Introduction to Google Cloud Platform

Lak Lakshmanan
Tech Lead, Big Data + ML

Welcome to the first module of our Big Data Fundamentals course. It provides an introduction to Google Cloud Platform.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

In this module, we'll examine the infrastructure behind Google Cloud Platform -- or GCP -- which was originally built to power Google's own applications and is now available to you.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

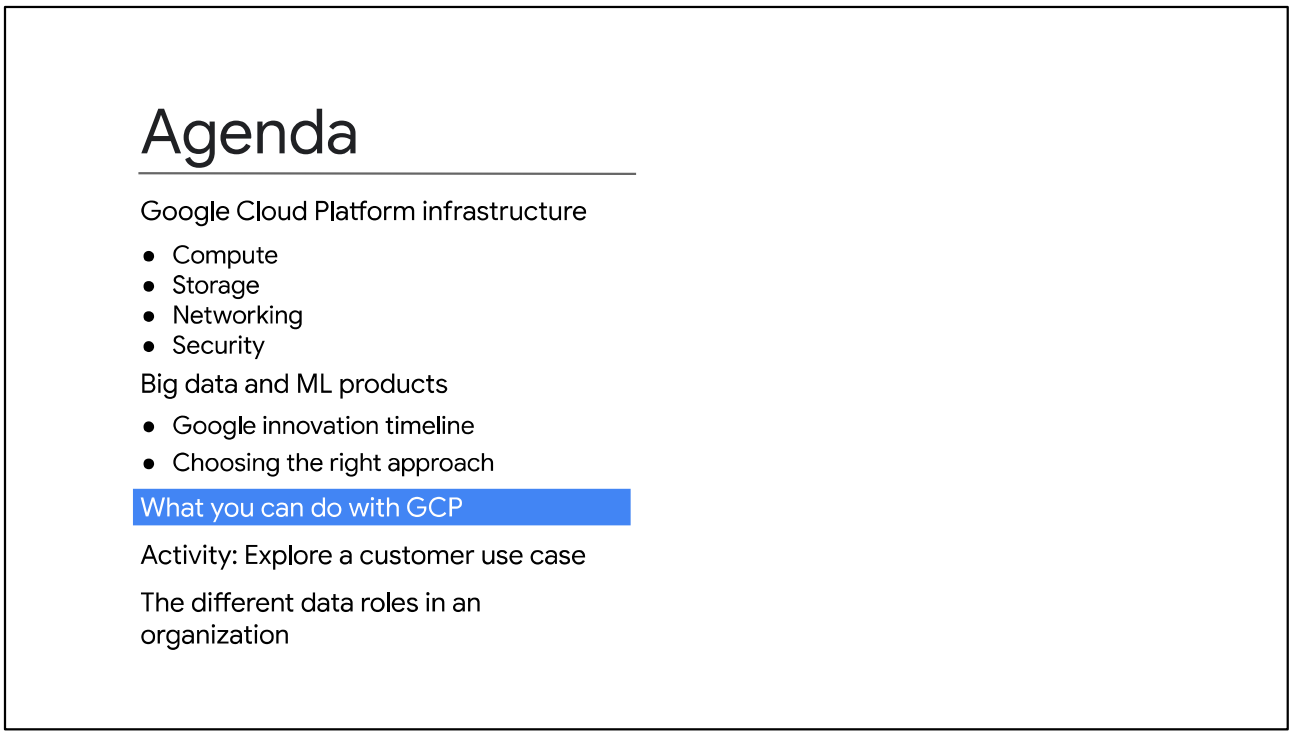
- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

Then we'll cover the big data and ML products that are built on top of that infrastructure and when you should choose which products for your solution architecture.



After that, my favorite part of this course -- learning from other customers who are using Google Cloud by exploring their use cases and getting inspired to solve those similar challenges for own teams and projects.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

You'll learn where you can look up and reference case studies of Google Cloud Platform customers by industry and/or product. Then, you will examine their solution architecture in a short activity.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

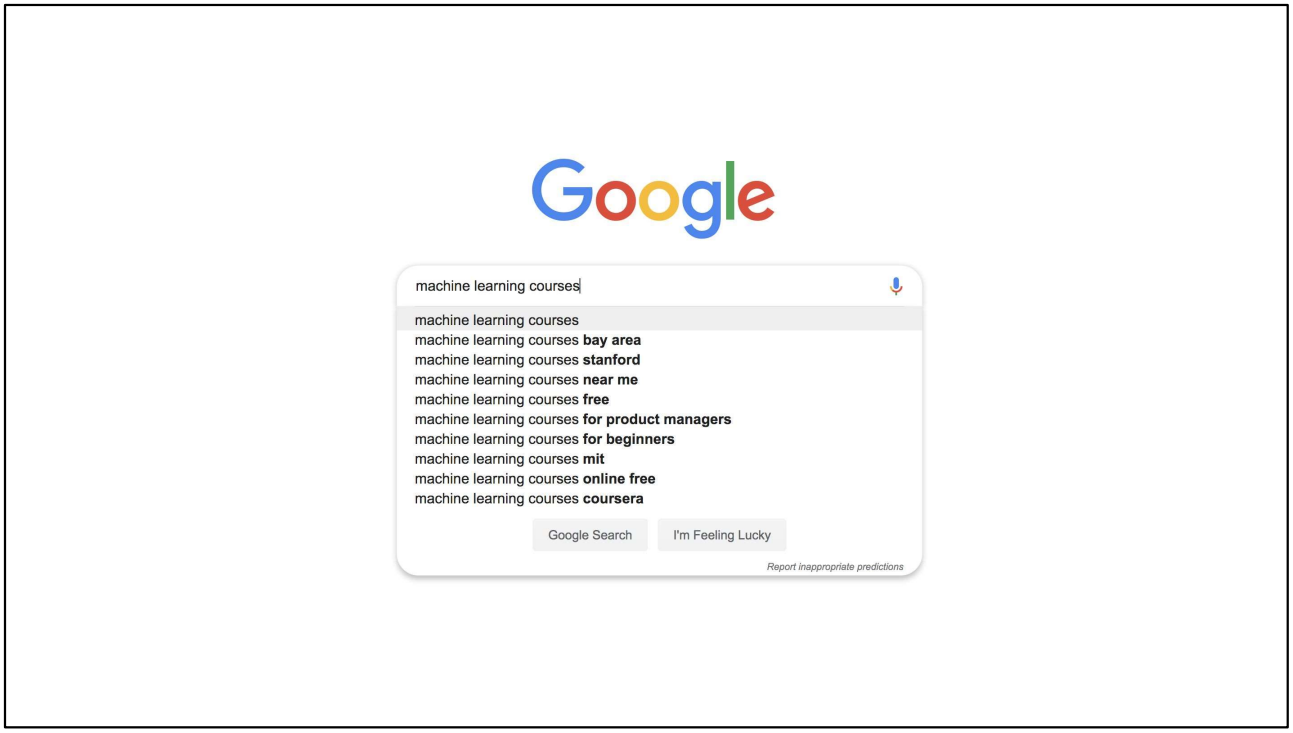
- Google innovation timeline
- Choosing the right approach

What you can do with GCP

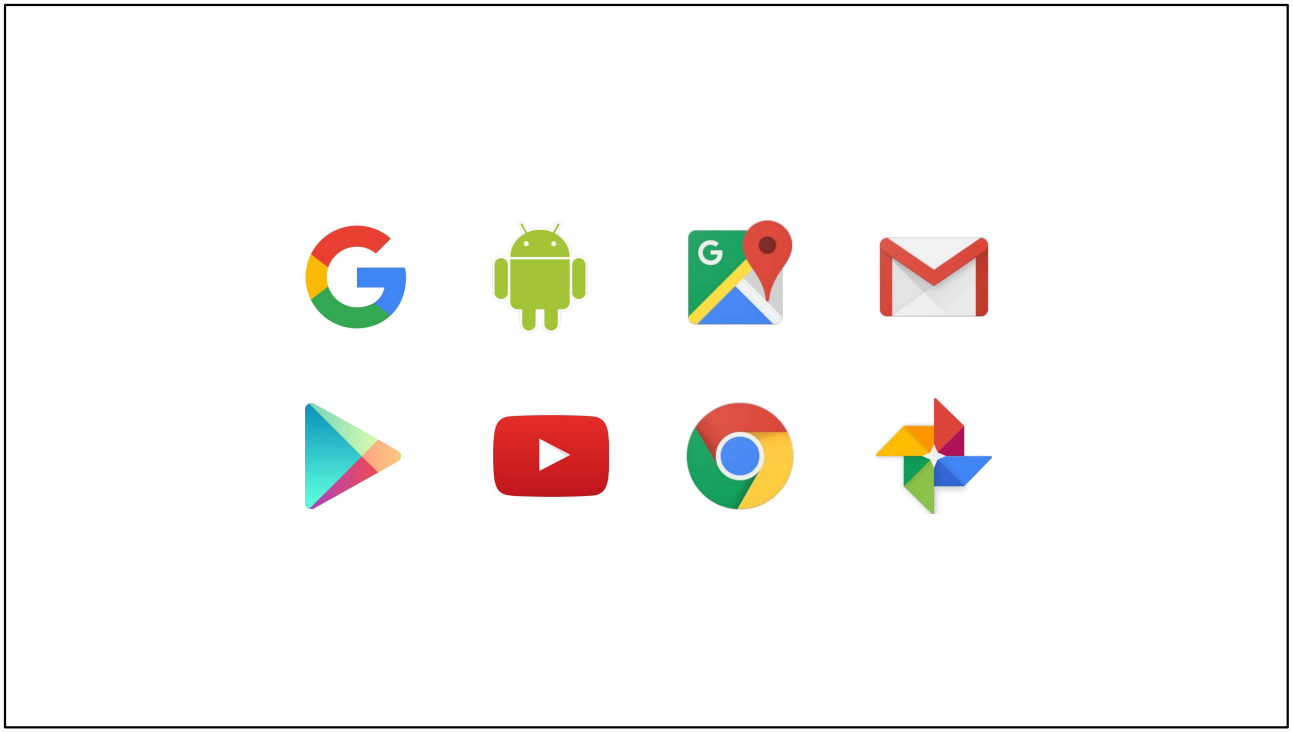
Activity: Explore a customer use case

The different data roles in an organization

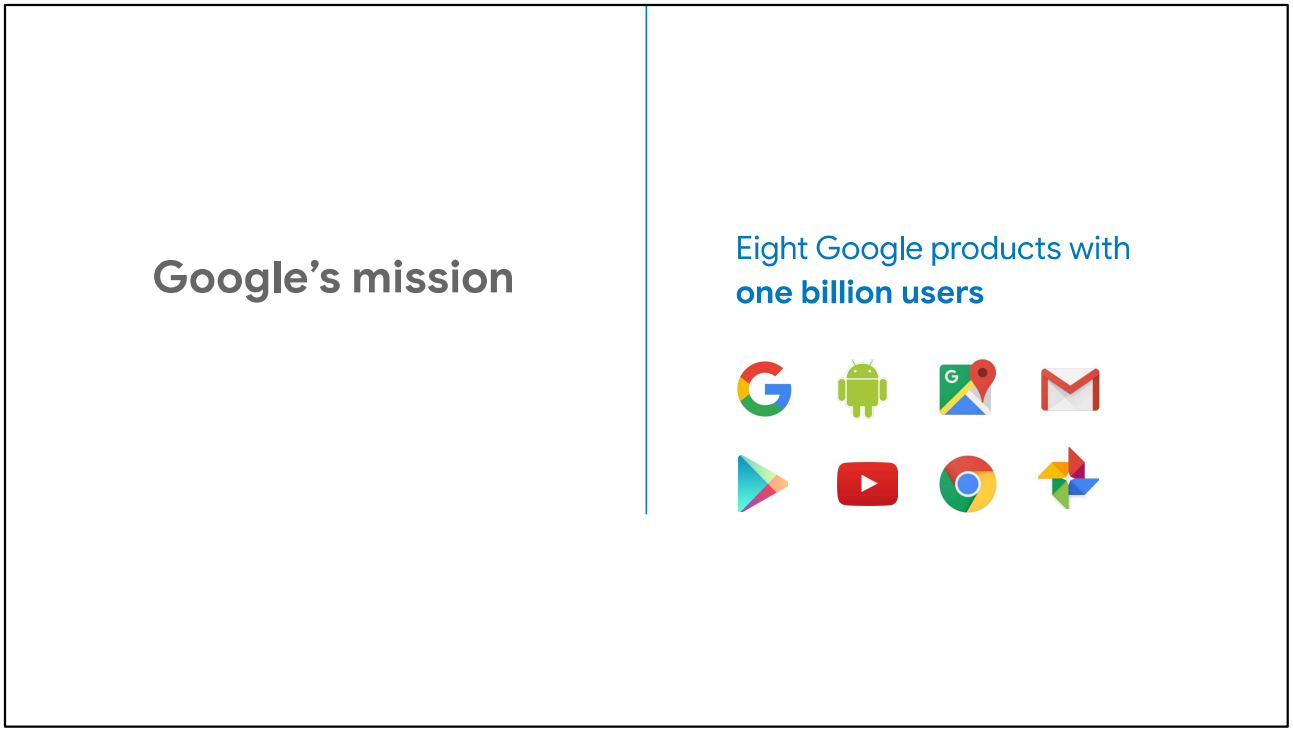
Building the right team structure is critical to solving these big data challenges. We'll explore the different types of roles and personas for building a successful big data team within your organization.



Consider for a second the impact Google Search has on our daily lives with timely and relevant responses.



Now, think of other Google products—Gmail, Maps, Chrome, YouTube, Android, Play, Drive, and Photos.



Each of these products has over 1 billion monthly users. Google had to develop the infrastructure to ingest, manage, and serve all the data from these applications, and to do so with a growing user base and data requirements that are constantly evolving.

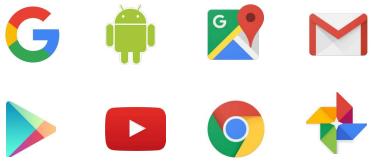
8 products with a billion users …

Actually, there is an 9th Google product that has a billion end-users -- Google Cloud, except that it's your users; the end-users served by Google Cloud customers such as Home Depot and Spotify; Twitter and New York Times; Colgate-Palmolive and Go-JEK.

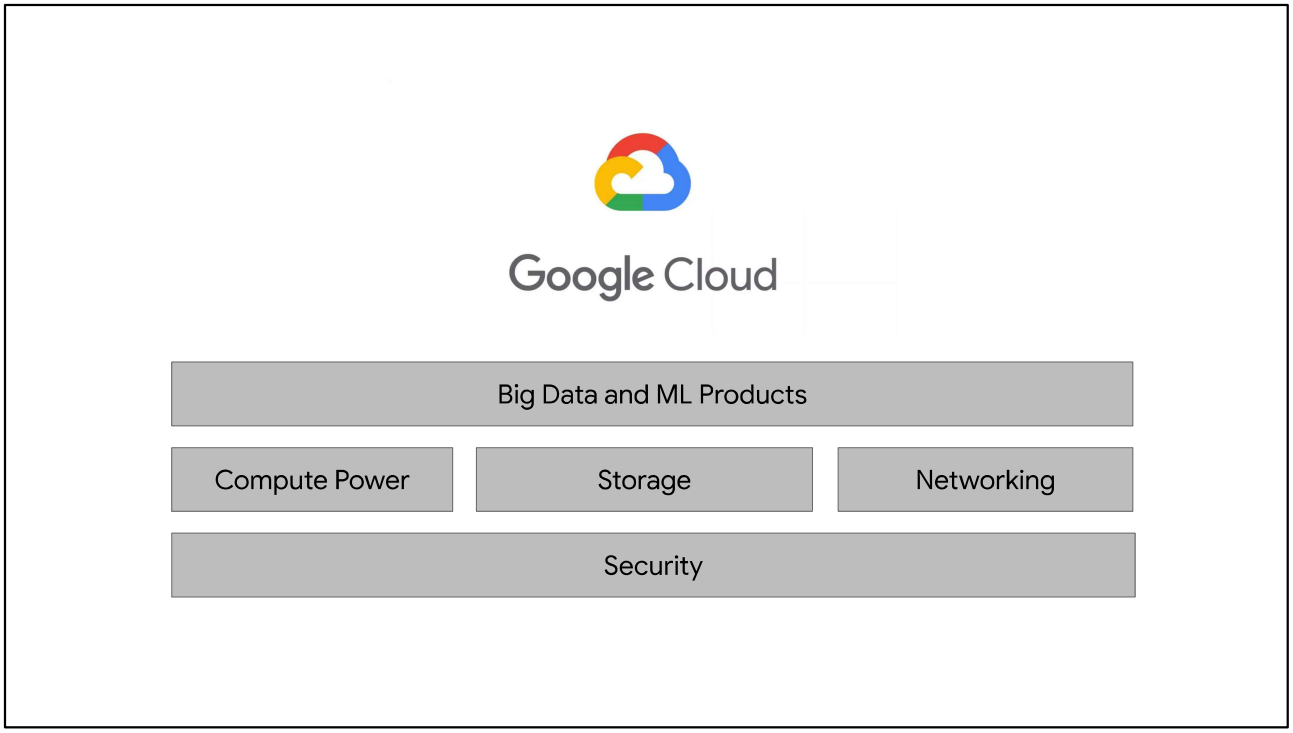
Google's mission

Organize the world's information and make it universally accessible and useful.

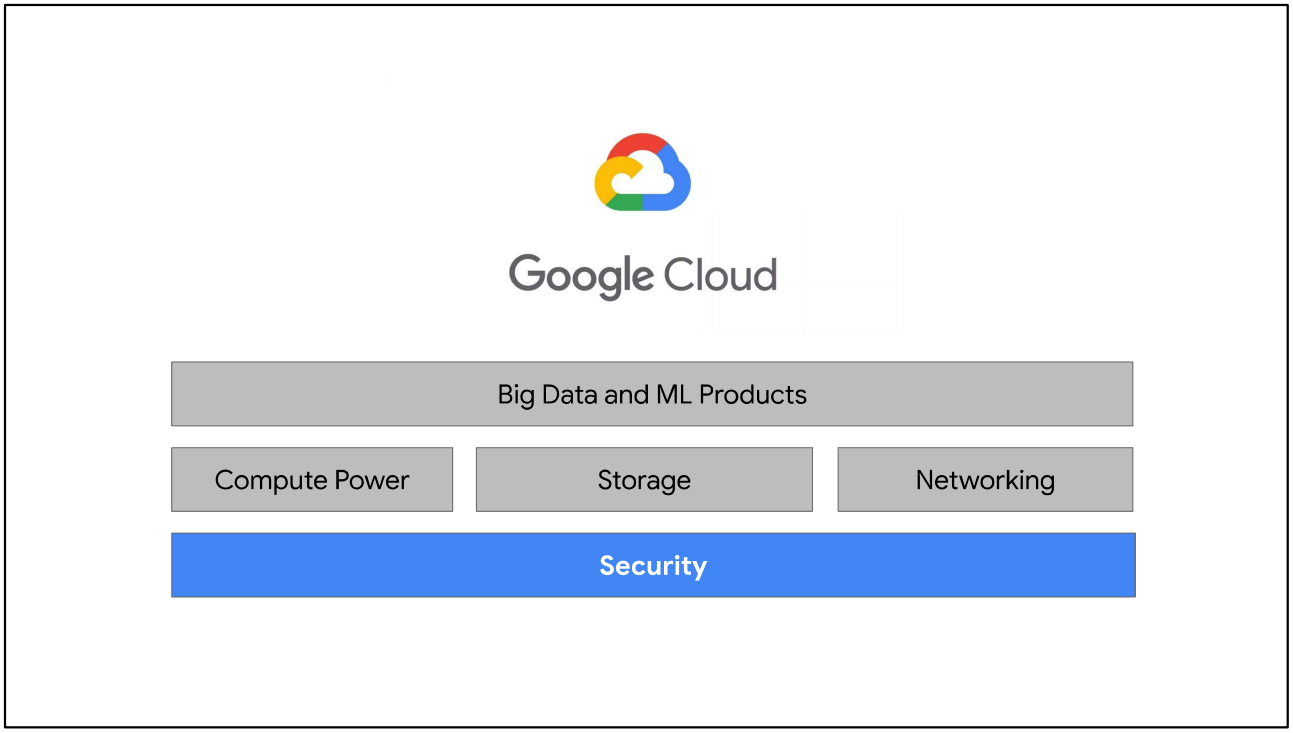
Eight Google products with **one billion users**



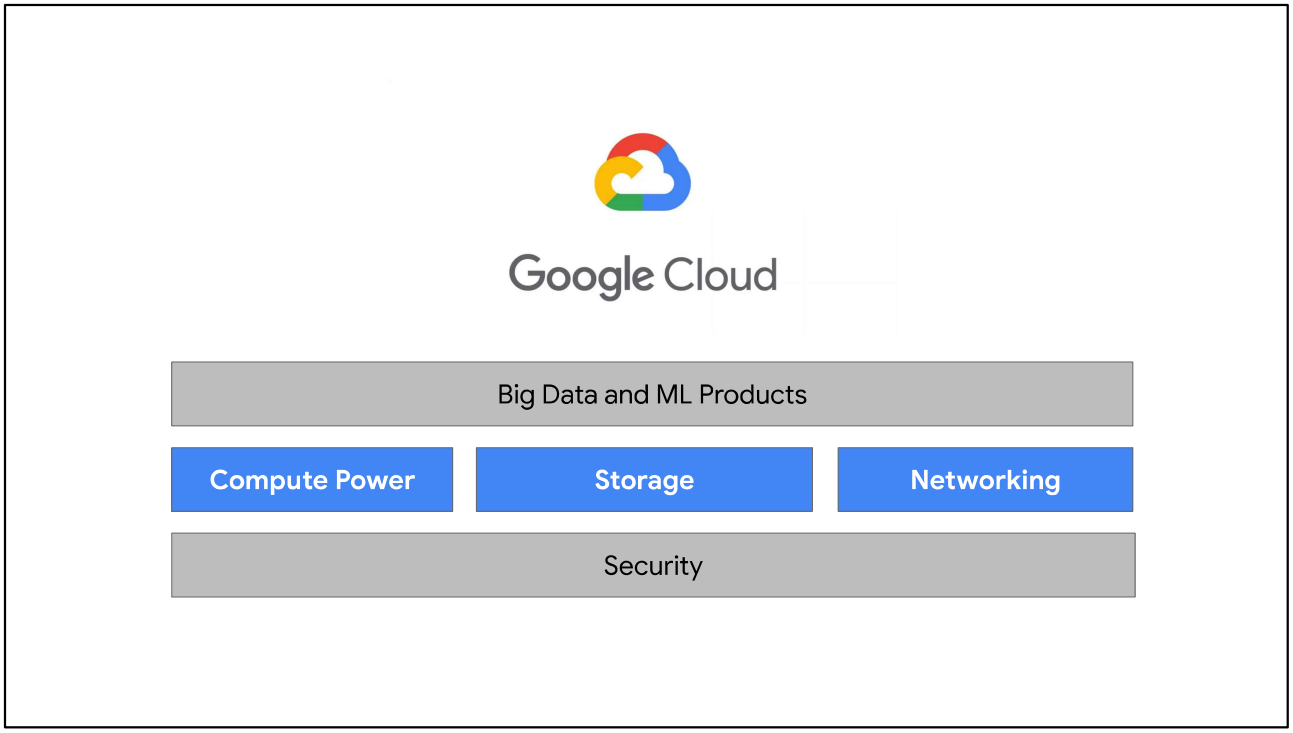
Let's look at the building blocks behind Google's big data infrastructure and how you can leverage it with Google Cloud.



