



Exploring and Preparing your Data with BigQuery

Evan Jones

Again, I'm Evan Jones, one of the course designers for Data to Insights. I've been teaching data analysis for over ten years. My life at Google before developing courses like this one was in Google Finance, where we built pretty fun machine learning models to predict and optimize expenses here at Google. And I'm thrilled that Google has made their internal petabyte-scale data analysis tools available to the world, through Google Cloud. It's that platform that we're going to be using to explore and derive insights using their big data tools.



Course agenda




- 01 Introduction to Data on Google Cloud
- 02 Big Data Tools
- 03 Exploring your Data with SQL in BigQuery
- 04 Pricing

Let's take a quick look at the agenda of topics we're going to cover. First, we'll start with the basics of Google Cloud, and why letting the cloud handle your compute and storage needs enables massive scalability.


After the fundamentals of cloud, we'll go into the big data tools, available to you as a data analyst.

Third is where we'll start coding in SQL, or the Structured Query Language.

Fourth, we'll explore the BigQuery pricing model for query processing and data storage.



Course agenda




- 05 Cleaning and Transforming Data
- 06 Storing and Exporting Data
- 07 Ingesting New Datasets
- 08 Visualization Basics with Looker Studio


Next stop is a discussion on dirty data and how we can clean it up with SQL or a new UI tool.

Sixth and seventh on this list is how you can create and store your own data sets of BigQuery, from your queries or from external data sources.

We'll close here with an introduction to visualization and how to create reports.



Course agenda




- 09 Joining and Merging Datasets
- 10 Advanced Clauses and Functions
- 11 Schema Design and Nested Data Structures
- 12 Advanced Visualization with Looker Studio


Moving on to some of the more advanced topics we're going to cover, you're going to look at joins and unioning your datasets together in BigQuery, as some of the more advanced statistical functions and user-defined functions you may not have seen before.

Afterwards is one of my favorite sections on how repeated fields in Arrays work within BigQuery's nested data structures.

Again here we'll close with some more advanced data visualization tips.



Course agenda



13 Optimizing for Performance

14 Data Access

In these last sections, we'll walk through one of the most popular topics which is troubleshooting query and dataset performance.

Lastly, before wrapping up, we'll close this course series with a critical topic of data security and access control.

Introduction to Data on Google Cloud

01 Analytics challenges faced by data analysts

02 Big data on-premises versus in the cloud

03 Real-world use cases

04 Google Cloud project basics



So first and foremost, we're going to take a look at those challenges that are faced by data analysts. So let's just jump right into those.

Data analysts face query, infrastructure, and storage challenges



My queries are taking way **too long** to run and is stalling my analysis.



We're a data department, not an **infrastructure** department. Maintaining and upgrading our own servers is unsustainable.



We can only **afford to store a subset** of the data our business generates.



I have no easy way to **combine and query** all the data I've collected.



My on premises clusters **aren't scaling** with my analysis.



We don't have a **central data** analytics warehouse or set of tools.

So if you run any queries in your life, particularly like when I was learning database processing in school. My instructors and teachers would say, hey, run this one query and then you can go to the bathroom or do whatever you need to do while your query is running, right? So upper left you see that queries are taking too long, like potentially stalling your analysis.

Or what about if I wanted to combine 15 data sources and query all of them. And I want to do that within a reasonable amount of time. A lot of times, that was hard to do.

And in the middle say, it wasn't a querying problem it was actually an infrastructure problem. I'm a data analyst, a data scientist, I'm not a hardware purchasing department. I don't know about buying servers, and storing multiple versions of hard drives that are redundant in case of a hard drive product fails. And I have to maintain the network of all of my data as it relates to processing my queries and accessing the data where that's stored. I don't want to deal with any of that kind of infrastructure, right? But I have to, as a necessary evil if I want to be a big data shop, right?

Or if you're using, say, like Hadoop on your clusters, you're managing your clusters but you've had this amazing capital outlay to get this awesome processing cluster. But now you're punished by your own success because now your clusters can't scale. Because your organization says, you did such an amazing job, now we have ten times the data, can your clusters handle it or do you need to buy more and kind of keep expanding out your ever growing infrastructure empire? And again, it's how much of the business of building infrastructure do you want to be in versus spending

that opportunity cost of infrastructure versus writing out those amazing queries or those machine learning models to get those insights.

Next, is a pretty apparent one, which is just to get cost. So maybe you have a ton of data, you have a torrent of data but you literally can't afford to just process all of it. Just because performance wise, it's prohibitive on your machines and you can only create a few columns. Or it's just the monetary cost, which processing that much data and storing that much data is just prohibitive.

And last but not least, if you have no central place where you can just dump all this data into like a staging area or an analytics warehouse, that could be a problem as well.

These are a lot of the same exact problems that Google had kind of growing up, right? And faced with a torrent of search indexing data and adds volume data. The necessary problems that Google as a big data organization had to solve. And we'll see exactly how they did that, and the benefits of technology and time that have evolved to create a lot of these cool Google Cloud tools, the big data tools like BigQuery.

Introduction to Data on Google Cloud

01 Analytics challenges faced by data analysts

02 Big data on-premises versus in the cloud

03 Real-world use cases

04 Google Cloud project basics



So let's compare a little bit about, all right, so maybe you don't have Google BigQuery yet. Maybe you have something that's on-premises.

