groups, service types, role groups, role types, the severity of the alert, priority, and specific health tests. A host or service or role group is a collection of hosts or services or roles respectively to which the same alert rule applies.

Some of the use cases for Alert email routing are as follows:

- You can create a role group that includes all the secondary roles such as Data Nodes. When configuring an alert rule for this role group, you can include the email addresses of the experts of these roles so that they receive email notifications only about the critical alerts of the secondary roles. You can even silence the alerts of the non-critical roles of this group and react more quickly to the critical alerts.
- You can create a service group that includes all the storage services such as HDFS and HBase. When configuring an alert rule for this service group, you can include the email addresses of your storage experts who should receive email notifications only about the alerts of the storage services.

- You can create a host group that includes the hosts that are in the same rack such as Rack1. When configuring an alert rule for this host group, you can include the email addresses of the cluster administrator of these hosts who should receive email notifications only about the alerts of Rack1.
- If the cluster operator does not need email notifications about a health test that is often in a bad state and is not critical for the operation, then the operator can define an alert rule to silence the alerts of that health test.

You can configure the following parameters for Alert email routing:

- User profiles - A user profile defines a user or a user group who should receive email alerts.
- Host groups - A host group defines a subset of hosts. You can use this group as one logical unit for email alerting.
- Service grouos - A service group defines a subset of services. You can use this group as one logical unit for email alerting.
- Role groups - A role group defines a subset of roles. You can use this group as one logical unit for email alerting.
- Alert rules - An alert rule ensures certain hosts or services or roles or health test alerts are sent to only selected email addresses.

Configuring user profiles for email alerts

You can configure a user profile using Cloudera Manager. A user profile defines a user or a user group who should receive email alerts. You can create any number of user profiles.

Before you begin

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Log into Cloudera Manager as an Administrator.
2. Go to Clusters Cloudera Management Service Configuration .
3. Select Alert Publisher from the SCOPE drop-down menu and select Main from the CATEGORY drop-down menu.
4. Locate the User profiles for email alerting property or search for it by typing its name in the Search box.
5. Configure the User profiles for email alerting property. Click ⊕ Add Another to add a new user profile.
6. Enter a unique name for the user profile in the Key field.
7. Enter the comma-separated list of email addresses in the Value field.
8. Click Save Changes at the bottom of the page to save your settings.
9. Restart the Alert Publisher role. For more information, see *Starting, Stopping, and Restarting Role Instances*.

Related Information

Starting, Stopping, and Restarting Role Instances

Configuring host groups for email alerts

You can configure a host group using Cloudera Manager. A host group defines a subset of hosts and alerts generated by hosts are sent to the cluster administrator who manages these hosts. You can create any number of host groups.

Before you begin

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Log into Cloudera Manager as an Administrator.
2. Go to Clusters Cloudera Management Service Configuration .

3. Select Alert Publisher from the SCOPE drop-down menu and select Main from the CATEGORY drop-down menu.
4. Locate the Host groups for email alerting property or search for it by typing its name in the Search box.
5. Configure the Host groups for email alerting property. Click ⊕ Add Another to add a new host group.
6. Enter a unique name for the host group in the Key field.
7. Enter the comma-separated list of host names in the Value field.

   Important: If you do not enter a valid host name, then you might encounter a validation error. In case of an empty list, all managed hosts are considered as a host group.

8. Click Save Changes at the bottom of the page to save your settings.
9. Restart the Alert Publisher role. For more information, see *Starting, Stopping, and Restarting Role Instances*.

Related Information

Starting, Stopping, and Restarting Role Instances

Configuring service groups for email alerts

You can configure a service group using Cloudera Manager. A service group defines a subset of services for which you need email alerts. You can create any number of service groups.

Before you begin

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Log into Cloudera Manager as an Administrator.
2. Go to Clusters Cloudera Management Service Configuration .
3. Select Alert Publisher from the SCOPE drop-down menu and select Main from the CATEGORY drop-down menu.
4. Locate the Service groups for email alerting property or search for it by typing its name in the Search box.
5. Configure the Service groups for email alerting property. Click ⊕ Add Another to add a new service group.
6. Enter a unique name for the service group in the Key field.
7. Enter the comma-separated list of service names in the Value field.

   Important: If you do not enter a valid service name, then you might encounter a validation error. In case of an empty list, all managed services are considered as a service group.

8. Click Save Changes at the bottom of the page to save your settings.
9. Restart the Alert Publisher role. For more information, see *Starting, Stopping, and Restarting Role Instances*.

Related Information

Starting, Stopping, and Restarting Role Instances

Configuring role groups for email alerts

You can configure a role group using Cloudera Manager. A role group defines a subset of roles and their alerts are sent to the experts who manage these roles. You can create any number of role groups.

Before you begin

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Log into Cloudera Manager as an Administrator.
2. Go to Clusters Cloudera Management Service Configuration .
3. Select Alert Publisher from the SCOPE drop-down menu and select Main from the CATEGORY drop-down menu.
4. Locate the Role groups for email alerting property or search for it by typing its name in the Search box.
5. Configure the Role groups for email alerting property. Click ⊕ Add Another to add a new role group.
6. Enter a unique name for the role group in the Key field.
7. Enter the comma-separated list of role names in the Value field.

   Important: If you do not enter a valid role name, then you might encounter a validation error. In case of an empty list, all managed roles are considered as a role group.

8. Click Save Changes at the bottom of the page to save your settings.
9. Restart the Alert Publisher role. For more information, see *Starting, Stopping, and Restarting Role Instances*.

Related Information

Starting, Stopping, and Restarting Role Instances

Configuring alert rules for email alerts

You can configure an alert rule using Cloudera Manager. An alert rule ensures certain hosts or services or roles or health test alerts are sent to only selected email addresses. You can create any number of alert rules.

Before you begin

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Log into Cloudera Manager as an Administrator.
2. Go to Clusters Cloudera Management Service Configuration .
3. Select Alert Publisher from the SCOPE drop-down menu and select Main from the CATEGORY drop-down menu.
4. Locate the Alert rules for email alerting property or search for it by typing its name in the Search box.
5. Configure the Alert rules for email alerting property. Click ⊕ Add Another to add a new alert rule.
6. Select the severity level of the alert from the list of severity levels (Warning, Critical, or Both).

   Important:

   For Health Alert Threshold, you must select the Concerning parameter to ensure the email alerts are sent with the warning severity.

   To select this parameter, from the Cloudera Manager home page, navigate to Clusters Cloudera Management Console Configuration and in the search bar, type Health Alert Threshold.

7. Enter the priority value for the alert rule in the Priority field.

   Important: The alert rule with the lowest priority value is processed first. The priority value must be a non-negative integer.

8. Specify the values in the User Profiles, Clusters, Hosts, Services, Roles, Host Groups, Service Groups, Service Types, Role Groups, Role Types, and Health Tests fields as needed. You must enter the values atleast in one of these fields to create a matching alert rule.

   Important: If the User Profiles field remains empty, then no users receive email notifications about the alerts. You can enter any number of predefined user profiles in the User Profiles field. If there is more than one user profile, then email alerts are sent to more addresses.

   You must set the Alert: Mail From Address parameter according to the mail server hostname. Otherwise, the emails may not sent to an external email service.
9. Click Save Changes at the bottom of the page to save your settings.
10. Restart the Alert Publisher role. For more information, see *Starting, Stopping, and Restarting Role Instances*.

Related Information

Starting, Stopping, and Restarting Role Instances

Enabling Configuration Change Alerts

You can set configuration change alerts to be service-wide, or on specific roles for the service.

About this task

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Click a service, role, or host.
2. Click the Configuration tab.
3. Click the Monitoring category.
4. Check the Enable Configuration Change Alerts checkbox.
5. Enter a Reason for change, and then click Save Changes to commit the changes.
6. Click the Cloudera Manager logo to return to the Home page.
7. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Enabling HBase Alerts

You can enable region or Hbck alerts for the HBase service.

About this task

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Go to the HBase service.
2. Click the Configuration tab.
3. Select Scope *HBase service_name* (Service-Wide) .
4. Click the Monitoring category.

5. Set one of the region or Hbck alerts:
   - Hbck Region Error Count
   - Hbck Error Count
   - Hbck Alert Error Codes
   - Hbck Slow Run
   - Region Health Canary Slow Run
   - Canary Unhealthy Region Count
   - Canary Unhealthy Region Percentage
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Enabling Health Alerts

You can enable alerts when the health of a role or service crosses a threshold.

About this task

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Select Clusters*cluster_name*service_name or open the page for a role.
2. Click the Configuration tab.
3. Select Scope*role_name* or *service_name* (Service-Wide).
4. Click the Monitoring category.
5. Check the Enable Health Alerts for this Role or Enable Service Level Health Alerts checkbox, depending on whether you are configuring a role or a service.
6. Enter a Reason for change, and then click Save Changes to commit the changes.
7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Modifying the Health Threshold

You can configure the threshold for when a health alert is raised.

About this task

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Select AdministrationAlerts.
2. Click 🔧 to the right of Health Alert Threshold.
3. Select ScopeEvent Server.
4. Click the Main category.
5. Select the Bad or Concerning option.
6. Enter a Reason for change, and then click Save Changes to commit the changes.

7. Click the Cloudera Manager logo to return to the Home page.
8. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Alerts Transitioning Out of Alerting Health Threshold

You can configure an alert when a service or role instance transitions from an alerting to a non-alerting health threshold.

About this task

Minimum Required Role: Limited Cluster Administrator (also provided by Full Administrator and Cluster Administrator)

Procedure

1. Select AdministrationAlerts.
2. Click 🔧 to the right of Alert on Transitions out of Alerting Health.
3. Select Scope*role_name* or *service_name* (Service-Wide).
4. In the category Event Server Default Group, check the Alert on Transitions out of Alerting Health checkbox.
5. Enter a Reason for change, and then click Save Changes to commit the changes.
6. Click the Cloudera Manager logo to return to the Home page.
7. Click the ⏻ icon that is next to any stale services to invoke the cluster restart wizard.

Configuring Log Alerts

You can configure an alert when a daemon emits a log message that matches a specified regular expression. For more information, see *Configuring Log Alerts*.

Related Information

Configuring Log Alerts

Configuring Alert Delivery

You can configure alerts to be delivered by email or sent as SNMP traps.

If you choose email delivery, you can add to or modify the list of alert recipient email addresses. You can also send a test alert email.

Note: If alerting is enabled for events, you can search for and view alerts in the Events tab, even if you do not have email notification configured.

Related Information

Managing Alerts

Triggers

A *trigger* is a statement that specifies an action to be taken when one or more specified conditions are met for a service, role, role configuration group, or host. The conditions are expressed as a tsquery statement, and the action to be taken is to change the health for the service, role, role configuration group, or host to either Concerning (yellow) or Bad (red).

Triggers can be created for services, roles, role configuration groups, or hosts. Create a trigger by doing one of the following:

- Directly editing the configuration for the service, role (or role configuration group), or host configuration.
- Clicking Create Trigger on the drop-down menu for most charts. Note that the Create Trigger command is not available on the drop-down menu for charts where no context (role, service, and so on) is defined, such as on the HomeStatus tab.

  Important: Because triggers are a new and evolving feature, backward compatibility between releases is not guaranteed at this time.

- Use the Create Trigger expression builder.

The Structure of Triggers

A trigger is defined by a JSON formatted string that includes four parts:

- Name
- Expression
- Stream threshold
- Whether or not the trigger should be enabled

Each of the four parts of a trigger is described in the following sections.

Name (required)

A trigger's name must be unique in the context for which the trigger is defined. That is, there cannot be two triggers for the same service or role with the same name. Different services or different roles can have triggers with the same name.

Expression (required)

A trigger expression takes the form:

IF (*CONDITIONS*) DO *HEALTH_ACTION*

When the conditions of the trigger are met, the trigger is considered to be firing. A condition is any valid tsquery statement. In most cases conditions employ stream filters to filter out streams below or above a certain threshold. For example, the following tsquery can be used to retrieve the streams for DataNodes with more than 500 open file descriptors:

```
SELECT fd_open WHERE roleType=DataNode AND last(fd_open) > 500
```

The stream filter used here, last(fd_open) > 50, is composed of four parts:

- A scalar producing function "last" that takes a stream and returns its last data point
- A metric to operate on

  Note: For more information about metric expression functions, see Metric Expression Functions.

- A comparator
- A scalar value

Other scalar producing functions are available, like min or max, and they can be combined to create arbitrarily complex expressions:

```
last(moving_avg(fd_open)) >= 500
```

See the tsquery documentation for more details.

Conditions can be combined using the logical operators AND and OR. For example, here is a trigger expression with two conditions:

```
IF ((SELECT fd_open WHERE roleType=DataNode AND last(fd_open) > 500) OR (SEL
ECT fd_open WHERE roleType=NameNode AND last(fd_open) > 500)) DO health:bad
```

A condition is met if it returns more than the number of streams specified by the streamThreshold (see below). A trigger fires if the logical evaluation of all of its conditions results in a met condition. When a trigger fires, two actions can be taken: health:concerning or health:bad. These actions change the health of the entity on which the trigger is defined.

Stream Threshold (optional)

The stream threshold determines the number of streams that need to be returned by the tsquery before the condition is met. The default is 0; that is, if the tsquery returns any results the condition will be met. For example if the stream threshold is set to 10 and the condition is SELECT fd_open WHERE roleType=DataNode AND last(fd_open) > 500 the condition will be considered met only if there are at least 10 DataNodes that have more than 500 file descriptor opened, so at least 10 streams were returned by the tsquery.

Enabled (optional)

Whether the trigger is enabled. The default is true, (enabled).

Trigger Example

The following is a JSON formatted trigger that fires if there are more than 10 DataNodes with more than 500 file descriptors opened:

```
[{"triggerName": "sample-trigger", "triggerExpression": "IF (SELECT fd_open
WHERE roleType = DataNode and last(fd_open) > 500) DO health:bad", "streamTh
reshold": 10, "enabled": "true"}]
```

Related Information

tsquery Language

Creating a Trigger Using the Expression Editor

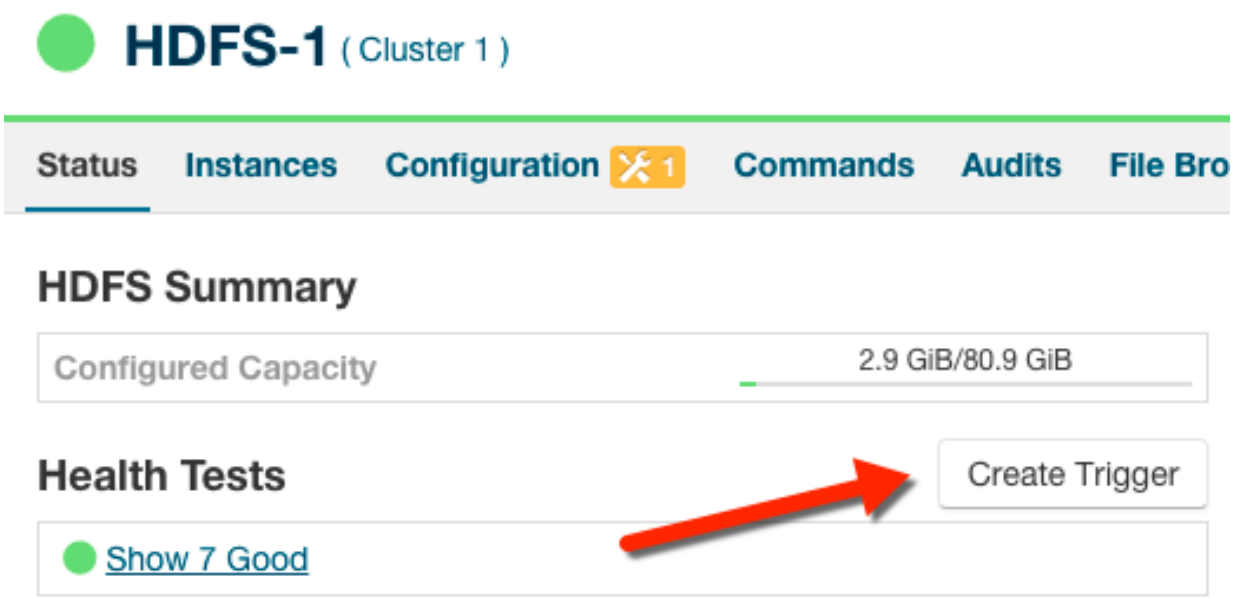
About this task

The Create New Trigger screen allows you to use a graphical editor to build the JSON string that defines a trigger. You can use the expression editor section to build the tsquery statement, or you can edit the tsquery statement manually. Triggers use the tsquery language to create trigger expressions.

Procedure

To create a trigger using the expression editor:

1. Go to a service, role, role configuration group, or host configuration page and click the Create Trigger button next to the Health Test section.



   The Create New Trigger screen displays.

   As you build the trigger, the actual query text displays to the right, along with a preview of a chart returned by the query.
2. Enter a name for the trigger in the Name field.
3. Build the metric expressions:
   a. Select the function to use in your expression, either Last, Min, or Max.
   b. Select the metric by typing its name in the Metric field. A list of available metrics displays as you type.
   c. Select the operator, either >, >=, =, <, or <=.
   d. Enter the value to use for the comparison in the Value field.
   e. (Optional) Click the + icon to add additional expressions. Additional expressions are added to the query using the logical operator AND.
4. (Optional) Create a predicate for the query. Under Attribute Conditions, click the + icon to add an attribute condition.

   A set of fields displays that you use to build an expression for the predicate.

   a. Type the attribute name in the Attribute field. A list of attributes displays as you type.
   b. Select the operator, either = or RLIKE.
   c. Enter the value for the comparison in the Value field.
   d. (Optional) Click the + icon to add additional expressions. Additional expressions are added to the predicate using the logical operator AND.
5. Select an Action from the drop-down menu to define the action taken when the trigger fires:
   - Mark as bad (red)
   - Mark as concerning (yellow)
6. Enter a value for the Stream Threshold. Leave the value set to 0 to include all streams; enter an integer to set the number of streams required to meet the condition.
7. Select Enabled to enable the trigger. If you disable the trigger, it does not run.
8. (Optional) Select Suppressed. A suppressed trigger still runs but does not impact the health display of the owning entity.

9. Verify your expression:

   In the area to the right of the expression builder, in the Preview section, the expression you have built displays. A chart also displays the result of the query. Click Show Filtered Streams to see all streams. Click Hide Filtered Streams to hide streams that do not meet the stream threshold.

   You can edit your trigger using the fields in the expression builder, or you can click the Edit Manually link to display a text box in which you can manually edit the trigger. Click Use Editor to return to the expression builder.

   Important: If you select Edit Manually, changes you make manually do not appear in the expression builder when you click Use Editor.

10. Click Create Trigger to save your trigger.

Related Information

tsquery Language

Metric Expressions

Predicates

Editing, Deleting, Suppressing, or Deleting a Trigger

Procedure

1. Go to the service, role, role configuration group, or host configuration page where the trigger was created. (For example: select ClustersHDFS.)
2. In the Health Tests section, click the trigger name. (You may need to click a Show ... link to expand the list of triggers.)

