Cloudera Runtime 7.2.16

Using Data Analytics Studio

Date published: 2020-02-28
Date modified: 2023-01-11

https://docs.cloudera.com/

Legal Notice

Contents

Compose queries

You can write and edit queries using the query composer.

Procedure

1. Select a database from the left panel.

   The Tables column displays all the tables in the database.
2. Search for the required tables, if needed.

   Click on a table name to view the columns within the table. Use the Filter field to enter text to further refine your search for the required column.
3. Enter the query in the worksheet tab.

   If your database contains more than or equal to 10,000 columns, press Ctrl + Spacebar on your keyboard to enable the auto-complete pop-up while you type the query.

   The auto-complete pop-up appears as you type if the number of columns in your database is less than 10,000.

   Each worksheet is identified by a unique name and you can add worksheets using the plus icon. As you start entering the query, the intuitive composer suggests SQL commands, keywords, and table columns according to the query.
4. Perform the desired operation on the query:
   - Click Execute to run the query. Alternatively, you can use the keyboard shortcut to execute the query: press Ctrl + Enter on Windows and control + return on XOS. Make sure that you press the keyboard shortcut while you are in the query editor.
   - Click Save As to save the worksheet with a different name.
   - Click Visual Explain to view the query details in the form of a block diagram.
   - Select Show Results if you want to view the results of the query.
   - Select Download Results to download the query results.
   - Click Saved tab to view saved queries and edit them.

Manage queries

Search for existing queries, and refine your search results based on various filters. You can also save the search for future use.

DAS uses a simple SQL query language that enables users familiar with SQL to query the data.

Searching queries

You can search for queries and see the list of queries that have been searched. You can refine your search results on the basis of parameters such as status of the query, queue to which the query belongs, the user of the query, tables read for the query, and tables written for the query, and execution modes.

Procedure

1. Enter your search query.
2. Click the time range drop down to select the from and to dates. You can also click to select one of the quick ranges provided in the list.
3. In the Refine tab, click the plus icon to further refine your search result. All options to refine your search appear. Select the required parameters and click Apply.

4. In the Actions column, click the pencil icon to open and edit the query in Composer view.
5. In the Actions column, click the i icon to view the query details.

Refining query search using filters

You can further refine your query search results using various filters.

Procedure

1. Click + in the Refine tab.
2. Select the filters from the list by clicking each filter.
3. Click Apply.

Note: The total number of items in every facet, such as User, Queue, Tables Read, etc. is displayed in parentheses next to the facet title in the Refine section. For optimum resource utilization, DAS does not refresh the result set every time you filter items. For example, if there are three users: hive, admin, and admin1, and you select only admin and click Apply, DAS does not show User (1). It still displays User (3).

Saving the search results

You have an option to save the filters with which you refined the queries, along with the result set for future use. It is easy to load, optionally modify, and run queries from the saved query list.

Click the save icon to save your search and the results. The saved query search names get listed under the SEARCHES dropdown list.

Compare queries

You can compare two queries to know how each query is performing in terms of speed and cost effectiveness. DAS compares various aspects of the two queries, based on which you can identify what changed between the executions of those two queries, and you can debug performance-related issues between different runs of the same query.

About this task

The query comparison report provides you a detailed side-by-side comparison of your queries, including recommendations for each query, metadata about the queries, visual explain for each query, query configuration, time taken at each stage of the query execution, and Directed Acyclic Graphs (DAG) information and DAG counters.

To compare two queries:

Procedure

1. Sign in to DAS and click Queries.
2. Under ACTIONS, click the Compare button to select a query.

   The selected query is displayed on the comparison toolbar.
3. Next, select the query that you want to compare by clicking the Compare button.

   The selected query is displayed on the comparison tool bar next to the query that you initially selected.
4. Click COMPARE.

   The comparison report is displayed.
5. To remove any query from the comparison toolbar, click x.
6. To change the order of the queries that are being compared, click the Swap queries button on the comparison toolbar.

View query details

For a given query, you can view the optimization recommendations, query details, the visual representation of the explain plan, Hive configurations, the time taken to compile, build and run the query, and the DAG information.