Reasons why Google Cloud is used for data analysis

- Storage is cheap
- Focus on queries, not Infrastructure
- Massive scalability

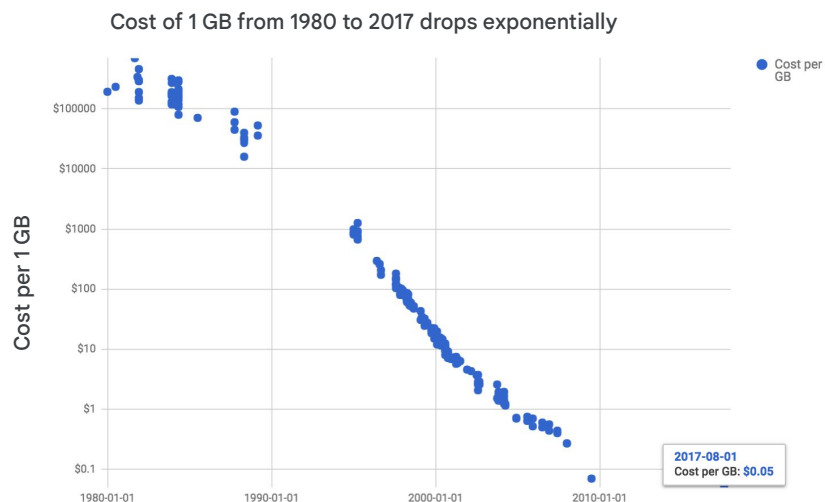
And why you should potentially move to the cloud, Google Cloud, over a potentially an on-premises solution you have.

So storing data and storing objects on Cloud Storage is super cheap as you can see.

And again, as we mentioned in the New York City cab example you can focus on building awesome queries and wowing your audience, instead of wowing your electricity bill by paying for rooms and rooms full of servers.

And again, I joke about this stuff, but it's only in the last 10, 15 years that a lot of this has become available. That you as a data analyst can leverage the power of literally, just planet-scale data centers that Google has built for its own operations and then made available through Google Cloud to you all. And that speaks to the point of massive scalability is it's built to Google scale.

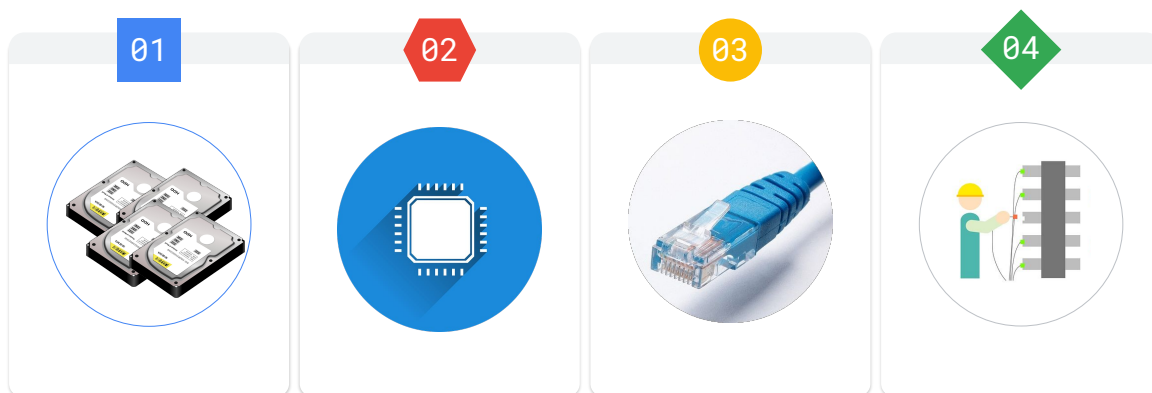
The cost of 1 GB of storage has dropped dramatically



Google Cloud

So this is just a very fun graph to show. If you're trying to process even just gigabytes of data back in the day, back in the 80s, you're looking to pay upwards of \$100,000 per gigabyte. Whereas now it's \$0.05 to equivalently buy one of these hard drives online. And of course, buying terabyte hard drives now is, consumers can buy them now. But again, that's not the only piece that goes into processing big data. Yeah, storing data is fantastic. You need to have a ton of storage. But that's not the only piece to actually process big data.

Traditional big data platforms require an investment in infrastructure



Google Cloud

So what are the other pieces that are involved?

So you got your hard drives, great, you're sure storing the data.

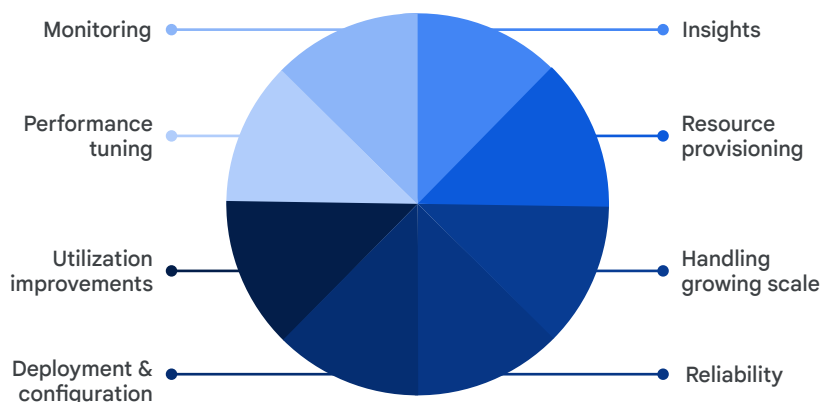
But if you're going to ask computation questions like say I wanted to do the process, how fast those New York City taxi cabs are moving. So I need to do it like the time over the distance. And it's going to be a computation that I need to run. So I gotta have some kind of processing power that I can then attach to my hard drives. So okay, I need a server.

And then okay, well, in order to talk to these servers we need some kind of networking. That's that step three there. It needs to be fast. But if my data analyst team is in London and my servers are located somewhere else, say home base is somewhere else in the world. That speed to reach that data could be throttling our network, as well.

