Big Data Connector for Google BigQuery

Qlik Connector for Google BigQuery

By Stretch

Stretch | Kompagnistræde 21 | Copenhagen | +45 26 28 78 44 | jonathan.hvid@stretch.dk

Introduction

The Big Data Connector for Google BigQuery is a Qlik Sense connector developed by Stretch, based on our collaboration with multiple customers and their needs for at simple and efficient way of loading large datasets, 100 million+ rows, from Google BigQuery into their Qlik Sense solutions.

The identified requirements for the connectors was

- Support loads of 100 million+ rows
- Achieve loads speeds allowing for multiple daily full reloads of such datasets
- Utilize google service-account for authentication
- Support multiple billing account
- Support cross project data loads
- Be easy to use

The current version of the connector meets all the above requirements.

This document outlines the overall the achieve speed of the connector

Performance

The performance test was carried out on a Google Cloud Services Account and an instance of Qlik Sense Enterprise running on an Azure virtual machine.

The connector was tested in its two modes of operation: Direct and fast mode.

Test setup

Google Cloud Services

- Services
  - BigQuery
  - Cloud storage
- Authentication
  - Service Account – Json key file
- Dataset
  - Based in Googles sample data, extended by unique custom key
  - 5 Columns
  - 1.08 billion rows
  - 68.04 GB

Qlik Sense Enterprise

- Qlik Sense Enterprise – April 2019
- Windows server 2016 – fully Patched
- Server is joined a 2016 windows domain
- Qlik Sense running as domain service account
- Hosted on Microsoft Azure (to ensure cross internet communication)
- Azure F8S_v2 Server
  - 16 GB ram
  - 8 Threads
  - Clock Speed 2.4 – 3.1 (Turbo boost)
  - SSD Storage
- Domain user with root admin used for testing

Connector Configuration

- Uncompressed files were used
- Works threads were limited to 3 threads

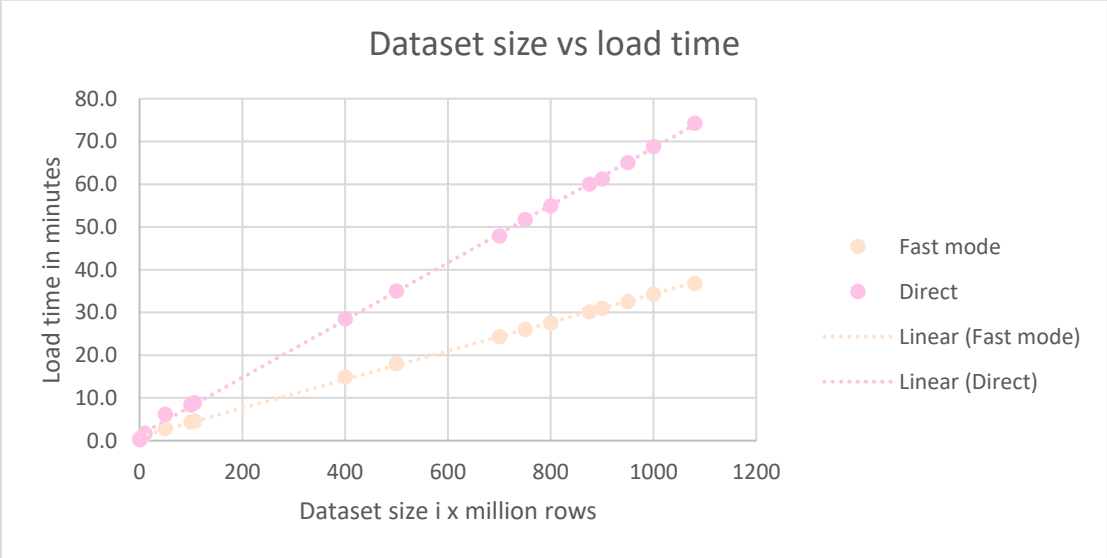
Performance

The performance test was conduction on the above listed setup. The dataset is a representative sample dataset available from Google. The dataset was extended to one billion rows by a unique key and some data transformation. The different dataset sizes were simulated by including a limit statement in the BigQuery query within Qlik Sense.

