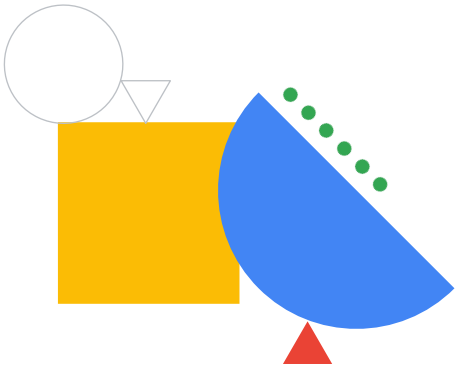
Google Cloud

Big Data and Machine Learning on Google Cloud



Module 1

Google Cloud Big Data and Machine Learning Fundamentals

Introduction

Welcome to the first module of the Big Data and Machine Learning Fundamentals course! This module lays the foundation to the next four modules,

- Data engineering
  - M.2: Data engineering for streaming data
  - M.3: Big data with BigQuery
- Machine learning
  - M.4: Machine learning options
  - M.5: The Machine Learning workflow with Vertex AI

Agenda

Google Cloud infrastructure

- Compute
- Storage

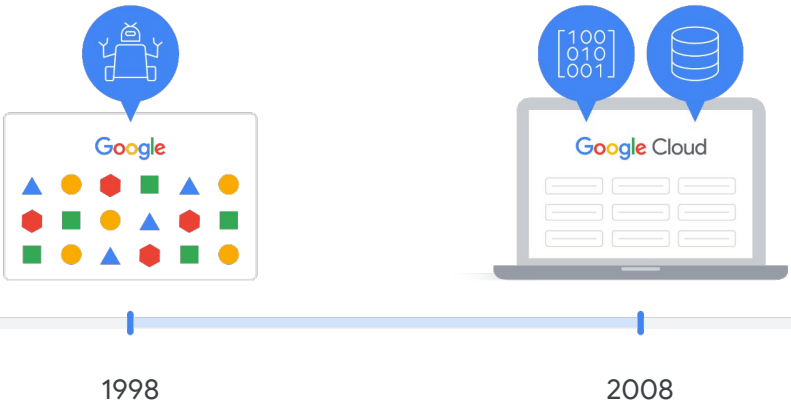
Data/ML products

- The history of big data and ML products
- Big data and ML product categories
- Customer example
- Hands-on lab

Here you'll explore the Google infrastructure through compute and storage, and see how innovation has enabled big data and machine learning capabilities.

- After that, you'll explore the history of big data and ML products, which will help you understand the relevant product categories.
- And to put it all together, you'll see an example of a customer who adopted Google Cloud for their big data and machine learning needs.
- Finally, you'll get hands-on practice using big data tools to analyze a public dataset.

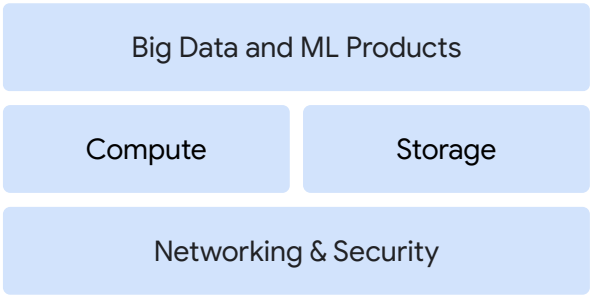
Google has years of experience with data and AI



Google has been working with data and artificial intelligence since its early days as a company in 1998.

Ten years later in 2008, the Google Cloud Platform was launched to provide secure and flexible cloud computing and storage services.

The Google Cloud infrastructure



You can think of the Google Cloud infrastructure in terms of three layers.

- At the base layer is **networking and security**, which lays the foundation to support all of Google's infrastructure and applications.
- On the next layer sit **compute** and **storage**. Google Cloud separates, or decouples, as it's technically called, compute and storage so they can scale independently based on need.
- And on the top layer sit the **big data and machine learning products**, which enable you to perform tasks to ingest, store, process, and deliver business insights, data pipelines, and ML models.

And thanks to Google Cloud, these tasks can be accomplished without needing to manage and scale the underlying infrastructure.

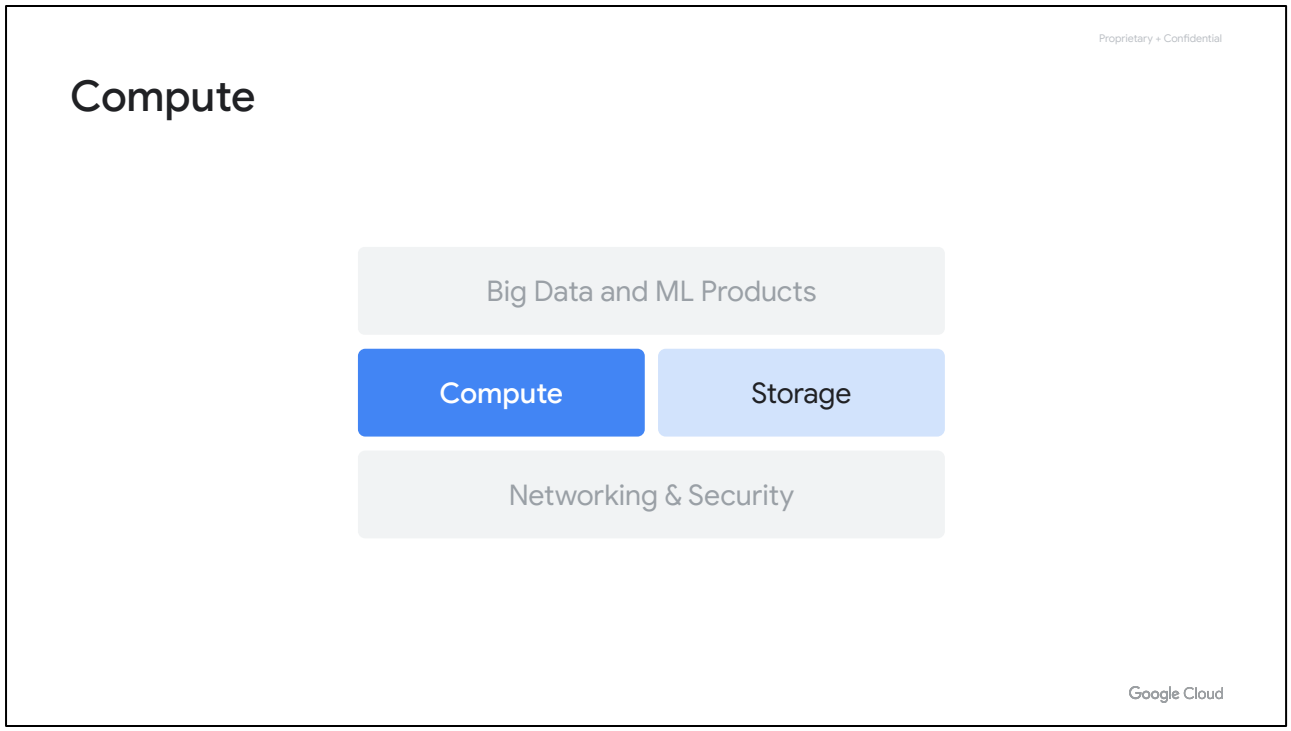
Focus of this course

cloud.google.com/training

This course focuses on the middle layer, compute and storage, and the top layer, big data and machine learning products.

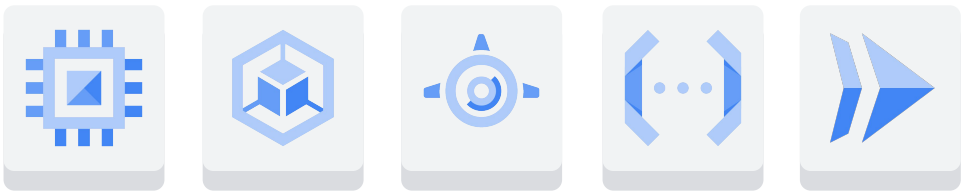
Networking and security fall outside of the focus of this course, but if you're interested in learning more you can explore cloud.google.com/training for more options.

Compute



Let's focus our attention on the middle layer of the Google Cloud infrastructure, **compute** and **storage**. We'll begin with **compute**.

Google Cloud computing services



Compute Engine | Google Kubernetes Engine | App Engine | Cloud Functions | Cloud Run

Organizations with growing data needs often require lots of compute power to run big data jobs. And as organizations design for the future, the need for compute power only grows.

Google offers a range of computing services, which includes: Compute Engine, Google Kubernetes Engine, App Engine, Cloud Functions, and Cloud Run.

