



Introducing Google Cloud



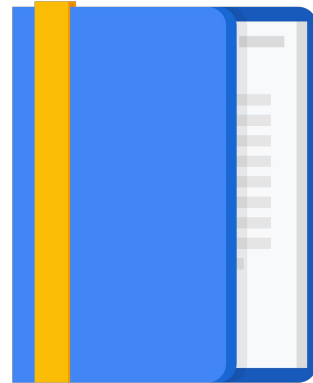
This course is designed to take your existing AWS knowledge and bring it into Google Cloud. You may find that throughout this course some topics seem very familiar to you. This is great; this shows that you have a clear understanding of how the different elements of a cloud platform work. Our goal is to cover the basics of cloud computing and make sure that everyone has a solid foundation in Google Cloud to build upon.

Agenda

Introduction to Google Cloud

Quiz

Resources

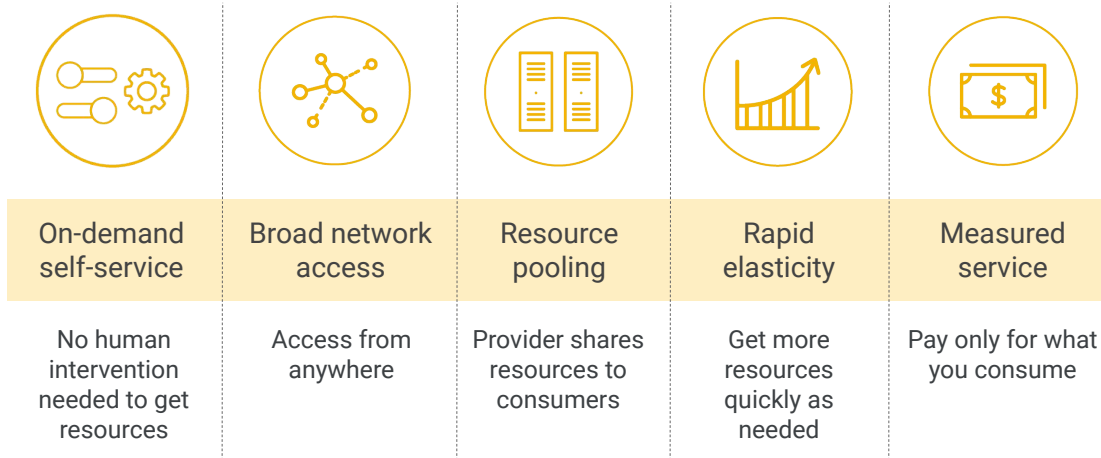


Google Cloud offers four main kinds of services: compute, storage, big data and machine learning. This course focuses mostly on the first two together with the topic of networking. After all, you can't use resources in the Cloud without Cloud Networking.

Periodically, we will check in to link back to products and services that you may have experience with and highlight the services in Google Cloud that match that functionality.

The Cloud is a great home for your applications and your data because it can free you from a lot of overhead chores. And Google Cloud gives you reasonably priced access to the same planet-scale infrastructure that Google runs on. What exactly is Google Cloud? How is it organized? And what makes it unique? In this module, we will orient you to the basics.

What is cloud computing?



Cloud computing has five fundamental attributes, according to the [definition of cloud computing](#) proposed by the United States National Institute of Standards and Technology.

First, customers get computing resources on-demand and self-service. Cloud-computing customers use an automated interface and get the processing power, storage, and network they need, with no need for human intervention.

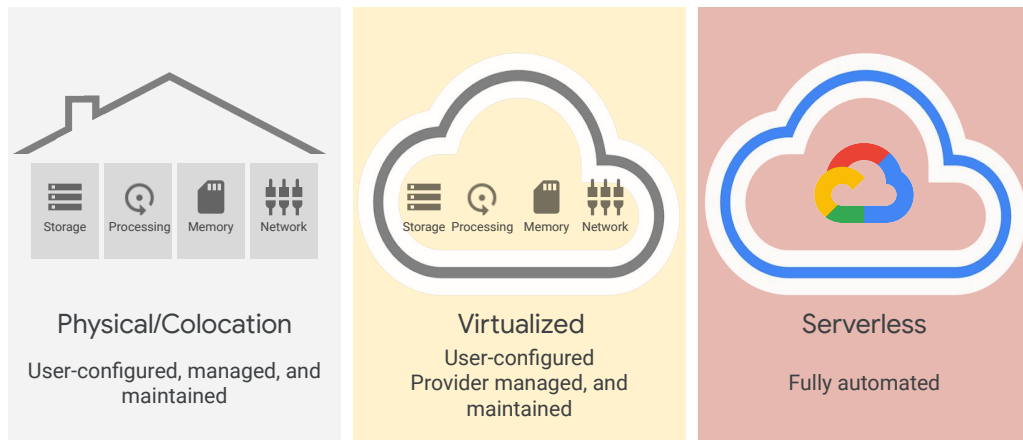
Second, they can access these resources over the network.

Third, the provider of those resources has a big pool of them, and allocates them to customers out of the pool. That allows the provider to get economies of scale by buying in bulk. Customers don't have to know or care about the exact physical location of those resources.

Fourth, the resources are elastic. Customers who need more resources can get more rapidly. When they need less, they can scale back.

And last, the customers pay only for what they use or reserve, as they go. If they stop using resources, they stop paying.

