Cloudera Streaming Analytics 1.3.0

SQL Stream Job Lifecycle

Date published: 2019-12-17
Date modified: 2021-03-25

CLOUDERA

https://docs.cloudera.com/

Legal Notice

Contents

Running SQL Stream jobs

Every time you execute an SQL statement in the SQL Stream console, it becomes a job and runs on the deployment as a Flink job. You can manage the running jobs using the Jobs tab on the UI.

About this task

There are two logical phases to run a job:

1. Parse: The SQL is parsed and checked for validity and then compared against the virtual table schema(s) for correct typing and key/columns.
2. Execution: If the parse phase is successful, a job is dynamically created, and runs on an open slot on your cluster. The job is a valid Flink job.

Before you begin

- Make sure that you have registered a data source.
- Make sure that you have created a Virtual Table Source.

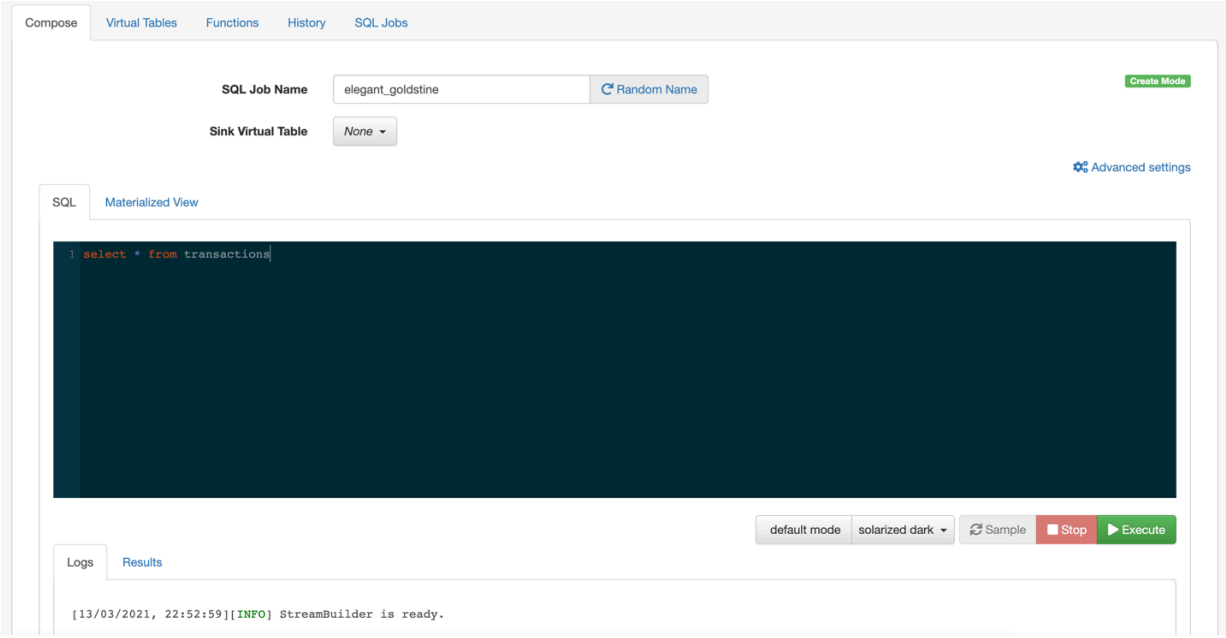
Procedure

1. Go to your cluster in Cloudera Manager.
2. Click on SQL Stream Builder from the list of Services.
3. Click on SQLStreamBuilder Console.
   The Streaming SQL Console opens up in a new window.
4. Provide a name for the SQL job.
   a) Optionally, you can click on the Random Name button to generate a name for the SQL job.
5. Select a Virtual Table Sink.
   a) Optionally, you can leave the sink to None.