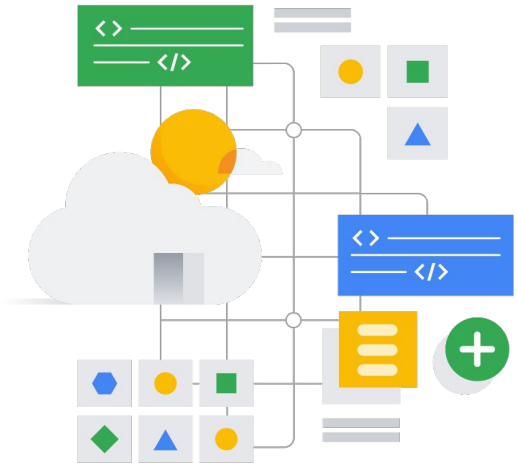
Compute Engine

- IaaS offering
- Compute
- Storage
- Network
- Maximum flexibility

Let's start with **Compute Engine**.

Compute Engine is an IaaS offering, or infrastructure as a service, which provides compute, storage, and network resources virtually that are similar to physical data centers. You use the virtual compute and storage resources the same as you manage them locally. Compute Engine provides maximum flexibility for those who prefer to manage server instances themselves.

Google Kubernetes Engine (GKE)



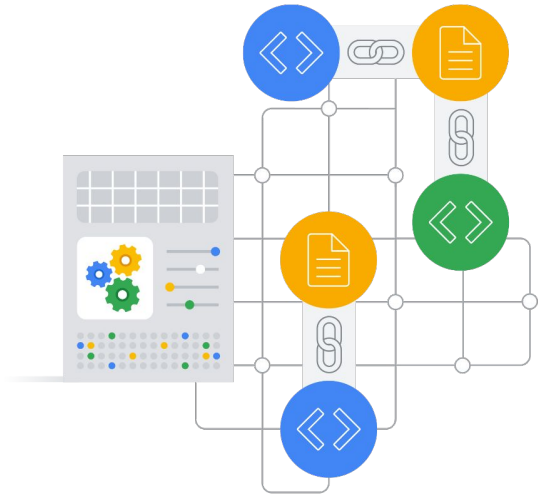
- Containerized applications
- Cloud environment

The second is **Google Kubernetes Engine**, or GKE.

GKE runs containerized applications in a cloud environment, as opposed to on an individual virtual machine, like Compute Engine. A container represents code packaged up with all its dependencies.

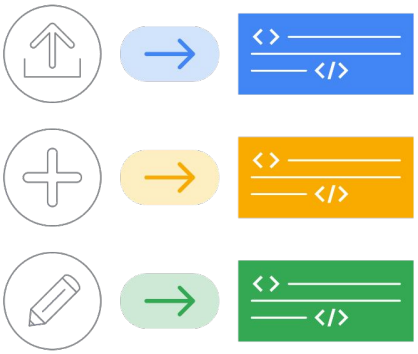
App Engine

- Fully managed PaaS offering
- Bind code to libraries
- Focused on application logic



The third computing service offered by Google is **App Engine**, a fully managed PaaS offering, or platform as a service. PaaS offerings bind code to libraries that provide access to the infrastructure application needs. This allows more resources to be focused on application logic.

Cloud Functions



- Executes code in response to events
- Functions as a service offering

The fourth is Cloud Functions, which executes code in response to events, like when a new file is uploaded to Cloud Storage. It's a completely serverless execution environment, which means you don't need to install any software locally to run the code and you are free from provisioning and managing servers. Cloud Functions is often referred to as functions as a service.

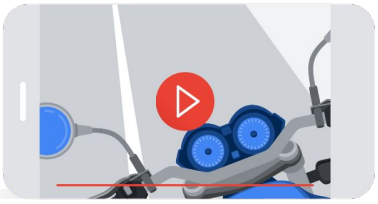
Cloud Run

- Fully managed platform
- Lets you focus on writing code
- Automatically scales up and down
- Charges you only for the resources you use

And finally there is **Cloud Run**, a fully managed compute platform that enables you to run request or event-driven stateless workloads without having to worry about servers. It abstracts away all infrastructure management so you can focus on writing code. It automatically scales up and down from zero, so you never have to worry about scale configuration.

Cloud Run charges you only for the resources you use so you never pay for over provisioned resources.

Example technology



Automatic
video stabilization

Stabilizes an unstable video to minimize movement

Let's understand how Google Photos relies on the compute capability provided by Google Cloud to implement their video stabilization.

Google Photos offers a feature called automatic video stabilization. This takes an unstable video, like one captured while riding on the back of a motorbike, and stabilizes it to minimize movement.

Let's look at an example of a technology that requires a lot of compute power.

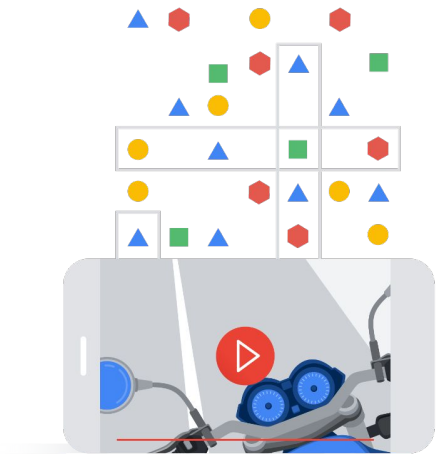
*[The one on the left has **Optical Image Stabilization** (OIS) and **Electronic Image Stabilization** (EIS) turned off, while the one on the right has both of those features turned on.]*

Youtube clip

Features require proper data

Automatic video stabilization

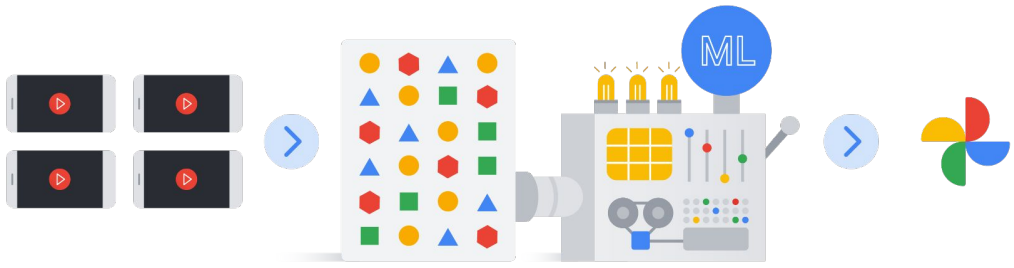
- Video itself
- Time series data on the camera's position
- Orientation from the onboard gyroscope
- Motion from the camera lens



For video stability to work as intended, you need the proper data. This includes the video itself, which is really a large collection of individual images, along with time series data on the camera's position and orientation from the onboard gyroscope, and motion from the camera lens.

A short video can require over a billion data points to feed the ML model to create a stabilized version. As of 2020, roughly 28 billion photos and videos were uploaded to Google Photos every week, with more than four trillion photos in total stored in the service.

