Google Cloud

Big Data Tools Overview

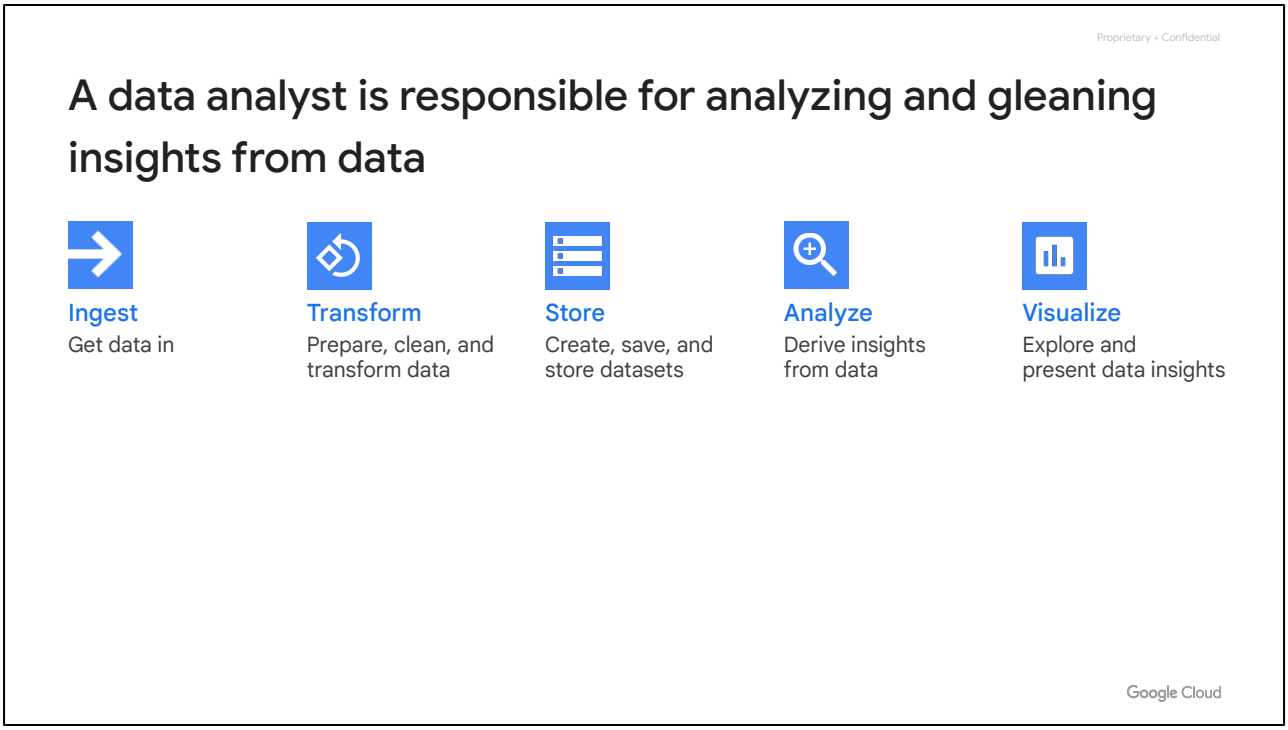
Evan Jones

In this next module, we will highlight the five common tasks of any data analyst and map those to their respective tools on Google Cloud.

Big data tools overview

01	Data analyst tasks and challenges, and Google Cloud data tools
02	9 fundamental BigQuery features
03	Google Cloud tools for analysts, data scientists, and data engineers

After that, we'll explore the BigQuery feature set itself and end with a discussion comparing data analysts, data scientists, and data engineers.



Okay, so before we get into the really cool part, which is showing you the useful big data tools on Google Cloud, we first have to talk about the data analyst tasks themselves as a whole.

So here are the five things that any data analyst worth their salt is going to perform.

You're going to ingest data, you're going to transform it, clean it up. All data is dirty data.

And then you're going to be creating some reporting data tables and storing that data for analysis, which is that fourth step. Finally, look how far we've come. It took four steps to actually take to get to the analysis portion where you're writing these cool, sophisticated queries to get insights from your data.

And then you're pairing that, potentially, with a visualization tool or platform to really make those insights shine and explain them to people.

Challenges in each task prevent data analysts from getting to scalable insights

Ingest
Get data in

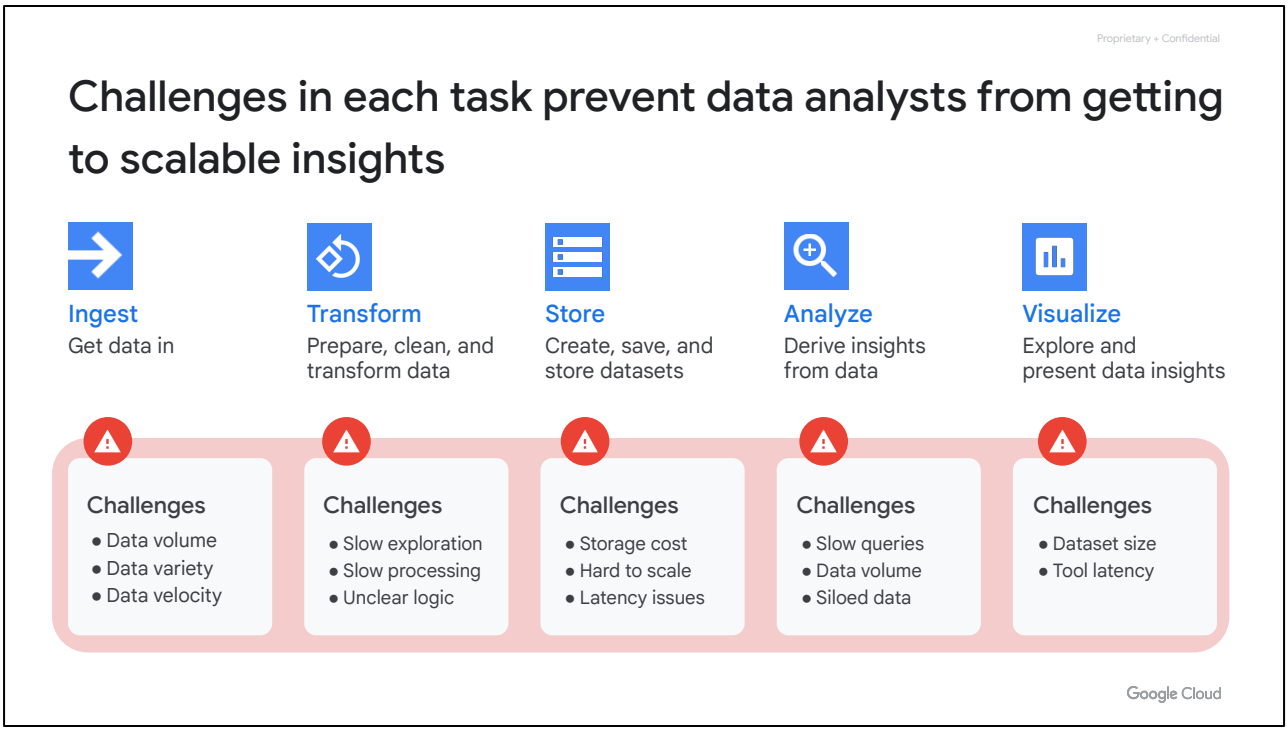
Transform
Prepare, clean, and transform data

Store
Create, save, and store datasets

Analyze
Derive insights from data

Visualize
Explore and present data insights

But the road is fraught with challenges. So at each of these different steps, as we saw with some of the challenges that organizations face, or data analysts have faced earlier on, each of these different steps has their own pitfalls.



So ingestion, you've got petabytes of data, it's going to bottleneck your tool. You don't even begin to imagine loading all of your data at once. So unfortunately, you're loading only in a sample, or you're looking only at a small amount of your data. So you can't really make amazing progress with loading all of your data in at once or it just takes forever.

