CHAPTER 38

Google BigQuery

BigQuery is a Google-managed data warehouse product that is highly scalable, fast, and optimized for data analytics with rudimentary in-built machine learning capabilities as part of the product offering. It is also one of Google's many serverless products. This means that you do not physically manage the infrastructure assets and the overhead responsibilities/costs. It is only used to solve the business use case, and it just works in a highly performant manner.

BigQuery is suited for storing and analyzing structured data. The idea of structured data is that it must have a schema that describes the columns or fields of the dataset. CSV or JSON files are examples of structured data formats. BigQuery differentiates itself from other relational databases in that it can store a collection of other fields (or columns) as a record type, and a particular field in a row can have more than one value. These features make BigQuery more expressive for storing datasets without the flat row constraint of relational databases.

Similar to relational databases, BigQuery organizes rows into *tables*, and are accessed using the familiar Structured Query Language (SQL) for databases. However, individual rows in a table cannot be updated by running a SQL Update statement. Tables can only be appended to or entirely re-written. Meanwhile, a group of tables in BigQuery is organized into *datasets*.

When a query is executed in BigQuery, it runs in parallel on thousands of cores. This feature greatly accelerates the performance of query execution and consequently the speed of gaining insights from your data. This ability for massive parallel execution is one of the major reasons individuals, companies, and institutions are migrating to BigQuery as their data warehouse of choice.

Also BigQueryML is a powerful platform for building machine learning models inside of BigQuery. The models take advantage of automated feature engineering and hyper-parameter optimization and are automatically updated based on changes to the underlying dataset. This feature is extremely powerful and lowers the threshold

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_38

of business intelligence and analytics personnel to more easily harness the predictive power of using machine learning for business forecasting and decision-making.

What BigQuery Is Not

As powerful and widely purposed as BigQuery is, it may not be properly suited for some use cases:

- BigQuery is not a replacement for a relational database. Some business use cases may involve a large number of table row updates; in such an instance, BigQuery is most likely not the data storage solution of choice, as relational databases are well suited for such highly transactional tasks. GCP offers the Cloud SQL and Cloud Spanner as parts of its managed relational products.
- BigQuery is not a NoSQL database. Data stored in BigQuery must have a schema. NoSQL is a schema-less data storage solution. GCP also has Cloud BigTable and Cloud Datastore, which are highly scalable and performant managed NoSQL products.

Getting Started with BigQuery

BigQuery can be accessed and used via a variety of ways; they include

- The BigQuery web UI
- The command-line tool, **‘bq’**
- The client API libraries for programmatic access

In this section, we will introduce BigQuery by working with the web UI, because it gives a graphical view of the datasets and tables within BigQuery and is good for quick execution of queries on the query engine.

To open BigQuery from the GCP dashboard, click the triple dash on the top-left corner and select **BigQuery** from the product section labeled **Big Data** as shown in Figure 38-1.