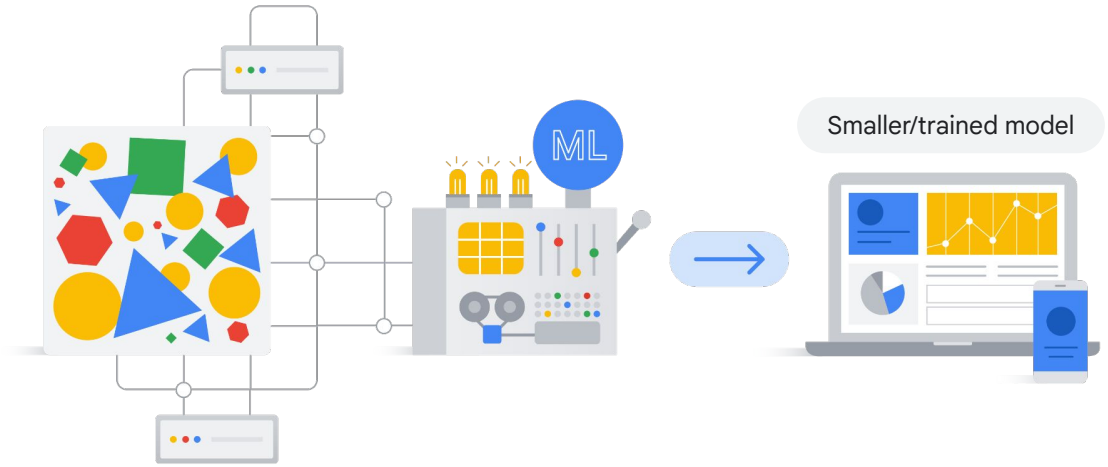
Features require proper training



Develop Train Serve

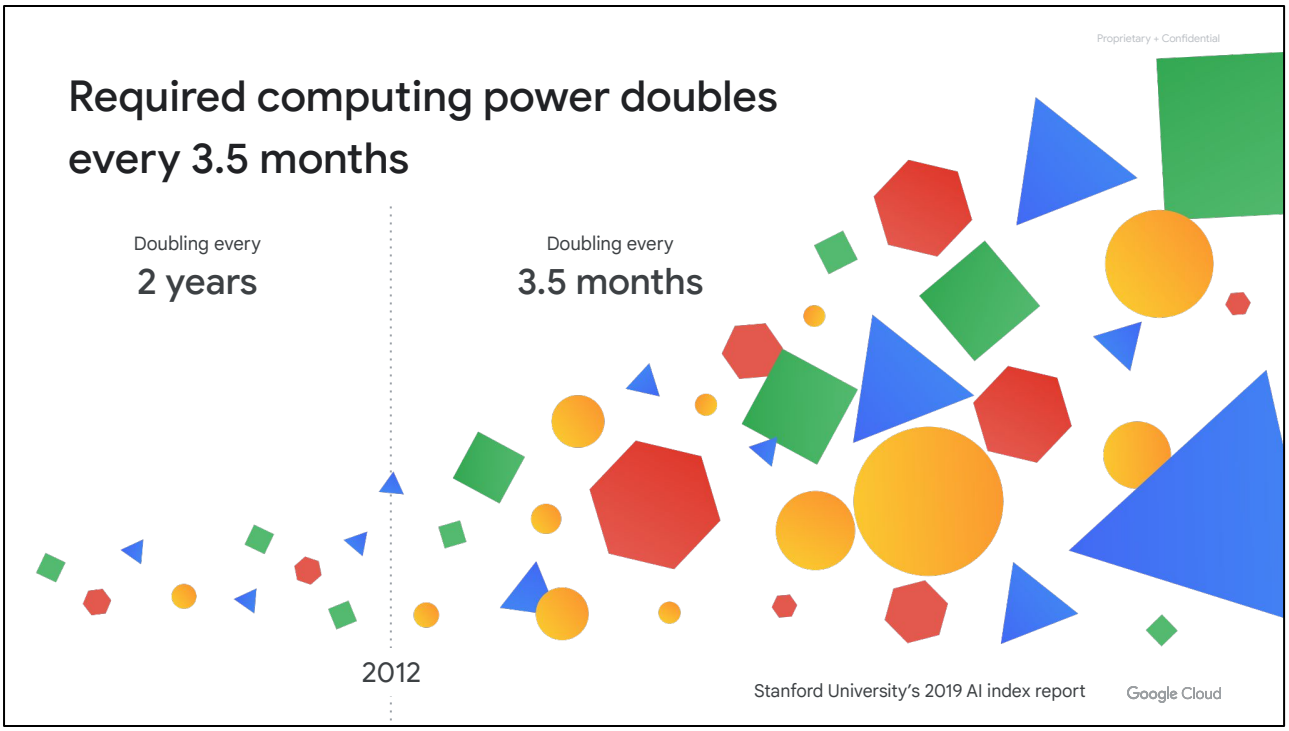
To ensure that this feature works as intended, and accurately, the Google Photos team needed to develop, train, and serve a high-performing machine learning model on millions of videos. That's a large training dataset!

ML models train on a vast network of data centers



Just as the hardware on a standard personal computer might not be powerful enough to process a big data job for an organization, the hardware on a smartphone is not powerful enough to train sophisticated ML models.

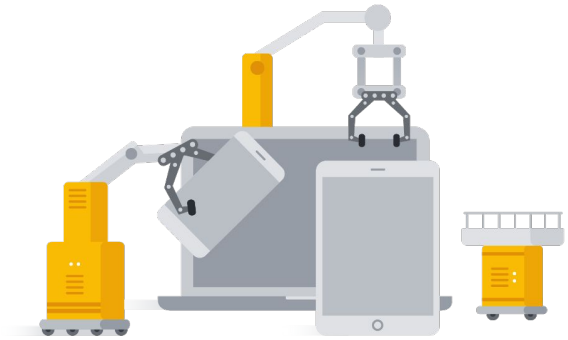
That's why Google trains production machine learning models on a vast network of data centers, only to then deploy smaller, trained versions of the models to the smartphone and personal computer hardware.



But where does all that processing power come from?

According to Stanford University's 2019 AI index report, before 2012, artificial intelligence results tracked closely with Moore's Law, with the required computing power used in the largest AI training runs doubling every two years. The report states that, since 2012, the required computing power has been doubling approximately every three and a half months.

Hardware limitations



CPU

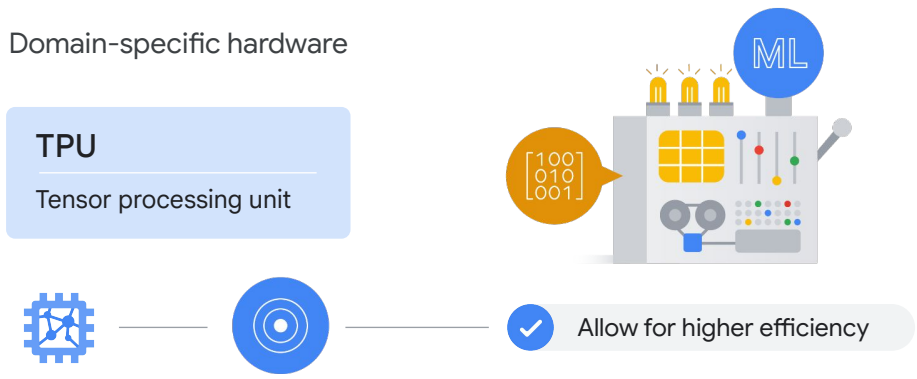
Central processing unit

GPU

Graphics processing unit

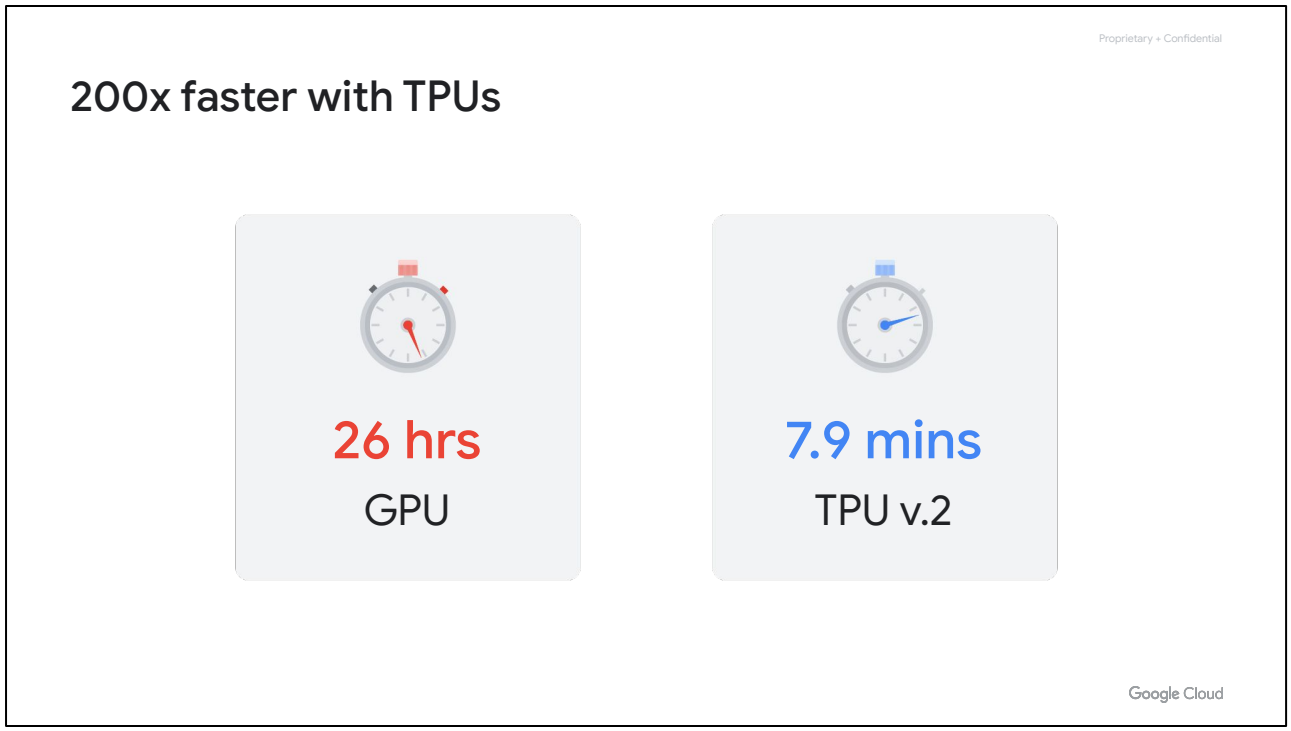
This means that hardware manufacturers have run up against limitations, and CPUs, which are central processing units, and GPUs, which are graphics processing units, can no longer scale to adequately reach the rapid demand for ML.

