# Google Cloud Platform Building Blocks

A

B

C

H

I

J

K

L

# A

## Access Context Manager

Access Context Manager allows GCP organization administrators to define fine-grained, attribute based access control for projects and resources in GCP.

Administrators first define an access policy, which is an organization-wide container for access levels and Access Zones.

Access levels describe the necessary requirements for requests to be honored. Examples include:

- **Device type and operating system**
- IP address
- **User identity**

Access zones define sandboxes of resources which can freely exchange data within the zone, but are not allowed to export data outside of it. Use of access zones is currently by invitation only. If your application is not whitelisted, any attempt to call access zone APIs will result in an error.

## AirFlow Apache

Schedule and monitor workflows as directed acyclic graphs (DAGs) of tasks. The airflow scheduler executes your tasks on an array of workers while following the specified dependencies.
The user interface lets visualize pipelines running in production, monitor progress, and troubleshoot issues.

## Andromeda

Google Cloud's software-defined network virtualization stack

## Anycast

Anycast is a network addressing and routing methodology in which a single destination address has multiple routing paths to two or more endpoint destinations. Routers will select the desired path on the basis of number of hops, distance, lowest cost, latency measurements or based on the least congested route. Anycast networks are widely used for content delivery network(CDN) products to bring their content closer to the end user.

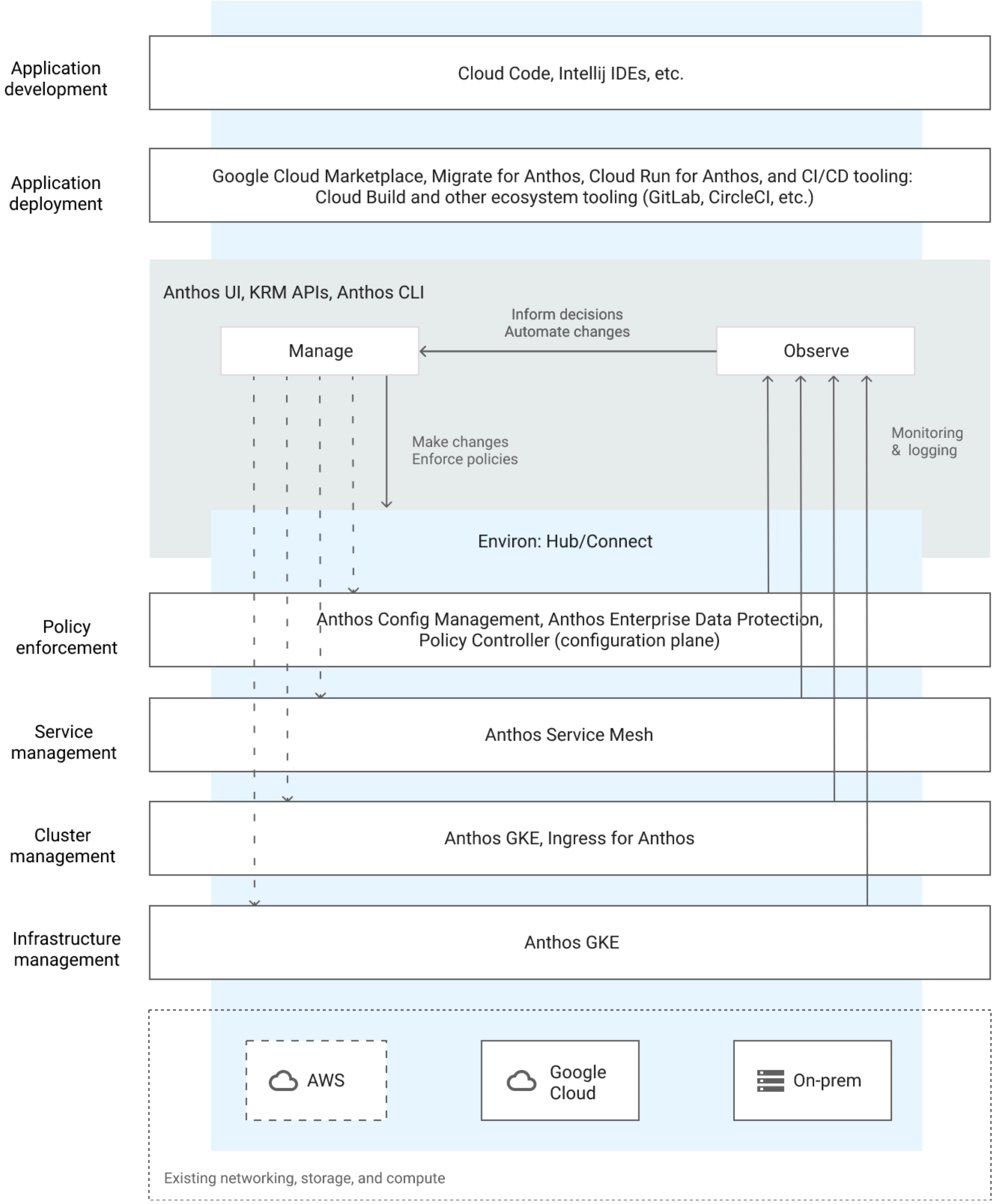
The Internet Protocol and other network addressing systems recognize five main addressing methods:

- Unicast addressing uses a one-to-one association between a sender and destination: each destination address uniquely identifies a single receiver endpoint.
- Broadcast uses a one-to-all association; a single datagram from one sender is routed to all of the possibly multiple endpoints associated with the broadcast address. The network automatically replicates datagrams as needed to reach all the recipients within the scope of the broadcast, which is generally an entire network subnet.
- Multicast addressing uses a one-to-many-of-many or many-to-many-of-many association; datagrams are routed simultaneously in a single transmission to many recipients. It differs from broadcast in that the destination address designates a subset, not necessarily all, of the accessible nodes.
- Anycast addressing is a one-to-one-of-many association where datagrams are routed to any single member of a group of potential receivers that are all identified by the same destination address. The routing algorithm selects the single receiver from the group based on least-expensive routing metric. In practice, this means that packets are routed to the topologically-nearest member of an anycast group.

- Geocast refers to the delivery of information to a group of destinations in a network identified by their geographical locations. It is a specialized form of multicast addressing used by some routing protocols for mobile ad hoc networks.

# Anthos

Anthos is an enterprise application management platform that provides a consistent development and operations experience for multi-cloud and on-premises environments.


Application development
Cloud Code, Intellij IDEs, etc.
Application deployment
Google Cloud Marketplace, Migrate for Anthos, Cloud Run for Anthos, and CI/CD tooling: Cloud Build and other ecosystem tooling (GitLab, CircleCI, etc.)
Anthos UI, KRM APIs, Anthos CLI
Inform decisions
Automate changes
Manage
Observe
Make changes
Enforce policies
Monitoring & logging
Environ: Hub/Connect
Policy enforcement
Anthos Config Management, Anthos Enterprise Data Protection, Policy Controller (configuration plane)
Service management
Anthos Service Mesh
Cluster management
Anthos GKE, Ingress for Anthos
Infrastructure management
Anthos GKE
AWS
Google Cloud
On-prem
Existing networking, storage, and compute

# API Analytics

Gain end-to-end visibility across APIs programs with the operational, developer engagement, and business metrics required to monitor, measure, and manage API programs. Get real-time information about the entire digital ecosystem including apps, consumption, API performance and usage metrics like traffic trends and spikes, latency, response times, and other custom criteria.

# Apigee API Platform

Apigee is a full lifecycle API management platform that enables API providers to design, secure, deploy, monitor, and scale APIs. Apigee sits in-line with runtime API traffic and enforces a set of out-of-the-box API policies, including key validation, quota management, transformation, authorization, and access control. API providers use the customizable developer portal to enable developers to consume APIs easily and securely, and to measure API performance and usage

# Apigee Sense

Apigee Sense works in conjunction with the Apigee Edge API Management Platform to give API teams a powerful weapon to protect APIs from attacks. Sense provides a layer of API security by identifying and alerting administrators to suspicious API behaviors. Administrators determine the response and apply corrective actions to maintain user experience and protect back-end systems. Apigee Sense takes the critical next step of automating remediation for future attacks.

# APIs, CLOUD APIS

Access Google Cloud Platform products from your code. Cloud APIs provide similar functionality to Cloud SDK and Cloud Console, and allow you to automate your workflows by using your favorite language. Use these Cloud APIs with REST calls or client libraries in popular programming languages.

# App Engine

Google App Engine enables you to build and host applications on the same systems that power Google applications. PaaS. Managed Service. Using the App Engine standard environment means that your application instances run in a sandbox, using the runtime environment of a supported language listed below.

For some languages, building an application to run in the standard environment is more constrained and involved, but your applications will have faster scale up times, they are sandboxed and the fee is cheaper (can even scale the App to 0). App Engine Standard takes seconds to scale.

# App Engine Flexible Environment

App Engine allows developers to focus on doing what they do best, writing code. Based on Google Compute Engine, the App Engine flexible environment automatically scales your app up and down while balancing the load. Microservices, authorization, SQL and NoSQL databases, traffic splitting, logging, versioning, security scanning, and content delivery networks are all supported natively. In addition, the App Engine flexible environment allows you to customize the runtime and even the operating system of your virtual machine using Dockerfiles. App Engine Flexible takes minutes to scale.

Use it when you need resources from GCP Compute Engine, when you need your own Docker containers or your language is one of these: Python, Java, Node.js, Go, Ruby, PHP, .NET.

Learn about the differences between the standard environment and the flexible environment.

# Armor, Cloud Armor

Cloud Armor: Google Cloud Armor offers a policy framework and rules language for customizing access to internet-facing applications and deploying defenses against denial of service attacks.

# AS Autonomous Systems ASN Autonomous Systems Number

Within the Internet, an autonomous system (AS) is a collection of connected Internet Protocol (IP) routing prefixes under the control of one or more network operators on behalf of a single administrative entity or domain that presents a common, clearly defined routing policy to the Internet.

Originally the definition required control by a single entity, typically an Internet service provider or a very large organization with independent connections to multiple networks, that adhere to a single and clearly defined routing policy, as originally defined in RFC 1771.[2] The newer definition in RFC 1930 came into use because multiple organizations can run Border Gateway Protocol (BGP) using private AS numbers to an ISP that connects all those organizations to the Internet. Even though there may be multiple autonomous systems supported by the ISP, the Internet only sees the routing policy of the ISP. That ISP must have an officially registered **autonomous system number (ASN).**

# Autoscaler

Managed instance groups offer autoscaling capabilities that allow you to automatically add or delete instances from a managed instance group based on increases or decreases in load. Autoscaling helps your applications gracefully handle increases in traffic and reduces cost when the need for resources is lower. You just define the autoscaling policy and the autoscaler performs automatic scaling based on the measured load.

Autoscaling works by scaling up or down your instance group. That is, it adds more instances to your instance group when there is more load (upscaling), and deletes instances when the need for instances is lowered (downscaling).

# AutoML

**Cloud AutoML** is a machine learning product suite that enables developers with limited machine learning expertise to provide their data sets and obtain access to quality trained models produced by Google's transfer learning and Neural Architecture Search (Google's technology for finding, generating, evaluating, and training numerous neural architectures to automatically select a solution for the customer's application):

- **Cloud AutoML Vision** is a simple and flexible machine learning service that lets businesses and developers with limited machine learning expertise train custom and scalable vision models for their own use cases.
- **Cloud AutoML Natural Language** enables customers to categorize input text into their own custom defined labels (supervised classification). Users can customize models to their own domain or use case.
- **Cloud AutoML Translation** is a simple and scalable translation solution that allows businesses and developers with limited machine learning expertise to customize the Google Neural Machine Translation (GNMT) model for their own domain or use-case.

# Availability

Availability, or better, high availability is when a system or application is always  accessible also in case of hardware problems. HA is really important with databases, with replicas and snapshots, but also with VMs, when there is the capability to detect problems and raise another VM with the same configuration and data.

# Availability Policies

Availability options determine how your VM instances behave when maintenance events occur. How to set an instance to live migrate when a maintenance event occurs.
When there are maintenance events such as hardware or software updates that require Google to move your VM to a different host machine, Google Compute Engine automatically manages the scheduling behavior for your instances. Compute Engine **live migrates** your VM instances if you configured the instance's availability policy to use live migration. This prevents your applications from experiencing disruptions during these events. Alternatively, you can also choose to terminate your instances during these events rather than live migrating them.

# Avro

Apache Avro™ is a data serialization system.

Avro provides:

- Rich data structures.
- A compact, fast, binary data format.
- A container file, to store persistent data.
- Remote procedure call (RPC).
- Simple integration with dynamic languages. Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation as an optional optimization, only worth implementing for statically typed languages.

# B

# Backend Bucket

Backend buckets allow you to use Google Cloud Storage buckets with HTTP(S) Load Balancing.

An HTTP(S) load balancer can direct traffic from specified URLs to either a backend bucket or a backend service. For example, the load balancer can send requests for static content to a Cloud Storage bucket and requests for dynamic content to a VM.

For example, you can have the load balancer send traffic with a path of `/static` to a storage bucket and all other requests to your instances.

# Baking, Image

**Manual:** You can create a simple custom image by creating a new VM instance from a public image, configuring the instance with the applications and settings that you want, and then creating a custom image from that instance. Use this method if you can configure your images from scratch manually rather than using automated baking or importing existing images.

You can create a simple custom image using the following steps:

1. Create an instance from a public image.
2. Connect to the instance.
3. Customize the instance for your needs.
4. Stop the instance.

5. Create a custom image from the boot disk of that instance. This process requires you to delete the instance but keep the boot disk.

**Automated:** Manual baking is an easy way to start if you have a small number of images, but large numbers of images become difficult to audit and manage. Packer is an open-source tool for making image creation more reproducible, auditable, configurable, and reliable. For more information on how to create an automated image-creation pipeline, see the Automated Image Builds with Jenkins, Packer, and Kubernetes solution. You can also use Packer as part of a Spinnaker pipeline to produce images that are deployed to clusters of instances.

# Billing

A billing account is used to define who pays for a given set of resources. A billing account includes a payment instrument, to which costs are charged, and access control that is established by Cloud Platform Identity and Access Management (IAM) roles.

A billing account can be linked to one or more projects. Project usage is charged to the linked billing account. Projects that are not linked to a billing account cannot use GCP services that aren't free.

# Billing to BigQuery

Tools for monitoring, analyzing and optimizing cost have become an important part of managing development. Billing export to BigQuery enables you to export your daily usage and cost estimates automatically throughout the day to a BigQuery dataset you specify. You can then access your billing data from BigQuery. You can also use this export method to export data to a JSON file.

Regular file export to CSV and JSON is also available. However, if you use regular file export, you should be aware that regular file export captures a smaller dataset than export to BigQuery. For more information about regular file export and the data it captures, see Export Billing Data to a File.

