Google Associate Cloud Engineer notes

31.5.2021 27-minute read
notes • GCP

Personal notes used as a general recap before the Associate Cloud Engineer exam

> Exam recently successfully passed

⭐ Important concepts are marked with a star
In most cases the original source is linked

The following are my personal notes, so I assume no responsibility or liability for any errors or omissions in the content
found an error or have a question? write to me

Index

- Index
- Glossary
- Gcloud Basics
- Network
  - Virtual Private Clouds
    - Quotas
  - Cloud Interconnect
  - Load balancer
    - Choose the Load Balancer
  - Firewall
  - Cloud Armor
- Storage
  - Storage Types
  - GCP storage services
- Compute services
  - **Compute engine (VM)**
  - **GKE - Google Kubernetes Engine**
  - **Cloud Run**
  - **App Engine**
  - **Cloud functions**
  - **Cloud endpoints**
  - Cloud Tasks
- GKE - Google Kubernetes Engine
- IAM
  - Hierarchy
- Projects
- Stackdriver
  - Quotas
- Cloud Source Repositories
- Billing
  - Pricing calculation
  - Billing monitor
- Cloud Deployment Manager
- Identity-Aware Proxy
- Cloud Logging
- Resources

Glossary

- Multi-Tier Application
  - In software engineering, multitier architecture (often referred to as n-tier architecture) or multilayered architecture is a client–server architecture in which presentation, application processing and data management functions are **physically separated**. - wiki
- Zone vs Regional
  - Regions are composed by zones (e.g. Region us-west1 is composed by zones us-west1-[a,b,c])- docs
  - All regions are at least 100 miles apart. - doc
- db transactional**:** do all the jobs or nothing (have roolback)
- OLAP vs OLTP - article
  - OLTP: OnLine Transaction Process (*Database)*
    - Purpose of an OLTP system is to handle data
  - OLAP: Online Analytical Processing (*Data Warehouse*)
    - Used for analysing the data
- Gcloud
  - manage Google Cloud Platform resources and developer workflow - doc
- Network address translation **(NAT)**

    - mapping an IP address space into another - wiki
    - used to map private IP (192.168.x.x) to external-public IP
    - how a NAT can manage multiple connections from devices inside the network, using the same external IP? (that is: how the NAT can forward the received packages?)
        - Each package have IP:port of the sender, NAT replace this with custom *port* linked with the internal device ip - reddit
- **Proxy**
    - client directs the request to the proxy server, which evaluates the request and performs the required network transactions - wiki
        - "In computer networks, a proxy is a middleman you've assigned to send and receive messages for you." - reddit
    - **Proxy vs nat**
        - "NAT works at the network layer while proxy at the application layer." - huawei
            - NAT is transparent to various applications
            - proxy must resort to the IP address of the proxy server specified in application programs
- Global resources - doc
    - GCP resources accessible by any resource in any zone within the same project.
    - ⭐ Some global resources:
        - **Images**: used by any instance or disk resource in the same project as the image
        - **Snapshots**
        - VPC network
        - Routes
- Google Front End Service - doc
    - When a service wants to make itself available on the Internet, it can register itself with an infrastructure service called the Google Front End (GFE)