   Note: The Sink Virtual Table is an optional argument to a job that specifies the destination for the continuous results. If you select None, the results are not sent to a sink, but sampled to the screen or to a materialized view.

6. Add a SQL query to the SQL window.

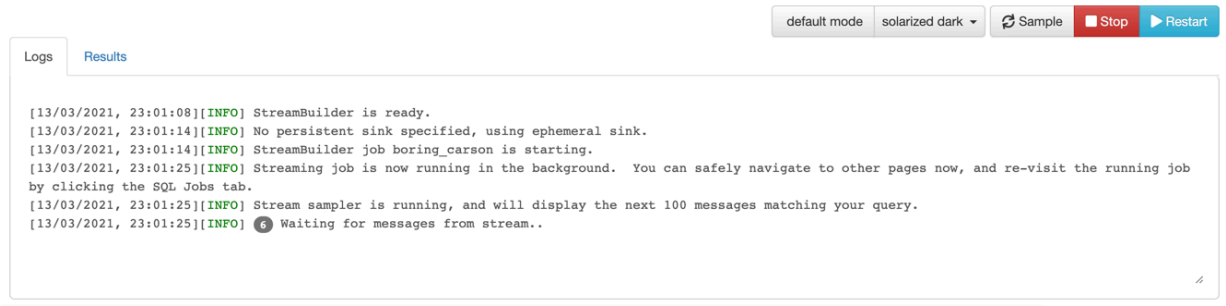


Note: You cannot start a job without adding a SQL statement in the SQL editor window. In case, there are no SQL statements provided and you click on the Execute button, the following error message is displayed: You must provide a SQL query.

When starting a job, the number of slots consumed on the specified cluster is equal to the Parallelism setting. The default is 1 slot. To change the parallelism setting, click Advanced settings.

7. Click Execute.

   The Logs window updates the status of SSB.



8. Click Results to check the sampled data.

   These results are only samples, not the entire result of the new stream being created from the output of the query. The entire result set is sent to the Sink Virtual Table and/or a Materialized View.

Note: As SSB is querying unbounded data streams, you need to click on the Stop button to stop the job execution.

Results

A job is generated that performs the SQL continuously on the stream of data from the Source Virtual Table, and pushes the results to a Sink Virtual Table.

Related Information

Data Sources in SSB

Using Virtual Tables

Monitoring SQL Stream jobs

Stopping, restarting and editing SQL jobs

As a SQL Stream job processes streaming data, you need to stop the job to finish the process. You can restart a SQL Stream job after stopping it. In case you need to update or change the configurations set for a SQL Stream job, you can restart it. You can navigate through the job life cycle using the Streaming SQL Console.

Stopping a SQL Stream job

1. Select Console from the left hand menu.
2. Click on the SQL Jobs tab.
3. Click on the red stop button for the job you would like to stop.

Restarting a SQL Stream job

1. Select Console from the left hand menu.
2. Click on the SQL Jobs tab.
3. Click on the select box below the State column.
4. Select Cancelled or Failed to see stopped jobs.
5. Select the job you would like to restart.
6. Select the Details tab at the bottom.
7. Select the Edit Selected Job button.

   The SQL window in Edit Mode appears.
8. Select Restart to restart the job.

Editing a SQL Stream job

1. Select Console from the left hand menu.
2. Click on the SQL Jobs tab.
3. Select the job you would like to edit.
4. Select the Details tab at the bottom.
5. Select the Edit Selected Job button.

   This will bring up the SQL window in Edit Mode.
6. Edit the Target Deployment, Sink Virtual Table and the SQL itself as needed.
7. Select Restart to restart the job.

   The job will be stopped and restarted.

Sampling data for a running job

You can sample data from a running job. This is useful if you want to inspect the data to make sure the job is producing the results you expect.

Procedure

1. Select Console on the main menu.
2. Click on the SQL Jobs tab.
3. Select the job you would like to edit.
4. Select the Details tab at the bottom.