# BigQuery, Cloud BigQuery

BigQuery is Google's serverless, highly scalable, enterprise data warehouse designed to make all your data analysts productive at an unmatched price-performance. Because there is no infrastructure to manage, you can focus on analyzing data to find meaningful insights using familiar SQL without the need for a database administrator.

# BigQuery Data Transfer Service

The BigQuery Data Transfer Service automates data movement from SaaS applications to Google BigQuery on a scheduled, managed basis. Your analytics team can lay the foundation for a data warehouse without writing a single line of code. BigQuery Data Transfer Service initially supports Google application sources like Adwords, DoubleClick Campaign Manager, DoubleClick for Publishers and YouTube.

# Bigtable, Cloud Bigtable

Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, enabling you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.

Cloud Bigtable is exposed to applications through multiple client libraries, including a supported extension to the Apache HBase library for Java. As a result, it integrates with the existing Apache ecosystem of open-source Big Data software.

https://cloud.google.com/bigtable/docs/performance

# Blue Green deployment

Blue-green deployment is a technique that reduces downtime and risk by running two identical production environments called Blue and Green.

At any time, only one of the environments is live, with the live environment serving all production traffic. For this example, Blue is currently live and Green is idle.

# Bq Big Query cli

Command line for Big Query (load, export, query...). Command Reference
To examine the schema of a specific table, run

```
bq show projectId:datasetId.tableId
```

Use `bq help` to get detailed information about the bq command-line tool

```
bq help query
```

To run a query, run the command `bq query "sql_statement"`

```
bq query "SELECT word, SUM(word_count) as count FROM publicdata:samples.shakespeare
WHERE word CONTAINS 'raisin' GROUP BY word"
```

List

```
bq ls
```

Use the bq mk command to create a new dataset named babynames

```
bq mk babynames
```

The `bq load` command creates or updates a table and loads data in a single step

```
bq load babynames.names2010 yob2010.txt name:string,gender:string,count:integer
```

Run bq show to see the schema:

```
bq show
```

Run the `bq rm` command to remove the `babynames` dataset

```
bq rm -r babynames
```

Data Formats:

- Cloud Storage:
  - CSV
  - JSON (newline delimited only)
  - Avro
  - Parquet
  - ORC
  - Cloud Datastore exports

  - Cloud Firestore exports
- Readable data source (such as your local machine):
  - CSV
  - JSON (newline delimited only)
  - Avro
  - Parquet
  - ORC

# Billing API

You can configure Billing on Google Cloud Platform (GCP) in a variety of ways to meet different needs.
**GCP resources** are the fundamental components that make up all GCP services, such as Google Compute Engine virtual machines (VMs), Google Cloud Pub/Sub topics, Google Cloud Storage buckets, and so on. For billing and access control purposes, resources exist at the lowest level of a hierarchy that also includes projects and an organization.
**Projects:** All lower level resources are parented by projects, which are the middle layer in the hierarchy of resources. You can use projects to represent logical projects, teams, environments, or other collections that map to a business function or structure. Any given resource can only exist in one project.
An **organization** is the top of the hierarchy of resources. All resources that belong to an organization are grouped under the organization node, to provide insight into and access control over every resource in the organization.

For more information on projects and organizations, see the Cloud Resource Manager documentation.
A billing account can be linked to one or more projects. Project usage is charged to the linked billing account. Projects that are not linked to a billing account cannot use GCP services that aren't free.

# Bucket

Buckets are the basic containers that hold your data. Everything that you store in Cloud Storage must be contained in a bucket. You can use buckets to organize your data and control access to your data, but unlike directories and folders, you cannot nest buckets. Because there are limits to bucket creation and deletion, you should design your storage applications to favor intensive object operations and relatively few buckets operations.

Part of Cloud Storage.

# C

# Canary update

The Instance Group Updater feature allows you to perform canary updates, so that you can test your updates on a random subset of instances before fully committing to the update.

A canary update is an update that is applied to a partial number of instances in the instance group. Canary updates let you test new features or upgrades on a subset of instances, instead of rolling out a potentially disruptive update to all your instances. If an update is not going well, you only need to roll back a small number of instances, minimizing the disruption for your users. From the perspective of the server, a canary update is the same as a standard rolling update, except that the number of instances that should be updated is less than the total size of the instance group. Like a standard rolling update, a canary update is disruptive to the instances affected; that is, the affected instances are deleted and replaced by new VM instances during the update.

# Carrier Peering

Cloud Interconnect extends your on-premises network to Google's network through a highly available, low latency connection. You can use Google Cloud Interconnect - Dedicated (Dedicated Interconnect) to connect directly to Google or use Google Cloud Interconnect - Partner (Partner Interconnect) to connect to Google through a supported service provider.

**Direct Peering**: Google allows you to establish a direct peering connection between your business network and Google's. With this connection you will be able to exchange Internet traffic between your network and Google's at one of our broad-reaching Edge network locations. Direct peering with Google is done by exchanging BGP routes between Google and the peering entity. After a direct peering connection is in place, you can use it to reach all of Google's services including the full suite of Google Cloud Platform products. Carrier peering allows you to obtain enterprise-grade network services that connect your infrastructure to Google by using a service provider.

**Google Cloud Interconnect**: Cloud Interconnect offers enterprise-grade connections to Google Cloud Platform using Google Services for Dedicated Interconnect, Partner Interconnect and Cloud VPN. This solution allows you to directly connect your on-premises network to your Virtual Private Cloud.

**Carrier Peering:** When connecting to Google through a service provider, you can get connections with higher availability and lower latency, using one or more links. Work with your service provider to get the connection you need.

**CDN Interconnect** allows select CDN providers to establish direct interconnect links with Google's edge network at various locations.

# Case Studies Cloud Architect

Mountkirk Games → games, noSQL, backend + analytics

- **Backend:**
  transactional database user profiles and game state (current MySQL)
  Time series database service for future analysis
  hardened Linux distro (enhanced for security)
- **Analytics:**
  Process data on the fly or late because of slow mobile networks (Dataflow)
  queries to access at least 10 TB of historical data (BigQuery)

Dress4win → web-based company, wardrobe, advertising, e-commerce, referrals, and a freemium app

- **Current:** MySQL, Redis, Web Application servers, Apache Hadoop/Spark, RabbitMQ, Jenkins
- **Needs**:
  Scalability, security, optimize
  non-production environments
  automation framework for provisioning resources
  failover of the production environment to cloud
  CI/CD
  multiple private connections between the production data center and cloud environment

JencoMart → retailer (Amazon.com)

- **Current:** LAMP (Linux, Apache, MySQL and PHP), Oracle user profiles, PostgreSQL database, SAN
- **Needs**:
  Optimize for capacity during peak periods
  Migration → Modify applications for the cloud

Decrease latency in Asia + green

TerramEarth → SMART heavy equipment for the mining and agricultural industries (20M operative)

- **Current:** each equip get 120 fields/sec stored locally (accessed for analysis when a vehicle is serviced via maintenance port)
  200,000 cellular network, direct data collection → 9TB/day
  Single Datacenter Linux and Windows-based systems → gzip CSV files → DWH 3 weeks old
  Python

- **Needs**:
  stock replacement parts and reduce unplanned downtime Support the dealer network
  decrease latency increase security
  Use customer and equipment data to anticipate customer need

# CBT

The `cbt` tool is a command-line interface for performing several different operations on Cloud Bigtable. It is written in Go using the Go client library for Cloud Bigtable. Source code for the `cbt` tool is available in the GitHub repository GoogleCloudPlatform/google-cloud-go. The `cbt` tool is available as a Cloud SDK component.

# Cloudevents

CloudEvents is a specification for describing event data in a common way. CloudEvents seeks to dramatically simplify event declaration and delivery across services, platforms, and beyond!
CloudEvents is a new effort and it's still under active development. However, its working group has received a surprising amount of industry interest, ranging from major cloud providers to popular SaaS companies. The specification is now under the Cloud Native Computing Foundation.

# CDN, Cloud CDN

Google Cloud CDN leverages Google's globally distributed edge points of presence to accelerate content delivery for websites and applications served out of Google Compute Engine and Google Cloud Storage. Cloud CDN lowers network latency, offloads origins, and reduces serving costs. Once you've set up HTTP(S) Load Balancing, simply enable Cloud CDN with a single checkbox.

# Cloud Native

Cloud native is a term used to describe container-based environments. Cloud-native technologies are used to develop applications built with services packaged in containers, deployed as microservices and managed on elastic infrastructure through agile DevOps processes and continuous delivery workflows.
The Cloud Native Computing Foundation builds sustainable ecosystems and fosters communities to support the growth and health of cloud native open source software.

# Coldline

Nearline and Coldline offer ultra low-cost, **highly-durable, highly available archival storage**. Coldline is ideal for cold storage - data your business expects to touch less than once a year. For warmer storage, choose Nearline:

data you expect to access less than once a month, but possibly multiple times throughout the year. Both options are available across all GCP regions and provide **unparalleled sub-second access speeds with a consistent API**.

# Composer, Cloud

Workflow orchestration service that lets you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers. Built on Apache Airflow open source project and operated using the Python programming language. Built-in integration with BigQuery, Dataflow, Dataproc, Datastore, Cloud Storage, Pub/Sub, Cloud ML Engine.

# Compute Engine

Google Compute Engine delivers virtual machines running in Google's innovative data centers and worldwide fiber network. Compute Engine's tooling and workflow support enable scaling from single instances to global, load-balanced cloud computing. Compute Engine's VMs boot quickly, come with high-performance persistent and local disk options, and deliver consistent performance. Our virtual servers are available in many configurations, including predefined sizes, and options to create Custom Machine Types optimized for your specific needs. Flexible pricing and automatic sustained use discounts make Compute Engine the leader in price/performance.

# Connection Draining

When **Connection Draining** is enabled, Auto Scaling will wait for outstanding requests to complete before terminating instances

# Console

Use the Cloud Platform Console to run your application on Google's infrastructure. You can provision, configure, and manage Google Cloud Platform products and Developer APIs.

# Container Builder / Cloud Build

Cloud Build is a service that executes your builds on Google Cloud Platform's infrastructure.

Cloud Build can import source code from a variety of repositories or cloud storage spaces, execute a build to your specifications, and produce artifacts such as Docker containers or Java archives.

You can write a build config to provide instructions to Cloud Build on what tasks to perform. You can configure builds to fetch dependencies, run unit tests, static analyses, and integration tests, and create artifacts with build tools such as docker, gradle, maven, bazel, and gulp.

Cloud Build executes your build as a series of build steps, where each build step is run in a Docker container. Executing build steps is analogous to executing commands in a script.

1. Prepare your application code and any needed assets.
2. Create a build config file in YAML or JSON format, which contains instructions for Cloud Build.
3. Submit the build to Cloud Build.
4. Cloud Build executes your build based on the build config you provided.
5. If applicable, any built images are pushed to Container Registry.

https://cloud.google.com/cloud-build/docs/quickstart-docker

# Container Registry

Container Registry is a private container image registry that runs on Google Cloud Platform. Container Registry supports Docker Image Manifest V2 and OCI image formats.

Many people use Dockerhub as a central registry for storing public Docker images, but to control access to your images you need to use a private registry such as Container Registry.

You can access Container Registry through secure HTTPS endpoints, which allow you to push, pull, and manage images from any system, VM instance, or your own hardware. Additionally, you can use the Docker credential helper command-line tool to configure Docker to authenticate directly with Container Registry.

# D

# Data Catalog

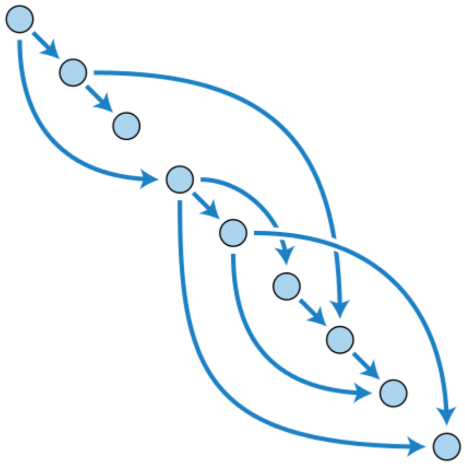
GCP metadata management service to discover, understand, and manage their data. BigQuery, Storage, Pub/Sub direct integration.
Search interface for data discovery, a flexible and powerful cataloging system for capturing both technical and business metadata, and a strong security and compliance foundation with Cloud Data Loss Prevention (DLP) and Cloud Identity and Access Management (IAM) integrations

# Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data.

# Daemon

In multitasking computer operating systems, a **daemon** is a computer program that runs as a background process, rather than being under the direct control of an interactive user. For example, syslogd is the daemon that implements the system logging facility, and sshd is a daemon that serves incoming SSH connections.



# DAG Directed acyclic graph

A DAG Directed acyclic graph is a finite directed graph with no directed cycles. That is, it consists of finitely many vertices and edges, with each edge directed from one vertex to another, such that there is no way to start at any vertex $v$ and follow a consistently-directed sequence of edges that eventually loops back to $v$ again. Equivalently, a DAG is a directed graph that has a topological ordering, a sequence of the vertices such that every edge is directed from earlier to later in the sequence.

# Data Loss Prevention API

The Google Data Loss Prevention API helps you understand and manage sensitive data. It provides fast, scalable classification and optional redaction for sensitive data elements like credit card numbers, names, social security numbers, passport numbers, US and selected international driver's license numbers, phone numbers, and more.

# Data Pipeline

In computing, a pipeline, also known as a data pipeline, is a set of data processing elements connected in series, where the output of one element is the input of the next one. The elements of a pipeline are often executed in parallel or in time-sliced fashion.

# Data Studio

Business Intelligence tools.

# Dataflow, Cloud Dataflow

Google Cloud Dataflow is a fully managed service for strongly consistent, parallel data-processing pipelines. It provides an SDK for Java with composable primitives for building data-processing pipelines for batch or continuous processing. This service manages the life cycle of Google Compute Engine resources of the processing pipeline(s). It also provides a monitoring user interface for understanding pipeline health.

# Datalab, Cloud Datalab

Google Cloud Datalab is an interactive tool for exploration, transformation, analysis and visualization of your data on Google Cloud Platform. It runs in your cloud project and enables you to write code to use other Big Data and storage services using a rich set of Google-authored and third party libraries.
demo

# Dataprep, Cloud Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.
demo

# Dataproc, Cloud Dataproc

Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing. It provides management, integration, and development tools for unlocking the power of rich open source data processing tools. With Cloud Dataproc, you can create Spark/Hadoop clusters sized for your workloads precisely when you need them.

# Datastore, Cloud Datastore

Google Cloud Datastore is a fully managed, schemaless, non-relational datastore [Mongo, DynamoDB like]. It provides a rich set of query capabilities, supports atomic transactions, and automatically scales up and down in response to load. It can scale to support an application with 1,000 users or 10 million users with no code changes.