Not to mention the army of folks like database administrators and server technicians as part of step four, that you need to actually maintain this massive amount of growing infrastructure that you have, as well. So why do it, right?

Typical big data processing

Time to understanding



Google Cloud

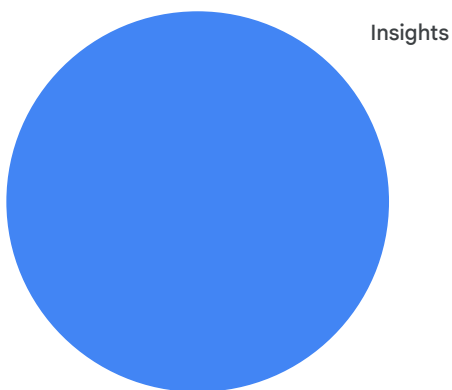
So here's another kind of scary pie chart. So you run in the door as a data analyst, you run in the door as an analyst. And you really want to spend your time creating those insights. But at the same time, it's a small shop and you have your own cluster that you need to maintain. And as part of that, you've been drafted into the apprenticeship for your hardware team as well. Because in order to maintain the awesome querying on the boxes that you have, you need to also monitor them in case hard drives fail. You need to performance tune them for specially written queries or jobs that you have running.

Any type of software updates that comes out, you need to manage and install those, make sure they're all backwards compatible. Provisioning of the resources that you have. So say you have a marketing team that also wants to query using your cluster, as well. But you have a data analysis team that takes higher priority. You have to manage all that, as well.

Not to mention that say next year, you did such an amazing job that your CEO wants to say okay, you guys have done a phenomenal as data analysts. We're going to give you 10 times or 100 times more data. And you'll need to process that, as well, with your growing and growing infrastructure. So it becomes a battle again, of growing of that infrastructure versus growing deeper into more and more sophisticated analysis techniques and things like building machine learning models as well. So it becomes almost unsustainable to keep building and building and building and maintaining infrastructure if your core focus wants to be on data analysis.

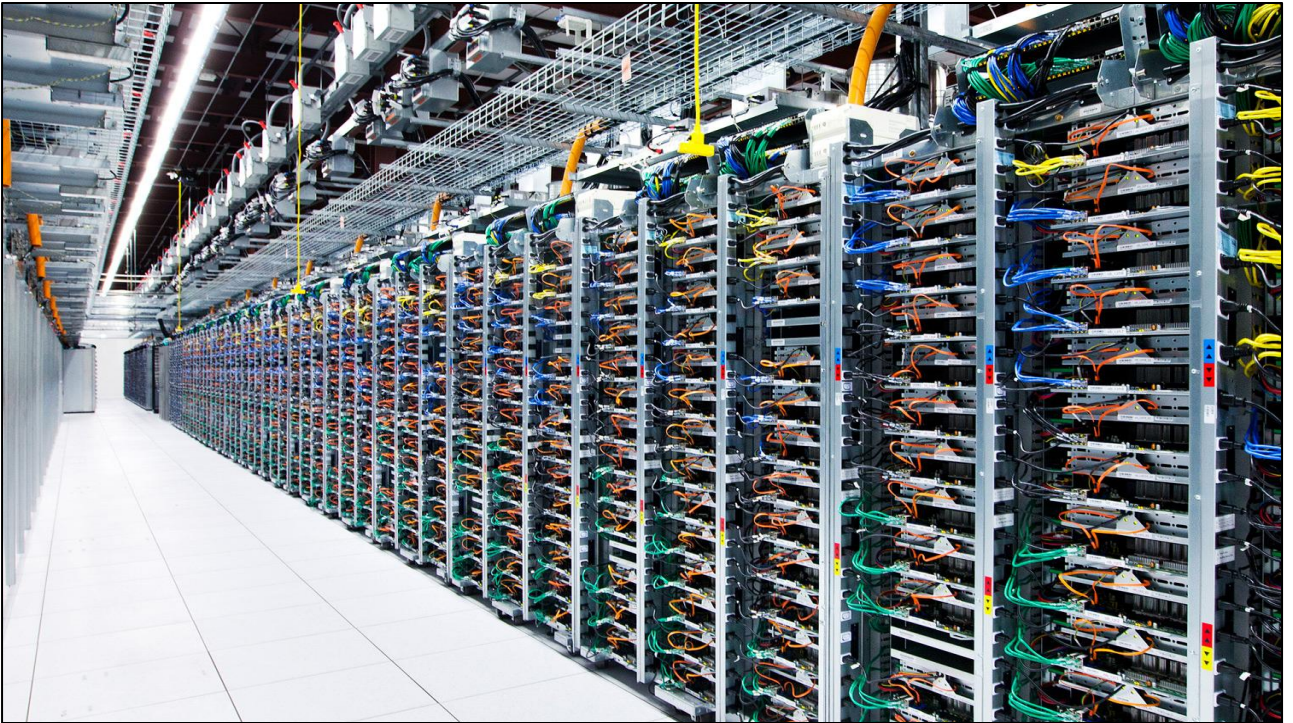
Big data with Google: Focus on insights, not infrastructure

Time to understanding



Google Cloud

And what the benefit that you get, of course with Google Cloud is you don't have to focus on an infrastructure. If you want to write amazing queries as you saw, you just write and learn some basic SQL. By the end of this course series you'll be proficient in SQL if you're not already familiar with it. And then you can just type that query into the query editor and click Run. That infrastructure already exists, right?