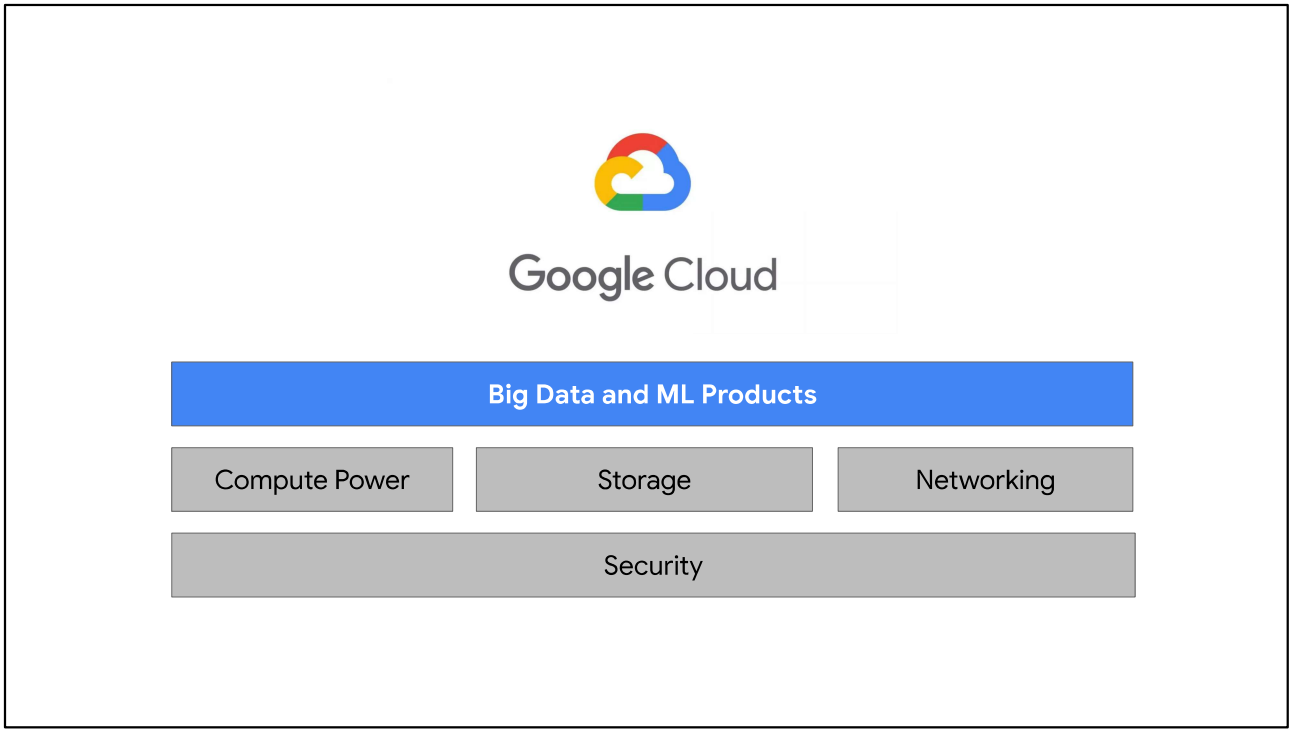
There are four fundamental aspects of Google's core infrastructure and a top layer of products and services that you'll interact with most often.



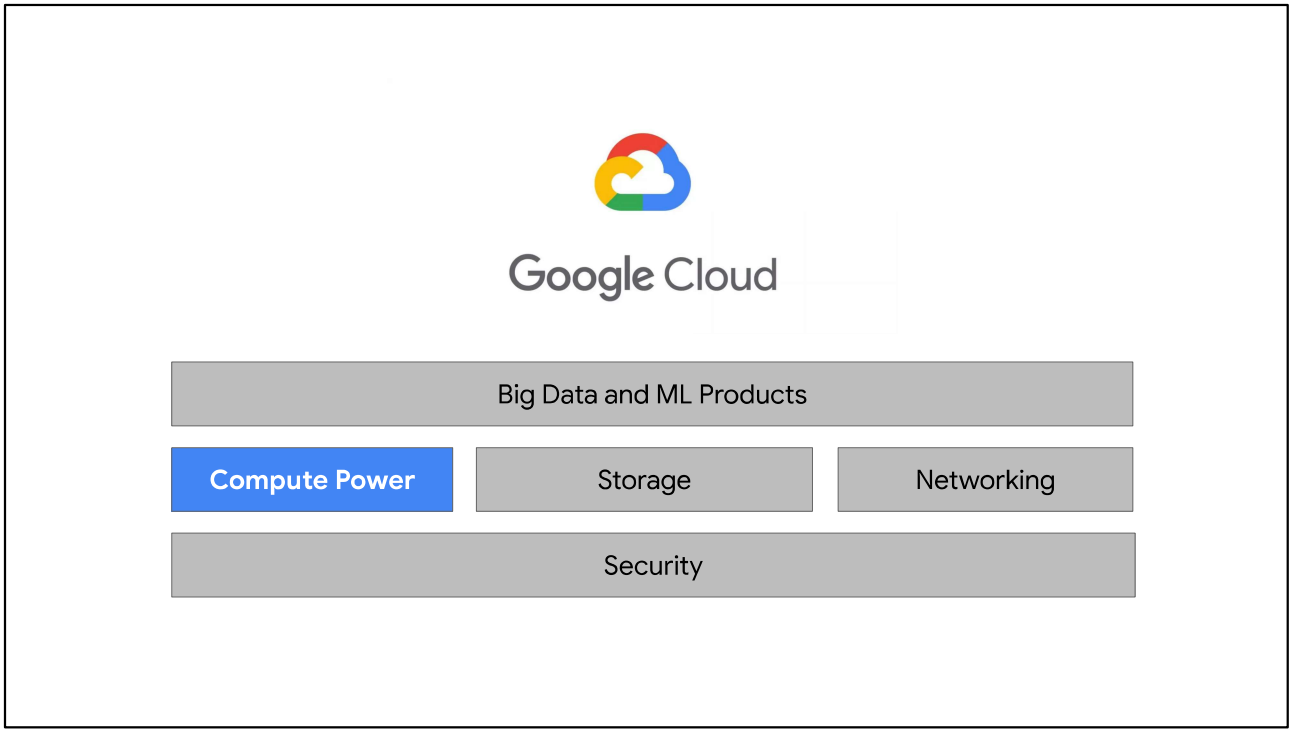
The base layer that covers all of Google's applications (and therefore Google Cloud's, too) is security.



On top of that are compute, storage, and networking. These allow you to process, store, and deliver those business-changing insights, data pipelines, and ML models.



Finally, while running your big data applications on bare metal virtual machines is possible, Google has developed the top layer of big data and ML products to abstract away a lot of the hard work of managing and scaling that infrastructure for you.



A growing data organization like yours will need lots of compute power to run big data jobs -- especially as you design for the future to outpace your growth in new users and data for the next decade.

Fused Video Stabilization on the Pixel smartphone with ML

Data sources:

- Image frames (stills from video)
- Phone gyroscope
- Lens motion

Let's start with an example illustrating how Google uses its own compute power.

Google Photos has recently been introducing smart features like this one for automatic video stabilization, for when the camera is shaky as you see here on the left. What data sources do you think are needed as inputs to the model?

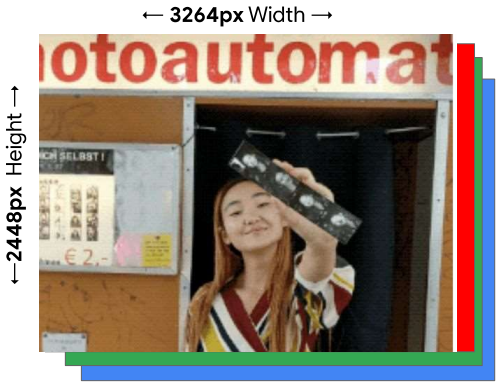
You need the video data itself, which is essentially lots of individual images called *frames* at a timestamp ordering. But we need more contextual data than just the video itself, right?

Absolutely. We need time series data on the camera's position and orientation from the onboard gyroscope and motion on the camera lens.

So … how much video data are we talking about for the Google Photos' ML model to compute and stabilize these videos?

https://ai.googleblog.com/2018/03/behind-motion-photos-technology-in.html

A single high-res image represents millions of data points to learn



3 “Layers” in depth for Red Blue Green

8 Megapixel resolution

3264 (w) x 2448 (h) x 3 (RGB) =

23,970,816
data points per image*

* ML training time affected by availability of compute resources

If you consider the total number of floating point values representing a single frame of high-res video, it’s the product of the number of channel layers multiplied by the area of each layer, which, with modern cameras, can easily be in the millions. An 8-megapixel camera creates images with 8 million pixels each. Multiply that by 3 channel layers, and you get over 23 million data points per image frame. And there are 30 frames per second of video. You can quickly see how a short video becomes over a billion data points to feed into the model.

1.2 billion photos and videos are uploaded to Google Photos every day.

Total size of over 13 PB of photo data.

Tour of Google's data center

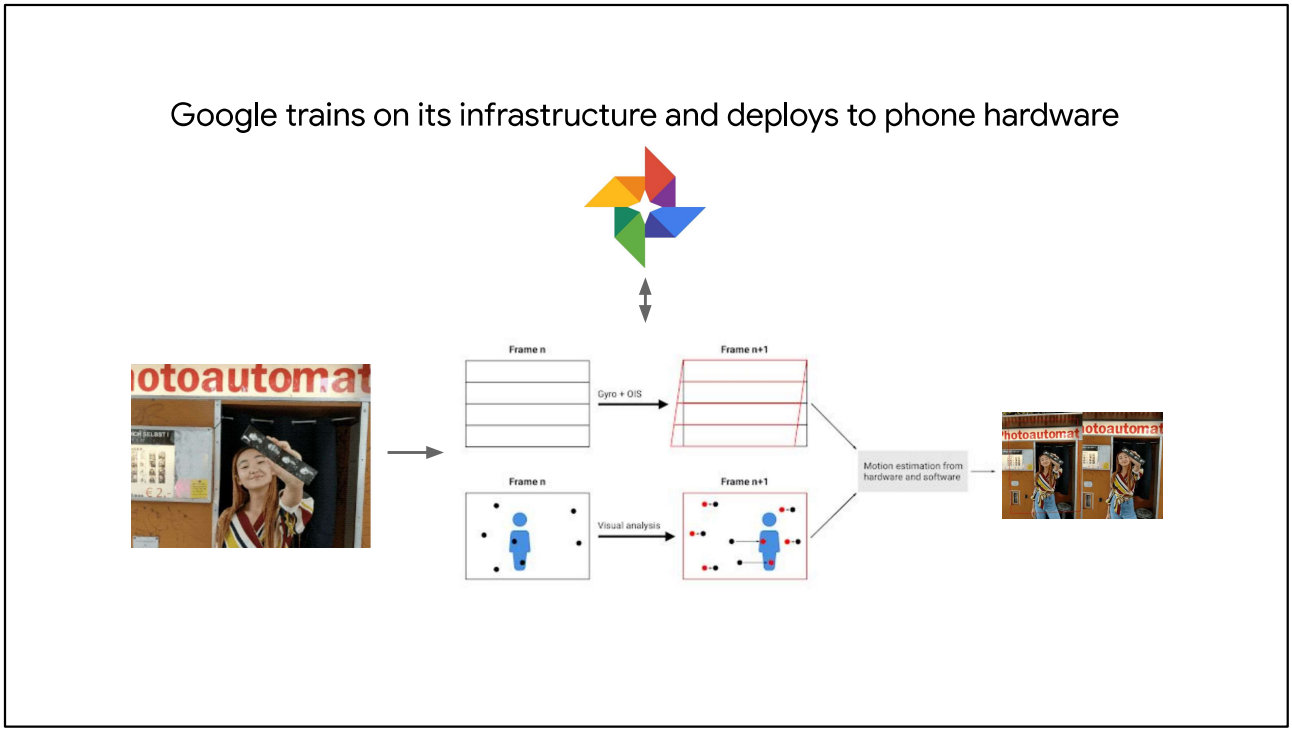
YouTube

(1PB or 400 hours of video uploaded every minute)

And from 2018 estimates, roughly 1.2 billion photos and videos are uploaded to the Google Photos service every day. This is 13+ PB of photo data in total.

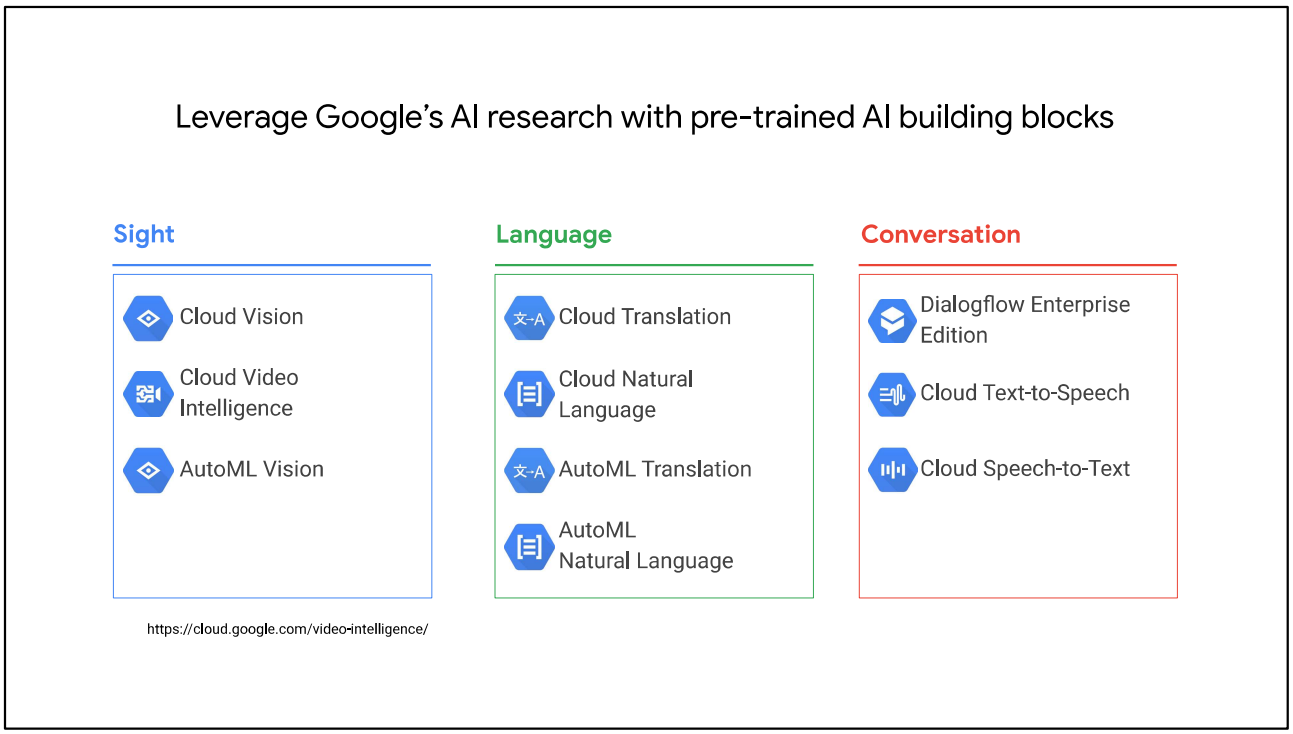
For YouTube -- which also has ML models for video stabilization and other models for automatically transcribing audio -- you're looking at over 400 hours of video uploaded every minute. That 60 PB every hour.

https://en.wikipedia.org/wiki/Google_Photos
https://www.youtube.com/watch?time_continue=10&v=x5rHog6RnNQ
https://ai.googleblog.com/2017/11/fused-video-stabilization-on-pixel-2.html



But it's not just about the size of each video in pixels -- the Google Photos team needed to develop, train, and *serve* a high-performing ML model on millions of videos to ensure the model is accurate. That's the training dataset for this simple feature.

Just as your laptop hardware may not be powerful enough to process a big data job for your organization, a phone's hardware is not powerful enough to train sophisticated ML models. Google trains its production ML models on its vast network of data centers and then deploys smaller trained versions of the models to the hardware on your phone for predictions on new video.



A common theme throughout this course is that when Google makes breakthroughs in AI research, it continues to invest in new ways to expose these as fully trained models for everyone. You can therefore **leverage Google's AI research** with pre-trained AI building blocks.

For example, if you're a company producing movie trailers and quickly want to detect labels and objects in thousands of trailers to build a movie recommendation system, you could use the **Cloud Video Intelligence** API instead of building and training your own custom model. There are other fully trained models for **language** and for **conversation**, too.

You'll learn and practice using these AI building blocks later in this course.

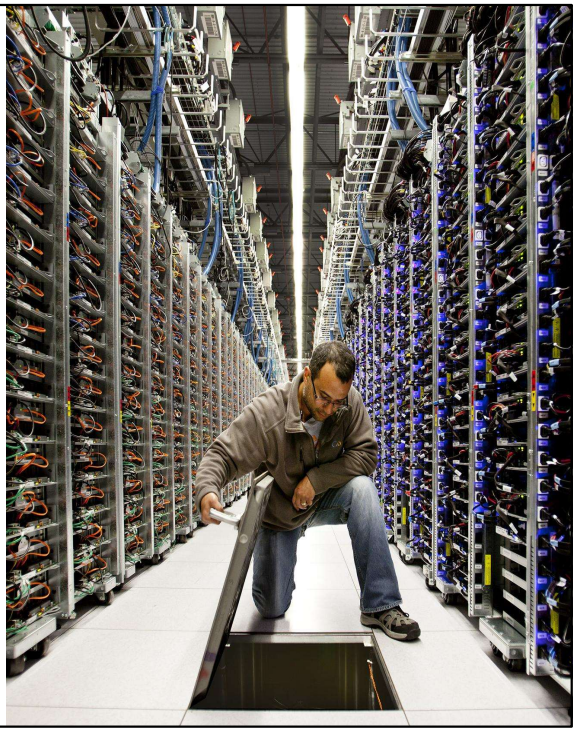
Getting back to the Google Photos ML story...

https://cloud.google.com/video-intelligence/

"[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

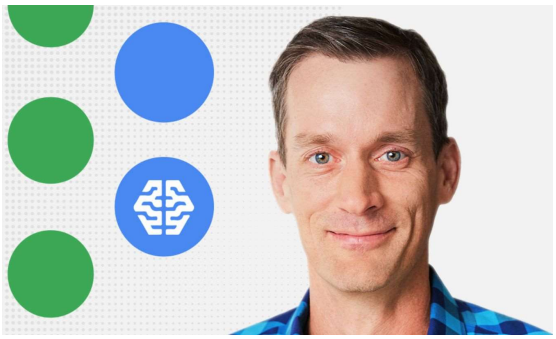


...running that many sophisiticated ML models on large structured and unstructured datasets for Google's own products required a massive investment in computing power.

That is why Wired said "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."

In essence, Google has been doing distributed computing for over 10 years for its applications and now has made that compute power available to you through Google Cloud.

WIRED Article:
https://www.wired.com/2012/10/ff-inside-google-data-center/



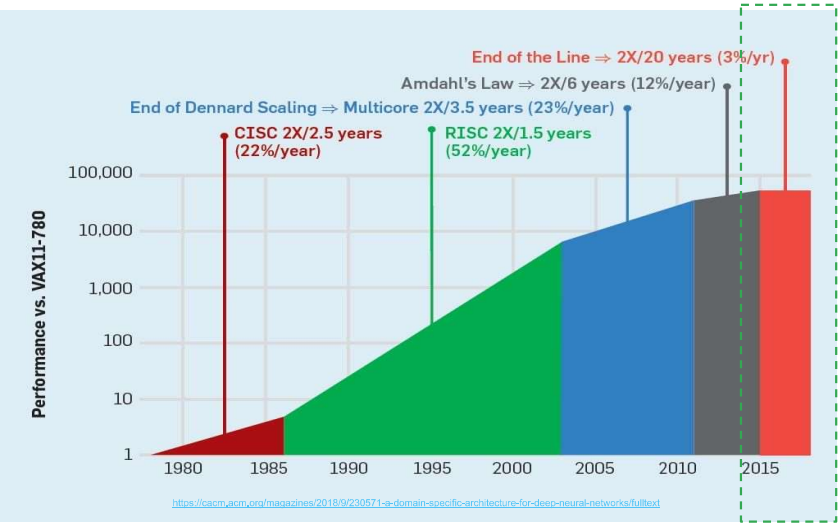
“If everyone spoke to their phone for 3 minutes, we’d exhaust all available computing resources”

— Jeff Dean, 2014

But simply scaling the raw number of servers in Google’s data centers isn’t enough.

Here's an interesting rough calculation by Jeff Dean, who leads Google's AI division. He realized years ago that if everybody wanted to use voice search on their phones, and used it for only 3 minutes, we would need to double our computing power.

Will Moore's Law save us?



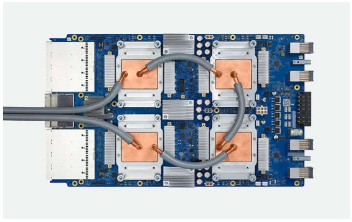
Historically, compute problems like this could be addressed through Moore's Law. Moore's Law was a trend in computing hardware that described the rate at which computing power doubled. For years, computing power was growing so rapidly that you could simply wait for it to catch up to the size of your problem.

Although computing was growing rapidly even as recently as 8 years ago, in the past few years, growth has slowed dramatically as manufacturers run up against fundamental limits. Compute performance has reached a plateau.

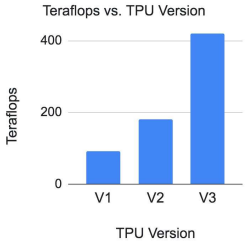
One solution is to limit the power consumption of a chip.

Tensor Processing Units (TPUs) are specialized ML hardware

Cloud TPU v2
180 teraflops
64-GB High Bandwidth Memory (HBM



Cloud TPU v3
420 teraflops
128-GB HBM



You can do that by building *application-specific chips,* or ASICs.

