Module 7

Ingesting New Datasets into Google BigQuery

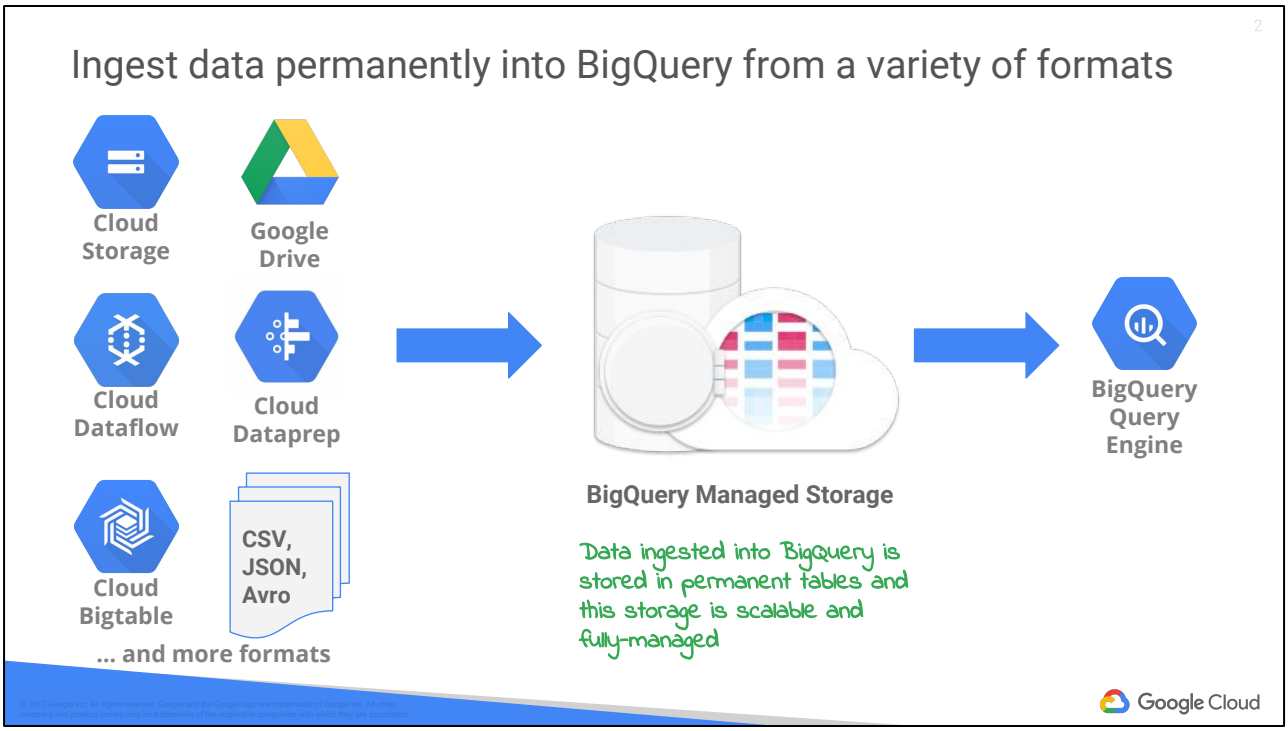
In this module we will:

- Query from External Data Sources
- Avoid Data Ingesting Pitfalls
- Ingest New Data into Permanent Tables
- Discuss Streaming Inserts