How did we get here? Where are we going?



The first wave of the trend towards cloud computing was colocation. Colocation gave users the financial efficiency of renting physical space, instead of investing in data center real estate.

Virtualized data centers of today, the second wave, share similarities with the private data centers and colocation facilities of decades past. The components of virtualized data centers match the physical building blocks of hosted computing—servers, CPUs, disks, load balancers, and so on—but now they are virtual devices. Virtualization does provide a number of benefits: your development teams can move faster, and you can turn capital expenses into operating expenses. With virtualization you still maintain the infrastructure; it is still a user-controlled/user-configured environment.

About 10 years ago, Google realized that its business couldn't move fast enough within the confines of the virtualization model. So Google switched to a container-based architecture—a fully automated, elastic third-wave cloud that consists of a combination of automated services and scalable data. Services automatically provision and configure the infrastructure used to run applications.

Today Google Cloud makes this third-wave cloud available to Google customers.

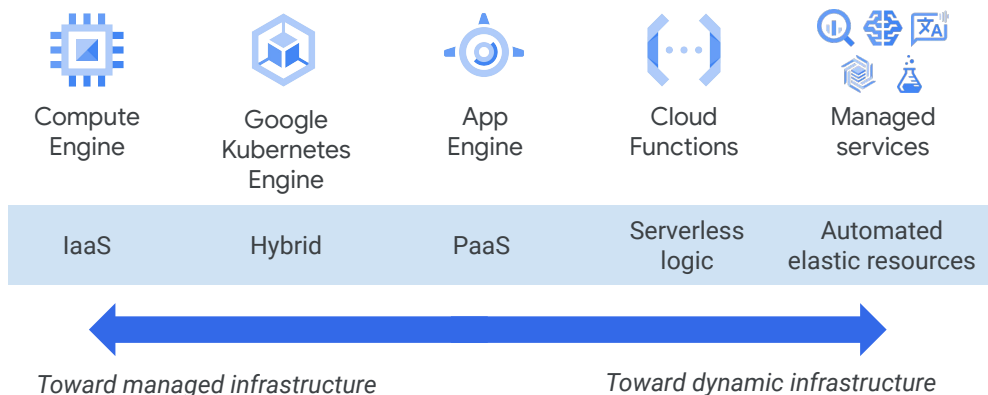
Every company is a data company



Google believes that, in the future, every company—regardless of size or industry—will differentiate itself from its competitors through technology. Largely, that technology will be in the form of software. Great software is centered on data. Thus, every company is or will become a data company.

Google Cloud provides a wide variety of services for managing and getting value from data at scale.

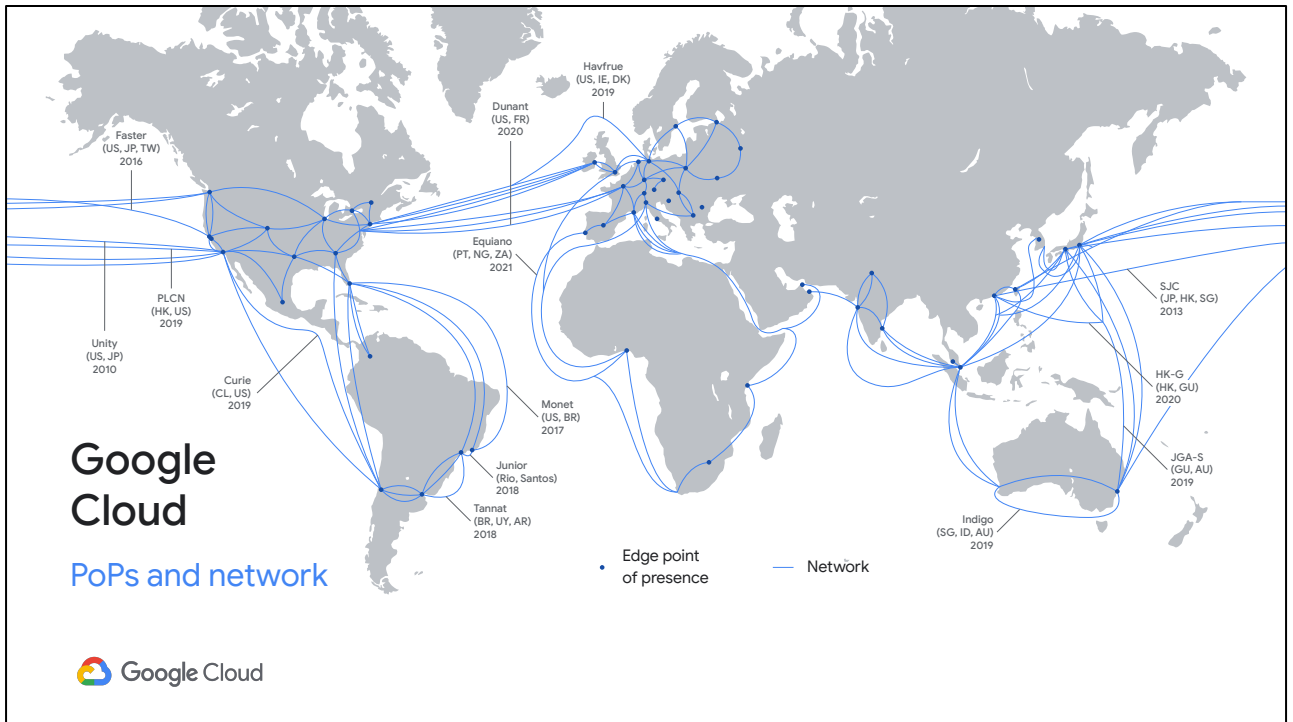
Google Cloud computing architectures meet you where you are



Virtualized data centers brought you infrastructure as a service (IaaS) and platform as a service (PaaS) offerings. IaaS offerings provide you with raw compute, storage, and network, organized in ways familiar to you from physical and virtualized data centers. PaaS offerings, on the other hand, bind your code to libraries that provide access to the infrastructure your application needs, thus allow you to focus on your application logic.