Transactions
Entity groups

An entity group consists of a root entity and all of its descendants. Applications typically use entity groups to organize highly related data. For example, an application could use an entity group to store data about one product, or one user profile. For information about consistency levels and performance considerations when you use entity groups, see Transactions and entity groups

Strong Consistency on Reading Entity Values and Indexes
In Cloud Datastore, there are only two APIs that provide a strongly consistent view for reading entity values and indexes: (1) the lookup by key method and (2) the ancestor query. If application logic requires strong consistency, then the developer should use one of these methods to read entities from Cloud Datastore.

# Deployment Manager

Google Cloud Deployment Manager allows you to specify all the resources needed for your application in a declarative format using yaml. You can also use Python or Jinja2 templates to parameterize the configuration and allow reuse of common deployment paradigms such as a load balanced, auto-scaled instance group. Treat your configuration as code and perform repeatable deployments.
A configuration describes all the resources you want for a single deployment. A configuration is a file written in YAML syntax that lists each of the resources you want to create and its respective resource properties. A configuration must contain a resources: section followed by the list of resources to create.

# Dialogflow Enterprise Edition

**Dialogflow Enterprise Edition** is a development suite for voice and text conversational apps including chatbots. Dialogflow is cross-platform and can connect to your own apps (on the web, Android, iOS, and IoT) or existing platforms (e.g., Actions on Google, Facebook Messenger, Slack). Dialogflow Enterprise Edition is the paid enterprise tier of Dialogflow provided under the Google Cloud Platform Terms of Service. The free tier of Dialogflow (Dialogflow Standard Edition) is not offered via the Google Cloud Platform Terms of Service and is provided under the Dialogflow Standard Edition Terms of Service.

# Docker Containers

Software containers are a convenient way to run your applications in multiple isolated user-space instances. You can run containers on either Linux or Windows Server 2016 public VM images. Containers allow your applications to run with fewer dependencies on the host virtual machine and run independently from other containerized applications that you deploy to the same virtual machine instance. These characteristics make containerized applications more portable, easier to deploy, and easier to maintain at scale.

Docker and rkt are two popular container technologies that allow you to easily run containerized applications.

# DNS, Cloud DNS

Google Cloud DNS is a high-performance, resilient, global Domain Name System (DNS) service that publishes your domain names to the global DNS in a cost-effective way.

DNS is a hierarchical distributed database that lets you store IP addresses and other data, and look them up by name. Google Cloud DNS lets you publish your zones and records in the DNS without the burden of managing your own DNS servers and software. RESTful API to publish and manage DNS records for your applications and services.

# DNS Load Balancing

DNS load balancing is the practice of configuring a domain in the Domain Name System (DNS) such that client requests to the domain are distributed across a group of server machines.
DNS load balancing relies on the fact that most clients use the first IP address they receive for a domain. In most Linux distributions, DNS by default sends the list of IP addresses in a different order each time it responds to a new client, using the round-robin method. As a result, different clients direct their requests to different servers, effectively distributing the load across the server group.

Unfortunately, this simple implementation of DNS load balancing has inherent problems that limit its reliability and efficiency. Most significantly, DNS does not check for server or network outages or errors, and so always returns the same set of IP addresses for a domain even if servers are down or inaccessible.

Another issue arises because resolved addresses are usually cached, by both intermediate DNS servers (called resolvers) and clients, to improve performance and reduce the amount of DNS traffic on the network. Each resolved address is assigned a validity lifetime (called its time-to-live, or TTL), but long lifetimes mean that clients might not learn about changes to the group of servers in a timely fashion, and short lifetimes improve accuracy but lead to the increased processing and DNS traffic that caching is meant to mitigate in the first place.

This is for normal DNS Services. Cloud DNS Services improve that.

**GCP Cloud DNS** uses anycast to serve your managed zones from multiple locations around the world for high availability. Requests are automatically routed to the nearest location, reducing latency and improving authoritative name lookup performance for your users.

**Amazon Route 53** let choose a routing policy:

- Simple routing policy – Use for a single resource that performs a given function for your domain, for example, a web server that serves content for the example.com website.
- Failover routing policy – Use when you want to configure active-passive failover.
- Geolocation routing policy – Use when you want to route traffic based on the location of your users.
- Geoproximity routing policy – Use when you want to route traffic based on the location of your resources and, optionally, shift traffic from resources in one location to resources in another.
- Latency routing policy – Use when you have resources in multiple AWS Regions and you want to route traffic to the region that provides the best latency.
- Multivalue answer routing policy – Use when you want Route 53 to respond to DNS queries with up to eight healthy records selected at random.
- Weighted routing policy – Use to route traffic to multiple resources in proportions that you specify.

**Azure Traffic Manager** let choose a routing policy too:

- Priority: Select Priority when you want to use a primary service endpoint for all traffic, and provide backups in case the primary or the backup endpoints are unavailable.
- Weighted: Select Weighted when you want to distribute traffic across a set of endpoints, either evenly or according to weights, which you define.
- Performance: Select Performance when you have endpoints in different geographic locations and you want end users to use the "closest" endpoint in terms of the lowest network latency.
- Geographic: Select Geographic so that users are directed to specific endpoints (Azure, External, or Nested) based on which geographic location their DNS query originates from. This empowers Traffic Manager customers to enable scenarios where knowing a user's geographic region and routing them based on that is important. Examples include complying with data sovereignty mandates, localization of content & user experience and measuring traffic from different regions.
- Multivalue: Select MultiValue for Traffic Manager profiles that can only have IPv4/IPv6 addresses as endpoints. When a query is received for this profile, all healthy endpoints are returned.
- Subnet: Select Subnet traffic-routing method to map sets of end-user IP address ranges to a specific endpoint within a Traffic Manager profile. When a request is received, the endpoint returned will be the one mapped for that request's source IP address.

# E

## ECMP

Equal-cost multi-path routing (ECMP) is a routing strategy where next-hop packet forwarding to a single destination can occur over multiple "best paths" which tie for top place in routing metric calculations. Multi-path routing can be used in conjunction with most routing protocols, because it is a per-hop decision limited to a single router. It can substantially increase bandwidth by load-balancing traffic over multiple paths; however, there may be significant problems in deploying it in practice

## Encryption at Rest

- Google Cloud Platform encrypts customer content stored at rest, without any action required from the customer, using one or more encryption mechanisms. There are some minor exceptions, noted further in this document.
- Data for storage is split into chunks, and each chunk is encrypted with a unique data encryption key. These data encryption keys are stored with the data, encrypted with ("wrapped" by) key encryption keys that are exclusively stored and used inside Google's central Key Management Service. Google's Key Management Service is redundant and globally distributed.
- Data stored in Google Cloud Platform is encrypted at the storage level using either AES256 or AES128.
- Google uses a common cryptographic library, Tink, to implement encryption consistently across almost all Google Cloud Platform products. Because this common library is widely accessible, only a small team of cryptographers needs to properly implement and maintain this tightly controlled and reviewed code.

## Encryption in Transit

- Google encrypts and authenticates all data in transit at one or more network layers when data moves outside physical boundaries not controlled by Google or on behalf of Google. Data in transit inside a physical boundary controlled by or on behalf of Google is generally authenticated but not necessarily encrypted.
- Depending on the connection that is being made, Google applies default protections to data in transit. For example, we secure communications between the user and the Google Front End (GFE) using TLS.
- Google Cloud customers with additional requirements for encryption of data over WAN can choose to implement further protections for data as it moves from a user to an application, or virtual machine to virtual machine. These protections include IPsec tunnels, Gmail S/MIME, managed SSL certificates, and Istio.

## Encryption Keys

Cloud Storage always encrypts your data on the server side, before it is written to disk, at no additional charge. Besides this standard behavior, there are additional ways to encrypt your data when using Cloud Storage:

- *Customer-supplied encryption keys*: You can create and manage your own encryption keys for server-side encryption, which act as an additional encryption layer on top of the standard Cloud Storage encryption.
- *Customer-managed encryption keys*: You can generate and manage your encryption keys using Cloud Key Management Service, which act as an additional encryption layer on top of the standard Cloud Storage encryption.Your encryption keys are stored within Cloud KMS. The project that holds your encryption keys can then be independent from the project that contains your buckets, thus allowing for better separation of duties.

  When you apply a customer-managed encryption key to an object, Cloud Storage uses the key when encrypting:

    - The object's data.
    - The object's CRC32C checksum.
    - The object's MD5 hash.

    Cloud Storage uses standard server-side keys to encrypt the remaining metadata for the object, including the object's name. This allows you to read and update general metadata, as well as list and delete objects, without needing the customer-managed encryption key. However, to perform any of these actions, you must have sufficient permission to do so.

- *Client-side encryption*: encryption that occurs before data is sent to Cloud Storage. Such data arrives at Cloud Storage already encrypted but also undergoes server-side encryption.

# Endpoint

An endpoint is an individual IP/port pair fronting a service that can handle requests. Any given service can have zero or more endpoints. The endpoint may be a VM, container, load balancer, or other entity capable of handling the requests. An example would be a cluster of user-managed Redis servers.

Endpoints can have optional metadata, in the form of key:value pairs, that can be used by clients. For example, a Redis service may have metadata like replica:server. Metadata can be used to store URLs. You can use a service such as jq to pull the new URL out of the response.

# Endpoints Cloud Endpoints

Google Cloud Endpoints is a tool that helps you to develop, deploy, secure and monitor your APIs running on Google Cloud Platform.

# Environ (Anthos)

Environs are a Google Cloud concept for logically organizing clusters and other resources, letting you use and manage multi-cluster capabilities and apply consistent policies across your systems.

# Egress traffic

Egress traffic is network traffic that begins inside of a network and proceeds through its routers to a destination somewhere outside of the network. For example, an email message that is considered egress traffic will travel from a user's workstation and pass through the enterprise's LAN routers before it is delivered to the Internet to travel to its final destination.

# F

# Failover

Failover is switching to a redundant or standby computer server, system, hardware component or network upon the failure or abnormal termination of the previously active application,[1] server, system, hardware component, or network. Failover and switchover are essentially the same operation, except that failover is automatic and usually operates without warning, while switchover requires human intervention.

# Filestore

Cloud Filestore is a scalable and highly available shared file service fully managed by Google (disk multiple Vms like AWS EFS). Cloud Filestore provides persistent storage ideal for shared workloads. It is best suited for enterprise applications requiring persistent, durable, shared storage which is accessed by NFS or requires a POSIX compliant file system.

# Firestore

Cloud Firestore is a NoSQL document database for storing, syncing, and querying data for mobile and web apps. Its client libraries provide live synchronization and offline support, while its security features and integrations with Firebase and Google Cloud Platform accelerate building serverless apps.

Cloud **Firestore** is the next major version of Cloud **Datastore** and a re- branding of the product. Taking the best of Cloud **Datastore** and the Firebase Realtime Database, Cloud **Firestore** is a NoSQL document database built for automatic scaling, high performance, and ease of application development.

# Firewall Rule

Google Cloud Platform (GCP) firewall rules let you allow or deny traffic to and from your virtual machine (VM) instances based on a configuration you specify. GCP firewall rules are applied at the virtual networking level, so they provide effective protection and traffic control regardless of the operating system your instances use.

Every VPC network functions as a distributed firewall. While firewall rules are defined at the network level, connections are allowed or denied on a per-instance basis. You can think of the GCP firewall rules as existing not only between your instances and other networks, but between individual instances within the same network.

| Priority | Direction | Action | Enforcement | Target | Source | Destination | Protocols, Ports |
|---|---|---|---|---|---|---|---|
| Integer from 0 to 65535, inclusive; default 1000. | ingress | Either allowor deny. | Either enabled(default) or disabled. | Instances receiving traffic from the source.<br>One of the following:<br>• All instances in the VPC network<br>• Instances by service account<br>• Instances by network tag | One of the following:<br>• Range of IPv4 addresses; default is any (0.0.0.0/0)<br>• Instances by service account<br>• Instances by network tag | Destination is not specified separately for ingressrules. The target defines the destination. | Specify a protocol or protocol and a port.<br>If not set, the rule applies to all protocols. |
| Integer from 0 to 65535, inclusive; default 1000. | egress | Either allowor deny. | Either enabled(default) or disabled. | Instances sending traffic to the destination.<br>One of the following:<br>• All instances in the VPC network<br>• Instances by service account<br>• Instances by network tag | Source is not specified separately for egress rules. The target defines the source. | Any network or a specific range of IPv4 addresses; default is any (0.0.0.0/0). | Specify a protocol or protocol and a port.<br>If not set, the rule applies to all protocols. |

https://cloud.google.com/vpc/docs/firewalls#ingress_cases

https://cloud.google.com/vpc/docs/firewalls#egress_cases

# Forwarding Rule

[**Load Balancers**] Forwarding Rules map the IP address for your load balancer to the Target Proxy that will handle the requests. First we will need to create our IP address though. We will need a global, rather than regional, IP address for our HTTP load balancer.

```
gcloud compute addresses create my-address --global
gcloud compute forwarding-rules create my-https-forwarding-rule --global
--address 123.123.123.123 --ip-protocol TCP --port-range 443
--target-https-proxy my-https-proxy
```

# Cloud Functions

Google Cloud Functions is a serverless execution environment for building and connecting cloud services. With Cloud Functions you write simple, single-purpose functions that are attached to events emitted from your cloud infrastructure and services. Your function is triggered when an event being watched is fired. Your code executes in a fully managed environment. There is no need to provision any infrastructure or worry about managing any servers.