Tensor Processing Unit (TPU)



To help overcome this challenge, in 2016 Google introduced the **Tensor Processing Unit**, or **TPU**. TPUs are Google's custom-developed **application-specific** integrated circuits (ASICs) used to accelerate machine learning workloads.

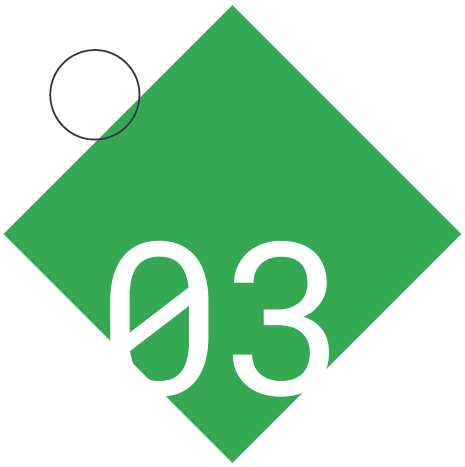
TPUs act as **domain-specific** hardware, as opposed to **general-purpose** hardware with CPUs and GPUs. This allows for higher efficiency by tailoring architecture to meet the computation needs in a domain, such as the matrix multiplication in machine learning.



With TPUs, the computing speed increases more than 200 times.

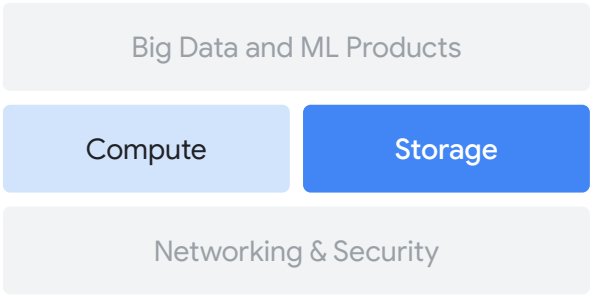
This means that instead of waiting 26 hours for results with a single state-of-art GPU, you'll only need to wait for 7.9 minutes for a full Cloud TPU v.2 pod to deliver the same results.

Cloud TPUs have been integrated across Google products, and this state-of-the-art hardware and supercomputing technology is available with Google Cloud products and services.



Storage

Storage

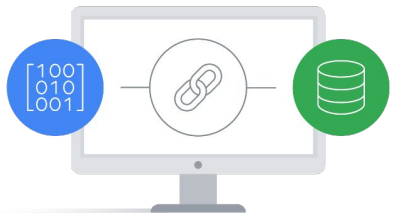


Now that we've explored compute and why it's needed for big data and ML jobs, let's now examine storage.

Compute and storage are decoupled

Decoupled

Cloud computing



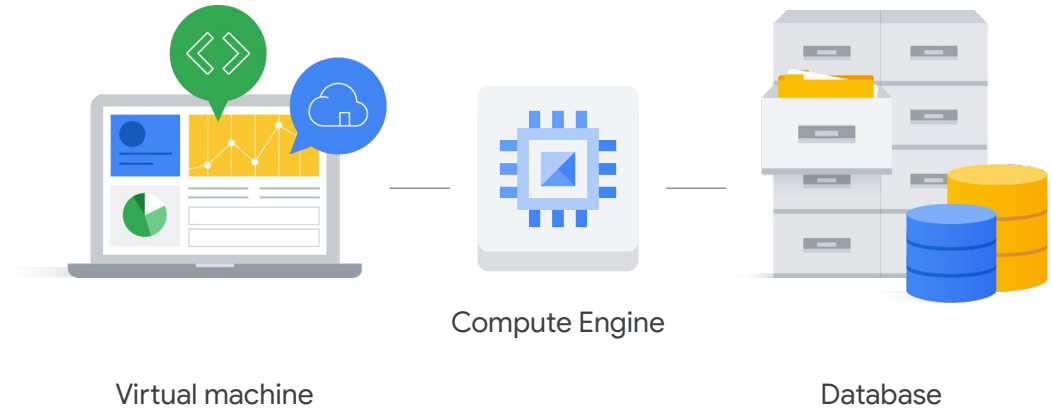
Coupled

Desktop computing

For proper scaling capabilities, compute and storage are decoupled. This is one of the major differences between cloud computing and desktop computing.

With cloud computing, processing limitations aren't attached to storage disks.

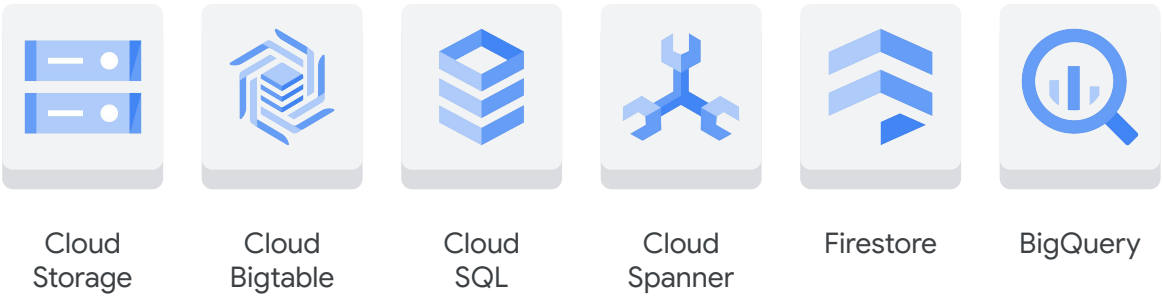
Applications require a database and storage solution



Most applications require a database and storage solution of some kind.

With Compute Engine, for example, which was mentioned previously, you can install and run a database on a virtual machine, just as you would do in a data center.

Google Cloud database and storage services



These include:

- Cloud Storage
- Cloud Bigtable
- Cloud SQL
- Cloud Spanner, and
- Firestore
- BigQuery

The goal of these products is to reduce the time and effort needed to store data. This means creating an elastic storage bucket directly in a web interface or through a command line, for example on Google Cloud Storage.

Storage offerings

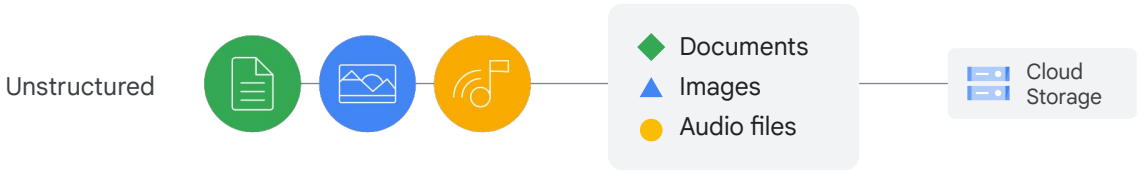
Google Cloud

- Relational databases
- Non-relational databases
- Worldwide object storage

Google Cloud offers relational and non-relational databases, and worldwide object storage.

Choosing the right option to store and process data often depends on the data type that needs to be stored and the business need.

Unstructured data → Cloud Storage



Let's start with unstructured versus structured data.

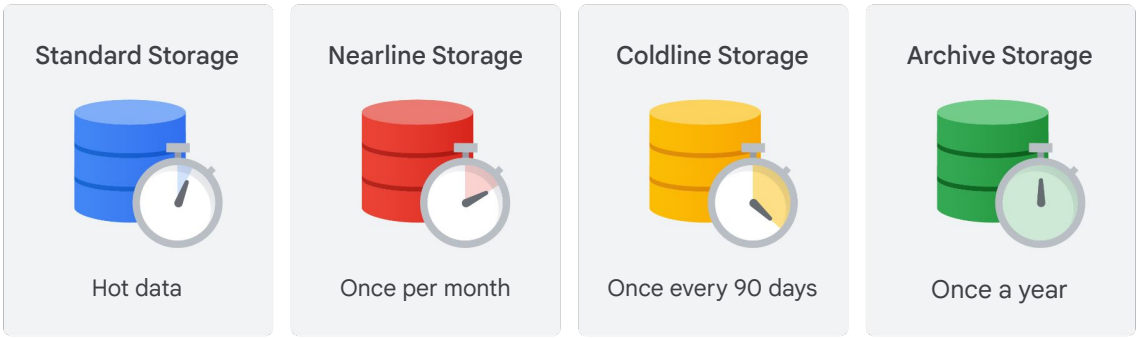
Unstructured data is information stored in a non-tabular form such as documents, images, and audio files. Unstructured data is usually suited to **Cloud Storage**, but **BigQuery** now offers the capability to store unstructured data as well.