In the IaaS model, you pay for what you allocate. In the PaaS model, you pay for what you use.

As cloud computing has evolved, the momentum has shifted toward managed infrastructure and managed services.

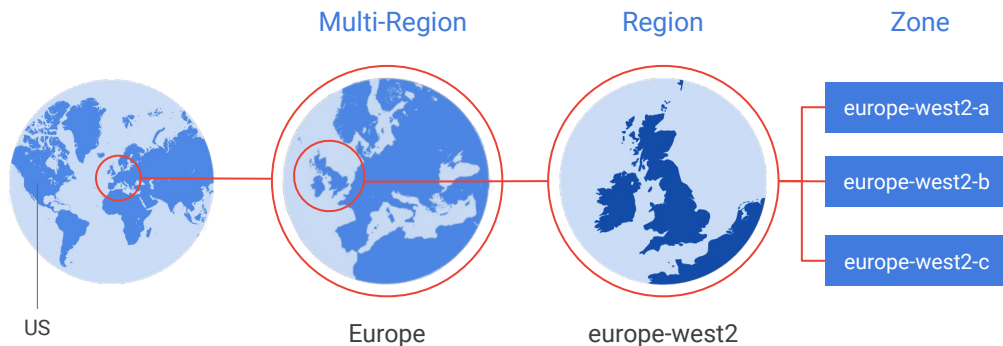


According to some publicly available estimates, Google's network carries as much as 40% of the world's internet traffic every day. Google's network is the largest network of its kind on Earth. Google has invested billions of dollars over the years to build it.

It is designed to give customers the highest possible throughput and lowest possible latencies for their applications.

The network interconnects 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's edge caching nodes respond to users requests from an Edge Network location that will provide the lowest latency.

Google Cloud is organized into regions and zones



Regions and zones

[Regions](#) are independent geographic areas that consist of [zones](#). Locations within regions tend to have round-trip network latencies of under 5 milliseconds on the 95th percentile.

A zone is a deployment area for Google Cloud resources within a region. Think of a zone as a single failure domain within a region. In order to deploy fault-tolerant applications with high availability, you should deploy your applications across multiple zones in a region to help protect against unexpected failures.

To protect against the loss of an entire region due to natural disaster, you should have a disaster recovery plan and know how to bring up your application in the unlikely event that your primary region is lost.

For more information on the specific resources available within each location option, see Google's [Global Data Center Locations](#).

Google Cloud's services and resources can be [zonal](#), [regional](#), or [managed by Google across multiple regions](#). For more information on what these options mean for your data, see [geographic management of data](#).

Zonal resources

Zonal resources operate within a single zone. If a zone becomes unavailable, all of the zonal resources in that zone are unavailable until service is restored.

- Compute Engine VM instance resides within a specific zone.