Cloud Functions provides a connective layer of logic that lets you write code to connect and extend cloud services. Listen and respond to a file upload to Cloud Storage, a log change, or an incoming message on a Cloud Pub/Sub topic. Cloud Functions augments existing cloud services and allows you to address an increasing number of use cases with arbitrary programming logic. Cloud Functions have access to the Google Service Account credential and are thus seamlessly authenticated with the majority of Google Cloud Platform services, including Cloud Vision API, as well as many others. In addition, Cloud Functions are supported by numerous Google Cloud client libraries, which further simplify these integrations.
For more information on creating triggers and associating them with your functions, see Events and Triggers

## G

# gcloud Reference

The Cloud SDK is a set of tools for Cloud Platform. It contains gcloud, gsutil, and bq, which you can use to access Google Compute Engine, Google Cloud Storage, Google BigQuery, and other products and services from the command-line. You can run these tools interactively or in your automated scripts.
In addition to running gcloud commands from the command line, you can also run them from scripts or other automations — for example, when using Jenkins to drive automation of Google Cloud Platform tasks.

```
gcloud logging read

gcloud app appengine

gcloud auth - manage oauth2 credentials for the Google Cloud SDK

gcloud bigtable - manage your Cloud Bigtable storage

gcloud builds - create and manage builds for Google Cloud Build

gcloud components - list, install, update, or remove Google Cloud SDK components

gcloud composer - create and manage Cloud Composer Environments
Cloud Composer is a managed Apache Airflow service that helps you create, schedule,
monitor and manage workflows
```

```
gcloud compute - create and manipulate Google Compute Engine resources

gcloud config - view and edit Cloud SDK properties

gcloud container - deploy and manage clusters of machines for running containers

gcloud dataflow manage Google Cloud Dataflow jobs

gcloud dataproc - create and manage Google Cloud Dataproc clusters and jobs

gcloud datastore - manage your Cloud Datastore indexes

gcloud debug - commands for interacting with the Cloud Debugger

gcloud deployment-manager - manage deployments of cloud resources

gcloud dns - manage your Cloud DNS managed-zones and record-sets

gcloud docker - enable Docker CLI access to Google Container Registry

gcloud domains - manage domains for your Google Cloud projects (custom domains)

gcloud endpoints - create, enable and manage API services

gcloud firebase - work with Google Firebase

gcloud functions - manage Google Cloud Functions

gcloud iam - manage IAM service accounts and keys

gcloud iot - manage Cloud IoT resources

gcloud kms - manage cryptographic keys in the cloud

gcloud logging read

gcloud ml - use Google Cloud machine learning capabilities (vision speech..)

gcloud ml-engine - manage Cloud ML Engine jobs and models

gcloud organizations - create and manage Google Cloud Platform Organizations

gcloud projects - create and manage project access policies

gcloud pubsub - manage Cloud Pub/Sub topics and subscriptions

gcloud redis - manage Cloud Memorystore Redis resources

gcloud services - list, enable and disable APIs and services

gcloud source - cloud git repository commands

gcloud spanner - command groups for Cloud Spanner

gcloud sql - create and manage Google Cloud SQL databases

gcloud topic - gcloud supplementary help
gcloud version - print version information for Cloud SDK components
```

## Google Front Ends (GFEs)

Software-defined, distributed systems that are located in Google POPs and perform global load balancing in conjunction with other systems and control planes

# Google Transfer Appliance

Transfer Appliance is a rackable high capacity storage server that you set up in your datacenter. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage. Choose from 100TB or 480TB's of raw capacity per appliance to move your data to Google Cloud quickly.

# Gradle App Engine Plugin

This Gradle plugin provides tasks to build and deploy Google App Engine applications

- Using Gradle and the App Engine Plugin (standard environment)
- Using Gradle and the App Engine Plugin (flexible environment)

# gsutil

gsutil is a Python application that lets you access Cloud Storage from the command line. You can use gsutil to do a wide range of bucket and object management tasks, including:

- Creating and deleting buckets.
- Uploading, downloading, and deleting objects.
- Listing buckets and objects.
- Moving, copying, and renaming objects.
- Editing object and bucket ACLs.

For a complete list of guides to completing tasks with gsutil, see Cloud Storage How-to Guides.

gsutil Quickstart shows you how to set up a Google Cloud Platform project, enable billing, install gsutil, and run basic commands with the tool.

```
gsutil acl set  get ch

gsutil cat -h gs://bucket/meeting_notes/2012_Feb/*.txt

//Concatenate a sequence of objects into a new composite object
gsutil compose obj1 obj2

//config - Obtain credentials and create configuration file
gsutil config -f   --Create a token with full-control access for storage resources:

//cors - Get or set a CORS JSON document for one or more buckets

cp - Copy files and objects

defacl - Get, set, or change default ACL on buckets
```

```
defstorageclass - Get or set the default storage class on buckets

du - Display object size usage

hash - Calculate file hashes

iam - Get, set, or change bucket and/or object IAM permissions

kms - Configure Cloud KMS encryption

label - Get, set, or change the label configuration of a bucket

lifecycle - Get or set lifecycle configuration for a bucket

logging - Configure or retrieve logging on buckets

ls - List providers, buckets, or objects
gsutil ls -l gs://bucket/*.txt

mb - Make buckets

rsync - Synchronize content of two buckets/directories

setmeta - Set metadata on already uploaded objects

signurl - Create a signed url - es uploading a plain text file via HTTP PUT
gsutil signurl -m PUT -d 1h -c text/plain <private-key-file> gs://<bucket>/<obj>

stat - Display object status

test - Run gsutil unit/integration tests (for developers)

update - Update to the latest gsutil release

version - Print version info about gsutil

versioning - Enable or suspend versioning for one or more buckets

web - Set a main page and/or error page for one or more buckets
```

# H

## Hardware Security Module

Google Cloud Hardware Security Module is a cloud-hosted key management service that lets you protect encryption keys and perform cryptographic operations within a managed HSM service. You can generate, use, rotate, and destroy various symmetric and asymmetric keys.

## HASH

A hash function is any function that can be used to map data of arbitrary size to fixed-size values. The values returned by a hash function are called hash values, hash codes, digests, or simply hashes. The values are used to index a fixed-size table called a hash table.

A good hash function satisfies two basic properties: 1) it should be very fast to compute; 2) it should minimize duplication of output values (collisions).
Hash functions are also used to build caches for large data sets stored in slow media. A cache is generally simpler than a hashed search table, since any collision can be resolved by discarding or writing back the older of the two colliding items

# HBase – Apache HBase

Apache HBase is an open-source, distributed, versioned, non-relational database modeled after Google's Bigtable: A Distributed Storage System for Structured Data by Chang et al. Just as Bigtable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

https://www.tutorialspoint.com/hbase/hbase_quick_guide.htm

# Hadoop

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation

# Health-checks

A health checker polls instances at specified intervals. Instances that do not respond successfully to a specified number of consecutive probes are marked as `UNHEALTHY`. No new connections are sent to such instances, though existing connections are allowed to continue. The health checker continues to poll unhealthy instances. If an instance later responds successfully to a specified number of consecutive probes, it is marked `HEALTHY` again and can receive new connections.

- HTTP health checks
- HTTPS health checks
- HTTP/2 health checks
- TCP health checks
- SSL (TLS) health checks

# Helm

Helm is a tool for managing packages of pre-configured Kubernetes resources.
It creates objects called CHARTS for K8s configuration with variables and automatic deploying and management. Useful for automation and smart duplication of deployments
Here GCP Example
Use Helm to:

- Find and use popular software packaged as Helm charts to run in Kubernetes
- Share your own applications as Helm charts
- Create reproducible builds of your Kubernetes applications
- Intelligently manage your Kubernetes manifest files
- Manage releases of Helm packages

# Hive

The **Apache Hive™** data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage and queried using SQL syntax.

Built on top of **Apache Hadoop™**, Hive provides the following features:

- Tools to enable easy access to data via SQL, thus enabling data warehousing tasks such as extract/transform/load (ETL), reporting, and data analysis.
- A mechanism to impose structure on a variety of data formats
- Access to files stored either directly in **Apache HDFS™** or in other data storage systems such as **Apache HBase™**
- Query execution via Apache Tez™, Apache Spark™, or MapReduce
- Procedural language with HPL-SQL
- Sub-second query retrieval via Hive LLAP, Apache YARN and Apache Slider.

Hive provides standard SQL functionality, including many of the later SQL:2003 and SQL:2011 features for analytics.

Hive's SQL can also be extended with user code via user defined functions (UDFs), user defined aggregates (UDAFs), and user defined table functions (UDTFs).

# I

# IAM

**Cloud Identity & Access Management (Cloud IAM)** provides administrators the ability to manage cloud resources centrally by controlling who can take what action on specific resources.

In Cloud IAM, you grant access to **members**. Members can be of following types:

- Google account
- Service account
- Google group
- G Suite domain
- Cloud Identity domain

You cannot assign a permission to the user directly; instead you grant them a role. When you grant a role to a user, you grant them all the permissions that the role contains.

# Identity-Aware Proxy

**Google Cloud Identity-Aware Proxy** is a tool that helps control access, based on a user's identity and group membership, to applications running on Google Cloud Platform

# IoT Core

Google Cloud IoT Core is a fully managed service that allows you to easily and securely connect, manage, and ingest data from internet connected devices. It permits utilization of other Google Cloud services for collecting, processing, analyzing, and visualizing IoT data in real time.
See IoT Solution

# Images

Use operating system images to create boot disks for your instances. You can use one of the following image types:

- **Public images** are provided and maintained by Google, open-source communities, and third-party vendors. By default, all projects have access to these images and can use them to create instances.
- **Custom images** are available only to your project. You can create a custom image from on-prem environment, import virtual disks boot disks and other images. Then, use the custom image to create an instance. Copy one image to another image using either the `gcloud` tool

You can use most public images at no additional cost, but there are some premium images that do add additional cost to your instances. Custom images that you import to Compute Engine add no cost to your instances, but do incur an image storage charge while you keep your custom image in your project.

# Impala Apache

Apache Impala is the open source, native analytic database for Apache Hadoop. Impala is shipped by Cloudera, MapR, Oracle, and Amazon.

Impala raises the bar for SQL query performance on Apache Hadoop while retaining a familiar user experience. With Impala, you can query data, whether stored in HDFS or Apache HBase – including SELECT, JOIN, and aggregate functions – in real time. Furthermore, Impala uses the same metadata, SQL syntax (Hive SQL), ODBC driver, and user interface (Hue Beeswax) as Apache Hive, providing a familiar and unified platform for batch-oriented or real-time queries. (For that reason, Hive users can utilize Impala with little setup overhead.)

# Ingress traffic

Ingress traffic is network traffic that originates from outside of the network's routers and proceeds toward a destination inside of the network. For example, an email message that is considered ingress traffic will originate somewhere outside of a enterprise's LAN, pass over the Internet and enter the company's LAN before it is delivered to the recipient.

# Instance Groups

You can create and manage groups of virtual machine (VM) instances so that you don't have to individually control each instance in your project. Compute Engine offers two different types of instance groups: **managed** and **unmanaged** instance groups.
**A managed instance group** uses an instance template to create a group of identical instances. You control a managed instance group as a single entity.

- A zonal managed instance group, which contains instances from the same zone.
- A regional managed instance group, which contains instances from multiple zones across the same region.

**Unmanaged instance groups** are groups of dissimilar instances that you can arbitrarily add and remove from the group. Unmanaged instance groups do not offer autoscaling, rolling update support, or the use of instance templates so Google recommends creating managed instance groups whenever possible. Use unmanaged instance groups only if you need to apply load balancing to your pre-existing configurations or to groups of dissimilar instances.

If you must create a group of dissimilar instances that do not follow an instance template, see Unmanaged Instance Groups.

## Internet Key Exchange (IKE)

Internet Key Exchange (IKE, sometimes IKEv1 or IKEv2, depending on version) is the protocol used to set up a security association (SA) in the IPsecprotocol suite. IKE builds upon the Oakley protocol and ISAKMP.IKE uses X.509 certificates for authentication – either pre-shared or distributed using DNS(preferably with DNSSEC) – and a Diffie–Hellman key exchange to set up a shared session secret from which cryptographic keys are derived.

**In other words are the techniques for Keys exchange in IPSEC…**remember only that they need a shared session secret

# IPsec VPN

In computing, Internet Protocol Security (IPsec) is a secure network protocol suite that authenticates and encrypts the packets of data sent over an IPv4 network. The initial IPv4 suite was developed with so few security provisions that the IP version was incomplete, open or left for further research development. IPsec includes protocols for establishing mutual authentication between agents at the beginning of a session and negotiation of cryptographic keys to use during the session. IPsec can protect data flows between a pair of hosts (host-to-host), between a pair of security gateways (network-to-network), or between a security gateway and a host (network-to-host).
**A VPN tunnel connects two VPN gateways** and serves as a virtual medium through which encrypted traffic is passed. Two VPN tunnels must be established to create a connection between two VPN gateways: Each tunnel defines the connection from the perspective of its gateway, and traffic can only pass once the pair of tunnels is established. A Cloud VPN tunnel is always associated with a specific Cloud VPN gateway resource.

# ISTIO

Istio makes it easy to create a network of deployed services with load balancing, service-to-service authentication, monitoring, and more, with few or no code changes in service code. You add Istio support to services by deploying a special sidecar proxy throughout your environment that intercepts all network communication between microservices, then configure and manage Istio using its control plane functionality, which includes:

- Automatic load balancing for HTTP, gRPC, WebSocket, and TCP traffic.
- Fine-grained control of traffic behavior with rich routing rules, retries, failovers, and fault injection.
- A pluggable policy layer and configuration API supporting access controls, rate limits and quotas.
- Automatic metrics, logs, and traces for all traffic within a cluster, including cluster ingress and egress.
- Secure service-to-service communication in a cluster with strong identity-based authentication and authorization.

The main component of ISTIO are:
Pilot
Citadel
Mixer

# J

# Jupyter notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

# Jenkins on GCP

Jenkins is a self-contained, open source automation server which can be used to automate all sorts of tasks related to building, testing, and delivering or deploying software.

Jenkins can be installed through native system packages, Docker, or even run standalone by any machine with a Java Runtime Environment (JRE) installed.

https://jenkins.io/

# JWT - JSON Web Tokens

JWT Handbook
JSON Web Token (JWT) is an open standard (RFC 7519) that defines a compact and self-contained way for securely transmitting information between parties as a JSON object. This information can be verified and trusted because it is digitally signed. JWTs can be signed using a secret (with the HMAC algorithm) or a public/private key pair using RSA or ECDSA.

Although JWTs can be encrypted to also provide secrecy between parties, we will focus on signed tokens. Signed tokens can verify the integrity of the claims contained within it, while encrypted tokens hide those claims from other parties. When tokens are signed using public/private key pairs, the signature also certifies that only the party holding the private key is the one that signed it.

## K

# K3s

See Kubernetes small

# Kafka

Apache Kafka is a distributed streaming platform. A streaming platform has three key capabilities:

- Publish and subscribe to streams of records, similar to a message queue or enterprise messaging system.
- Store streams of records in a fault-tolerant durable way.
- Process streams of records as they occur.
- Kafka is generally used for two broad classes of applications:
  - Building real-time streaming data pipelines that reliably get data between systems or applications
  - Building real-time streaming applications that transform or react to the streams of data

Same as Pub/Sub, and GCP has Confluent Cloud on GCP. Confluent Cloud is a fully-managed streaming service based on Apache Kafka. Led by the creators of Kafka—Jay Kreps, Neha Narkhede and Jun Rao—Confluent provides enterprises with a real-time streaming platform built on a reliable, scalable ecosystem of products that place Kafka at their core

# Key Management Service

Cloud KMS is a cloud-hosted key management service that lets you manage cryptographic keys for your cloud services the same way you do on premises. You can generate, use, rotate, and destroy AES256, RSA 2048, RSA 3072, RSA 4096, EC P256, and EC P384 cryptographic keys. Cloud KMS is integrated with Cloud IAM and Cloud Audit Logging so that you can manage permissions on individual keys and monitor how these are used. Use Cloud KMS to protect secrets and other sensitive data that you need to store in Google Cloud Platform.