A series of sixteen different dataset load was tested, ranging from 100.000 rows (6.3 MB) to 1 billion rows (68 GB)

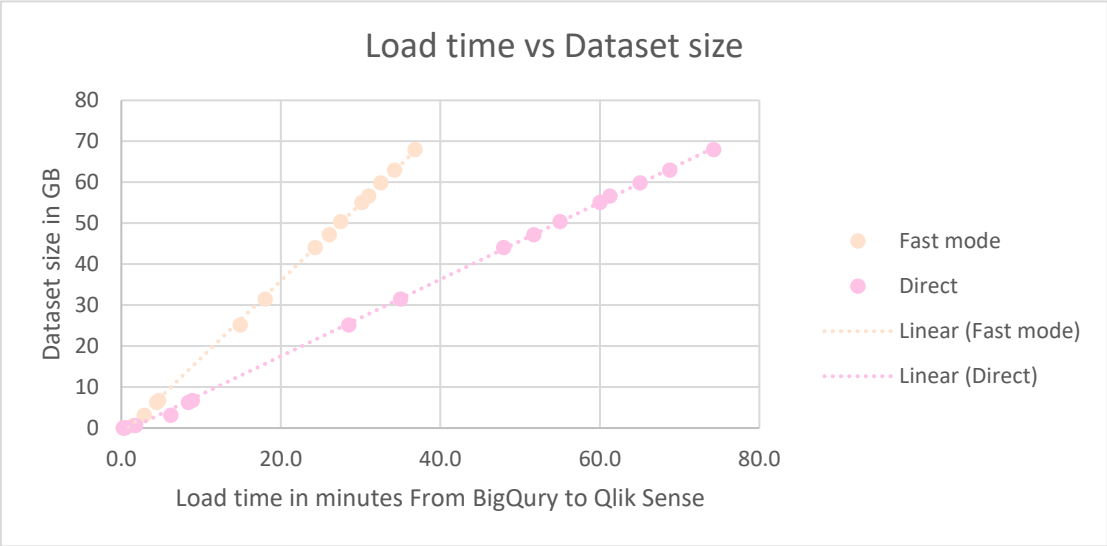
The load time was recorded from starting of the load script until the connection was closed. Saving and indexing the app was excluding, since these operations are very dependent to system resource and performance.

Number of rows vs. load time



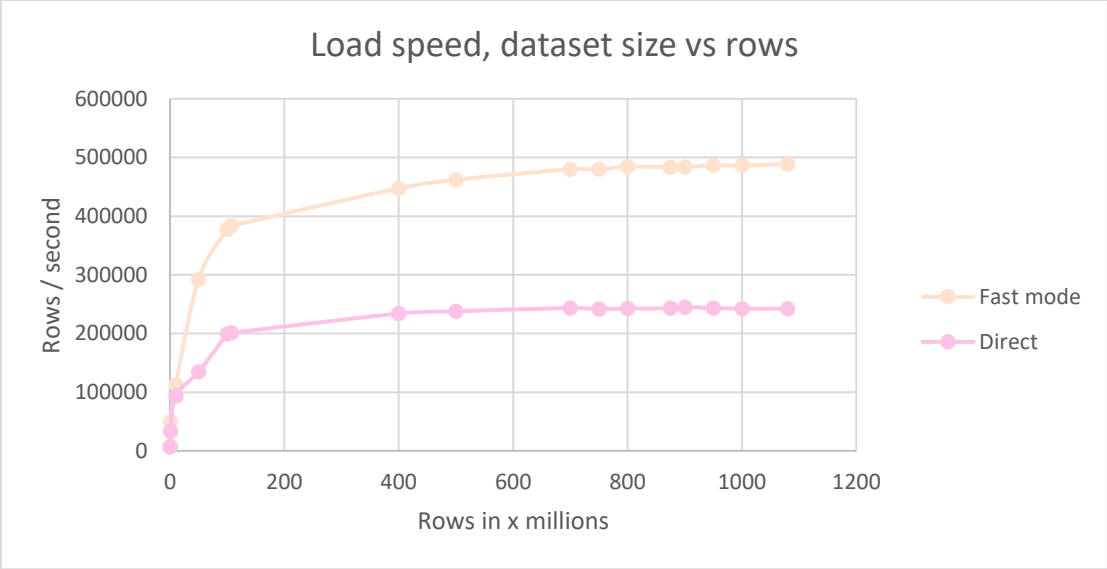
This graph shows the number of rows in the data set vs. the load time to Qlik Sense. As seen from the graph the connector scales very well with the size of the datasets, it is close to linier. It can also be noted that a 900 million rows dataset is loaded within an hour I direct mode and within a half an hour in fast mode. This performance allows for multiple daily reloads of very large datasets.