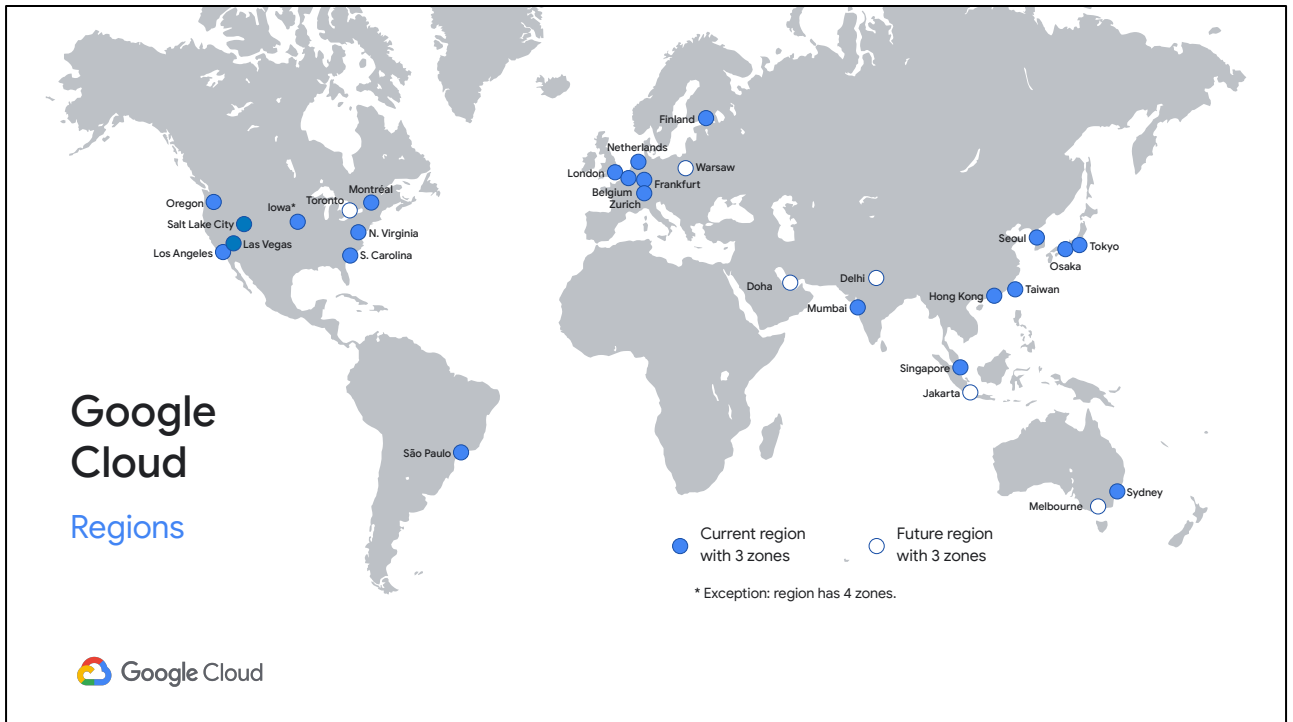
Regional resources

Regional resources are deployed with redundancy within a region. This gives them higher availability relative to zonal resources.

Multi-regional resources

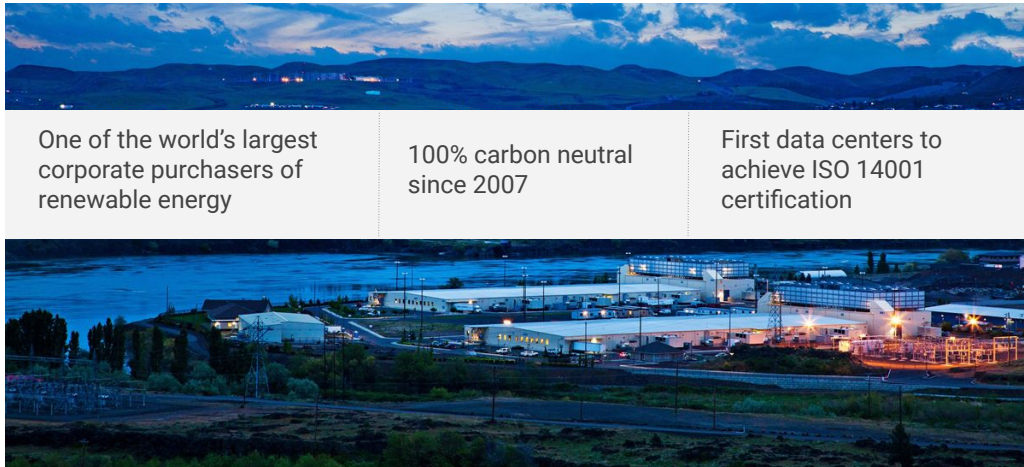
A few Google Cloud services are managed by Google to be redundant and distributed within and across regions. These services optimize availability, performance, and resource efficiency. As a result, these services require a trade-off on either latency or the consistency model. These trade-offs are documented on a product-specific basis. The following services have one or more multi-regional deployments in addition to any regional deployments:

- App Engine and its features
- Firestore
- Cloud Storage
- BigQuery



As of mid-2020, Google Cloud has 24 regions and 73 zones with more to come.

Google is committed to environmental responsibility



This image shows Google's data center in Hamina, Finland. The facility is one of the most advanced and efficient data centers in the Google fleet. Its cooling system, which uses sea water from the Bay of Finland, reduces energy use and is the first of its kind anywhere in the world.

Google is one of the world's largest corporate purchasers of wind and solar energy. Google has been 100% carbon neutral since 2007, and will shortly reach 100% renewable energy sources for its data centers.

The virtual world is built on physical infrastructure, and all those racks of humming servers use vast amounts of energy. Together, all existing data centers use roughly 2% of the world's electricity. So Google works to make data centers run as efficiently as possible. Google's data centers were the first to achieve ISO 14001 certification, a standard that maps out a framework for improving resource efficiency and reducing waste.

Regions and zones are different in AWS

Google Cloud

Each region is composed of zones in close proximity to other zones.

Google offers services that are multi-regional, global, or zonal.

AWS

Each region is composed of multiple Availability Zones in close proximity to other zones.

Most services are regional or zonal; Cloudfront is global.



Google and AWS both use regions as a way to provide cloud services to customers. One difference is that Google also uses zones to provide data center services, and every region will have at least 3 zones.

AWS uses clusters of data centers called *Availability Zones* as a way to provide high availability. Every region will have at least two availability zones.

Google Cloud and AWS use PoPs in different ways

Google Cloud

Uses PoPs to provide Cloud CDN and to deliver built-in edge caching for services such as App Engine and Cloud Storage.

AWS

Uses PoPs to provide the CDN service Amazon CloudFront and to deliver built-in edge caching for services such as Lambda@Edge.



Google Cloud and AWS both have points of presence (PoPs) located in many more locations around the world. These points of presence locations help cache content closer to end users. However, each platform uses their respective points of presence locations in different ways.

Google Cloud uses points of presence to provide Cloud CDN and to deliver built-in edge caching for services such as App Engine and Cloud Storage.

AWS uses points of presence to provide the content delivery network service Amazon CloudFront and for edge caching services like Lambda at the edge.