Gcloud Basics

```
# Init of the tool
gcloud init

# Gcloud structure
$ gcloud compute instances  list
# |------base--| |--who--| |-what-|

$ gcloud components install  kubectl # exception
# |------base-----| |-what-| |-who-|

# Set the project
gcloud config set project PROJECT-NAME

# Bucket Versioning (NB: is gsutil)
gsutil versioning set (on|off) gs://<bucket_name>...
gsutil versioning get gs://<bucket_name>...

# List VM
gcloud compute instances list [--zones] [--format json]

# Create VM with boot disk
gcloud compute instances create VM_NAME \
    --source-snapshot=BOOT_SNAPSHOT_NAME \
    --boot-disk-size=BOOT_DISK_SIZE \
    --boot-disk-type=BOOT_DISK_TYPE \
    --boot-disk-device-name=BOOT_DISK_NAME

# Install components (e.g. kubectl, minikube, kustomize, bq)
gcloud components list
gcloud components install PRODUCT

# Set a default Region
gcloud config set compute/region europe-west1

# Create Compute Engine persistent disks
gcloud compute disks create my-disk-1 my-disk-2 --zone=us-east1-a

# Resize a cluster nodes
gcloud container clusters resize sample-cluster --num-nodes=2

# Add IAM policy binding
gcloud projects add-iam-policy-binding example-project-id-1  --member='user:test-user@gmail.com' --role='roles/editor'

# Delete `default` VPC (NB: start with 'compute')
gcloud compute networks delete defaulta

# Create VPC
gcloud compute networks create

# gcloud Wide Flags
--account         # GCP user account to use for invocation
--project         # The Google Cloud Platform project ID to use for this invocation
--billing-project # project that will be charged quota for operations performed
--configuration   # The configuration to use for this command invocation
--flags-file      # A YAML or JSON file that specifies a --flag:value dictionary
--flatten         # Use to "flatten" resources list
--format          # Set the format for printing command output resources
--log-http        # Log all HTTP server requests and responses to stderr
```

```
--trace-token     # Token used to route traces of service requests for investigation of issues
--verbosity
--quiet
--impersonate-service-account

# List VPC networks
gcloud compute networks list

# List existing clusters for running containers
gcloud container clusters list

# Describe cluster image info (NB: is gcloud not kubectl)
gcloud container images describe gcr.io/myproject/myimage:tag
```

Network

Virtual Private Clouds

- A Virtual Private Cloud (VPC) network is a virtual version of a physical network, implemented inside of Google's production network, using Andromeda.
  - Or: An on-demand configurable pool of shared resources allocated within a public cloud environment - wiki
- VPC network consists of one or more useful IP range partitions *called subnets*
- **Networks** and **subnets** are different resources in Google Cloud - doc
  - VPC networks do not have any IP address ranges associated with them
- *VPCs are global* resources and *subnets* within that VPC *are regional resources* - wiki
  - VPC in *auto-mode* create one subnet for each region
  - CIDR range: smaller the number after the slash, the more addresses are available
- Shared VPC
  - Allows an organization to connect resources from multiple projects to a common VPC - gcp
    - Each resource can communicate with each other using internal IPs from that network
  - ⭐ Usage: designate a project as a *host project* and attach other *service projects* to it
- **VPC Network peering**
  - Used to connect two VPC regardless they belong to same project or same organization - doc
  - Key proprieties - doc
    - VPC Network Peering works with Compute Engine, GKE, and App Engine flexible environment.
      - ⭐ Note *App engine standard* is missing (flixible present)
- Alias IP ranges
  - Used to assign multiple IP to a VM
    - useful if the resource host multiple services and you want to assign at each service a different IP (useful for GKE pods)
- ***subnets***
  - Each VPC network consists of one or more useful IP range partitions called subnets - doc
  - Each subnet is associated with a **region**
  - Show the default subnets:
    - `gcloud compute networks subnets list --network default`
- ⭐ Number of available regions and zones
  - As of Q1 2020, Google Cloud Platform is available in **25 regions** and **77 zones**
- **Routes** - doc
  - system-generated default route:
    - Priority of 1000 and target 0/0
    - path out of the VPC network, including the path to the internet
    - standard path for Private Google Access
- ⭐ **Private Google Access** - doc
  - Allow VM without external IP to communicate with Google APIs and services
- A network must have at least one subnet before you can use it.
  - Could be created with:

```
# Create the VPC network:
gcloud compute networks create NETWORK \
    --subnet-mode=auto \ # auto or custom
    --bgp-routing-mode=DYNAMIC_ROUTING_MODE \ # global or regional
    --mtu=MTU # maximum transmission unit size

# List VPC networks
gcloud compute networks list
```

- Projects can contain multiple VPC networks.
  - new projects start with a default network (an auto mode VPC network) that has one subnetwork (subnet) in each region.
- Auto vs custom mode
  - Auto: one subnet from each region is automatically created within it
  - Custom: no subnets are automatically created
- ⭐ How many VPC networks can we create? - the default is 5

```
# You can list quotas using `describe`
$ gcloud compute project-info describe | grep -B 3 -A 3 NETWORK
  metric: SNAPSHOTS
  usage: 0.0
- limit: 5.0 # <-- [!]
  metric: NETWORKS
  usage: 1.0
- limit: 100.0
  metric: FIREWALLS
--
  metric: SSL_CERTIFICATES
  usage: 0.0
- limit: 100.0
  metric: SUBNETWORKS
  usage: 27.0
- limit: 10.0
  metric: TARGET_TCP_PROXIES
```