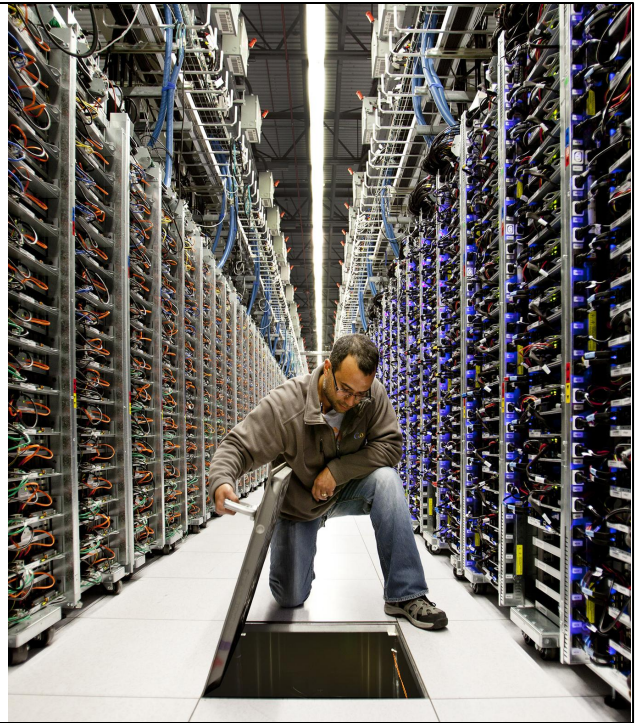


Speaking of infrastructure, this is what you're running it against. So this is not to kind of make this like a scary image. You don't have to maintain these servers, if you had to guess, right? What would this be? And I hope that the ethernet cables here are actually color coordinated with Google's logo. That would be pretty awesome. But if you haven't guess this, this is one of Google's data centers. So Google itself had to deal with processing sheer massive amounts of data. Google's mission is to democratize and get fundamental access to the information of the entire world, right, to organize the information in the entire world. And storing all the world's information is, my guess, needs a ton of those resources, as we mentioned earlier, right, the hard drives, massive data centers, and everything that goes into that as well. So it's a problem that Google had to solve.

"[Google's] ability to build, organize, and operate a huge network of servers and fiber-optic cables with an efficiency and speed that rocks physics on its heels.

This is what makes Google Google: its physical network, its thousands of fiber miles, and those many thousands of servers that, in aggregate, add up to the **mother of all clouds.**"

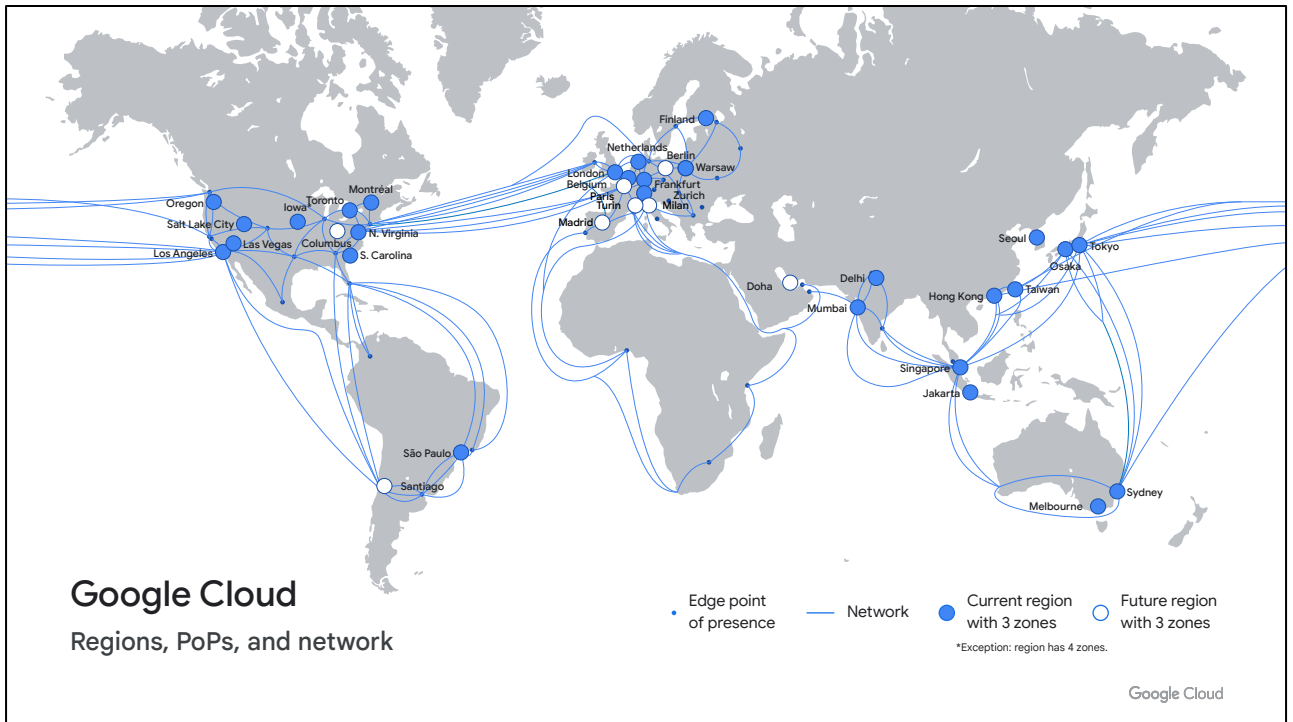
- Wired



And that's exactly what makes Google, Google. So this is the networking capacity, the fiber optic cables, all the servers and all the army of engineers that make all of this run.

And then thankfully, for all of us data analysts, Google has then opened that up through Google Cloud as a service model. So if you wanted to leverage the power of Google's infrastructure, you can do so through the cloud.

