

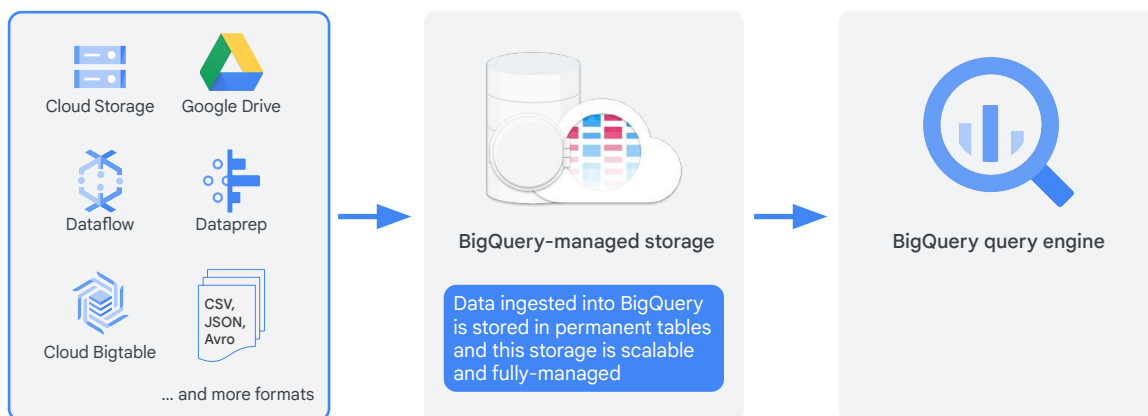
Ingesting New Datasets into BigQuery

Evan Jones

Now so far, we've only queried data sets that already exist within BigQuery. The next logical step after you're finished with all these courses is to load your own datasets in the BigQuery and analyze them.

So, that's why in this module, we'll cover how you can load extra node data into BigQuery, and create your very own datasets. First, let's cover the difference between loading data into BigQuery versus querying it directly from an external data source.

Ingest data permanently into BigQuery from a variety of formats



Google Cloud

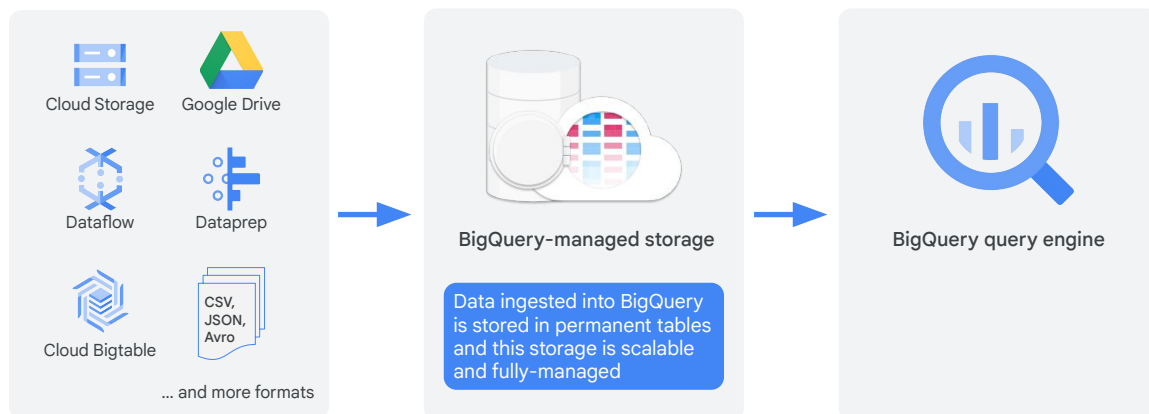
As you can see on the left, there's a lot of different file format types and even systems that you can actually ingest and grab data from, and then load permanently into BigQuery-managed storage.

So, to name a few of very common staging areas Google Cloud Storage, we could have your massive CSV files stored into Cloud Storage buckets, which is very common, or Dataflow jobs, your data engineering team has set up these beautiful pipelines. And as part of one of the steps in the pipelines, you can have that data write out or materialize itself into a BigQuery table for analysis. That's very common.