Load time vs. dataset size

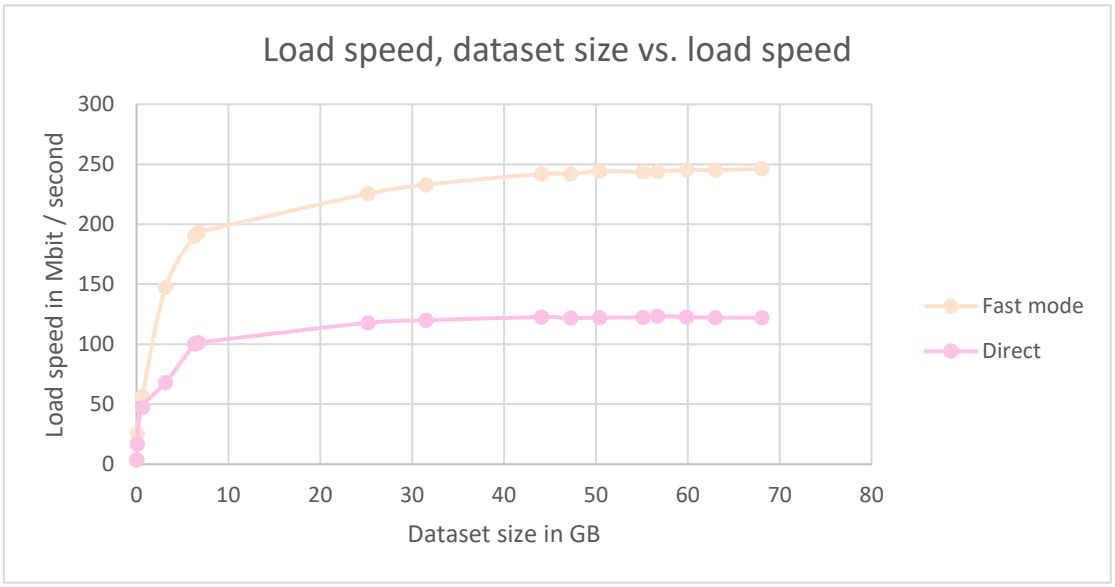


This graph shows the load time in minutes vs. the size of the dataset in GB. As expected, the graphs show near linear scaling of performance.

Resulting load speed vs. dataset size

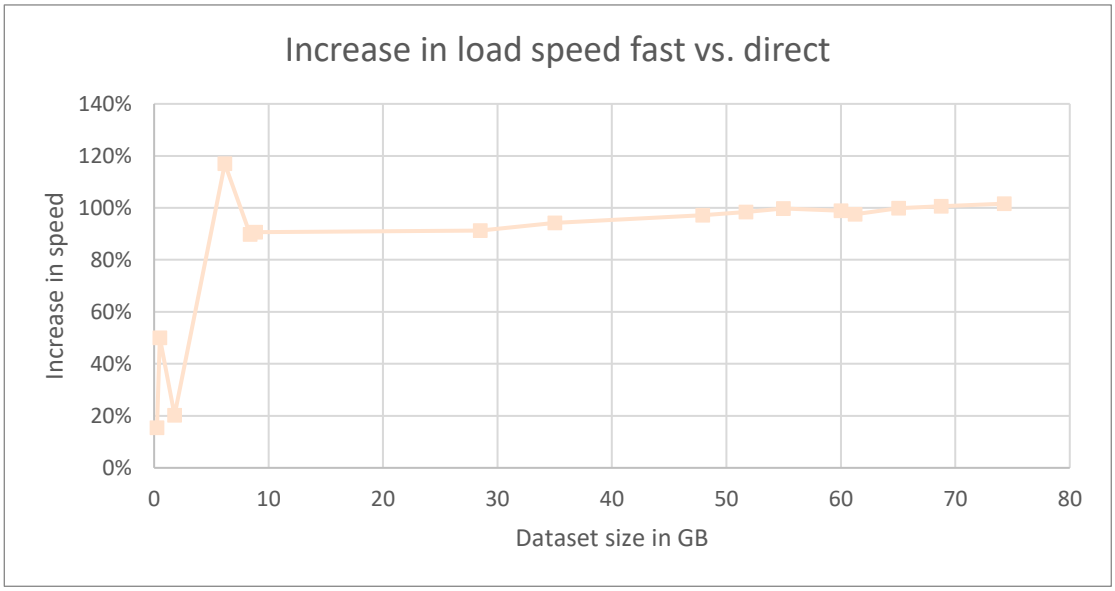


This graph shows the load speed in rows per second for the different dataset size. As seen the connector is more efficient with larger datasets due to its ability to spread the load over multiple worker threads. It is also seen that the load speed hits a plateau around 245.000 rows / second, which seems to be the speed limit to which Qlik Sense can ingest data from a connector.



