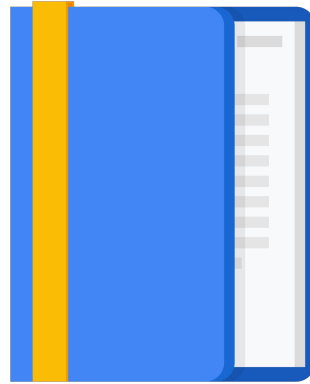# Google Cloud

GCP Fundamentals: Core Infrastructure

Introducing Google Cloud Platform

Agenda

Introduction to Google Cloud Platform

Quiz

What is cloud computing?

On-demand self-service — No human intervention needed to get resources

Broad network access — Access from anywhere

Resource pooling — Provider shares resources to customers

Rapid elasticity — Get more resources quickly as needed

Measured service — Pay only for what you consume

Google Cloud

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.
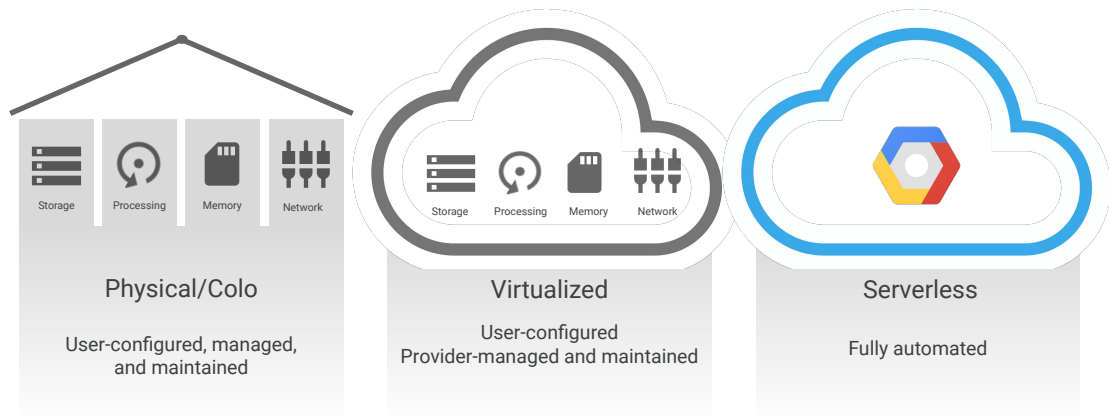
Second, they can access these resources over the network.

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?

Physical/Colo	Virtualized	Serverless
Storage Processing Memory Network	Storage Processing Memory Network	
User-configured, managed, and maintained	User-configured Provider-managed and maintained	Fully automated

Google Cloud

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 Platform makes this third-wave cloud available to Google customers.
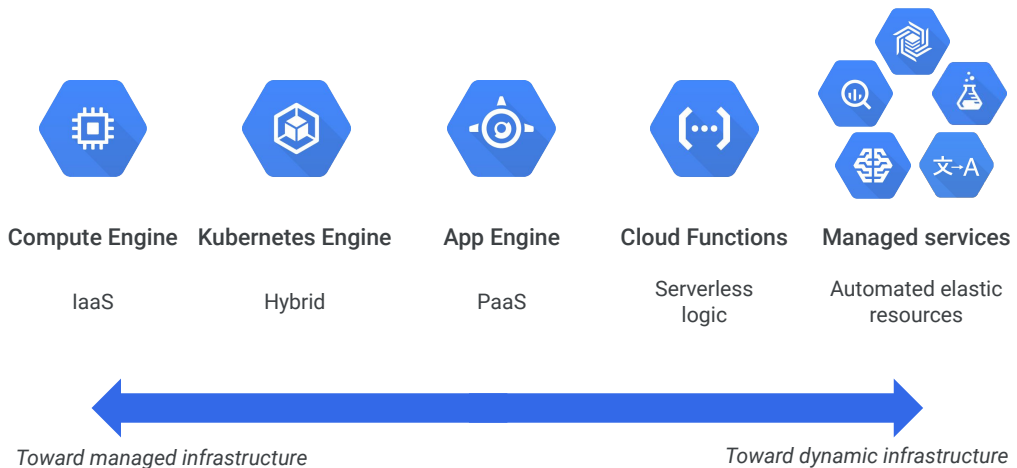
Every company is a data company

Google Cloud

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.