Cloud Storage is a fully managed scalable service

Cloud Storage is a managed service for storing unstructured data. Cloud Storage is a service for storing your objects in Google Cloud. An object is an immutable piece of data consisting of a file of any format. You store objects in containers called buckets. All buckets are associated with a project, and you can group your projects under an organization. Each project, bucket, and object in Google Cloud is a resource in Google Cloud, as are things such as Compute Engine instances. After you create a project, you can create Cloud Storage buckets, upload objects to your buckets, and download objects from your buckets.

A few examples include serving website content, storing data for archival and disaster recovery, and distributing large data objects to end users via Direct Download.

Cloud Storage primary storage classes

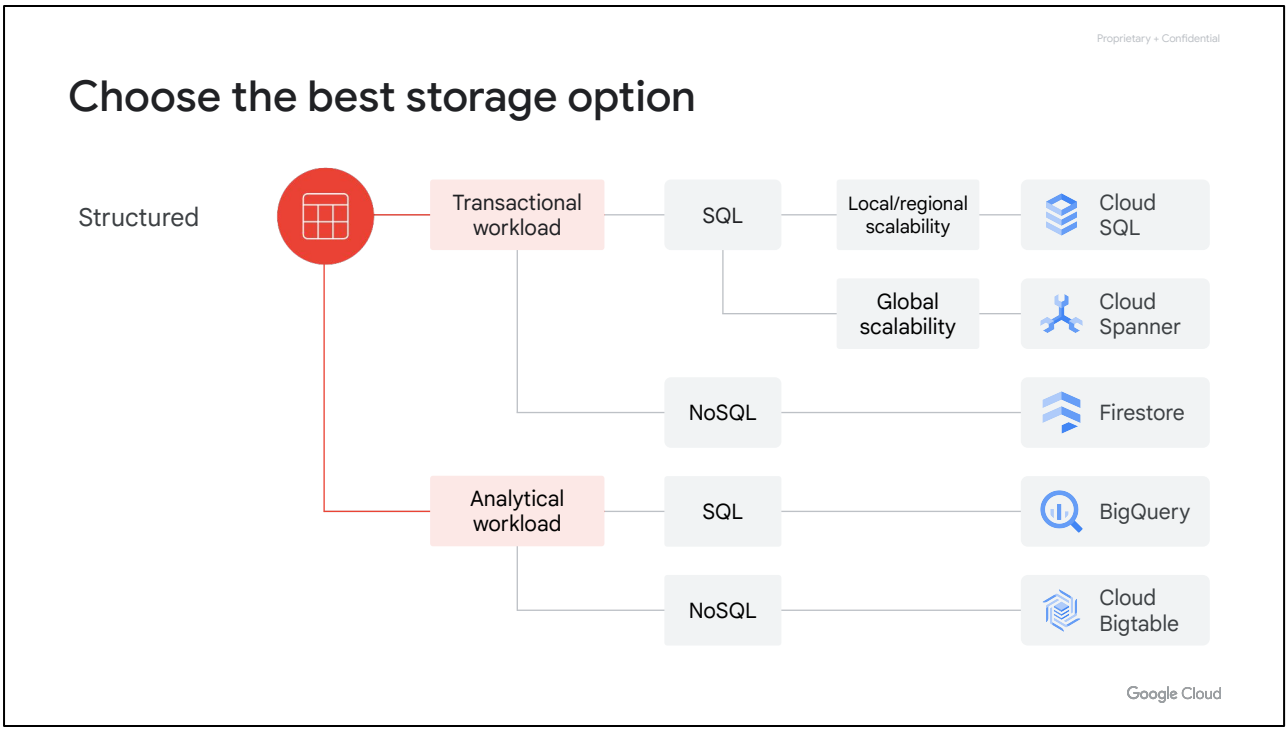


Cloud Storage has four primary storage classes.

- The first is **Standard Storage**. Standard Storage is considered best for frequently accessed, or "hot," data. It's also great for data that is stored for only brief periods of time.
- The second storage class is **Nearline Storage.** This is best for storing infrequently accessed data, like reading or modifying data once per month or less, on average. Examples include data backups, long-tail multimedia content, or data archiving.
- The third storage class is **Coldline Storage**. This is also a low-cost option for storing infrequently accessed data. However, as compared to Nearline Storage, Coldline Storage is meant for reading or modifying data, at most, once every 90 days.
- The fourth storage class is **Archive Storage**. This is the lowest-cost option, used ideally for data archiving, online backup, and disaster recovery. It's the best choice for data that you plan to access less than once a year, because it has higher costs for data access and operations and a 365-day minimum storage duration.

Structured data

Alternatively, there is **structured data**, which represents information stored in tables, rows, and columns.



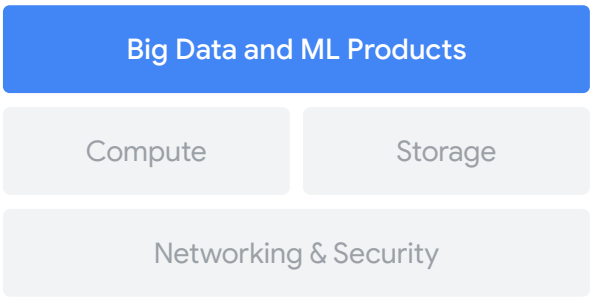
Structured data comes in two types: **transactional** workloads and **analytical** workloads.

- **Transactional workloads** stem from Online Transaction Processing systems, which are used when fast data inserts and updates are required to build row-based records. This is usually to maintain a system snapshot. They require relatively standardized queries that impact only a few records.
  - So, if your data is transactional and you need to access it using **SQL**, then Cloud SQL and Cloud Spanner are two options.
    - **Cloud SQL** works best for **local to regional scalability**,
    - while **Cloud Spanner**, it best to scale a database **globally**.
  - If the transactional data will be accessed **without SQL**,
    - **Firestore** might be the best option. Firestore is a transactional No-SQL, document-oriented database.

- Then there are **analytical workloads**, which stem from Online Analytical Processing systems, which are used when entire datasets need to be read. They often require complex queries, for example, aggregations.
  - If you have analytical workloads that require **SQL** commands, **BigQuery** is likely the best option. BigQuery, Google's data warehouse solution, lets you analyze petabyte-scale datasets.
  - Alternatively, **Cloud Bigtable** provides a scalable **NoSQL** solution for analytical workloads. It's best for real-time, high-throughput applications that require only millisecond latency.

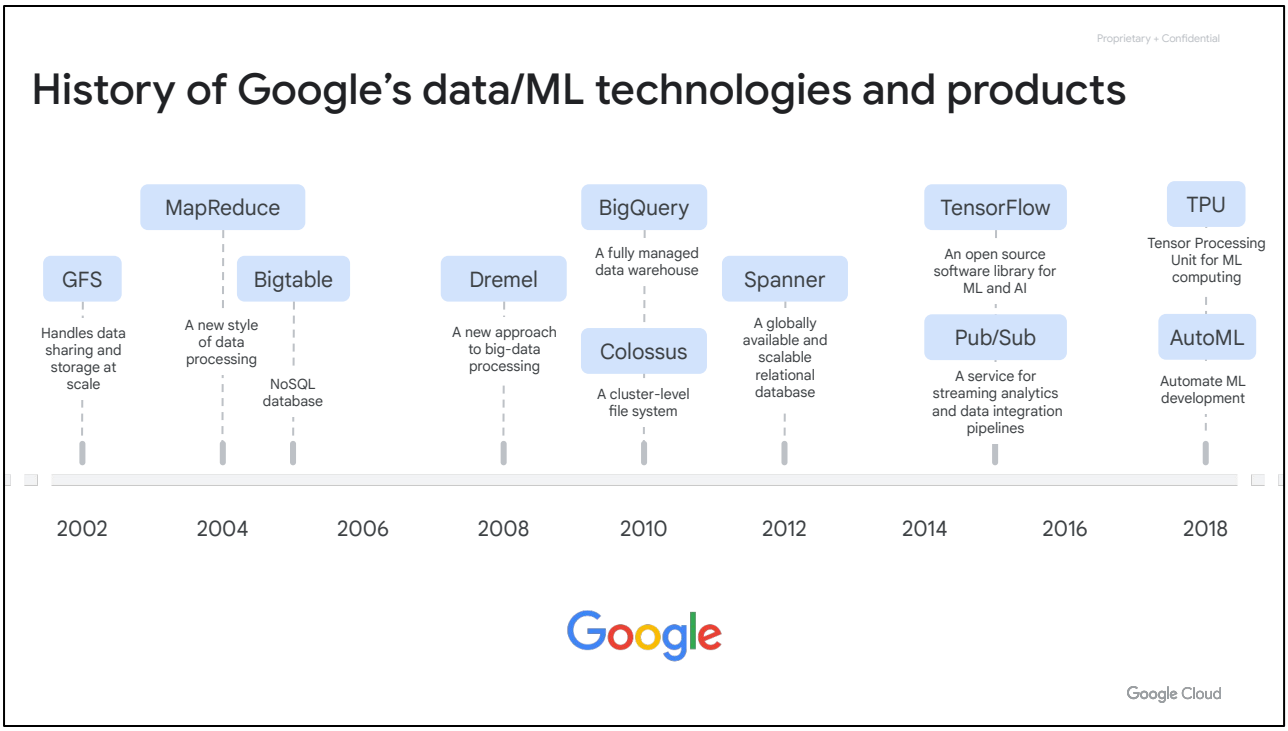
The history of big data and ML products

Big data and machine learning products



The final layer of the Google Cloud infrastructure that is left to explore is big data and machine learning products.