Google Cloud's points of presence connect to data centers through Google-owned fiber. This unimpeded connection means that Google Cloud-based applications have fast, reliable access to all of the services on Google Cloud.

Summary of region and zones terminology

Concept	Google Cloud term	AWS term
Cluster of data centers and services	Region	Region
Abstracted data center	Zone	Availability Zone
Edge caching	PoP (multiple services)	PoP (multiple services)



To summarize let's look at the terminology associated with region and zones. Both Google Cloud and AWS products use the term *region* to define a cluster of data centers and services that are relatively close to each other.

Data center services and availability can be abstracted into Zones in Google Cloud, which are the equivalent of Availability Zones in AWS.

Google Cloud uses POPs to deliver built-in edge caching for multiple services, such as App Engine and Cloud Storage. AWS delivers edge caching in a similar way.

Google offers customer-friendly pricing

Billing in sub-hour increments	Discounts for sustained use	Discounts for committed use	Discounts for preemptible use	Custom VM instance types
For compute, data processing and other services.	Automatically applied to virtual machine use over 25% of a month.	Pay less for steady, long-term workloads.	Pay less for interruptible workloads.	Pay only for the resources you need for your application.



Google was the first major cloud provider to deliver [per-second billing](#) for its Infrastructure-as-a-Service compute offering, Compute Engine. Per-second billing is offered for users of Compute Engine, Google Kubernetes Engine (container infrastructure as a service), Dataproc (the open-source Big Data system Hadoop as a service), and App Engine flexible environment VMs (a platform as a service).

Compute Engine offers automatically applied [sustained-use discounts](#), which are automatic discounts that you get for running a virtual-machine instance for a significant portion of the billing month. Specifically, when you run an instance for more than 25% of a month, Compute Engine automatically gives you a discount for every incremental minute you use for that instance.

[Custom virtual machine types](#) allow Compute Engine virtual machines to be fine-tuned for their applications, so that you can tailor your pricing for your workloads.

Try the [online pricing calculator](#) to help estimate your costs.

Open APIs and open source means that customers can leave

Open APIs; compatibility with open-source services



Cloud Bigtable



Dataproc

Open source for a rich ecosystem



TensorFlow



Kubernetes



Forseti Security

Multi-vendor-friendly technologies



Operations



Google
Kubernetes Engine



Google gives customers the ability to run their applications elsewhere if Google becomes no longer the best provider for their needs.

This includes:

- Using Open APIs. Google services are compatible with open-source products. For example, Cloud Bigtable, a horizontally scalable managed database: Bigtable uses the Apache HBase interface, which gives customers the benefit of code portability. Another example: Dataproc offers the open-source big data environment Hadoop as a managed service.
- Google publishes key elements of its technology, using open-source licenses, to create ecosystems that provide customers with options other than Google. For example, TensorFlow, an open-source software library for machine learning developed inside Google, is at the heart of a strong open-source ecosystem.
- Google provides interoperability at multiple layers of the stack. Kubernetes and Google Kubernetes Engine give customers the ability to mix and match microservices running across different clouds. Google Cloud's operations suite lets customers monitor workloads across multiple cloud providers.

Security is designed into Google's technical infrastructure

Layer	Notable security measures (among others)
Operational security	Intrusion detection systems; techniques to reduce insider risk; employee U2F use; software development practices
Internet communication	Google Front End; designed-in Denial of Service protection
Storage services	Encryption at rest
User identity	Central identity service with support for U2F
Service deployment	Encryption of inter-service communication
Hardware infrastructure	Hardware design and provenance; secure boot stack; premises security



Hardware design and provenance: Both the server boards and the networking equipment in Google data centers are custom-designed by Google. Google also designs custom chips, including a hardware security chip that is currently being deployed on both servers and peripherals.

Secure boot stack: Google server machines use a variety of technologies to ensure that they are booting the correct software stack, such as cryptographic signatures over the BIOS, bootloader, kernel, and base operating system image.

Premises security: Google designs and builds its own data centers, which incorporate multiple layers of physical security protections. Access to these data centers is limited to only a very small fraction of Google employees. Google additionally hosts some servers in third-party data centers, where we ensure that there are Google-controlled physical security measures on top of the security layers provided by the data center operator.

Encryption of inter-service communication: Google's infrastructure provides cryptographic privacy and integrity for remote procedure call ("RPC") data on the network. Google's services communicate with each other using RPC calls. The infrastructure automatically encrypts all infrastructure RPC traffic which goes between data centers. Google has started to deploy hardware cryptographic accelerators that will allow it to extend this default encryption to all infrastructure RPC traffic inside Google data centers.

User identity: Google's central identity service, which usually manifests to end users as the Google login page, goes beyond asking for a simple username and password. The service also intelligently challenges users for additional information based on risk factors such as whether they have logged in from the same device or a similar location in the past. Users also have the option of employing second factors when signing in, including devices based on the Universal 2nd Factor (U2F) open standard

Encryption at rest: Most applications at Google access physical storage indirectly via storage services, and encryption (using centrally managed keys) is applied at the layer of these storage services. Google also enables hardware encryption support in hard drives and SSDs.

Google Front End ("GFE"): Google services that want to make themselves available on the Internet register themselves with an infrastructure service called the Google Front End, which ensures that all TLS connections are ended using correct certificates and following best practices such as supporting perfect forward secrecy. The GFE additionally applies protections against Denial of Service attacks.