Second, transforming the data. It's slow going. Perhaps you have to either rely on another team, a data engineering team, to write sophisticated pipelines to transform your data. And you wish there was an easier way to either write it yourself, or some kind of cool tool that'll help you build these things up in just a little bit of an easier way. And that was a clear spoiler alert for one of the tools you're going to be learning on the next slide.

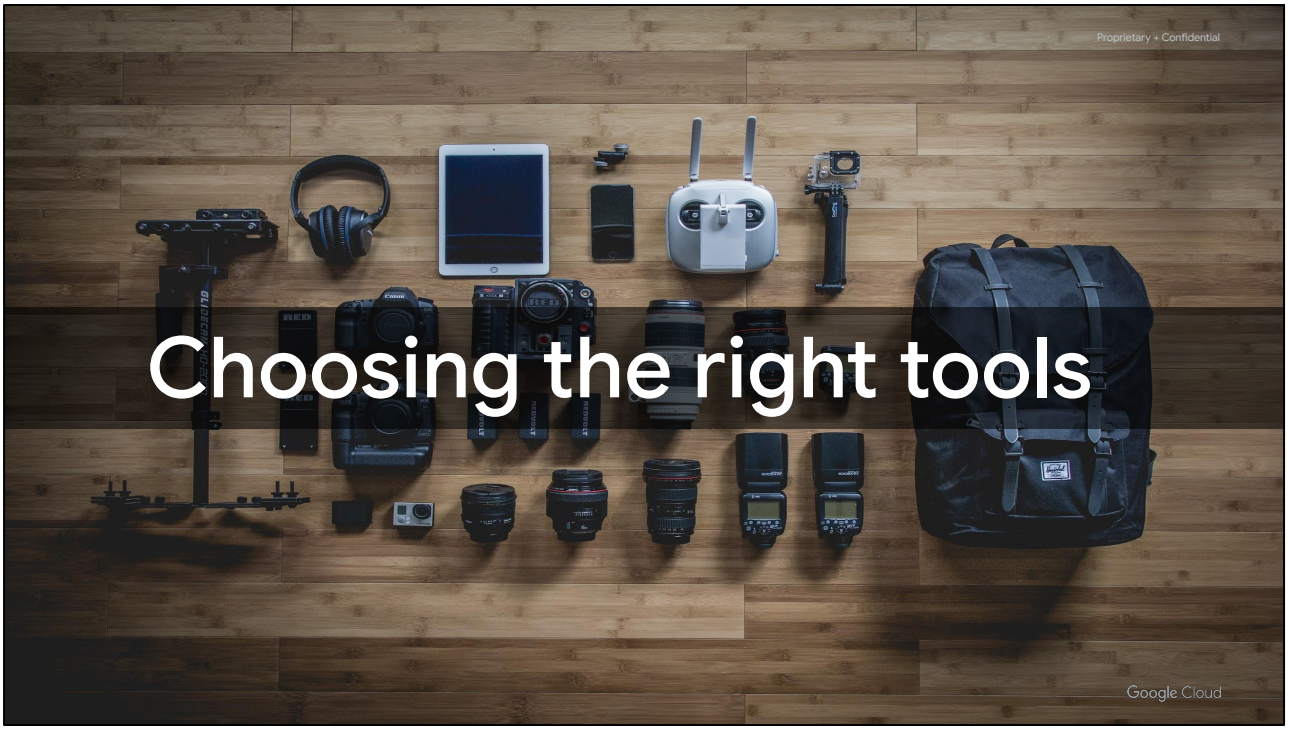
So onto storage. Scaling up the amount of data that you need to store, as we mentioned before, has been a problem for organizations that have managed their own hardware internally or relied on things that aren't as inherently scalable as relying on Google Cloud analysis.

Your queries are bottlenecking, your data is in many different places and there's no easy way to mash it together.

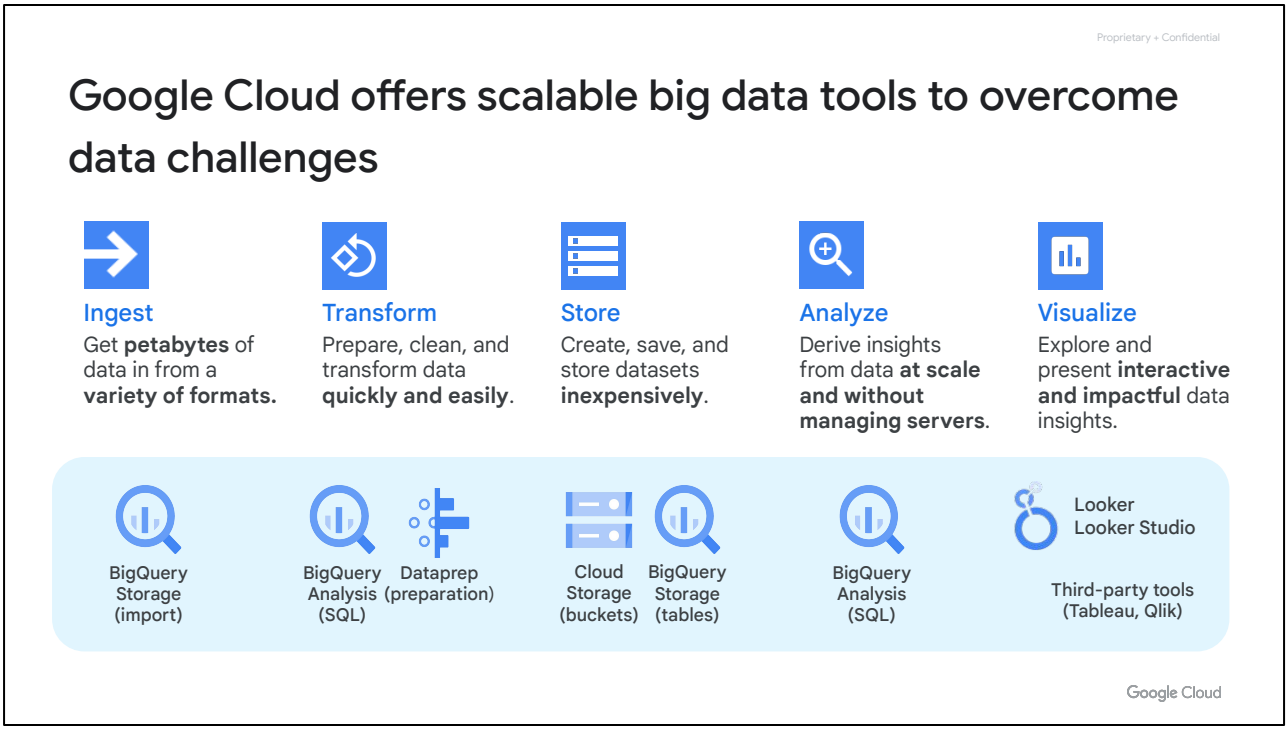
Visualizing your insights, you have amazing insights that you want to show. And as soon as you go to present it to your stakeholders and your peers, your tool starts to lag. You want to filter down and drill down into a particular insight and you have a

30-minute meeting. And, unfortunately, it takes the tool ten minutes to load and drill down into that insight. And then it's, you've lost the audience's attention by that point as well.

Let's see where the Google Cloud can step in. So here's the right tools for scalability, and this will allow you to address and overcome a lot of these challenges. So ingestion, Google Cloud, BigQuery in particular, is a petabyte-scale data analytics platform.



So here's the right tools…



…for scalability, and this will allow you to address and overcome a lot of these challenges. So ingestion, Google Cloud, BigQuery in particular, is a petabyte-scale data analytics platform.

And one of the great things that we're going to cover in the ingestion part, or the pricing lab that you're going to do, is actually importing data into BigQuery in batch form is free, which is great.