Quotas

Some number from doc

- Maximum number of secondary IP ranges per subnet - 30
- Maximum number of connections to a single VPC network - 25
- Maximum number of VM instances - 15,000 per network

Cloud Interconnect

- Cloud Interconnect extends your on-premises network to Google's network through a highly available, low latency connection. - docs
- Note: you can connect to GCP in three ways - docs
  - Cloud VPN
    - ⭐ Cloud VPN is under *Hybrid Connectivity*
      - During setup, you can specify the *Google Compute Engine VPN gateway*
  - Cloud Interconnect
  - Cloud Router
- To access only Google Workspace or supported Google APIs:
  - **Direct Peering**
    - direct peering connection between your and Google's edge network
  - **Carrier Peering**
    - sing a service provider to obtain enterprise-grade network services that connect your infrastructure to Google.
- Other connections
  - CDN Interconnect
    - third-party Content Delivery Network (CDN) providers to establish direct peering links with Google's edge network
- Network tier
  - You can specify which network use for the connections - doc
    - After selected a default tier, you could always choose which use at deployment time
  - Two kind of tiers
    - Premium tier: use high performing G. networks
    - Standard tier: cheaper, use standard internet networks

Load balancer

- Eli5 Load Balancer - serve the request to machines less busy
- Allows you to put your resources behind a single IP address that is externally accessible or internal to your Virtual Private Cloud (VPC) network - gcp
- Anycast = single destination IP address has multiple routing paths to two or more endpoint destinations - wiki
- CLI

```
# Create a forwarding rule to direct network traffic to a load balancer
gcloud compute forwarding-rules create
```

Choose the Load Balancer

Based on docs

- **Internal Load Balancer**

  distribute traffic to instances inside of Google Cloud
  choose based on Traffic type

  - Internal HTTP(S) Load Balancing - docs
    - Regional
  - Internal TCP/UDP Load Balancing - docs
    - Regional

- **External Load Balancer**

  distribute traffic coming from the internet to your VPC
  choose based on zone and traffic type
  if UDP traffic: use External TCP/UDP

  - External HTTP(S) - docs
    - Global
  - SSL *Proxy* - docs

        - Global
    - TCP *Proxy* - docs
        - Global
    - External Network TCP/UDP - docs
        - Regional

- Note
    - Secure Sockets Layer
        - SSL operates directly on top of the transmission control protocol (TCP) - source - eli5
        - SSL can use TCP [1] to transport SSL records, and so SSL relies on TCP as a service - source
    - Health Check
        - Load Balancer can use health checking mechanisms - docs

Firewall

- ⭐ Each VPC network implements a distributed virtual firewall- doc
- let you allow or deny connections to or from your VM instances - docs
    - You must specify VPC and applies to incoming (ingress) or outgoing (egress) connection, not both
- Every network has **two** implied firewall rules that permit outgoing connections and block incoming connections.
- ⭐ Default rules: link
    - Allow connection between VM inside same network, and ICMP
- ⭐ Always blocked traffic - doc
    - Egress traffic to TCP destination port 25 (SMTP)
        - "TCP port 25 is frequently blocked by Internet Service Providers (ISPs), as an anti-spam technique since it's used in MX spamming and abuse of open proxy/relay machines." -web
    - Protocols other than TCP, UDP, ICMP, IPIP, AH, ESP, SCTP, and GRE to external IP addresses of Google Cloud resources
- CLI

```
# Create a Compute Engine firewall rule
gcloud compute firewall-rules create [NAME] [--network=SUBNET_NAME] --destination-ranges[CIDR_RANGE] [--direction]
```

Cloud Armor

- Help protect your applications and websites against denial of service and web attacks. - doc
- DDoS protection, hybrid and multicloud support, IP-based and geo-based access, Adaptive protection (custom ML model trained)