GCP computing architectures meet you where you are

Compute Engine	Kubernetes Engine	App Engine	Cloud Functions	Managed services
IaaS	Hybrid	PaaS	Serverless logic	Automated elastic resources

Toward managed infrastructure   *Toward dynamic infrastructure*

Google Cloud

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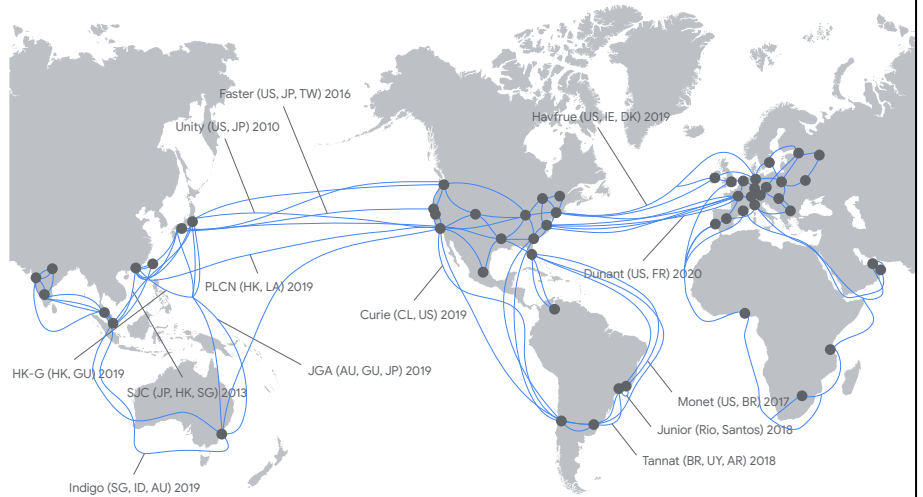
In the IaaS model, you pay for what you allocate. In the PaaS model, you pay for what you use.

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

Google Cloud Platform

134 points of presence and 13 subsea cable investments around the globe

● Edge points of presence

— Network

Faster (US, JP, TW) 2016
Unity (US, JP) 2010
Havfrue (US, IE, DK) 2019
PLCN (HK, LA) 2019
Dunant (US, FR) 2020
Curie (CL, US) 2019
HK-G (HK, GU) 2019
JGA (AU, GU, JP) 2019
SJC (JP, HK, SG) 2013
Monet (US, BR) 2017
Junior (Rio, Santos) 2018
Tannat (BR, UY, AR) 2018
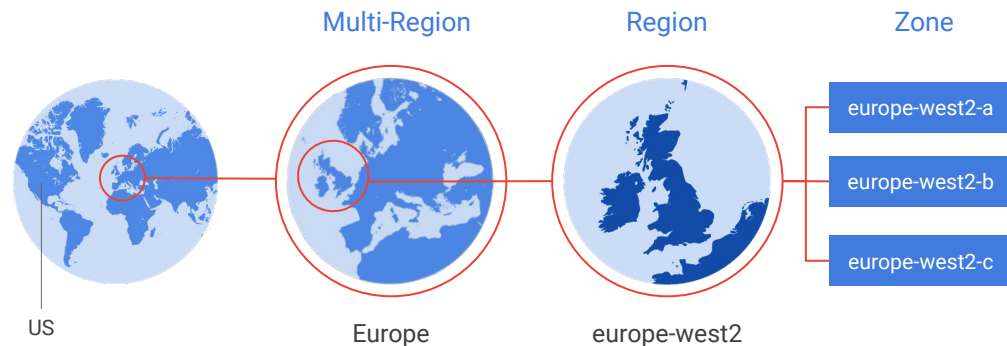Indigo (SG, ID, AU) 2019

Google Cloud

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 responds to the user's request from an Edge Network location that will provide the lowest latency. Google's edge caching network sites content close to end users to minimize latency.

GCP is organized into regions and zones

| Multi-Region | Region | Zone |
| US | Europe | europe-west2 | europe-west2-a / europe-west2-b / europe-west2-c |

Google Cloud

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 Platform 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 Platform'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.

- Google 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 Platform 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:
- Google App Engine and its features
- Google Cloud Datastore
- Google Cloud Storage
- Google BigQuery

At the time of this writing, Google Cloud Platform has 20 regions, with more to come.

Google is committed to environmental responsibility

100% carbon neutral since 2007

One of the world's largest corporate purchasers of renewable energy

First data centers to achieve ISO 14001 certification

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.