To view the query details, you can either click on a particular query or the i icon under the Actions column on the **Queries** page.

Viewing the query recommendations

The **Recommendations** tab provides information about how you can improve the performance of the query and optimize its cost.

Viewing the query details

The **Query Details** tab provides information such as, the Hive query ID of the query, the user who executed the query, the start time, the end time, the total time taken to run the query, the tables that were read and written, the application ID, the DAG ID(s), the session ID, the thread ID, and the queue against which the query was run.

Click the Edit button in the **Qury Details** section to edit and rerun the query.

Viewing the visual explain for a query

The **Visual Explain** tab provides a graphical representation of the query execution plan. The explain plan can be read from right to left. It provides details about every stage of the query operation.

Viewing the Hive configurations for a query

The **Configs** tab provides the Hive configuration details for the query. You can search for a particular configuration by entering the configuration name in the search field.

Viewing the query timeline

The **Timeline** tab shows the time taken by every stage of the query execution.

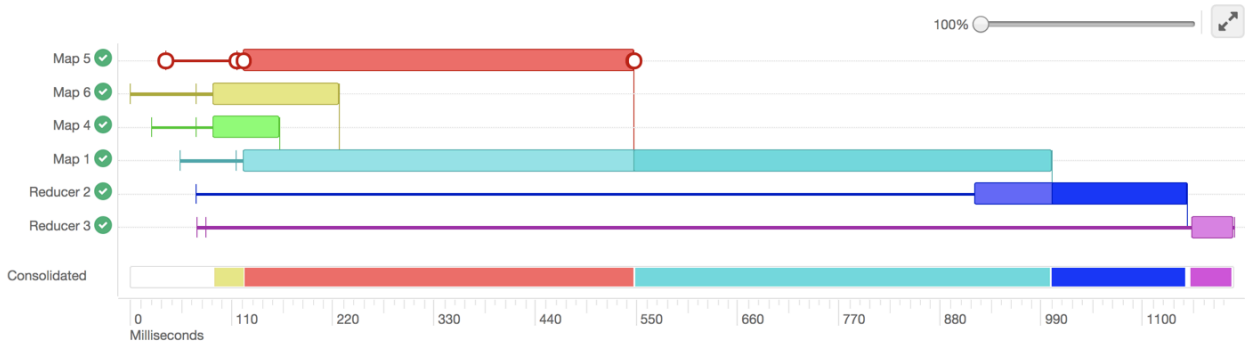
The first stage is Pre-execution and DAG construction. It is executed on the Hive engine. It constitutes the time taken to compile, parse, and build the DAG for the next phase of the query execution. In the next stage of query execution, the DAG generated in Hive is submitted to Tez engine for execution. The DAG Runtime shows the time taken by the Tez engine to run the DAG. In the post-execution stage, the HDFS files are moved or renamed.

Viewing the task-level DAG information

The DAG Info tab shows the Directed Acyclic Graph of the vertices against time. Each mapping and reducing task is a vertex.

The following image shows the DAG swimlane:

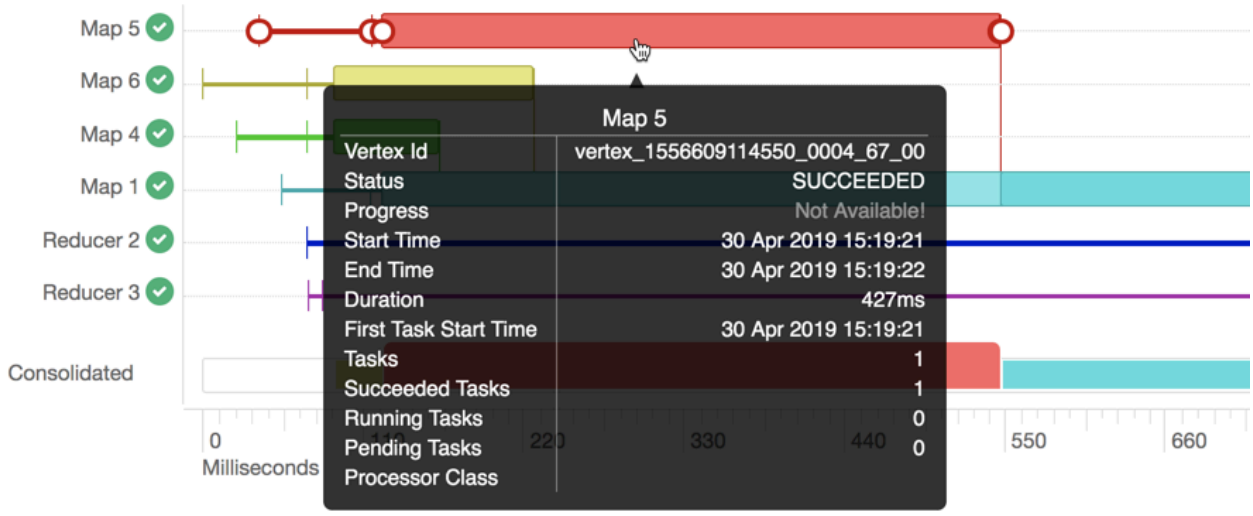
Figure 1: DAG swimlane in DAS



Each horizontal bar of the swimlane represents the total time taken by the vertex to complete. The vertical lines indicate the time when the vertex initialized, the time when the vertex started, the time when the first task started, the time when the last task completed, and the time when the vertex finished its execution. When you mouseover the vertical line, the bubble displays the stage of the vertex execution and provides a timestamp.

To know more about a particular vertex, hover the mouse pointer anywhere on the horizontal bar, as shown in the following image:

Figure 2: DAG swimlane: Viewing information about a vertex

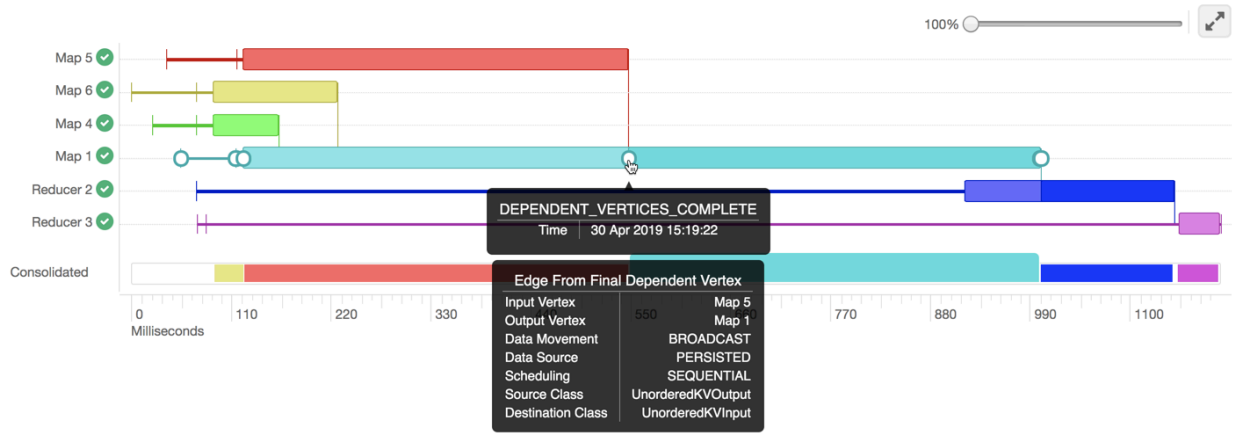


The following details can help you to view the time taken by a particular task and debug the query:

- Vertex ID: It is a unique identifier for a particular vertex.
- Status: Indicates whether the query executed successfully or not.
- Progress: Indicates the progress of the vertex, usually for long-running queries.
- Start time: Indicates when a particular vertex started.
- End time: Indicates when the particular vertex ended.
- Duration (in milliseconds): Indicates the total time taken by the vertex to complete its execution.
- First task start time: Indicates when the first task within that vertex started its execution.
- Tasks: Indicates the total number to tasks executed in a particular vertex.
- Succeeded tasks: Indicates the number of tasks that were executed successfully within that vertex.
- Running tasks: Indicates the tasks that are still running.
- Pending tasks: Indicates the tasks that have not yet started their execution.
- Processor class: It is the Hive processor for Tez that forms the vertices in Tez and processes the data. For example, org.apache.hadoop.hive.ql.exec.tez.ReduceTezProcessor, org.apache.hadoop.hive.ql.exec.tez.MapTezProcessor.