Transforming your data, so say you wanted to write some simple SQL. You can just do that directly inside of BigQuery. Or if you didn't even want to write any SQL, one of the cool labs we're going to do later on is using a tool called Cloud Data Prep where you can chain together, through a graphical user interface, a neat visual flow of how you want to process the data. So say you wanted to drag and drop a deduplication and then parse this particular field. You can do that visually, and you'll get a lot of practice with that as part of this course.

Storing data, again, we mentioned it a lot, Google Cloud Storage, inexpensive. BigQuery itself, you're going to see in the pricing lab, it's as of the time of this recording, is $0.02 per gigabyte per month. If the data is there for a long time, that storage cost is cut in half.

Analysis, that's really where BigQuery shines. I'm going to really go into the nine core parts of its feature set shortly. And this is managing scale, right, fully managed. No dev ops, managing it without you managing your servers, just write cool SQL.

Last but not least, visualization tools. Looker Studio is a free visualization tool that can sit on top of BigQuery. You can then let BigQuery processing do all the hard, heavy lifting. Enterprise customers who upgrade to Looker Studio Pro receive support and expanded administrative features. Tableau or QlikView are alternative third-party visualization tools.

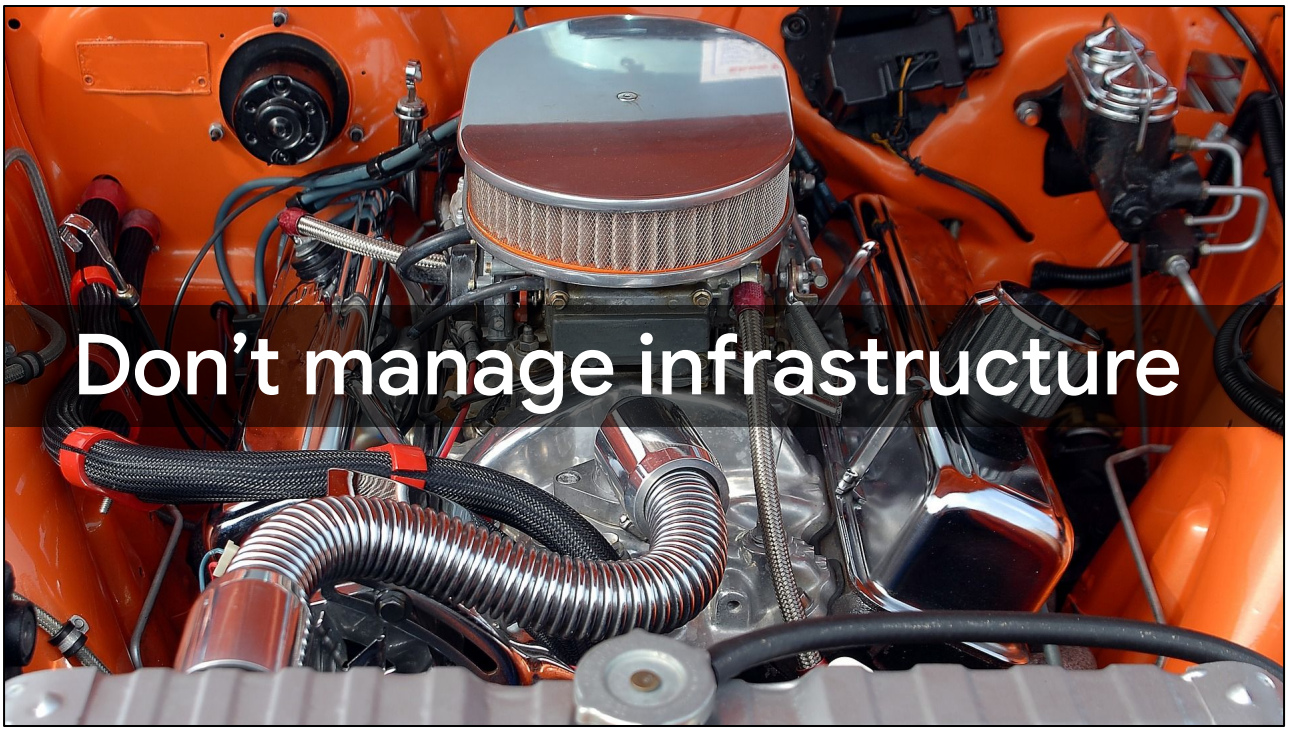
Big data tools overview

01 Data analyst tasks and challenges, and Google Cloud data tools

02 9 fundamental BigQuery features

03 Google Cloud tools for analysts, data scientists, and data engineers

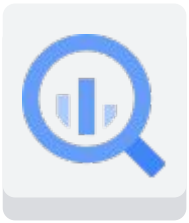
Okay, let's quickly cover the nine core features of BigQuery and then explain a little bit of the difference between data scientists, data analysts, and data engineers. And we'll get you launched into your next lab.



So BigQuery, the core tenet is, don't manage the infrastructure yourself. So if we haven't made this clear enough, if you don't want to be buying hard drives and managing that hardware especially. Even if you have it perfect today, a year from now when you're processing half that data or ten times that data, let the platform scale for you.

And it allows you to focus on finding those insights yourself. Become really, really good, not as the jack of all trades, in managing your hardware, writing the queries, and doing the job of ten different people. Become extremely deep and proficient at mining those insights. And if you wanted to double down and continue to take these additional courses and pick up how to do machine learning. It's, again, along this ramp of really focusing on all those insights instead of infrastructure.

BigQuery is a petabyte-scale data analytics warehouse



BigQuery

1. Fully-managed data warehouse	2. Reliable	3. Economical
No-ops, petabyte-scale	Backed by Google data centers	Pay only for the processing and storage you use

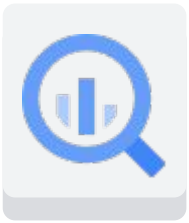
Okay, so BigQuery in a nutshell. So nine key points here.

So it's the fully-managed petabyte-scale data warehouse.

It's, as you saw with the pictures, backed by Google data centers.

The economies of the cloud mean that you pay for only what you consume, plus the cost of storage if you're creating those permanent tables.