We'll examine the evolution of data processing frameworks through the lens of product development. Understanding the chronology of products can help address typical big data and machine learning challenges.



Historically speaking, Google experienced challenges related to big data quite early–mostly with large datasets, fast-changing data, and varied data. This was the result of needing to index the World Wide Web.

And as the internet grew, Google needed to invent new data processing methods.

- So, in 2002, Google released the **Google File System**, or GFS. GFS was designed to handle data sharing and petabyte storage at scale. It served as the foundation for Cloud Storage and also what would become the managed storage functionality in BigQuery.

- A challenge that Google was facing around this time was how to index the exploding volume of content on the web. To solve this, in 2004 Google wrote a report that introduced **MapReduce**. MapReduce was a new style of data processing designed to manage large-scale data processing across big clusters of commodity servers.

- As Google continued to grow, new challenges arose, specifically with recording and retrieving millions of streaming user actions with high throughput. The solution was the release in 2005 of **Cloud Bigtable**, a high-performance NoSQL database service for large analytical and operational workloads.

With MapReduce available, some developers were restricted by the need to write code to manage their infrastructure, which prevented them from focusing on application

logic.
As a result, from 2008 to 2010, Google started to move away from MapReduce as the solution to process and query large datasets.

- So, in 2008, **Dremel** was introduced. Dremel took a new approach to big-data processing by breaking the data into smaller chunks called shards, and then compressing them.

Dremel then uses a query optimizer to share tasks between the many shards of data and the Google data centers, which processed queries and delivered results. The big innovation was that Dremel autoscaled to meet query demands.

Dremel became the query engine behind BigQuery.

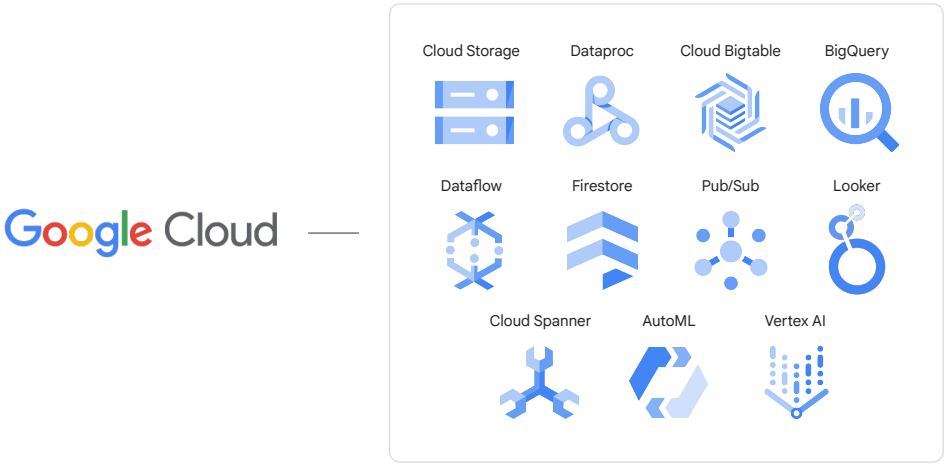
Google continued innovating to solve big data and machine learning challenges. Some of the technology solutions released include:

- **Colossus**, in 2010, which is a cluster-level file system and successor to the Google File System.
- **BigQuery**, in 2010 as well, which is a fully-managed, serverless data warehouse that enables scalable analysis over petabytes of data. It is a Platform as a Service (PaaS) that supports querying using ANSI SQL. It also has built-in machine learning capabilities. BigQuery was announced in May 2010 and made generally available in November 2011.
- **Spanner**, in 2012, which is a globally available and scalable relational database.
- **Pub/Sub**, in 2015, which is a service used for streaming analytics and data integration pipelines to ingest and distribute data.

And **TensorFlow**, also in 2015, which is a free and open source software library for machine learning and artificial intelligence.

- 2018 brought the release of the Tensor Processing Unit, or **TPU**, which you'll recall from earlier, and
- **AutoML**, as a suite of machine learning products.
- The list goes on till **Vertex AI**, a unified ML platform released in 2021.

A robust big data and ML product line



And it's thanks to these technologies that the big data and machine learning product line is now robust.

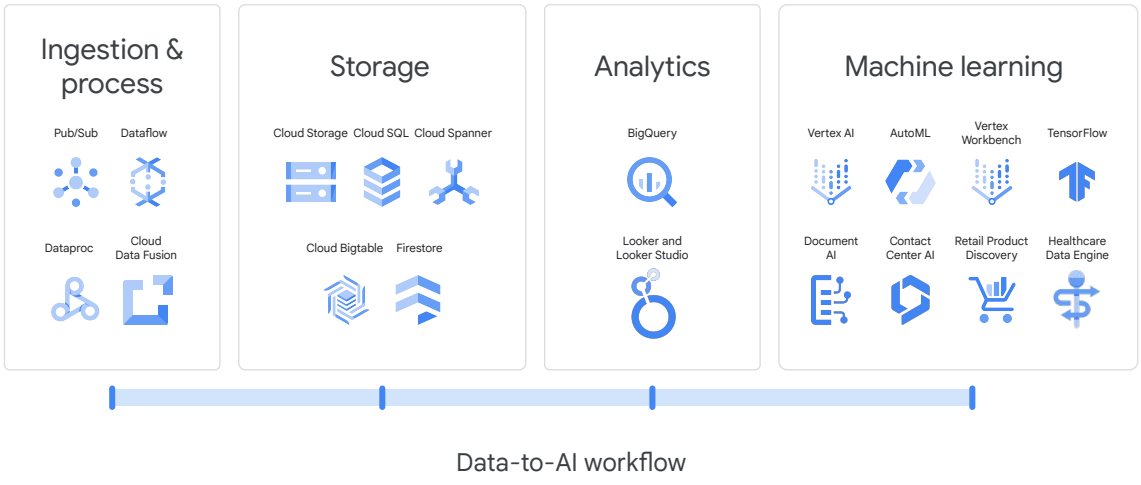
This includes:

- Cloud Storage
- Dataproc
- Cloud Bigtable
- BigQuery
- Dataflow
- Firestore
- Pub/Sub
- Looker
- Cloud Spanner
- AutoML, and
- Vertex AI, the unified platform

These products and services are made available through Google Cloud, and you'll get hands-on practice with some of them as part of this course.

Big data and ML product categories

Big data and ML product categories

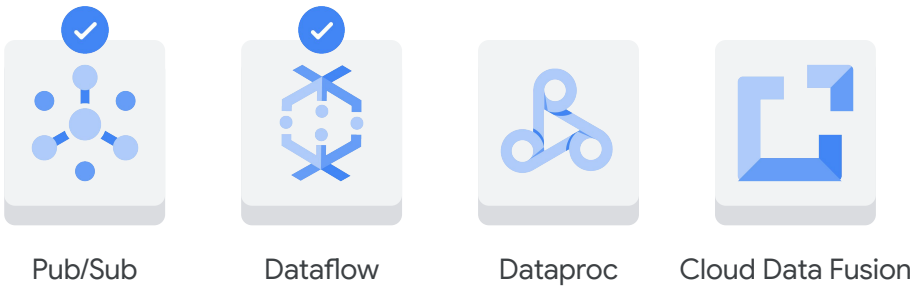


As we explored previously, Google offers a range of big data and machine learning products. So, how do you know which is best for your business needs?

Let's look closer at the list of products, which can be divided into four general categories along the data-to-AI workflow: **ingestion and process, storage**, **analytics**, and **machine learning**.

Understanding these product categories can help narrow down your choice.