And as you saw as one of the UI layers for Dataflow, that Dataprep tool that you got a lot of practice with the last course, does exactly that. It will invoke that materialization step for a Dataflow and then write that out to BigQuery-managed storage.

Other Google Cloud tools, big data tools like Cloud Bigtable, you can export or copy that data from Bigtable into a BigQuery-managed storage.

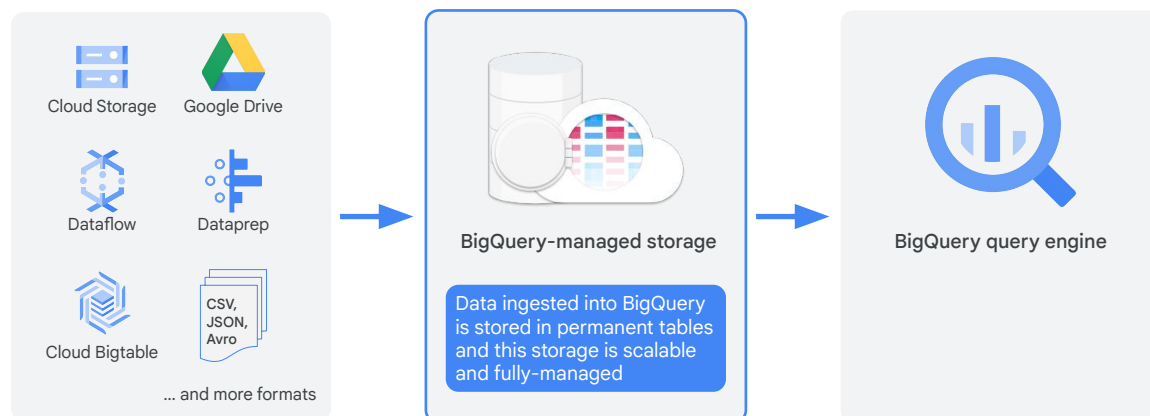
Ingest data permanently into BigQuery from a variety of formats



And of course, you can manually upload through your desktop or a file browser ingest those tables into BigQuery-managed storage.

So, why do we keep mentioning the word managed?

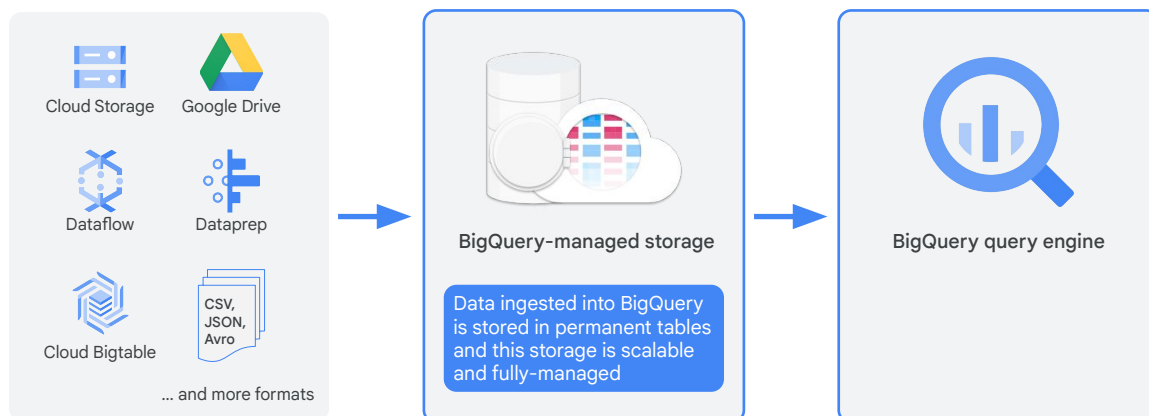
Ingest data permanently into BigQuery from a variety of formats



So that big concept or that big icon that you see there in the middle is a key core component of the BigQuery service.