Google designed new types of hardware specifically for ML. The *Tensor Processing Unit*, or TPU, is an ASIC specifically optimized for ML, and it has more memory and a faster processor for ML workloads than traditional CPUs or GPUs. Google has been working on the TPU for several years and has made it available to other businesses like yours for really big and challenging ML problems.

https://cloud.google.com/tpu/

TPUs enable faster models and more iterations

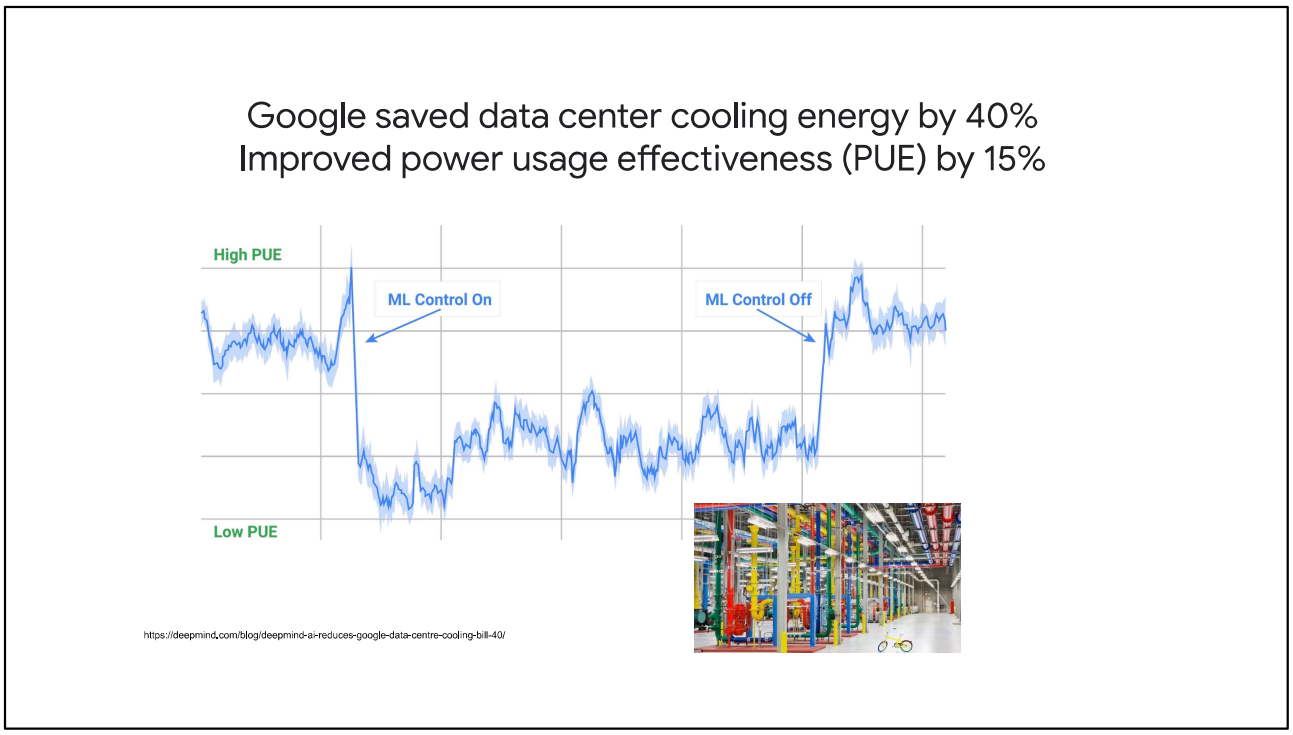
"**Cloud TPU Pods** have transformed our approach to visual shopping by delivering a **10X speedup** over our previous infrastructure. We used to spend months training a single image recognition model, whereas now we can train much more accurate models in a few days on Cloud TPU Pods.

We've also been able to take advantage of the additional memory the TPU pods have, allowing us to process many more images at a time. This **rapid turnaround time enables us to iterate faster** and deliver improved experiences for both eBay customers and sellers."

— **Larry Colagiovanni**
VP of New Product Development, eBay

One such business is eBay**.** They use Cloud TPU Pods to deliver faster inferences to their users by a factor of 10 -- **a 10x speed up.** The decrease in model training time has also led to faster model experimentation; ML model training and feature engineering is one of the most time-consuming parts of any ML project**.**

eBay's VP of new product development remarked that the additional memory of the TPU pods enabled them to improve their **turnaround time and iterate faster.**

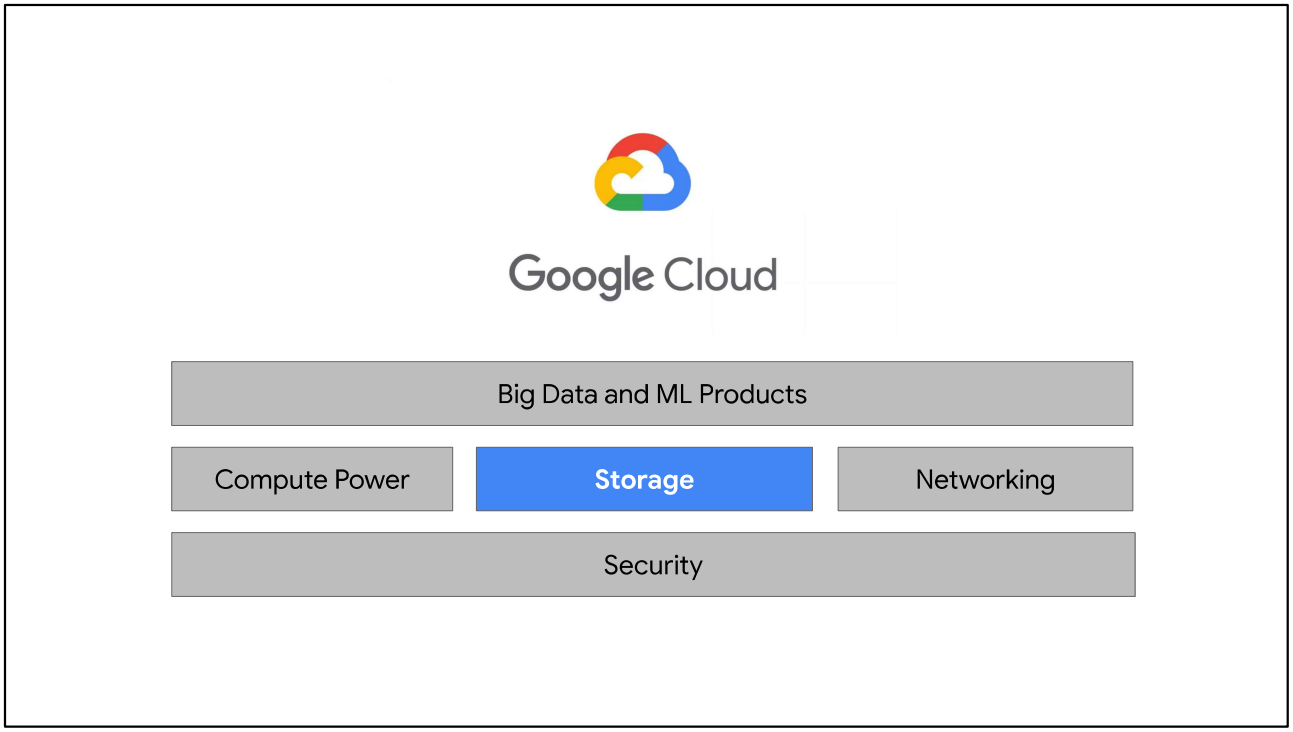


One last example of compute power at Google, and an inside look at our culture of thinking 10X, is how our teams used ML to boost Google's own data center efficiency. The potential impact for ML was there, considering the number of data centers that Google has to keep cooled and powered, and we were already collecting streaming sensor data for our existing monitoring platforms.

Engineers at Alphabet's Deepmind company saw this as an opportunity to ingest that sensor data and train a machine learning model to optimize cooling better than humans and existing systems could. The model they implemented reduced the cooling energy used by 40% and boosted the overall power effectiveness by 15%. I find this example particularly inspirational because it's a machine learning model trained on ML-specialized hardware in a data center, telling the data center how hot it can run the ML-specialized hardware that the model is training on. Powerful stuff.

Later in the course you'll see a demo on how you can setup a streaming data ingestion pipeline for your IoT devices in less than an hour.

https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42542.pdf



In the demo at the end of the previous section, I copied the ingested and transformed files out of the compute instance into Cloud Storage. This way, we could stop our compute instance and retain access to the data.

While we've discussed the demand for a great amount of computing power for today's big data and ML jobs, we still need a place to store all the data that is generated. As with my demo, this needs to be separate from the compute instances so that we can analyze it, transform it, and feed it into our models.

This is one major way that cloud computing differs from desktop computing. Compute and storage are independent. You don't want to think of disks attached to the compute instance as the limit of how much data you can process.

1.2 billion photos and videos are uploaded to Google Photos every day.

Total size of over 13 PB of photo data.

Tour of Google's data center

YouTube

(1PB or 400 hours of video uploaded every minute)

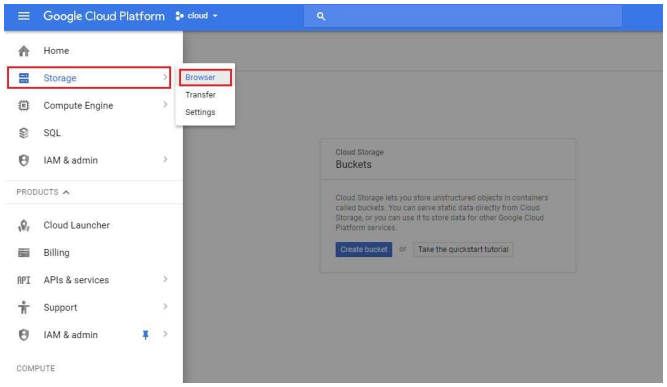
Recall that Google had to solve this distributed storage problem itself for massive datasets like Google Photos, YouTube and Gmail.

Getting your data into your solution and transforming it for your purposes should be your first priority. In the roles and team structure discussions that I will talk about later in this module, I will talk about the need for data engineers to build data pipelines before you can build ML models from that data. And when the pipeline is built, the job is not done.
Once the data is in your system, data engineers have to replicate data, back it up, scale it, and remove it as needed—all at scale.

https://en.wikipedia.org/wiki/Google_Photos
https://www.youtube.com/watch?time_continue=10&v=x5rHog6RnNQ
https://ai.googleblog.com/2017/11/fused-video-stabilization-on-pixel-2.html

Creating a Cloud Storage bucket for your data is easy

UI



Cloud Storage

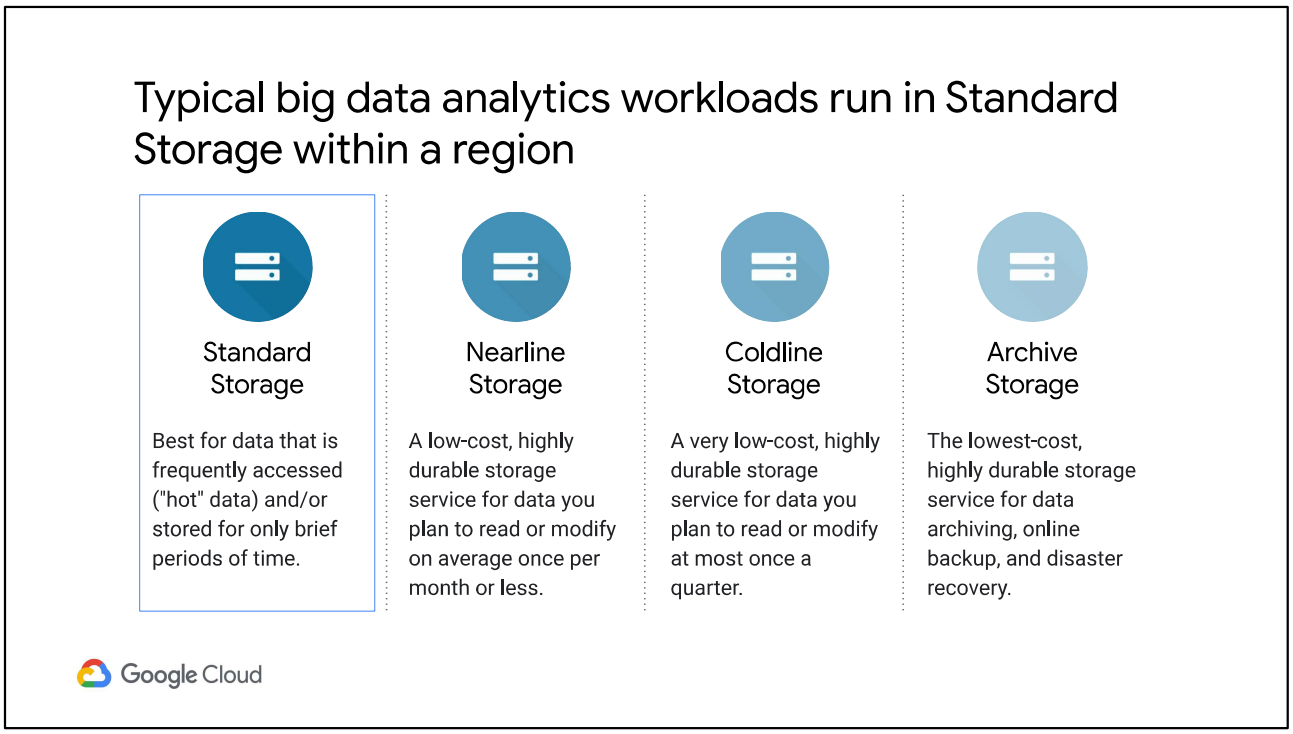
CLI

```
gsutil mb -p [PROJECT_NAME] -c [STORAGE_CLASS]
-l [BUCKET_LOCATION] gs://[BUCKET_NAME]/
```

Instead of managing the storage infrastructure yourself, you can use Google Cloud Storage, which is a durable, global, file system.

Creating an elastic storage bucket is as simple as using the web UI in console.cloud.google.com or the Google Storage Utility function (gsutil) in the command line interface. As an additional level of flexibility, you can choose what type of storage class you want for your data.

https://cloud.google.com/storage/docs/creating-buckets



There are four storage classes for you to choose from based on your data needs.

- Standard Storage
- Nearline Storage
- Coldline Storage, and
- Archive Storage

Regardless of which one you choose, all classes have multi-region, dual-region, and region location options.
They differ based on access speed and cost. Standard Storage is the fastest, and Archive Storage is the least expensive.
A good tradeoff, for example, is to use Nearline storage for data you might access only monthly.

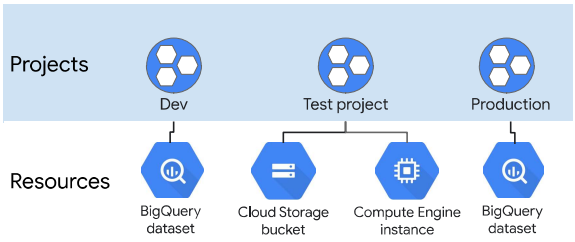
For data analysis workloads, it's common to use a Standard Storage bucket within a region for staging your data.
Why do I say "within a region"? That's because you need the data to be available to your data processing computing resources, which will often be within a single region.
Co-locating your resources maximizes the performance for data-intensive computations and can reduce network charges.

Cloud Storage buckets and Compute Engine are examples of Google Cloud *resources.* Let's cover some of the account management logistics that you need in order to use cloud resources.

Starting from the the most granular objects, you see that resources, like your Cloud Storage bucket or Compute Engine instance, belong to specific projects.

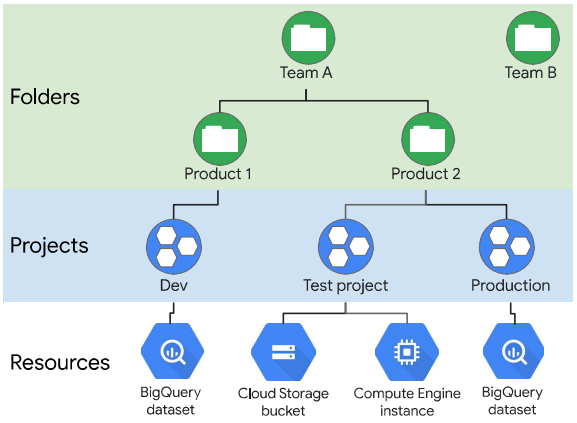
Bucket names have to be globally unique and GCP assigns you a project id that is globally unique too.

Google Cloud Platform
resource hierarchy



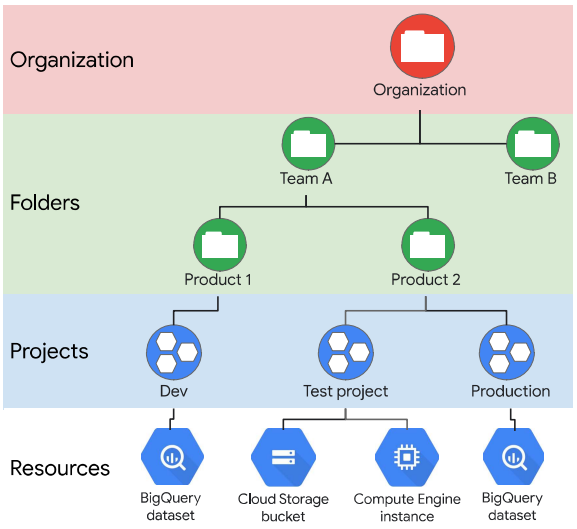
A project is the base-level organizing entity for creating and using resources and services and managing billing, APIs, and permissions.

Zones and regions *physically* organize the GCP resources you use, and projects *logically* organize them. Projects can be easily created, managed, deleted, or even recovered from accidental deletions.



Google Cloud Platform resource hierarchy

Folders are another logical grouping you can have for collections of projects. Having an organization is required to use folders, and the organization is the root node of the entire GCP hierarchy.

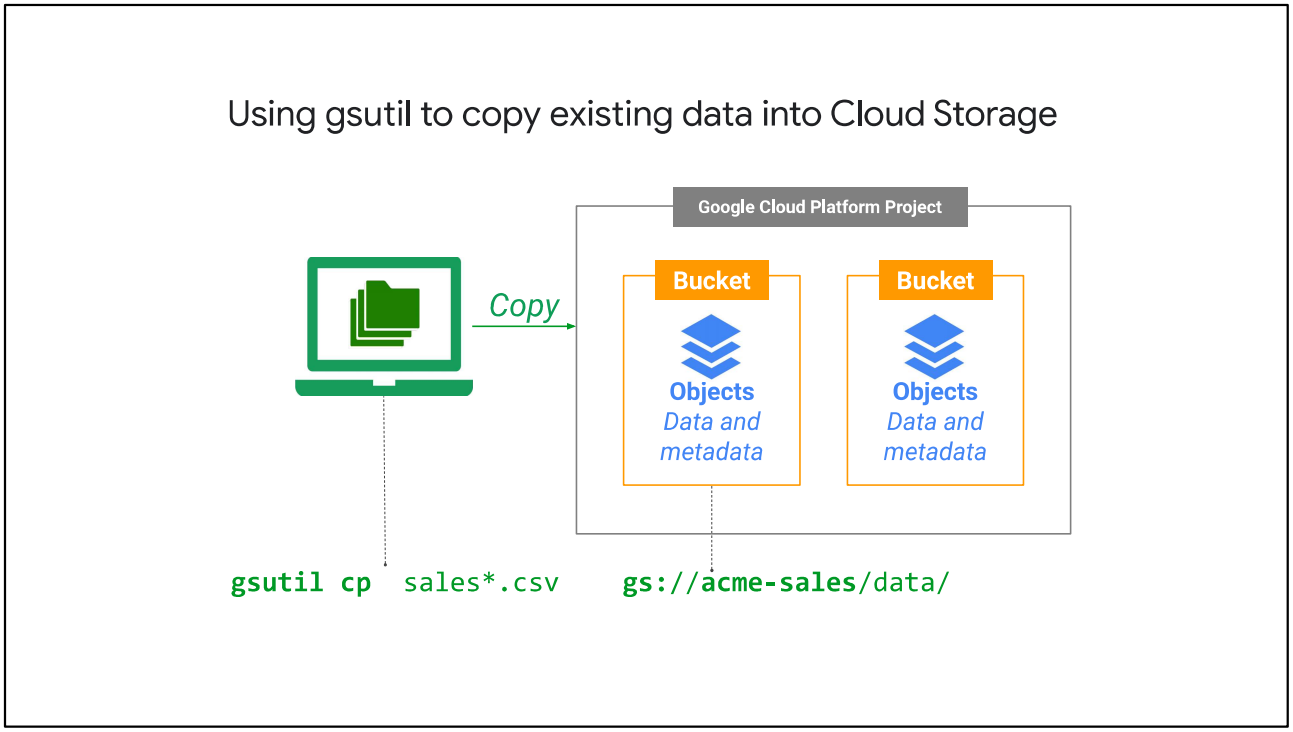


Google Cloud Platform resource hierarchy

Lastly, while not required, an organization is quite useful because it will allow you to set policies that apply throughout your enterprise.

Cloud Identity and Access Management, also called IAM or I-A-M, lets you fine-tune access control to all the GCP resources you use. You define IAM policies that control user access to resources.

Remember if you want to use folders, you must have an organization.



Now that you have a Cloud Storage bucket created, how do you get your data on the cloud and work with the data once it's there in the bucket?

In the demo, I used gsutil commands -- specifically we can use "cp" for copy and specify a target bucket location.

If you spin up a Compute Engine instance, gsutil is already available. On your laptop, you can download the Google Cloud SDK to get gstuil. gsutil uses a familiar UNIX command syntax.

https://cloud.google.com/storage/docs/creating-buckets
https://cloud.google.com/storage/docs/overview

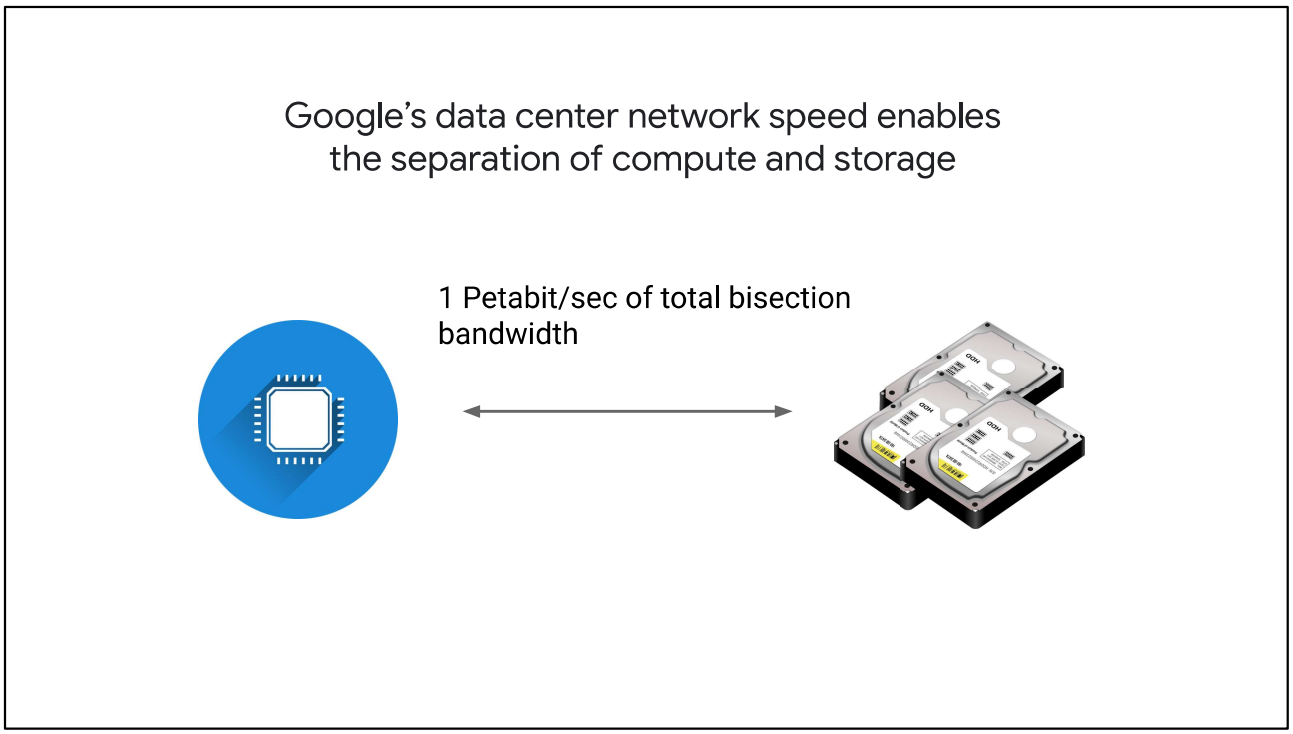
So far, we have looked at compute and at storage.
The third part of Google Cloud infrastructure is Networking.