# Knative

Knative is an essential set of components to build and run serverless applications on Kubernetes. Knative offers features like scale-to-zero, autoscaling, in-cluster builds, and eventing framework for cloud-native applications on Kubernetes. Whether on-premises, in the cloud, or in a third-party data center, Knative codifies the best practices shared by successful real-world Kubernetes-based frameworks.
Main parts: Build Serving Eventing (based on Cloudevents)

# Kubeflow

Kubeflow ⟶ Kubernetes + Tensorflow ⟶ simple, portable and scalable Video

- Easy, repeatable, portable deployments on a diverse infrastructure (for example, experimenting on a laptop, then moving to an on-premises cluster or to the cloud)
- Deploying and managing loosely-coupled microservices
- Scaling based on demand

# Kubernetes

Kubernetes has grown into the most popular solution to manage containerized workloads anywhere. Providing automated container orchestration and efficient machine management, Kubernetes improves your reliability and reduces the time and resources attributed to DevOps.
Kubernetes Engine supports the common Docker container format.
With GKE On-Prem, Google will offer a consistent Kubernetes experience for your applications across on-premises and the cloud. Using GKE On-Prem, you get a reliable, efficient, and secured way to run Kubernetes clusters, anywhere

# Kubernetes Secrets

Kubernetes Secrets let you store and manage sensitive information, such as passwords, OAuth tokens, and ssh keys.
A Kubernetes secret is a simple object that's stored securely (e.g. encrypted at rest) by the orchestrator and can contain arbitrary data in key-value format.
The value is base64 encoded, so we can also store binary data like certificates. Kubernetes makes it easy to consume secrets by letting you simply mount them onto your container, either as env var — not recommended — or as a file.

# Kubernetes SMALL

**K3s**

K3s runs on any Linux distribution without any additional external dependencies or tools. It is marketed by Rancher as a lightweight Kubernetes offering suitable for edge environments, IoT devices, CI pipelines, and even ARM devices, like Raspberry Pi's. K3s achieves its lightweight goal by stripping a bunch of features out of the Kubernetes binaries (e.g. legacy, alpha, and cloud-provider-specific features), replacing docker with containerd, and using sqlite3 as the default DB (instead of etcd). As a result, this lightweight Kubernetes only consumes 512 MB of RAM and 200 MB of disk space. K3s has some nice features, like Helm Chart support out-of-the-box.

K3s can do multiple node Kubernetes cluster. However, due to technical limitations of SQLite, K3s currently does not support High Availability (HA), as in running multiple master nodes. The K3s team plans to address this in the future.

**Kind**

Kind (Kubernetes-in-Docker), as the name implies, runs Kubernetes clusters in Docker containers. This is the official tool used by Kubernetes maintainers for Kubernetes v1.11+ conformance testing. It supports multi-node clusters as well as HA clusters. Because it runs K8s in Docker, kind can run on Windows, Mac, and Linux.
Kind is optimized first and foremost for CI pipelines, so it may not have some of the developer-friendly features of other offerings.

**Desktop Docker**

Docker for Mac/Windows now ships with a bundled Kubernetes offering.
Kubernetes versions are tightly coupled with the Docker version (i.e. Docker stable channel ships with K8s v1.10. If you want K8s v1.13, you need to switch to Docker edge channel).
Not as easy to destroy and start a new K8s cluster. AFAIK, you would have to disable Kubernetes and re-enable it through the Docker desktop app preferences.

# Kubernetes Engine

Kubernetes Engine is a managed, production-ready environment for deploying containerized applications. Launched in 2015, Kubernetes Engine builds on Google's experience of running services like Gmail and YouTube in containers for over 12 years. Kubernetes Engine allows you to get up and running with Kubernetes in no time, by completely eliminating the need to install, manage, and operate your own Kubernetes clusters.
Google Kubernetes Engine is a powerful cluster manager and orchestration system for running your Docker containers. Kubernetes Engine schedules your containers into the cluster, keeps them healthy and manages them automatically based on requirements you define (such as CPU and memory). It's based on Kubernetes, the leading open-source container orchestration system giving you the flexibility to take advantage of on-premises, hybrid, or public cloud infrastructure.
Regional Service: https://cloud.google.com/kubernetes-engine/docs/concepts/regional-clusters

# L

# Lift and shift

Lift and shift is a strategy for moving an application or operation from one environment to another – without redesigning the app. In the lift-and-shift approach, certain workloads and tasks can be moved from on-premises storage to the cloud, or data operations might be transferred from one data center to another.

# Live Migration

Compute Engine offers live migration to keep your virtual machine instances running even when a host system event occurs, such as a software or hardware update. Compute Engine live migrates your running instances to another host in the same zone rather than requiring your VMs to be rebooted. This allows Google to perform maintenance that is integral to keeping infrastructure protected and reliable without interrupting any of your VMs.

Live migration keeps your instances running during:

- Regular infrastructure maintenance and upgrades.
- Network and power grid maintenance in the data centers.
- Failed hardware such as memory, CPU, network interface cards, disks, power, and so on. This is done on a best-effort basis; if a hardware fails completely or otherwise prevents live migration, the VM crashes and restarts automatically and a `hostError` is logged.
- Host OS and BIOS upgrades.
- Security-related updates, with the need to respond quickly.
- System configuration changes, including changing the size of the host root partition, for storage of the host image and packages.

Live migration does not change any attributes or properties of the VM itself. The live migration process just transfers a running VM from one host machine to another host machine within the same zone. All VM properties and attributes remain unchanged, including internal and external IP addresses, instance metadata, block storage data and volumes, OS and application state, network settings, network connections, and so on.

# Load Balancing

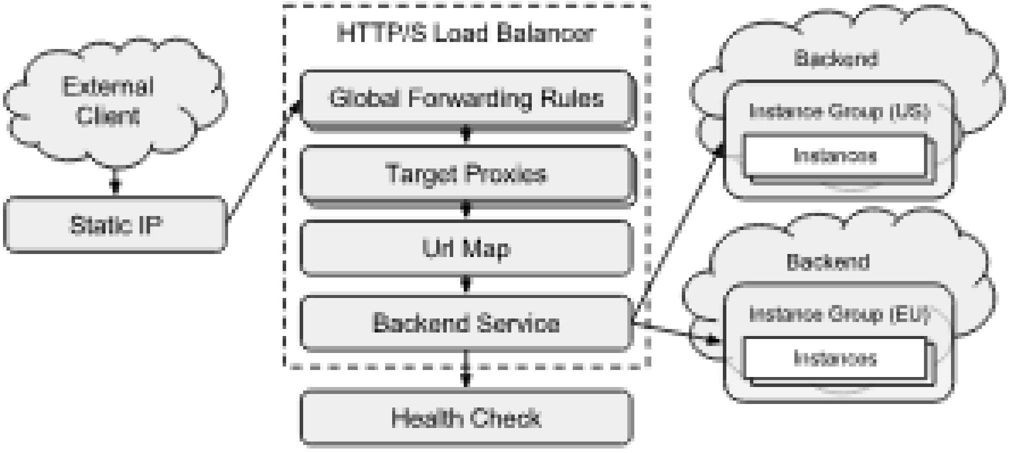
Google Cloud Platform Load Balancing gives you the ability to distribute load-balanced compute resources in single or multiple regions, to meet your high availability requirements, to put your resources behind a single anycast IP and to scale your resources up or down with intelligent Autoscaling. Cloud Load Balancing is fully integrated with Cloud CDN for optimal content delivery.
Using Cloud Load Balancing, you can serve content as close as possible to your users, on a system that can respond to over 1 million queries per second. Cloud Load Balancing is a fully distributed, software defined, managed service. It is not instance or device based, so you do not need to manage a physical load balancing infrastructure.
**Google global load balancing** is implemented entirely in software, done by Google Front Ends (GFEs). The GFEs are distributed globally and load balance traffic in sync with each other by working with Google's other software-defined systems and global control plane.
**Google regional load balancing** is implemented entirely in software. Your instances are in a single GCP region and traffic is distributed to instances within a single region.
https://www.ianlewis.org/en/google-cloud-platform-http-load-balancers-explaine
https://cloud.google.com/load-balancing/docs/https/adding-a-backend-bucket-to-content-based-load-balancing

Types of Load Balancers:

| Load balancer | Traffic type | Global/Regional | External/Internal | External Ports for Load Balancing |
|---|---|---|---|---|
| HTTP(S) | HTTP or HTTPS | Global | External | HTTP on 80 or 8080; HTTPS on 443 |
| SSL Proxy | TCP with SSL offload | Global | External | 25, 43, 110, 143, 195, 443, 465, 587, 700, 993, 995, 1883, and 5222 |
| TCP Proxy | TCP without SSL offload. Does not preserve client IP addresses | Global | External | 25, 43, 110, 143, 195, 443, 465, 587, 700, 993, 995, 1883, 5222 |
| Network TCP/UDP | TCP/UDP without SSL offload. Preserves client IP addresses. | Regional | External | Any |

|  |  |  |  |  |
|---|---|---|---|---|
| Internal TCP/UDP | TCP or UDP | Regional | Internal | Any |

# Load Balancing Http(s)



A complete HTTP load balancer is structured as follows:

1. A global forwarding rule directs incoming requests to a target HTTP proxy.
2. The target HTTP proxy checks each request against a URL map to determine the appropriate backend service for the request.
3. The backend service directs each request to an appropriate backend based on serving capacity, zone, and instance health of its attached backends. The health of each backend instance is verified using an HTTP health check, an HTTPS health check, or an HTTP/2 health check. If the backend service is configured to use an HTTPS or HTTP/2 health check, the request will be encrypted on its way to the backend instance.
4. Sessions between the load balancer and the instance can use the HTTP, HTTPS, or HTTP/2 protocol. If you use HTTPS or HTTP/2, each instance in the backend services must have an SSL certificate.

**HTTPS**

- An HTTPS load balancer uses a target HTTPS proxy instead of a target HTTP proxy.
- An HTTPS load balancer requires at least one signed SSL certificate installed on the target HTTPS proxy for the load balancer.
- The client SSL session terminates at the load balancer.
- HTTPS load balancers support the QUIC transport layer protocol.

# Load Balancing Internal

Internal Load Balancing enables you to run and scale your services behind a private load balancing IP address that is accessible only to instances internal to your Virtual Private Cloud (VPC).

# Load Balancing Network

Use Network Load Balancing to balance the load on your systems based on incoming IP protocol data, such as address, port, and protocol type.

Network Load Balancing uses forwarding rules that point to target pools, which list the instances available for load balancing and define which type of health check that should be performed on these instances. See Setting Up Network Load Balancing for more information.

Network Load Balancing is a regional, non-proxied load balancer. You can use it to load balance UDP traffic, and TCP and SSL traffic on ports that are not supported by the SSL proxy and TCP proxy load balancers.

A Network load balancer is a pass-through load balancer (direct server return(DSR), direct routing). It does not proxy connections from clients, that is (link):

- The IP packets are forwarded unmodified to the VM, there is no address or port translation
- The VM thinks that the load balancer IP is one of its own IPs
- Is fast

# Load Balancing SSL Proxy

Google Cloud SSL Proxy Load Balancing terminates user SSL (TLS) connections at the load balancing layer, then balances the connections across your instances using the SSL or TCP protocols. Cloud SSL proxy is intended for non-HTTP(S) traffic. For HTTP(S) traffic, HTTP(S) load balancing is recommended instead.

SSL Proxy Load Balancing supports both IPv4 and IPv6 addresses for client traffic. Client IPv6 requests are terminated at the load balancing layer, then proxied over IPv4 to your backends. Load balancing service that can be deployed globally. You can deploy your instances in multiple regions, and the load balancer automatically directs traffic to the closest region that has capacity.

# Load Balancing TCP Proxy

Google Cloud Platform (GCP) TCP Proxy Load Balancing allows you to use a single IP address for all users around the world. GCP TCP proxy load balancing automatically routes traffic to the instances that are closest to the user.

Note that global load balancing requires that you use the Premium Tier of Network Service Tiers, which is the default tier. Otherwise, load balancing is handled regionally.

Cloud TCP Proxy Load Balancing is intended for non-HTTP traffic. For HTTP traffic, HTTP Load Balancing is recommended instead. For proxied SSL traffic, use SSL Proxy Load Balancing.

TCP Proxy Load Balancing supports both IPv4 and IPv6 addresses for client traffic. Client IPv6 requests are terminated at the load balancing layer, then proxied over IPv4 to your backends.

# Local Disks

Local SSDs are physically attached to the server that hosts your VM instance. Local SSDs have higher throughput and lower latency than standard persistent disks or SSD persistent disks. The data that you store on a local SSD persists only until the instance is stopped or deleted.

Because Local SSDs are located on the physical machine where your virtual machine instance is running, they can be created only during the instance creation process. Local SSDs cannot be used as boot devices.

## M

# Machine Learning Glossary

Useful glossary about general machine learning terms, plus terms specific to TensorFlow

# Machine Learning Engine

Cloud Machine Learning Engine is a managed service that enables you to easily build machine learning models with the powerful TensorFlow framework. It provides scalable training and prediction services that work on large scale datasets.
Cloud ML Engine offers training and prediction services, which can be used together or individually. Cloud ML Engine is a proven service used by enterprises to solve problems ranging from identifying clouds in satellite images, ensuring food safety, and responding four times faster to customer emails.
TensorFlow, scikit-learn and XGBoost host your trained models on Cloud ML Engine so that you can send them prediction requests and manage your models and jobs using the GCP services.

# Maglev

Distributed systems for Network Load Balancing

# Managed services

Unmanaged Services need to be monitored and maintained to some extent by trained personnel in order to work properly. Compute Engine is partially unmanaged.

With managed services you don't have to care about provision and maintenance, it is all automated or done by experts. You have only to set up procedures or code. Cloud Run, Functions and Cloud Datastore are managed.

# Maven App Engine Plugin

# Memorystore

Cloud Memorystore provides a fully managed in-memory data store service to build application caches or provide sub-millisecond data access. Cloud Memorystore is a scalable and highly available Redis service fully managed by Google.

# Mobile App

The Google Cloud Console mobile app gives you a convenient way to discover, understand, and respond to production issues. Monitor and make changes to Cloud Platform resources from your iOS and Android device. Manage Cloud Platform resources such as projects, billing, Google App Engine apps, and Google Compute Engine VMs. Receive and respond to alerts helping you quickly address production-impacting issues.

## N

# Natural Language

Google Cloud Natural Language provides powerful natural language understanding as an easy to use API. This API enables application developers to answer the following questions:
1) What are the entities referred to in the block of text?;

2) What is the sentiment (positive or negative) for this block of text?;
3) What is the language of this block of text?; and
4) What is the syntax for this block of text (including parts of speech and dependency trees)? Users can call this API by passing in a block of text or by referring to a document in Google Cloud Storage.

# Nearline

Nearline and Coldline offer ultra low-cost, **highly-durable, highly available archival storage**. Coldline is ideal for cold storage - data your business expects to touch less than once a year. For warmer storage, choose Nearline: data you expect to access less than once a month, but possibly multiple times throughout the year. Both options are available across all GCP regions and provide unparalleled **sub-second access speeds with a consistent API**.

# Network Service Tiers

**Network Service Tiers**: Network Service Tiers enable you to select different quality networks (tiers) for outbound traffic to the internet: the Standard Tier primarily utilizes third party transit providers while the Premium Tier leverages Google's private backbone and peering surface for egress.