   A page displays showing the query and chart for the trigger. Click Show Filtered Streams to see all streams. Click Hide Filtered Streams to hide streams that do not meet the stream threshold.
3. Click the Actions drop-down menu and select one of the following actions:
   - Edit Trigger

     A page opens where you can edit the query. Click Save Trigger to save your changes.
   - Disable Trigger or Enable Trigger
   - Suppress Trigger or Unsuppress Trigger
   - Delete Trigger

Cloudera Manager Trigger Use Cases

Cloudera Manager allows you to monitor cluster performance. Some indicators require timely attention to keep your data safe. Triggers let you track occurrence and severity of issues so that you can fix problems before they result in system failures.

The conditions you create for your trigger can be quite complex, but they do not need to be in order to be useful. This topic describes two triggers that alert you when you are approaching capacity limits for your cluster.

Creating a Trigger for Memory Capacity

About this task

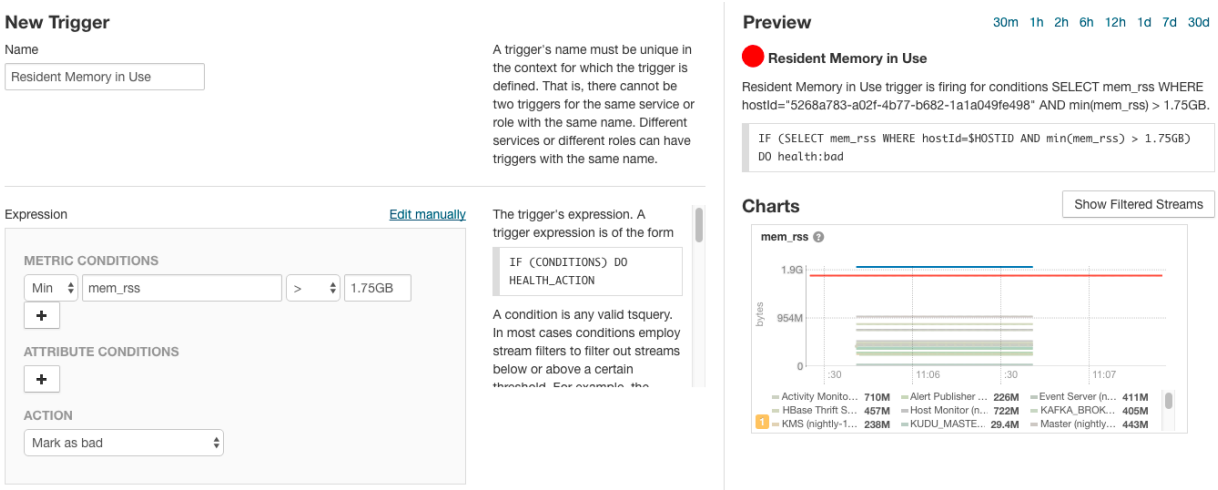
A common use case is to monitor memory usage, and trigger a warning if your system is approaching its upper limit.

Procedure

To create a memory usage trigger:

1. In Cloudera Manager, go to the Hosts page.
2. Click a link in the Name column to open a host status page.
3. In the Health Tests section, click Create Trigger.
4. On the New Trigger page, enter the name Resident Memory In Use.
5. In Expression, set these metric conditions:
   a. Scalar Function: Min.
   b. Metric: mem_rss.
   c. Comparator: > (greater than).
   d. Scalar Value: 1.75GB. (This is a low value for demonstration purposes, so that it will trigger the action. In practice, use a value that more accurately reflects the memory limits of your cluster.)
6. Set Action to Mark as bad.

   As you work, the Preview shows the resulting chart and current status of your host.



7. Scroll down and choose whether to apply this trigger to All hosts.
8. Click Create Trigger.

Creating a Trigger for CPU Capacity

Procedure

1. In Cloudera Manager, go to the Hosts page.
2. Click a link in the Name column to open a host status page.
3. In the Health Tests section, click Create Trigger.
4. On the New Trigger page, enter the name CPU Capacity.
5. In Expression, set these metric conditions.
   - Scalar Function: Min.
   - Metric: cpu_percent.
   - Comparator: > (greater than).
   - Scalar Value: 90.
6. Set Action to Mark as concerning.
7. Click Create Trigger.

Results

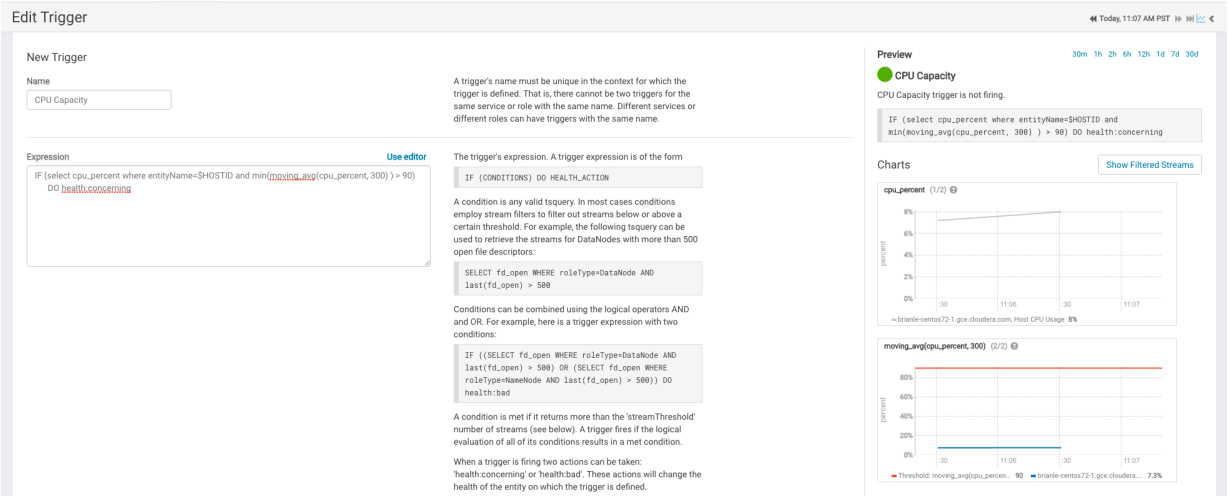
This trigger fires whenever the CPU percentage exceeds 90%. However, just exceeding 90% of available CPU resources is not necessarily a bad thing. What would be of more concern is if CPU resources were to consistently exceed 90% over an extended period. You can modify the trigger to evaluate the average CPU usage over time.

To modify the trigger to capture high CPU usage in a five minute window, do the following.

1. In Cloudera Manager, go to the Hosts page.
2. Click a link in the Name column to open the host status page.
3. In the Health Tests section, click the Show *n* Good link.
4. Click the link for CPU Capacity.
5. Choose Actions Edit Trigger .

   The metric expression function for average over time, moving_avg, is not available from the pop-up menu in the editor. You can edit the expression directly using tsquery language.
6. Above the Expression editor, click Edit manually.
7. Revise the expression as follows.

```
IF (select cpu_percent where entityName=$HOSTID and min(moving_avg(cpu_p
ercent, 300) ) > 90)
      DO health:concerning
```



8. Click Save Trigger.

Related Information

Triggers

tsquery Language

Cloudera Manager Metrics

Lifecycle and Security Auditing

An *audit event* is an event that describes an action that has been taken for a cluster, host, license, parcel, role, service or user.

Minimum Required Role: Auditor (also provided by Full Administrator)

Cloudera Manager records cluster, host, license, parcel, role, and service *lifecycle events* (activate, create, delete, deploy, download, install, start, stop, update, upgrade, and so on), user *security-related events* (add and delete user, login failed and succeeded), and provides an audit UI and API to view, filter, and export such events. The Cloudera

Manager audit log does not track the progress or results of commands (such as starting or stopping a service or creating a directory for a service), it just notes the command that was executed and the user who executed it. To view the progress or results of a command, follow the procedures in the topic *Viewing and Running Recent Commands*.

In CDP, Ranger performs auditing against the data access policies defined for each service. Ranger provides an audit UI and API to view, filter, and export service access audit logs. For information about auditing features, see the Ranger Auditing documentation.

Related Information

Viewing Running and Recent Commands

Viewing Audit Events

You can view audit events for a cluster, service, role, or host.

Object	Procedure
Cluster	**1.** Click the Audits tab on the top navigation bar.
Service	**1.** Click the Clusters tab on the top navigation bar. **2.** Select a service. **3.** Click the Audits tab on the service navigation bar.
Role	**1.** Click the Clusters tab on the top navigation bar. **2.** Select a service. **3.** Click the Instances tab on the service navigation bar. **4.** Select a role. **5.** Click the Audits tab on the role navigation bar.
Host	**1.** Click the Hosts tab on the top navigation bar. **2.** Select a host. **3.** Click the Audits tab on the host navigation bar.

Audit event entries are ordered with the most recent at the top.

Audit Event Properties

The following properties can appear in an audit event entry:

- Date - Date and time the action was performed.
- Command - The action performed.
- Source - The object affected by the action.
- User - The name of the user that performed the action.
- IP Address - The IP address of the client that initiated the action.
- Host IP Address - The IP address of the host on which the action was performed.
- Service - The name of the service on which the action was performed.
- Role - The name of the role on which the action was performed.

Filtering Audit Events

You filter audit events by selecting a time range and adding filters.

You can use the Time Range Selector or a duration link (

) to set the time range. When you select the time range, the log displays all events in that range. The time it takes to perform a search will typically increase for a longer time range, as the number of events to be searched will be larger.

Related Information

Time Line

Adding a Filter

You can filter audit events by a chosen property.

Procedure

Add the filter:

- Click the icon that displays next to a property when you hover in one of the event entries. A filter containing the property, operator, and its value is added to the list of filters at the left and Cloudera Manager redisplays all events that match the filter.
- Click the Add a filter link. A filter control is added to the list of filters.
  a. Choose a property in the drop-down list. You can search by properties such as Username, Service, Command, or Role. The properties vary depending on the service or role.
  b. If the property allows it, choose an operator in the operator drop-down list.
  c. Type a property value in the value text field. To match a substring, use the like operator and specify % around the string. For example, to see all the audit events for files created in the folder /user/joe/out specify Source l ike %/user/joe/out%.
  d. Click Search. The log displays all events that match the filter criteria.
  e. Click ✚ to add more filters and repeat steps a through d.

Removing a Filter

You can remove a filter and repeat the event search with different filter criteria.

Procedure

1. Click the ➖ at the right of the filter. The filter is removed.
2. Click Search. The log displays all events that match the filter criteria.

Downloading Audit Events

You can download audit events in CSV formats.

Procedure

1. Specify desired filters and time range.
2. Click the Download CSV button.
   A file with the following fields is downloaded: service, username, command, ipAddress, resource, allowed, time stamp, operationText. The structure of the resource field depends on the type of the service:
   - HDFS - A file path
   - Hive, Hue, and Impala - *database*:*tablename*
   - HBase - *table family*:*qualifier*

   For Hive, Hue, and Impala query and load commands, operationText is the query string.

HDFS Service Audit Log

```
service,username,command,ipAddress,resource,allowed,timestamp
hdfs1,cloudera,setPermission,10.20.187.242,/user/hive,false,"2013-02-09T00
:59:34.430Z"
hdfs1,cloudera,getfileinfo,10.20.187.242,/user/cloudera,true,"2013-02-09T00
:59:22.667Z"
```

```
hdfs1,cloudera,getfileinfo,10.20.187.242,/,true,"2013-02-09T00:59:22.658Z"
```

In this example, the first event access was denied, and therefore the allowed field has the value false.

Charting Time-Series Data

Cloudera Manager enables you to enter a query for a time series, chart the time-series data, group (facet) individual time series if your query produced multiple time series, and save the results as a dashboard.

The following sections have more details on the terminology used, how to query for time-series data, displaying chart details, editing charts, and modifying chart properties.

Terminology

The list below describes the terminology used when creating Charts.

Entity

A Cloudera Manager component that has metrics associated with it, such as a service, role, or host.

Metric

A property that can be measured to quantify the state of an entity or activity, such as the number of open file descriptors or CPU utilization percentage. For a list of all the metrics supported by Cloudera Manager, see the Cloudera Manager Metrics reference documentation.

Time Series

A list of (time, value) pairs that is associated with some (entity, metric) pair such as, (datanode-1, fd_open), (*hostname*, cpu_percent). In more complex cases, the time series can represent operations on other time series. For example, (datanode-1 , cpu_user + cpu_system).

Facet

A display grouping of a set of time series. By default, when a query returns multiple time series, they are displayed in individual charts. Facets allow you to display the time series in separate charts, in a single chart, or grouped by various attributes of the set of time series.

Related Information

Cloudera Manager Metrics

Building a Chart with Time-Series Data

Use the Chart Builder to build a chart with time-series data.

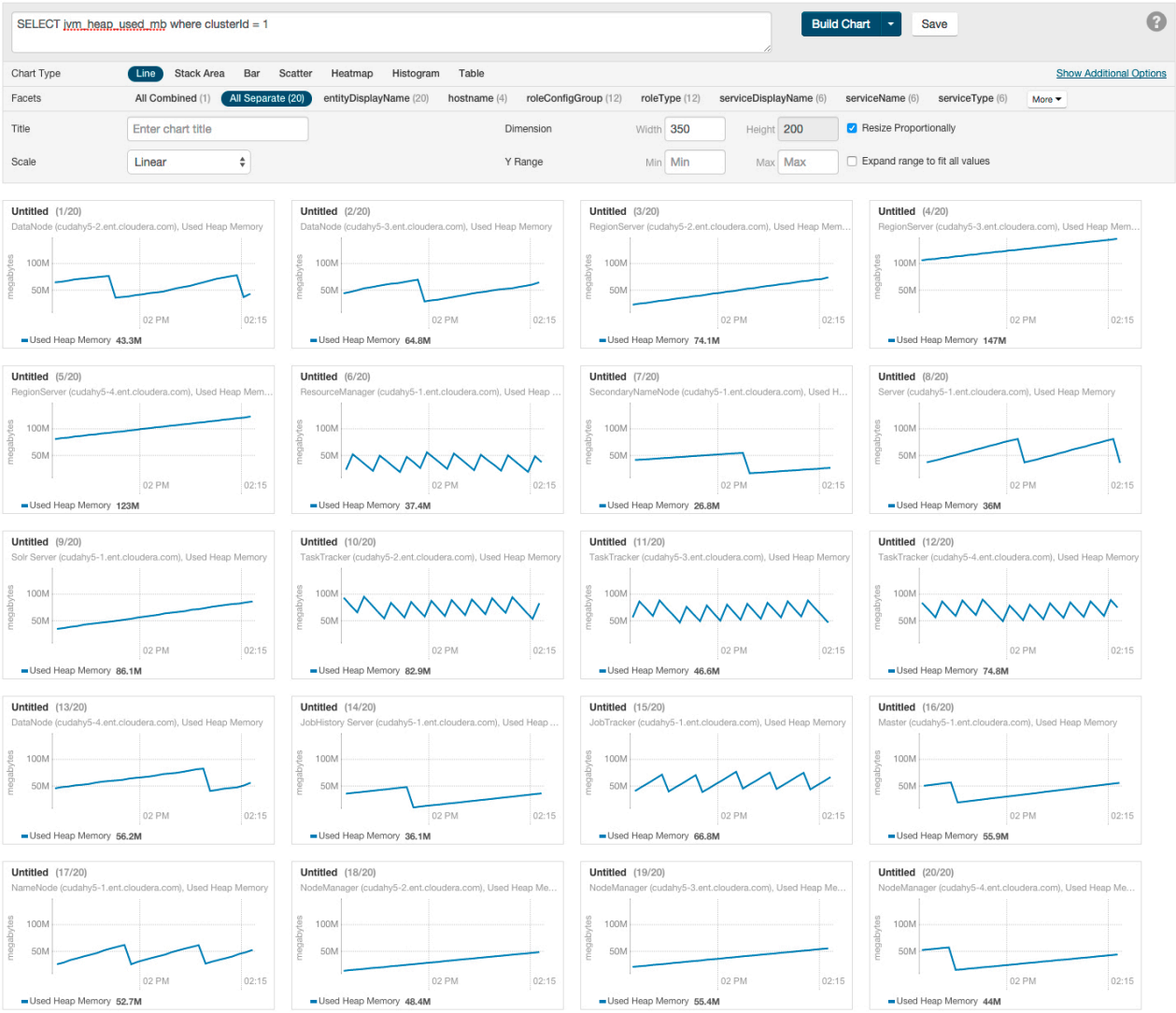
Procedure

1. Select Charts Chart Builder .
2. Display time series in one of the following ways.
   - Select a recently used statement: Click the to the right of the Build Chart button to display a list of recently run statements and select a statement. The statement text displays in the text box and the chart(s) that display that time series will display.
   - Select from the list of Chart Examples:
     a. Click the question mark icon to the right of the Build Chart button to display a list of examples with descriptions.
     b. Click Try it to create a chart based on the statement text in the example.
   - Type a new statement: Press Spacebar in the text box. tsquery statement components display in a drop-down list. These suggestions are part of type ahead, which helps build valid queries. Scroll to the desired component

and click Enter. Continue choosing query components by pressing Spacebar and Enter until the tsquery statement is complete.

Example

For example, the query SELECT jvm_heap_used_mb where clusterId = 1 could return a set of charts like the following:



Configuring Time-Series Query Results

A time-series query returns one or more time series or scalar values. By default a maximum of 250 time series will be returned.

Before you begin

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Procedure

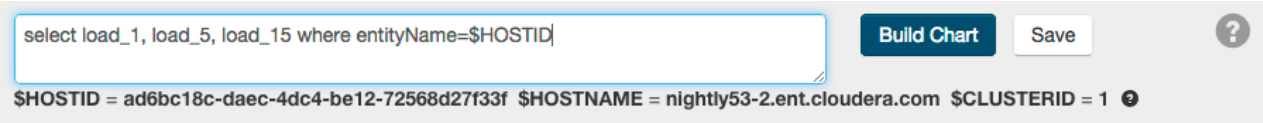
1. To change this value, select Administration Settings .

2. In the Advanced category, set the Maximum Number Of Time-Series Streams Returned Per Time-Series Query or the Maximum Number of Time-Series Streams Returned Per Heatmap property.
3. Click Save Changes.

Using Context-Sensitive Variables in Charts

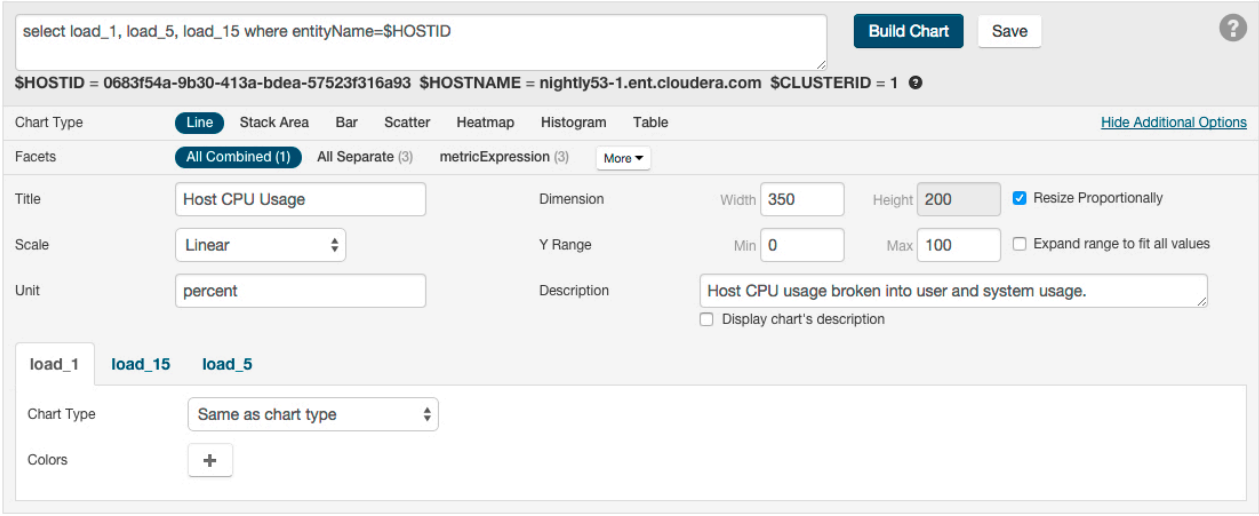
When editing charts from a service, role or host status or charts page, or when adding a chart to a status page, a set of context-sensitive variables (each beginning with '$') will be displayed below the query box on the Chart Builder page.