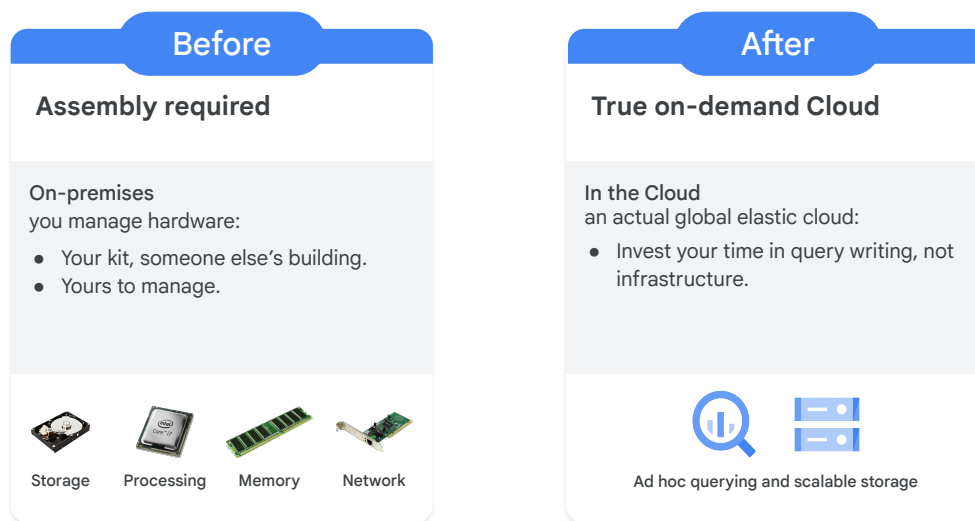
And that's the main benefit. So the cloud will run on Google infrastructure.



Speaking of Google infrastructure, here's a very exciting map that I always like to show. So when I first joined Google and I saw this, I was like wait a minute, you can't show that. That's the data centers where Google has all of its operations, and that sort of thing, as well. And thankfully they have good security there, as well. But yeah, this is all public knowledge.

So Google, in order to deliver these search results around the world and these funny cat videos for YouTube, necessitated a massive planet-scale network and operations to deliver this kind of data as well. Google Cloud is built on that.

So if you wanted to do big data analysis and your analyst's in London, and you had your clients in Singapore and your data needed to be everywhere, in effect, then you could leverage Google Cloud and the scalable storage in your data centers and the network that it has, to make that all happen at scale.



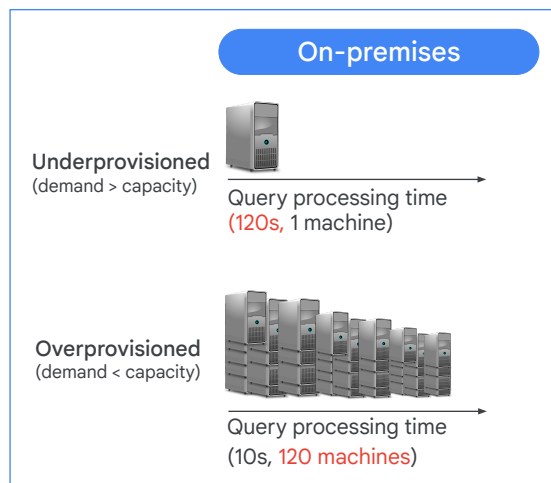
So here's kind of a general summary if we didn't scare you enough.

You have to assemble your own hardware if you're doing it the old assembly way, right. You have to get your storage and your processing power, and how much you actually going to store in memory versus persistent disk. And build up that great networking of yourself.

Or you could use the mother of all data centers and Google Cloud to make that happen for you. And really, the main point here is to get your entire organization to have almost no switching costs to get on to something like Google Cloud. And just really excite your analyst team and evangelize your folks who are interested and curious about data, but didn't want to install a bunch of things on their machine, or worry about learning so many new different technologies, when they can just focus on writing simple queries and executing those inside of BigQuery.

And that's the formation of this course as well, is the last step is a little bit of a knowledge in SQL and getting familiar with what some of these big data tools are. And once you have that familiarity, you can take that back to your organizations and basically say, hey, it's not so bad. Here are the tools. Here are some sample queries. Let's give it a whirl.

Google Cloud enables on-demand scalability



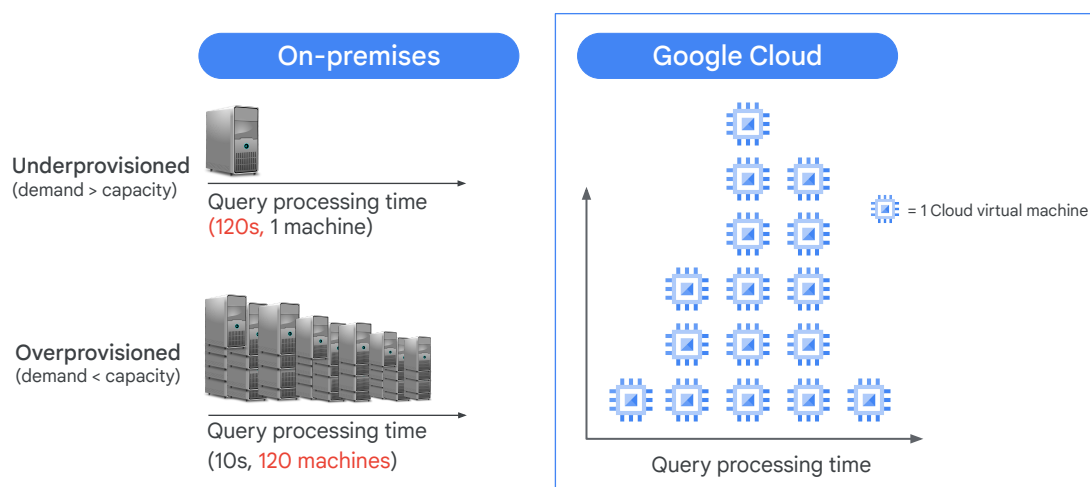
Google Cloud

One of the key words when it comes to talking about Google Cloud of course, is that scalability. So here again, if you did it yourself and you wanted to buy the servers and manage the infrastructure you can run into two potential issues here on the left. So on-premises you can have over capacity or under capacity.