# O

# Object Storage

Object storage is a computer data storage architecture that manages data as objects, as opposed to other storage architectures like file systems which manages data as a file hierarchy, and block storage which manages data as blocks within sectors and tracks. Each object typically includes the data itself, a variable amount of metadata, and a globally unique identifier.

# OLAP (Database)

**Online analytical processing**, or **OLAP**, is an approach to answering multi-dimensional analytical (MDA) queries swiftly in computing.[1] OLAP is part of the broader category of business intelligence, which also encompasses relational databases, report writing and data mining.[2] Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM),[3] budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture.[4] The term *OLAP* was created as a slight modification of the traditional database term online transaction processing (OLTP).

# OLTP (Database)

Systems (DBMS) that manage transaction-oriented applications, typically for data entry and retrieval transaction processing.

# Oozie

Oozie is a workflow scheduler system to manage Apache Hadoop jobs.

Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions.

Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability.

Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).

Oozie is a scalable, reliable and extensible system.

# Organization

The Organization resource represents an organization (for example, a company) and is the root node in the GCP resource hierarchy. The Organization resource is the hierarchical ancestor of project resources and Folders. The IAM access control policies applied on the Organization resource apply throughout the hierarchy on all resources in the organization.

# OS-level virtualization

Containers, in short.
OS paradigm in which the kernel allows the existence of multiple isolated user space instances.
Such instances, called containers (Solaris, Docker), Zones (Solaris), virtual private servers (OpenVZ), partitions, virtual environments (VEs), virtual kernel (DragonFly BSD), or jails (FreeBSD jail or chroot jail),[1] may look like real computers from the point of view of programs running in them. A computer program running on an ordinary operating system can see all resources (connected devices, files and folders, network shares, CPU power, quantifiable hardware capabilities) of that computer. However, programs running inside of a container can only see the container's contents and devices assigned to the container.

# P

# Paxos

Paxos is a family of protocols for solving consensus in a network of unreliable processors. Consensus is the process of agreeing on one result among a group of participants. This problem becomes difficult when the participants or their communication medium may experience failures.
The Paxos family of protocols includes a spectrum of trade-offs between the number of processors, number of message delays before learning the agreed value, the activity level of individual participants, number of messages sent, and types of failures.

# Persistent Disk

Persistent disks are available as either standard hard disk drives (HDD) or solid-state drives (SSD). For more general information about persistent disks and the types of persistent disks that are available, read the persistent disks overview.

different storage solutions that are available for your Compute Engine instances.

- Zonal standard persistent disk and Zonal SSD persistent disk: Efficient, reliable block storage.
- Regional persistent disk and regional SSD persistent disk: Regional block storage replicated in two zones.
- Local SSD: High performance transient local block-storage.
- Cloud storage buckets: Affordable object storage.

If you are not sure which option to use, the most common solution is to add a persistent disk to your instance.

You can attach up to 16 (128 in beta) independent persistent disks to most instances, but instances with shared-core machine types or custom machine types with less than 3.75 GB of memory are limited to a maximum of 4 persistent disks.

If you attach a persistent disk to multiple instances, all of those instances must attach the persistent disk in read-only mode. It is not possible to attach the persistent disk to multiple instances in read-write mode. If you need to share dynamic storage space between multiple instances, connect your instances to Cloud Storage or create a network file server.

Regional persistent disks have storage qualities that are similar to both standard and SSD persistent disks. However, regional persistent disks provide durable storage and replication of data between two zones in the same region; you can failover your workload running on regional persistent disks to another zone using the force-attach command. The force-attach command allows you to attach the regional persistent disk to a standby VM instance even if the disk cannot be detached from the original VM due to its unavailability.

**Local SSDs** are physically attached to the server that hosts your virtual machine instance. Local SSDs have higher throughput and lower latency than standard persistent disks or SSD persistent disks. The data that you store on a local SSD persists only until the instance is stopped or deleted. Each local SSD is 375 GB in size, but you can attach up to eight local SSD devices for 3 TB of total local SSD storage space per instance.

# Preemptible virtual machine

Preemptible VMs are highly affordable, short-lived compute instances suitable for batch jobs and fault-tolerant workloads. Preemptible VMs offer the same machine types and options as regular compute instances and last for up to 24 hours. If your applications are fault-tolerant and can withstand possible instance preemptions, then preemptible instances can reduce your Google Compute Engine costs significantly.

# Projects, Host and Shared

In a Shared VPC scenario, the host project contains a common Shared VPC network usable by VMs in service projects. With Shared VPC, the VLAN attachments and Cloud Routers for an interconnect need to be created only in the Shared VPC host project. Because VMs in the service projects use the Shared VPC network, Service Project Admins do not need to create other VLAN attachments or Cloud Routers in the service projects themselves.
See also Shared VCP

# Proxy Server

In computer networks, a **proxy server** is a server (a computer system or an application) that acts as an intermediary for requests from clients seeking resources from other servers.A client connects to the proxy server, requesting some service, such as a file, connection, web page, or other resource available from a different server and the proxy server evaluates the request as a way to simplify and control its complexity. Proxies were invented to add structure and encapsulation to distributed systems.[2] Today, most proxies are **web proxies**, facilitating access to content on the World Wide Web, providing anonymity and may be used to bypass IP address blocking.
A proxy server that passes unmodified requests and responses is usually called a gateway or sometimes a *tunneling proxy*.

# Pub/Sub, Cloud Pub/Sub

Google Cloud Pub/Sub is designed to provide reliable, many-to-many, asynchronous messaging between applications. Publisher applications can send messages to a “topic” and other applications can subscribe to that topic to receive the messages. By decoupling senders and receivers, Google Cloud Pub/Sub allows developers to communicate between independently written applications.

# Q

## QUIC protocol

HTTPS Load Balancing supports the QUIC protocol in connections between the load balancer and the clients. QUIC is a transport layer protocol that provides congestion control similar to TCP and security equivalent to SSL/TLS for HTTP/2, with improved performance. QUIC allows faster client connection initiation, eliminates head-of-line blocking in multiplexed streams, and supports connection migration when a client's IP address changes.

# R

## Redundancy

Redundancy is the duplication of resources. This is used in case of a failover: there is active and passive redundancy. Active is automated, like with load balancers. Passive is manual.

## Region  Zone and Global

A region is a specific worldwide geographical location where you can host your resources with any Cloud Vendor.
Each region has one or more zones.
Resources that live in a zone, such as virtual machine instances or zonal persistent disks, are referred to as zonal resources.
Other resources, like static external IP addresses, are regional.
Other resources, such as images, are global resources that can be used by any other resources across any location

## Replication

Replication is when a resource is duplicated and continuously updated with changes from the master node.
So they synchronize state between them.

## RDP

Remote Desktop Protocol (RDP) is a proprietary protocol developed by Microsoft, which provides a user with a graphical interface to connect to another computer over a network connection. The user employs RDP client software for this purpose, while the other computer must run RDP server software.

Clients exist for most versions of Microsoft Windows (including Windows Mobile), Linux, Unix, macOS, iOS, Android, and other operating systems. RDP servers are built into Windows operating systems; an RDP server for Unix and OS X also exists. By default, the server listens on TCP port 3389 and UDP port 3389.

## Resources global regional zonal

Google Cloud Platform (GCP) resources are hosted in multiple locations worldwide. These locations are composed of regions and zones within those regions. Putting resources in different zones in a region provides isolation from many types of infrastructure, hardware, and software failures. Putting resources in different regions provides an even higher degree of failure independence. This allows you to design robust systems with resources spread across different failure domains.

# Resource Manager

Google Cloud Platform provides resource containers such as Organizations, Folders, and Projects, that allow you to group and hierarchically organize other Cloud Platform resources. This hierarchical organization lets you easily manage common aspects of your resources such as access control and configuration settings. The **Resource Manager** service enables you to programmatically manage these resource containers.

# Roles

You can grant permissions by granting roles to a user, a group, or a service account.

- **Primitive roles**, which include the Owner, Editor, and Viewer roles that existed prior to the introduction of Cloud IAM
- **Predefined roles**, which provide granular access for a specific service and are managed by GCP
- **Custom roles**, which provide granular access according to a user-specified list of permissions

https://cloud.google.com/iam/docs/granting-changing-revoking-access

# RAM disks

Google Compute Engine instances have high-performance, enterprise-class memory that you can use to run your applications. You can allocate some of this memory to create a RAM disk with exceptionally low latency and high throughput. RAM disks work well when your application expects a file system structure and cannot simply store its data in system memory. RAM disks alone do not provide any storage redundancy or flexibility, so it is best to use RAM disks in combination with other instance storage options.

RAM disks share instance memory with your applications. If your instances do not have enough memory to contain RAM disks and your applications, create instances with high-memory machine types or upgrade your existing instances to add more memory.

# Router Google Cloud Router

Google Cloud Router enables dynamic Border Gateway Protocol (BGP) route updates between your VPC network and your non-Google network.

# RPO: Recovery Point Objective

Recovery Point Objective (RPO) describes the interval of time that might pass during a disruption before the quantity of data lost during that period exceeds the maximum allowable threshold or “tolerance.”

# RTO: Recovery Time Objective

RTO is the answer to the question: “How much time did it take to recover after notification of business process disruption?“
RTO designates the amount of “real time” that MAY pass without BIG Damage between disruption and the restore of normal business operations.

# RUN - Cloud RUN

Cloud Run is a fully managed compute platform that automatically scales stateless containers. Cloud Run is serverless: it abstracts away all infrastructure management.
There are:

- fully managed Cloud Run
- Cloud Run on Anthos, which supports both Google Cloud and on-premises environments.

Cloud Run is built upon an open standard, Knative, enabling the portability of your applications.

# S

# Schema Auto-Detection Big Query

Schema auto-detection is available when you load data into BigQuery, and when you query an external data source.

When auto-detection is enabled, BigQuery starts the inference process by selecting a random file in the data source and scanning up to 100 rows of data to use as a representative sample. BigQuery then examines each field and attempts to assign a data type to that field based on the values in the sample.

# Scalability

Scalability is the capability of an application or a system to handle a huge volume of workload and expand or decrease in response to an increased traffic or system resources requests.
In Cloud is usually achieved by making groups of resources managed by a balancing system (load balancer, usually).

# SDK

Google Cloud SDK is a set of tools that you can use to manage resources and applications hosted on Google Cloud Platform. These include the gcloud, gsutil, and bq command line tools. The gcloud command-line tool is downloaded along with the Cloud SDK; a comprehensive guide to gcloud can be found in gcloud Overview.

Additionally, gcloud reference documents all of the gcloud CLI's functionality.
You can download Cloud Client Libraries for supported languages.

# Secret management with Cloud KMS

Applications often require access to small pieces of sensitive data at build or run time. These pieces of data are often referred to as *secrets*. Secrets are similar in concept to configuration files, but are generally more sensitive, as they may grant access to additional data, such as user data.

Applications often require access to small pieces of sensitive data at build or run time. These pieces of data are often referred to as *secrets*. Secrets are similar in concept to configuration files, but are generally more sensitive, as they may grant access to additional data, such as user data.

This topic describes some of the main concepts of secret management. It also provides guidance on how you can use Google Cloud Key Management Service for secret management.

Note: Cloud KMS does not directly store secrets. It can encrypt secrets that you store elsewhere.

Several options exist for managing secrets. Some common ways of storing secrets include using:

- Code or binaries
- A deployment manager
- A secret volume in a container
- Metadata of a VM
- A storage system

# Security Key Enforcement

2-step verification with a security key uses cryptography to provide two-way verification: it makes sure you're logging into the service you originally registered the security key with, and the service verifies that it's the correct security key as well. This provides superior protection to text-message verification.
G Suite, Google Cloud Platform, and Cloud Identity admins and users enrolled in the Advanced Protection Program have access to sensitive data and systems. While security keys are recommended for all users for stronger protection against phishing, enforcing security keys for admins and other high-value users should be the first step.

Titan Security Keys are built with a secure element that includes firmware engineered by Google to verify the integrity of the key and implement FIDO A virtual VPN gateway running in GCP managed by Google, using a configuration you specify in your project. Each Cloud VPN gateway is a regional resource using a regional external IP address. A Cloud VPN gateway can connect to an on-premises VPN gateway or another Cloud VPN gateway.standards to work with many popular devices, browsers, and services. Titan Security Keys are available on the Google Store

# Security Scanner

**Google Cloud Security Scanner** is a web application security scanner that enables developers to easily check for a subset of common web application vulnerabilities in websites built on App Engine and Compute Engine.

# Serverless

Serverless computing is a cloud computing execution model in which the cloud provider runs the server, and dynamically manages the allocation of machine resources. Pricing is based on the actual amount of resources consumed by an application, rather than on pre-purchased units of capacity.
Serverless computing can simplify the process of deploying code into production. Scaling, capacity planning and maintenance operations may be hidden from the developer or operator. Serverless code can be used in conjunction with code deployed in traditional styles, such as microservices. Alternatively, applications can be written to be purely serverless and use no provisioned servers at all.

# Serial Console

Enable interactive access to an instance's serial console to debug boot and networking issues, troubleshoot malfunctioning instances, interact with the GRand Unified Bootloader (GRUB), and perform other troubleshooting tasks.

A virtual machine instance has four virtual serial ports. Interacting with a serial port is similar to using a terminal window, in that input and output is entirely in text mode and there is no graphical interface or mouse support. The instance's operating system, BIOS, and other system-level entities often write output to the serial ports, and can accept input such as commands or answers to prompts. Typically, these system-level entities use the first serial port (port 1) and serial port 1 is often referred to as the serial console.

By default, you can call the getSerialPortOutput method to read information that your instance has written to its serial ports, but you cannot write information for your instance to read. However, if you run into problems accessing your instance through SSH or need to troubleshoot an instance that is not fully booted, you can enable interactive access to the serial console, which lets you connect to and interact with any of your instance's serial ports. For example, you can directly run commands and respond to prompts in the serial port.