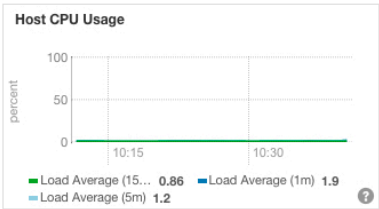
For example, you might see variables similar to those in the query below:



Notice the $HOSTNAME portion of the query string. $HOSTNAME is a variable that will be resolved to a specific value based on the page before the query is actually issued. In this case, $HOSTNAME will become nightly53-2.ent.cloudera.com.

The chart below shows an example of the output of a similar query.





Context-sensitive variables are useful since they allow portable queries to be written. For example the query above may be on the host status page or any role status page to display the appropriate host's swap rate. Variables cannot be used in queries that are part of user-defined dashboards since those dashboards have no service, role or host context.

Chart Properties

By default, the time-series data retrieved by the tsquery is displayed on its own chart, using a Line style chart, a default size, and a default minimum and maximum for the Y-axis. You can change the chart type, facet the data, set the chart scale and size, and set X- and Y-axis ranges.

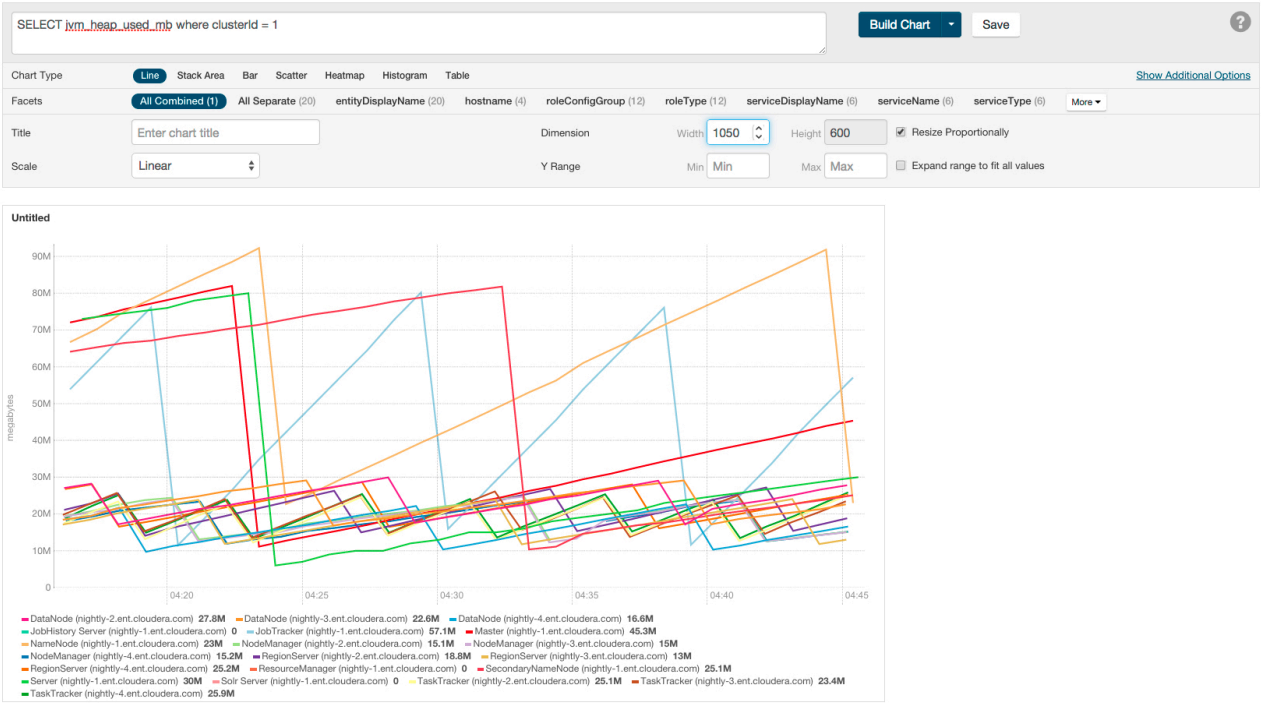
Changing Scale

You can set the scale of the chart to linear, logarithmic, and power.

Changing Dimensions

You can change the size of your charts by modifying the values in the Dimension fields. They change in 50-pixel increments when you click the up or down arrows, and you can type values in as long as they are multiples of 50. If you have multiple charts, depending on the dimensions you specify and the size of your browser window, your charts may appear in rows of multiple charts. If the Resize Proportionally checkbox is checked, you can modify one dimension and the other will be modified automatically to maintain the chart's width and height proportions.

The following chart shows the same query as the previous chart, but with All Combined selected (which shows all time series in a single chart) and with the Dimension values increased to expand the chart.



Changing Axes

You can change the Y-axis range using the Y Range minimum and maximum fields.

The X-axis is based on clock time, and by default shows the last hour of data. You can use the Time Range Selector or a duration link (30m 1h 2h 6h 12h 1d 7d 30d) to set the time range.

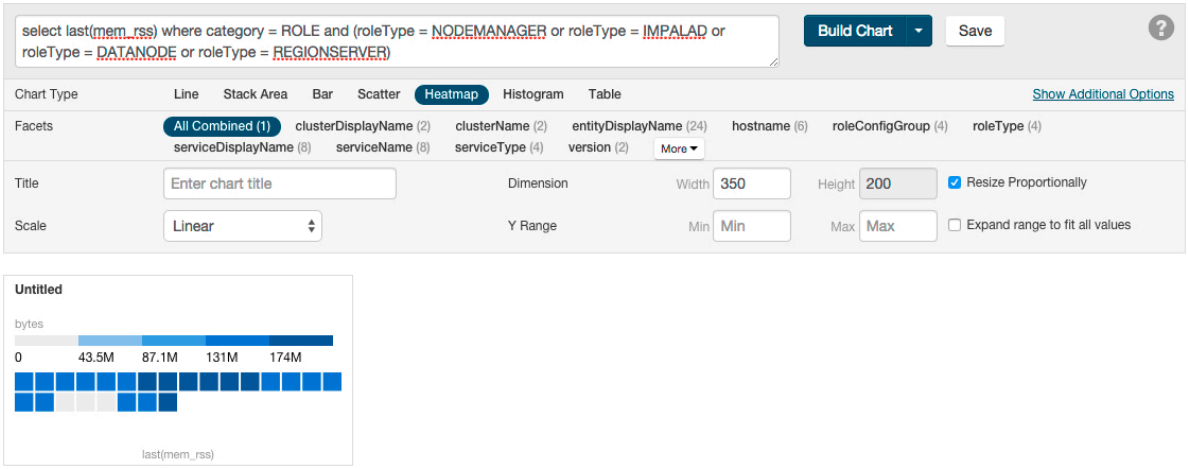
Changing the Chart Type

You can change a chart from one type to another.

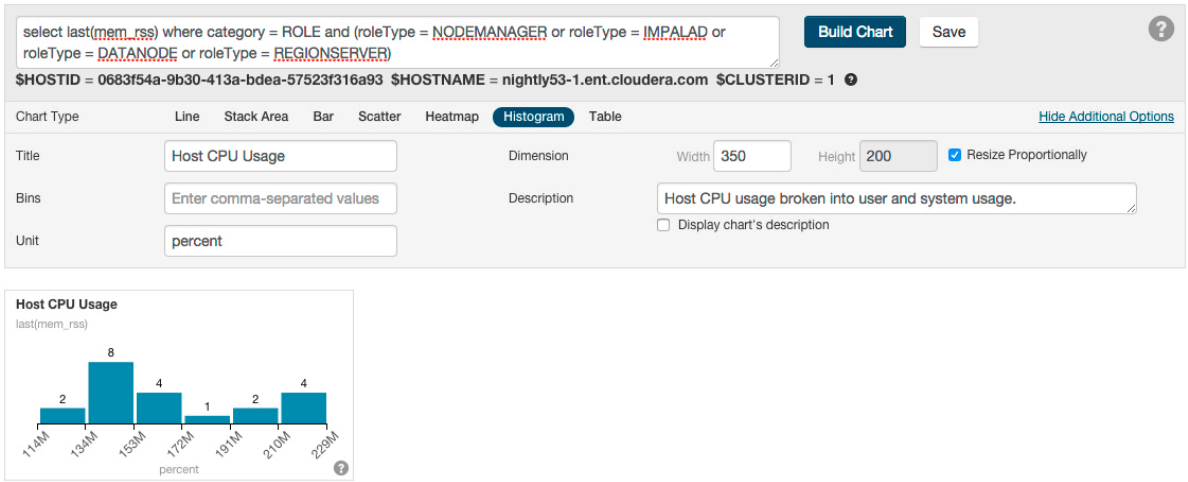
Procedure

1. Click ChartsChart Builder to access the chart, or click the chart on the Cloudera Manager home page and then View Entity Chart.

2. Click one of the chart types:
   - Line - Displays the points in the time series as continuous line.
   - Stack Area - Displays the points in the time series as continuous line and the area under the line filled in.
   - Bar - Displays each the value of the metric averaged over a second as a bar.
   - Scatter - Displays the points in the time series as dots.
   - Heatmap - Displays a metric thermometer and grid of colored squares. The thermometer displays buckets that represent a range of metric values and a color coding for the bucket. Each square represents an entity and the color of the square represents the value of a metric within a range. The following heatmap shows the last value of the resident memory for the NodeManager, ImpalaD, DataNode, and RegionServer roles.



   - Histogram - Displays the time series values as a set of bars where each bar represents a range of metric values and the height of the bar represents the number of entities whose value falls within the range. The following histogram shows the number of roles in each range of the last value of the resident memory.



   - Table - Displays the time series values as a table with each row containing the data for a single time value.

Note: Heatmaps and histograms render charts for a single point as opposed to time series charts that render a series of points. For queries that return time series, Cloudera Manager will generate the heatmap or histogram based on the last recorded point in the series, and will issue the warning: "Query returned more than one value per stream. Only the last value was used." To eliminate this warning, use a scalar returning function to choose a point. For example, use select last(cpu_percent) to use the last point or sele ct max(cpu_percent) to use the maximum value (in the selected time range).

Related Information

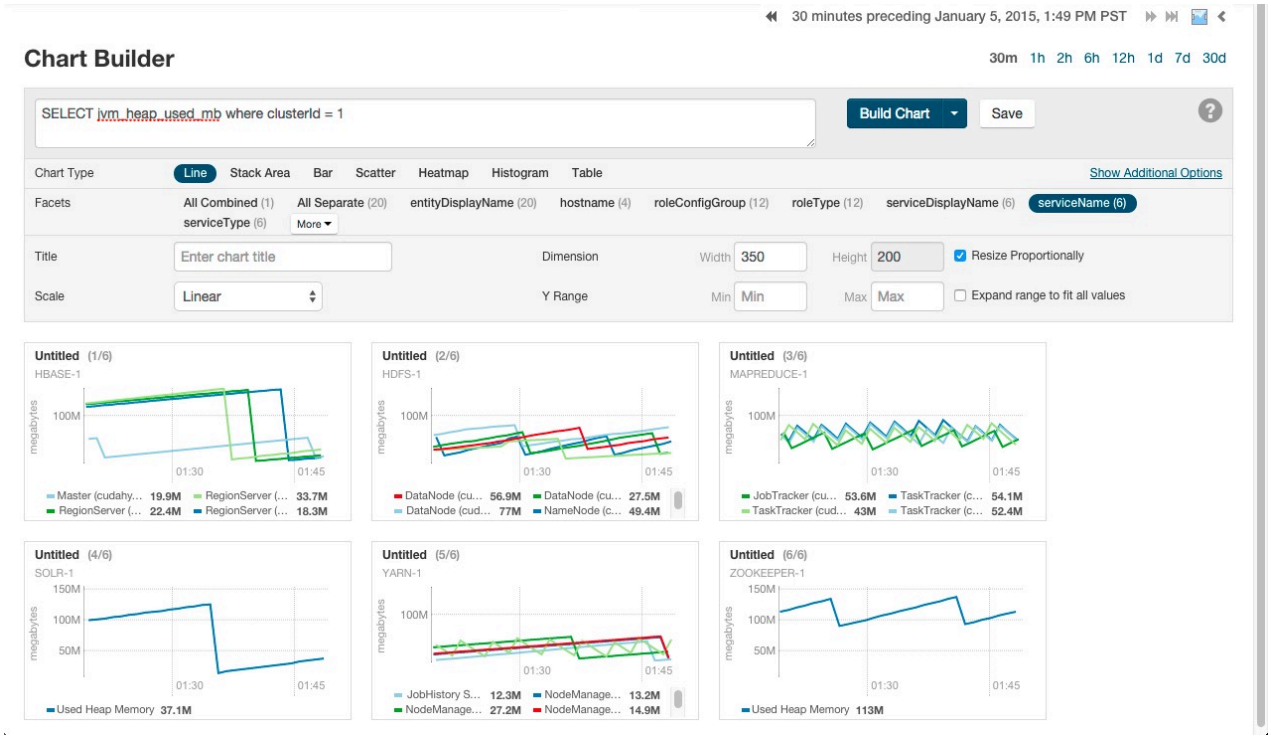
Metric Expressions

Grouping (Faceting) Time Series

A time-series plot for a service, role, or host may actually be a composite of multiple individual time series. Using facets, you can combine time series based their attributes.

For example, the query SELECT jvm_heap_used_mb where clusterId = 1 returns time-series data for the JVM heap used. Each time series has hostname, role type, metric, and entity name attributes. By default each attribute is displayed all on a single chart.

To change the organization of the chart data, click one of the facets in the facet section in the upper part of the screen. The number in parentheses indicates how many charts will be displayed for that facet. As shown in the image below if the serviceName facet is selected for the JVM heap query, the time series is grouped into six charts, one chart each for each service name. The charts for service types with multiple roles contain multiple lines (for example, HBase, HDFS) while services that have only one role (for example, ZooKeeper) contain just a single line. When a chart contains multiple lines, each entity is identified by a different color line.



Displaying Chart Details

You can interact with a chart to display various chart details.

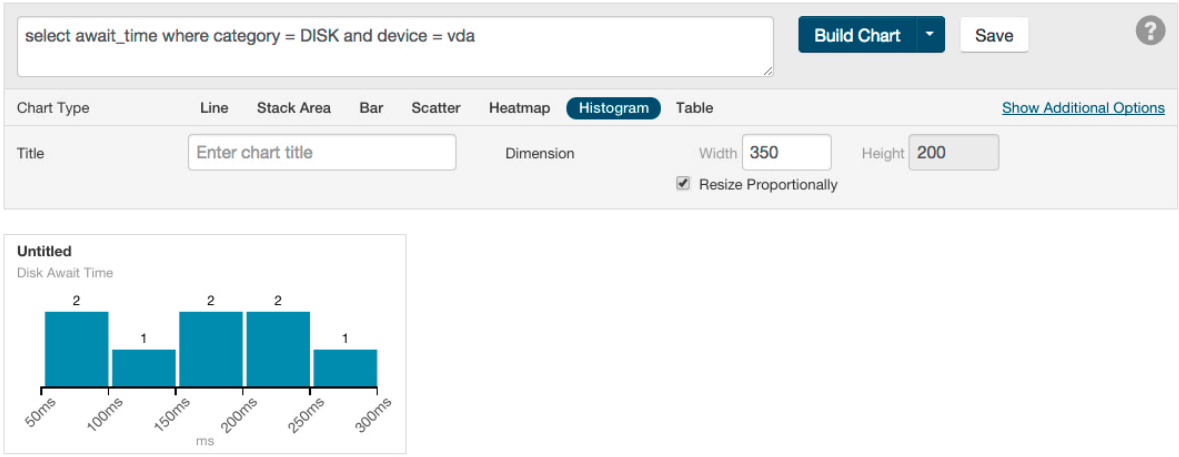
When you move your mouse over a chart, its background turns gray, indicating that you can act upon it.

- Moving the mouse to a data point on a line, stack area, or bar chart shows the details about that data point in a pop-up tooltip.
- Click a line, stack area, scatter, or bar chart to expand it into a full-page view with a legend for the individual charted entities as well more fine-grained axes divisions.
  - If there are multiple entities in the chart, you can
    - Check and uncheck the legend item to hide or show the time series for the entities on the chart.

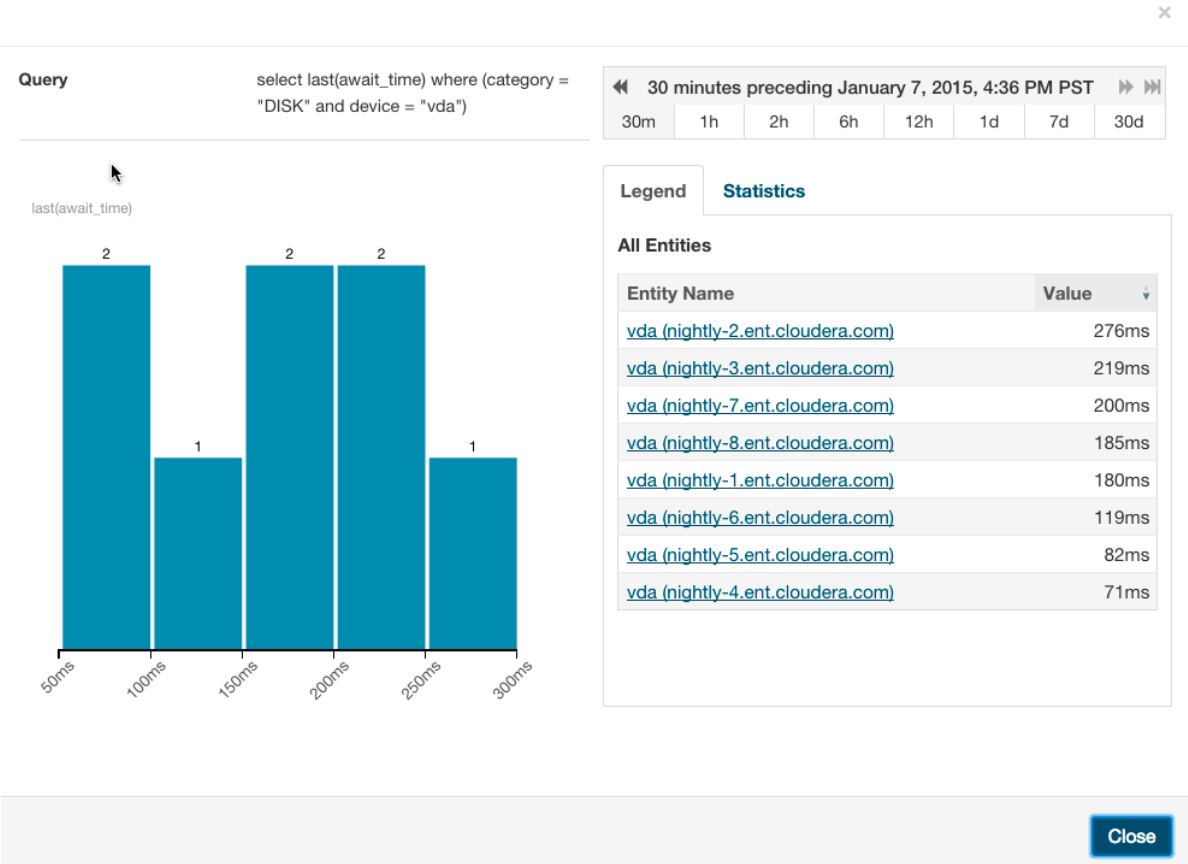
    - If there are service, role, or host instances in the chart, click the View link to display the instance's Status page.
  - Click the Close button to return to the regular chart view.