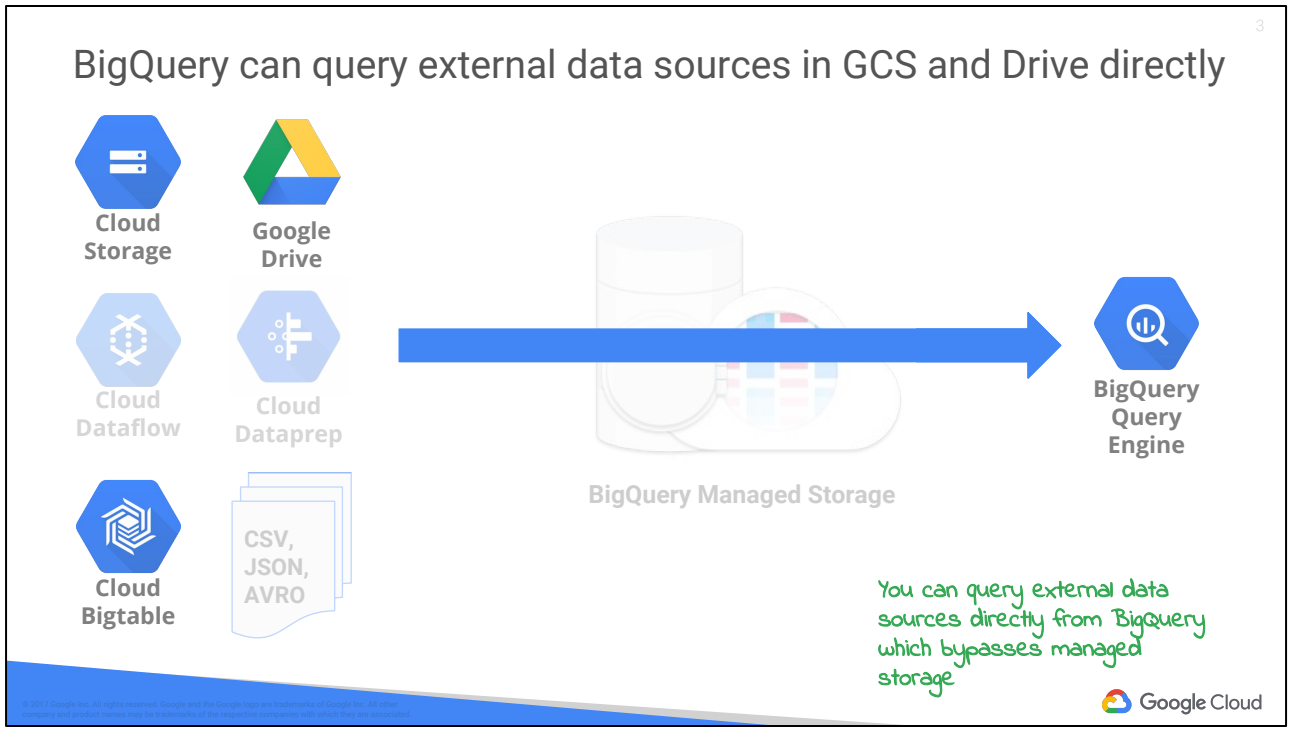
Google Cloud

So far we have only queried datasets that already exist within BigQuery. The next logical step after you're finished with this course is practicing loading your own datasets into BigQuery and analyzing them. In this module we will cover how you can load data into BigQuery and create your own datasets.

Let's first cover the difference between loading data into BigQuery versus querying it directly from an external data source.



BigQuery can ingest datasets from a variety of different formats. Once inside BigQuery native storage, it is fully managed by the BigQuery team here at Google (replication, backups, scaling out size, and more).



You also have the option of querying external data sources directly and bypassing BigQuery managed storage.

Pitfalls: Querying from External Data Sources Directly

Limitations

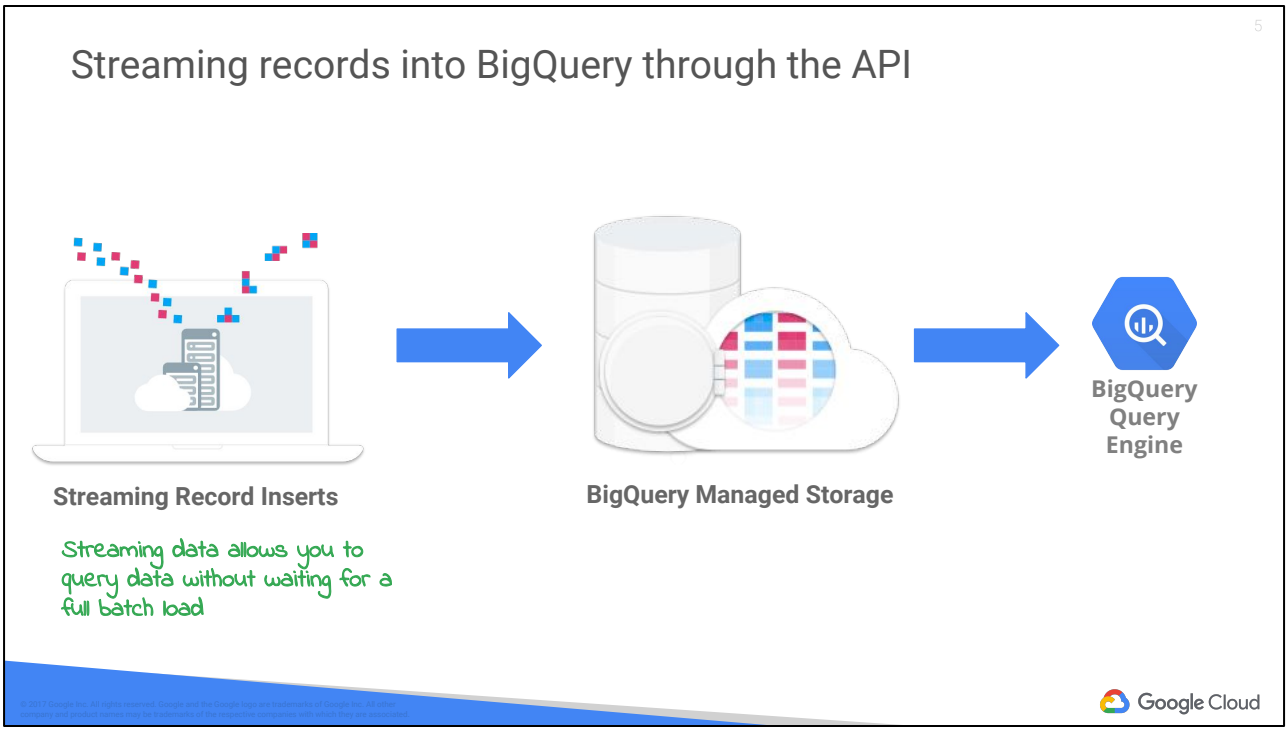
- Strong Performance Disadvantages
- Data Consistency not Guaranteed
- Can't Use Table Wildcards (cool feature we will introduce shortly)