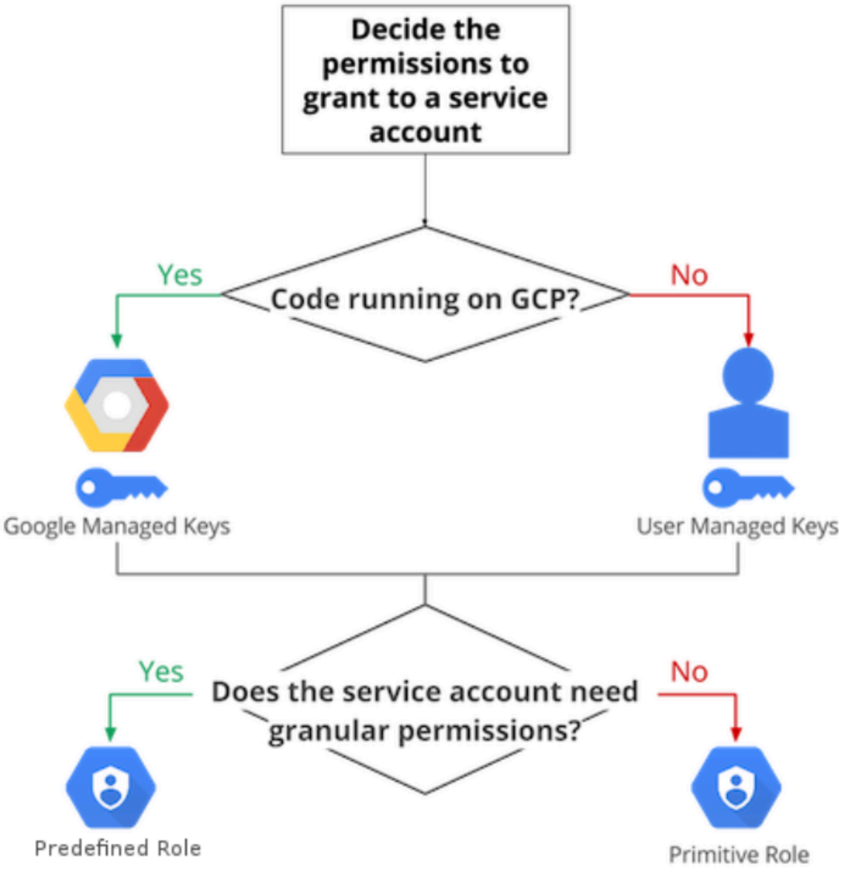
# Service

A service is a collection of endpoints (IP/ports) that provide a set of behaviors. Clients look up a service by its name and then connect to the endpoints that provide that service. Services can also have optional metadata (key-value pairs) associated with them (for example, use_https:true).
A service must belong to a namespace. Each service name must be unique within that namespace.

# Service accounts

A service account is an identity that an instance or an application can use to run API requests on your behalf. This identity is used to identify applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. You can create a service account and grant the service account access to the Cloud Storage API. Then, you would update your application code to pass the service account credentials to the Cloud Storage API. Your application authenticates seamlessly to the API without embedding any secret keys or user credentials in your instance, image, or application code.
You can use service accounts to create instances and other resources. If you create a resource using a service account, that resource is then owned by the creating service account. You can also change the service account of an existing instance.

*Access Scopes*
Access scopes are the legacy method of specifying permissions for your instance. Before the existence of IAM roles, access scopes were the only mechanism for granting permissions to service accounts. Although they are not the primary way of granting permissions now, you must still set up access scopes when configuring an instance to run as a service account.
In addition to setting access scopes, you must grant the correct IAM roles to a service account to determine the level of access the account has.
Default: https://www.googleapis.com/auth/cloud-platform
**Service Account Keys**
To allow a user to manage service account keys, grant the Service Account Key Admin role (roles/iam.serviceAccountKeyAdmin)
To use a service account outside of the Google Cloud Platform (on other platforms or on premise), you must establish the identity of the service account. Public/private key pairs will let you do that.
You can create a service account key using the GCP Console, the gcloud tool, the serviceAccounts.keys.create() method, or one of the client libraries.
When you create a key, your new public/private key pair is generated and downloaded to your machine; it serves as the only copy of the private key. You are responsible for storing the private key securely. Take note of its location and ensure the key is accessible to your application; it needs the key to make authenticated API calls.
You can list the service account keys for a service account using the GCP Console, the gcloud tool, the serviceAccount.keys.list() method, or one of the client libraries.


Decide the permissions to grant to a service account
Code running on GCP?
Yes
No
Google Managed Keys
User Managed Keys
Does the service account need granular permissions?
Yes
No
Predefined Role
Primitive Role

# Service Directory

Service Directory is a single place to publish, discover, and connect to services Google Cloud, multi-cloud, and on-premises environments and can scale up to thousands of services and endpoints.
Works with Cloud DNS. Service Directory zones allow services to be made available on Virtual Private Cloud (VPC). With Service Directory, you can register all of your services in a single place and resolve them via HTTP, gRPC, and DNS.
You can create a universal service name that works across Google Cloud products, like App Engine and GKE. You can make these services available over DNS.

# Service Mesh

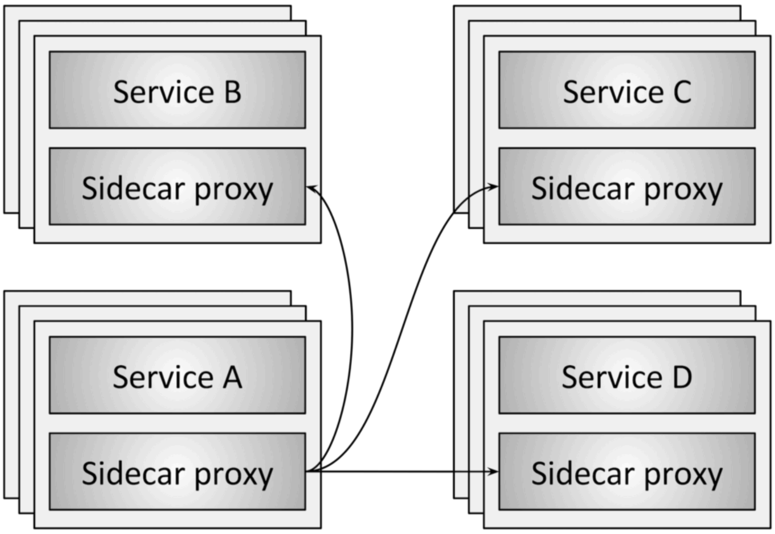
A service mesh is the network of microservices of a set of applications and the interactions between them. As a service mesh grows in size and complexity, it can become harder to understand and manage. Its requirements can include discovery, load balancing, failure recovery, metrics, and monitoring. A service mesh also often has more complex operational requirements, like A/B testing, canary rollouts, rate limiting, access control, and end-to-end authentication.

GCP uses ISTIO, which offers load balancing, service-to-service authentication, monitoring, and advanced discovery.
ISTIO adds a special sidecar proxy to services that intercepts all network communication between microservices.
Main features:

- Automatic load balancing for HTTP, gRPC, WebSocket, and TCP traffic.
- Fine-grained control of traffic behavior with rich routing rules, retries, failovers, and fault injection.
- A pluggable policy layer and configuration API supporting access controls, rate limits and quotas.

- Automatic metrics, logs, and traces for all traffic within a cluster, including cluster ingress and egress.
- Secure service-to-service communication in a cluster with strong identity-based authentication and authorization.

# Service Mesh Control and Data Plane



In a service mesh each service instance is colocated with a sidecar network proxy, that **is the DATA PLANE** and performs these actions:

- Health checking
- Routing
- Load balancing
- Authentication and authorization
- Observability

So, the data plane is responsible for conditionally translating, forwarding, and observing every network packet that flows to and from a service instance

Instead, **the control plane** provides policy and configuration for all of the running data planes in the mesh. Does not touch any packets/requests in the system but takes a set of isolated stateless sidecar proxies and turns them into a distributed system.
**So it is configuration and control of the overall system.**
Control Plane sets policy that will eventually be enacted by the data plane.

- Data planes: Linkerd, NGINX, HAProxy, Envoy, Traefik
- Control planes: Istio, Nelson, SmartStack

Istio is the control plane and Envoy is the data plane

# Shared VCP

Shared VPC allows an organization to connect resources from multiple projects to a common VPC network, so that they can communicate with each other securely and efficiently using internal IPs from that network. When you use Shared VPC, you designate a project as a host project and attach one or more other service projects to it. The VPC networks in the host project are called Shared VPC networks. Eligible resources from service projects can use subnets in the Shared VPC network.
Shared VPC lets organization administrators delegate administrative responsibilities, such as creating and managing instances, to Service Project Admins while maintaining centralized control over network resources like subnets, routes, and firewalls.

# Shell

Google Cloud Shell provides you with command-line access to your cloud resources directly from your browser. You can easily manage your projects and resources without having to install the Google Cloud SDK or other tools on your system. With Cloud Shell, the Cloud SDK gcloud command-line tool and other utilities you need are always available, up to date and fully authenticated when you need them.

# Shutdown Scripts

Create and run shutdown scripts that execute commands right before an instance is terminated or restarted. This is useful if you rely on automated scripts to start up and shut down instances, allowing instances time to clean up or perform tasks, such as exporting logs, or syncing with other systems.

Shutdown scripts are especially useful for instances in a managed instance group with an autoscaler. If the autoscaler shuts down an instance in the group, the shutdown script runs before the instance stops and the shutdown script performs any actions that you define. The script runs during the limited shutdown period before the instance stops. For example, your shutdown script might copy processed data to Cloud Storage or backup any logs.

Shutdown scripts function very similarly to startup scripts. Much of the documentation for startup scripts also applies for shutdown scripts.

# Sinks (Stackdriver Logs Groups)

To create an export sink, click the **Create Export** button at the top of the **Logs Exports** page. You can also access this button at the top of the Logs Viewer page.

# Snapshots

Snapshots are different from public images and custom images, which are used primarily to create instances or configure instance templates. Snapshots are useful for periodic backup of the data on your persistent disks, and you can use snapshots to create a custom image when needed. You can create snapshots from persistent disks even while they are attached to running instances.
Snapshots are incremental and automatically compressed, so you can create regular snapshots on a persistent disk faster and at a much lower cost than if you regularly created a full image of the disk.

https://cloud.google.com/compute/docs/disks/restore-and-delete-snapshots

# SOCKS

A **SOCKS** server is a general purpose **proxy** server that establishes a TCP connection to another server on behalf of a client, then routes all the traffic back and forth between the client and the server. It works for any kind of network protocol on any port. **SOCKS** Version 5 adds additional support for security and UDP.

# Source Repositories

Cloud Source Repositories provides **Git version control** to support collaborative development of any application or service, including those that run on App Engine and Compute Engine.

# SPIFFE

SPIFFE is a set of open-source standards for securely identifying software systems in dynamic and heterogeneous environments.
Systems that adopt SPIFFE can easily and reliably mutually authenticate wherever they are running.
SPIFFE is a set of open-source specifications for a framework capable of bootstrapping and issuing identity to services across heterogeneous environments and organizational boundaries. The heart of these specifications is the one that defines short lived cryptographic identity documents – called SVIDs via a simple API. Workloads can

then use these identity documents when authenticating to other workloads, for example by establishing a TLS connection or by signing and verifying a JWT token.

# Spanner, Cloud Spanner

Cloud Spanner is a fully managed, mission-critical relational database service. It is designed to provide a scalable online transaction processing (OLTP) database with high availability and strong consistency at global scale.

*Differences between Cloud Spanner and CloudSQL*

The Main difference is that Spanner is horizontally scalable whereas Cloud SQL is not.

for Cloud SQL you can select machine type, type of hard disk and size, region and zone. Maximal data throughput through network is 2000 MB/s. You are limited with having everything on one server, so that's basically your limit.

Efficiency of Spanner depends on the other hand on number of nodes which are used. with every node throughput increases so you can scale horizontally by just adding nodes (just change one digit in settings, that's all. Each node has 2 TB of storage and possibility of 10000 QPS of reads or 2000 QPS of writes.

Speed also depends on database schema and modeling.

Regarding pricing, Spanner is more expensive. With CloudSQL it depends on instance type so it's possible to adjust based on the needs.

Details DBA CAP

# Spark - Apache Spark

Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

# Speech

**Google Cloud Speech-to-Text** allows developers to convert audio to text by applying powerful neural network models in an easy to use API.

# SPINNAKER

**Spinnaker**: open-source, multi-cloud continuous delivery platform that helps you release software changes with high velocity and confidence.

# SQL Cloud SQL

Google Cloud SQL is a web service that allows you to create, configure, and use relational databases that live in Google's cloud. It is a fully-managed service that maintains, manages, and administers your databases, allowing you to focus on your applications and services.

# SQL-proxy

The Cloud SQL Proxy works by having a local client, called the proxy, running in the local environment. Your application communicates with the proxy with the standard database protocol used by your database. The proxy uses a secure tunnel to communicate with its companion process running on the server.

# Sqoop

Apache Sqoop(TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.

# Stackdriver

Google Stackdriver is an integrated monitoring, logging, and diagnostics hosted solution that helps you gain insight into applications that run on Google Cloud Platform and other public cloud platforms. Stackdriver helps you keep your cloud-powered applications fast and available. Stackdriver includes Monitoring, Logging, Error Reporting, Debugger, Profiler, and Trace components.

# Stackdriver Debugger

Stackdriver Debugger is a feature of Google Cloud Platform that lets you inspect the state of an application, at any code location, without stopping or slowing down the running app. Stackdriver Debugger makes it easier to view the application state without adding logging statements.

You can use Stackdriver Debugger with any deployment of your application, including test, development, and production. The debugger adds less than 10ms to the request latency only when the application state is captured. In most cases, this is not noticeable by users.

For Java, Phyton, Go, Nodejs, Ruby, PHP

# Stackdriver Error Reporting

Stackdriver Error Reporting aggregates and displays errors produced in your running cloud services.
Supported languages are Go, Java, .NET, Node.js, PHP, Python, and Ruby. To report errors from Android and iOS client applications, we recommend setting up Firebase Crash Reporting.
Stackdriver Error Reporting is Generally Available for Google Cloud Functions and Google App Engine standard environment and is a Beta feature for Google App Engine flexible environment, Google Compute Engine, and AWS EC2.

Reporting errors from your application can be achieved by logging application errors to Google Stackdriver Logging or by calling an API endpoint. The setup process depends on your platform; please refer to the setup guides.

Information in Stackdriver Error Reporting is retained for 30 days.
Error Reporting displays errors for the currently-selected GCP Console project. It does not support Stackdriver Workspaces.

# Stackdriver Logging

Stackdriver Logging is part of the Stackdriver suite of products in Google Cloud Platform (GCP). It includes storage for logs, a user interface called the Logs Viewer, and an API to manage logs programmatically. Logging lets you read and write log entries, search and filter your logs, export your logs, and create logs-based metrics.

Log entries are held in Stackdriver Logging for a limited time known as the retention period. After that, the entries are deleted. If you want to keep your log entries longer, export them outside of Stackdriver Logging.
The retention periods for different types of logs are listed in the Logging Quota Policy.
An advanced logs filter is an expression in the Logging filter language. It is used in the Logs Viewer and the Stackdriver Logging API to select log entries, such as those from a particular VM instance or those arriving in a particular time period with a particular severity level.
Monitored resources: Examples are individual Compute Engine VM instances, individual Amazon EC2 VM instances, database instances, and so on. For a complete listing of monitored resource types, see Monitored Resources and Services.

https://cloud.google.com/logging/docs/logs-based-metrics/

# Stackdriver Monitoring

Stackdriver Monitoring provides visibility into the performance, uptime, and overall health of cloud-powered applications. Stackdriver collects metrics, events, and metadata from Google Cloud Platform, Amazon Web Services, hosted uptime probes, application instrumentation, and a variety of common application components including Cassandra, Nginx, Apache Web Server, Elasticsearch, and many others. Stackdriver ingests that data and generates insights via dashboards, charts, and alerts. Stackdriver alerting helps you collaborate by integrating with Slack, PagerDuty, HipChat, Campfire, and more.