BigQuery is a petabyte-scale data analytics warehouse



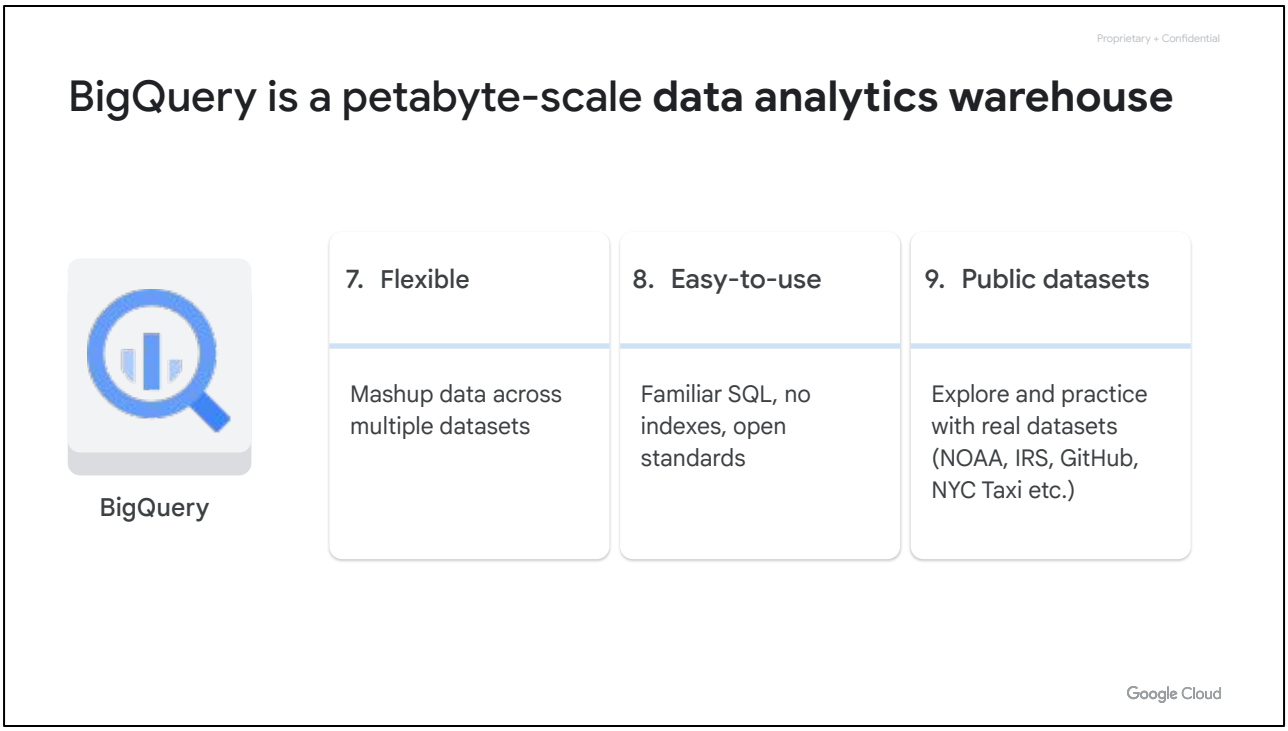
BigQuery

4. Secure	5. Auditable	6. Scalable
Role ACLs, data encrypted in transport and at rest	Every transaction logged and queryable	Highly parallel processing model means fast queries

Next up, security. It is access controlled and you actually, within your Google Cloud project, can manage access for members and groups and whoever needs to actually have access to your datasets as well. And we'll cover a lot of those in one of the data access modules as part of the third course. So your data itself, if you're more concerned with how Google is storing it, is encrypted in transport in at rest for your data centers and replicated.

Each of the different queries is actually logged as a separate transaction. If you're familiar with Cloud Monitoring, you can actually monitor all of those queries that are going through and then look back through for history as well.

And the big benefit, as we mentioned, is scalable, BigQuery. So you can process multiple queries in parallel, actually, up to 50 concurrent queries at the exact same time, going on at once.



Last points, with a little bit of SQL you can use joins and unions and bring your data together across many different datasets and really break apart those data silos.

Number eight, the reason why you're, hopefully, taking this course is you want to get better with your SQL, because BigQuery loves when you write excellent awesome SQL to get those insights out of it. And another interesting point, if you really like the behind the scenes architecture, is BigQuery actually has no indexes and it even actually has no keys, if you're familiar with database terminology. So, again, as part of an analytics warehouse, that architecture and performance discussion is something that's very, very fun to dive into. And, again, it's built on open standards as well.

Explore around and find pre-built queries and examples online of over 50 different datasets that are available for you to explore as well. And so, as I like to say, good artists copy, great artists steal. So if you find a query that looks good that somebody else has written, steal it, modify it, make it work for you.

Three ways to interface with BigQuery

01

Web UI

Build, validate, and run queries quickly through the Web UI.

This will be our primary focus for this course.

02

Command-Line Interface (CLI)

Use Cloud Shell or the Google Cloud SDK (gcloud) to interact through a terminal.

```
bq mk [DATASET_ID]
```

03

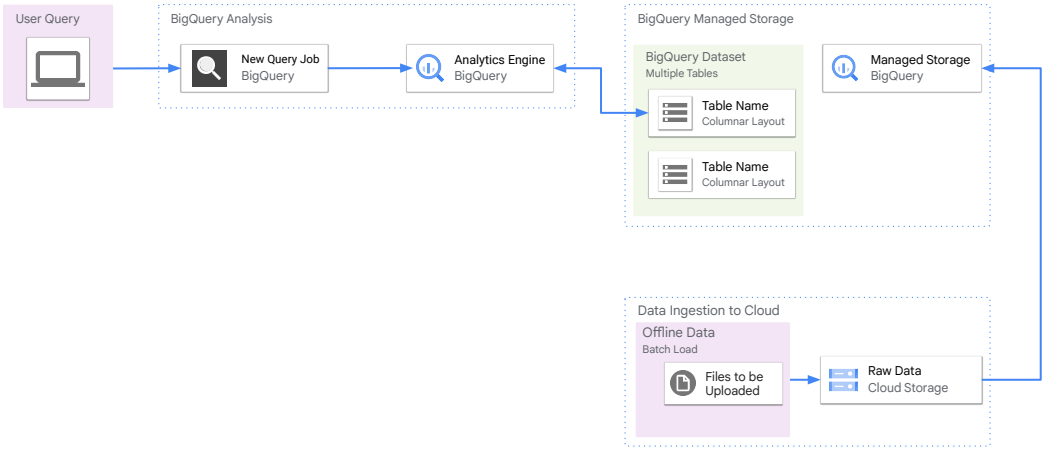
REST API

Programmatically run queries using languages like Java and Python over HTTP.

```
GET
https://www.googleapis.com/bigquery/v2/projects/projectId/queries/jobId
```

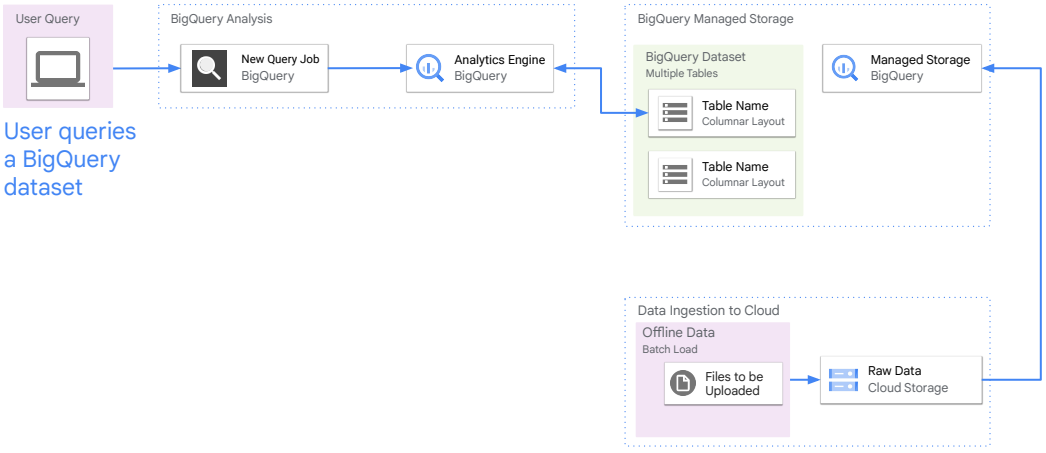

All right, so we mentioned the three different ways to access BigQuery. Primarily going to be focused on the web UI for this course series. You can also access it through the command line, which is great. And then, much like if you wrote any other application, you can get your queries ran over the web through RESTful APIs.

Creating and querying datasets: BigQuery terminology



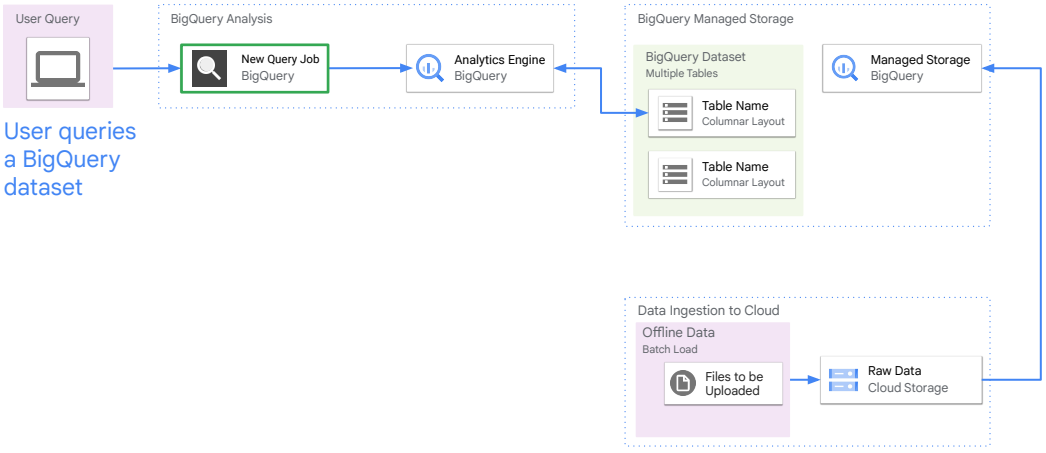
Just a quick architecture diagram here to kind of get a lot of these terms cleared up. So, starting with the left.