Storage

Storage Types

Source from Google cloud

- **Block storage**
    - Is the traditional storage type for Vm - netapp
        - Stores chunk of raw data linearly in constant size blocks
    - G. product = Persistent disk, Local SSD
- **Object storage**
    - World-wide storage and retrieval of any amount of data at any time - docs
        - Link a key (e.g. URL) to the specific object and there is no hierarchy - reddit
    - G. product = GCS
- **Cache**
    - In-memory data with faster access, data are wiped with the Vm reboot (if not stored), can suffer of cache invalidation
    - G. product = RAM - Local SSD - Memorystore

GCP storage services

- **Cloud Datastore / firestore**
    - Highly scalable NoSQL *document* database, transactional, terabytes+
    - ⭐ Firestore ****is the newest version of Datastore
    - Accelerate development of *mobile*, web, and IoT apps with direct connectivity to the database - doc
    - ⭐ Datastore use GQL language - doc
    - CLI

```
# To export all kinds in the exampleNs namespace in the exampleProject project to the exampleBucket
gcloud datastore export gs://exampleBucket --namespaces='exampleNs' --project='exampleProject'
```

- Filestore - doc
    - Managed NFS Network File System (NFS) - docs
        - Allowing client computer to access files over a computer network much like local storage is accessed - wiki
    - ⭐ Filestore vs GCS
        - "(filestore) provide high-performance file storage capabilities to applications running on Compute Engine and Kubernetes Engine instances" - so

- Memorystore - doc
  - Basically, managed redis service
  - Reduce latency with scalable, secure, and highly available in-memory service for Redis and Memcached.
    - 100% compatible with open source Redis and Memcached
  - ⭐ Max size: 300 GB - doc
  - CLI - doc

```
# Create a Memorystore for Redis instance:
gcloud redis instances create myinstance --size=2 --region=us-central1 \
    --redis-version=redis_5_0
```

- **BigTable**
  - Scalable NoSQL database for large analytical and operational workloads.
  - low latency, no transactional, wide-column store, no SQL-like queries, expose Apache HBase API, petabytes+
  - Structure
    - instance - doc
      - container for up to 4 Bigtable clusters
      - Instances have one or more clusters, located in different zones
    - Clusters - doc
      - Bigtable service in a specific location
    - Nodes - doc
      - compute resources that Bigtable uses
      - Each cluster in an instance has 1 or more nodes
  - CLI - docs

```
# Install the cbt tool
gcloud components install cbt

# Create an instance
cbt createinstance <instance-id> <display-name> <cluster-id> <zone> <num-nodes> <storage-type>

# Create a table
cbt createtable <table-id>

# Count rows in a table
cbt count <table-id>
```

- **Cloud Storage (gcs)**
  - objects are immutable, object versioning**,** petabytes+
  - access types: Uniform (recommended), fine-grained (use deprecated ACL) - doc
  - GCS best practices - doc
  - Sharing and collaboration - doc
  - ⭐ Signed URLs
    - URL that provides limited permission and time to make a request - doc
      - “allowing users without credentials to perform specific actions on a resource”
    - Each signed URL is associated to a service account
    - The most common requests for signed URLs are object uploads and downloads
  - Storage classes - doc
    - Cold options:
      - Nearline Storage: read or modify on average *once per month* or less
      - Coldline Storage: read or modify at most *once a quarter*
      - Archive Storage: less than once a year
    - ⭐Change storage class
      - Note: is a *rewriting* process, you don’t change the “original” bucket class
      - ⭐ Note: use *gsutil* not gcloud
        - The gsutil command is used only for Cloud Storage. - so

```
# Rewrites a source object to a destination object.
gsutil -m rewrite -s coldline gs://your-bucket/**

# Create a bucket
gsutil mb gs://BUCKET_NAME
```

  - **Locations**
    - Region: Lowest latency within a single region
    - Multi-regions: Highest availability across largest area
    - Dual-regions: High availability and low latency across 2 regions
    - ⭐ There is no one-step solution for moving objects from being regional to multi-regional - stack
  - ⭐Versioning
    - When use versioning, latest object version is called *live version* - doc
- **Cloud SQL**