- Heatmap - Clicking a square in a heatmap displays a line chart of the time series for that entity.
- Histogram -
  - Mousing over the upper right corner of a histogram and clicking opens a pop-up containing the query that generated the chart, an expanded view of the chart, a list of entity names and links to the entities whose metrics

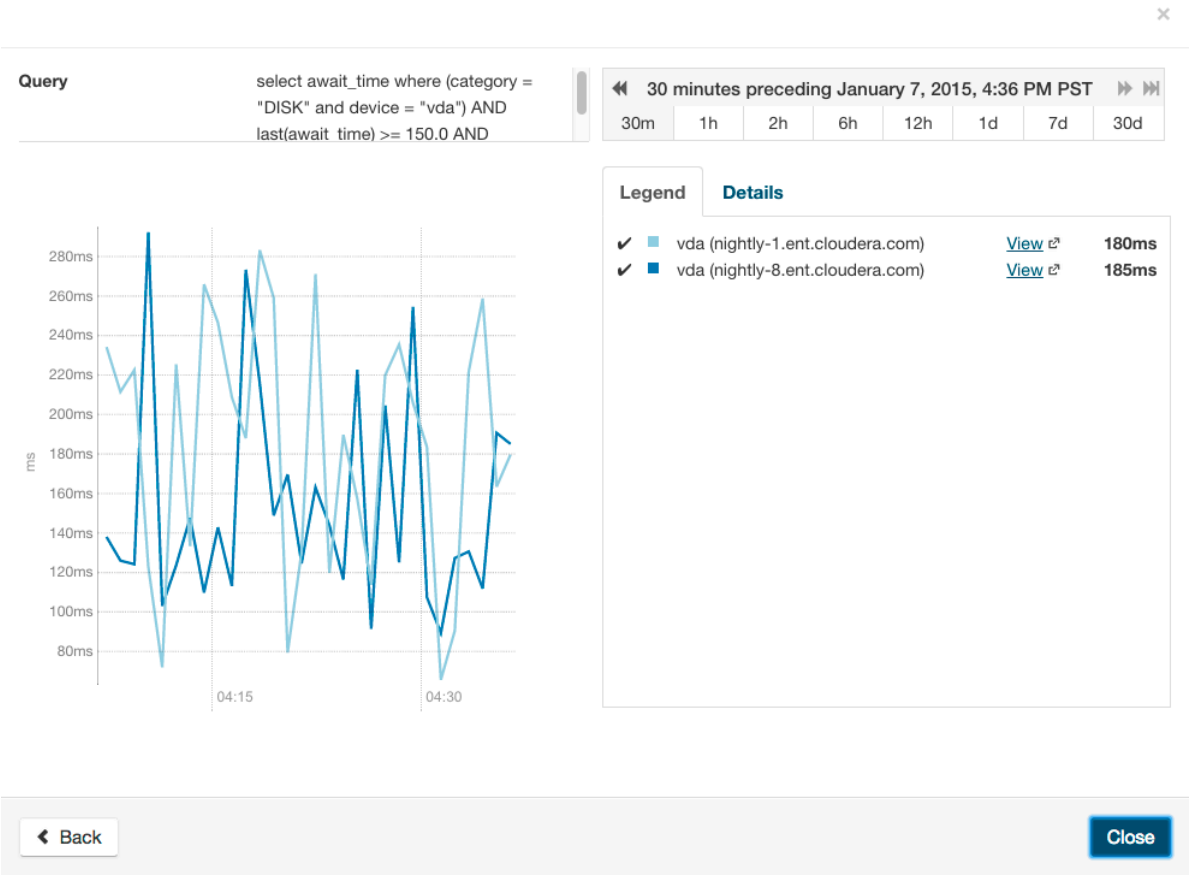
are represented by the histogram bars, and the value of the metric for each entity. For example, clicking the following histogram



displays the following:



- Clicking a bar in the expanded histogram displays a line chart of the time series from which the histogram was generated:



Clicking the < Back link at the bottom left of the line chart returns to the expanded histogram.

Editing a Chart

You can edit a chart from the custom dashboard and save it back into the same or another existing dashboard, or to a new custom dashboard.

About this task

Editing a chart only affects the copy of the chart in the current dashboard – if you have copied the chart into other dashboards, those charts are not affected by your edits.

Procedure

1. Move the cursor over the chart, and click the gear icon at the top right.
2. Click Open in Chart Builder. This opens the Chart Builder page with the chart you selected already displayed.
3. Edit the chart's select statement and click Build Chart.

Related Information

Dashboard Types

Saving a Chart

After you edit a chart, you can save it to a new or existing custom dashboard.

Before you begin

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

Users with Viewer, Limited Operator, or Operator user roles can edit charts and view the results, but cannot save them to a dashboard.

Procedure

1. Modify the chart's properties and click Build Chart.
2. Click Save to open the Save Chart dialog box, and select one of the following:
   - Update chart in current dashboard: <name of current dashboard>.
   - Add chart to another dashboard.
   - Add chart to a new custom dashboard.
3. Click Save Chart.
4. Click View Dashboard to go to the dashboard where the chart has been saved.

   Saving a chart only affects the copy of the chart in the dashboard where you save it – if you have previously copied the chart into other dashboards, those charts are not affected by your edits.

What to do next

You can also arrange and resize the charts on a page by clicking the Edit Layout link. See Moving and Resizing Charts on page 145.

Related Information

Saving Charts to a New Dashboard

Saving Charts to an Existing Dashboard

Obtaining Time-Series Data Using the API

You can obtain time-series data using the Cloudera Manager API. For details about using a tsquery statement to obtain time-series data, see the /timeseries API documentation at http:// cmServerHost:7180/static/apidocs/path__timeseries.html.

Procedure

1. Click ChartsChart Builder to access the chart, or click the chart on the Cloudera Manager home page and then click View Entity Chart.
2. To see the API call that returns the time-series data for an existing chart, click the blue down-arrow at the upper-right corner of the chart and click Export JSON.
   A new web browser window opens, displaying the time-series data in JSON format. The query string of the URL for that window displays the API call that retrieved the time-series data.

Dashboards

A *dashboard* is a set of charts. This section covers creating, configuring, and managing dashboards.

Related Information

Charting Time-Series Data

Dashboard Types

A *default dashboard* is a predefined set of charts that you cannot change.

In a default dashboard you can:

- Display chart details.

- Edit a chart and then save back to a new or existing custom dashboard.

A *custom dashboard* contains a set of charts that you can change. In a custom dashboard you can:

- Display chart details.
- Edit a chart and then save back to a new or existing custom dashboard.
- Save a chart, make any modifications, and then save to a new or existing dashboard.
- Remove a chart.

When you first display a page containing charts it has a custom dashboard with the same charts as a default dashboard.

Related Information

Displaying Chart Details
Editing a Chart
Saving Charts to a New Dashboard
Saving Charts to an Existing Dashboard
Saving a Chart
Chart Properties
Removing a Chart from a Custom Dashboard

Creating a Dashboard

When you create a dashboard, you specify a name and optionally a duration.

Procedure

1. Do one of the following:
   - Select Charts New Dashboard .
   - Select Charts Manage Dashboards and click Create Dashboard.
   - Save a chart to a new dashboard.
2. Specify a name and optionally a duration.

   Conventional dashboard names follow the patterns given by one of the following Java regular expressions:

   ```
   Pattern.compile("^(.+):(\\d):" + STATUS_VIEW_SUFFIX + "$");
   ```

   ```
   Pattern.compile("^(.+):" + STATUS_VIEW_SUFFIX + "$");
   ```

   Examples of expected names are: HOST:STATUS_VIEW, MGMT:STATUS_VIEW, or HDFS:5:STATUS_VIEW. If the dashboard name does not match the expected pattern, a warning will be displayed in the server log.
3. Click Create Dashboard.

Related Information

Saving Charts to a New Dashboard

Managing Dashboards

You can create, clone, edit, export, import, and remove dashboards.

To manage dashboards, select Charts Manage Dashboards .

- Create Dashboard - create a new dashboard.
- Clone - clones an existing dashboard.
- Edit - edit an existing dashboard.
- Export - exports the specifications for the dashboard as a JSON file.
- Import Dashboard - reads an exported JSON file and recreates the dashboard.
- Remove - deletes the dashboard.

Configuring Dashboards

You can change the time scale of a dashboard, switch between default and custom dashboards, and reset a custom dashboard.

Setting the Time Scale of a Dashboard

By default the time scale of a dashboard is 30 minutes. To change the time scale, click a duration link

30m 1h 2h 6h 12h 1d 7d 30d at the top-right of the dashboard.

Setting the Dashboard Type

To set the dashboard type, click and select one of the following:

- Custom - displays a custom dashboard.
- Default - displays a default dashboard.
- Reset - resets the custom dashboard to the predefined set of charts, discarding any customizations.

Saving Charts to Dashboards

You can save the charts and their configurations (type, dimension, and y-axis minimum and maximum) to a new dashboard or to an existing dashboard.

Minimum Required Role: Configurator (also provided by Cluster Administrator, Limited Cluster Administrator , and Full Administrator)

If your tsquery statement resulted in multiple charts, those charts are saved as a unit (either to a new or existing dashboard). You cannot edit the individual plots in that set of charts, but you can edit the set as a whole. A single edit button appears for the set that you saved — typically on the last chart in the set.

You can edit a copy of the individual charts in the set, but the edited copy does not change the original chart in the dashboard from which it was copied.

Saving Charts to a New Dashboard

You can save new or existing charts to a new dashboard.

Procedure

1. Optionally, modify the chart properties.
2. If the chart was created with the Chart Builder, optionally type a name for the chart in the Title field.
3. Do one of the following:
   - New chart: Click Save.
   - Existing chart: Move the cursor over the chart, and click the ⚙ icon at the top right.
4. Optionally, edit the chart name.
5. Select the Add chart to a new custom dashboard option.
6. Enter a dashboard name.
7. Click Save Chart.
   The new dashboard appears on the menu under the top-level Charts tab.

Related Information

Chart Properties

Saving Charts to an Existing Dashboard

You can save new or existing charts to an existing dashboard.

Procedure

1. Optionally, modify the chart properties.
2. If the chart was created with the Chart Builder, optionally type a name for the chart in the Title field.
3. Do one of the following:
   - New chart: Click Save.
   - Existing chart: Move the cursor over the chart, and click the icon at the top right.
4. Optionally, edit the chart name.
5. Select the Add chart to an existing custom or system dashboard option.
6. Select a dashboard from the Dashboard Name drop-down list.
7. Click Save Chart. The chart is added (appended) to the dashboard you select.

Related Information

Chart Properties

Adding a New Chart to the Custom Dashboard

You can add new charts to the custom dashboard on the Status tab of a service, host, or role.

Procedure

1. Click and select one of the following:
   - Add From Charts Library: Displays the charts page.
     - Select one or more charts.
   - Add From Chart Builder: Displays the Add Chart To *Dashboard* page, with variables preset for the specific service, role, or host where you want to add the dashboard.
     a. Click the question mark icon to the right of the Build Chart button and select a metric from the List of Metrics, type a metric name or description into the Basic text field, or type a query into the Advanced field.
     b. Click Build Chart. The charts that result from your query are displayed, and you can modify their chart type, combine them using facets, change their size and so on.
2. Click Add.

Note: If the query you've chosen has resulted in multiple charts, all the charts are added to the dashboard as a set. Although the individual charts in this set can be copied, you can only edit the set as a whole.

Removing a Chart from a Custom Dashboard

You can remove a chart from a custom dashboard.

Procedure

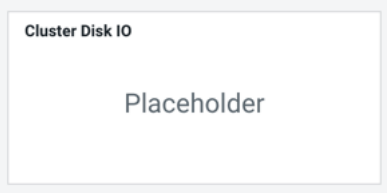
1. Move the cursor over the chart and click the icon at the top right.
2. Click Remove.

Moving and Resizing Charts

You can move or resize the charts on a dashboard, the cluster home page, or a service page.

Procedure

1. Click the Edit Layout link.
   The dashboard changes to an editing view where each chart's display is replaced with a placeholder box that is the same size as the actual chart.



2. Change the size of a chart by dragging the lower-right corner of the chart.
3. Drag charts to change their relative positions.
4. Click Finish Editing to save the new layout.

tsquery Language

The tsquery language is used to specify statements for retrieving time-series data from the Cloudera Manager time-series datastore.

Below are some common queries using the tsquery language:

- Retrieve time series for all metrics for all DataNodes:

```
select * where roleType=DATANODE
```

- Retrieve cpu_user_rate metric time series for all DataNodes:

```
select cpu_user_rate where roleType=DATANODE
```

- Retrieve the jvm_heap_used_mb metric time series divided by 1024 and the jvm_heap_committed metric time series divided by 1024 for all roles running on the host named "my host":

```
select jvm_heap_used_mb/1024, jvm_heap_committed_mb/1024 where category=
ROLE and hostname="my host"
```

- Retrieve the jvm_total_threads and jvm_blocked_threads metric time series for all entities for which Cloudera Manager collects these two metrics:

```
select jvm_total_threads,jvm_blocked_threads
```

Collections

A *collection* is a type of data. The field can take the value:

- ENTITY_DATA -
- IMPALA_QUERIES -
- IMPALA_QUERY_DETAILS -
- YARN_APPLICATIONS -

tsquery Syntax

A tsquery statement has a specific structure.

SELECT [*metric expression*]FROM *collection* WHERE [*predicate*]

Note the following properties of tsquery statements:

- The statement select * is invalid.

- Tokens are case insensitive. For example, Select, select, and SeLeCt are all equivalent to SELECT.
- Multiple statements can be concatenated with semi-colons. Thus example 3 in the *tsquery Language* topic can be written as:

```
select jvm_heap_used_mb/1024 where category=ROLE and hostname=myhost; se
lect jvm_heap_commited_mb/1024 where category=ROLE and hostname=myhost
```

- The metric expression can be replaced with an asterisk (*), as shown in example 1 of the *tsquery Language* topic. In that case, all metrics that are applicable for selected entities, such as DATANODE in example 1, are returned.
- The collection can be omitted.
- The predicate can be omitted, as shown in example 4 of the *tsquery Language* topic. In such cases, time series for all entities for which the metrics are appropriate are returned. For this query you would see the jvm_new_threads metric for NameNodes, DataNodes, TaskTrackers, and so on.

Related Information

Metric Expressions

Predicates

Metric Expressions

A *metric expression* generates the time series. It is a comma-delimited list of one or more metric expression statements.

A *metric expression statement* is the name of a metric, a metric expression function, or a scalar value, joined by one or more metric expression operators.

See the FAQ which answers questions concerning how to discover metrics and use cases for scalar values. For a list of all the supported metrics, see the *Cloudera Manager Metrics* reference documentation.

Metric expressions support the binary operators: +, -, *, /.

Here are some examples of metric expressions:

- jvm_heap_used_mb, cpu_user, 5
- 1000 * jvm_gc_time_ms / jvm_gc_count
- total_cpu_user + total_cpu_system
- max(total_cpu_user)

Related Information

Metric Expression Functions

FAQ

Cloudera Manager Metrics

Metric Expression Functions

Metric expressions support the functions listed in the following table. A function can return a time series or a scalar computed from a time series. Functions that return scalars must be used for heatmap charts.

Function	Returns Scalar?	Description
avg(*metric expression*)	N	Computes a simple average for a time series.
count_service_roles()	Y	Returns the number of roles. There are three variants of this function: • count_service_roles(roleType, roleState) - Returns the number of roles of the specified roleType and roleState . For example, count_service_roles(datanode, running) returns the number of running DataNodes. • count_service_roles(roleType) - Returns the number of roles with the specified role Type. • count_service_roles() - Return the number of roles. For example, select events_critical where count_service_roles() > 100 returns the event_critical metric when the number of roles is greater than 100.

Function	Returns Scalar?	Description
dt(*metric expression*)	N	Derivative with negative values. The change of the underlying metric expression per second. For example: dt(jvm_gc_count).
dt0(*metric expression*)	N	Derivative where negative values are skipped (useful for dealing with counter resets). The change of the underlying metric expression per second. For example: dt0(jvm_gc_time_ms) / 10.
getClusterFact(string factName, double defaultValue)	Y	Retrieves a fact about a cluster. Currently supports one fact: numCores. If the number of cores cannot be determined, defaultValue is returned.
getHostFact(string factName, double defaultValue)	Y	Retrieves a fact about a host. Currently supports one fact: numCores. If the number of cores cannot be determined, defaultValue is returned. For example, select dt(total_cpu_user) / getHostFact(numCores, 2) where category=HOST divides the results of dt(total_cpu_user) by the current number of cores for each host. The following query computes the percentage of total user and system CPU usage each role is using on the host. It first computes the CPU seconds per second for the number of cores used by taking the derivative of the total user and system CPU times. It normalizes the result to the number of cores on the host by using the getHostFact function and multiplies the result by 100 to get the percentage. `select dt0(total_cpu_user)/getHostFact(numCores,1)*100, dt0(total_cpu_system)/getHostFact(numCores,1)*100 where category=ROLE and clusterId=1`
greatest(*metric expression*, *scalar metric expression*)	N	Compares two metric expressions, one of which one is a scalar metric expression. Returns a time series where each point is the result of evaluating max(*point*, *scalar metric expression*).
integral(*metric expression*)	N	Computes the integral value for a stream and returns a time-series stream within which each data point is the integral value of the corresponding data point from the original stream. For example, select integral(maps_failed_rate) will return the count of the failed number of maps.
counter_delta(*metric expression*)	N	Computes the difference in counter value for a stream and returns a time-series stream within which each data point is the difference in counter value of the corresponding data point from the counter value of previous data point in the original stream. For example: select counter_delta(maps_failed_rate) returns the count of the failed number of maps. This method is more accurate than the integral() function. However there are a few caveats: • This function is only implemented for single time-series streams. For streams of cross-entity aggregates, continue to use the integral() function. • If you apply this method for time-series streams which was created using a version of Cloudera Manager older than 5.7, Cloudera Manager fills in the older data points using the integral() function.
last(*metric expression*)	Y	Returns the last point of a time series. For example, to use the last point of the cpu_percent metric time series, use the expression select last(cpu_percent).
least(*metric expression*, *scalar metric expression*)	N	Compares two metric expressions, of which one is a scalar metric expression. Returns a time series where each point is the result of evaluating min(*point*, *scalar metric expression*).
max(*metric expression*)	Y	Computes the maximum value of the time series. For example, select max(cpu_percent).
min(*metric expression*)	Y	Computes the minimum value of the time series.
moving_avg(*metric expression*, time_window_sec)	N	Computes the moving average for a time series over a time window time_window_sec specified in seconds (2, 0.1, and so on)
stats(*metric expression*, *stats name*)	N	Some time-series streams have additional statistics for each data point. These include rollup time-series streams, cross-entity aggregates, and rate metrics. The following statistics are available for rollup and cross-entity aggregates: max, min, avg, std_dev, and sample. For rate metrics, the underlying counter value is available using the "counter" statistics. For example, stats(fd_open_across_datanodes, max) or stats(swap_out_rate, counter).
sum(*metric expression*)	Y	Computes the sum value of the time-series.