# Stackdriver Trace

Stackdriver Trace is a distributed tracing system that collects latency data from your applications and displays it in the Google Cloud Platform Console. You can track how requests propagate through your application and receive detailed near real-time performance insights. Stackdriver Trace automatically analyzes all of your application's traces to generate in-depth latency reports to surface performance degradations, and can capture traces from all of your VMs, containers, or Google App Engine projects. A Zipkin collector is also available, which allows Zipkin tracers to submit data to Stackdriver Trace. Projects running on Google App Engine are automatically captured.

# Stackdriver Workspaces

A Workspace is a tool for monitoring resources contained in one or more GCP projects or AWS accounts. Each Workspace can have between 1 and 100 **monitored projects**, including one or more GCP projects and any number of AWS accounts. You can have as many Workspaces as you wish, but GCP projects and AWS accounts cannot be monitored by more than one Workspace.

A Workspace contains the custom dashboards, alerting policies, uptime checks, notification channels, and group definitions that you use with your monitored projects. A Workspace can access metric data from its monitored projects, but the metric data and log entries remain in the individual projects.

# Startup Scripts

Create and run your own startup scripts on your virtual machines to perform automated tasks every time your instance boots up. Startup scripts can perform many actions, such as installing software, performing updates, turning on services, and any other tasks defined in the script. You can use startup scripts to easily and programmatically customize your virtual machine instances, including on new instances at creation time.

# Storage, Cloud Storage

Google Cloud Storage is a RESTful service for storing and accessing your data on Google's infrastructure. The service combines the performance and scalability of Google's cloud with advanced security and sharing capabilities.

In Cloud Storage, you create a bucket to store your data. A bucket has three properties that you specify when you create it: a globally unique name, a location where the bucket and its contents are stored, and a default storage class for objects added to the bucket.

Integrate storage into your apps with a single unified API

Optimize price/performance across four storage classes with Object Lifecycle Management

Access data instantly from any storage class

Designed for secure and durable storage

| | | | |
|---|---|---|---|
| Multi-Regional Storage | • >99.99% typical monthly availability<br>• 99.95% availability SLA*<br>• Geo-redundant | Storing data that is frequently accessed ("hot" objects) around the world, such as serving website content, streaming videos, or gaming and mobile applications. | $0.026 |
| Regional Storage | • 99.99% typical monthly availability<br>• 99.9% availability SLA*<br>• Lower cost per GB stored<br>• Data stored in a narrow geographic region<br>• Redundant across availability zones | Storing frequently accessed in the same region as your Google Cloud DataProc or Google Compute Engine instances that use it, such as for data analytics. | $0.02 |
| Nearline Storage | • 99.9% typical monthly availability<br>• 99.0% availability SLA*<br>• Very low cost per GB stored<br>• Data retrieval costs<br>• Higher per-operation costs<br>• 30-day minimum storage duration | Data you do not expect to access frequently (i.e., no more than once per month). Ideal for back-up and serving long-tail multimedia content. | $0.010 |

| | | | |
|---|---|---|---|
| Coldline Storage | • 99.9% typical monthly availability<br>• 99.0% availability SLA*<br>• Lowest cost per GB stored<br>• Data retrieval costs<br>• Higher per-operation costs<br>• 90-day minimum storage duration | Data you expect to access infrequently (i.e., no more than once per year). Typically this is for disaster recovery, or data that is archived and may or may not be needed at some future time. | $0.007 |

# Storage Transfer Service

Storage Transfer Service transfers data from an online *data source* to a *data sink*. Your *data source* can be an Amazon Simple Storage Service (Amazon S3) bucket, an HTTP/HTTPS location, or a Cloud Storage bucket. Your *data sink* (the destination) is always a Cloud Storage bucket.

You can use Storage Transfer Service to:

- Backup data to a Cloud Storage bucket from other storage providers.
- Move data from a Multi-Regional Storage bucket to a Nearline Storage bucket to lower your storage costs.

Storage Transfer Service performs a data transfer with a *transfer operation*. Transfer operations are scheduled and configured through a *transfer job*. Storage Transfer Service has options that make data transfers and synchronization between data sources and data sinks easier. For example, you can:

- Schedule one-time transfer operations or recurring transfer operations.
- Delete existing objects in the destination bucket if they don't have a corresponding object in the source.
- Delete source objects after transferring them.
- Schedule periodic synchronization from data source to data sink with advanced filters based on file creation dates, file-name filters, and the times of day you prefer to import data.

By default, Storage Transfer Service copies a file from the data source if the file doesn't exist in the data sink or if it differs between the version in the source and the sink. The default is also to retain files in the source after the transfer operation.

# T

# Target Pools

A Target Pool resource defines a group of instances that receive incoming traffic from forwarding rules. When a forwarding rule directs traffic to a target pool, Google Cloud Load Balancing picks an instance from these target pools based on a hash of the source IP and port and the destination IP and port. See the Load distribution algorithm for more information about how traffic is distributed to instances.

# Tasks - Cloud Tasks

Queues for App Engine
Cloud Tasks lets you separate out pieces of work that can be performed independently, outside of your main application flow, and send them off to be processed, asychronously, using handlers that you create. These independent pieces of work are called tasks.
The Cloud Tasks service is designed for asynchronous work. It does not provide strong guarantees around the timing of task delivery and is **therefore unsuitable for interactive applications where a user is waiting for the result**.

# Terraform

Terraform is a tool (devops) for building, changing, and versioning infrastructure with code.
Configuration files describe to Terraform the components needed to run a single application or your entire datacenter.
**Terraform plan** command is used to create an execution plan. Terraform determines what actions are necessary to achieve the desired state specified in the configuration files.

# Test Lab

Google Cloud Test Lab enables you to test **mobile applications** using physical and virtual devices in the cloud. It runs instrumentation tests and script-less robotic tests on a matrix of device configurations, and reports detailed results to help improve the quality of your mobile app.

# Tools for Android Studio

Cloud Tools for Android Studio is a set of tools for the Android Studio IDE that help you develop your Android applications and deploy them on Google Cloud Platform. Firebase

# Tools for Eclipse

Cloud Tools for Eclipse is a Google-sponsored open source plugin that supports Google Cloud Platform development inside the Eclipse IDE.

# Tools for IntelliJ

Using Cloud Tools for IntelliJ you can easily deploy Java backends for your cloud apps to the Google App Engine standard and flexible environments. You can run and test the backend locally, and when you're finished developing, you can deploy your project live from within IntelliJ IDEA. If there are problems in production, you can debug your live cloud backend using Stackdriver Debugger without halting or slowing down the application.

# Tools for PowerShell

Cloud Tools for PowerShell lets you script, automate, and manage your Windows workloads running on Cloud Platform. Using PowerShell's powerful scripting environment, customize your cloud workflows using the Windows tools you're already familiar with.

## Tools for Visual Studio

## Traffic Migration/Splitting

Manage how much traffic is received by a version of your application by migrating or splitting traffic.

**Traffic migration** smoothly switches request routing, gradually moving traffic from the versions currently receiving traffic to one or more versions that you specify.

**Traffic splitting** distributes a percentage of traffic to versions of your application. You can split traffic to move 100% of traffic to a single version or to route percentages of traffic to multiple versions. Splitting traffic to two or more versions allows you to conduct A/B testing between your versions and provides control over the pace when rolling out features.

Remember: Traffic splitting is applied to URLs that do not explicitly target a version. For example, the following URLs split traffic because they target all the available versions within the specified service:

- [MY_PROJECT_ID].appspot.com - Distributes traffic to versions of the default service.
- [MY_SERVICE].[MY_PROJECT_ID].appspot.com - Distributes traffic to versions of the MY_SERVICE service.

For more information, see How Requests are Routed.

To manually perform traffic migration and splitting from the GCP Console, see Migrating Traffic and Splitting Traffic.

## Translation

**Google Cloud Translation (and Google Cloud Translation v2 or any subsequent general availability version/release)** is a RESTful API that automatically translates text from one language to another language (e.g. French to English). You can use the API to programmatically translate text in your webpages or apps.

# U

## url Maps

HTTP(S) Load Balancing allows you to direct traffic to different instances based on the incoming URL. For example, you can send requests for `http://www.example.com/audio` to one backend service, which contains instances configured to deliver audio files, and requests for `http://www.example.com/video` to another backend service, which contains instances configured to deliver video files.
Create a UrlMaps resource:

```
"hostRules": [
    {
      "description": string,
      "hosts": [
        string
      ],
      "pathMatcher": string
    }
```

V

# Video Intelligence

Google Cloud Video Intelligence makes videos searchable, and discoverable, by extracting metadata with an easy to use REST API. You can now search every moment of every video file in your catalog. It quickly annotates videos stored in Google Cloud Storage, and helps you identify key entities (nouns) within your video; and when they occur within the video. Separate signal from noise, by retrieving relevant information within the entire video, shot-by-shot, -or per frame-.

# Virtual Machine/Instances

An *instance* is a virtual machine (VM) hosted on Google's infrastructure. You can create an instance by using the Google Cloud Platform Console or the `gcloud` command-line tool.
Compute Engine instances can run the public images for Linux and Windows Server that Google provides as well as private custom images that you can create or import from your existing systems. You can also deploy Docker containers, which are automatically launched on instances running theContainer-Optimized OS public image.

You can use a set of predefined machine types or by creating your own custom machine types.

# Virtual Private Cloud

A VPC network, sometimes just called a "network," is a virtual version of a physical network, like a data center network. It provides connectivity for your Compute Engine virtual machine (VM) instances, Kubernetes Engine clusters, App Engine Flex instances, and other resources in your project.
Provides a private network topology with IP allocation, routing, and network firewall policies to create a secure environment for your deployments.

# VPC Shared

Shared VPC allows an organization to connect resources from multiple projects to a common VPC network, so that they can communicate with each other securely and efficiently using internal IPs from that network.
Shared VPC lets organization administrators delegate administrative responsibilities, such as creating and managing instances, to Service Project Admins while maintaining centralized control over network resources like subnets, routes, and firewalls.

# VPN, Cloud VPN

Cloud VPN securely connects your on-premises network to your Google Cloud Platform (GCP) Virtual Private Cloud (VPC) network through an IPsec VPN connection. Traffic traveling between the two networks is encrypted by one VPN gateway, then decrypted by the other VPN gateway. This protects your data as it travels over the Internet.

To create a virtual private network (VPN), see Choosing a VPN Routing Option.

# VPN gateway

**Cloud VPN gateway**

A virtual VPN gateway running in GCP managed by Google, using a configuration you specify in your project. Each Cloud VPN gateway is a regional resource using a regional external IP address. A Cloud VPN gateway can connect to an on-premises VPN gateway or another Cloud VPN gateway.

**On-premises VPN gateway**
The VPN gateway not in GCP, connected to a Cloud VPN gateway, can be a physical device in your data center or a physical or software-based VPN offering in another cloud provider's network. Cloud VPN instructions are written from the point of view of your VPC network, so the “on-premises gateway” is the gateway connecting to Cloud VPN.

**VPN tunnel**
A VPN tunnel connects two VPN gateways and serves as a virtual medium through which encrypted traffic is passed. Two VPN tunnels must be established to create a connection between two VPN gateways: Each tunnel defines the connection from the perspective of its gateway, and traffic can only pass once the pair of tunnels is established.

## Video Intelligence

**Google Cloud Video Intelligence** makes videos searchable, and discoverable, by extracting metadata with an easy to use REST API. It quickly annotates videos stored in Google Cloud Storage, and helps you identify key noun entities of your video and when they occur within the video.

## Vision

**Google Cloud Vision** enables developers to understand the content of an image by encapsulating powerful machine learning models in an easy to use API. It quickly classifies images into thousands of categories (e.g., "sailboat", "lion", "Eiffel Tower"), detects individual objects and faces within images, and finds and reads printed words contained within images. You can build metadata on your image catalog, moderate offensive content, or enable new marketing scenarios through image sentiment analysis. You can also analyze images uploaded in the request and integrate with your image storage on Google Cloud Storage.

# Y

## Yarn - Yet Another Resource Negotiator

The fundamental idea of YARN is to split up the functionalities of resource management and job scheduling/monitoring into separate daemons. The idea is to have a global ResourceManager (RM) and per-application ApplicationMaster (AM). An application is either a single job or a DAG of jobs.

The ResourceManager and the NodeManager form the data-computation framework. The ResourceManager is the ultimate authority that arbitrates resources among all the applications in the system. The NodeManager is the per-machine framework agent who is responsible for containers, monitoring their resource usage (cpu, memory, disk, network) and reporting the same to the ResourceManager/Scheduler.

# W

## Wireshark

Wireshark is the world’s foremost network protocol analyzer. It lets you see what’s happening on your network at a microscopic level. It is the de facto (and often de jure) standard across many industries and educational institutions.

Wireshark is a free and open source packet analyzer. It is used fornetwork troubleshooting, analysis, software and communications protocol development, and education. Originally named Ethereal, the project was renamed Wireshark in May 2006 due to trademark issues.

Wireshark on Cloud runs on Amazon Web Services (AWS) and Google Cloud Platform (GCP) is used for network troubleshooting, analysis, software and communications protocol development and education.

Wireshark is owned by Wireshark (https://www.wireshark.org/) and they own all related trademarks and IP rights for this software.

# Windows on Compute Engine

You can run your Windows applications on Google Compute Engine and take advantage of many benefits available to virtual machine instances such as reliable storage options, the speed of the Google network, and Autoscaling.

Compute Engine provides several tools to help bring your Windows applications and services to the cloud:

- Use Windows Server images to create instances with a basic Windows environment upon which you can build your applications. For Windows Server 2016 and 2012 R2 images, you can select from either the Windows Server with Desktop Experience or Windows Server Core configurations.
- Use SQL Server images to start instances that have Windows Server with SQL Server preinstalled. Pay for both Windows Server and SQL Server licenses only when you use them. Windows Server images receive per-second billing and SQL Server images receive per-minute billing.
- Run .NET applications on your Compute Engine instances.
- Deploy Active Directory to your instances and bring your domain services to the cloud.
- Run IIS web servers to host your web content on Windows instances.
- If you have existing licenses for SQL Server or other applications that run in a Windows environment, use your existing Microsoft application licenses through the Microsoft License Mobility program.

To get started, try the Windows quickstart, create a Windows Server instance, or create an instance with SQL Server preinstalled. Connecting to a Windows Instance

http-load-balancing-iis

## Z

# Zone Region and Global

A region is a specific worldwide geographical location where you can host your resources with any Cloud Vendor.
Each region has one or more zones.
Resources that live in a zone, such as virtual machine instances or zonal persistent disks, are referred to as zonal resources.
Other resources, like static external IP addresses, are regional.
Other resources, such as images, are global resources that can be used by any other resources across any location