- Relational SQL db, transaction, replica service, terabyte+
- MySQL, PostgreSQL, and (ms) SQL Server
- ⭐ CLI

```
# Updates the settings of a Cloud SQL instance
gcloud sql instances patch [NAME] [--backup-start-time] [--backup-location]

# Commands for working with backups of Cloud SQL instances
gcloud sql backups [create/delete/describe/list/restore]
```

- **Cloud Spanner**
  - Relational SQL db, horizontal scaling, petabytes+
- **BigQuery**
  - Datalake for data warehousing (OLAP), analyze the data
  - CLI

```
# Import data to bq
bq load --autodetect --source_format=FORMAT DATASET.TABLE PATH_TO_SOURCE
```

Compute services

Compute engine (VM)

- IaaS - VM on demand
- IP address - doc
  - To locate the *external* (and internal) VM IP, you should use:
    `gcloud compute instances list`
  - VM, if allowed, can receive and external IP that is mapped to its internal
- ⭐ Quotas - doc
  - You have two quotas: *project* and *regional*
    - Project:
      - Cap for a specific project, check with:
        `gcloud compute project-info describe --project PROJECT_ID`
    - Regional
      - VM quotas are managed at the regional level
        `gcloud compute regions describe REGION`
  - Note:
    "Quotas do not guarantee that resources are always available"
- Storage:
  - Local SSDs - doc
    - Best performances: physically attached to the server that hosts your VM instance
    - 375 GB in size each
    - can attach 24 for 9 TB totals
- shutdown scripts - doc
  - have a limited amount of time to finish running
    (Preemptible instances: 30s)
  - You can directly provide the script from Console using *shutdown-script* metadata key
- **Instance groups**
  - collection of VM that you can manage as a single entity - doc
  - Two kinds
    - **Managed instance groups (MIGs)**
      - ⭐ multiple identical VMs, workloads scalable and highly available
      - create VM from *instance template* and optional *Stateful configuration* (e.g. disks)
      - Two types:
        - zonal MIG
          - deploys instances to a single zone
        - regional MIG
          - deploys instances to multiple zones across the same region
    - **Unmanaged instance groups**
      - can contain heterogeneous instances, you need to manage it
      - do not offer autoscaling, autohealing, rolling update support, multi-zone support
      - ⭐"Use unmanaged instance groups if you need to apply load balancing to groups of heterogeneous instances, or if you need to manage the instances yourself." - doc
- Save money
  - Preemptible
    - ⭐ Preemptible instances can't live migrate to a regular VM instance - doc
  - Committed use discounts
    - If some workload will be (almost) always required, you can commit some VM for 1 up to 3 years and receive a discount
- Shielded VM - doc
  - ⭐verifiable integrity of your Compute Engine VM instances, prevent malware or rootkits
  - use of Secure Boot, virtual trusted platform module (vTPM)-enabled Measured Boot, and integrity monitoring.
- Cloud Console
  - ⭐To restart a VM, you use a button named *reset*
  - You can order VM instances by Labels, status, zone, in use by, IP,
  - ⭐ You can filter VM instances by Labels, status, Member of managed instance group, IP, VM proprieties
- GPU
  - ⭐ You must set your GPU instances to stop for host maintenance events - doc
- Snapshot
  - ⭐ Compute Engine uses incremental snapshots so that each snapshot contains only the data that has changed since the previous snapshot. - doc

        - To make snapshot you should get the *Compute Storage Admin role* - doc
            - Permissions to create, modify, and delete disks, images, and snapshots.
- Disks
    - The disks are *regional*, and you can enable *regional replication*
        - Regional: the disk will be replicated synchronously across two zones within a region
    - Use case: Copy a VM from a zone to another in the same region
        - *gcloud* to copy the the disk to the new zone, then create a new VM from that disk
        - More on this: Moving an instance between zones - doc

GKE - Google Kubernetes Engine

- Run containerized application on managed environment
- build on top of Compute engine
- Regional cluster
    - You specify one cluster, and GCP replicate the settings on all the zones
    - Problem: pay resources multiplied per the number of zones available
- Multi zone cluster
    - Can choose more than one zone of a region
        - You save money (e.g. can choose 2 zones instead of 3)
            - And you don't need to maintain same nodes on both zones,
              in case of zone failure, this led to a potential
    - Problem: the master is only on the primary zone, and is alone. If the zone die, the master die
- Auto provisioning - doc
    - GCP try to understand the resources required for a pod, and create on-demand a nodepool with enough resources to accomplish the pod
        - Adapt the nodepool on demand on pod requirements
- Binary Authorization
    - Deploy only trusted containers on Google Kubernetes Engine.