Creating and querying datasets: BigQuery terminology



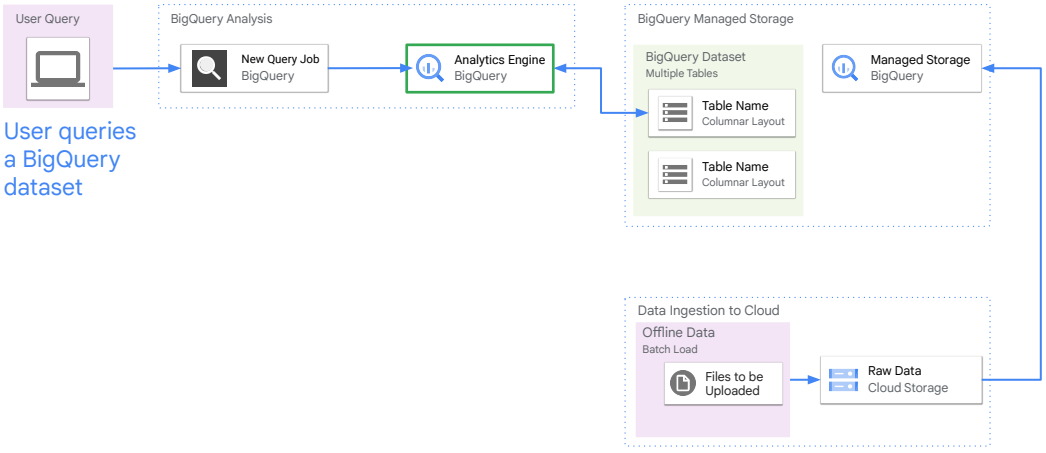
User will query BigQuery.

Creating and querying datasets: BigQuery terminology



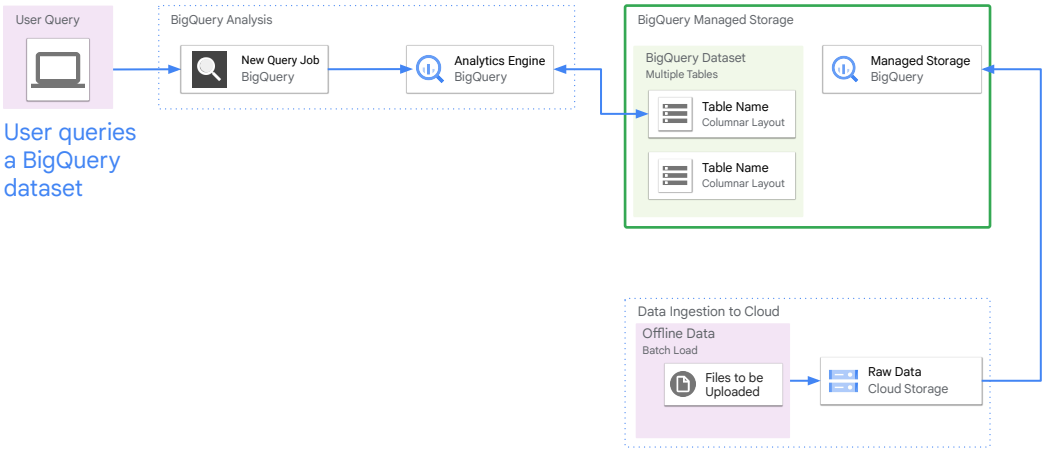
A unit of work in BigQuery itself is called a job. We'll revisit the job when we talk about BigQuery pricing later on.

Creating and querying datasets: BigQuery terminology



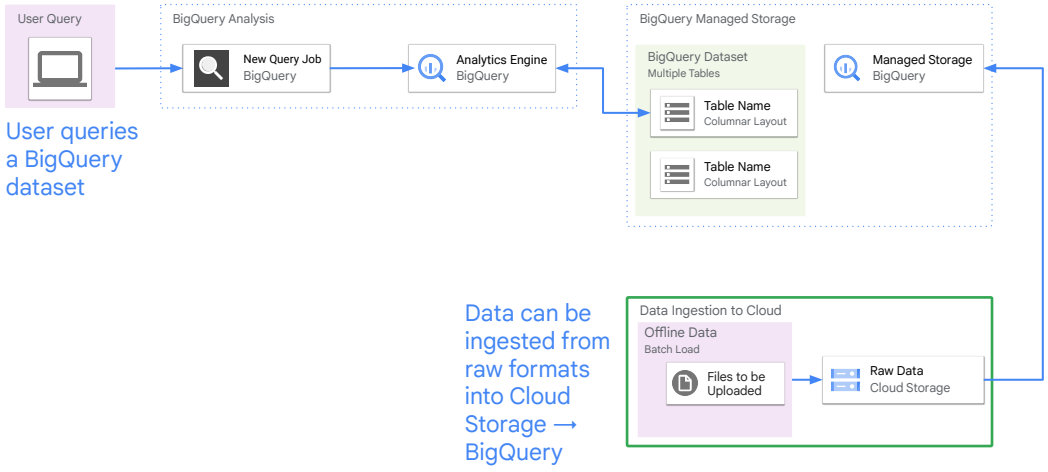
Jobs run on a very fast analytics engine that was developed internally at Google,

Creating and querying datasets: BigQuery terminology



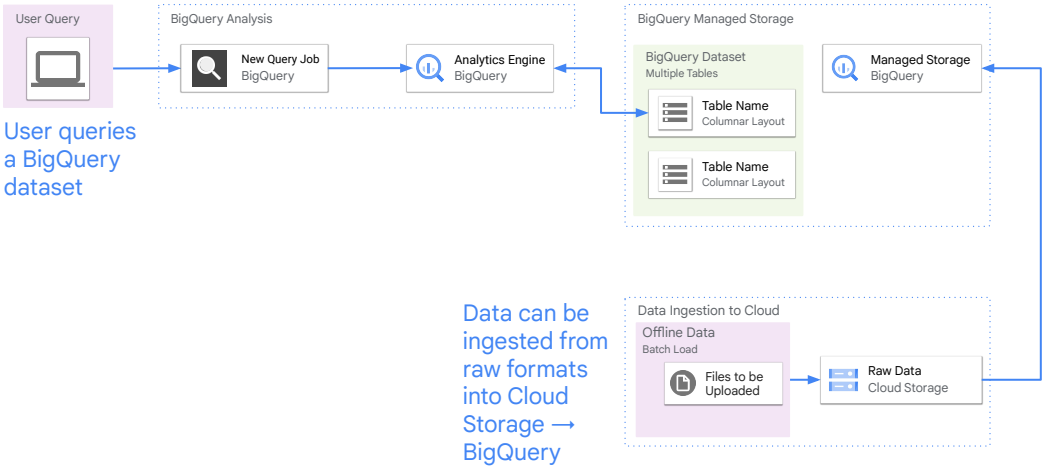
and then made available as a service through BigQuery. And then those query jobs are then mapped to the underlying data, which is fully managed behind the scenes in those tables.

Creating and querying datasets: BigQuery terminology



And then, walking back the other way, all the way at the bottom there, you can ingest data into something like Google Cloud storage if you wanted to. Or directly into BigQuery if you wanted to, and then have that be available for analysis.