The vertical lines connecting two vertices denote the dependency of a vertex on another vertex, as shown in the following image:

Figure 3: DAG swimlane showing dependent vertices

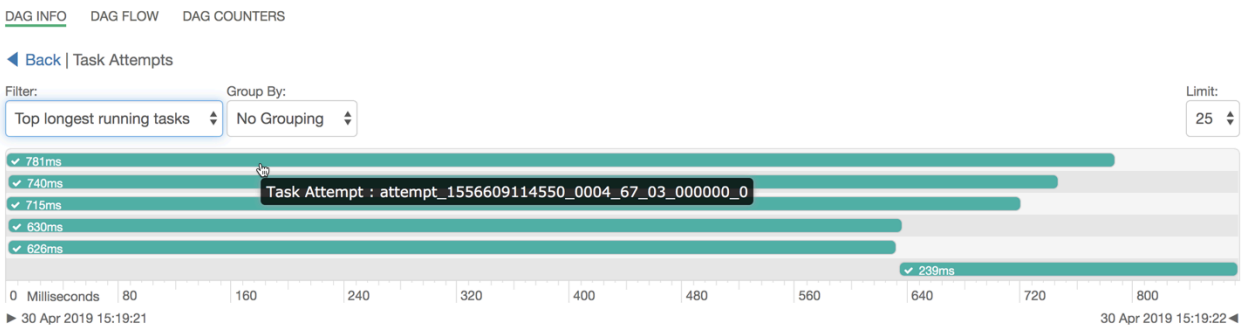


In the above example, Map 1 depends on the results of Map 5. Map 1 will finish its execution only when Map 5 finishes its execution successfully. Similarly, Reducer 2 depends on Map 1 to complete its execution.

The consolidated timeline shows the percentage of time each vertex took to complete executing. You can increase or decrease the scale of the timeline axis by moving the slider on the top right corner of the **DAG Info** section.

To further debug, click anywhere on the horizontal bar. This takes you to the **Task Attempts** section. The **Task Attempts** section shows the number of times a certain task was attempted. Each task attempt has a unique task attempt ID, as shown in the following image:

Figure 4: DAG info: Viewing task attempt



You can filter the result set by:

- Top longest running tasks: Used to filter the task that took the most time to complete.
- Errored tasks: Used to filter tasks that stopped because of some error.
- Tasks which started last: Used to filter the tasks which started late.

You can also group the result set either by tasks, containers in which the tasks were running, or the nodes on which the tasks were running.

For a query has more than one DAG ID associated with it, you can view its DAG by selecting the DAG ID from the dropdown.

DAS also allows you to download detailed logs of a task attempt. The log provides information about the container in which the task was running. To download the logs, click anywhere on the task attempt. On the **Details** pop-up, click Open log in new tab.