Cloud Run

- In a nutshell: you give Google's Cloud Run a Docker container containing a webserver.
  Google will run this container and create an HTTP endpoint. - medium
- ⭐ Could be easily confused with App engine (in particular with App engine flex)
    - here some reddit useful comments
        - "AppEngine can only be deployed to a single region."
        - Cloud run allows you to deploy a service to every region within a single project making your API truly global, all within a single project.

App Engine

- ⭐ Basics:
    - One Application per project
    - Application can contain multiple *Services:*
      logical components that can securely share App Engine features and communicate
    - Each Service change create a new *Version*
    - Each *Version* run on a machine called *Instance*
        - Resident Instances:
          even when you've scaled to zero, these instances will still be alive. - medium
        - Dynamic instances:
          create instances with automatic scaling - doc
- PaaS - Run code in the cloud without worry about the infrastructure
- you tell Google how your app should be run.
  The App Engine will create and run a container from these instructions.
    - e.g. specify a *app.yml* with:

      runtime: nodejs12
      entrypoint: node build/server.js
- Basic features
    - Can't write on local disk, you can test the app locally, fit well with microservice architecture
    - ⭐ Only one App Engine per Project - doc
    - ⭐ Limits - doc
        - Maximum services per application: 5
        - Maximum versions per application: 5
        - Maximum instances per version with manual scaling: 20
- **environments** - docs
    - Standard environment
        - code in specific version of specific languages, faster startup (sec)
    - Flexible environment
        - provide your docker, slower startup (min)
- ⭐ **Locations**
    - App Engine is regional, You cannot change an app's region after you set it. - doc
- **Services**
    - You can deploy multiple services on one App Engine inside a single project using the *service*
    - "An App Engine app is made up of a single application resource that consists of one or more services." - doc
        - "Within each service, you deploy versions of that service"
    - **Limits**
        - Maximum services per app - Free App 5 - Paid App 105
        - Maximum versions per app - Free App 15 - Paid App 210

- Usage

```
# Create an App Engine application (Region required)
gcloud app create

# Deploy to App Engine
gcloud app deploy [YAML]

# Deploy but not use the new version
gcloud app deploy [YAML] --no-promote

# Sets the traffic split of versions across a service or a project.
gcloud app services set-traffic [SERVICE] --split-by [cookie, ip, random] --splits [\proportion of traffic should go t

# Migrate traffic to new service
gcloud app services set-traffic [SERVICE] --migrate [\attempt to automatically migrate traffic from the previous versi
```

  - If split with cookie, cookie name is GOOGAPPUID - doc

- ⭐ Scaling - doc
  - Basic scaling**:**
    App Engine attempts to keep your cost low, even though that may result in higher latency as the volume of incoming requests increases
  - Automatic Scaling:
    each instance in your app has its own queue for incoming requests. Appengine automatically handle the increasing load

- ⭐ Instance
  - The instance class determines the amount of memory and CPU available to each instance - doc (like VM instance type)

- ⭐ HTTP url - doc
  - *https://**PROJECT_ID**.**REGION_ID**.r.appspot.com*

- app.yaml - doc
  - You configure your App Engine app's settings in the app.yaml file.
  - Some interesting yaml keys are:
    - *api_version*: Required
    - *default_expiration*: Sets a global default cache period for all static file handlers for an application
    - *env_variables:* define environment variables
    - *includes*: include other the configuration file
    - *instance_class*
    - *libraries:* - deprecated, use *requirements.txt* to specify Python lib
    - *threadsafe*: required, Configures your application to use concurrent requests
    - *version*: - better configure with CLI
    - ***automatic_scaling***
      - [min/max]_instances
      - ⭐ *max_concurrent_requests*: n^ of concurrent requests an automatic scaling instance can accept before the scheduler spawns a new instance (Default: 10, Maximum: 80).
      - *max_idle_instances*: maximum number of idle instances that App Engine should maintain for this version.
    - **basic_scaling**
      - max_instances: ⭐ Note *min_instances* value doesn't exist!
      - *idle_timeout*: instance will be shut down this amount of time after receiving its last request