Related Information

Charting Time-Series Data
Time Series Attributes

Predicates

A *predicate* limits the number of streams in the returned series.

A predicate can take one of the following forms:

- *time_series_attribute operator value*, where
  - *time_series_attribute* is one of the supported attributes.
  - *operator* is one of = and rlike
  - *value* is an attribute value subject to the following constraints:
    - For attributes values that contain spaces or values of attributes of the form *xxx*Name such as displayName, use quoted strings.
    - The value for the rlike operator must be specified in quotes. For example: hostname rlike "host[0-3]+.*".
    - *value* can be any regular expression as specified in regular expression constructs in the Java Pattern class documentation.
- *scalar_producing_function*(*metric_expression*) *comparator number*, where
  - *scalar_producing_function* is any function that takes a time series and produces a scalar. For example, min or max.
  - *metric_expression* is a valid metric expression. For example, total_cpu_user + total_cpu_system.
  - *comparator* is a comparison operator: <, <=, =, !=, >=, >.
  - *number* is any number expression or a number expression with units. For example, 5, 5mb, 5s are all valid number expressions. The valid units are:
    - Time - ms (milliseconds), s (seconds), m (minutes), h (hours), and d (days).
    - Bytes -b (bytes), kb or kib (kilobytes), mb or mib (megabytes), gb or gib (gigabytes), tb or tib (terabytes), and pb or pib (petabytes)
    - Bytes per second - Bytes and Time: bps, kbps, kibps, mbps, mibps, and so on. For example, 5 kilobytes per second is 5 kbps.
    - Bytes time - Bytes and Time combined: bms, bs, bm, bh, bd, kms, ks, and so on. For example, 5 kilobytes seconds is 5 ks or 5 kis.

You use the AND and OR operators to compose compound predicates.

Example Statements with Compound Predicates

1. Retrieve all time series for all metrics for DataNodes or TaskTrackers.

```
select * where roleType=DATANODE or roleType=TASKTRACKER
```

2. Retrieve all time series for all metrics for DataNodes or TaskTrackers that are running on host named "myhost".

```
select * where (roleType=DATANODE or roleType=TASKTRACKER) and hostname=
myhost
```

3. Retrieve the total_cpu_user metric time series for all hosts with names that match the regular expression "host[0-3]+.*"

```
select total_cpu_user where category=role and hostname rlike "host[0-3]+
.*"
```

Example Statements with Predicates with Scalar Producing Functions

1. Return the entities where the last count of Java VM garbage collections was greater than 10:

```
select jvm_gc_count where last(jvm_gc_count) > 10
```

2. Return the number of open file descriptors where processes have more than 500Mb of mem_rss:

```
select fd_open where min(mem_rss) > 500Mb
```

Related Information

Time Series Attributes

Metric Expression Functions

Java Pattern

Discovering possible predicates

You can use *Predicates* (‘where’ clauses) to specify filters on a wide range of aspects when querying metrics. In some cases, this large selection makes it challenging to choose the right filters for the intended purpose. This topic describes a way to find out more about applicable filters and values in the context of a specific query.

In a tsquery, the main way to filter results is by defining conditions on time series entity attributes. Like metrics, the set of applicable attributes are determined by the entity type(s) being queried. The Charting Time-Series Data topic describes this data model in more detail.

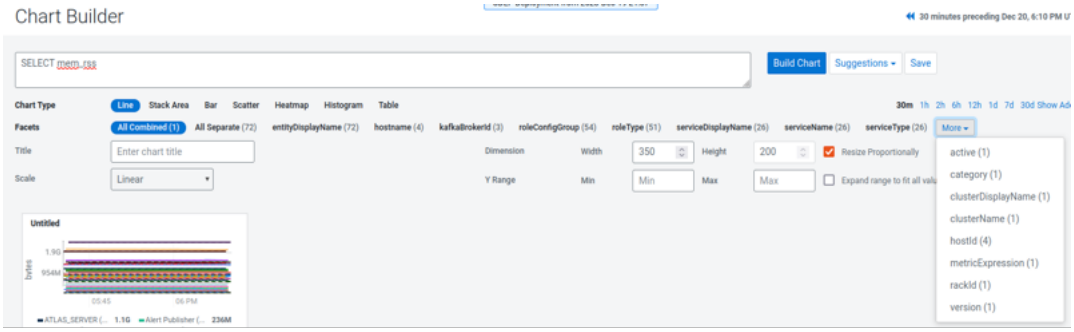
You can find more information in the following sections of the tsquery Language topic:

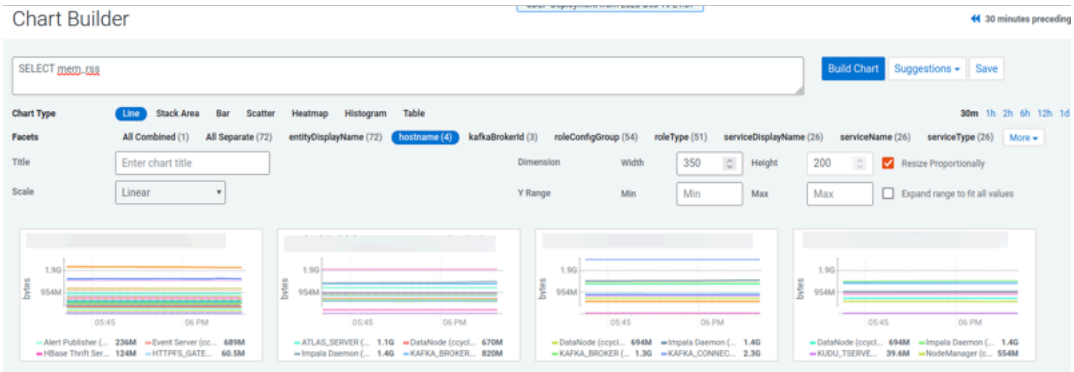
- Predicates: explains what can be used in ‘where’ clauses, including and in addition to attributes.
- Time Series Attributes: a subset of the attributes you can use, with a description for each.
- Time Series Entities and their Attributes: a subset of the time series entity types available, with the applicable attributes listed.

Important:

The tables of entities and attributes cannot be complete for any given cluster, as service integrations (including third-party ones) can extend both. To show the attributes applicable to a specific query, along with the values they can have, on a running cluster, you can enter it into Chart Builder without the ‘where’ clause.

The chart facets displayed are the attributes, while selecting a facet shows one chart for each currently valid value. For more information, see Charting Time-Series Data and the following figures:





In addition to the above approaches, you can also use the Cloudera Manager API to request the same information from a running cluster as follows:

- Mapping of entity types to the metrics they have, as a JSON document: GET /timeseries/schema (The response is large, approximately 50 MB.) This further clarifies the relationship between entity types, attributes, and metrics, while providing insight into the metrics applicable to each entity type.
- A detailed explanation of attributes (similar to the Time Series Attributes table): GET /timeseries/entityTypeAttributes.
- Definitions of the entity types referred to in the schema document, along with applicable attribute names (similar to the Time Series Entities and their Attributes table): GET /timeseries/entityTypes

The results obtained from the API include all entity types and attributes tracked by the running cluster, including those added by service integrations.

Filtering by Day of Week or Hour of Day

You can add an expression to the predicate of a tsquery statement that limits the stream to specified days of the week or to a range of hours in each day.

By Day – Limits the stream to selected days of the week.

The day in () expression takes an argument with a comma-separated list of days of the week, enclosed in parentheses. The days of the week are numbered 1 through 7; 1 = Monday, 2 = Tuesday, and so on. Use the following syntax:

```
day in (#, #, ...)
```

For example, the following expression limits the stream to events that occurred only on weekdays:

```
day in (1,2,3,4,5)
```

By Hour – Limits the stream to a range of hours each day.

The hour in expression takes an argument with a range of hours separated by a colon and enclosed in square brackets. Valid values are integers 0–23:

```
hour in [#:#]
```

For example, the following expression limits the stream to events that occur only between 9:00 a.m. and 5:00 p.m.:

```
hour in [9:17]
```

Add the day or time range expression after the WHERE clause. Do not use the AND keyword. For example:

```
select fd_open where category = ROLE and roleType = SERVICEMONITOR day in (1
,2,3,4,5)
```

You can also combine day in and hour in expressions. Always put the day expression before the hour expression. The following example limits the stream to weekdays between 9:00 a.m. and 5:00 p.m.:

```
select fd_open where category = ROLE and roleType = SERVICEMONITOR day in (1
,2,3,4,5) hour in [9:17]
```

Time Series Attributes

You can use time series attributes when you build a predicate.

Attribute names and most attribute values are case insensitive. displayName and serviceType are two attributes whose values are case sensitive.

Name	Description
active	Indicates whether the entities to be retrieved must be active. A nonactive entity is an entity that has been removed or deleted from the cluster. The default is to retrieve only active entities (that is, active=true). To access time series for deleted or removed entities, specify active=false in the query. For example: `SELECT fd_open WHERE roleType=DATANODE and active=false`
agentName	A Flume agent name.
applicationName	One of the Cloudera Manager monitoring daemon names.
cacheId	The HDFS cache directive ID.
category	The category of the entities returned by the query: CLUSTER, DIRECTORY, DISK, FILESYSTEM, FLUME_SOURCE, FLUME_CHANNEL, FLUME_SINK, HOST, HTABLE, IMPALA_QUERY_STREAM, NETWORK_INTERFACE, ROLE, SERVICE, USER, YARN_APPLICATION_STREAM, YARN_QUEUE. Some metrics are collected for more than one type of entity. For example, total_cpu_user is collected for entities of category HOST and ROLE. To retrieve the data only for hosts use: `select total_cpu_user where category=HOST` The ROLE category applies to all role types (see roleType attribute). The SERVICE category applies to all service types (see serviceType attribute). For example, to retrieve the committed heap for all roles on host1 use: `select jvm_committed_heap_mb where category=ROLE and hostname="host1"`
clusterDisplayName	The user-defined display name of a cluster.
clusterName	The cluster ID. To specify the cluster by its display name, use the clusterDisplayName attribute.
componentName	A Flume component name. For example, channel1, sink1.
device	A disk device name. For example, sda.
entityName	A display name plus unique identifier. For example: HDFS-1-DATANODE-692d141f436ce70aac080aedbe83f887.
expired	A Boolean that indicates whether an HDFS cache directive expired.
groupName	A user group name.
hbaseNamespace	The name of the HBase namespace.
hostId	The canonical identifier for a host in Cloudera Manager. It is unique and immutable. For example: 3d645222-2f7e-4895-ae51-cd43b91f1e7a.
hostname	A hostname.
hregionName	The HBase region name. For example, 4cd887662e5c2f3cd5dd227bb03dd760.
hregionStartTimeMs	Milliseconds from UNIX epoch since Cloudera Manager monitoring started collecting metrics for the HBase region.

Name	Description
htableName	The name of an HBase table.
iface	A network interface name. For example, eth0.
logicalPartition	A Boolean indicating whether or not the disk is a logical partition. Applies to disk entity types.
mountpoint	A mount point name. For example, /var, /mnt/homes.
nameserviceName	The name of the HDFS nameservice.
ownerName	The owner username.
partition	A partition name. Applies to partition entity types.
path	A filesystem path associated with the time-series entity.
poolName	A pool name. For example, hdfs cache pool, yarn pools.
queueName	The name of a YARN queue.
rackId	A Rack ID. For example, /default.
roleConfigGroup	The role group that a role belongs to.
roleName	The role ID. For example, HBASE-1-REGIONSERVER-0b0ad09537621923e2b460e5495569e7.
roleState	The role state: BUSY, HISTORY_NOT_AVAILABLE, NA, RUNNING, STARTING, STOPPED, STOPPING, UNKNOWN
roleType	The role type: ACTIVITYMONITOR, AGENT, ALERTPUBLISHER, BEESWAX_SERVER, CATALOGSERVER, DATANODE, EVENTSERVER, FAILOVERCONTROLLER, HBASE_INDEXER, HBASERESTSERVER, HBASETHRIFTSERVER, HIVEMETASTORE, HIVESERVER2, HOSTMONITOR, HTTPFS, HUESERVER, IMPALAD, JOBHISTORY,JOBTRACKER, JOURNALNODE, KT_RENEWER, LLAMA, MASTER, NAVIGATOR, REGIONSERVER, SERVICEMONITOR, NAMENODE, NODEMANAGER, REPORTSMANAGER, SECONDARYNAMENODE, SERVER, SOLR_SERVER, SQOOP_SERVER, STATESTORE, TASKTRACKER.
rollup	The time-series store table rollup type.
schedulerType	The scheduler type associated with the pool service.
serviceDisplayName	The user-defined display name of a service entity.
serviceName	The service ID. To specify a service by its display name use the serviceDisplayName attribute.
serviceState	The service state: HISTORY_NOT_AVAILABLE, NA, RUNNING, STARTING, STOPPED, STOPPING, UNKNOWN
serviceType	The service type: ACCUMULO,FLUME, HDFS, HBASE, HIVE, HUE, IMPALA, KS_INDEXER, MAPREDUCE, MGMT, OOZIE,SOLR, SPARK,SQOOP,YARN, ZOOKEEPER.
solrCollectionName	The Solr collection name. For example, my_collection.
solrReplicaName	The Solr replica name. For example, my_collection_shard1_replica1.
solrShardName	The Solr shard name. For example, shard1.
systemTable	A boolean indicating whether the HBase table is a system table or not.
tableName	The name of a table.
userName	The name of the user.
version	The version of the cluster. The value can be any of the supported Cloudera Runtime major versions.

Time Series Entities and their Attributes

The following table shows the entities and associated attributes that can appear in the predicate ("where" clause) of a tsquery statement.

Entity	Attributes
All Roles	roleType, hostId, hostname, rackId, serviceType, serviceName
All Services	serviceName, serviceType, clusterId, version, serviceDisplayName, clusterDisplayName

Entity	Attributes
Agent	roleType, hostId, hostname, rackId, serviceType, serviceName, clusterId, version, agentName, serviceDisplayName, clusterDisplayName
Cluster	clusterId, version, clusterDisplayName
Directory	roleName, hostId, path, roleType, hostname, rackId, serviceType, serviceName, clusterId, version, agentName, hostname, clusterDisplayName
Disk	device, logicalPartition, hostId, rackId, clusterId, version, hostname, clusterDisplayName
File System	hostId, mountpoint, rackId, clusterId, version, partition, hostname, clusterDisplayName
Flume Channel	serviceName, hostId, rackId, roleName, flumeComponent, roleType, serviceType, clusterId, version, agentName, serviceDisplayName, clusterDisplayName
Flume Sink	serviceName, hostId, rackId, roleName, flumeComponent, roleType, serviceType, clusterId, version, agentName, serviceDisplayName, clusterDisplayName
Flume Source	serviceName, hostId, rackId, roleName, flumeComponent, roleType, serviceType, clusterId, version, agentName, serviceDisplayName, clusterDisplayName
HDFS Cache Pool	serviceName, poolName, nameserviceName, serviceType, clusterId, version, groupName, ownerName, serviceDisplayName, clusterDisplayName
HNamespace	serviceName, namespaceName, serviceType, clusterId, version, serviceDisplayName, clusterDisplayName
Host	hostId, rackId, clusterId, version, hostname, clusterDisplayName
HRegion	htableName, hregionName, hregionStartTimeMs, namespaceName, serviceName, tableName, serviceType, clusterId, version, roleType, hostname, roleName, hostId, rackId , serviceDisplayName, clusterDisplayName
HTable	namespaceName, serviceName, tableName, serviceType, clusterId, version, serviceDisplayName, clusterDisplayName
Network Interface	hostId, networkInterface, rackId, clusterId, version, hostname, clusterDisplayName
Rack	rackId
Service	serviceName, serviceType, clusterId, serviceDisplayName
Solr Collection	serviceName, serviceType, clusterId, version, serviceDisplayName, clusterDisplayName
Solr Replica	serviceName, solrShardName, solrReplicaName, solrCollectionName, serviceType, clusterId, version, roleType, hostId, hostname, rackId, roleName, serviceDisplayName, clusterDisplayName
Solr Shard	serviceName, solrCollectionName, solrShardName, serviceType, clusterId, version, serviceDisplayName, clusterDisplayName
Time Series Table	tableName, roleName, roleType, applicationName, rollup, path
User	userName
YARN Pool	serviceName, queueName, schedulerType

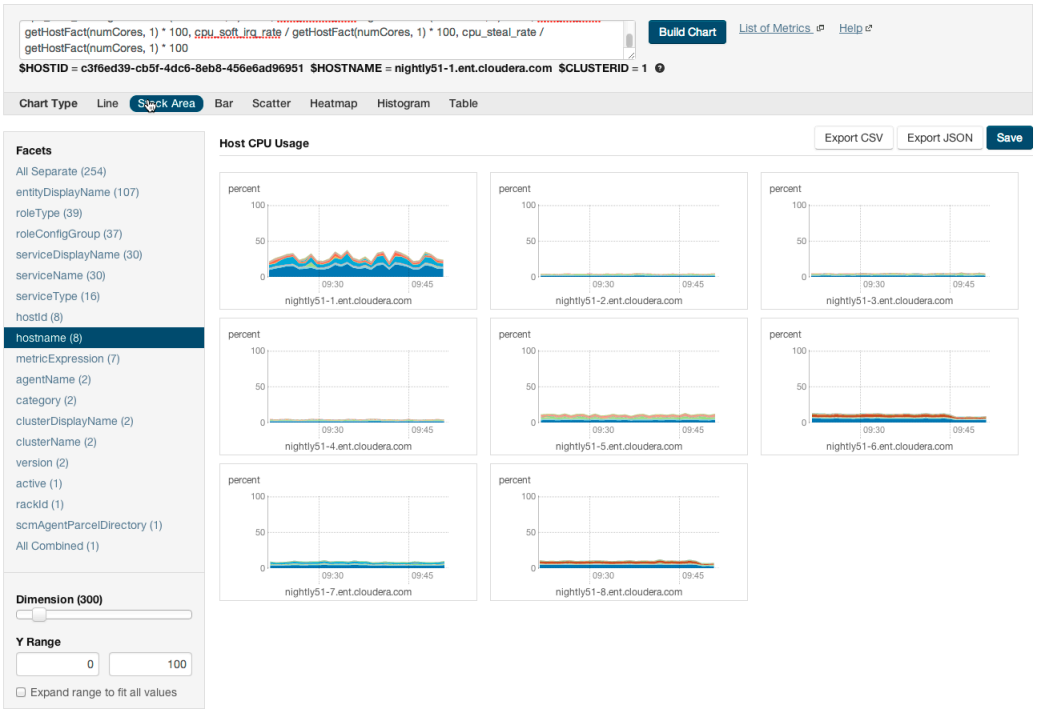
FAQ

The following topic covers frequently asked questions about charting time-series data.