So the first one is you don't have enough machines to process that amazing query. So you're running that New York Taxi cab 14 gigabyte query and you have a small machine and you're going to be sacrificing time. So it didn't cost you as much, because you only have one machine. But you're going to sacrifice time. So it's going to take much longer to process that query.

Or say you're one of the opposite, say you blew your entire IT budget for the year, and you bought the best of all the machines. And you're paying for the electricity cost to run those machines and the updates of the software, and the initial capital expenditure outlay to get that. And your query is mega fast, right. So your query processing time is really fast, but you're still paying for the massive amount of computing power. And the interesting point to note here is if you have a server and you have your storage, your hard disk platters, your persistent disk and your CPU all in the same box, right, that's a server, you're paying for electricity to the CPU even if you're not using any queries, right. So you're still burning electricity just to keep that data stored on persistent disk, to keep that data alive even when you're not using that for querying.

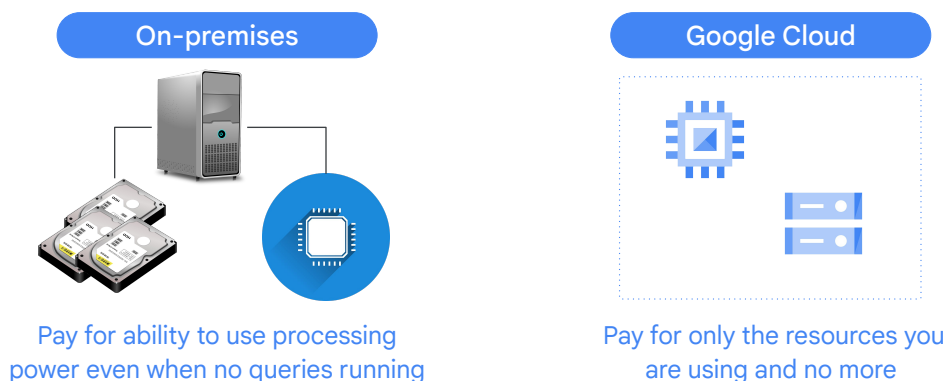
Google Cloud enables on-demand scalability



Google Cloud

Contrast that with Google Cloud on the right and the resources are there when you need them. And they go away when you don't. So you process a query. Google realizes, wow, this is a mega query that you're processing, which is awesome, right. So you've written some really complex analytical query. Google will automatically ramp up these cloud virtual machines behind the scenes. And then when your query is done, it'll ramp them down. And you only pay, in BigQuery's case, for the amount of data you're processing. Fully managed behind the scenes, you're not setting up virtual machines and processing these data yourself. All this is handled behind the scenes. And you just need to write the little bit of SQL and run that in the interface.

Separation of storage and computing power enables efficient resource allocation



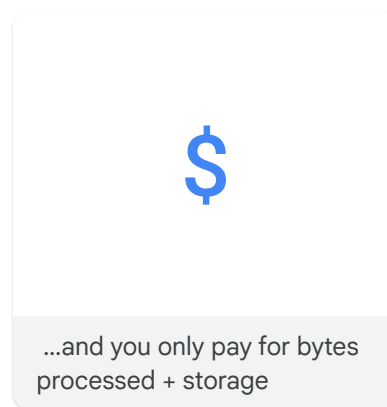
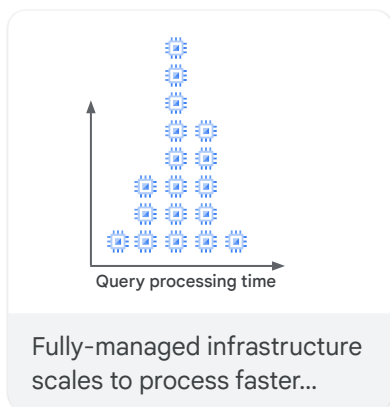
Google Cloud

So I mentioned kind of in passing, one of those key enabling features of Google Cloud. And that is the decoupling of storage and compute power.

So on-premises both those are co-located right on that server. Within Google Cloud, if you want to store data but you don't want to process it yet, say you want to store in stage 10 petabytes of data. On-premises you'd be dumping that into many, many, many, many, many, many hard drives and paying for the electricity cost to store all that.

Google Cloud, you're just paying for what you use. If you want to store it, great, awesome, it's there. Whenever you want to run compute power on it, then you're charged again for the amount of bytes that you processed. But storage and computing power are two separate concepts, which enables you to optimize for both. More efficient resource allocation is the key take away.

Key takeaway: BigQuery scales automatically and you only pay for what you use



The first key takeaway is that BigQuery in particular, will scale automatically to meet the demands of your query, and you're just going to pay for the amount of querying data that you process, and if you're storing data on BigQuery, the amount of storage on persistent disk that you're going to actually store behind the scenes.

But again, the key benefits here is that it's fully managed. You don't have to worry about the replication of the data behind the scenes. You're not paying for the electricity to keep the data centers running, or for the new software updates for the latest and greatest version of database software that comes out. All those updates are coming to you automatically from the BigQuery team.