Creating and querying datasets: BigQuery terminology



But the key takeaway from this slide is at the top you have the BigQuery Analytics engine in that one box, and then you also have the BigQuery Managed Storage. See you have this powerful query engine, and you also have this replicated scalable storage for all your data that is being stored. So it's actually two technologies, or two services in one.

BigQuery is actually two services in one

BigQuery

BigQuery Managed Storage

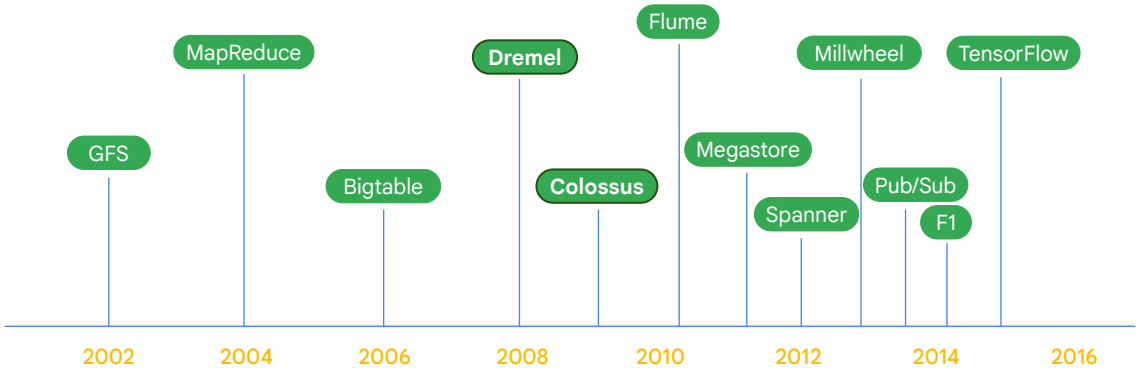
Fully-managed and ***scalable data storage*** that is based on the same technology that stores Google's product data (ads, gmail etc.)

BigQuery Analysis

Fast massively parallel ***SQL Engine*** based on Google's own internal Dremel query engine technology

So, Google BigQuery is that managed storage piece, which is scalable and it's the same technology that stores a lot of Google's product data, right? Think ads, Google email service, Gmail,. But it's also that really lightning fast analytics engine, SQL engine, and it's built on the massive evolution of Google technologies over time. The relentless march, if you will, to keep performing better and better. because Google is naturally incentivized because of the amount, massive amounts of data that it has

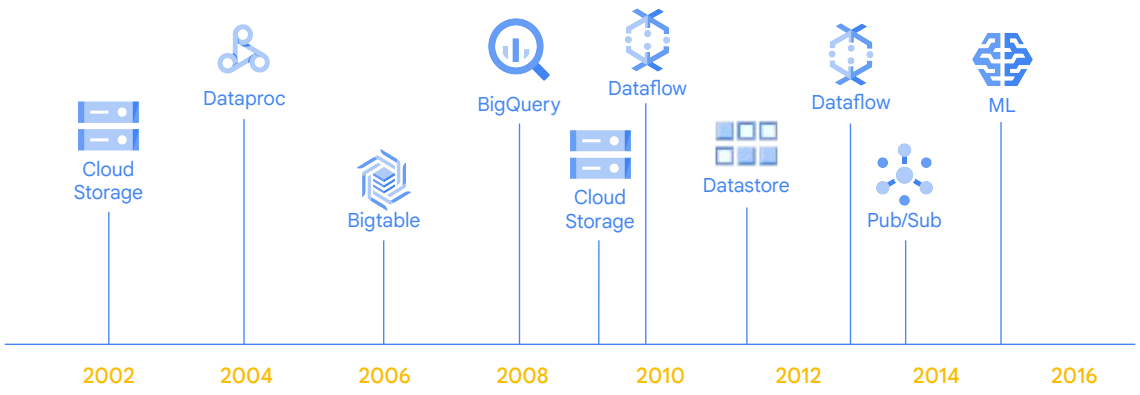
Google innovates data technologies



Google Research Publications referenced are available here: http://research.google.com/pubs/papers.html
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2009
http://research.google.com/pubs/pub35290.html

So Google loves to innovate data technologies. So there is a lot that are focused on here. One of the words that may immediately look familiar to those who have been around the big data block for a while is MapReduce. So in 2004, Google Research actually came out with a white paper that became MapReduce, and then open-sourced it, which was then used as the foundation for Hadoop, which is that massive parallel-processing, right? Bits of data mapped with tasks, and then processing all that in parallel. Not content with that, in 2008, it released the Dremel white paper which is processing queries over smaller chunks of data but doing it massively in parallel, and having that done through SQL. And that, the Dremel technology, plus Colossus, which is the massive hard drive in the Cloud, those two technologies form the basis…

Google Cloud opens up that innovation to you



Google Research Publications referenced are available here: http://research.google.com/pubs/papers.html
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2009
http://research.google.com/pubs/pub35290.html

….of what was then BigQuery and Google Cloud Storage as well. So as you can see, Google has opened up those technologies to you as part of the Google Cloud, and continues to innovate. If the other technologies here interest you, Dataflow, again, is one of those data engineering tools where you can build those massive data pipelines, ingest streaming data, and batch data and then dump it into BigQuery.

If machine learning is your game, learning things like TensorFlow as part of additional courses, is also one of those great technologies that's available through Google Cloud as well. And if your ultimate end result is to get to machine learning, stick around for the third course in this course series where we'll cover a lot of the initial introductions to some of the tools.

Google Research Publications referenced are available here:
http://research.google.com/pubs/papers.html
The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2009 http://research.google.com/pubs/pub35290.html