Google's high-quality private network, petabit-bisectional bandwidth, edge points of presence are combined using state-of-the-art software-defined networking to deliver a powerful solution.

First, the private network.

Google has laid thousands of miles of fiber-optic cable that crosses oceans with repeaters (to amplify optical signals) and, as you can see in this amusing gif, it's shark proof!

Google's data centers around the world are interconnected by this private Google network, which, by some publicly available estimates, carries as much as 40% of the world's internet traffic every day. This is the largest network of its kind on Earth, and it continues to grow.



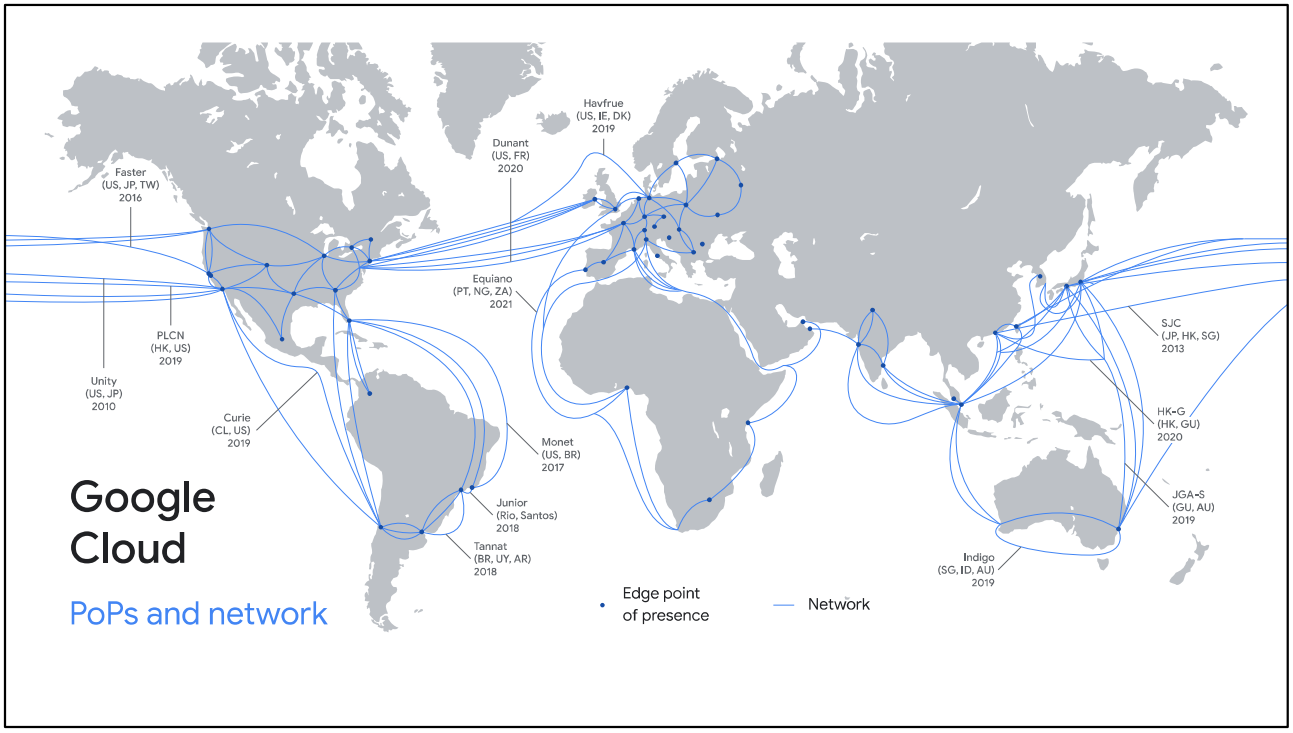
Second, the petabit bisectional bandwidth.

One of the themes we will discuss is the separation of compute and storage. You no longer need to do everything on a single machine or single cluster of machines with their own dedicated storage.

Why? Well if you have a fast-enough network, you can perform computations on data located elsewhere, like many distributed servers. Google's Jupiter network can deliver enough bandwidth to allow 100,000 machines to communicate with any other machine in the datacenter at over 10 Gbs.

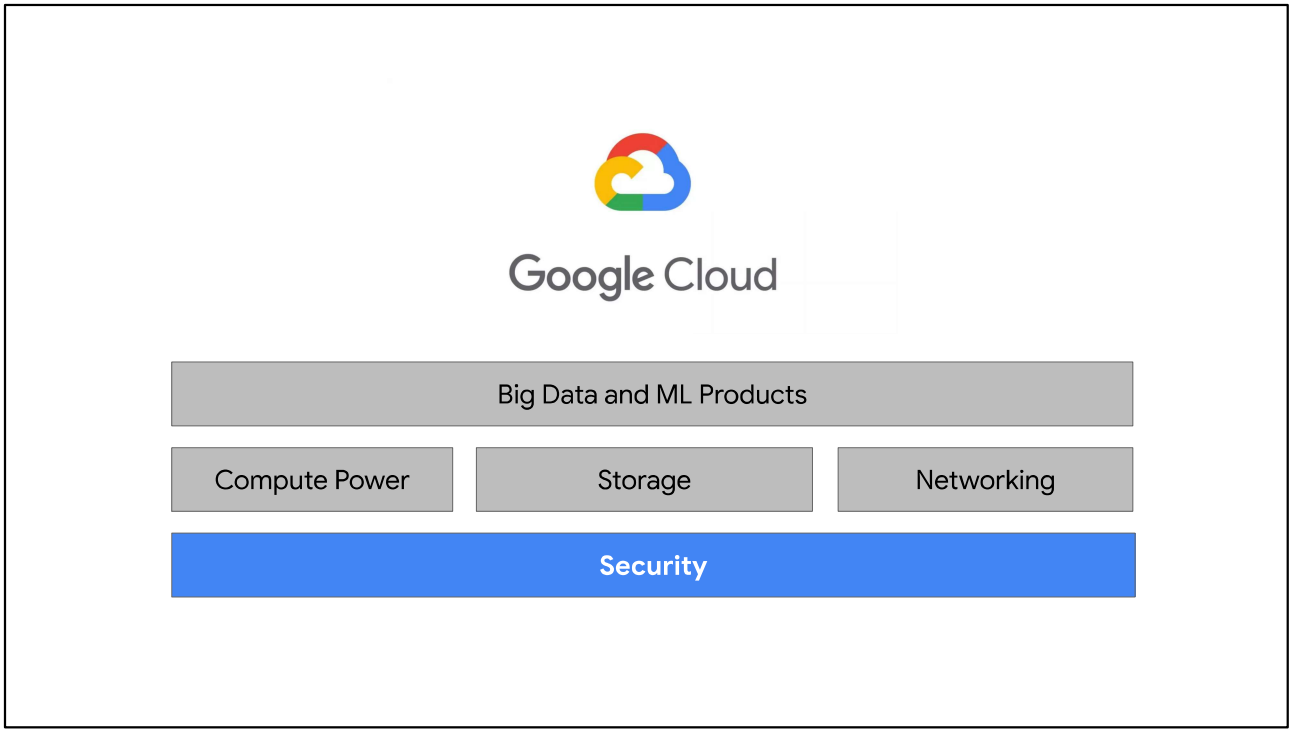
This full-duplex bandwidth means that locality within the cluster is not important. If every machine can talk to every other machine at 10 Gbps, racks don't matter for data analytics and ML training.

https://cloud.google.com/blog/products/gcp/bigquery-under-the-hood



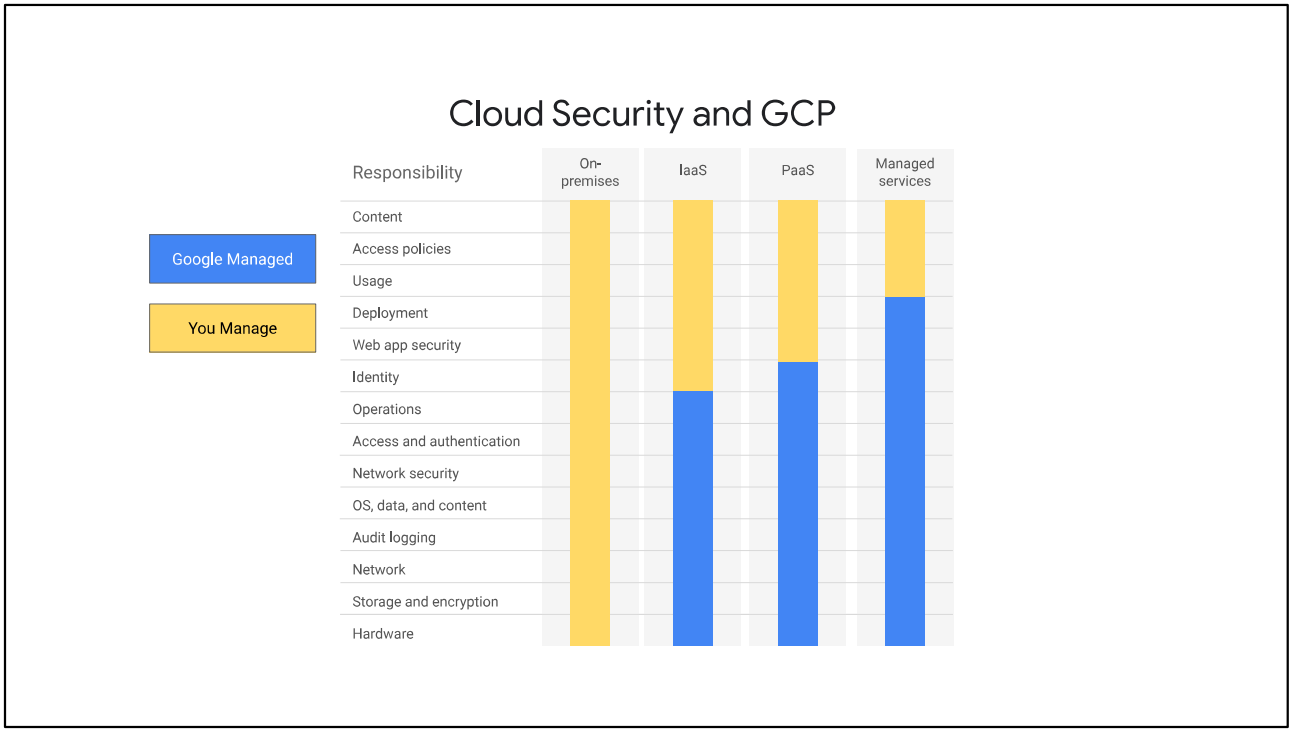
But you need to ingest data, probably from around the world. You need to serve out the results of your analytics and predictions. This is where edge points of presence come in.

The network interconnects with the public internet at more than 90 internet exchanges and more than 100 points of presence worldwide. When an internet user sends traffic to a Google resource, Google responds to the user's request from an Edge Network location that will provide the lowest delay or latency. Google's edge caching network places content close to end users to minimize latency. Your applications in GCP, like your machine learning models, can take advantage of this edge network too.



The last piece of core infrastructure underpinning your data pipelines and ML models is Google-grade security.

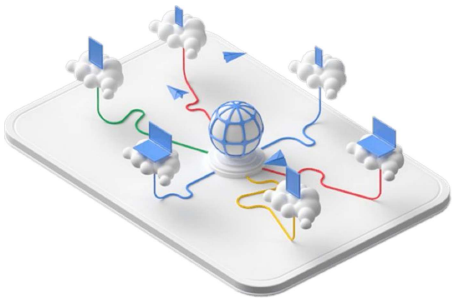
When you build an application on your on-premises infrastructure, you're responsible for the entire stack's security: from the physical security of the hardware and the premises in which they are housed, through the encryption of the data on disk, the integrity of your network, and all the way up to securing the content stored in those applications.



But when you move an application to GCP, Google handles many of the lower layers of security, like the physical security of the hardware and its premises, the encryption of data on disk, and the integrity of the physical network. Because of its scale, Google can deliver a higher level of security at these layers than most customers could afford to on their own.

The upper layers of the security stack, including the securing of data, remain your responsibility. Google provides tools like Cloud IAM to help you implement the policies you define at these layers.

Communications to Google Cloud are encrypted in transit

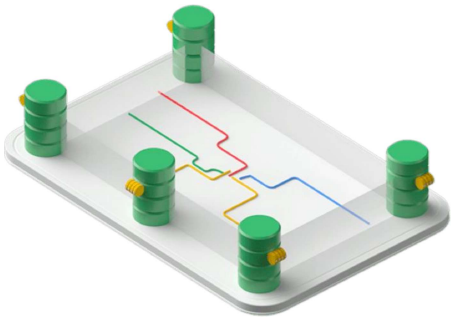


- In-transit encryption
- Multiple layers of security
- Backed by Google security teams 24/7

Communications over the Internet to our public cloud services are encrypted in transit. Google's network and infrastructure have multiple layers of protection to defend our customers against denial-of-service attacks.

https://cloud.google.com/security/encryption-in-transit/resources/encryption-in-transit-whitepaper.pdf
https://cloud.google.com/security/infrastructure/

Stored data is encrypted at rest and distributed



- Data automatically encrypted at rest
- Distributed for availability and reliability

Stored data is automatically encrypted at rest and distributed for availability and reliability. This helps guard against unauthorized access and service interruptions.

https://cloud.google.com/security/infrastructure/
https://cloud.google.com/security/encryption-at-rest/default-encryption/resources/encryption-whitepaper.pdf
https://cloud.google.com/security/encryption-in-transit/resources/encryption-in-transit-whitepaper.pdf

Spotlight: BigQuery granular control over data access

- BigQuery table data encrypted with keys (and those keys are also encrypted)
- Monitor and flag queries for anomalous behavior
- Limit data access with authorized views

One specific product I'll highlight here that you'll see a lot of in this course is BigQuery, Google Cloud's petabyte-scale analytics data warehouse. Data in a BigQuery table is encrypted using a data encryption key. Then, even those data-encryption keys are encrypted with key-encryption keys for additional security. This is known as envelope encryption. BigQuery also allows you to provide your own encryption keys should you wish.

Inside BigQuery, you also can monitor your team's BigQuery usage and running queries and proactively limit data at row and column level to specified groups. We'll cover BigQuery as a service in greater detail later.

https://cloud.google.com/bigquery/docs/customer-managed-encryption
https://cloud.google.com/bigquery/docs/share-access-views

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

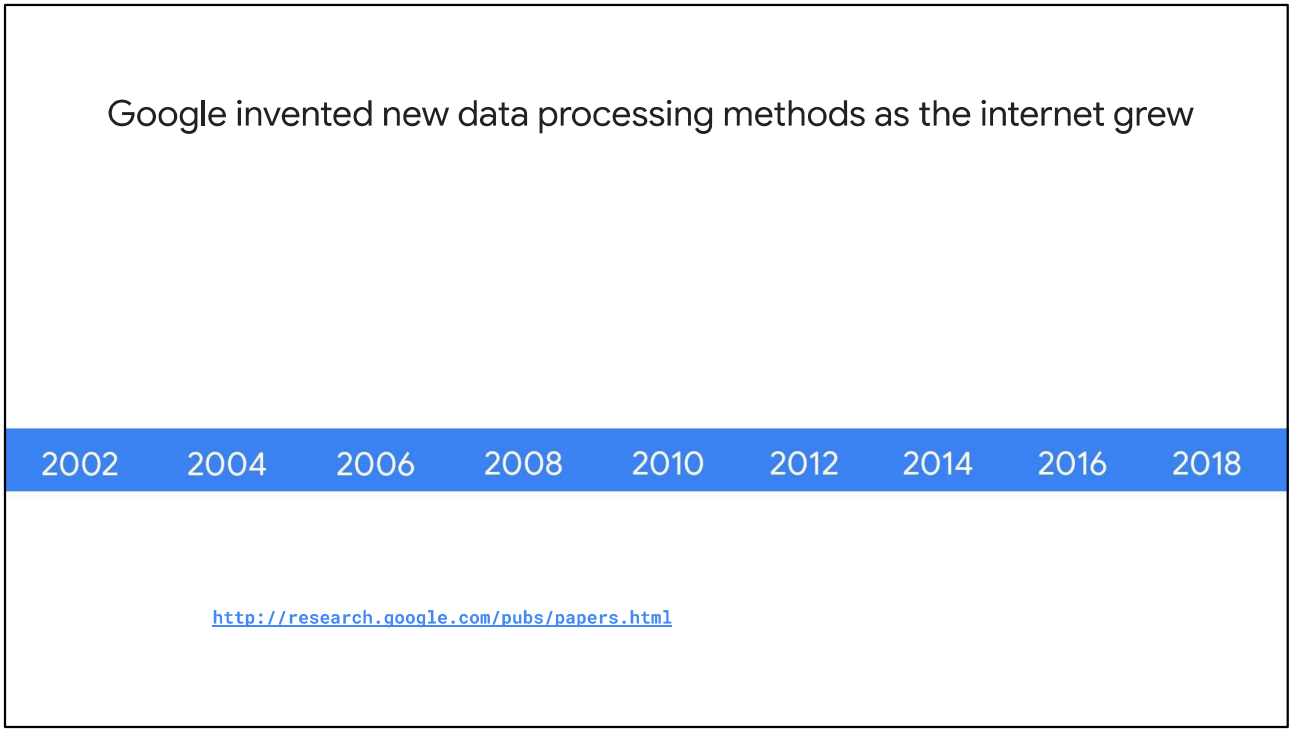
- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

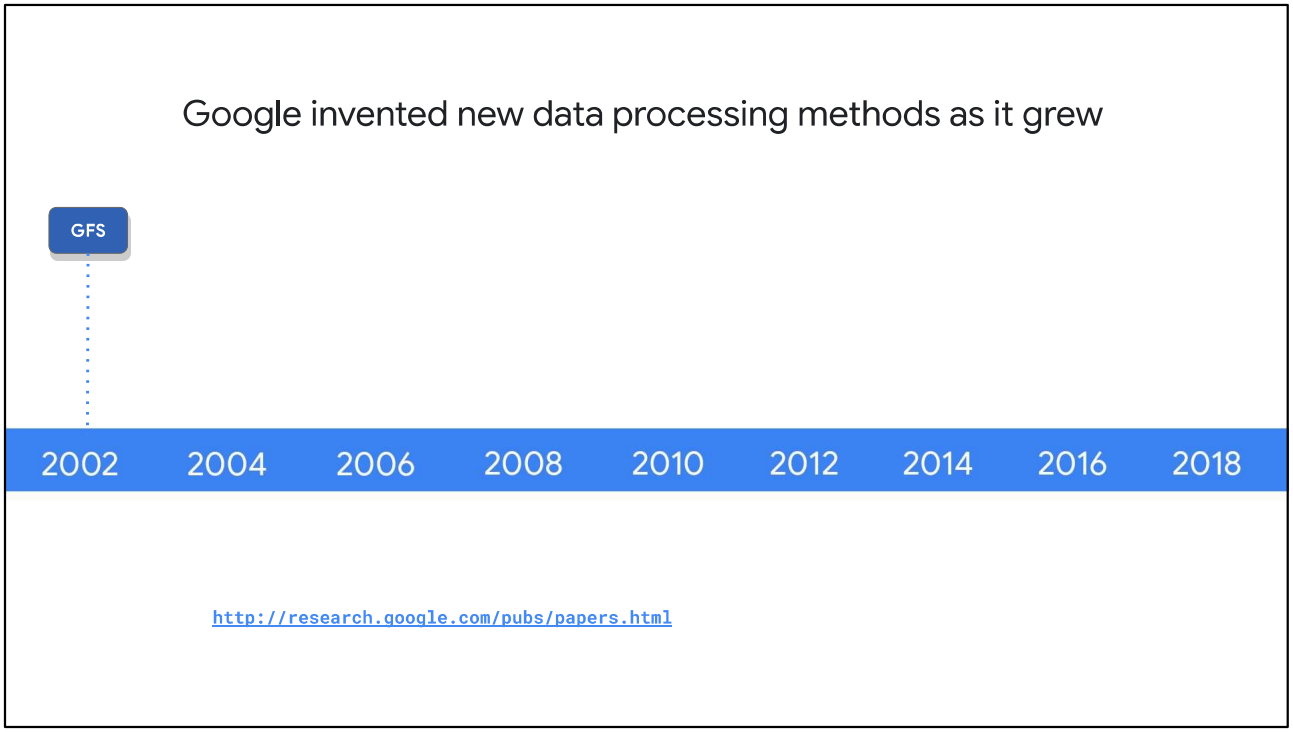
The different data roles in an organization

So far, we have talked about low-level infrastructure: compute, storage, networking, and security. However, as a data engineer or data scientist or data analyst, you will typically work with higher-level products. So, let's talk about the big data and ML products that form Google Cloud Platform.



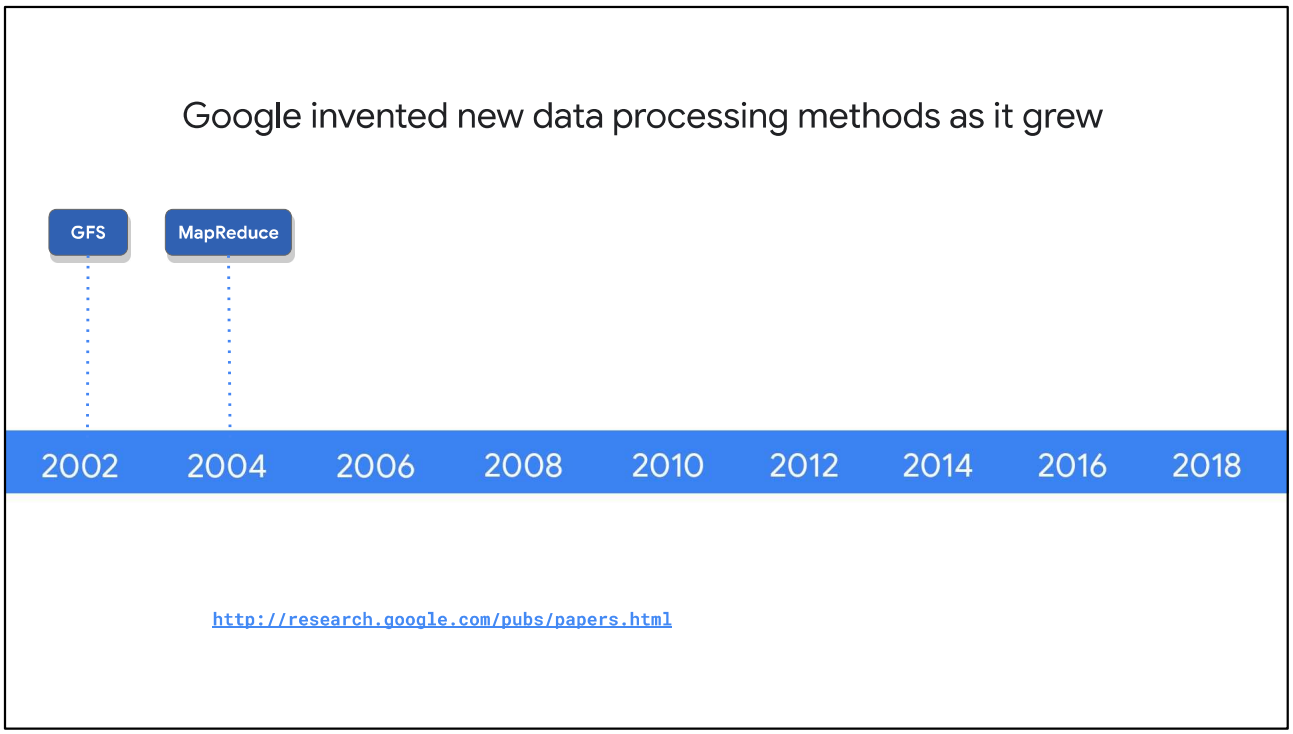
One of the interesting things about Google is that historically, we have faced issues related to large datasets, fast changing data, and varied data -- what is commonly called Big Data -- earlier than the rest of the world. Having to index the world wide web will do that … and so, as the internet grew, Google invented new data processing methods.