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 Cloud

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

Google 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 Google 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 mean customers can leave

| Open APIs; compatibility with open-source services | Open source for a rich ecosystem | Multi-vendor-friendly technologies |

Cloud Bigtable

Cloud Dataproc

TensorFlow   Kubernetes

Forseti Security

Google Stackdriver

Kubernetes Engine

Google Cloud

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, Google Cloud Bigtable, a horizontally scalable managed database: Bigtable uses the Apache HBase interface, which gives customers the benefit of code portability. Another example: Google Cloud 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 Stackdriver 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

Google Cloud

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 terminated 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 Platform?

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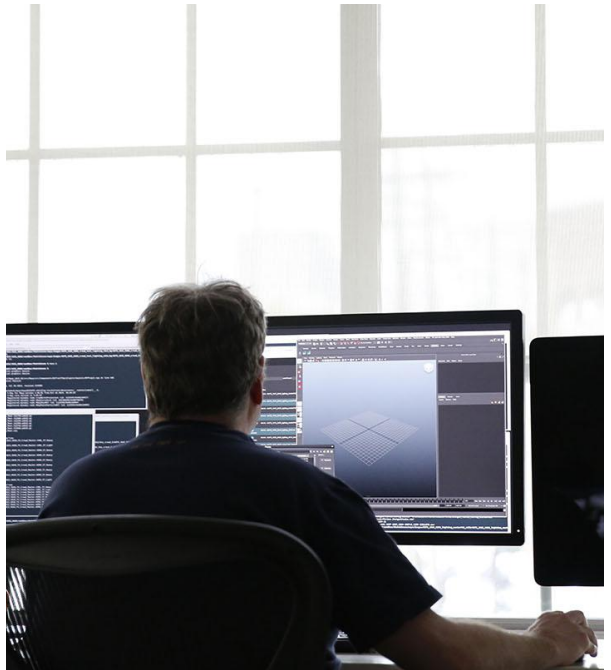
Google Cloud

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

Google Cloud Platform offers a range of compute services

Compute

Compute Engine   Kubernetes Engine   App Engine   Cloud Functions
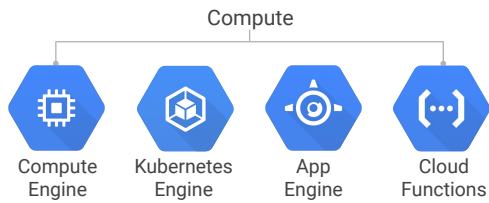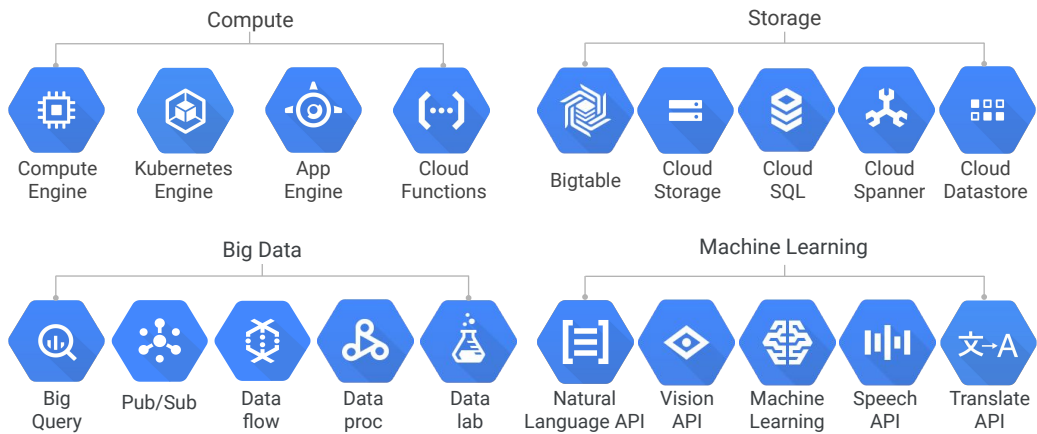
Google Cloud

Google Cloud Platform'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 Platform offers a range of storage services

Compute					Storage				
Compute Engine	Kubernetes Engine	App Engine	Cloud Functions		Bigtable	Cloud Storage	Cloud SQL	Cloud Spanner	Cloud Datastore

Google Cloud

This course will examine each of Google Cloud Platform'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 Platform offers services for getting value from data

This course also examines the function and purpose of Google Cloud Platform'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

