Quick guide to loading data into SQream

SQream Technologies

Version 2019.2

Table of Contents

Quick guide to loading data into SQream

Overview

Importing your data into SQream is easy. There are 3 different methods:

- **COPY FROM:** quickly inserts bulk data from local csv files into a single table
- **External Tables:** can import a new external table in the current database under a specific schema.
- **Data Streaming:** streams data directly from a source such as a different database or from a remote server. The source data can be any file, application, or database that you can access with one of the SQream connectors.

General Notes

- Wherever csv file type is mentioned, it includes any delimiter-separated values file type as defined in RFC 4180.
- **Data Streaming:** Although you cannot perform data transformations and manipulations as part of the actual data streaming, you can perform a full range of transformations at the source API before streaming.
- The maximum row length is 10239 bytes.

Let's compare import methods to see the advantages and limitations of each method.

Comparing import methods

The following table summarizes the advantages and limitations of each method.

Command	Advantages	Limitations
COPY FROM	• Easy to use • Can skip rows, such a header row as part of command • Can perform date format transformations as part of command • Error handling options	• Local data only (or accessible locally via NFS) • Cannot perform any data manipulations other than date format
External Tables	• Most powerful method • Full range of data transformations and manipulations	• Local data only (or accessible locally via NFS) • No error handling options
Data Streaming	• Can access remote data • Can insert directly from another database	• Requires external data manipulations before streaming • Requires 3rd party tool for ETL with SQream connectors or coding

So which import method is the right one for you?

Choosing the import method that is right for you

To select the method that is right for you, answer the following 3 questions:

- What format is your data in?
- Where is your data?
- Do you need to perform data transformation while importing your data?

What format is your data in?

csv file

csv files are supported by all 3 methods.

Apache Parquet file

Apache Parquet files are supported only by the **External Table** method.

All other file types

If your data is not in either a csv file or an Apache Parquet file, you must use the **Data Streaming** method to import your data.

Other data sources

If you are importing from a different data source, such as a database or streaming software, you must use the **Data Streaming** method to import your data.

Where is your data?

Local

If your data is in a local file location, you can use any method.

Remote

If your data is in a remote location, you must use the **Data Streaming** method

Do you need to manipulate the data in any way while importing?

Simple data transformations can be performed as part of the import process. Any SQL or user-defined function can be applied to the data or to a specified column during the SQL INSERT.

For example, you might want to sort your data during import. The sorted data would result in more efficient queries.

NOTE: Data transformations will result in longer import times.

No

If you don't need to do any data transformation, you can choose any method.

Yes

- If you do need to do some data transformation, use the **External Table** method
- You can also use the **Data Streaming** method if you perform the transformation using external software before streaming.
- **Copy From** does not support data transformation (with the exception of date format).

Limited data handling with Copy From command line options

The **Copy From** method provides command line options that let you easily do the following:

- Date Format: specify the date format
- Skip Rows: you can skip a specified number of rows at the beginning of the file (such as header row) and also stop import after a specified number of rows
- Error Handling: continue to import up to a specified number of errors

Which method is right for you?

Command	csv file	Parquet file	Other file type (including other data sources)	Remote file location?	Data transformation?
COPY FROM	✓	×	×	×	× (date format only)
External Tables	✓	✓	×	×	✓
Data Streaming	✓	✓	✓	✓	× (must perform in source API)