Google Cloud

Limitations

External data source limitations include the following:

- BigQuery does not guarantee data consistency for external data sources. Changes to the underlying data while a query is running can result in unexpected behavior.
- Query performance for external data sources may not be as high as querying data in a native BigQuery table. If query speed is a priority, load the data into BigQuery instead of setting up an external data source. The performance of a query that includes an external data source depends on the external storage type. For example, querying data stored in Google Cloud Storage is faster than querying data stored in Google Drive. In general, query performance for external data sources should be equivalent to reading the data directly from the external storage.
- You cannot use the `TableDataList` JSON API method to retrieve data from tables that reside in an external data source. For more information, see Tabledata: list.
- You cannot run a BigQuery job that exports data from an external data source.
- You cannot reference an external data source in a wildcard table query.



Stream records into BigQuery

- Using the API tabledata().insertAll() method
- Max row size: 1MB
- Max throughput: 100,000 per second per project
- Max rows per request: 10,000 (batching recommended)

Event Logging

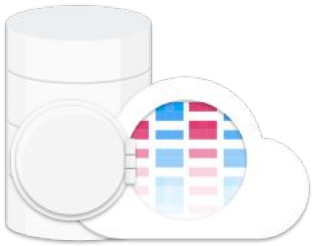
One example of high volume event logging is event tracking. Suppose you have a mobile app that tracks events. Your app, or mobile servers, could independently record user interactions or system errors and stream them into BigQuery. You could analyze this data to determine overall trends, such as areas of high interaction or problems, and monitor error conditions in real-time.

Real-time dashboards and queries

In certain situations, streaming data into BigQuery enables real-time analysis over transactional data. Since streaming data comes with a possibility of duplicated data, ensure that you have a primary, transactional data store outside of BigQuery.

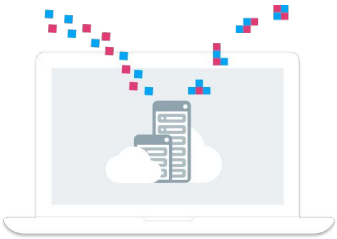
https://cloud.google.com/bigquery/streaming-data-into-bigquery

Summary: Ingest new datasets into BigQuery managed storage



Data stored permanently in BigQuery is fully-managed (performance, backups, redundancy).

Load new data from a variety of formats



Setup streaming ingestion into BigQuery through APIs

Google Cloud

As you have seen, there are a variety of ways to get your new data into BigQuery. We covered loading and storing this data as new permanent tables (which have the benefit of being fully-managed). We also looked at querying external data directly and why this may be useful for one-time Extract Transform Load jobs. Lastly, you can stream individual records into BigQuery through an API.

Next up is our lab where we will practice creating new datasets, loading external data into tables, and running queries on this new data.

Links:
Loading Data into BigQuery: https://cloud.google.com/bigquery/loading-data

Image (data blocks) cc0: https://cloud.google.com/data-transfer/

Lab 6
Ingesting and Querying New Datasets

Google Cloud

Ingesting and Querying New Datasets

In this lab, you will ingest new data sources into Google BigQuery and learn how to query external data sources directly.

Google Cloud

Image (data cloud) cc0: https://cloud.google.com/data-transfer/