You're probably thinking, "How can I make sure I don't accidentally run up a big GCP bill?" GCP 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.
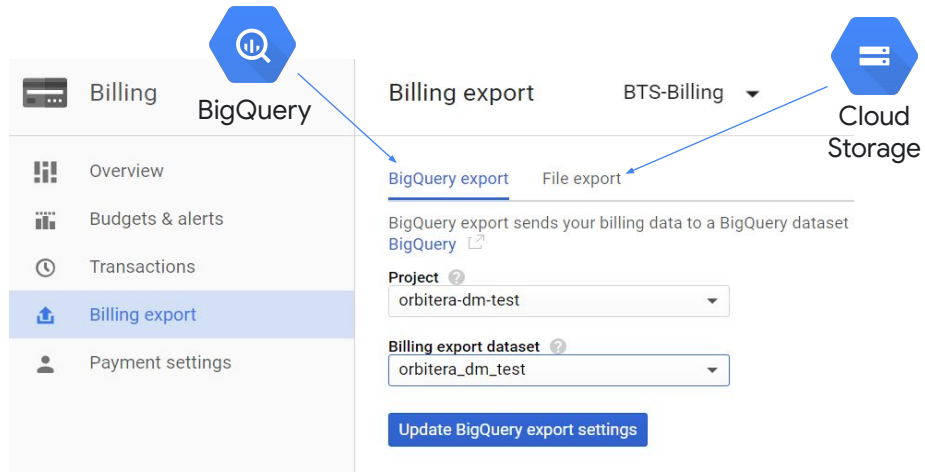
To be notified when costs approach your budget limit, you can create an **alert**. For example, with a budget limit of $20,000 and an alert set at 90%, you'll receive a notification alert when your expenses reach $18,000. Alerts are generally set at 50%, 90% and 100%, but can also be customized.

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

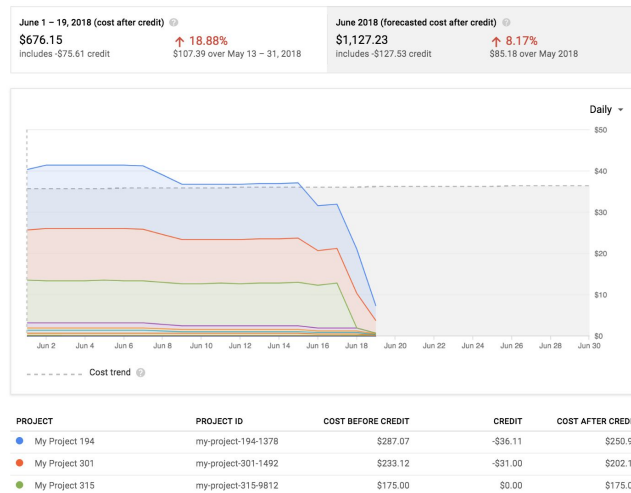# Billing export allows you to store detailed billing information

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

June 1 – 19, 2018 (cost after credit) ⓘ		June 2018 (forecasted cost after credit) ⓘ	
$676.15	↑ 18.88%	$1,127.23	↑ 8.17%
includes -$75.61 credit	$107.39 over May 13 – 31, 2018	includes -$127.53 credit	$85.18 over May 2018

Daily ▾

- - - - - - - - - Cost trend ⓘ

PROJECT	PROJECT ID	COST BEFORE CREDIT	CREDIT	COST AFTER CREDIT
● My Project 194	my-project-194-1378	$287.07	-$36.11	$250.96
● My Project 301	my-project-301-1492	$233.12	-$31.00	$202.12
● My Project 315	my-project-315-9812	$175.00	$0.00	$175.00

And Reports is a visual tool in the 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

GCP 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 GCP 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 GCP 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 GCP 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 Platform

Quiz
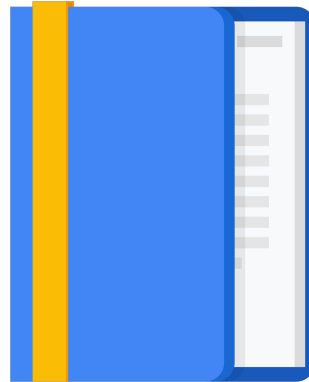
Question #1

Name some of Google Cloud Platform's pricing innovations.

Question #1

Name some of Google Cloud Platform's pricing innovations.

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

Question #2

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

Question #2

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

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

More resources

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

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

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

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

Google Cloud Platform solutions http://cloud.google.com/solutions/