Thousands, and Millions and Billions … Oh My!!!!!

written by DaltonRuer | May 18, 2016

When most people think of Qlik they think of our patented Qlik Indexing Engine having all of your data in memory. I love demonstrating the lightning fast speeds and responsiveness with hundreds of millions of rows of data. More and more recently though I'm getting the smiles mixed with "That's awesome but can you handle billions of rows of data?"

C'mon really????? Billions of rows of data? Gosh that's an awful lot of data. I'm afraid.

Just kidding even that much data doesn't scare me.

In fact it thrills me.

[Tweet "Gives me goose bumps to think about the kind of decisions that can be made when that much data is made available to the analysts and the decision makers."]

Gives me goose bumps to think about the kind of decisions that can be made when that much data is made available to the analysts and the decision makers. It also provides an opportunity for me to discuss one of the least known features that Qlik offers. It's called Direct Discovery and it allows you to consume even billions of rows of data.

Direct Discovery

Direct Discovery is a two step process. In step 1 Qlik reads enough information to allow the end user to select a cohort. Step 2 then uses the primary key information for that cohort to go back to the massive data store and read all of the details live.

Oh wait you want an example? More details? Well since you asked so nicely.

Typically with Qlik you would read all of your data from the source with a command like:

SQL Select {my fields} from {some table};

It would bring all of the data back, perform our Qlik magic on it to compress it in memory and you would be off to the races. With Direct Discovery the query is different and uses a different syntax. You start with something like this:

```
DIRECT QUERY

DIMENSION

record_id

procedure
```

When the data load encounters that Qlik actually issues 2 separate commands to the source:

1. Select distinct record_id
2. Select distinct procedure

Why? Because it’s easier and faster of course. The data source only has to prepare a minimal amount of records. Your network only has to transmit a minimum amount of data. Finally Qlik only has to read a minimum amount of data.

The final part of the syntax would be something like:

```
DETAIL
admitting_diagnosis,
diagnosis,
codenum,
icd9,
daynum,
rn,
type_of_admission
from surgery_events;
```

The fields that you identify in the DETAIL section of the command are usable immediately within Qlik despite the fact that it doesn't actually retrieve the data for them. You can see the field names in the data model viewer they just show as having 0 rows of data. You can see the fields in a field list. You can add the fields to charts. There just isn't any real data for them. Yet anyway.

Your application is then designed to allow the end user to select a cohort using the DIMENSION fields in some way and then Qlik will go and retrieve the data live from the data source for that cohort.

I've had so much fun working with Teradata Aster lately that it only made sense for me to use my Teradata database as a data source. It provides a robust, high performance and highly reliability storage mechanism for those with massive amounts of data. In the video I use the command above to extract the dimensions, select a cohort of patients, then allow Qlik to extract the data live. Just for fun I also utilize the Aster Management Console to show you the commands that Teradata processes from Qlik to further solidify how it all works. Kind

of the extra step you'd expect from me.

You want more don't you?

The ferocious appetite in you to consume massive amounts of data wants more information doesn't it? You can check out all of the details on the Qlik Sense help page for Direct Discovery:

http://help.qlik.com/en-US/sense/2.2/Subsystems/Hub/Content/DirectDiscovery/access-large-data-sets-with-direct-discovery.htm

The following post contains a fantastic PDF document explaining even more including some nifty variables you can use like the one I documented in the video:

https://community.qlik.com/docs/DOC-6917

Yes you can even use the Direct Discovery feature for cases where you want closer to real time information from smaller sets of data. You know those situations where you only have a few hundred million rows of data but you still need the functionality of pulling live rather than having pre-loaded all of the detailed data.