Google: Big Query

Data Warehouse and Management

Members:
Divya Thakkar
Jaxson
Sidney

What is BigQuery?

BigQuery is Google's serverless cloud data warehouse. It does not have any infrastructure so it is easy to manage using SQL and without the need of a data administrator. It reports with the in-memory BI engine so you create dashboards and reports quickly. You can build, using simple SQL operational machine learning solutions and or carry out geospatial analysis. BigQuery captures and analyzes data up to 1 TB of data and store of 10GB of data free for each month in real time.

BigQuery is an enterprise data warehouse. It solves the problem of the time consuming and the expensive process of querying massive datasets without the hardware and infrastructure. It is fully managed and can be deployed without any resources, infrastructures or disks.

There's three main ways to interact with BigQuery.

- Loading and exporting
- Querying and viewing
- Managing data

To use the interactions, use:

- The BigQuery web UI in the console
- The classic web UI
- The command-line tool
- The Rest API or client libraries

You can load data into the BigQuery storage where you can export it if you want it back out. You can also set a table as an external data source. This allows you to query data that is stored outside of BigQuery.

You can view the data you load by viewing or querying it in your tables. You can run interactive queries, run batch queries, create views, or use partitioned tables to query a part of your data.

You can manage the data in BigQuery by listing projects, jobs datasets, and tables. You can get information about jobs, datasets and tables. You can define,update, patch or delete datasets and tables. You can also manage table partitions with BigQuery.

Key Features:

BigQuery has three key features:

- BigQuery ML
- BigQuery Bi Engine
- BigQuery GIS

BigQuery ML:
In simple SQL, this enables data analysts to query and build operational ML models on a huge scale structure or semi structured data in a short amount of time.

Features:

- It builds tests and operationalize custom machine learning models using SQL language
- It creates machine learning models in minutes directly without widespread data sampling
- At a petabyte scale in a fraction of costs and time, it can do product recommendations, segmentations and predictions.

The functionality is available using:

- The BigQuery web UI
- BQ command-line tool
- BQ rest API
- Or a jupyter notebook or business intelligence platform

It requires extensive programming and knowledge for machine learning on datasets. BQ ML, for short, makes use of machine learning through SQL tools and skills. BQ ML can be used to evaluate ML models in BigQuery by Analysts without needing to export data to any external applications.

Supported Models for BigQuery ML:

- Linear regression for forecasting;
- Binary Logistic regression for clarification
- Multicast logistic regression for clarification
- K-means clustering for data segmentation(beta)
- TensorFlow model importing.

Advantages of BQ ML:

- There's a few advantages of using BQ ML over other systems. One advantage for example is that the models are accessed by BQ using SQL so there is no need to know other languages. BQ ML also is different from the other options because of its speed and efficiency of any processes and everything is done in BigQuery.

Pricing:

- The charges are based on how much data is used to train a model. Or the queries you run against the data.

BigQuery Bi Engine:
What is it?
BigQuery Bi Engine is a beta service in BigQuery. It's a fast-in-memory analysis that lets users analyze huge and very complex datasets interactively. It integrates with tools like Data Studio, Looker, or Google Sheets.

Features :

- Enables real-time dashboarding
- Simplify BI architecture
- Optimize visual analytics over big datasets.

The advantages of BigQuery Bi Engine is that it's really fast. It also has a simplified architecture so it is easy to use. With that it is also very easy to configure with only a few steps. The only disadvantage is that it is in beta.

Limitations:

- Bi engine is available only on Data Studio. The queries against BQ views are not optimized by BI engine. The max limit is 10 GB of data. In Data Studio, custom queries are not optimized.

Pricing:

- Pricing is based on regular pricing for the over all BigQuery service

BigQuery Geographic Information Systems(GIS)
What is it?

BigQuery GIS is BigQuery's service that combines serverless a architecture with the native support for geospatial analysis.

Features:

- It uses SQL and PostGIS functions
- Analyses huge geospatial data using BigQuery's speed.
- It can use Google Earth Engine for a faster analysis.

Limitations:

- BQ GIS geography functions are only available in standard SQL.
- Only the BigQuery Client library for Python is the only thing that supports GEOGRAPHY data type. If you want to use other client libraries you need to convert GEOGRAPHY values to string with the ST_ASTEXT or ST_ASGEOJSON. Using ST_asText stores only one value. The data is annotated as STRING if you convert it to WKT.

Pricing:

- Using BQ GIS, charges are based on how much data is stored in the table and the queries you run against the data.

Security:

- BigQuery encrypts all data automatically before it is written to any disks. It is decrypted

in the same way when read by an authorized user. Google manages the key encryption keys by default but you can also use customer managed encryption keys and encrypt individual values in a table.

Default Encryption at rest:

- Each data and meta data in BigQuery is encrypted with Advanced Encryption Standard or (AES). Each encryption key is encrypted by a regularly updated and rotated set of master keys.

Customer-Managed encryption Keys:

- If you do not want Google to manage your key, you can also use a Cloud Key Management Service instead(CKMS or Cloud KMS). It uses AES-256 KEYS to encrypt your data.

Encryption of Individual values in a table:

- Encrypting individual values in a table in BigQuery uses Authenticated Encryption with Associated Data or AEAD functions. It is based on AES encryption.

Conclusion:

- BigQuery is Google's answer to Amazon's AWS and to Microsoft's Azure. It’s a very robust SaaS service that is very fast thanks to Google's huge resource pool. The pricing is adequate to the normal price for such services.

References:

- https://cloud.google.com/bigquery/