Denial of Service ("DoS") protection: The sheer scale of its infrastructure enables Google to simply absorb many DoS attacks. Google also has multi-tier, multi-layer DoS protections that further reduce the risk of any DoS impact on a service running behind a GFE.

Intrusion detection: Rules and machine intelligence give operational security engineers warnings of possible incidents. Google conducts Red Team exercises to measure and improve the effectiveness of its detection and response mechanisms.

Reducing insider risk: Google aggressively limits and actively monitors the activities of employees who have been granted administrative access to the infrastructure.

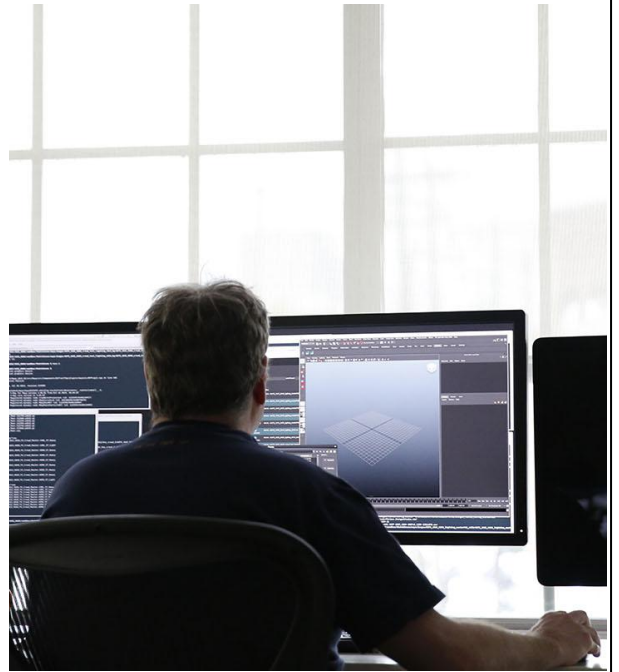
Employee U2F use: To guard against phishing attacks against Google employees, employee accounts require use of U2F-compatible Security Keys.

Software development practices: Google employs central source control and requires two-party review of new code. Google also provides its developers libraries that prevent them from introducing certain classes of security bugs. Google also runs a Vulnerability Rewards Program where we pay anyone who is able to discover and inform us of bugs in our infrastructure or applications.

For more information about Google's technical-infrastructure security, see <https://cloud.google.com/security/security-design/>

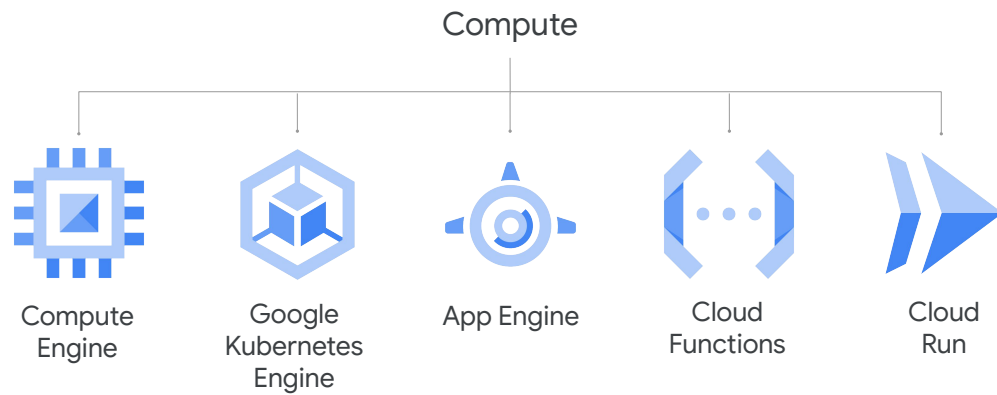
Why choose Google Cloud?

Google Cloud enables developers to **build**, **test**, and **deploy** applications on Google's **highly secure**, **reliable**, and **scalable** infrastructure.



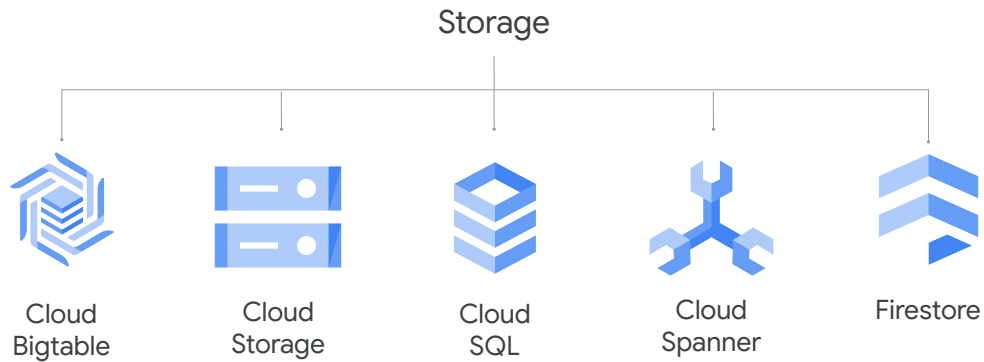
Google Cloud lets you choose from computing, storage, big data/machine learning, and application services for your web, mobile, analytics, and backend solutions.