Ingestion and process



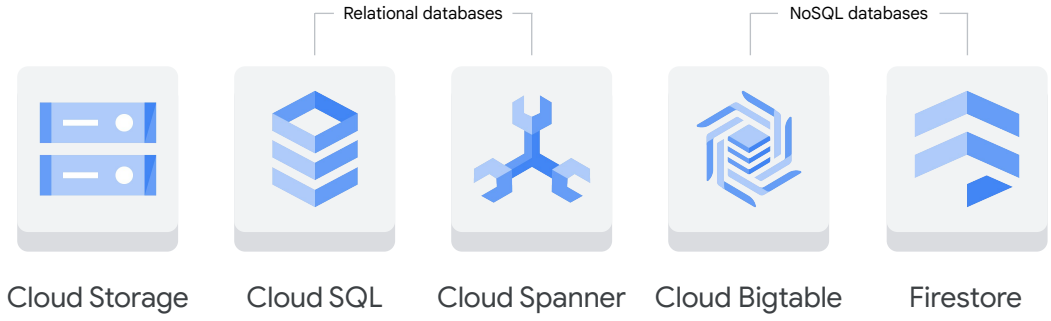
To be explored more later in this course

The first category is **ingestion and process**, which include products that are used to digest both real-time and batch data. The list includes:

- Pub/Sub
- Dataflow
- Dataproc
- Cloud Data Fusion

You'll explore how Dataflow and Pub/Sub can ingest streaming data later in this course.

Storage

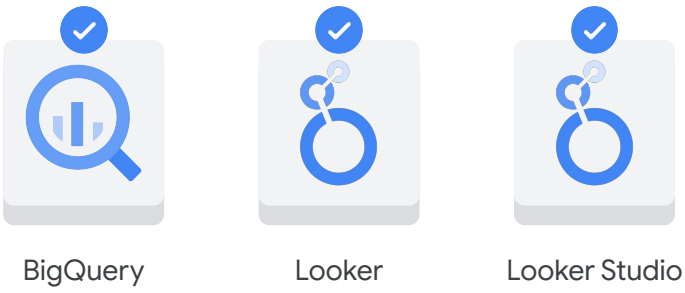


The second product category is data **storage**, and you'll recall from earlier that there are five storage products:

- Cloud Storage
- Cloud SQL
- Cloud Spanner
- Cloud Bigtable, and
- Firestore

Cloud SQL and Cloud Spanner are relational databases, while Bigtable and Firestore are NoSQL databases.

Analytics



To be explored more later in this course.

The third product category is **analytics**. The major analytics tool is BigQuery. BigQuery is a fully managed data warehouse that can be used to analyze data through SQL commands.

In addition to BigQuery, you can analyze data and visualize results using:

- Looker, and
- Looker Studio

You'll explore BigQuery, Looker, and Looker Studio in this course.

Machine learning | ML development platforms

To be explored more later in this course.

And the final product category is **machine learning, or ML**. ML products include both the ML development platform and the AI solutions:

The primary product of the ML development platform is Vertex AI, which includes:

- AutoML,
- Vertex AI Workbench, and
- TensorFlow

You'll explore Vertex AI and AutoML in this course.

Machine learning | AI solutions

Document AI

Contact Center AI

Retail Product Discovery

Healthcare Data Engine

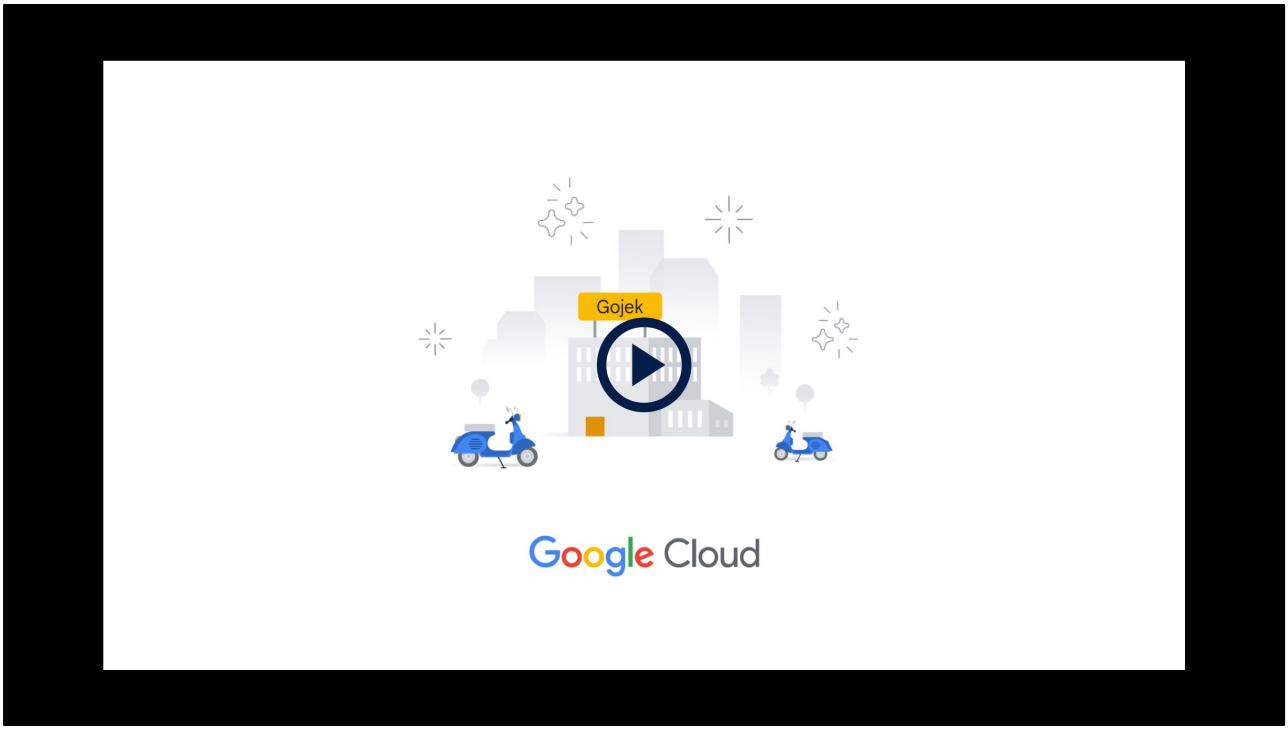
AI solutions are built on the ML development platform and include state-of-the-art products to meet both horizontal and vertical market needs. These include:

- Document AI
- Contact Center AI
- Retail Product Discovery, and
- Healthcare Data Engine

These products unlock insights that only large amounts of data can provide. We'll explore the machine learning options and workflow together with these products in greater detail later.

Customer example: Gojek

With many big data and machine learning products options available, it can be helpful to see an example of how an organization has leveraged Google Cloud to meet their goals.



Trainer note: During this section, play the Gojek video to the learners. Please reach out to yoannalong@google.com for access if you don't have the appropriate permission to do so. The video is hyperlinked to the slide image and will open in a new window.

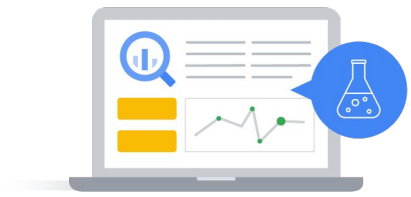
If you run into technical issues, use the following 8 slides to facilitate this section. They are set to skip by default.

SAY: In this section, you'll learn about a company called **Gojek** and how they were able to find success through Google Cloud's data engineering and machine learning offerings.

Lab: Exploring a BigQuery public dataset

Now it's time for some hands-on practice with one of the big data and machine learning products that was introduced in earlier–BigQuery.

Hands-on practice with BigQuery



1. Querying a public data set
2. Creating a custom table
3. Loading data into a table
4. Querying a table

In this lab, you'll use BigQuery to explore a public dataset.

You'll practice:

- Querying a public data set
- Creating a custom table
- Loading data into a table, and
- Querying a table

Summary

Summary

- Google Cloud infrastructure
- Big data and ML product history and categories

This brings us to the end of the first module of the Big Data and Machine Learning Fundamentals course. Before we move forward, let's review what we've covered so far.

You began by exploring the Google Cloud infrastructure through three different layers.

The Google Cloud infrastructure

At the base layer is **networking and security**, which makes up the foundation to support all of Google's infrastructure and applications.

On the next layer sit **compute** and **storage**. Google Cloud decouples compute and storage so they can scale independently based on need.