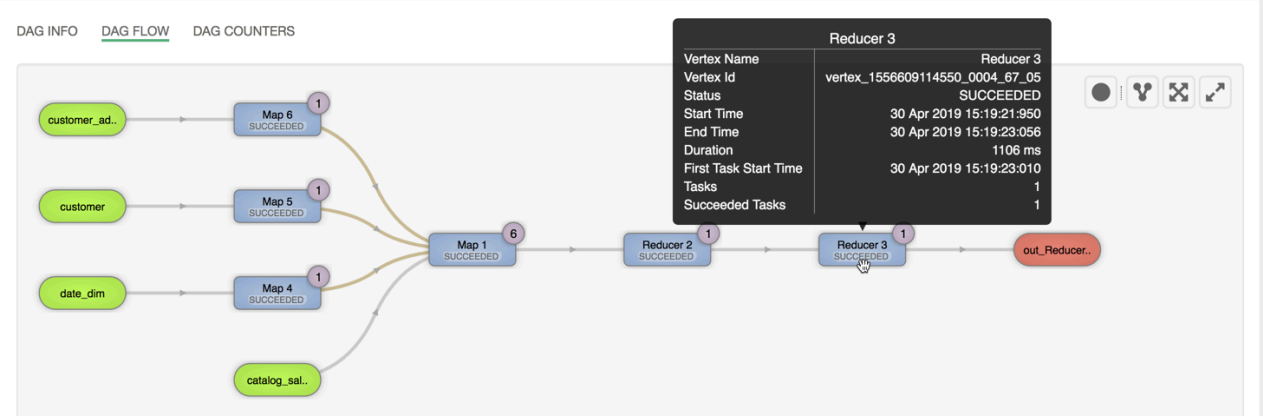
Note: Currently, the task-level logs are disabled for LLAP.

Viewing the DAG flow

The **DAG Flow** tab displays the Directed Acyclic Graph that is created by the Hive engine in the form of a flowchart.

The following image shows the DAG flow in the form of a flowchart:

Figure 5: DAG flowchart



Here, the inputs to vertices Map 4, Map 5, Map 6, and Map 1 are the tables displayed in green boxes. Next, Map 1 depends on the result set generated by Map 4, Map 5, and Map 6. Map 1 then passes its result as an input to Reducer 2. When the Reducer 2 finishes its execution, the results are passed on to Reducer 3. Reducer 3 is the last vertex in the DAG flow. After the Reducer 3 successfully completes its execution, the query output is written to a file in HDFS.

There are a few options to change the layout of the DAG flow. You can hide the input and the output nodes to view only the task vertices by clicking the Toggle source/sink visibility icon. You can switch between the horizontal and vertical orientation by clicking the Toggle orientation icon.

Viewing the DAG counters

The DAG counters provide a way to measure the progress or the number of operations that occur within a generated DAG. Counters are used to gather statistics for quality control purposes or problem diagnosis.

The DAG counters provide details, such as:

- Number of bytes read and written
- Number of tasks that initiated and ran successfully
- Amount of CPU and memory consumed

Viewing the Tez configurations for a query

The DAG Configurations tab in the **DAG Info** section on the **Query Details** page provides the Tez configuration details for the query that has a DAG associated with it. You can search for a particular configuration by entering the configuration name in the search field.

Enabling the DAS Event Processor

Data Analytics Studio (DAS) is installed when you create a Data Hub cluster with the Data Engineering template. In 7.2.16, you can run queries using DAS, but you cannot see the query history because the Event Processor has been disabled by default. You can enable the DAS Event Processor by setting the das.event-pipeline.enabled property to true in Cloudera Manager.

Procedure

1. Log in to CDP Management Console as an Administrator.
2. Go to Data Hub Environment Cloudera Manager Clusters Data Analytics Studio Configuration and add the following configuration in the Data Analytics Studio Eventprocessor Advanced Configuration Snippet (Safety Valve) for conf/props/eventprocessor_extra.properties field:

   Key: das.event-pipeline.enabled

   Value: true
3. Click Save Changes.
4. Restart the DAS service.

Manage databases and tables

DAS enables you to completely manage and administer databases and tables. You can create and drop databases and tables, manage columns and partitions, edit tables, view storage information, view detailed information about the databse and the tables within it, view table and column statistics, and preview tables.

Using the Database Explorer

From the Database Explorer, you can select a database to manage it, create a new database, or drop a database.

Once you select a database, the tables list gets refreshed to list all the tables from the selected database.

You can perform the following actions using the Database Explorer:

- Select a database
- Create a database
- Drop a database
- Search for a table within a database
- Create a table

Searching tables

After selecting a database, the tables tab is refreshed to list all the tables in the database.

To search for a table, start by entering the name of the table in the Search box. The list gets refreshed as you enter your search string to narrow the list down to your desired table.