This graph shows the resulting average data load speed in Mbit per second as calculated from the dataset size. It shows the same plateau, around 120 Mbit/s.

Mode performance



For dataset above, teen mio. rows, the fast mode provides a speed increase in the order of 100% or a factor x 2 to the already fast direct mode.

Connector resource footprint

From our testing, the footprint of the connector is around in direct mode is 2-4 thread with 90-100% utilization (dependent on configuration) and 1-2 treads with 75-100% utilization for the Qlik Sense engine process. The fast mode utilizes most of the CPU resources.

This relatively small system footprint, in direct mode, relative to a normal system of Qlik Sense deployment, allows for multiple concurrent load without extending the load time for individual instances of the connector or consuming all the resource of the loading node.

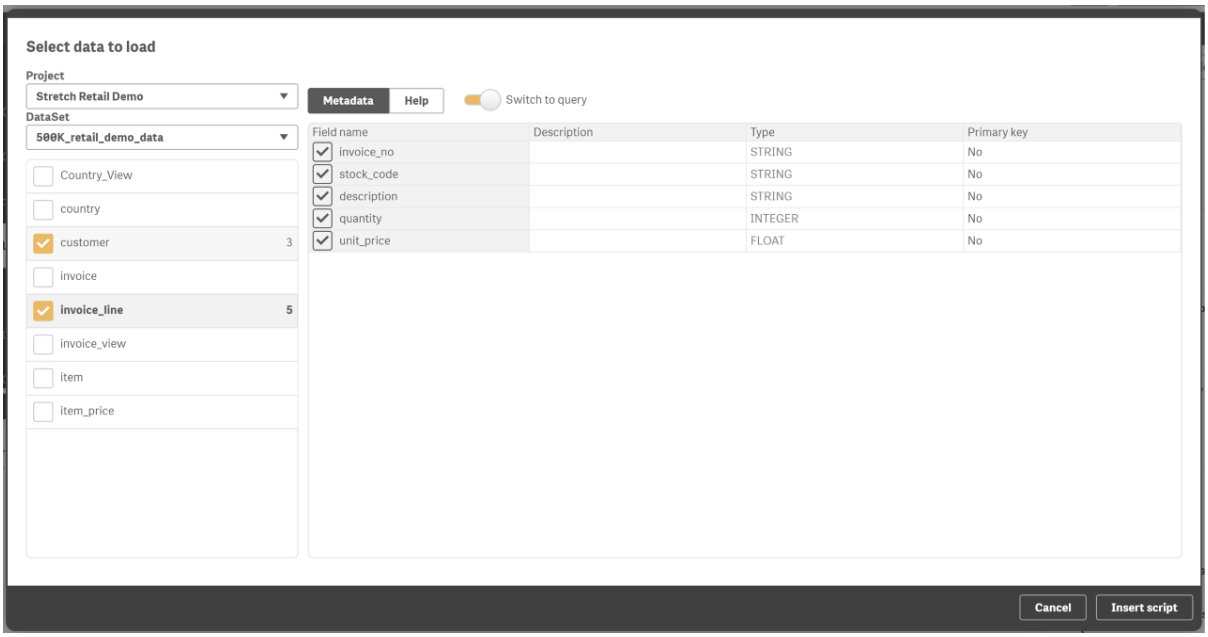
The testing of the connector showed a constant memory usage across the different data set sizes. Precise testing of the footprint is difficult due the nature of .net garbage collection, as it primarily is trigged when the system is low on virtual memory, which rarely was the case on the test system.

Scaling with multiple simultaneous loads

Test on a customer system, the load time for three simultaneous loads in direct mode, was identically, to a single load. It was estimated that maybe one or two more loads could be run without effecting the performance, but this would utilize the whole server and was note tested at this time. This was testing on a 32 threads system.

Easy to use

The connector can be used with an all graphic user interface to make it easy for users to quickly get an overview of their data and simply pick out that, which they want to import.



More information or a Free Trial

Video on how to use the connector:

https://www.youtube.com/watch?v=MqsIQOo1nVE&t=35s

Link to the connector on Qlik Market:

https://market.qlik.com/solutions/Big_Data_Connector_for_Google_BigQuery

FOR LICENSE OR A FREE TRIAL, PLEASE CONTACT YOUR STRETCH REPRESENTATIVE AT:

Martin Sahlin, CEO

martin.sahlin@stretch.dk

+46 701 45 65 49

+45 25 17 17 59

or

Jonathan Hvid, Head of Sales

jonathan.hvid@stretch.dk

+45 26 28 78 44