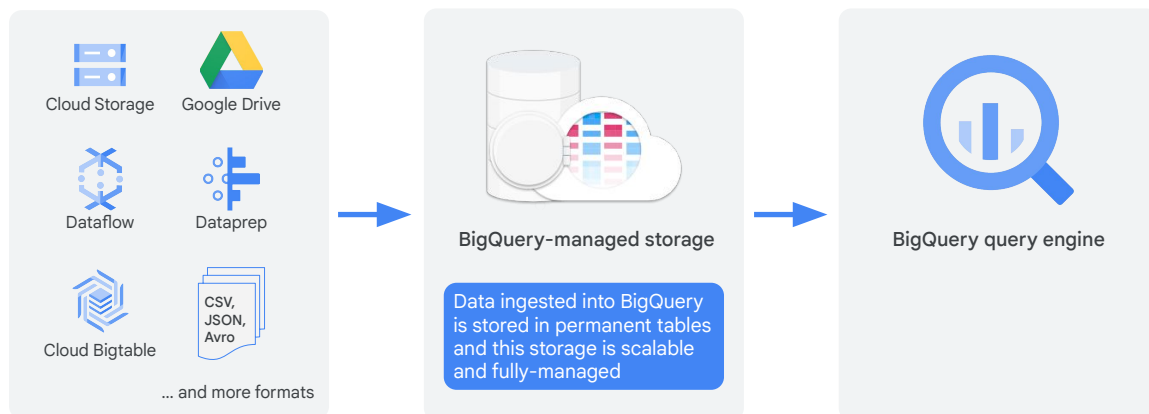
As we mentioned in one of the earlier courses, BigQuery is two components.

Ingest data permanently into BigQuery from a variety of formats



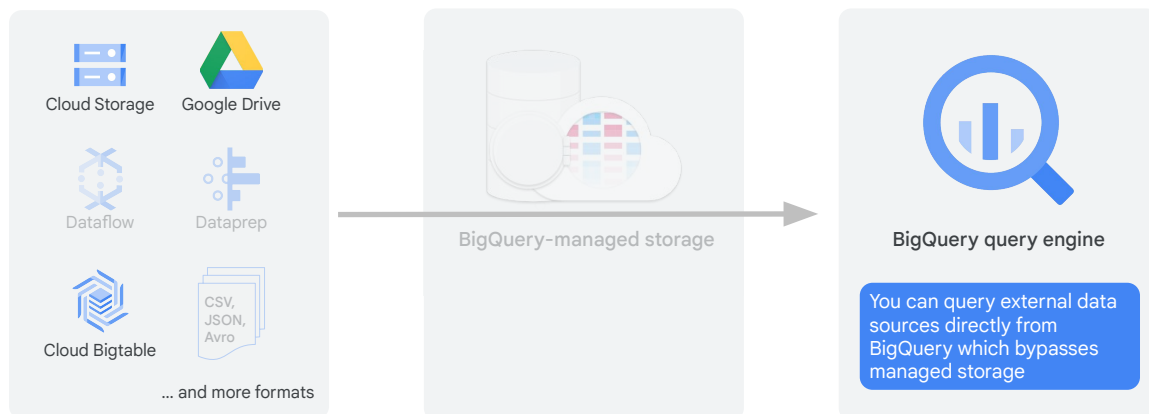
It's the query engine that will process your queries and it's also the data management piece behind the scenes that handles and stores and optimizes all of your data. So things like caching your data, storing it into column format and compressing those columns, which we're going to talk a little bit more about in the advanced course on the architecture of BigQuery, and expanding the data and making sure that it's replicated, and all these things that are traditional, like a database administrator wouldn't handle for you, the BigQuery team here at Google manages that for you behind the scenes.