Click the Refresh icon to refresh the list of tables.

Managing tables

Using the Database Explorer UI, you can view information about table columns, partitions, storage, and metadata, as well as preview a few rows from the table. This enables you to ensure that your table contains exactly the information that you intend it to.

For each table, the following tabs provide respective details:

- Columns: You can view the details of each column of the table. You can also search for columns using the Search box.
- Partitions: You can view the details of columns that are partitions in the table. You can also search from the list of columns.
- Storage Information: You can view the storage information such as input format, output format, if the table is compressed, number of buckets, buck columns, and so on.
- Detailed Information: You can view details such as the name of the database, the owner of the table, the created time, the last accessed time, table type, and so on.
- Statistics: You can view the table statistics and column statistics.
- Data Preview: You can preview a few rows from the table.

Creating tables

You can create a new table or upload an existing table to add it to a particular database.

Procedure

1. On the Database Explorer tab, click the + icon in the Tables section.

   The **Create Table** screen appears.
2. Enter the name of the table.
3. Click Add New Column to add columns to the table.
4. For each column, specify the following detail:
   a) Enter the name of the column.
   b) Select the data type of the column from the drop down list.
   c) Click Advanced to add advanced details of each column.
5. Click Advanced tab of the **Create Table** page.
   a) Select Transactional if you want the table to be a transactional table.
   b) Click Add Location if you want to specify a location for the table.
   c) Select a file format from the drop down list of Add File Format section.
   d) Click Add Row Format to specify details for the rows.
6. Click Properties to add key-value properties of the table.
7. Click Create.

Uploading tables

You can upload existing tables in CSV, JSON, or XML format to create new tables in the database. You can use the uploaded tables as other tables, run queries, and generate reports on them like other tables.

Procedure

1. Click Upload Table.

2. Enter the file format details in the Select File Format section.
   a) File type: Select the file type from the drop down list.
   b) Field Delimiter: This appears if you select CSV as the file format. Enter a delimiter character such as comma.

      Note: Pick a delimiter that is not part of the column data.

   c) Escape Character: This appears if you select CSV as the file format. Enter an escape character such as \.
   d) Quote Character: This appears if you select CSV as the file format. Enter a quote character.
   e) If the first row is header, click to select Is first row header checkbox.
   f) If the file contains endlines, click to select Contains endlines checkbox.
3. Select the source of the table file.
   a) Upload from HDFS: Select this option to upload files from HDFS.
   b) Upload from Local: Select this option to upload files from the local node.

      Note: Uploading tables from Amazon S3 is not supported in the CDP 1.2 release of Cloudera Runtime.

   The table details automatically appear as preview. Click Preview if the table details do not load automatically. DAS updates the name of the table, names of the columns, and the data types of each column.
4. Modify the names of the table and columns as needed.
5. Click the Advanced tab to specify the following additional options:
   a) Select Transactional if you want the table to be a transactional table.
   b) Click Add Location if you want to specify a location for the table.
   c) Select a file format from Add File Format.
   d) Click Add Row Format to specify details for the rows.
6. Click Properties to add key-value properties of the table.
7. Click Create.

Editing tables

You can edit tables to add new columns, update existing columns, and make changes to the properties of the table.

Procedure

1. On the **Database Explorer** tab, select a database.
2. From the list of tables, select the table that you want to edit.
3. From the Table view, click the Actions menu.
4. Click Edit.
5. Edit the details of existing columns.
6. Click Add New Column to add columns to the table.
7. For each column, specify the following details:
   a) Enter the name of the column.
   b) Select the data type of the column from the drop down list.
   c) Click Advanced to add more details for each column.
8. Click the Advanced tab on the **Create Table** page to specify the following additional options:
   a) Select Transactional if you want the table to be a transactional table.
   b) Click Add Location if you want to specify a location for the table.
   c) Select a file format from the drop down list of Add File Format section.
   d) Click Add Row Format to specify details for the rows.
9. Click Properties to add or edit key-value properties of the table.

10. Click Edit to save the changes you made to the table.

Deleting tables

You can delete a table using the Database Explorer.