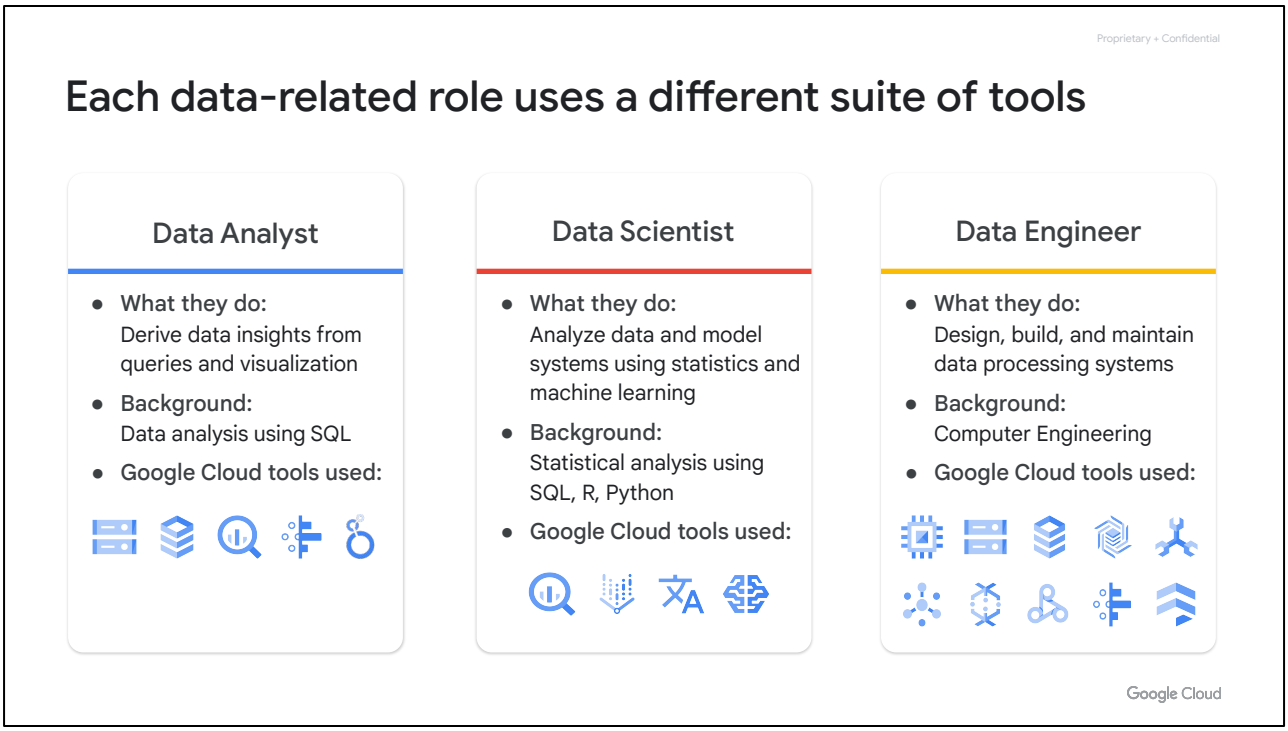
Big data tools overview

01 Data analyst tasks and challenges, and Google Cloud data tools

02 9 fundamental BigQuery features

03 Google Cloud tools for analysts, data scientists, and data engineers

So last but not least, I did mention machine learning and there's no discussion in machine learning that you can't have without explaining the difference between what is a Data Scientist do versus what a Data Engineer does versus a Data Analyst? So let's jump into a quick discussion to that.

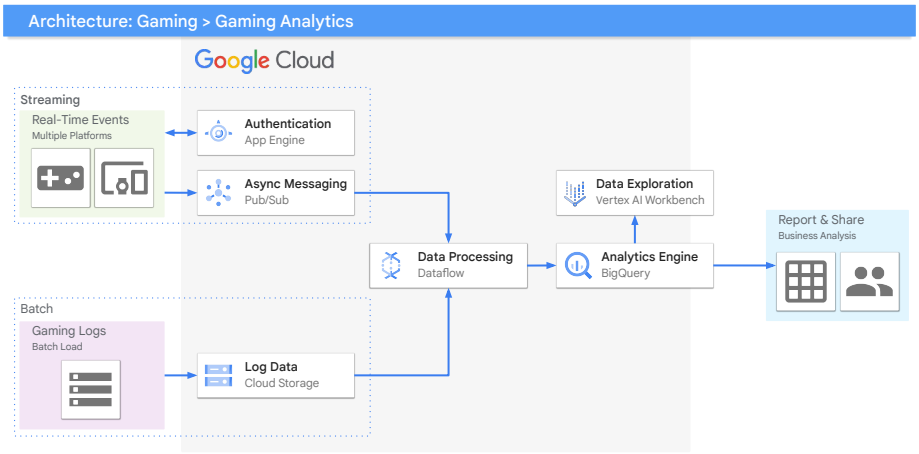


All right. So Data Analysts, you are the target audience for this course. So we do have a Data Engineering course after this one. So if these concepts and data really just excites you and you're like, "Man, I really want to build machine learning models", get through this course and then do an amazing job at learning a lot of these core concepts because all of these data concepts build on each other, one after another, after another. You can't build machine learning models without having first a fundamental understanding of processing your data and understanding how to sample it and understanding what is dirty data, right? So, largely a merge of technologies between the three with kind of the key differences between the rules, right? Data Analysts, like those of you taking this course, you'll be writing sequel inside a BigQuery Web UI and visualizing that data.

Data Scientists will be using something like online collaborative notebook, like a IPython notebook, and you can have more of a statistics background, and you'll be building these machine learning models, right? And you could be using SQL, but in addition, other languages like R and Python.

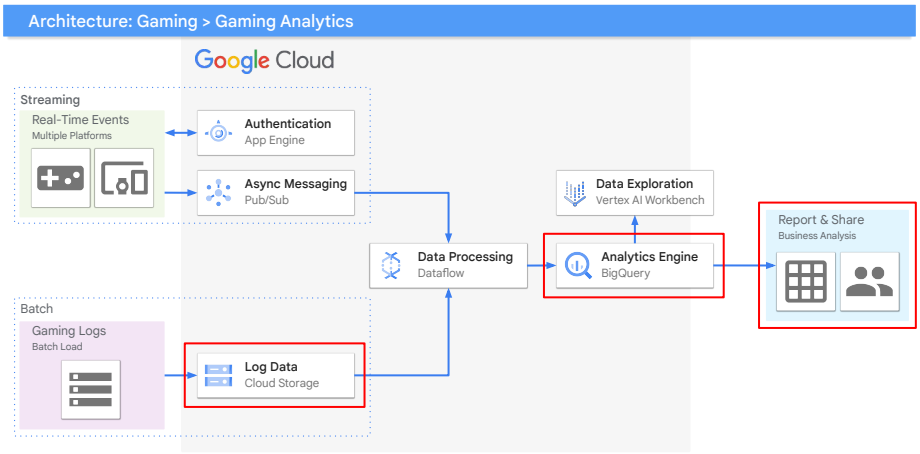
Data Engineers, by contrast, you're building these amazing pipelines of torrents of pair bytes of data, right? That are coming into these systems. So you have logs data that's being streamed from an online video game. And you just store that data somewhere and then you need a pipe that your data analysts in BigQuery for that to be available for analysis as well.

End-to-end gaming analytics example highlighting Google Cloud tools



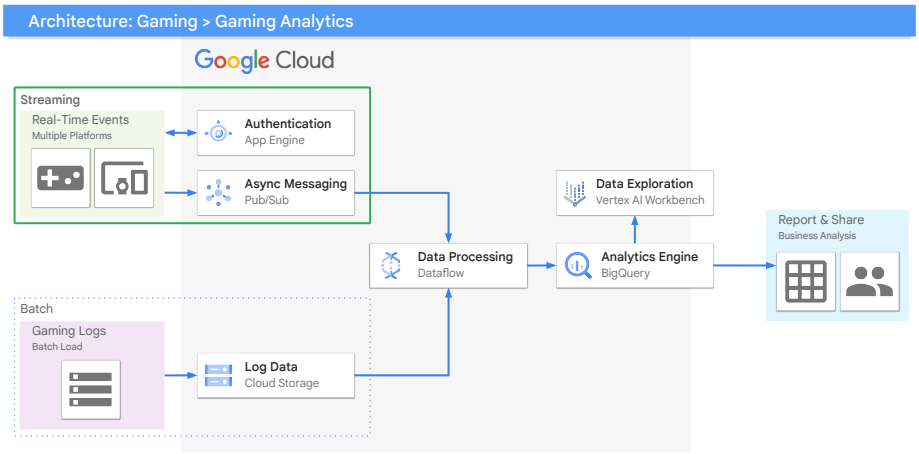
Let’s close with an example of online gaming analytics.

End-to-end gaming analytics example highlighting Google Cloud tools



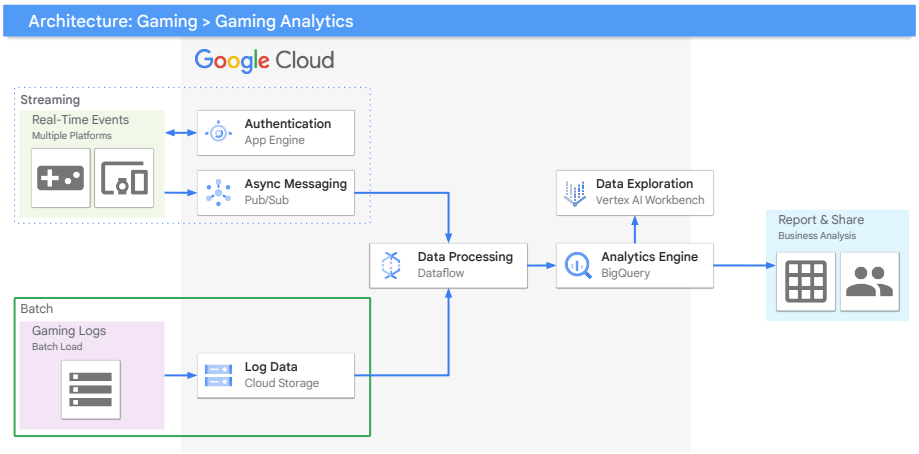
Highlighted in red is where your Data Analysts would focus their attention. So you have two different types of data.