In 2002, Google created



GFS, or the Google file system, to handle sharding and storing petabytes of data at scale. GFS is the foundation for Cloud Storage and also what would become BigQuery managed storage.

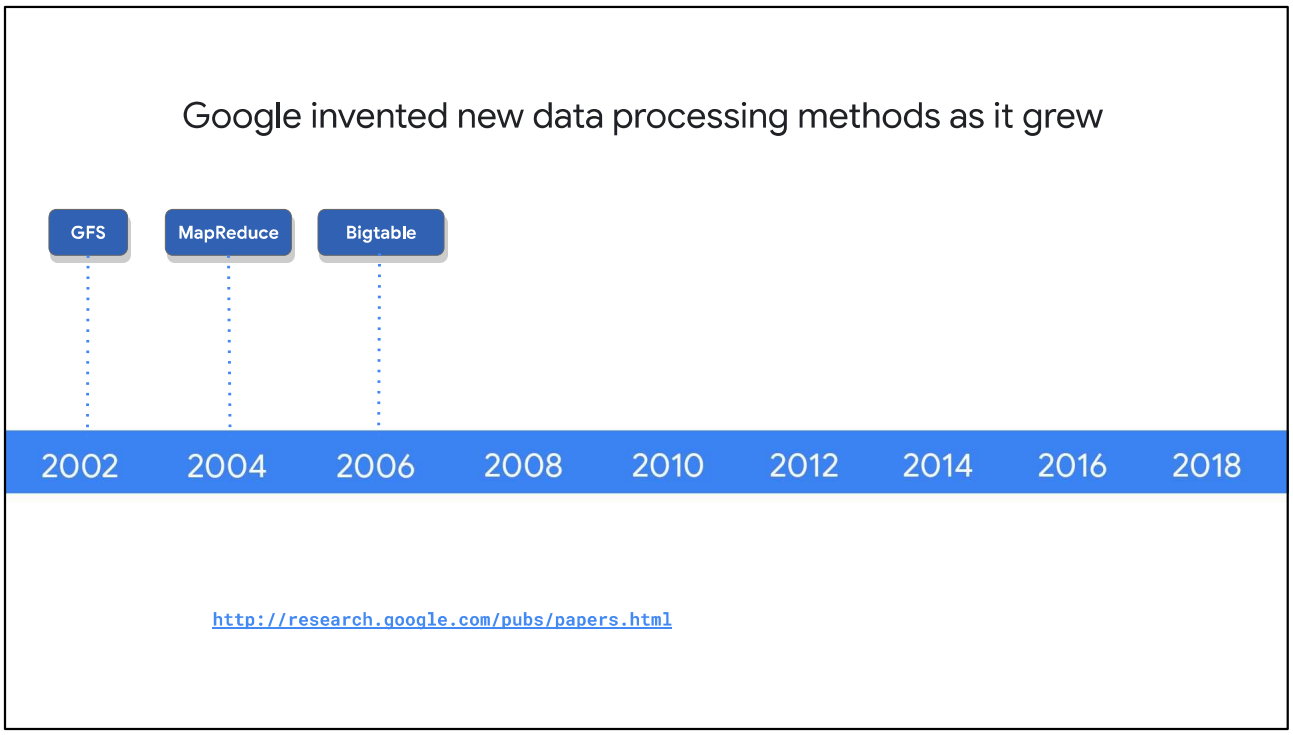
One of Google's next challenges was to figure out how to index the exploding volume of content on the web. To solve this, in 2004 Google invented a new style of data processing known as



MapReduce to manage large-scale data processing across large clusters of commodity servers. MapReduce programs are automatically parallelized and executed on a large cluster of commodity machines.

A year after Google published the white paper describing the MapReduce framework, Doug Cutting and Mike Cafarella created Apache Hadoop. Hadoop has moved far beyond its beginnings in web indexing and is now used in many industries for a huge variety of tasks that all share the common theme of variety, volume, and velocity of structured and unstructured data.

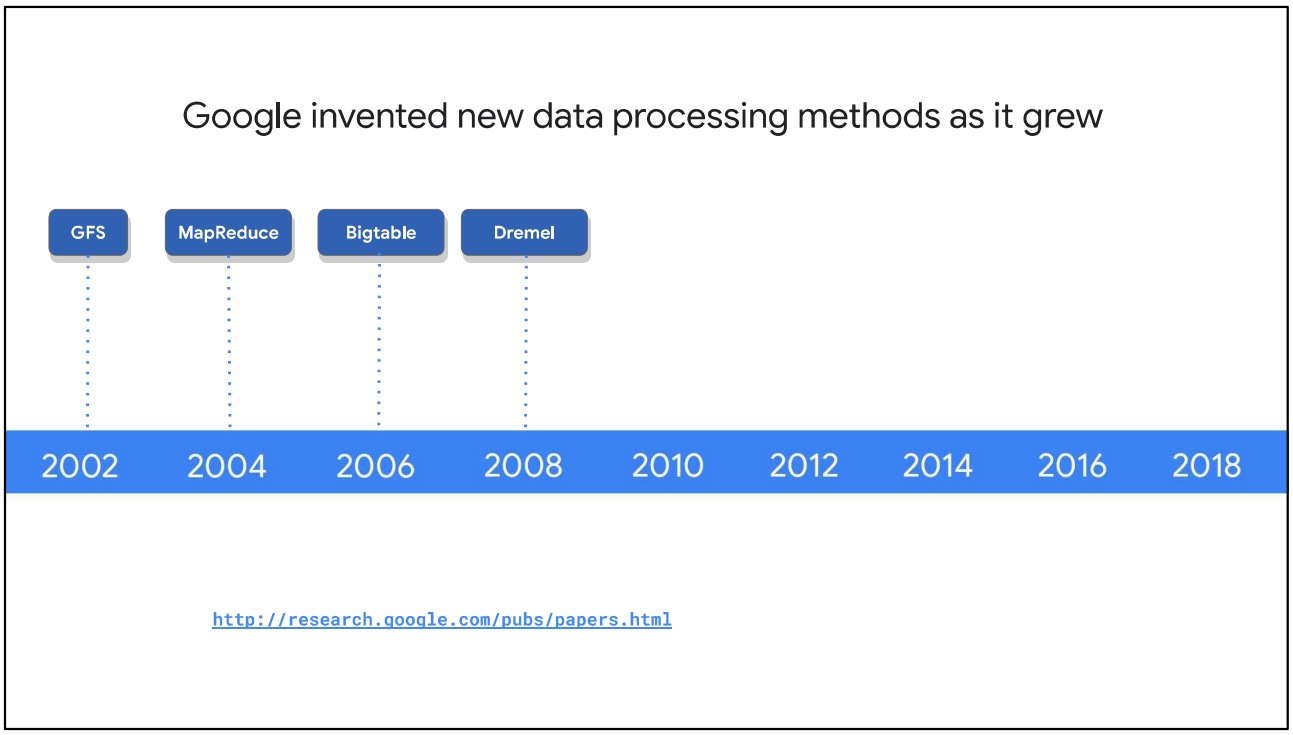
As Google's needs grew, we faced the problem of recording and retrieving millions of streaming user actions with high throughput. That became



Cloud Bigtable, which was an inspiration behind HBase/MongoDB.

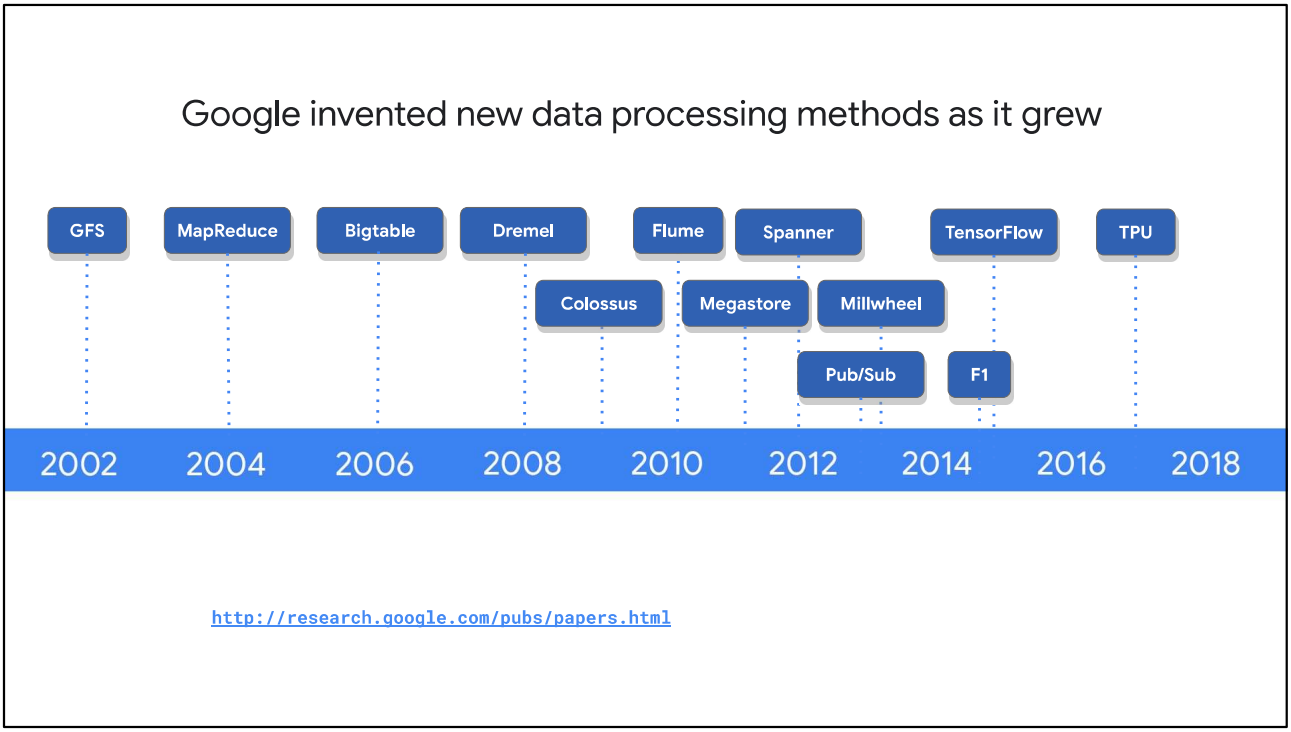
An issue with MapReduce is the need for developers to have to write code to manage lots of infrastructure. They couldn't just focus on their application logic.

So, in 2008 through 2010, Google started to moved away from MapReduce to process and query large datasets.

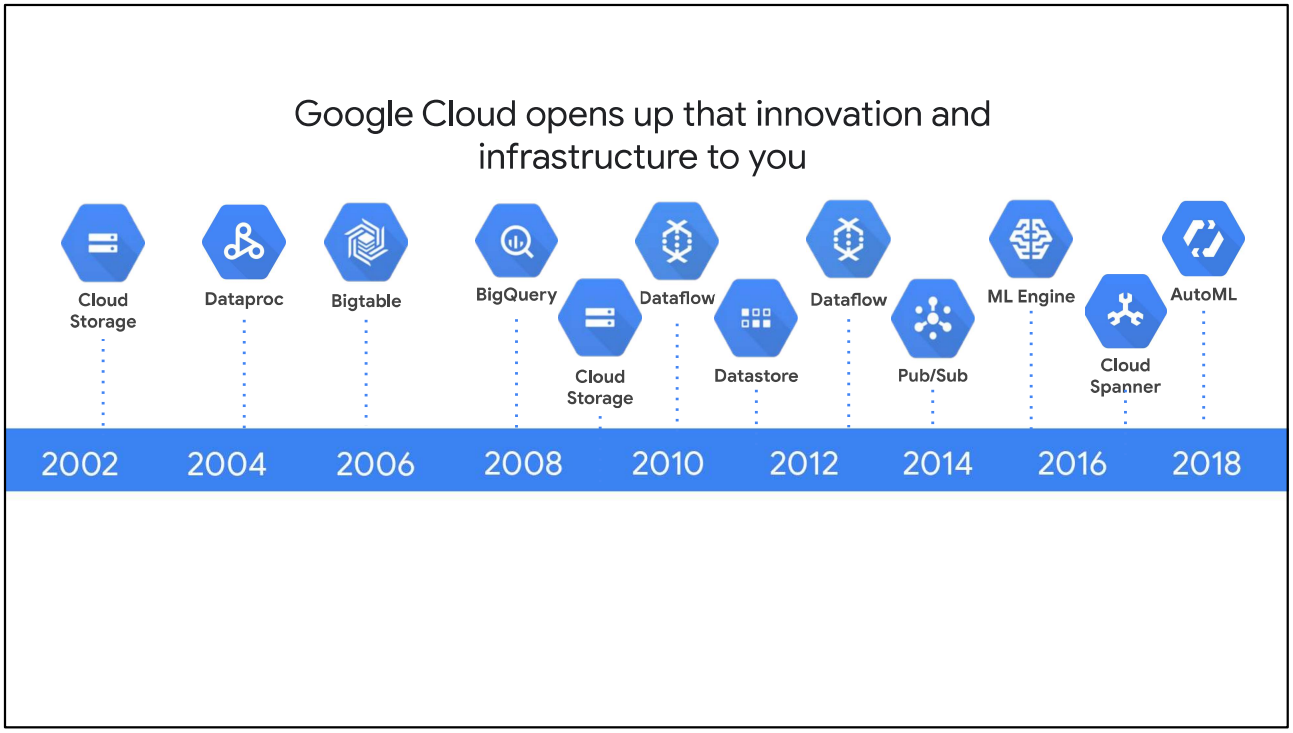


Dremel took a new approach to big data processing. Dremel breaks data into small chunks called shards and compresses them into a columnar format across distributed storage. It then uses a query optimizer to farm out tasks between the many shards of data and the Google data centers full of commodity hardware to process the query in parallel and deliver the results. The big leap forward was the service auto-manages data imbalances and communication between workers and autoscales to meet your query demand.

As you will soon see, Dremel became the query engine behind BigQuery.



Google continued to innovate to solve its big data and ML challenges and created Colossus as a next generation distributed data store, Spanner as a planet-scale relational database,
flume and millwheel for data pipelines, Pub/Sub for messaging, and TensorFlow for machine learning plus the specialized TPU hardware we saw earlier.



The good news for you is Google has opened up these innovations as products and services for you to leverage as part of the Google Cloud Platform.

You'll practice working in your labs with these very tools part of this fundamentals course.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

Choosing which big data and ML products are the right mix for your solution is a critical skill to learn. Later on in this module you'll get the opportunity to examine the architecture of real Google Cloud customers for inspiration.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

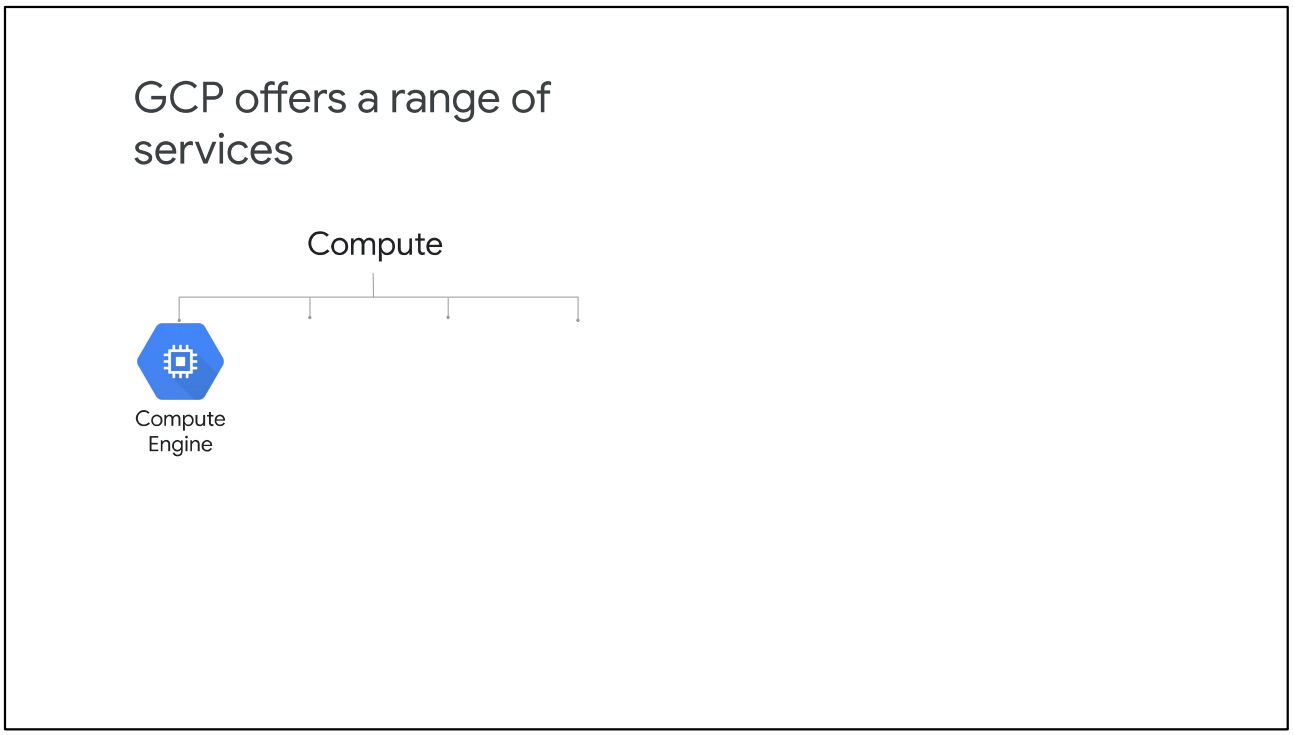
- Google innovation timeline
- Choosing the right approach

What you can do with GCP

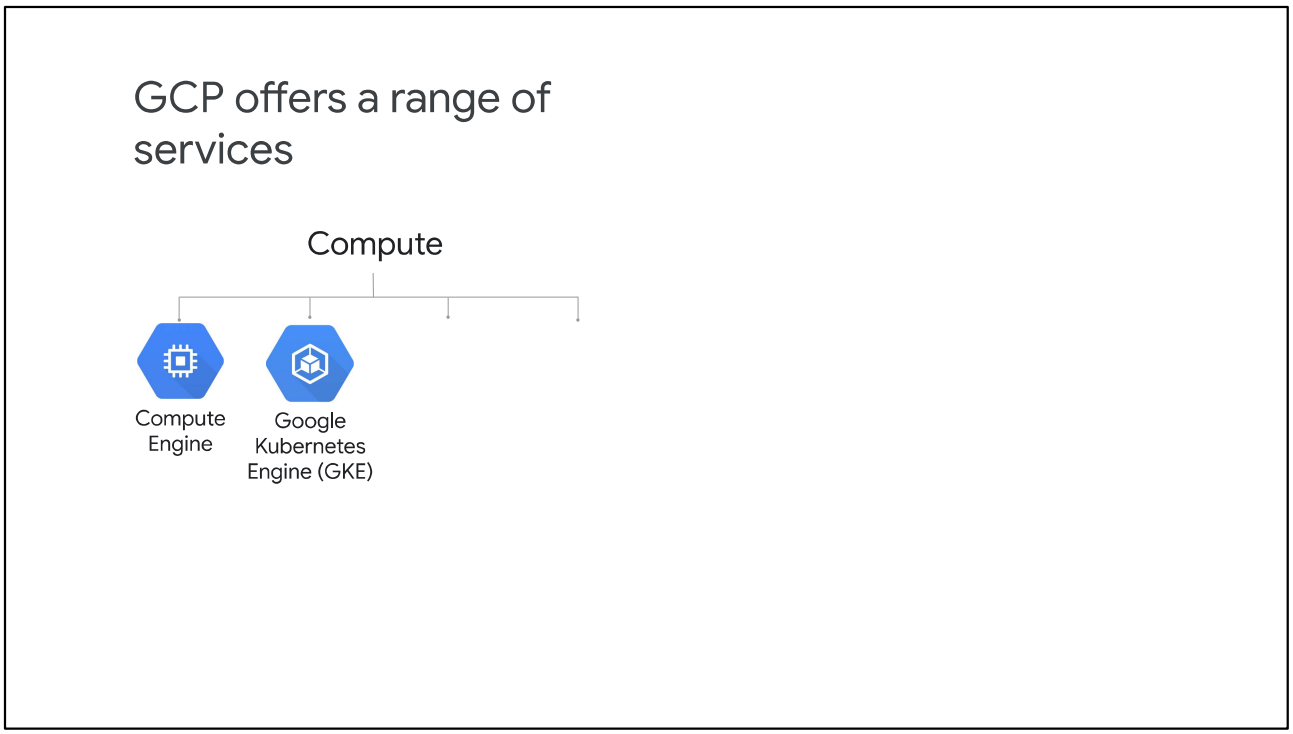
Activity: Explore a customer use case

The different data roles in an organization

Let's review the options available to you for compute and storage services so you can better interpret those use cases later.



The service that might be most familiar to newcomers is Compute Engine, which lets you run virtual machines on demand in the cloud. It's Google Cloud's Infrastructure-as-a-Service solution. It provides maximum flexibility for people who prefer to manage server instances themselves.