Procedure

1. On the Database Explorer tab, select the database.
2. From the list of tables, select the table that you need to delete.
3. From the Table view, click the Actions menu.
4. Click Delete.
5. In the dialog box that appears, click Confirm.

Managing columns

Using the Columns tab, you can view the details of each column of the table.

You can also search for columns using the Search box.

To search for a column, enter the name of the column in the search box and click Regex Search.

Managing partitions

On the Partitions tab, you can view the details of columns that are partitions in the table.

You can also search from the list of columns.

To search for a column that is a partition, enter the name of the column in the search box and click Regex Search.

Viewing storage information

Using the Storage Information tab, you can view the storage information of the table such as input format, output format, if the table is compressed, number of buckets, number of columns, and more details.

To search for particular details, enter the name in the search box and click Regex Search.

Viewing detailed information

Using the Detailed Information tab, you can view details such as the name of the database, the owner of the table, the created time, the last accessed time, table type, and more details.

To search for particular details, enter the name in the search box and click Regex Search.

Viewing table and column statistics

On the Statistics tab, you can view table statistics and column statistics.

To view the statistics of each column, click Show. The following details are displayed:

- Min
- Max
- Number of Nulls
- Distinct Count

- Average Column Length
- Max Column Length
- Number of True
- Number of False

If a message appears that the statistics might be stale, select Include Columns and click Recompute at the top of the page.

Previewing tables using Data Preview

On the Data Preview tab, you can preview a few rows from the table.

Manage reports

As a database administrator, you can view which columns and tables are used for joins and make changes to the data layout to optimize the performance of the query.

Note: The view reports feature is not available in the DAS-Lite version.

Viewing the Read and Write report

The Read and Write report displays a list of tables belonging to the selected database along with the read count, write count, and other details for each table.

To view the Read and Write report, click Reports Read and Write Report . The **Relations** page is displayed.

The left side lists the tables within the selected database along with the read count, write count, and so on. The right side shows these details in the form of an entity relationship diagram.

You can switch between the databases using the Database dropdown menu.

To further refine your search based on time, select the time period from the dropdown. After you select the time period, the corresponding “to” and “from” dates representing the selection are displayed along with the time zone. The time zone that is displayed is that of the DAS server.

For each table, select one of the following counts from the drop down list of counts:

- Projection Count
- Aggregation Count
- Filter Count
- Join Count

Viewing the Join report

The Join report provides you the details of joins between all the tables in a database.

To view the join report, click Reports Join Report . You can change the database from the dropdown list to view the join diagram for that database.

You can switch between the databases using the Database dropdown menu.

To refine your search based on time, select the time period from the dropdown. After you select the time period, the corresponding “to” and “from” dates representing the selection are displayed along with the time zone. The time zone that is displayed is that of the DAS server.

You can further refine your join diagram using the following join algorithms:

- Join
- Hash Join
- Merge Join
- Lateral View Join
- Sorted Merge

The diagram gets refreshed based on your selection.

If you hover over a particular column, then all the joins with the selected column are highlighted. If you hover over a particular connection line, then the number of times the two columns were joined (also known as the join count) is displayed. The join count that is displayed is for the time period that you select.

For every table present on the Join report, DAS displays the database in which that table belongs. If you select a particular database from the dropdown menu, then the Join report predominantly displays all the joins originating from the tables within the selected database to the tables belonging to other databases.

Disabling the reporting feature

Generating reports in Data Analytics Studio (DAS) make database calls and may keep them open, utilizing resources. Because DAS is deprecated and you may not be actively using DAS, you can disable the reporting feature using Cloudera Manager.

Procedure

1. Log in to Cloudera Manager as an Administrator.
2. Go to Clusters DAS service Configuration and add the following in the Data Analytics Studio Eventprocessor Advanced Configuration Snippet (Safety Valve) for conf/props/eventprocessor_extra.properties field:

   Key: reporting.enabled

   Value: false

   To enable the reporting feature, set reporting.enabled to true.
3. Click Save Changes.
4. Restart the DAS service.