Ingest data permanently into BigQuery from a variety of formats



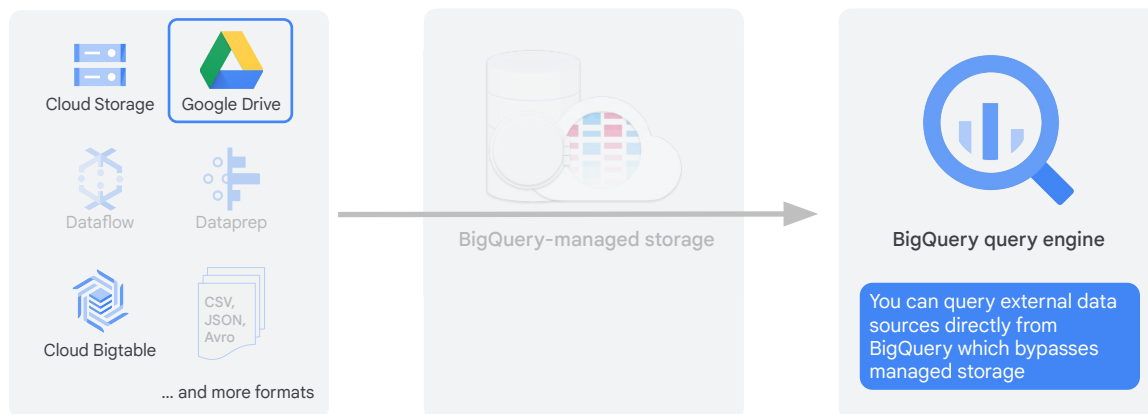
Why am I making such a big deal about managed storage? Because you might guess like, "Hey, all right. Cool. It's managed storage. I don't have to worry about that. When does my data never going to be in managed storage?" Right? And the answer is, it could quite possibly never even hit managed storage....

BigQuery can query external data sources in Cloud Storage and Drive directly



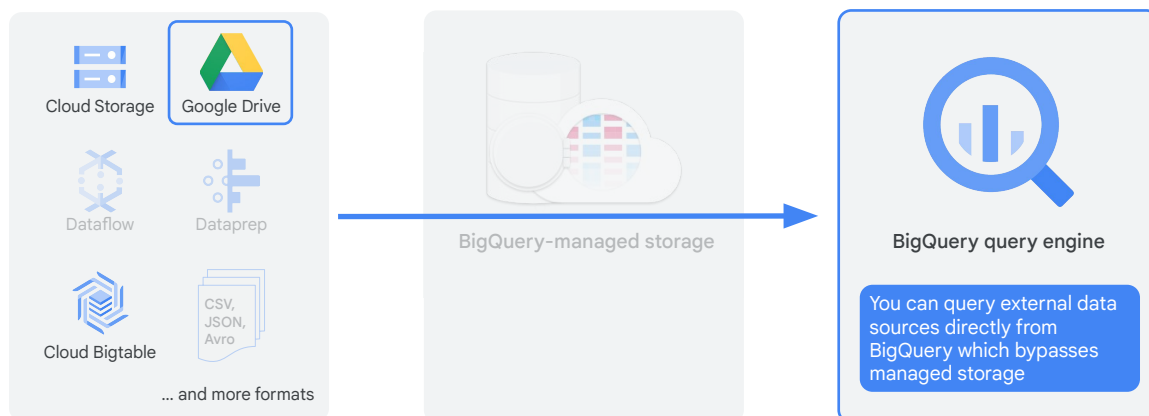
... if you connect directly to the external data source. This is like the mind-blowing concept, right? You can write a SQL query and that SQL query can be passed through and underline your actual data source.

BigQuery can query external data sources in Cloud Storage and Drive directly



It could be a Google Drive spreadsheet that someone is maintaining, and that data is not ingested and permanently stored inside of BigQuery. That's an extreme case because naturally you can see the caveats of relying on a collaborative spreadsheet as your system of record for a lot of your data. But this is a common occurrence for things like one-time extract transform, load jobs where you have a CSV that's stored in Cloud Storage, and you basically want to, instead of ingesting that data and storing that raw data inside a BigQuery, storing it in two places, a Cloud Storage and BigQuery, you instead query it, perform some preprocessing steps, clean it all up, and then at the end of that query, store the results of the query as a permanent table inside of BigQuery. So, that's one of the common use cases that I could think for creating or establishing this pointer or this external connection.




BigQuery can query external data sources in Cloud Storage and Drive directly



Now, as you see that big arrow over BigQuery-managed storage, you're just using the query engine. You get none of the performance advantages from BigQuery, the managed storage piece, and a lot of other drawbacks.