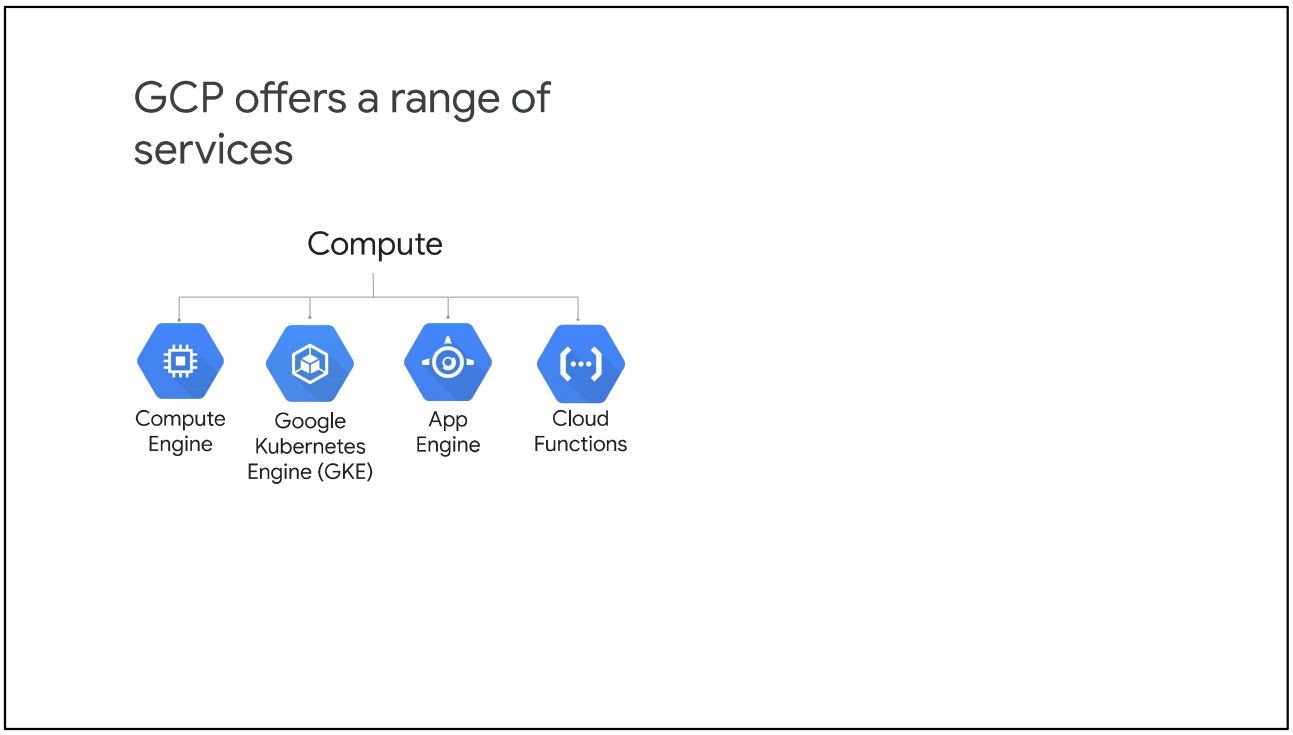


GKE, Google Kubernetes Engine, is different. Where Compute Engine is about individual machines running native code, GKE is about clusters running containers -- containers have code packaged up with all its dependencies.

GKE enables you to run containerized applications on a cloud environment that Google manages for you, under your administrative control. Containerization is a way to package code that's designed to be highly portable and to use resources very efficiently. Since most use cases involve multiple programs that need to execute, you need a way to orchestrate the containers running these separate programs. That's what Kubernetes does.

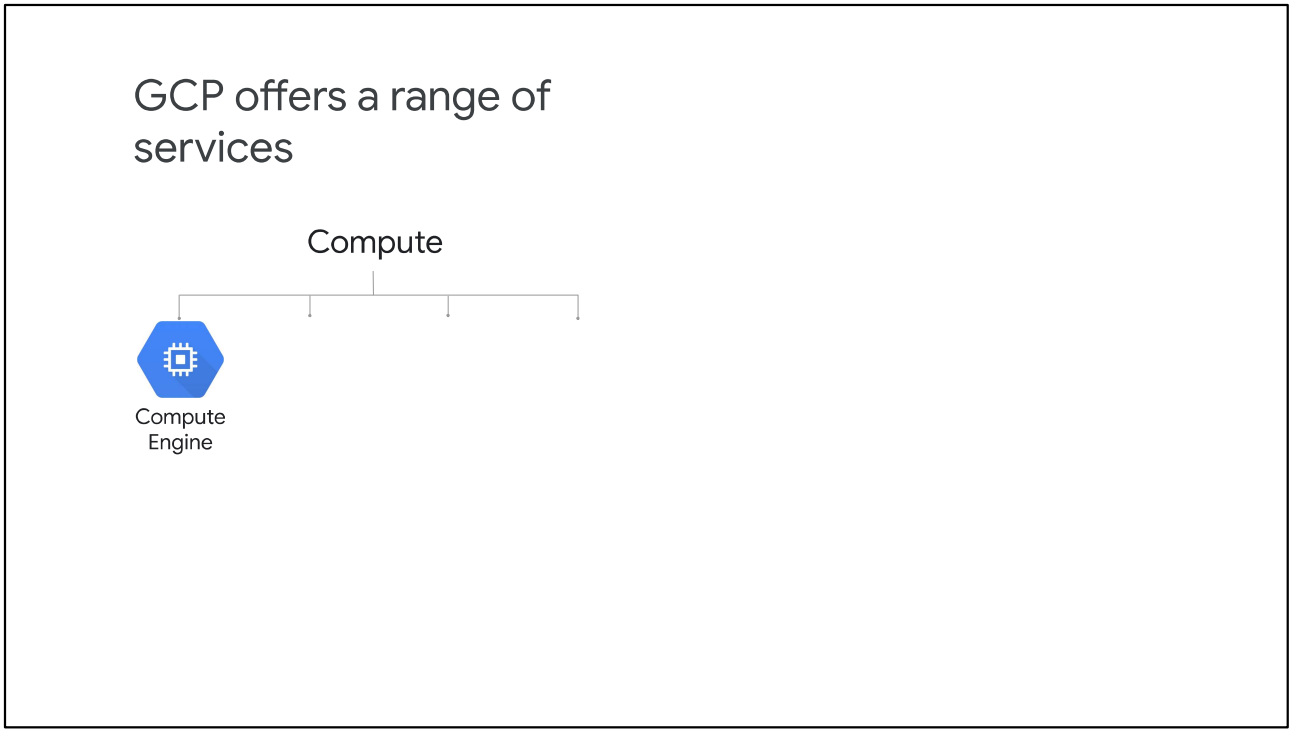
Kubernetes is a way to orchestrate code in containers. Although Kubernetes and GKE are outside the scope of this course, I'll link our cloud architecture specializations in the course resources.

App Engine is GCP's fully managed Platform-as-a-Service framework. That means it's a way to run code in the cloud without having to worry about infrastructure. You just focus on your code, and let Google deal with all the provisioning and resource management. You can learn a lot more about App Engine in the specialization "Developing Applications in Google Cloud Platform."



Cloud Functions is a completely serverless execution environment, or Functions-as-a-Service. It executes your code in response to events, whether those occur once a day or many times per second. Google scales resources as required, but you only pay for the service while your code runs.

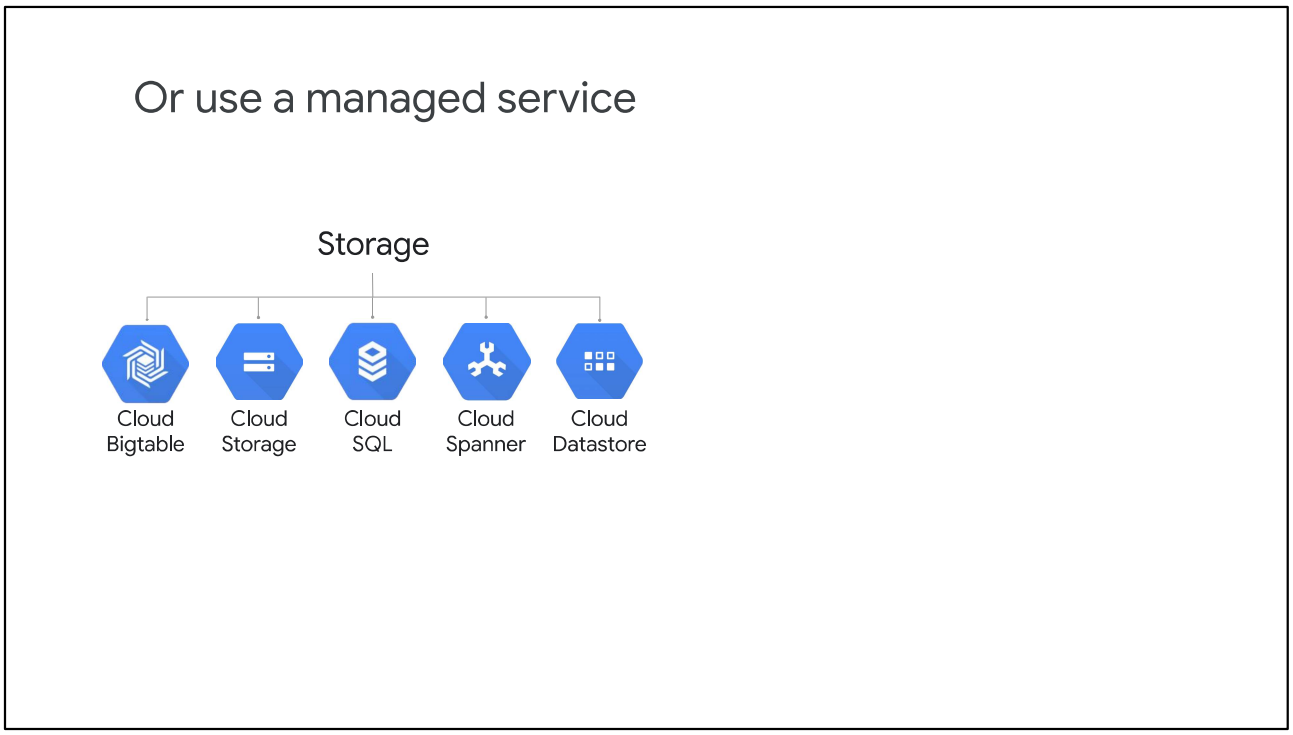
Typically, AppEngine is used for long-lived web applications that can autoscale from millions to billions of users. Cloud Functions are used for code that is triggered by an event such as a new file hitting Cloud Storage.



The fastest way to lift and shift your data workloads is by provisioning a VM and running your code. You'll experiment with this later on when you run Spark ML jobs on Cloud Dataproc, which spins up Compute Engine instances for your cluster.

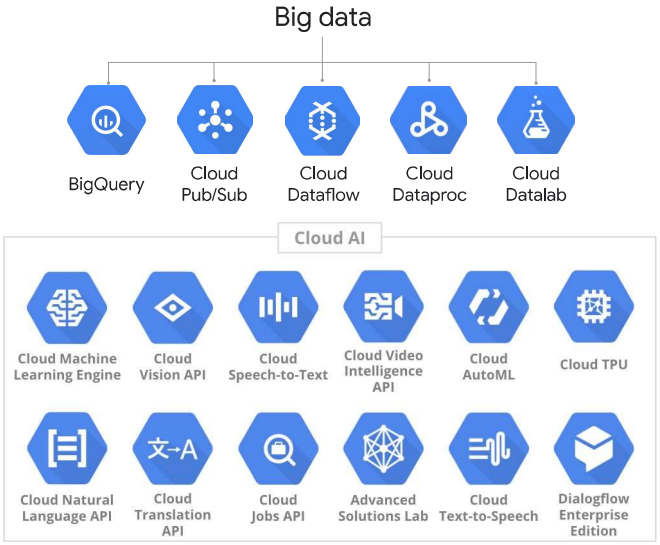
Build your own database solution

Most applications need a database of some kind. If you've built a cloud application, you can install and run your own database for it on a virtual machine in Compute Engine. You simply start up a virtual machine, install your database engine, and set it up, just like in your data center.

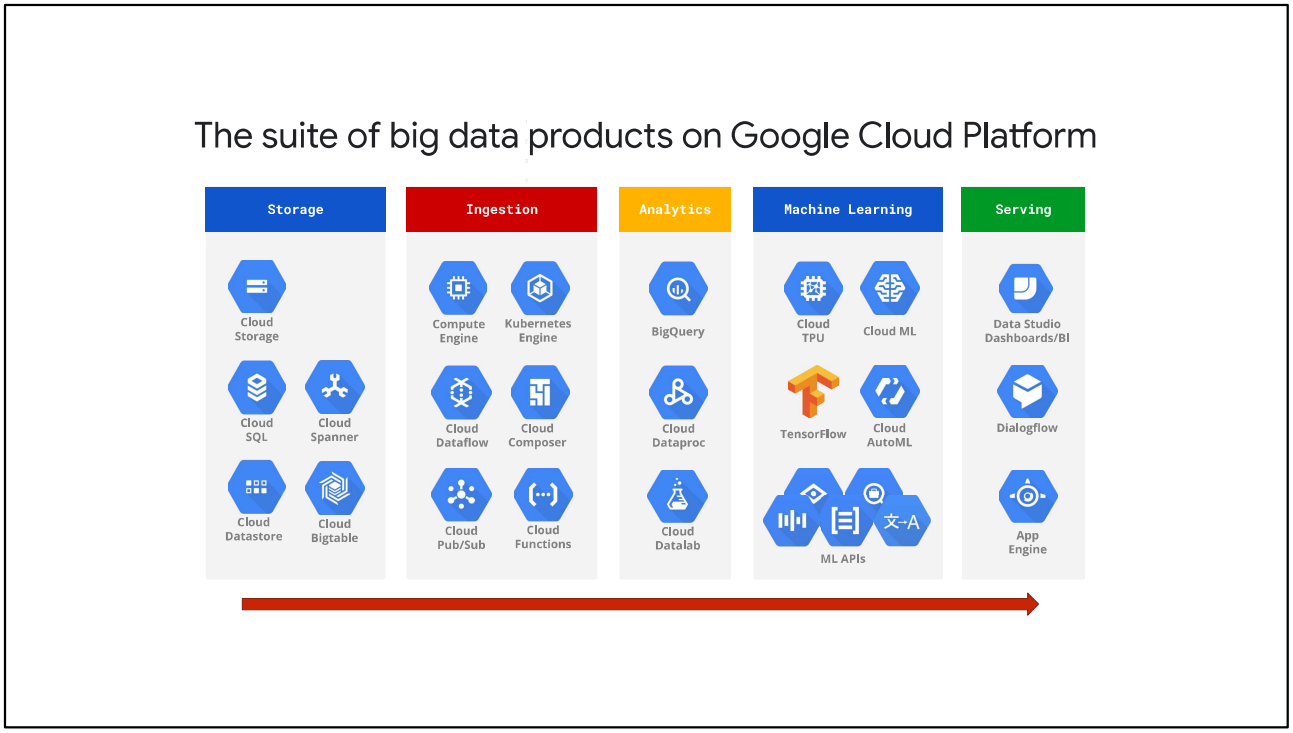


Alternatively, you can use Google's fully managed database and storage services. What all these -- Bigtable, Cloud Storage, Cloud SQL, Spanner, Datastore -- what all these have in common is that they reduce the work it takes to store all kinds of data. GCP offers relational and non-relational databases and worldwide object storage. You will learn more about these later in this course.

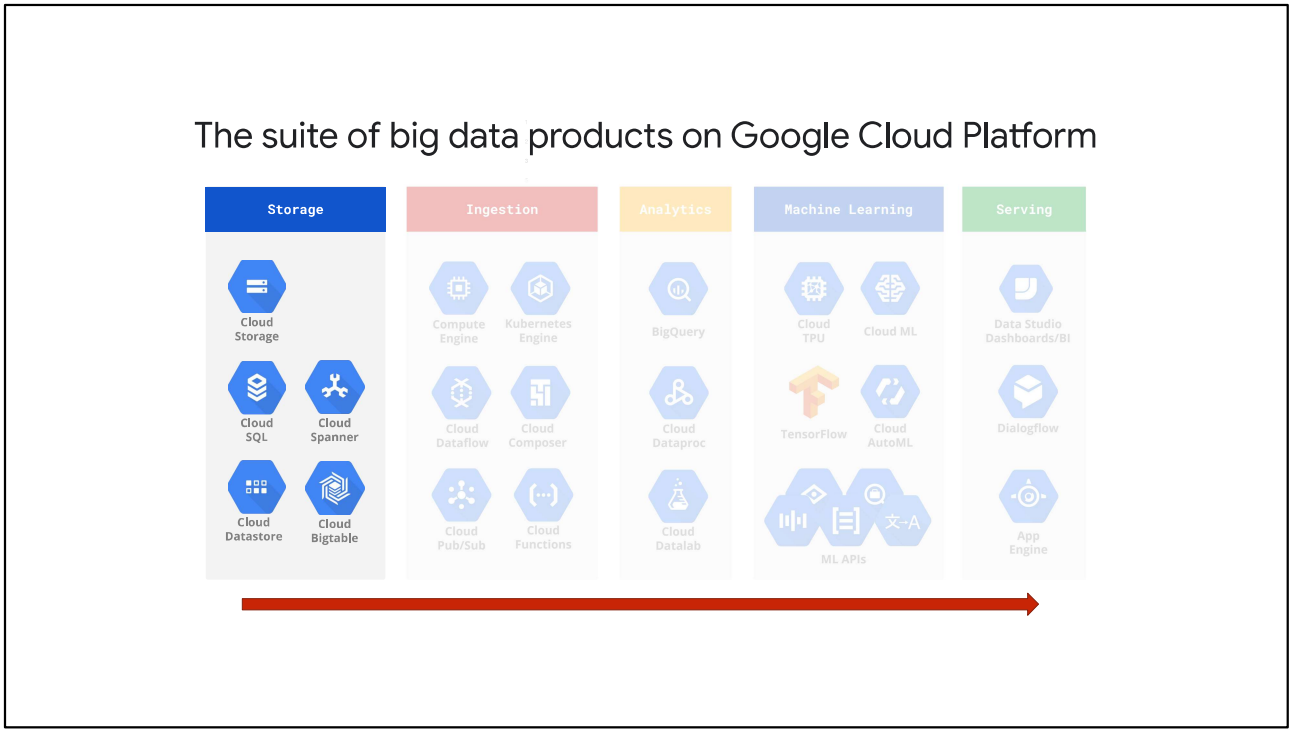
GCP offers a range of services



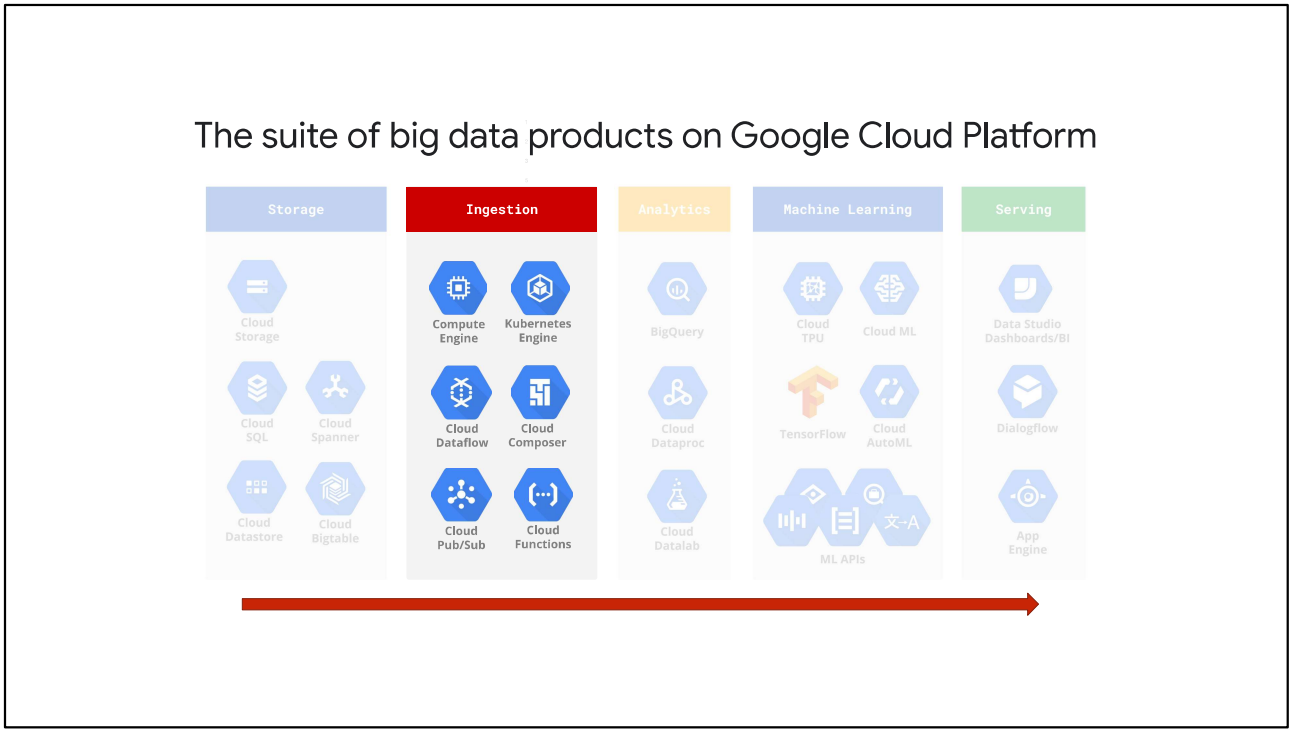
GCP also offers fully managed big data and machine learning services. Just as with storage and database services, you could build and implement these services yourself but why manage the infrastructure for compute and storage yourself when it can be fully-managed by Google Cloud?.



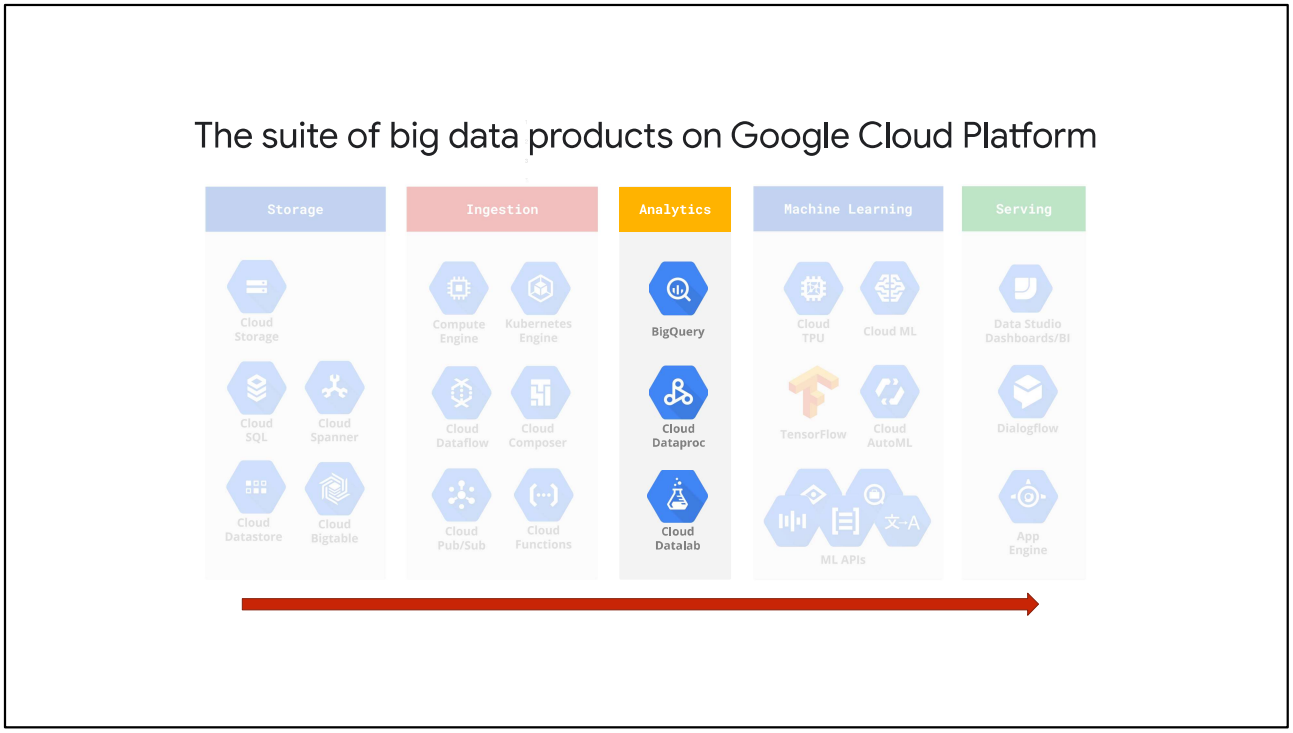
Here is the complete list of big data and ML products organized by where you would likely find them in a typical data processing workload.