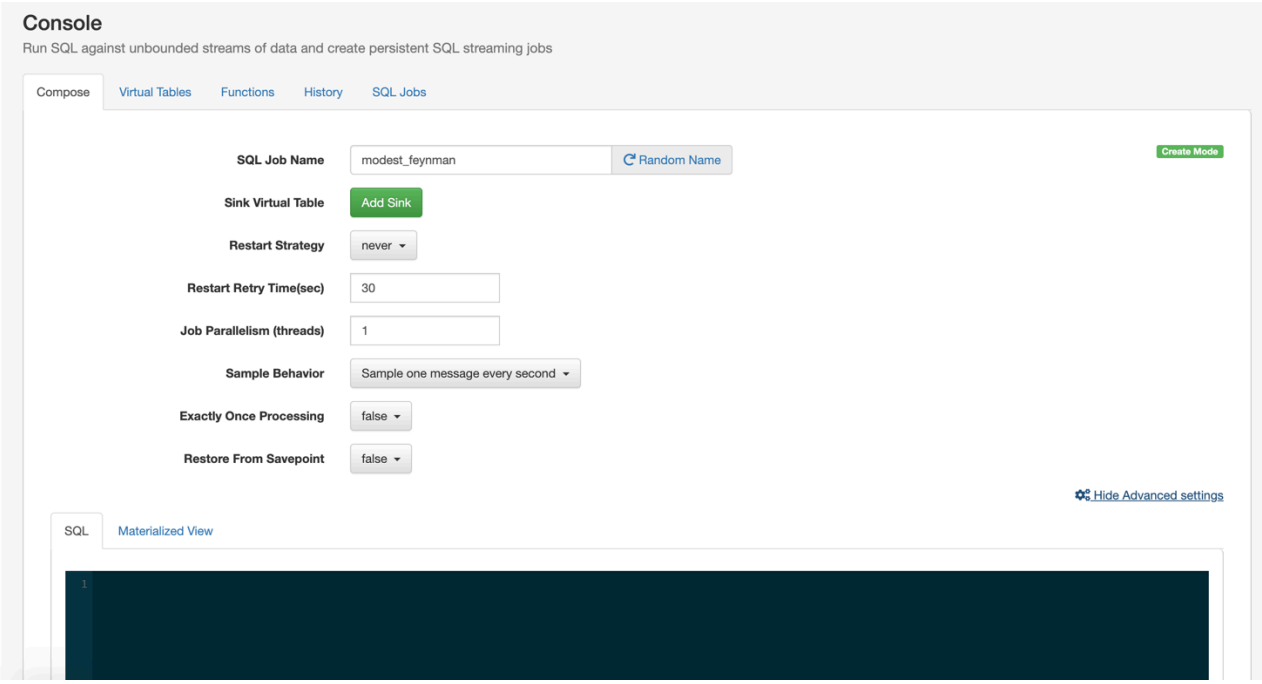
5. Select the Edit Selected Job button.
   The SQL window in Edit Mode appears.
6. Click the Sample button.

Results

Results will be sampled and displayed in the results window. If there is no data meeting the SQL query, sampling will give up after a few tries.

Advanced Job Management

When starting a job a number of more advanced features can be selected to configure various properties of the job. To configure these features select the Show Advanced Settings button on the Compose tab.



Restart strategy and restart retry time

The job will be restarted after Restart Retry Time seconds if set to Always. It will not be restarted if you select Stop from the SQL Jobs tab. If set to Never the job will not be restarted unless you select Restart from the Compose tab.

Job parallelism

The number of threads to start to process the job. Each thread consumes a slot on the cluster.

Sample interval

How often to sample data (in milliseconds) from the output stream. 1000ms is common and recommended.

Sink Exactly Once Semantics

Enabling exactly once processing of the data that is generated to the sink.

Restore From Savepoint

Enabling restoring from savepoint for the SQL job using the state in Flink.