Introduction to Data on Google Cloud

01 Analytics challenges faced by data analysts

02 Big data on-premises versus in the cloud

03 Real-world use cases

04 Google Cloud project basics



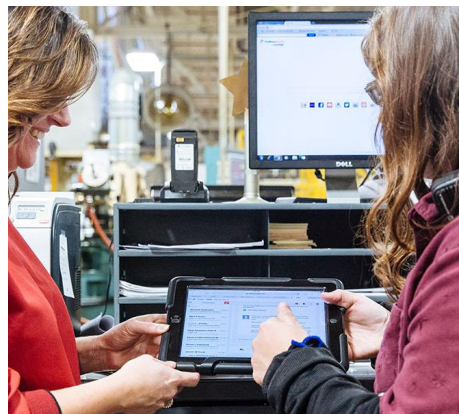
Okay let's run through two quick use cases and then I really want to get you started within Google Cloud itself. So I'm going to do a quick tour of that. I'm going to set you up with your sandbox accounts and then really get you into BigQuery and other tools to get you started. So let's run through these examples pretty quickly.

Store petabytes of data



[Our mission is] to make our data so intelligent it has the answer before the question is even asked. It was a stretch goal but essentially one that means we have to capture all the data we produce - both now and in the future.

Dan Nelson
Head of Data - Ocado



Google Cloud

First was a cool UK company that I found that was using Google Cloud called Ocado. If you're not familiar with them, they are a very large online grocer inside the UK. And they basically turned to Google Cloud, BigQuery in particular, when they said, wow, we're processing sheer massive amounts of data to handle all these online grocery transactions in addition to the actual transactions.

We need to pipe all that transactional data and all of the weather data and all of our distribution data into some central analytics reporting warehouse. Where we can just store it all there and then operate very, very quickly in an ad hoc analysis basis and query that data very quickly. And not have to worry about building this massive infrastructure ourselves. And that's exactly what they turned to BigQuery for. Is storing that data and then write these kind of analytical queries and get those insights back really, really quickly.

Focus on your business, not hardware



The less time that we can spend solving problems that are already solved, like scaling,... the more time and energy we can spend on turning our data into value.

Nicholas Harteau
VP Infrastructure - Spotify



Google Cloud

Another example is Spotify, which is a music streaming service and a whole music experience service. And they turned to Google Cloud and BigQuery in particular to basically, the same argument that we hear time and time again, you can spend your time as a data analysis, as a data scientist, on really thinking about those cool insights and those queries that you want to write, as opposed to managing for hardware failure or redundancy or backups or anything like that.

If you're a data engineer and that's your core job, more power to you. That's super fun and those jobs are definitely necessary, and Google itself has droves of engineers that make the platform, the service of BigQuery, happen. But if your goal is to glean insights for your business, perhaps better spent to write those awesome queries, and spend your time on thinking of more ingenious ways to tease insights out of your data than worrying about that infrastructure.

Introduction to Data on Google Cloud

01 Analytics challenges faced by data analysts

02 Big data on-premises versus in the cloud

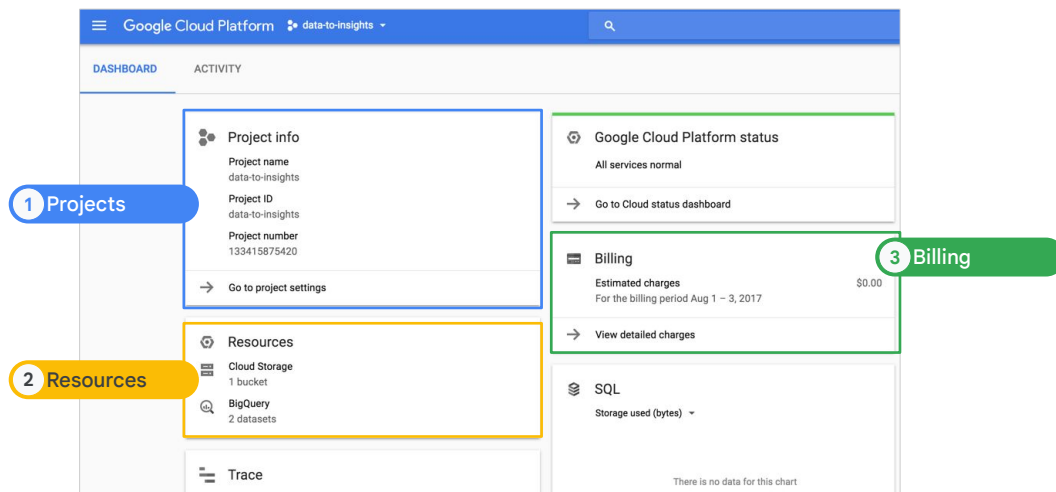
03 Real-world use cases

04 [Google Cloud project basics](#)



Okay. As promised, I want to keep this zippy. So, we're going to look at Google Cloud and a few screenshots here, and we're going to set you up with a sandbox account as part of your lab zero.

Navigate the Google Cloud using the dashboard



Here's the dashboard, as you're going to see in just a second. Three things that I want to call your attention to.

One, everything in Google Cloud, the umbrella resource that you're going to be using is called the **project**.

So everything is at the project level. You could have many different users for your projects and you can have many different **resources** that you're using as part of your project as well. So any type of access or BigQuery dataset you create is all under this project name.

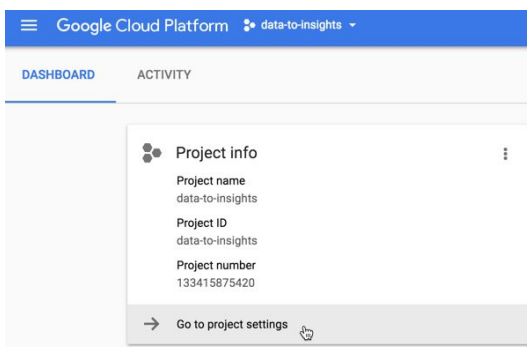
So, we have projects, you have resources that you're going to consume and then the **billing** where you'll be charged for these resources. So let's dive into each of those really quickly.