Google Cloud offers a range of compute services



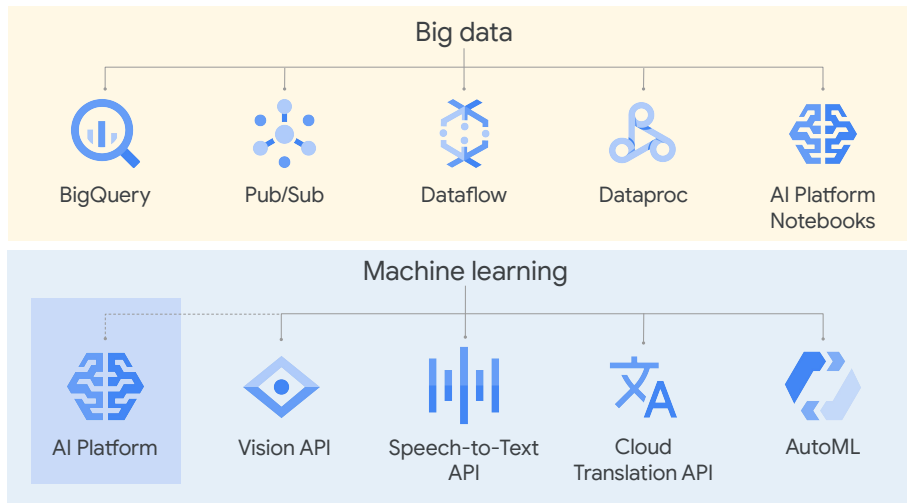
Google Cloud's products and services can be broadly categorized as Compute, Storage, Big Data, Machine Learning, Networking, and Operations/Tools. This course considers each of the compute services and discuss why customers might choose each.

Google Cloud offers a range of storage services



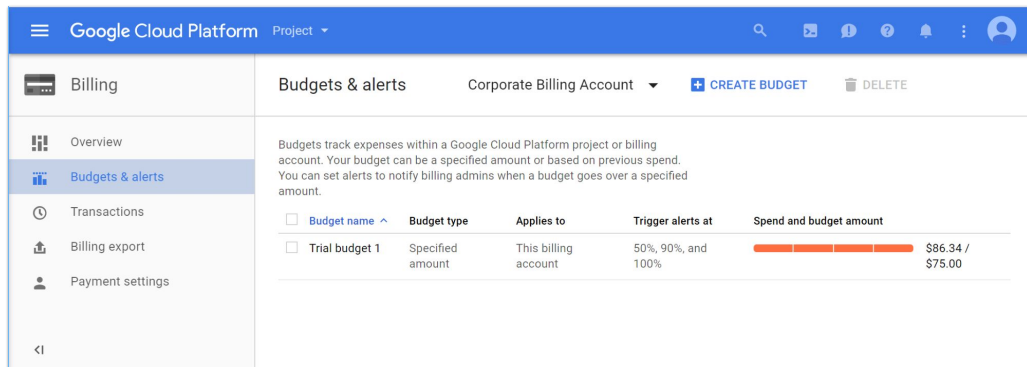
This course will examine each of Google Cloud's storage services: how it works and when customers use it. To learn more about these services, you can participate in the training courses in Google Cloud's [Data Analyst learning track](#).

Google Cloud offers services to get value from data



This course also examines the function and purpose of Google Cloud's big data and machine-learning services. More details about these services are also available in the training courses in Google Cloud's [Data Analyst learning track](#).

Budgets and alerts keep your billing under control



The screenshot shows the Google Cloud Platform Billing interface. The left sidebar contains a menu with options: Billing, Overview, Budgets & alerts (selected), Transactions, Billing export, and Payment settings. The main content area is titled 'Budgets & alerts' and shows a 'Corporate Billing Account' with a '+ CREATE BUDGET' button and a 'DELETE' button. Below this, there is a table with columns: Budget name, Budget type, Applies to, Trigger alerts at, and Spend and budget amount. The table contains one entry: 'Trial budget 1' with a 'Specified amount' type, applying to 'This billing account', triggering alerts at '50%, 90%, and 100%', and showing a spend of '\$86.34 / \$75.00'.

Budget name	Budget type	Applies to	Trigger alerts at	Spend and budget amount
Trial budget 1	Specified amount	This billing account	50%, 90%, and 100%	\$86.34 / \$75.00



You're probably thinking, "How can I make sure I don't accidentally run up a big Google Cloud bill?" Google Cloud provides four tools to help:

- Budgets and alerts,
- Billing export,
- Reports, and
- Quotas.

Let's look at Budgets and Alerts, first.

You can define **budgets** at the billing account level or at the project level. A budget can be a fixed limit, or it can be tied to another metric - for example, a percentage of the previous month's spend.