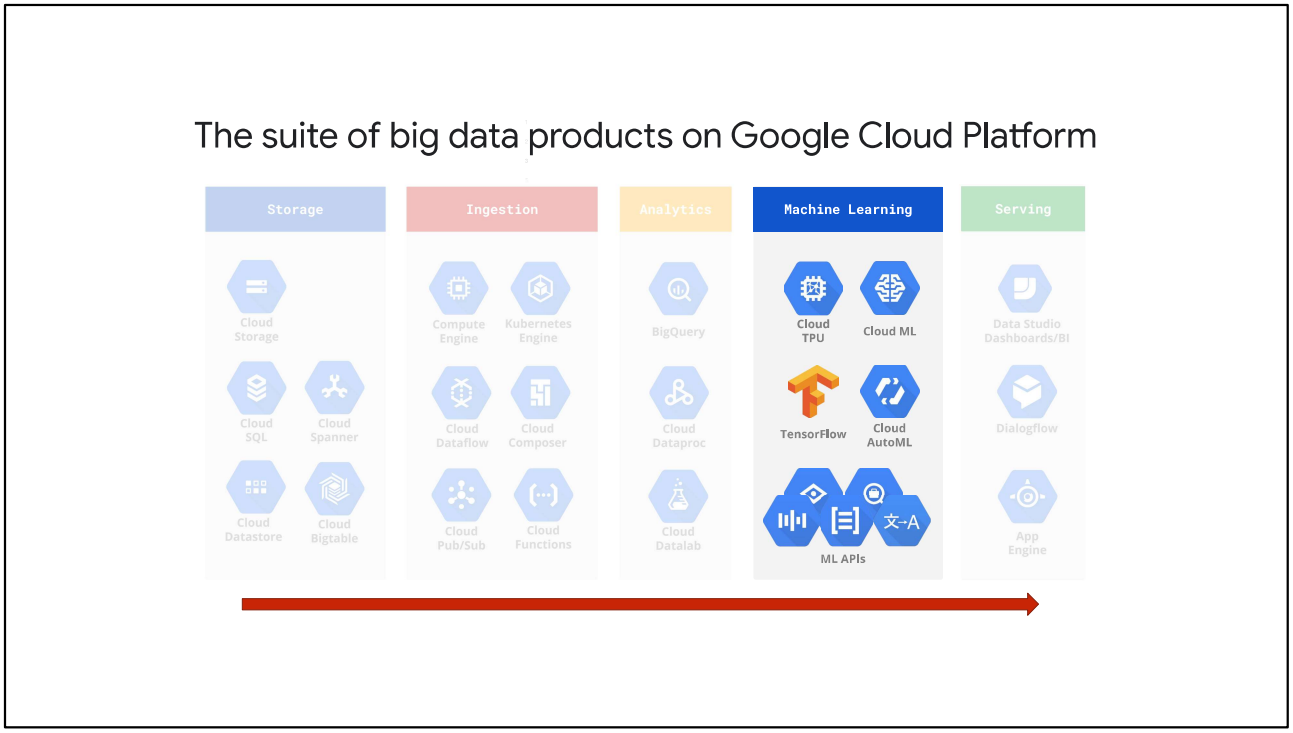
On the left you'll see the foundation to where your raw data is stored.



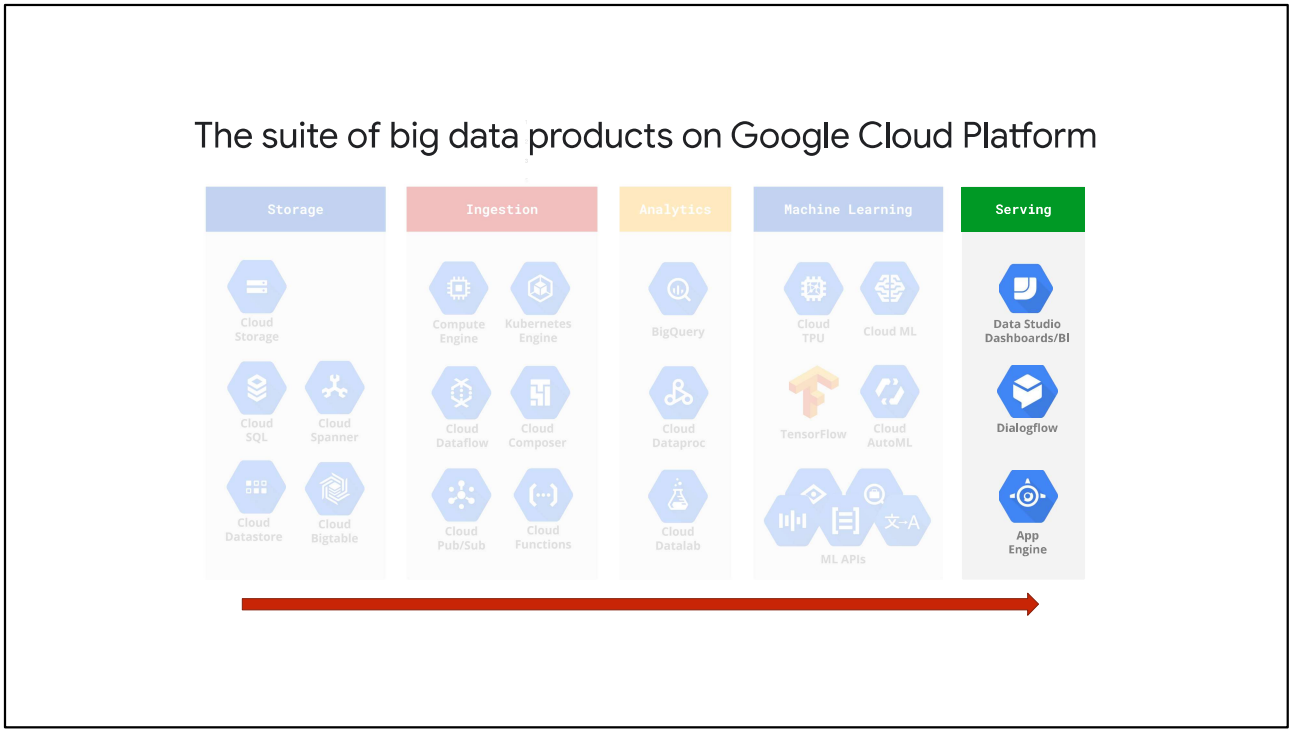
If your data isn't stored on GCP yet, you can ingest it using the tools you see next.



After your data is stored, you can analyze it using the tools in the third column



and run machine learning on it with the tools in the fourth column.



The last column is how you can serve your insights out to your users.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

Activity: Explore a customer use case

The different data roles in an organization

Now it's time to explore some of the cool big data and ML solutions that teams using Google Cloud have created. Then, you'll get a chance to find a use case and explore it in your activity.

Keller Williams uses AutoML Vision to automatically recognize common elements of house furnishings and architecture

Cloud AutoML Vision

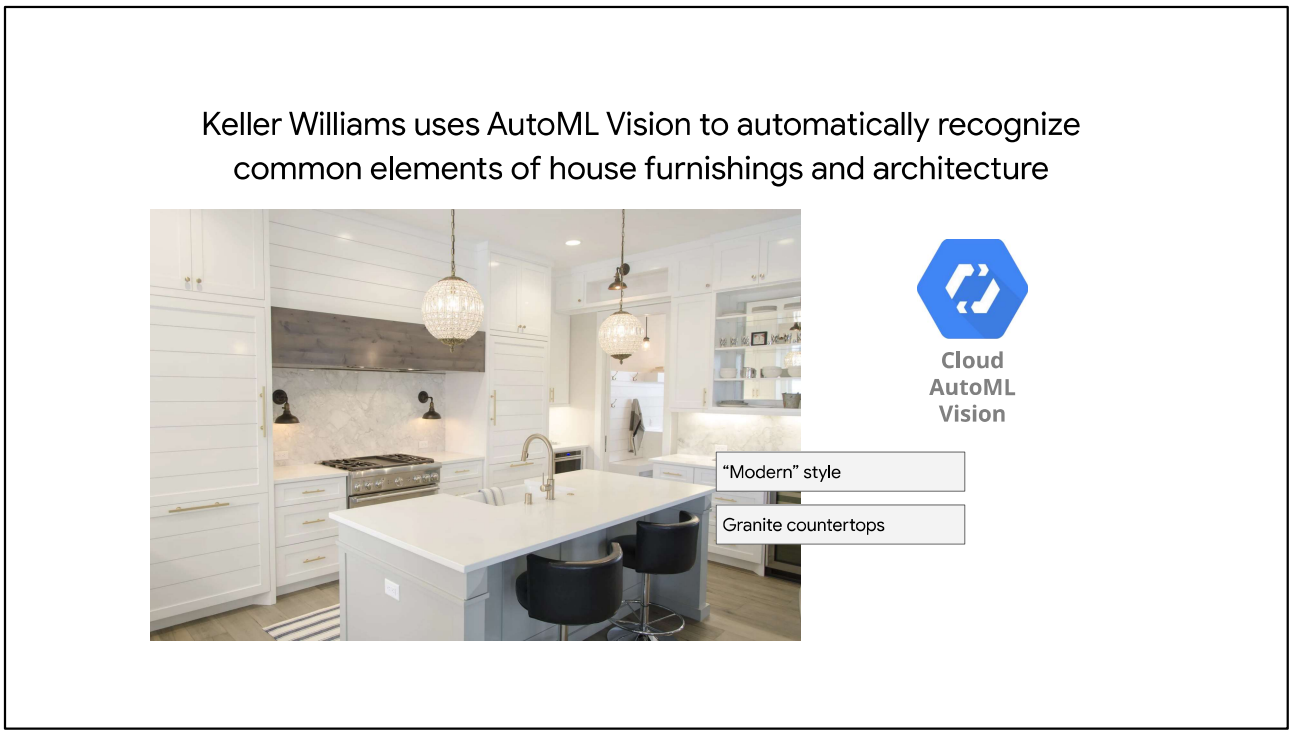
Keller Williams, a U.S. real estate company, uses AutoML Vision to automatically recognize specific features of houses like built-in bookcases. This helps agents get houses listed faster and buyers find houses that meet their needs.

Neil Dholakia, Chief Product Officer says “By training a custom model to recognize common elements of furnishings and architecture, customers can automatically search home listing photos for specific features like

Keller Williams uses AutoML Vision to automatically recognize common elements of house furnishings and architecture

Cloud AutoML Vision

‘granite countertops,’ or even more general styles

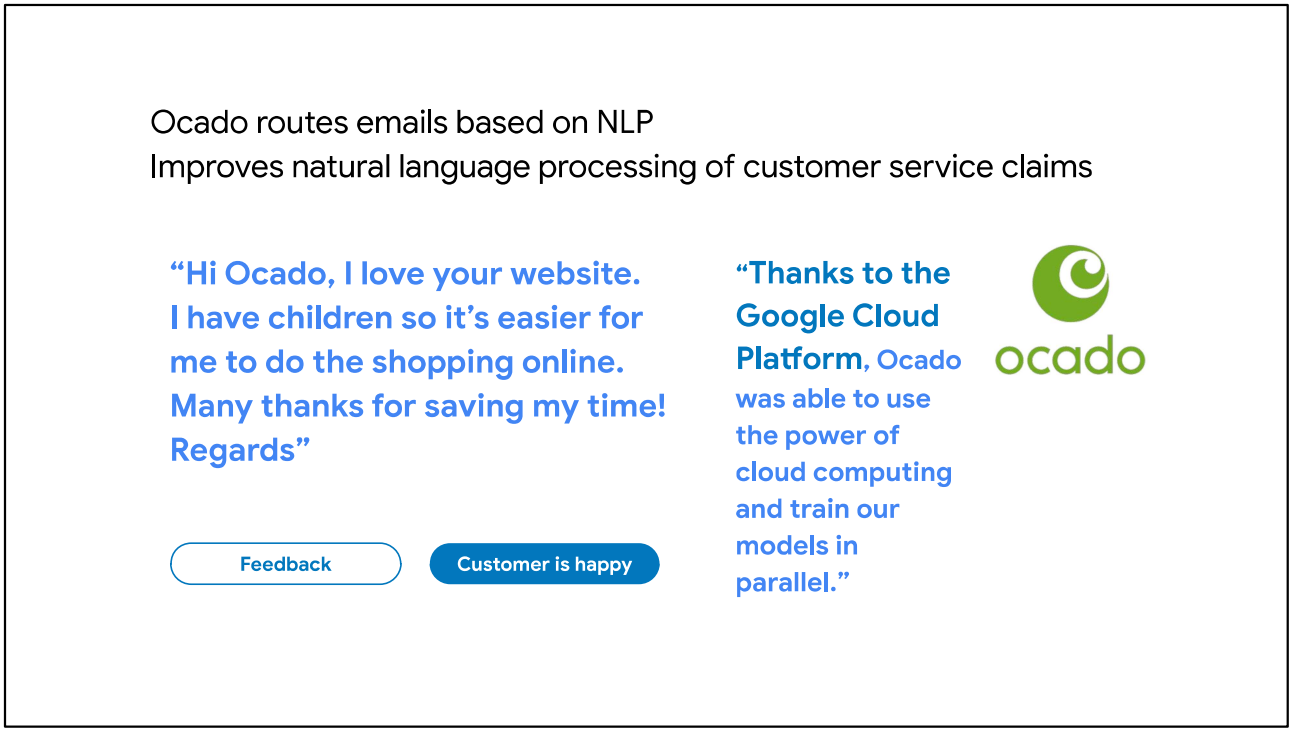


like 'modern.'"

This application of machine learning quickly allows Keller Williams realtors to record a video walkthrough of a new home and use the object detection capabilities of AutoML Vision to find and tag key aspects of the home that customers would want to search on.

A big benefit for their organization is that they already had many existing images and videos of home walkthroughs already. They simply fed them into the pre-built AutoML Vision model and customized it. All without writing a line of code. You'll learn more about AutoML Vision and practice creating models with it later in this course. [pause]

https://cloud.google.com/blog/products/gcp/empowering-businesses-and-developers-do-more-ai



Ocado, the U.K. online-only grocery supermarket, used ML to automatically route emails to the department that needs to process them, which avoids multiple rounds of reading and triaging.

With their old process, all the mails went to a central mailbox, where the email was read and then routed to the person or department that could handle it. This can’t scale, and led to long delays and poor user experience.

So, Ocado used machine learning, specifically the ability to process natural language, to discover customer sentiment and what each message was about, so they could route it immediately and automatically.

Kewpie uses ML to sort out the bad potatoes in baby food

Original process required humans to identify low-quality ingredients, which was expensive and stressful.

Machine learning was used to replicate the quality control process.

kewpie

One last use case.

Kewpie manufactures baby food. In this case, quality is not necessarily a matter of safety—because the food itself is safe—but discoloration can concern parents. So Kewpie turned to Google and our partner Brainpad to build a solution that leverages image recognition to detect low-quality potato cubes. The ML algorithm enabled them to free people from the tiring work of inspection and focus on other important work.

https://www.blog.google/products/google-cloud/how-ai-can-help-make-safer-baby-food-and-other-products/

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

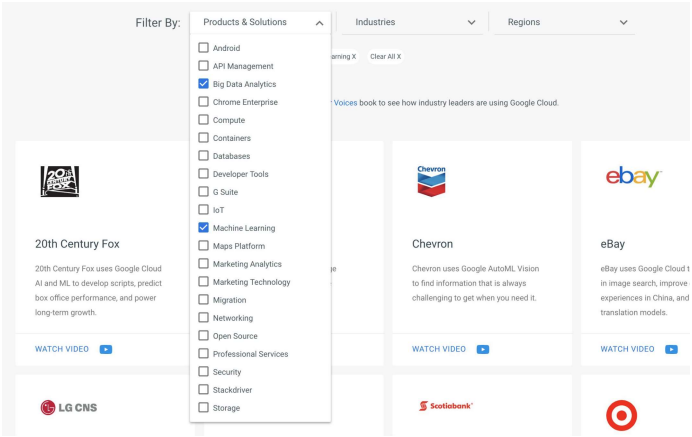
Activity: Explore a customer use case

The different data roles in an organization

Now it's your turn. One of the best ways to inspire and drive your team and projects to the cloud is to show your stakeholders examples from your industry; examples where someone has already succeeded with a solution.

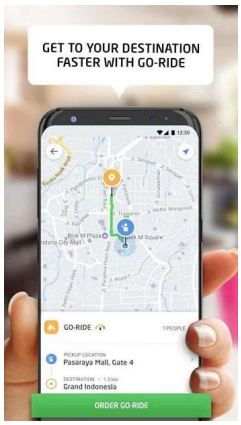
Your turn: Analyze real customer big data use cases

1. Navigate to cloud.google.com/customers/.
2. Filter Products & Solutions for Big Data Analytics.
3. Find an interesting customer use case.
4. Identify the key challenges, how they were solved with the cloud, and impact.



For this activity, navigate to cloud.google.com/customers and scroll down

GO-JEK brings goods and services to over
2 million families in 50 cities in Indonesia

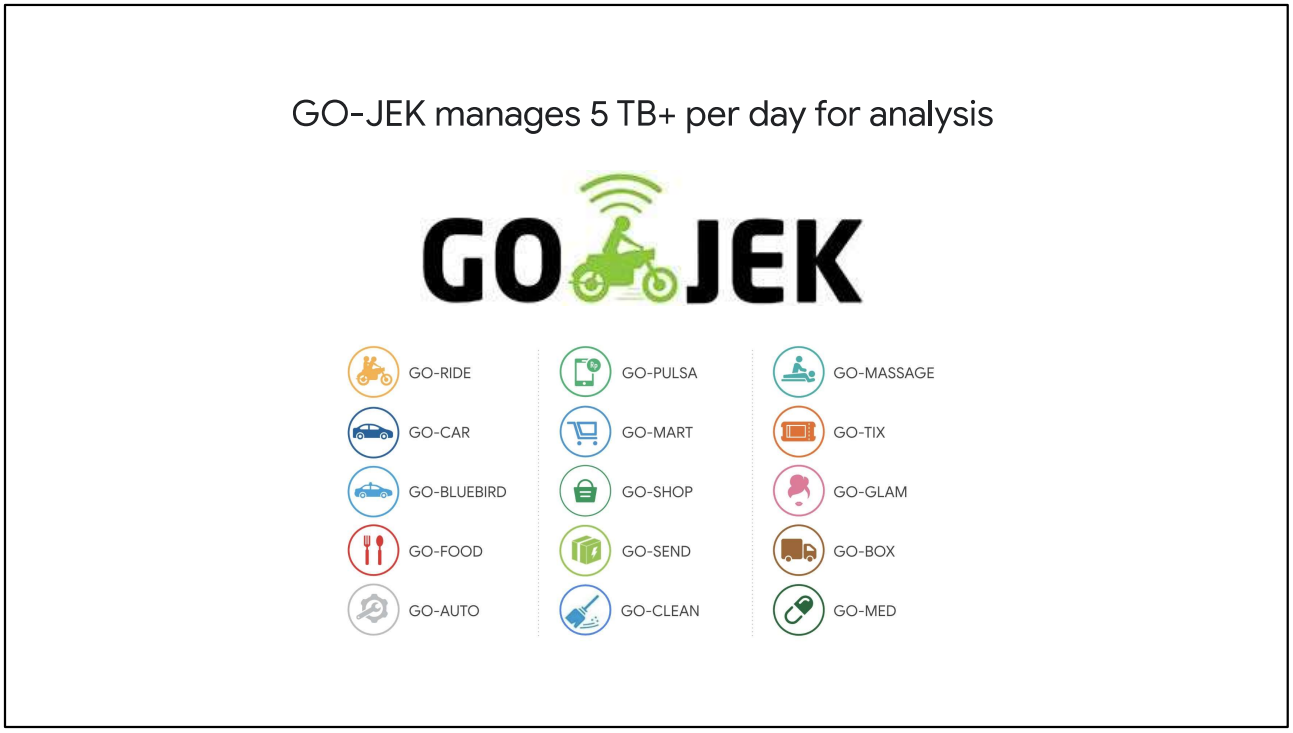


For our example we chose Go-JEK as they use a data engineering solution that maps nicely to the topics that we are going to cover as a part of this course.

GO-JEK is an Indonesia-based company that gives shared motorcycle rides, brings goods, and provides a wide variety of other services for over 2 million families across 50 cities in Indonesia.

https://cloud.google.com/customers/go-jek/

Their app has over 77 million downloads, and they are connected with over 150,000 merchants who sell through their delivery platform. And if you're interested in GIS data, they have over 1 million drivers delivering goods and giving rides across 50 cities.

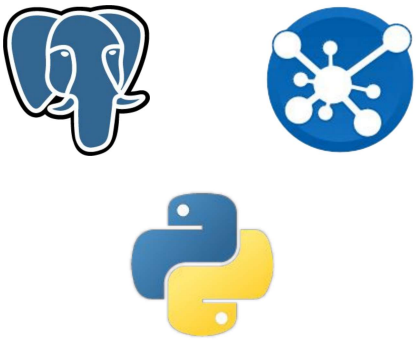


They manage more than 5 TB per day of data for analysis.

The CTO, Ajey Gore, gives this meaningful statistic:

“For example, we ping every one of our drivers every 10 seconds, which means 6 million pings per minute and 8 billion pings per day,” Gore says. “If you look at the scale and number of our customer interactions as well, we generate about 4 TB to 5 TB of data every day. We need to leverage this data to tell our drivers where demand from customers is strongest and how to get there.”

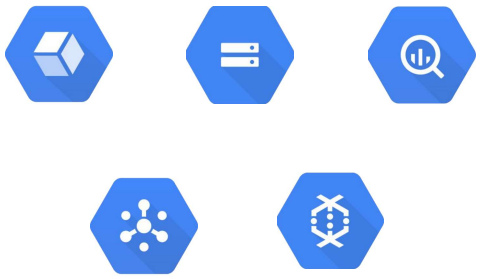
GO-JEK soon faced data scale and latency challenges



"Most of the reports are Day +1, so we couldn't identify the problems as soon as possible."

