Google Cloud Platform

- BigQuery Analytics Hub
- BigQuery
- BigQuery Working with Array Structures

BigQuery Analytics Hub

Analytics Hub is a mechanism for sharing datasets between BigQuery users. Google's official product documentation for Analytics Hub is here

Subscriber Workflow

- A dataset publisher creates and shares a Linked Dataset with the subscriber - this essentially becomes a symbolic link to the original dataset that the subscriber can access.
- In this model the person querying the data pays for it.

BigQuery

Google BigQuery is a data warehouse platform.

Schema

You can define table schemas via JSON documents which get ingested at the same time as your data using the `bq` tool.

Data Types

Data Types List from google documentation

Name	Data type	Description
Integer	INT64	Numeric values without fractional components
Floating point	FLOAT64	Approximate numeric values with fractional components
Numeric	NUMERIC	Exact numeric values with fractional components
BigNumeric	BIGNUMERIC	Exact numeric values with fractional components
Boolean	BOOL	TRUE or FALSE (case-insensitive)
String	STRING	Variable-length character (Unicode) data
Bytes	BYTES	Variable-length binary data
Date	DATE	A logical calendar date
Date/Time	DATETIME	A year, month, day, hour, minute, second, and subsecond
Time	TIME	A time, independent of a specific date
Timestamp	TIMESTAMP	An absolute point in time, with microsecond precision
Struct (Record)	STRUCT	Container of ordered fields each with a type (required) and field name (optional)

Name	Data type	Description
Geography	`GEOGRAPHY`	A pointset on the Earth's surface (a set of points, lines and polygons on the WGS84 reference spheroid, with geodesic edges)
JSON	`JSON`	Represents JSON, a lightweight data-interchange format

Differences between JSON and Record/Struct

JSON type allows you to ingest JSON without pre-defining the schema whereas a record/struct must be pre-defined and all the fields must be known in advance.

JSON fields are more fiddly to query and work with in general. It seems like you can't do things like UNNEST them.

Nested/Repeated Columns

To allow a column (or object) to repeat (e.g. to have an array of values) you must use `mode: NESTED` in your schema.

See Nested and repeated columns documentation

BigQuery Working with Array Structures

On a companies table where you can have multiple websites per company:

```
SELECT *
FROM
  project_namespace.dataset.companies as companies,
  UNNEST (companies.websites) as website
  LIMIT 1000
```

The above would produce a flattened table structure where company id is repeated per website.