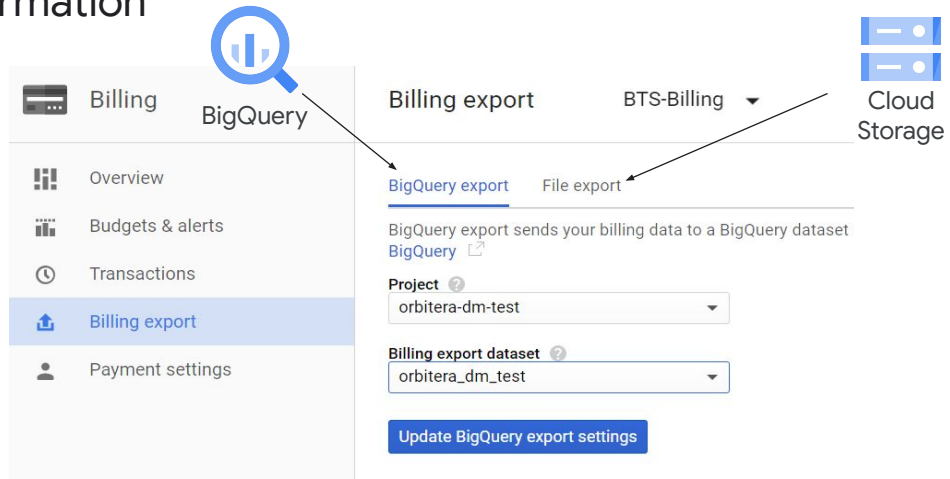
To be notified when costs approach your budget limit, you can create an **alert**. For example, with a budget limit of \$20,000 and an alert set at 90%, you'll receive a notification alert when your expenses reach \$18,000. Alerts are generally set at 50%, 90% and 100%, but can also be customized.

Source:

<https://cloud.google.com/billing/docs/>

<https://cloud.google.com/billing/docs/how-to/budgets>

Billing export allows you to store detailed billing information

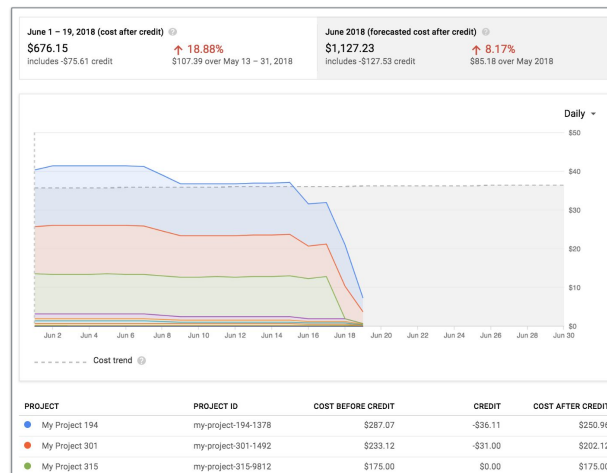


Billing export allows you to store detailed billing information in places where it is easy to retrieve for more detailed analysis, such as a BigQuery dataset or Cloud Storage bucket.

Source:

<https://codelabs.developers.google.com/codelabs/orbitera-gcp-billing/#1>

Reports is a visual tool to monitor expenditure



And Reports is a visual tool in the Cloud Console that allows you to monitor expenditure based on a project or services.

Source:

<https://cloud.google.com/billing/docs/how-to/reports>

Quotas are helpful limits



Rate quota

GKE API: 1,000 requests per 100 seconds

Allocation quota

5 networks per project

Many quotas are changeable



Google Cloud also implements quotas, which are designed to prevent the over-consumption of resources because of an error or a malicious attack, protecting both account owners and the Google Cloud community as a whole.

There are two types of quotas: rate quotas and allocation quotas. Both are applied at the project level.

Rate quotas reset after a specific time. For example, by default, the GKE service implements a quota of 1,000 calls to its API from each Google Cloud project every 100 seconds. After that 100 seconds, the limit is reset.

Allocation quotas govern the number of resources you can have in your projects. For example, by default, each Google Cloud project has a quota allowing it no more than 5 Virtual Private Cloud networks.

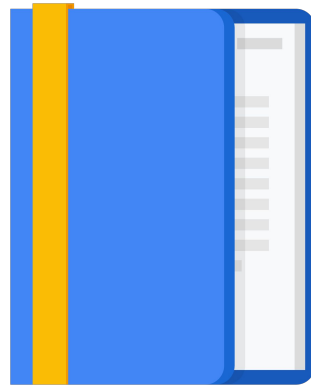
Although projects all start with the same quotas, you can change some of them by requesting an increase from Google Cloud Support.

Agenda

Introduction to Google Cloud

Quiz

Resources



Quiz 1

Name some of Google Cloud's pricing innovations.

Quiz 1

Name some of Google Cloud's pricing innovations.

1. Sub-hour billing
2. Sustained-use discounts
3. Compute Engine custom machine types

Quiz 2

Name some benefits of using Google Cloud other than its pricing.

Quiz 2

Name some benefits of using Google Cloud other than its pricing.

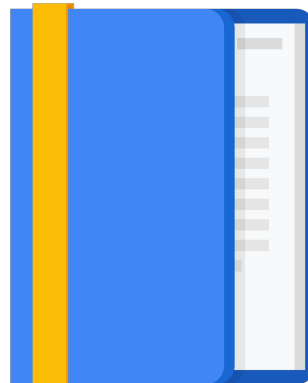
1. Commitment to environmental responsibility
2. Commitment to open-source technologies
3. Robust infrastructure

Agenda

Introduction to Google Cloud

Quiz

Resources



Resources

Why Google Cloud? <https://cloud.google.com/why-google/>

Pricing philosophy <https://cloud.google.com/pricing/philosophy/>

Data centers <https://www.google.com/about/datacenters/>

Google Cloud product overview <http://cloud.google.com/products/>

Google Cloud solutions <http://cloud.google.com/solutions/>