How do I compare information across hosts?

1. Click Hosts in the top navigation bar and click a host link.
2. In the Charts pane, choose a chart, for example Host CPU Usage and select ⚙ and then Open in Chart Builder.
3. In the text box, remove the where entityName=$HOSTID clause and click Build Chart.
4. In the Facets list, click hostname to compare the values across hosts.

5. Configure the time scale, minimums and maximums, and dimension. For example:



How do I compare all disk IO for all the DataNodes that belong to a specific HDFS service?

Use a query of the form:

```
select bytes_read, bytes_written where roleType=DATANODE and ser
viceName=hdfs1
```

replacing hdfs1 with your HDFS service name. Then facet by metricDisplayName and compare all DataNode byte_reads and byte_writes metrics at once.

When would I use a derivative function?

Some metrics represent a counter, for example, bytes_read. For such metrics it is sometimes useful to see the rate of change instead of the absolute counter value. Use dt or dt0 derivative functions.

When should I use the dt0 function?

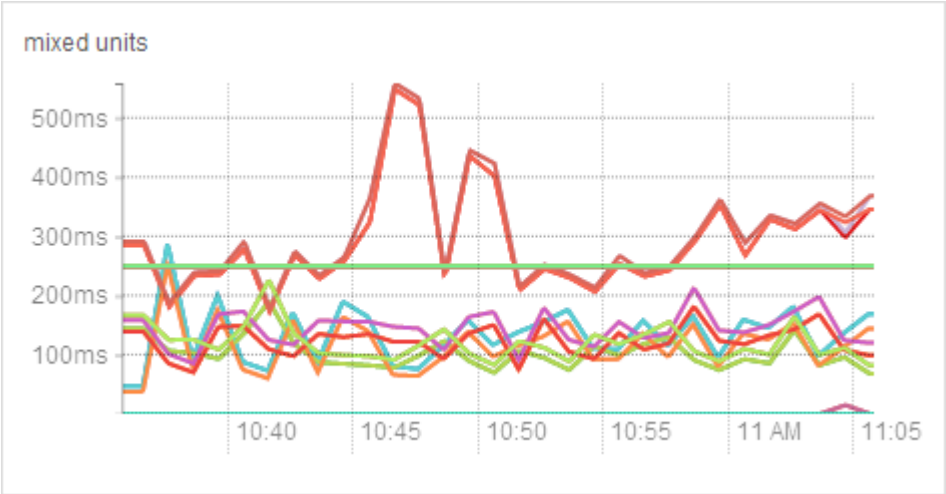
Some metrics, like bytes_read represent a counter that always grows. For such metrics a negative rate means that the counter has been reset (for example, process restarted, host restarted, and so on). Use dt0 for these metrics.

How do I display a threshold on a chart?

Suppose that you want to retrieve the latencies for all disks on your hosts, compare them, and show a threshold on the chart to easily detect outliers. Use the following query to retrieve the metrics and the threshold:

```
select await_time, await_read_time, await_write_time, 250 where
category=disk
```

Then choose All Combined (1) in the Facets list. The scalar threshold 250 will also be rendered on the chart:



I get the warning "The query hit the maximum results limit". How do I work around the limit?

There is a limit on the number of results that can be returned by a query. When a query results in more time-series streams than the limit a warning for "partial results" is issued. To circumvent the problem, reduce the number of metrics you are trying to retrieve or see the topic *Configuring Time-Series Query Results*.

You can use the rlike operator to limit the query to a subset of entities. For example, instead of

```
select await_time, await_read_time, await_write_time, 250 where
category=DISK
```

you can use

```
select await_time, await_read_time, await_write_time, 250 where
category=DISK and hostname rlike "host1[0-9]?.cloudera.com"
```

The latter query retrieves the disk metrics for ten hosts.

How do I discover which metrics are available for which entities?

- Type Select in the text box and then press Space or continue typing. Metrics matching the letters you type display in a drop-down list.
- Select Charts Chart Builder , click the question mark icon to the right of the Build Chart button and click the List of Metrics link
- Retrieve all metrics for the type of entity:

```
select * where roleType=DATANODE
```

Related Information

Grouping (Faceting) Time Series
Configuring Time-Series Query Results

Metric Aggregation

In addition to collecting and storing raw metric values, the Cloudera Manager Service Monitor and Host Monitor produce a number of aggregate metrics from the raw metric data.

Where a raw data point is a timestamp value pair, an aggregate metric point is a timestamp paired with a bundle of statistics including the minimum, maximum, average, and standard deviation of the data points considered by the aggregate.

Individual metric streams are aggregated across time to produce statistical summaries at different data granularities. For example, an individual metric stream of the number of open file descriptors on a host will be aggregated over time to the ten-minute, hourly, six-hourly, daily and weekly data granularities. A point in the hourly aggregate stream will include the maximum number of open file descriptors seen during that hour, the minimum, the average and so on. When servicing a time-series request, either for the Cloudera Manager UI or API, the Service Monitor and Host Monitor automatically choose the appropriate data granularity based on the time-range requested.

Cross-Time Aggregate Example

Consider the following fd_open raw metric values for a host:

```
9:00, 100 fds
9:01, 101 fds
9:02, 102 fds
. . .
9:09, 109 fds
```

The ten minutely cross-time aggregate point covering the ten-minute window from 9:00 - 9:10 would have the following statistics and metadata:

```
min: 100 fds
min timestamp: 9:00
max 109 fds
max timestamp 9:09
mean 104.5 fds
standard deviation: 3.02765 fds
count: 10 points
sample: 109 fds
sample timestamp: 9:09
```

The Service Monitor and Host Monitor also produce cross-entity aggregates for a number of entities in the system. Cross-entity aggregates are produced by considering the metric value of a particular metric across a number of entities of the same type at a particular time. For each stream considered, two metrics are produced. The first tracks statistics such as the minimum, maximum, average and standard deviation across all considered entities as well as the identities of the entities that had the minimum and maximum values. The second tracks the sum of the metric across all considered entities.

An example of the first type of cross-entity aggregate is the fd_open_across_datanodes metric. For an HDFS service this metric contains aggregate statistics on the fd_open metric value for all the DataNodes in the service. For a rack this metric contains statistics for all the DataNodes within that rack, and so on. An example of the second type of cross-entity aggregate is the total_fd_open_across_datanodes metric. For an HDFS service this metric contains the total number of file descriptors open by all the DataNodes in the service. For a rack this metric contains the total number of file descriptors open by all the DataNodes within the rack, and so on. Note that unlike the first type of cross-entity aggregate, this total type of cross-entity aggregate is a simple timestamp, value pair and not a bundle of statistics.

Cross-Entity Aggregate Example

Consider the following fd_open raw metric values for a set of ten DataNodes in an HDFS service at a given timestamp:

```
datanode-0, 200 fds
datanode-1, 201 fds
datanode-2, 202 fds
…
datanode-9, 209 fds
```

The cross-entity aggregate fd_open_across_datanodes point for that HDFS service at that time would have the following statistics and metadata:

```
min: 200 fds
```

```
min entity: datanode-0
max: 209 fds
max entity: datanode-9
mean: 204.5 fds
standard deviation: 3.02765 fds
count: 10 points
sample: 209 fds
sample entity: datanode-9
```

Just like every other metric, cross-entity aggregates are aggregated across time. For example, a point in the hourly aggregate of fd_open_across_datanodes for an HDFS service will include the maximum fd_open value of any DataNode in that service over that hour, the average value over the hour, and so on. A point in the hourly aggregate of total_fd_open_across_datanodes for an HDFS service will contain statistics on the value of the total_fd_open_across _datanodes for that service over the hour.

Presentation of Aggregate Data

Aggregate data points returned from the Cloudera Manager API appear as shown in this section.

A cross-time aggregate:

```
{
        "timestamp" : "2014-02-24T00:00:00.000Z",
        "value" : 0.014541698027508003,
        "type" : "SAMPLE",
        "aggregateStatistics" : {
          "sampleTime" : "2014-02-23T23:59:35.000Z",
          "sampleValue" : 0.0,
          "count" : 360,
          "min" : 0.0,
          "minTime" : "2014-02-23T18:00:35.000Z",
          "max" : 2.9516129032258065,
          "maxTime" : "2014-02-23T19:37:36.000Z",
          "mean" : 0.014541698027508003,
          "stdDev" : 0.17041289765265377
        }
      }
```

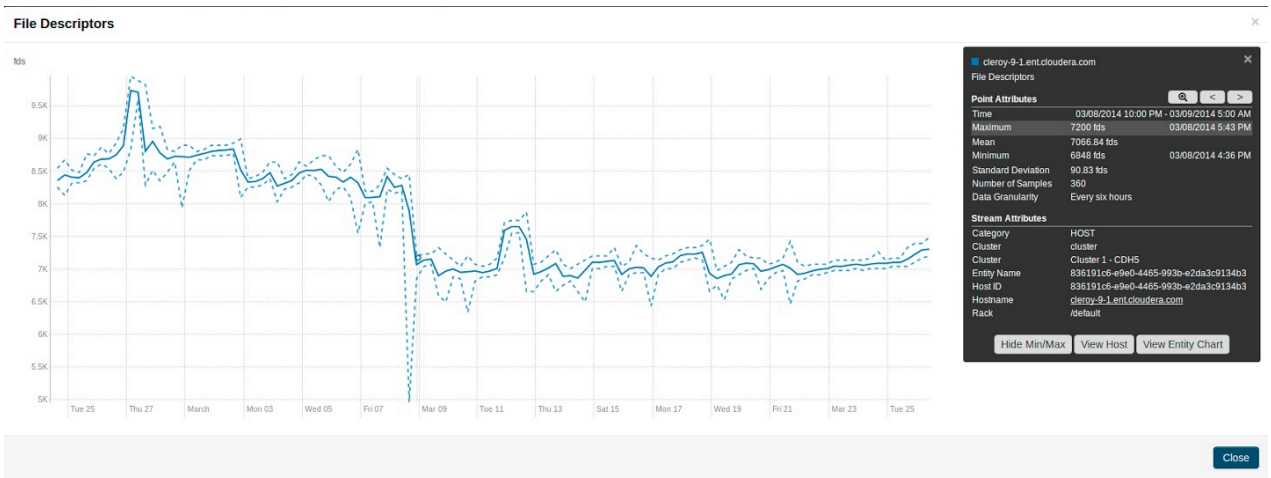
A raw cross-entity aggregate:

```
{
        "timestamp" : "2014-03-26T00:50:15.725Z",
        "value" : 3288.0,
        "type" : "SAMPLE",
        "aggregateStatistics" : {
          "sampleTime" : "2014-03-26T00:49:19.000Z",
          "sampleValue" : 7232.0,
          "count" : 4,
          "min" : 1600.0,
          "minTime" : "2014-03-26T00:49:42.000Z",
          "max" : 7232.0,
          "maxTime" : "2014-03-26T00:49:19.000Z",
          "mean" : 3288.0,
          "stdDev" : 2656.7549127961856,
          "crossEntityMetadata" : {
            "maxEntityDisplayName" : "cleroy-9-1.ent.cloudera.com",
            "minEntityDisplayName" : "cleroy-9-4.ent.cloudera.com",
            "numEntities" : 4.0
          }
        }
      }
```

A cross-time, cross-entity aggregate:

```
{
        "timestamp" : "2014-03-11T00:00:00.000Z",
        "value" : 3220.818863879957,
        "type" : "SAMPLE",
        "aggregateStatistics" : {
          "sampleTime" : "2014-03-10T22:28:48.000Z",
          "sampleValue" : 7200.0,
          "count" : 933,
          "min" : 1536.0,
          "minTime" : "2014-03-10T21:02:17.000Z",
          "max" : 7200.0,
          "maxTime" : "2014-03-10T22:28:48.000Z",
          "mean" : 3220.818863879957,
          "stdDev" : 2188.6143063503378,
          "crossEntityMetadata" : {
            "maxEntityDisplayName" : "cleroy-9-1.ent.cloudera.com",
            "minEntityDisplayName" : "cleroy-9-4.ent.cloudera.com",
            "numEntities" : 3.9787037037037036
          }
        }
      }
```

These differ from non-aggregate data points by having the aggregateStatistics structure. Note that the value field in the point structure will always be the same as the aggregteStatistics mean field. The Cloudera Manager UI presents aggregate statistics in a number of ways. First, aggregate statistics are made available in the hover detail and chart popover when dealing with aggregate data. Second, it is possible to turn on and turn off the display of minimum and maximum time-series streams in line charts of aggregate data. These streams are displayed using dotted lines and give a visual indication of the underlying metric values data range over the time considered, entities considered or both. These lines are displayed by default for single stream line charts of aggregate data. For all line charts this behavior can be turned on and turned off using the chart popover.



Accessing Aggregate Statistics Through tsquery

You can use the stats function to access aggregate statistics directly in tsquery.

For example, select stats(fd_open_across_datanodes, max) where category = service and serviceDisplayName = “my-hdfs-service” will return a single time-series stream containing the just the maximum statistic values from the fd_open_across_datanodes stream. The following statistics are available through the stats function: min, max, avg, std_dev, and sample.

Related Information

tsquery Language

Filtering Metrics

Metric Filters limit the amount of metric data reported to the Cloudera Manager server to avoid gaps in the reported data.

About this task

Metrics Filters allow you to limit the amount of metric data sent to the Cloudera Manager Service Monitor. In large clusters, some services, such as Kudu, send a high volume of non-essential metrics data to the Service Monitor, which can overload it, causing gaps in the data reported from these metrics in charts/dashboards and metrics queries, and potentially limiting the ability for Cloudera Manager to effectively monitor cluster health . To mitigate this problem, you can configure Metric Filters that limit the amount of data sent to the Service Monitor and Host Monitor. You can configure Metric Filters for any service deployed in a cluster.

You can configure the filter in several ways:

- Limit the collected metrics to only include those required for Health Tests.
- Limit the collected metrics to only include metrics used in the charts on the main page (‘Dashboard’) of the service.
- Include specific metrics - only the specified metrics will be collected for this service
- Exclude specific metrics - the specified metrics will not be collected for this service

Procedure

To configure a Metric Filter:

1. Log in to the Cloudera Manager Admin Console.
2. Navigate to the cluster where you want to add the filter.
3. Click ConfigurationMetric Filters.

   The Configuration page displays the Metric Filter parameter for all roles in the cluster.

   You can now choose whether to edit all the Metric Filters using the same values, or edit the filter for a specific role.
4. Do one of the following:
   - To configure the filter for a specific role, click the Edit Individual Values link and locate the parameter for the service. Follow the steps below.
   - To configure the same filter for all roles, edit the Default Group.
5. Select one of the following:
   - Include Only Health Test Metric Set
   - Include Only Default Dashboard Metric Set

6. To filter specific metrics:
   a) Select either Include or Exclude from the 'Include/Exclude Custom Metrics' drop-down list.
      If you selected either of the Metric sets, and select Include, the custom metrics will be added to the selected Metrics sets. If you select Exclude, they will be excluded from the selected sets.
   b) Click the "Plus" icon to add a metric.
   c) Enter the metric name.
   d) Click the "Plus" icon to add additional metrics.

   Note: You can find the names of specific metrics you would like to include or exclude either from the Cloudera Manager Metrics Reference (see link below) or from a chart that is displaying the required metrics in the Cloudera Manager Admin Console, as follows:

   a. Go to the Status page for a cluster or service.
   b. In a chart, click the gear icon and select Open in Chart Builder.

   The query displays at the top of the page. Metric names are part of the query. For example, the following query:

```
select dfs_capacity, dfs_capacity_used, dfs_capacity_used_non_hdfs
 where entityName=$SERVICENAME
```

   contains the following metric names:

   - dfs_capacity
   - Dfs_capacity_used
   - dfs_capacity_used_non_hdfs
7. Click Save Changes.
8. Refresh the cluster:
   a) Go the cluster's Status page.
   b) Click ActionsRefresh Cluster. This action does not restart the cluster.
      The cluster configuration is refreshed.

   Note: You can configure Cloudera Manager to automatically refresh the cluster after you update Metric configurations by enabling the Enable auto refresh for metric configurations parameter. This parameter is set at the role level, and, when set to true, Cloudera Manager refreshes only the role affected by a change to the metric filter or collection enable settings, not the entire cluster.

Results

Note: If you do not select any options, all metrics for the service will be sent to the Service Monitor and Host Monitor.

Related Information

Cloudera Manager Metrics Reference

Logs

The **Logs** page presents log information for Hadoop services, filtered by service, role, host, or search phrase as well log level (severity).

To configure logs, see the topic *Configuring Log Events*.

Related Information

Configuring Log Events

Viewing Logs

You can view logs that contain information such as warnings, errors, and more.

Procedure

1. Select Diagnostics Logs on the top navigation bar.
2. Click Search.
   The logs for all roles display. If any of the hosts cannot be searched, an error message notifies you of the error and the host(s) on which it occurred.

Logs List

Log results are displayed in a list.

The logs list contains the following columns:

- Host - The host where this log entry appeared. Clicking this link will take you to the **Host Status** page.
- Log Level - The log level (severity) associated with this log entry.
- Time - The date and time this log entry was created.
- Source - The class that generated the message.
- Message - The message portion of the log entry. Clicking View Log File displays the **Log Details** page, which presents a display of the full log, showing the selected message (highlighted) and the 100 messages before and after it in the log.

If there are more results than can be shown on one page (per the Results per Page setting you selected), Next and Prev buttons let you view additional results.

Related Information

Managing Hosts

Log Details

Filtering Logs

You filter logs by selecting a time range and specifying filter parameters.

About this task

You can use the Time Range Selector or a duration link (

) to set the time range. See for details. However, logs are, by definition, historical, and are meaningful only in that context. So the Time Marker, used to pinpoint status at a specific point in time, is not available on this page. The Now button (⏭) is available.

Procedure

1. Specify any of the log filter parameters:
   - Search Phrase - A string to match against the log message content. The search is case-insensitive, and the string can be a regular expression, such that wildcards and other regular expression primitives are supported.
   - Select Sources - A list of all the service instances and roles currently instantiated in your cluster. By default, all services and roles are selected to be included in your log search; the All Sources checkbox lets you select or

clear all services and roles in one operation. You can expand each service and limit the search to specific roles by selecting or clearing individual roles.
- Hosts - The hosts to be included in the search. As soon as you start typing a hostname, Cloudera Manager provides a list of hosts that match the partial name. You can add multiple names, separated by commas. The default is to search all hosts.
- Minimum Log Level - The minimum severity level for messages to be included in the search results. Results include all log entries at the selected level or higher. This defaults to WARN (that is, a search will return log entries with severity of WARN, ERROR, or FATAL only.
- Additional Settings
  - Search Timeout - A time (in seconds) after which the search will time out. The default is 20 seconds.
  - Results per Page - The number of results (log entries) to be displayed per page.

2. Click Search. The Logs list displays the log entries that match the specified filter.

Related Information

Time Line

Log Details

The **Log Details** page presents a portion of the full log, showing the selected message (highlighted), and messages before and after it in the log.