Projects

Projects organize and govern your activities in the cloud

1 Projects

- Navigate and launch cloud tools for your project by exploring the Products and Services menu.
- **Work collaboratively** by adding project users through IAM (Identity and Access Management).
- **Authorize Tools and Apps** through the API manager.



Google Cloud

So, projects organize all of your activities. So again, since this course is really focused on big data, this is largely going to be your BigQuery, your Dataprep, your Cloud Storage buckets, but this doesn't limit yourself there. If you want to go crazy and take this course and every other course that our Google Cloud team has created, you'll be using a lot more than just those few technologies. You eventually, be building up TensorFlow machine learning models and dealing with API authorization and dealing with apps, and that's all in the same dashboard interface. So once you learn this once, then you're good and you just keep plugging in more cool tools and technologies into it. And of course it's collaborative.

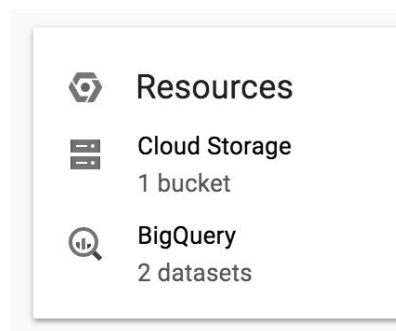
Resources

Resources are what you are using in the cloud

2 Resources

Commonly used by data analysts:

- **Storage in Cloud Storage**
 - Example: You use a bucket for uploading large CSV files to ingest later for analysis.
- **Datasets in BigQuery**
 - Example: You perform analysis on raw data and create a brand new dataset.



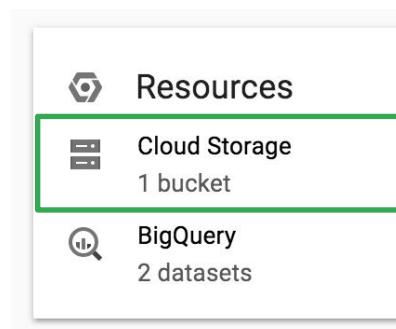
So, here's the tools in your toolkit. Two large ones that you can use as part of the specialization, you can be using Cloud Storage buckets and BigQuery datasets, as for data analysts.

Resources

The Cloud Storage bucket is your go-to for scalable storage

2 Resources

- Buckets are scalable containers that hold your data.
- You can create and upload files to your buckets within the Google Cloud console.



So Cloud Storage buckets, that's expandable container. It's a staging area - they can throw a ton of your data, CSV files, JSON files, whatever you want - get all that good stuff stored in Cloud Storage, and then you can ingest that into BigQuery.

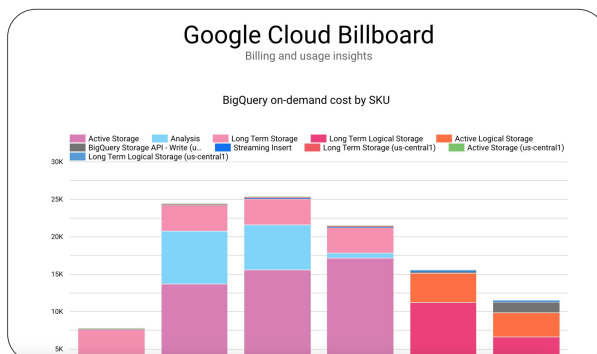
Billing

You are billed for the resources you use

3 Billing

Commonly used by data analysts:

- **Storage in Cloud Storage**
 - Billed for bucket storage
- **Datasets in BigQuery**
 - Billed for query processing
 - Billed for table storage



After this course, try exporting BigQuery logs using this [tutorial](#) to create a Looker Studio billing dashboard

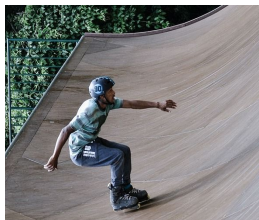
Ingest that into BigQuery in the form of datasets. And one of the great things about BigQuery is data loves data. So, you can get really meta with BigQuery and you can see how much your organization is actually using BigQuery and how many folks are running successful queries, failed queries, how much data do you actually scanning and processing, and you can visualize all that in this particular case is a Google Data Studio dashboard, and we'll cover how to create and visualize your insights in the second course of the course series in that dashboard.

But again, you're billed for those resources that you use and you can monitor those actual resources that you're using, and we'll cover the pricing of BigQuery, and how much you are going to be charged for processing those bytes of data in later modules as well, and how you can cost optimize and potentially set up those custom quotas if you worry about other users in your organization blowing your budget.

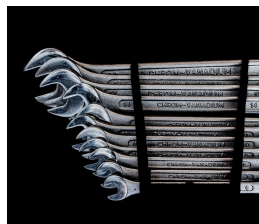
Module summary: Scale with Google Cloud



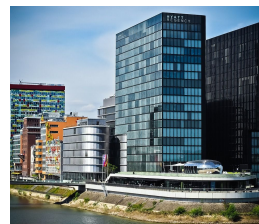
Overcome query speed, infrastructure, and cost challenges.



Efficiently scale your compute and storage needs.



Manage and monitor your project resources in one place.



Evangelize data analysis in your organization.

Wrapping up this module, let's review some of the key points about Google Cloud. We've covered some of the common challenges data analysts face and how the Cloud offers scalable, fully-managed tools for any data analyst to use. Now, in the following course modules, we'll introduce the actual specific tools and how they build on the compute and storage scalability of the Google Cloud.