With the success of their on-demand motorcycle ride service, GO-JEK faced the challenges when looking to scale their existing big data platform. Their management team stated, "most of the reports are produced one day later so we couldn't identify the problems as soon as possible"

GO-JEK migrated their data pipelines to GCP



- High performance scalability with minimal operational maintenance
- More granular data with high velocity and less latency (stream processing)
- The ability to solve business problems with real time data insights

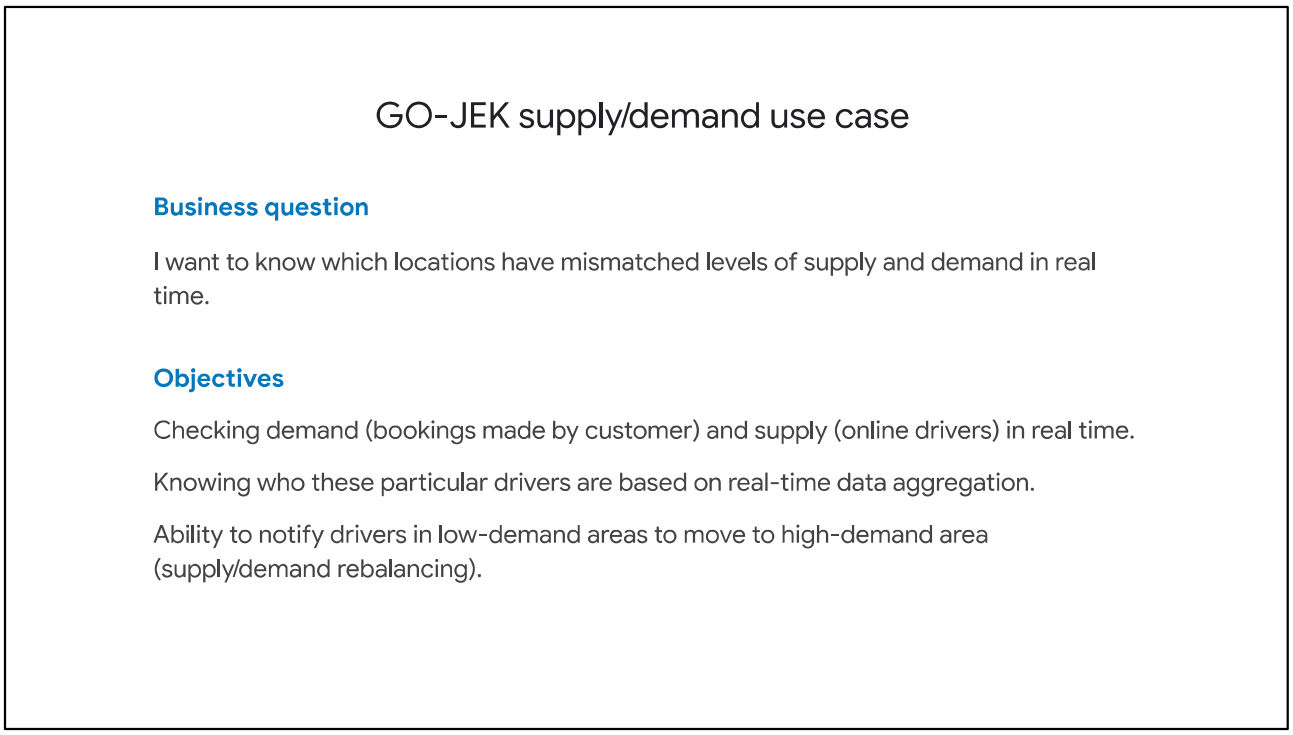
GO-JEK chose Google Cloud Platform and migrated their data pipelines to GCP for high performance with minimal day-to-day maintenance.

Their data engineering team uses Cloud Dataflow for streaming data processing and Google BigQuery for real-time business insights.

Their end-to-end architecture looks like this.

First, they ingest data from their mobile app, online, and IoT devices on vehicles (like GPS tracking for deliveries) into Cloud Pub/Sub. Then the data is brought into Cloud Dataflow for processing from Pub/Sub and a variety of other data sources to enrich the event data. Finally, after processing, the data is streamed into BigQuery as a data warehouse.

Keynote at NeXT 2018: https://www.youtube.com/watch?v=X1AwuBA2VlQ

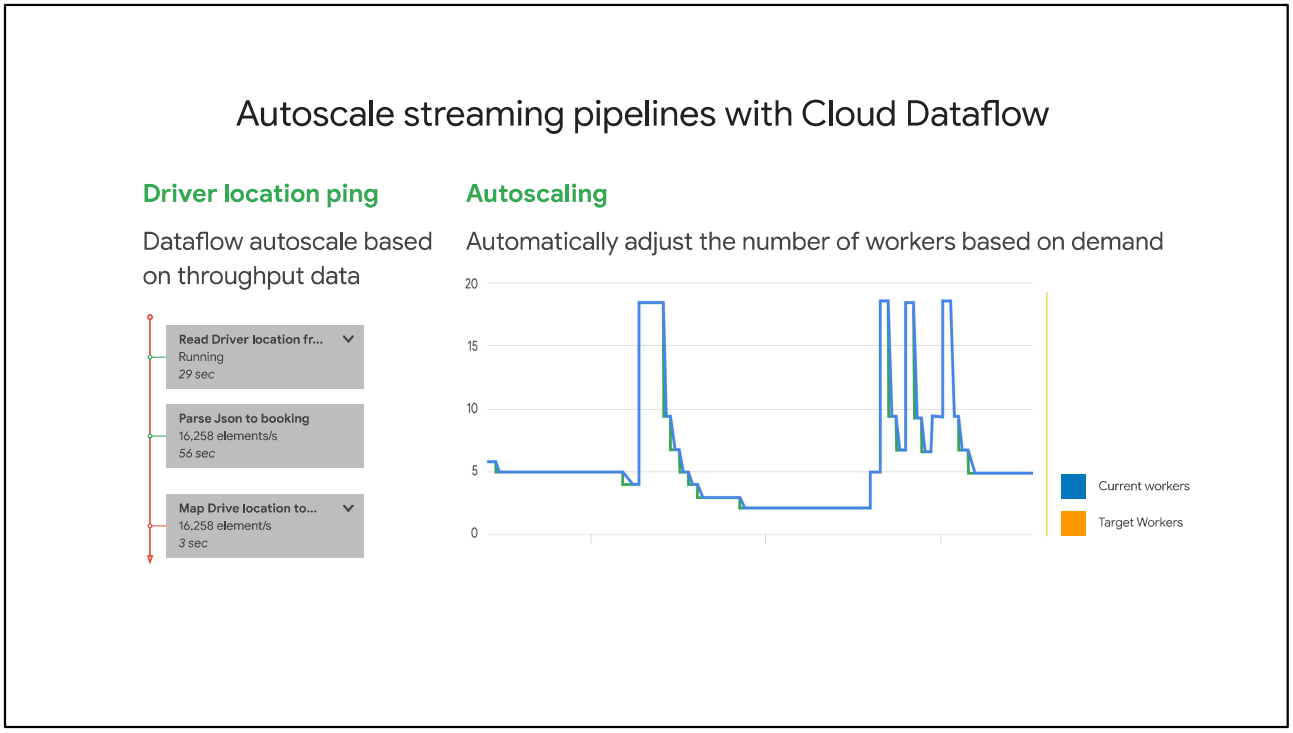


Here is an example of one of the problems the Go-JEK team solved.

The question was how could they quickly know which locations had too many or too few drivers to meet the demand of that area.

To solve this problem, the team needed to:

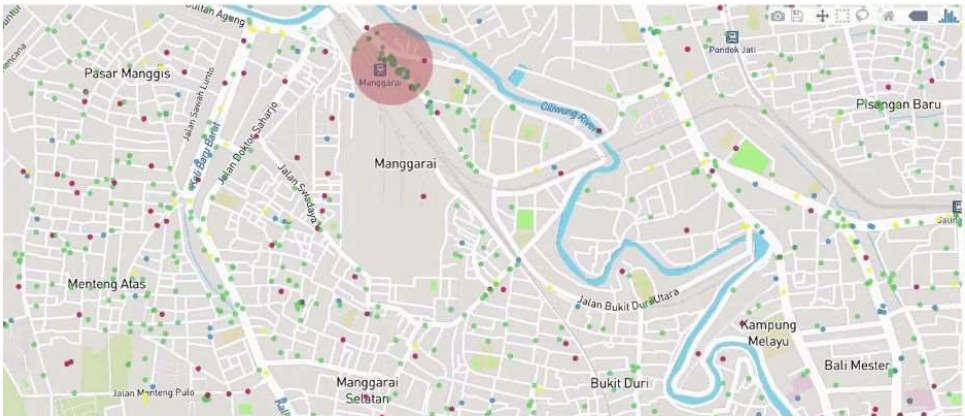
- Check the demand of bookings by customer against the supply of drivers in real time.
- Then the team needed to identify who these drivers are and finally notify them to re-route to higher demand areas.



How they achieved this was by building a streaming event data pipeline using Cloud Dataflow. Driver locations would ping out to pub/sub every 30 seconds and into Dataflow for processing. The pipeline then aggregates the supply pings from the drivers against the requests for bookings and connects to Go-JEKs notification system to alert drivers.

From a technology standpoint, the system needs to handle an arbitrarily high throughput of messages and scale up and down to process. Cloud Dataflow automatically manages the number of workers processing the pipeline to meet demand.

Visualize demand/supply mismatches with GIS data



The Go-Jek team is then able to visualize and highlight supply/demand mismatch areas for management reporting as you see in this example here. The green dots represent riders and new booking requests and the red dots are the drivers. You see the areas with the highest mismatches of supply and demand highlighted in red here like the train station which has many booking requests but few drivers.

The team can now actively monitor and ensure that they are sending drivers to the areas in highest demand which means faster booking times for riders and more fares for the drivers.

Agenda

Google Cloud Platform infrastructure

- Compute
- Storage
- Networking
- Security

Big data and ML products

- Google innovation timeline
- Choosing the right approach

What you can do with GCP

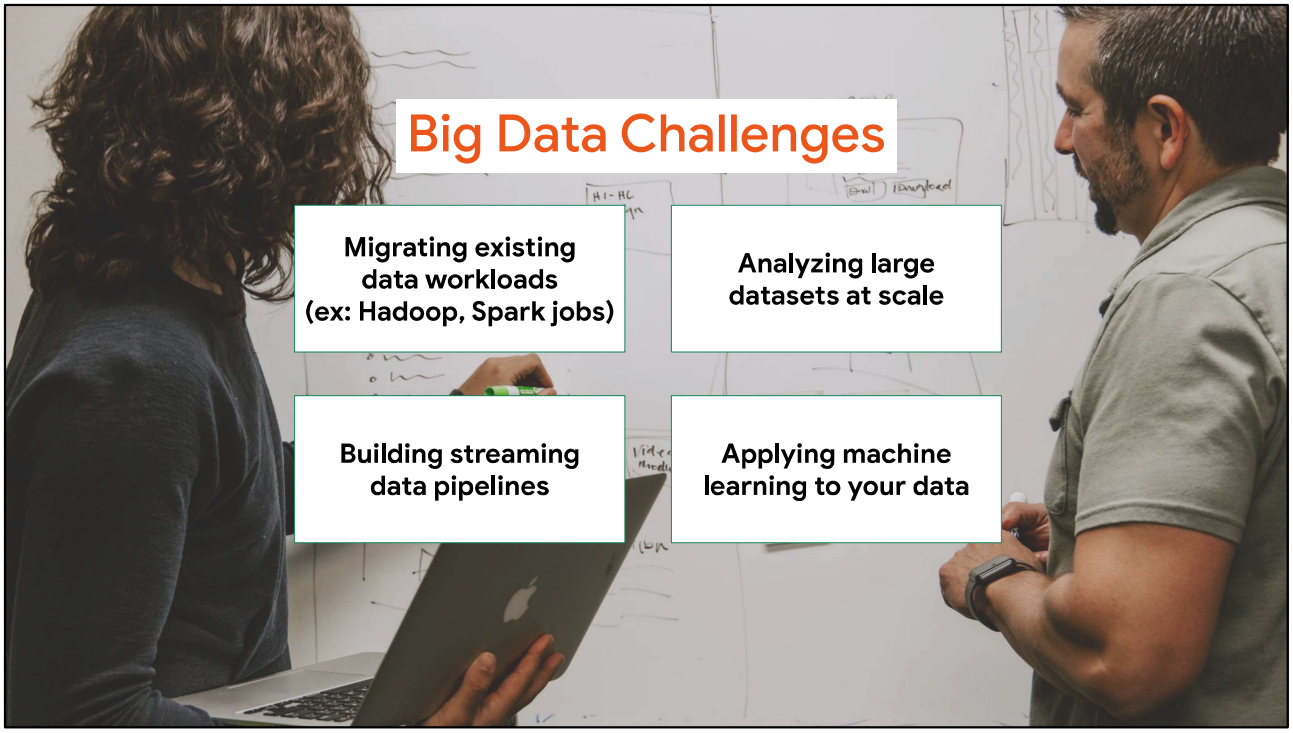
Activity: Explore a customer use case

The different data roles in an organization

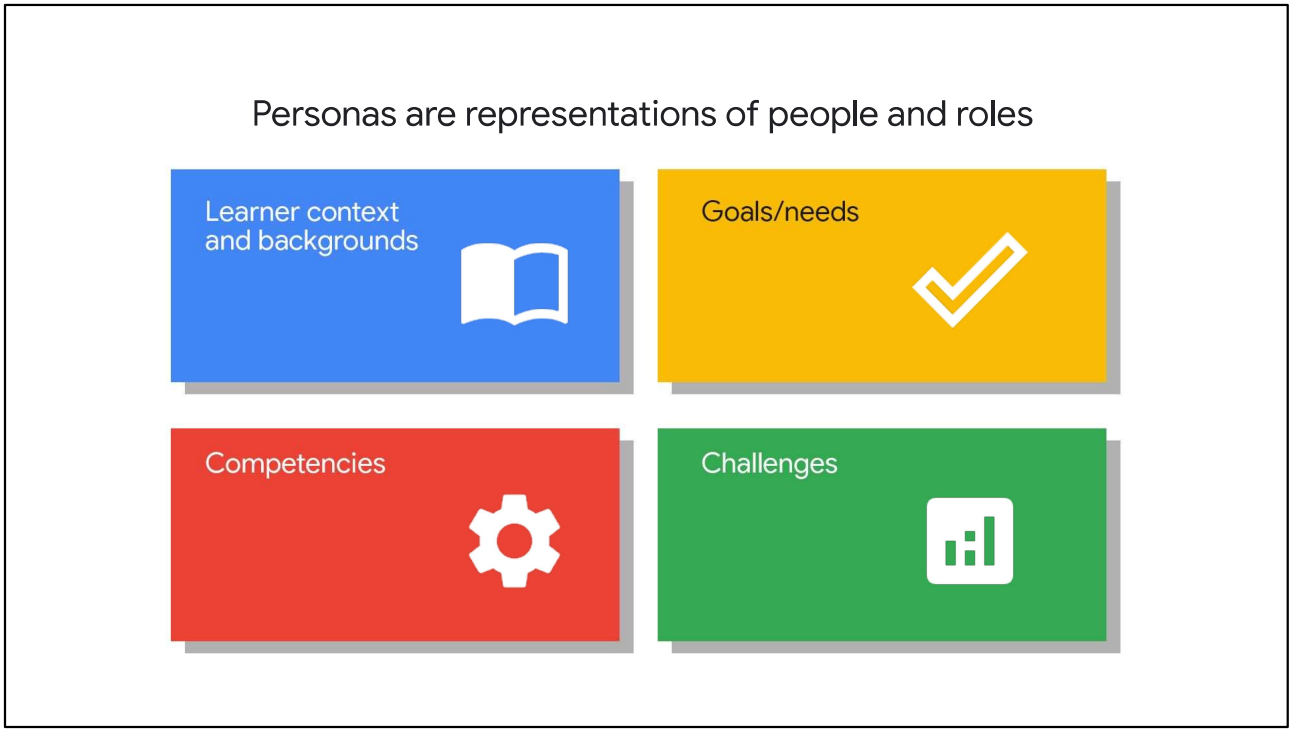
So far you've seen the infrastructure, the software, and the customers who are already using GCP. But the most critical factor to the success of your future big data and ML projects is your team itself. The people and the core skillsets required will make or break your next innovation.

As I have written in a blog post on the subject -- it's linked below -- a single person might have a combination of these roles, depending on the size of your organization. Your team size is one of the biggest drivers in whether you should hire for a specific skillset, upskill from within, or combine the two. [pause]

https://towardsdatascience.com/how-to-hire-a-machine-learning-team-b8055fff57f



Do you remember these big data challenges? Can you see how the roles would map to these?



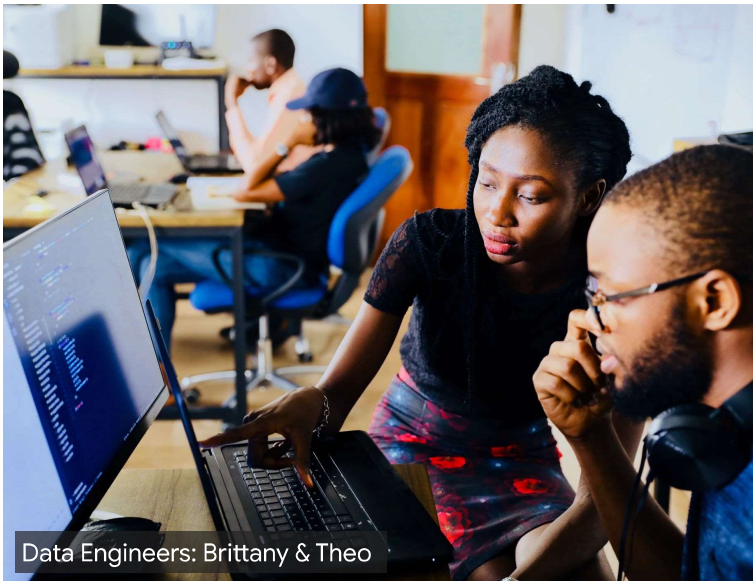
Within Google Cloud training, my team and I have thought about the different types of data science teams and roles that are using Google Cloud so that we can best tailor our data and ML courses and labs. One of the core challenges we face is how different types of users engage with our GCP big data and AI products.

We'll be using a few personas in this course, and their backgrounds, goals, and challenges might be similar to yours. Let's meet them now, and you'll see them again later.

https://medium.com/google-cloud/data-science-personas-cfda34a5d976

“Our CTO has challenged our data engineering team to find ways we can spend less on managing our on-prem cluster.

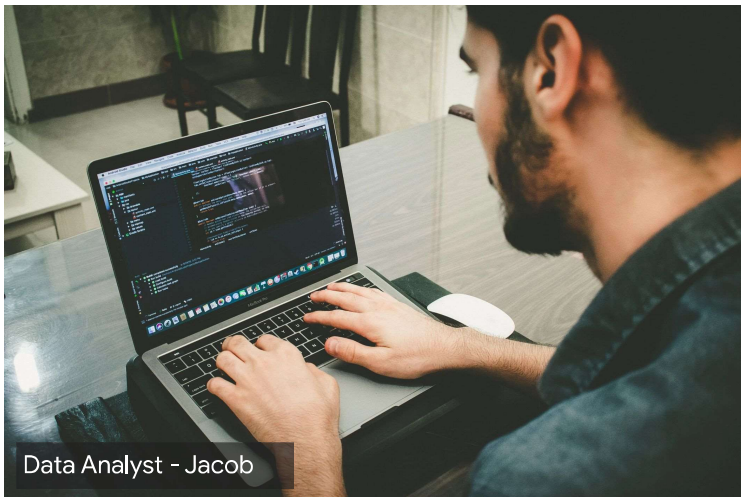
Right now, we just want to show her options that don’t require any code changes to our 100+ Hadoop jobs.”


Data Engineers: Brittany & Theo

Brittany and Theo lead their data engineering team in managing their Hadoop cluster for the organization’s data pipelines and compute jobs. Their organization was an early adopter of Hadoop for distributed computing back in 2007, and they actively ensure that the compute jobs are run and the cluster is well-maintained.

They say “Our CTO has challenged our data engineering team to find ways we can spend less on managing our on-prem cluster. Right now, we just want to show her options that don’t require any code changes to our 100+ Hadoop jobs.”

Brittany and Theo are data engineers who manage their company’s data platform.



“I’ve been asked to find a way to ingest and query my 5 TB of company data for fast insights... and I don’t have time to manage hardware.”

Jacob is a Data Analyst who has a background in building and querying his company’s mySQL transactional and reporting database. As the company grows, the reporting tables in his RDBMS are already starting to slow down, and users are reporting long query and dashboard loading times. He wants to find an easy path for scaling his company’s data reporting and not have to manage another data system as it grows.

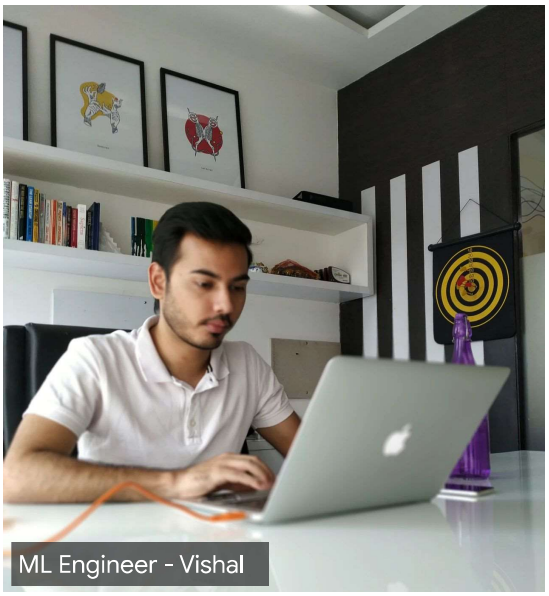
Jacob is a data analyst who wants to be able to derive insights from data and disseminate them with as little friction as possible.

“I really want to design our data pipelines for the future. For us that means lots and lots of streaming data from our IoT devices with low latency.”

Rebecca is a Data Engineer whose company specializes in harnessing data from *Internet of Things*, or IoT, devices.

She says “I really want to design our data pipelines for the future. For us that means lots and lots of streaming data from our IoT devices with low latency.”

Her team lead has asked her to come up with a plan to handle the expected 10X growth in streaming data volumes this year. She wants to future-proof her team’s pipelines but doesn’t want to spend hours manually scaling hardware up and down as streaming volume changes. Additionally, her business stakeholder team wants insights from all the IoT devices in the field on their dashboards with minimal delay.



“I pitched my team on the value ML can add, and I’ve got buy-in for a prototype.

What are some of the easiest ways I can see whether ML is feasible for my data?”

Vishal says “I pitched my team on the value ML can add, and I’ve got buy-in for a prototype. What are some of the easiest ways I can see whether ML is feasible for my data?”

Vishal is an Applied ML Engineer who has a background in building machine learning models in TensorFlow and keras. His team is growing rapidly, and he’s often asked by his leadership to assess the feasibility of ML for a wide variety of projects. He doesn’t have time to train and test all of the ideas with custom models, and he wants to empower his data analyst team by teaching them ML. [pause]

Do these personas sound familiar to your role and your team? Next, we’ll learn more about the Google Cloud Platform big data and machine learning approaches and solutions so that we can address each of these challenges.