The **Log Details** page shows you:

- The host
- The role
- The full path and name of the log file you are viewing.
- Messages before and after the one you selected.

The log displays the following information for each message:

- Time - the time the entry was logged
- Log Level - the severity of the entry
- Source - the source class that logged the entry
- Log Message

You can switch to display only messages or all columns using the buttons.

In addition, from the Log Details page you can:

- View the log entries in either expanded or contracted form using the buttons to the left of the date range at the top of the log.
- Download the full log using the Download Full Log button at the top right of the page.
- View log details for a different host or for a different role on the current host, by clicking the Change... link next to the host or role at the top of the page. In either case this shows a pop-up where you can select the role or host you want to see.

Viewing the Cloudera Manager Server Log

To help you troubleshoot problems, you can view the Cloudera Manager Server log. You can view the logs in the **Logs** page or in specific pages for the log.

Viewing Cloudera Manager Server Logs in the Logs Page

You can view Cloudera Manager Server logs in the **Logs** page.

Procedure

1. Select DiagnosticsLogs on the top navigation bar.
2. Next to Sources, select the Cloudera Manager Server checkbox and deselect the other options.
3. Adjust the search criteria and click Search.

Related Information

Logs

Viewing the Cloudera Manager Server Log

You can access the Cloudera Manager Server Log through the Diagnostics menu or the server host.

Procedure

1. Select DiagnosticsServer Log on the top navigation bar.
2. Optionally, you can view the Cloudera Manager Server log at /var/log/cloudera-scm-server/cloudera-scm-server .log on the Server host.

Viewing the Cloudera Manager Agent Logs

To help you troubleshoot problems, you can view the Cloudera Manager Agent logs. You can view the logs in the **Logs** page or in specific pages for the logs.

Viewing Cloudera Manager Agent Logs in the Logs Page

You can view the Cloudera Manager Agent logs in the **Logs** page.

Procedure

1. Select DiagnosticsLogs on the top navigation bar.
2. Click Select Sources to display the log source list.
3. Uncheck the All Sources checkbox.
4. Click ▸ to the left of Cloudera Manager and select the Agent checkbox.
5. Click Search.

Related Information

Logs

Viewing the Cloudera Manager Agent Log

You can view the Cloudera Manager Agent log through the **Hosts** page.

Procedure

1. Click the Hosts tab.
2. Click the link for the host where you want to see the Agent log.
3. In the Details panel, click the Details link in the Host Agent field.
4. Click the Agent Log link.

   You can also view the Cloudera Manager Agent log at /var/log/cloudera-scm-agent/cloudera-scm-agent.log on the Agent hosts.

Managing Disk Space for Log Files

All CDH cluster hosts write out separate log files for each role instance assigned to the host. Cluster administrators can monitor and manage the disk space used by these roles and configure log rotation to prevent log files from consuming too much disk space.

Related Information

Managing Role Groups

Disk Space Requirements

For each role assigned to a host, you should generally provision 2GB of disk space for log files. This recommendation is based on the default values of configuration properties that set the maximum log file size (200MB) and the maximum number of files (10). To calculate the disk space required for each host, multiply the configured maximum size of the log file by the configured maximum number of logs. Perform this calculation for each role on a host and add them together. (Note that Gateway roles do not generate log files.)

To determine the roles assigned to each host, open the Cloudera Manager Admin Console and go to HostsAll Hosts and expand the list of roles in the Roles column.

Managing Log Files

To manage log file configurations for all role instances of a service:

1. Go to *Service Name*Configuration.
2. Select CategoryLogs.
3. Edit the logging parameters.
4. Click Save Changes.

Note: You can also manage these configurations using Role Groups, which you can use to configure similar hosts with the same configuration values.

There are the parameters you use to manage log files:

Table 9: Log File Properties

Property	Description	Default Value
Role Type Max Log Size	Maximum size for a log file before the log file rolls over into a new file.	200 MB
Role Type Maximum Log File Backups	Maximum number of rolled-over log files to retain.	10
Role Type Log Directory	The path to the directory where the log files are saved.	/var/log/*log_file_name*
Role Type Logging Threshold (not available for all roles)	Logging level to limit the number of entries saved in the log file.	Depends on the role.

Reports

The Reports page lets you create reports about the usage of HDFS in your cluster—data size and file count by user, group, or directory. It also lets you report on the MapReduce activity in your cluster, by user.

To display the Reports page, select Clusters*Cluster name*Reports.

For users with the Administrator role, the Search Files and Manage Directories button on the Reports page opens a file browser for searching files, managing directories, and setting quotas.

If you are managing multiple clusters, or have multiple nameservices configured (if high availability or federation is configured) there will be separate reports for each cluster and nameservice.

See the following pages for more details:

Directory Usage Report

The directory usage report allows you to browse the HDFS filesystem in a way that is similar to the HDFS File Browser. However, the Directory Usage Report also allows you to sort the listings and select multiple items and perform actions on them.

Minimum Required Role: BDR Administrator (also provided by Full Administrator and Cluster Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

You can also view the last access time, the last modified time of any file in a directory, and the total size of all files in the directory. This usage information is updated on an hourly basis.

You can customize the report by adding filters. A number of preconfigured filters are available, and you can create a custom filter.

Related Information

The File Browser

Accessing the Directory Usage Report

You can access the directory usage report through the Clusters menu or the HDFS File Browser.

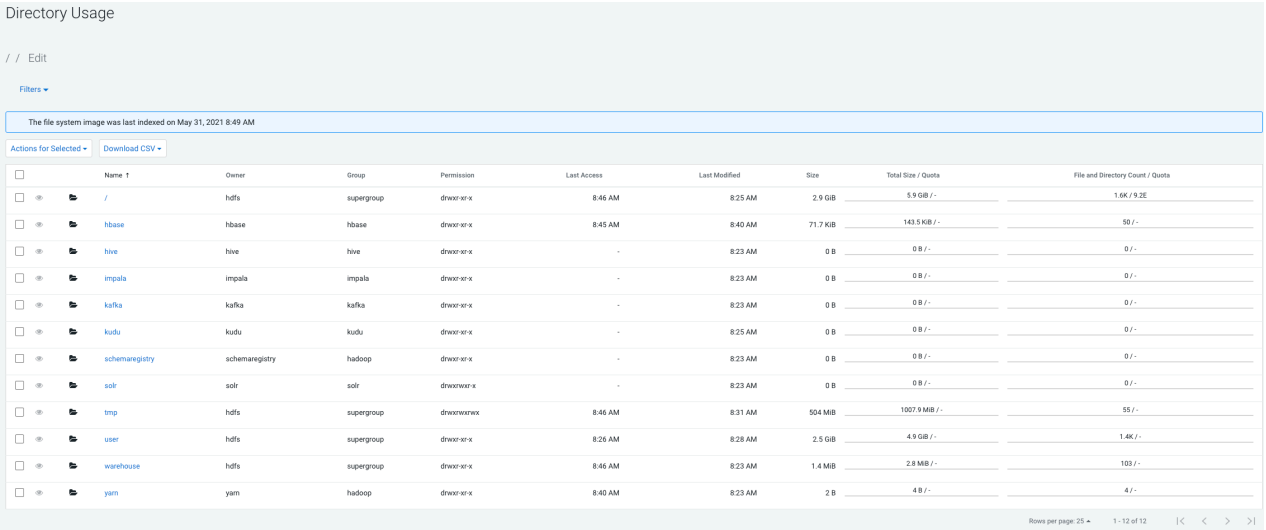
Procedure

1. Click Clusters *Cluster Name* Reports Directory Usage .
2. Optionally, you can access the report through the HDFS File Browser.
   a) Click Clusters *HDFS service* File Browser .
   b) Click the Directory Usage link located in the lower-right portion of the File Browser.

Using the Directory Usage Report

In the directory usage report, you can sort the display, view files and subdirectories in the directory, and more.

When you first open the report, the top level of the HDFS filesystem displays:



Click the Filters link to select options that limit the output by various criteria.

Click any column header to sort the display.

Click a directory name to view the files and subdirectories in the directory.

Click Download CSV to download a CSV version of the report. Select one of the following options:

- Top 1K Rows
- Top 10K Rows
- Top 100K Rows

Select one or more rows by checking the boxes on the left and then choose an action to perform on the selection from the Actions for Selected drop-down menu. You can select the following actions:

- Manage Quota – A dialog box opens in which you can set a quota for the number of files or disk space. These values are displayed in columns in the file listing.
- Include selected directories in disk usage reports – The selected directories appear in the disk usage reports.
- Exclude selected directories from disk usage reports – The selected directories do not appear in the disk usage reports.

Related Information

Disk Usage Reports

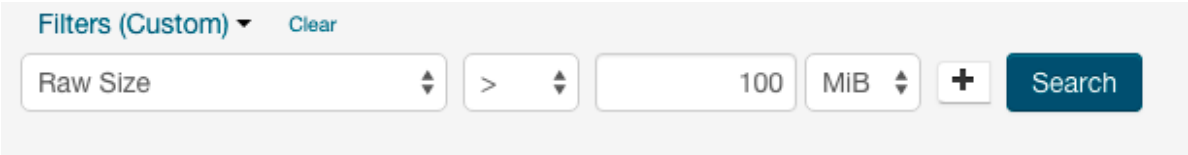
Filters

You can use filters to limit the display and to search for files.

Procedure

1. To apply filters to the directory usage report, click the Filters drop-down menu near the top of the page and select one of the following preconfigured filters:
   - Large Files
   - Large Directories
   - By Specific Owner
   - By Specific Group
   - Old Files
   - Old Directories
   - Files with Low Replication
   - Overpopulated Directories
   - Directories with Quotas
   - Directories Watched
2. To modify any of these filters, click the Customize link and select new criteria. Click Clear to revert to the preconfigured criteria for the filter.
3. Click the Search button to display the report with the filters applied.

4. You can also select Custom from the Filters drop-down menu to create a report in which you define the criteria. To create a custom report:
   a) Select any of the following criteria from the drop-down menu on the left:
      - Filename
      - Owner
      - Group
      - Path
      - Last Modified
      - Size
      - Diskspace Quota
      - Namespace Quota
      - Last Access
      - File and Directory Count
      - Replication
      - Parent
      - Raw Size
      - Size With Snapshot
      - Total Size
   b) Select an operator from the drop-down menu.
   c) Enter a value and units of measure for the comparison.
   d) Select the units of measure for the comparison from the drop-down menu. (Some criteria do not require units of measure.)
   e) Click the ✚ icon to add additional criteria.
   f) Click the Search button to display the directory usage report with the custom filter applied.



   The report changes to display the result of applying the filter. A new column, Parent is added that contains the full path to each file or subdirectory.

Disk Usage Reports

There are two types of disk usage reports: Current Disk Usage By Directory and Historical Disk Usage By Directory.

Procedure

1. To use these reports, select one or more directories to watch by clicking the 👁 icon for the directory.
2. Alternatively, you can select multiple directories, and then click Actions for Selected Include selected directories in disk usage reports .

Disk Usage Reports

The disk usage reports show HDFS disk usage statistics, either current or historical, by user, group, or directory.

The By Directory reports display information about the directories in the Watched list, so if you are not watching any directories there will be no results found for these reports. You can also specify which directories to watch by selecting them from the Directory Usage Report.

Related Information

Designating Directories to Include in Disk Usage Reports

Directory Usage Report

Viewing Current Disk Usage by User, Group, or Directory

The current disk usage reports show "current" disk usage in both chart and tabular form.

The data for these reports comes from the fsimage kept on the NameNode, so the data in a report will be only as current as when the last checkpoint was performed. Typically the checkpoint interval is (by default) once per hour, but if checkpoints are not being performed as frequently, the disk usage report may not be up to date. The disk usage report displays the current usage and does not account for deleted files that only exist in snapshots. These files are included in the usage information when you run the du command.

To create a disk usage report:

- Click the report name (link) to produce the resulting report.

Each of these reports show:

Bytes	The logical number of bytes in the files, aggregated by user, group, or directory. This is based on the actual files sizes, not taking replication into account.
Raw Bytes	The physical number of bytes (total disk space in HDFS) used by the files aggregated by user, group, or directory. This does include replication, and so is actually Bytes times the number of replicas.
File and Directory Count	The number of files aggregated by user, group, or directory.

Bytes and Raw Bytes are shown in IEC binary prefix notation (1 GiB = 1 * 230).

The directories shown in the Current Disk Usage by Directory report are the HDFS directories you have selected to include in directory usage reports. You can make these selections in the HDFS File Browser, or on the **Directory Usage** report.

The report data is also shown in chart format:

- Move the cursor over the graph to highlight a specific period on the graph and see the actual value (data size) for that period.
- You can also move the cursor over the user, group, or directory name (in the graph legend) to highlight the portion of the graph for that name.
- You can right-click within the chart area to save the whole chart display as a single image (a .PNG file) or as a PDF file. You can also print to the printer configured for your browser.

Related Information

Designating Directories to Include in Disk Usage Reports

Using the Directory Usage Report

Viewing Historical Disk Usage by User, Group, or Directory

You can use the historical disk usage reports to view disk usage over a time range you define. You can have the usage statistics reported per hour, day, week, month, or year.

Procedure

1. To create the report, click the report name (link) to produce the initial report. This generates a report that shows Raw Bytes for the past month, aggregated daily.
2. To change the report parameters, select the Start Date and End Date to define the time range of the report.
3. Select the Graph Metric you want to graph: bytes, raw bytes, or files and directories count.
4. In the Report Period field, select the period over which you want the metrics aggregated. The default is Daily. This affects both the number of rows in the results table, and the granularity of the data points on the graph.

5. Click Generate Report to produce a new report.

   As with the current reports, the report data is also presented in chart format, and you can use the cursor to view the data shown on the charts, as well as save and print them.

   For weekly or monthly reports, the Date indicates the date on which disk usage was measured.

   The directories shown in the Historical Disk Usage by Directory report are the HDFS directories you have set as watched directories.

Related Information

Designating Directories to Include in Disk Usage Reports

Downloading Reports as CSV and XLS Files

Any report can be downloaded to your local system as an XLS file (Microsoft Excel 97-2003 worksheet) or CSV (comma-separated value) text file.

About this task

To download a report, do one of the following:

Procedure

- From the main page of the Report tab, click CSV or XLS link next to in the column to the right of the report name
- From any report page, click the Download CSV or Download XLS buttons.
  Either of these opens the Open file dialog box where you can open or save the file locally.

Activity, Application, and Query Reports

The **Reports** page contains links for displaying metrics on the following types of activities in your cluster:

About this task

The Reports page contains links for displaying metrics on the following types of activities in your cluster:

- Disk usage
- MapReduce jobs
- YARN applications
- Impala queries
- HBase tables and namespaces

Procedure

1. To view the Reports page, click Clusters*ClusterName*Reports. You can generate a report to view aggregate job activity per hour, day, week, month, or year, by user or for all users.
2. Click the Start Date and End Date fields and choose a date from the date control.
3. In the Report Period drop-down, select the period over which you want the metrics aggregated. Default is Daily.
4. Click Generate Report.

   For weekly reports, the Date column indicates the year and week number (for example, 2013-01 through 2013-52). For monthly reports, the Date column indicates the year and month by number (2013-01 through 2013-12).

The File Browser

The File Browser tab on the HDFS service page lets you browse and search the HDFS namespace and manage your files and directories.

Minimum Required Role: BDR Administrator (also provided by Full Administrator and Cluster Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

The File Browser page initially displays the root directory of the HDFS file system in the gray panel at the top and its immediate subdirectories below. Click any directory to drill down into the contents of that directory or to select that directory for available actions.

Searching Within the File System

When you search within the file system, you can select from custom search criteria such as filename, owner, file size, and more.

To search the file system, click Custom report in the Reports section. The file and directory listings are taken from the fsimage stored on the NameNode, so the listings will be only as current as the last checkpoint. Typically the checkpoint interval is (by default) once per hour, but if checkpoints are not being performed as frequently, the listings may not be up to date.

To search the file system:

1. From the HDFS service page, select the File Browser tab.
2. Click Choose and do one of the following:
   - Select a predefined query. Depending on what you select, you may be presented with different fields to fill in or different views of the file system. For example, selecting Size will provide a choice of arithmetic operators and fields where you provide the size to be used as the search criteria.
     a. Select a property in the Choose... drop-down.
     b. Select an operator.
     c. Specify a value.
     d. Click ✚ to add another criteria (all of which must be satisfied for a file to be considered a match) and repeat the preceding steps.
3. Click the Generate Report button to generate a custom report containing the search results.

If you search within a directory, only files within that directory will be found. For example, if you browse /user and do a search, you might find /user/foo/file, but you will not find /bar/baz.

Enabling Snapshots

To enable snapshots for an HDFS directory and its contents, see the topic *Managing HDFS Snapshots*.

Setting Quotas

To set quotas for an HDFS directory and its contents, see *Setting HDFS Quotas*.

Designating Directories to Include in Disk Usage Reports

You can designate directories to include in a disk usage report.

Procedure

1. To add or remove directories from the directory-based Disk Usage reports, navigate through the file system to see the directory you want to add. You can include a directory at any level without including its parent.
2. Check the checkbox Include this directory in Disk Usage reports.

   As long as the checkbox is checked, the directory appears in the usage reports. To discontinue inclusion of the directory in Disk Usage reports, clear the checkbox.

Downloading HDFS Directory Access Permission Reports

For each HDFS service, you can download a report that details the HDFS directories a group has permission to access.

Before you begin

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

Procedure

1. In the Cloudera Manager Admin Console, click Clusters*ClusterName*Reports.
2. In the Directory Access by Group row, click CSV or XLS.
   The Download User Access Report pop-up displays.
3. In the pop-up, type a group and directory.
4. Click Download. A report of the selected type will be generated containing the following information – path, owner, permissions, and size – for each directory contained in the specified directory that the specified group has access to.

Cluster Support Tokens using Cloudera Manager

Cloudera Manager now provides Suppport Tokens for every cluster that it manages. When you create support cases on the Cloudera Support portal, you may be asked to provide a Support Token. The Support Token can be copied from the Cloudera Manager GUI into the Cloudera Support portal. The Support Tokens allow Cloudera to more accurately track usage and node-counts, particularly in the absence of a diagnostic bundle.

Ensure that you have the global visibility permission to access the Security Token and call the APIs.

There are two APIs available.

One API displays a ClusterName-SupportToken list that includes all the clusters and their SupportToken.

```
# ClouderaManagerResource
GET /cm/clusterSupportTokens
# Should be a list of tokens
{
"items": [
{
"clusterName": "Cluster 1",
"displayName": "Cluster 1",
"clusterSupportToken": "IWCnF5gY12JwnAAAAAAAAAAAAAAA8qj8"
},
{
"clusterName": "Cluster 2",
"displayName": "Cluster 2",
"clusterSupportToken": "PAtxTDK+50VxPAEAUAIAAAAAAAAZMe7E"
}
]
}
```

The second API takes ClusterName as the parameter and displays the corresponding SupportToken. The token provided is unique for each cluster.

```
# ClustersResource
GET /clusters/{clusterName}/clusterSupportToken
# should be one token
{
"clusterName": "Cluster 1",
"displayName": "Cluster 1",
"clusterSupportToken": "IWCnF5gY12JwnAAAAAAAAAAAAAAA8qj8"
}
```

Generating and displaying a unique Support Token for each managed cluster includes a truncated combination of Cloudera Managed Server UUID and Cluster ID. The Cluster ID will be used by Cloudera Support to validate the legitimacy and uniqueness of the environment against customer entitlements without the need to send a diagnostic bundle.