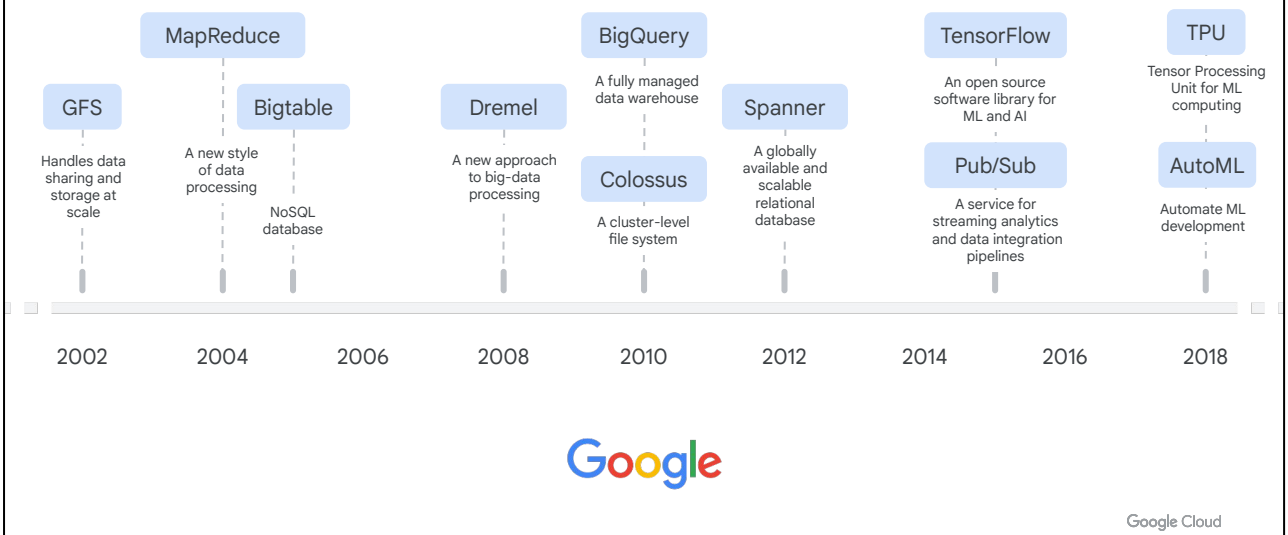
And on the top layer sit the **big data and machine learning products**.

Summary

- Google Cloud infrastructure
- Big data and ML product history and categories

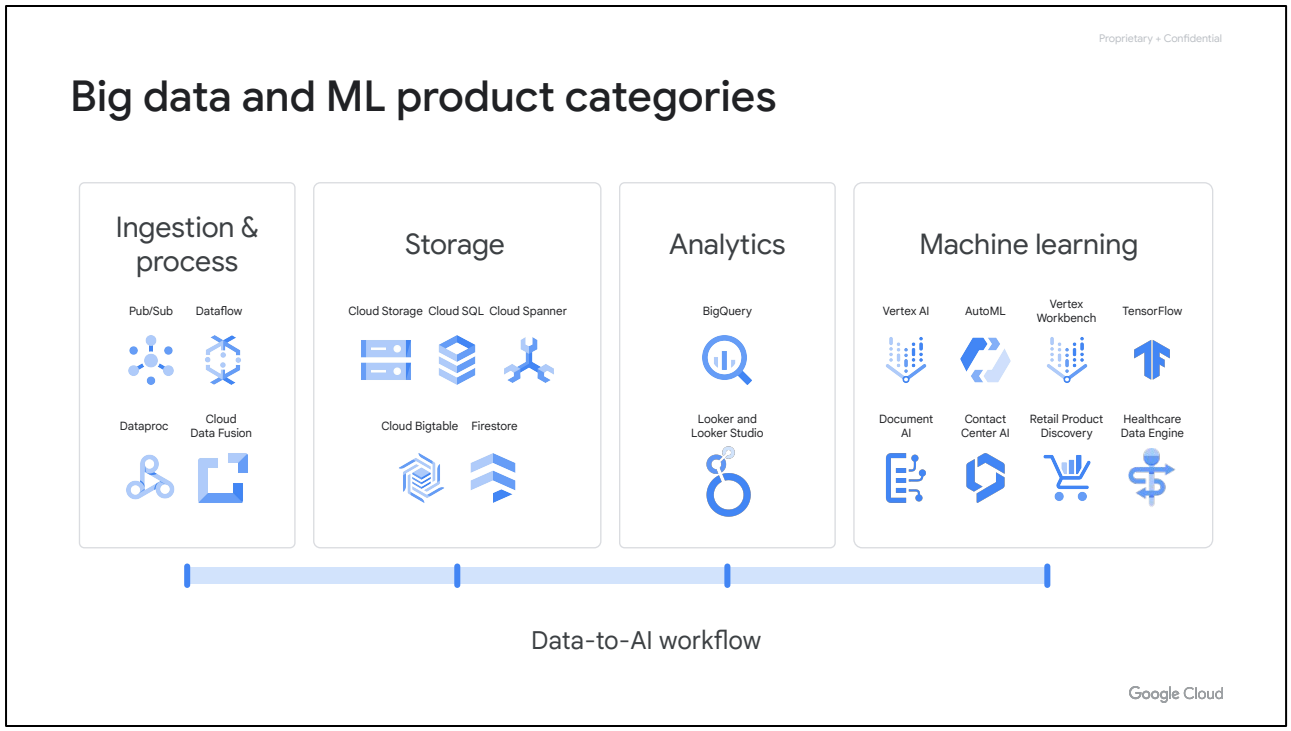
In the next section, you learned about the history of big data and ML technologies,

History of Google's data/ML technologies and products

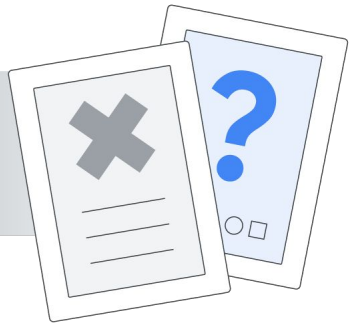


Google

Google has been working with data and artificial intelligence since its early days as a company, starting from GFS (Google File system), to BigQuery (Google's fully-managed data warehouse), and to TensorFlow (open source ML library), TPU (Tensor Processing Unit), and recently Vertex AI (a unified ML platform).



And finally explored the four major product categories: **Ingestion and process, storage**, **analytics**, and **machine learning**.



Quiz

Question #1

Question

Which Google hardware innovation tailors architecture to meet the computation needs on a domain?

A. CPUs (central processing units)

B. GPUs (graphic processing units)

C. TPUs (Tensor Processing Units)

D. DPUs (data processing units)

Question #1

Answer

Which Google hardware innovation tailors architecture to meet the computation needs on a domain?

A. CPUs (central processing units)

B. GPUs (graphic processing units)

C. TPUs (Tensor Processing Units)

D. DPUs (data processing units)

Question #2

Question

Which data storage class is best for storing data that needs to be accessed less than once a year?

A. Standard storage

B. Nearline storage

C. Coldline storage

D. Archive storage

Question #2

Answer

Which data storage class is best for storing data that needs to be accessed less than once a year?

A. Standard storage

B. Nearline storage

C. Coldline storage

D. Archive storage 

Question #3

Question

Compute Engine, Google Kubernetes Engine, App Engine, and Cloud Functions represent which type of services?

A. Database and storage

B. Compute

C. Networking

D. Machine learning

Question #3

Answer

Compute Engine, Google Kubernetes Engine, App Engine, and Cloud Functions represent which type of services?

A. Database and storage

B. Compute 

C. Networking

D. Machine learning

Question #4

Question

Cloud Storage, Cloud Bigtable, Cloud SQL, Cloud Spanner, and Firestore represent which type of services?

A. Machine learning

B. Networking

C. Database and storage

D. Compute

Question #4

Answer

Cloud Storage, Cloud Bigtable, Cloud SQL, Cloud Spanner, and Firestore represent which type of services?

A. Machine learning

B. Networking

C. Database and storage 

D. Compute

Question #5

Question

Pub/Sub, Dataflow, Dataproc, and Cloud Data Fusion align to which stage of the data-to-AI workflow?

A. Ingestion and process
B. Storage
C. Analytics
D. Machine learning

Question #5

Answer

Pub/Sub, Dataflow, Dataproc, and Cloud Data Fusion align to which stage of the data-to-AI workflow?

A. Ingestion and process 

B. Storage

C. Analytics

D. Machine learning

Question #6

Question

AutoML, Vertex AI Workbench, and TensorFlow align to which stage of the data-to-AI workflow?

A. Ingestion and process

B. Storage

C. Analytics

D. Machine learning

Question #6

Answer

AutoML, Vertex AI Workbench, and TensorFlow align to which stage of the data-to-AI workflow?

A. Ingestion and process
B. Storage
C. Analytics
D. Machine learning