End-to-end gaming analytics example highlighting Google Cloud tools



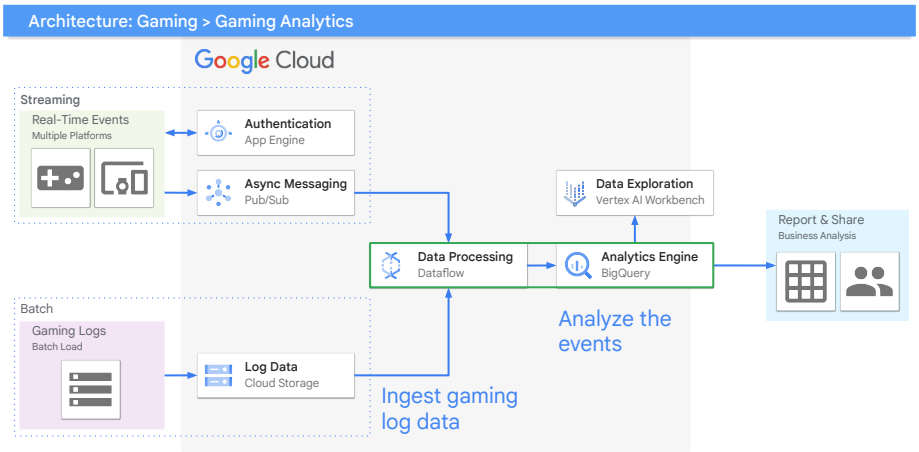
You have streaming events coming off a video game online.

End-to-end gaming analytics example highlighting Google Cloud tools



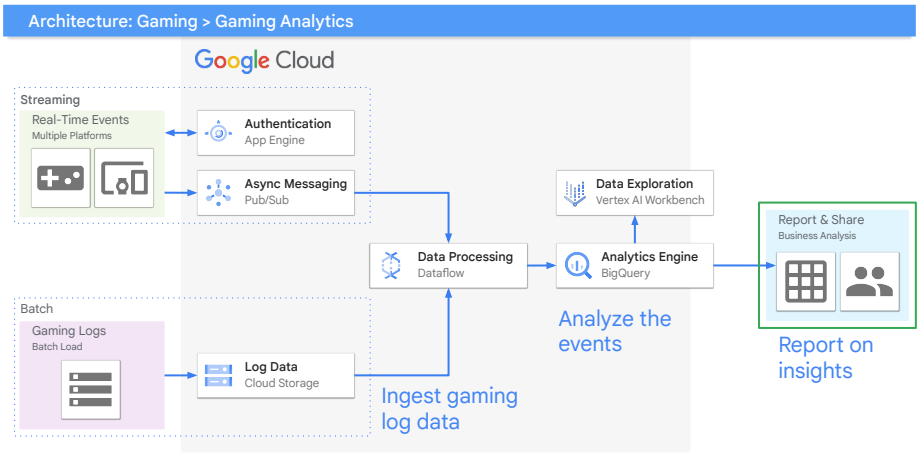
You also have massive logs of data that are loaded in batches to a Cloud Storage staging area through a bucket resource.

End-to-end gaming analytics example highlighting Google Cloud tools



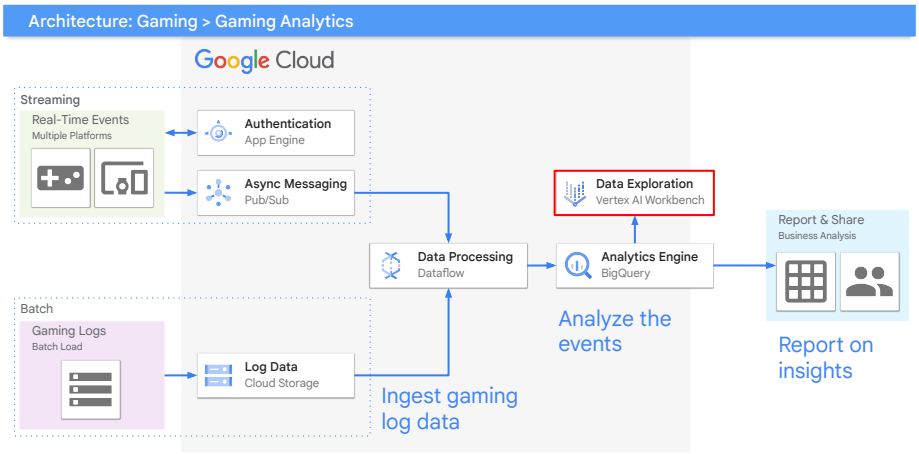
Data Engineers could use a service like Dataflow to build a pipeline. You could then pipe in massive amounts of data into a data warehouse like BigQuery to perform ad hoc queries.

End-to-end gaming analytics example highlighting Google Cloud tools



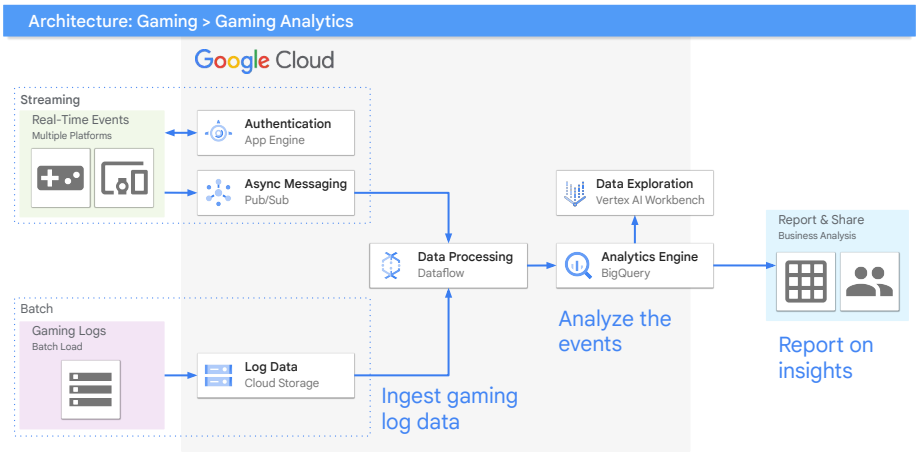
Further right of that, for your reports, you could use an Analytics tool like Looker Studio to analyze, explore, visualize, and present that information.

End-to-end gaming analytics example highlighting Google Cloud tools



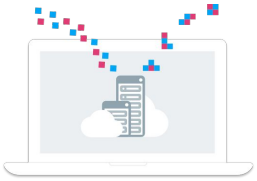
Or, if you're a Data Scientist, you can plug in Vertex AI as a layer on top of BigQuery and invoke those queries to pre-process your data to build something cool like a machine learning model.

End-to-end gaming analytics example highlighting Google Cloud tools



That completes a walkthrough of some of the tools in the toolbox. Remember, this is just an introduction to these big data tools for analysis. If you're interested in these technologies, I encourage you to explore them further.

Summary: Review data analyst tasks and tools



Reviewed data analyst tasks: ingest, transform, store, analyze, and visualize data.

Data analysts will use Cloud Storage, BigQuery, Dataprep, and Looker Studio.

Explored the nine features that make BigQuery a petabyte-scale data analytics warehouse.

Compared data analysts, data scientists, and data engineers.

To summarize what you've covered so far.

You looked at the lifecycle of Data Analyst tasks and mapped each of those tasks to the right tools to use on Google Cloud.

You also compared the data roles and tool sets used by Data Analysts, Data Scientists, and Data Engineers.

While this course is primarily targeted toward Data Analysts, it provides an insight into the more advanced tools and topics that are covered in greater depth in other courses like Data Engineering on Google Cloud.