Pitfalls: Querying from external data sources directly

Limitations:

-  Strong performance disadvantages
-  Data consistency not guaranteed
-  Can't use table wildcards

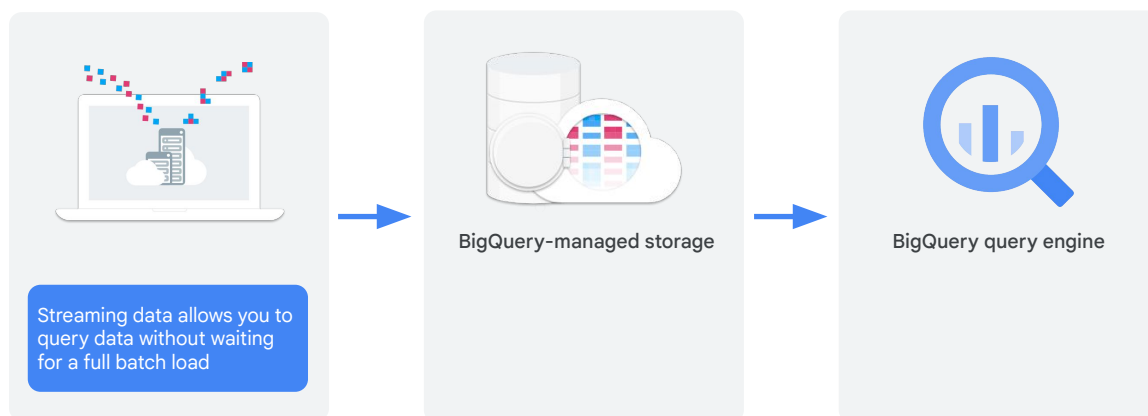
So, let's cover some of those limitations.

So, as performance disadvantages, there's a lot that goes into the special sauce of the BigQuery architecture behind-the-scenes, and what makes it much more performant to store your CSV data ingested permanently in a BigQuery as opposed to keeping it out on say a Google Spreadsheet or Google Cloud Storage. And a lot of those compression algorithms and the architecture of BigQuery, how it stores data in column format, we'll cover a lot in the architecture lecture coming up in the next course on advanced insights.

But one of the key things that should hopefully scare a lot of you away from using Google Spreadsheets as your source of truth for your underlying datastore could be data consistency. So, if you're writing out a BigQuery query as we mentioned in the previous slide, you have a BigQuery query that then is reaching out to a Google Drive spreadsheet. If you have folks that are editing that spreadsheet, the query doesn't necessarily know, "Well, hey, this is when I accessed it, at this particular timestamp, this is what the data was." If you have data in flux or in-flight, since it's not managed natively by BigQuery itself, there are few checks in place on whether or not the data that you're pulling was the data that you've expected when it was last updated in that particular spreadsheet as well.

And a lot of features that you can actually enable inside of BigQuery, like these table wildcards that we're going to discuss when we uncover unions and joins inside of our emerging data sets lecture, are unavailable outside of storing your data directly inside of BigQuery.

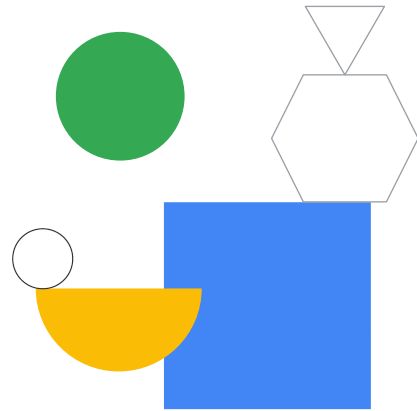
Streaming records into BigQuery through the API



We largely discussed batch loading a CSV or massive CSVs into BigQuery, but know that there is a streaming option available through the API where you can actually set it up, where you can ingest individual records at a time into BigQuery-managed storage and then run queries on those as well. So, the streaming API is well-documented and you guys can access that if you have a streaming or a new real-time data need for your application.

Lab Intro

Ingesting New Datasets into
BigQuery



Now it's time for us to ingest and query brand new data sources in BigQuery.

In this next lab, you'll practice loading data into BigQuery from external sources like Google Cloud Storage. You'll also learn how to set up an external data connection, but beware the caveats we discussed earlier.