For seeing the support token, as a user, you must have global visibility (UserAuthorities.USER authority). To know your role, navigate to UI Administration > Users & Roles > Roles for reference. Support token is not visible without an access to UserAuthorities.USER authority.

Note: Enabling diagnostic bundler is the simplest way to ensure that assets are kept up to date.

1. Log in to Cloudera Manager.
2. On the left navigation pane, click Support.
3. Click Support Tokens. Support Token dialog box appears.
4. Copy the Support Token.
5. Log in to the Cloudera Support portal.
6. Click Create Case.
7. If prompted, you can paste the Support Token into the case.

Sending Usage and Diagnostic Data to Cloudera

Cloudera Manager collects anonymous usage information and takes regularly-scheduled snapshots of the state of your cluster and automatically sends them anonymously to Cloudera. This helps Cloudera improve and optimize Cloudera Manager.

Minimum Required Role: Cluster Administrator (also provided by Full Administrator) This feature is not available when using Cloudera Manager to manage Data Hub clusters.

If you have a Cloudera Enterprise license, you can also trigger the collection of diagnostic data and send it to Cloudera Support to aid in resolving a problem you may be having.

Configuring a Proxy Server

Steps to configure a proxy server.

To configure a proxy server thorough which data and parcels are downloaded to and uploaded from the Cloudera Manager Server, do the following:

1. Select AdministrationSettings.
2. Click the Network category.
3. Configure proxy properties.
4. Enter a Reason for change, and then click Save Changes to commit the changes.

Managing Anonymous Usage Data Collection

Cloudera Manager sends anonymous usage information using Google Analytics to Cloudera. The information helps Cloudera improve Cloudera Manager. By default, anonymous usage data collection is enabled. You can disable this property.

Procedure

1. Select Administration Settings .
2. Under the Other category, set the Allow Usage Data Collection property.
3. Enter a Reason for change, and then click Save Changes to commit the changes.

Diagnostic Data Collection

To help with solving problems when using Cloudera Manager on your cluster, Cloudera Manager collects diagnostic data on a regular schedule, and automatically sends it to Cloudera.

By default Cloudera Manager is configured to collect this data weekly and to send it automatically. Cloudera analyzes this data and uses it to improve the software. If Cloudera discovers a serious issue, Cloudera searches this diagnostic data and notifies customers who might encounter problems due to the issue. You can schedule the frequency of data collection on a daily, weekly, or monthly schedule, or disable the scheduled collection of data entirely. You can also send a collected data set manually.

Automatically sending diagnostic data requires the Cloudera Manager Server host to have Internet access, and be configured for sending data automatically. If your Cloudera Manager Server does not have Internet access, you can manually send the diagnostic data.

Automatically sending diagnostic data might fail sometimes and return an error message of "Could not send data to Cloudera." To work around this issue, you can manually send the data to Cloudera Support.

What Data Does Cloudera Manager Collect?

Cloudera Manager collects and returns a significant amount of information about the health and performance of the cluster. It includes:

- Up to 1000 Cloudera Manager audit events: Configuration changes, add/remove of users, roles, services, and so on.
- One day's worth of Cloudera Manager events: This includes critical errors Cloudera Manager watches for and more.
- Data about the cluster structure which includes a list of all hosts, roles, and services along with the configurations that are set through Cloudera Manager. Where passwords are set in Cloudera Manager, the passwords are not returned.
- Cloudera Manager license and version number.
- Current health information for hosts, service, and roles. Includes results of health tests run by Cloudera Manager.
- Heartbeat information from each host, service, and role. These include status and some information about memory, disk, and processor usage.
- The results of running Host Inspector.
- One day's worth of Cloudera Manager metrics. If you are using a Cloudera trial version, host metrics are not included.
- A download of the debug pages for Cloudera Manager roles.
- For each host in the cluster, the result of running a number of system-level commands on that host.
- Logs from each role on the cluster, as well as the Cloudera Manager server and agent logs.
- Which parcels are activated for which clusters.
- Whether there's an active trial, and if so, metadata about the trial.

- Metadata about the Cloudera Manager Server, such as its JMX metrics, stack traces, and the database or host it's running with.
- HDFS or Hive replication schedules (including command history) for the deployment.
- Impala query logs.
- Cloudera Data Science Workbench collects aggregate usage data by sending limited tracking events to Google Analytics and Cloudera servers. No customer data or personal information is sent as part of these bundles.
- A copy of selected tables and columns from the Cloudera Manager database. This data is not collected by default. To configure inclusion of database tables, see :

Log support in Cloudera Manager for ECS cluster

You can manage your ECS cluster with additional log information that is provided as part of the Cloudera Manager support bundle.

Specific logs are captured when you install Embedded Container Service (ECS) on your Private Cloud Base cluster. Data that is captured include the ECS installation logs, RKE specifics pertaining to the ECS cluster, and Longhorn support bundle. For more information, see

How to generate a Longhorn Support Bundle

1. Login to the Cloudera Manager UI > Click Support > Send Diagnostic Data.
2. Select the radio button - Collect Diagnostic Data Only and click Collect Diagnostic Data.
3. Click the Download button when finished and expand the .zip file.

ECS installation logs

The ECS installation log details can be located under the folder named process-stdout-stderr/ which includes the stdout.log and /stderr.logstdout.log and /stderr.log files related to the ECS installation process. The included files follow the following pattern:

- *-docker-*
- *-ecs-*

RKE specific logs

RKE information pertaining to the ECS cluster is located in the rke-logs-(clusterName).tgz file. This file can be used to debug reported issues with the ECS cluster. Expanding rke-logs-(clusterName).tgz file results in a directory called "rke-logs-[your clustername].

Under the expanded directory is a file named {hostName}-{date}.tar.gz, general details about RKE logs are available.

Specific information related to other embedded entities within the {hostName}-{date}.tgz directories are detailed in the table and they are available under the rk2/ path.

Entity	Contains
Pod Logs (Under podlogs/)	Namespaces: kube-system, kube-public, cattle-system, cattle-alerting, cattle-logging, cattle-pipeline ingress-nginx cattle-prometheus, istio-system, longhorn-system, cattle-global-data, fleet-system, fleet-default, rancher-operator-system, cattle-monitoring-system, cattle-logging-system, cdp, and vault-system
Kubectl (under kubectl/)	List of nodes, pods, services, node descriptions, and version details.

Entity	Contains
Certificates (under certs/)	Agent and Server certificates.
Directories (under directories/)	• RKE2 agent directory • RKE2 server manifest directory • RKE2 server TLS directory
Crictl (under crictl/)	Multiple crictl command outputs

Within the longhorn-logs-{hostName}.gz file, the Longhorn log information is available. Expand the longhorn-logs-{hostName}.gz file.

Under the expanded directory, two new directories are available.

Specifically contains logs that contain files related to pod logs for all the workloads.

YAML - Files for Longhorn and Kubernetes

Configuring the Frequency of Diagnostic Data Collection

By default, Cloudera Manager collects diagnostic data on a weekly basis. You can change the frequency to daily, weekly, monthly, or never. If you are a Cloudera Enterprise customer and you set the schedule to never, you can still collect and send data to Cloudera on demand. If you are a Cloudera Enterprise customer and you set the schedule to never, data is not collected or sent to Cloudera.

Procedure

1. Select AdministrationSettings.
2. Under the Support category, click Scheduled Diagnostic Data Collection Frequency and select the frequency.
3. To set the day and time of day that the collection will be performed, click Scheduled Diagnostic Data Collection Time and specify the date and time in the pop-up control.
4. Enter a reason for the change and click Save Changes.

   You can see the current setting of the data collection frequency by viewing SupportScheduled Diagnostics: in the main navigation bar.

Configuring collection of Cloudera Manager table data

You can configure Cloudera Manager to include selected data from the Cloudera Manager database in Diagnostic Bundles.

About this task

Collection of Cloudera Manager table data is disabled by default. Including this data in the Diagnostic Bundles can help Cloudera Customer Support to diagnose problems. You can configure the maximum size of this data, and you can specify any Cloudera Manager database table to be excluded from the bundle.

Before you begin

Minimum Required User role Full Administrator.

Procedure

1. Log in to the Cloudera Manager Admin Console.
2. Go to AdministrationSettings.

3. To enable automatic collection of diagnostic data, select Send Diagnostic Data Automatically.

   If you do not select this, you can manually trigger sending the diagnostic data.
4. Select the Support category.
5. Select the Collect Cloudera Manager database copy with Diagnostic Data Bundle parameter.
6. To exclude specific database tables, add their names in the Cloudera Manager Database Table Names to skip property.

   Cloudera has excluded all known sensitive data from the table copies, including usernames, passwords, hashes and other data. If you have concerns about sensitive data being included, contact Cloudera support, who can assist you by providing the names of tables to exclude.
7. Enter the maximum size (in megabytes) of the copied database tables to be included in the Diagnostic Bundle.

   The default value is 1024 MB.

Results

Collection of table data is enabled. Note that the time required to collect all the data for a Diagnostic Bundles may increase by up to 8 minutes.

Specifying the Diagnostic Data Directory

You can configure the directory where collected diagnostic data is stored.

Procedure

1. Select AdministrationSettings.
2. Under the Support category, set the Diagnostic Data Bundle Directory to a directory on the host running Cloudera Manager Server. The directory must exist and be enabled for writing by the user cloudera-scm. If this field is left blank, the data is stored in /tmp.
3. Enter a reason for the change and click Save Changes.

Redaction of Sensitive Information from Diagnostic Bundles

By default, Cloudera Manager redacts known sensitive information from inclusion in diagnostic bundles. Cloudera Manager uses a set of standard rules to redact passwords and secrets. You can add additional redaction rules using regular expressions to specify data you want to be redacted from the bundles.

Procedure

1. To specify redaction rules for diagnostic bundles, go to Administration Settings and search for the Redaction Parameters for Diagnostic Bundles parameter.
   The edit screen for the property displays.
2. To add a new rule, click the ✚ icon. You can add one of the following:
   - Credit Card numbers (with separator)
   - Social Security Card numbers (with separator)
   - Email addresses
   - Custom rule (You must supply values for the Search and Replace fields, and optionally, the Trigger field.)
3. To modify a new rule, click the ❯ icon.

4. Edit the redaction rules as needed. Each rule has a description field where you can enter free text describing the rule and you can modify the following three fields:
   - Search - Regular expression to compare against the data. For example, the regular expression \d{4}[^\w]\d{4}[^\w]\d{4}[^\w]\d{4} searches for a credit card number pattern. Segments of data that match the regular expression are redacted using the Replace string.
   - Replace - String used to redact (obfuscate) data, such as a pattern of Xs to replace digits of a credit card number: XXXX-XXXX-XXXX-XXXX.
   - Trigger - Optional simple string to be searched before applying the regular expression. If the string is found, the redactor searches for matches using the Search regular expression. Using the Trigger field improves performance: simple string matching is faster than regular expression matching.
5. To delete a redaction rule, click the ➖ icon.
6. Click Save Changes.

Disabling the Automatic Sending of Diagnostic Data from a Manually Triggered Collection

If you do not want data automatically sent to Cloudera after manually triggering data collection, you can disable this feature. The data you collect will be saved and can be downloaded for sending to Cloudera Support at a later time.

Procedure

1. Select AdministrationSettings.
2. Under the Support category, uncheck the box for Send Diagnostic Data to Cloudera Automatically.
3. Enter a Reason for change, and then click Save Changes to commit the changes.

Manually Triggering Collection and Transfer of Diagnostic Data to Cloudera

To troubleshoot specific problems, or to re-send an automatic bundle that failed to send, you can manually send diagnostic data to Cloudera. Perform the following steps on the Cloudera Manager UI.

Procedure

1. Log in to the Cloudera Manager Admin Console.
2. You can perform either of the following steps to navigate to Send Diagnostic Data:
   - Go to Clusters and then select Send Diagnostic Data.
   - Go to Support and then select Send Diagnostic Data.
3. Change the System Identifier property.
4. Under the Other category, set the System Identifier property and click Save Changes.
5. Fill in or change the information here as appropriate:
   - Optionally, you can improve performance by reducing the size of the data bundle that is sent. Click Restrict log and metrics collection to expand this section of the form. The three filters, Host, Service, and Role Type, allow you to restrict the data that will be sent. Cloudera Manager will only collect logs and metrics for roles that match all three filters.
   - Select one of the following under Data Selection:
     - Select By Target Size to manually set the maximum size of the bundle. Cloudera Manager populates the End Time based on the setting of the Time Range selector. You should change this to be a few minutes after you observed the problem or condition that you are trying to capture. The time range is based on the timezone of the host where Cloudera Manager Server is running.

- Select By Date Range to manually set the Start Time and End Time to collect the diagnostic data. Click the Estimate button to calculate the size of the bundle based on the start and end times. If the bundle is too large, narrow the selection using the start and end times or by selecting additional filters.
- If you have a support ticket open with Cloudera Support, include the support ticket number in the field provided.

6. Depending on whether you have disabled automatic sending of data, do one of the following:

- Click Collect and Upload Diagnostic Data to Cloudera Support. A Running Commands window shows you the progress of the data collection steps. When these steps are complete, the collected data is sent to Cloudera.
- Click Collect Diagnostic Data only. A Command Details window shows you the progress of the data collection steps.

  a. In the Command Details window, click Download Result Data to download and save a zip file of the information.

  b. Send the data to Cloudera Support by doing one of the following:

  - Send the bundle using a Python script:

    1. Download the phone_home script.

    2. Copy the script and the downloaded data file to a host that has Internet access.

    3. Run the following command on that host:

    ```
    python phone_home.py --file downloaded data file
    ```

  - Attach the bundle to the SFDC case. Do not rename the bundle as this can cause a delay in processing the bundle.
  - Contact Cloudera Support and arrange to send the data file.

Troubleshooting Cluster Configuration and Operation

This section contains solutions to some common problems that prevent you from using Cloudera Manager and describes how to use Cloudera Manager log and notification management tools to diagnose problems.

Solutions to Common Problems

The table below describes solutions to common cluster configuration problems.

Symptom	Reason	Solution
Cloudera Manager		
The Cloudera Manager service will not be running as it exited in an unusual manner. Running service cloudera-scm-server status will print following message "cloudera-scm-server dead but pid file exists". The Cloudera Manager Server log file /var/log/cloudera-scm-server/cloudera-scm-server.log will have a stacktrace with "java.lang.OutOfMemoryError" logged.	Out of memory.	Examine the heap dump that the Cloudera Manager Server creates when it runs out of memory. The heap dump file is created in the /tmp directory, has file extension .hprof and file permission of 600. Its owner and group will be the owner and group of the Cloudera Manager server process, normally cloudera-scm:cloudera-scm.
You are unable to start service on the Cloudera Manager server, that is, service cloudera-scm-server start does not work and there are errors in the log file located at /var/log/cloudera-scm-server/cloudera-scm-server.log	The server has been disconnected from the database or the database has stopped responding or has shut down.	Go to /etc/cloudera-scm-server/db.properties and make sure the database you are trying to connect to is listed there and has been started.

Symptom	Reason	Solution
Logs include APPARENT DEADLOCK entries for c3p0.	These deadlock messages are cause by the c3p0 process not making progress at the expected rate. This can indicate either that c3p0 is deadlocked or that its progress is slow enough to trigger these messages. In many cases, progress is occurring and these messages should not be seen as catastrophic.	There are a variety of ways to react to these log entries. • You may ignore these messages if system performance is not otherwise affected. Because these entries often occur during slow progress, they may be ignored in some cases. • You may modify the timer triggers. If c3p0 is making slow progress, increasing the period of time during which progress is evaluated stop the log entries from occurring. The default time between Timer triggers is 10 seconds and is configurable indirectly by configuring maxAdministrativeTaskTime. • You may increase the number of threads in the c3p0 pool, thereby increasing the resources available to make progress on tasks.
Starting Services		
After you click the Start button to start a service, the Finished status does not display. This may not be merely a case of the status not getting displayed. It could be for a number of reasons such as network connectivity issues or subcommand failures.	The host is disconnected from the Server, as will be indicated by missing heartbeats on the Hosts tab.	• Look at the logs for the service for causes of the problem. • Restart the Agents on the hosts where the heartbeats are missing.
	Subcommands failed resulting in errors in the log file indicating that either the command timed out or the target port was already occupied	• Look at the log file at /var/log/cloudera-scm-server/cloudera-scm-server.log for more details on the errors. For example, if the port is already occupied you should see an "Address in use" error. • Go to the Hosts > Status tab. Click the Name of the host you want to inspect. Now go to the Processes tab and check the Stdout/Stderr logs to diagnose the cause of the failure. For example, if any binaries are missing or if Java could not be found.
After you click Start to start a service, the Finished status displays but there are error messages. The subcommands to start service components (such as JobTracker and one or more TaskTrackers) do not start.	A port specified in the Configuration tab of the service is already being used in your cluster. For example, the JobTracker port is in use by another process.	Enter an available port number in the port property (such as JobTracker port) in the Configuration tab of the service.
	There are incorrect directories specified in the Configuration tab of the service (such as the log directory).	Enter correct directories in the Configuration tab of the service.
Job is Failing	No space left on device.	One approach is to use a system monitoring tool such as Nagios to alert on the disk space or quickly check disk space across all systems. If you do not have Nagios or equivalent you can do the following to determine the source of the space issue: In the JobTracker Web UI, drill down from the job, to the map or reduce, to the task attempt details to see which TaskTracker the task executed and failed on due to disk space. For example: http://*JTHost*:50030/taskdetails.jsp?tipid=*TaskID* . You can see on which host the task is failing in the Machine column. In the NameNode Web UI, inspect the % used column on the NameNode Live Nodes page: http://*namenode*:50070/dfsnodelist.jsp?whatNodes=LIVE.
Send Test Alert and Diagnose SMTP Errors		

Symptom	Reason	Solution
You have enabled sending alerts from the Cloudera Manager Admin Console, however, Cloudera Manager does not seem to be sending any alerts. Using the Send Test Alert link under Administration Alerts shows success even though you do not receive an alert email.	There is possibly a mismatch of protocol or port numbers between your mail server and the Alert Publisher. For example, if the Alert Publisher is sending alerts to SMTPS on port 465 and your mail servers are not configured for SMTPS, you wouldn't receive any alerts.	Use the following steps to make changes to the Alert Publisher configuration: **1.** In the Cloudera Manager Admin Console, click the Cloudera Management Service. **2.** Click the Configuration tab. **3.** Select Scope Alert Publisher . **4.** Click the Main category. **5.** Change Alerts: Mail Server Protocol to smtp (or smtps). **6.** Click the Ports and Addresses category and change Alerts: Mail Server TCP Port to 25 (or to 465 for SMTPS) **7.** Enter a Reason for change, and then click Save Changes to commit the changes. **8.** Restart the Alert Publisher.

Related Information

maxAdministrativeTaskTime

numHelperThreads

Logs and Events

For information about problems, check the logs and events.

- Logs present log information for services, filtered by role, host, or keywords as well log level (severity).
- The topic *Viewing the Cloudera Manager Server Log* contains information on the server and host agents.
- The Events tab lets you search for and display events and alerts that have occurred within a selected time range filtered by service, hosts, or keywords.

Related Information

Logs

Viewing the Cloudera Manager Server Log

Events

Alerts