Cloud functions

- Function as a Service - Completely serverless execution environment
- use for (short) code that responds to events
- Cloud Functions (CF) vs App engine (AE) - stackoverflow
  - CF limited to Node.js, Python, Go, Java, .NET Core, and Ruby.
  - CF designed for standalone pieces
  - CF pay per call, AE call per time
  - ⭐ Cloud function for simple isolated functions, otherwise app engine
- ⭐ Settings
  - Runtime - doc
    - Triplet: Runtime (e.g. Python3.8), Base Image (e.g. Ubuntu), RuntimeID (e.g. python38)
  - Timeout - doc
    - Function execution time is limited by the timeout duration.
      Default 1m, max 9m.
  - Memory - doc
    - use the *–memory* flag
    - up to 4.096 MB, default 128MB
- Upload the code - doc
  - Inline editor (Cloud Console inline editor)
  - ZIP upload (with this code structure)
    - ZIP from Cloud Storage:
  - Cloud Source repository
- ⭐ CLI

```
# Deploy a function
gcloud functions deploy hello_get --runtime python38 --trigger-http --allow-unauthenticated

# Triggers available
--trigger-bucket # Every change in files in this bucket will trigger function execution.
--trigger-http   # Function will be assigned an endpoint
```

```
--trigger-topic  # Name of Pub/Sub topic
--trigger-event  # Specifies which action (storage, firebase...) should trigger the function
--trigger-resource # Specifies which resource from --trigger-event is being observed

# Delete the function
gcloud functions delete hello_get
```

Cloud endpoints

- "Develop, deploy, protect, and monitor your APIs with Cloud Endpoints." - doc
- Cloud Endpoints Frameworks: web framework for the App Engine standard *Python 2.7* and *Java 8* runtime environments - doc
- Control who has access to your API and validate every call with JSON Web Tokens and Google API keys
  - Integration with Auth0 and Firebase Authentication - for mobile apps
- You need to choose the *computing option* - table
  - Obviously, we have *cloud run* on this list
- **Apigee**
  - "Platform for developing and managing APIs"
  - Proxy to the real be for analytics, security, etc.

Cloud Tasks

Asynchronous task execution. - doc

- [asynchronously] - execution, dispatch and delivery of a large number of distributed tasks
- Your tasks can be executed on App Engine or any arbitrary HTTP endpoint
- Is similar to **Pub/Sub** - doc
  - Both Cloud Tasks and Pub/Sub can be used to implement message passing and asynchronous integration
  - Core difference: implicit vs. explicit invocation
    - Pub/sub support *impicit invocation*:
      Publishers do not need to know anything about their subscribers
    - Cloud task *explicit invocation:*
      a publisher specifies an endpoint where each message is to be delivered.

GKE - Google Kubernetes Engine

- Why Kubernetes?
  - YAML based, easy extensible, hybrid / multi cloud, microservices app
  - Resources
    - Why is Kubernetes getting so popular? - link
    - Why (and when) you should use Kubernetes - link
    - Do I Really Need Kubernetes? - link
    - "Let's use Kubernetes!" Now you have 8 problems - link
      - "…don't use k8s unless \you really need all that massive complexity…"
- Why GKE?
  - Services managed by Google:
    - Monitoring
    - Networking
    - Some Security management tasks
- Pod
  - can contain 1+ container(s)
- Node
  - are the "real" VM with specific hardware
- Master
  - Managed by Google, connected to nodes by a network peering
  - To reach the master (used to connect the *kubetl* command to a specific cluster) you need to pass the IAM
- Workloads
  - ⭐ k8s *deployments* are reported under *Workloads* GKE page - doc
- Anthos
  - "Anthos is basically GKE that can run on-premise." - eli5
  - "provides a consistent development and operations experience for cloud and on-premises environments" - doc
  - Used for hybrid and multi cloud
- **Configurations**
  - ReplicaSet
    - maintain a stable set of replica Pods running at any given time
  - **Deployments**
    - provides declarative updates for Pods and ReplicaSets. - doc
    - Deployments Vs ReplicaSet
      - "…we recommend using Deployments instead of directly using ReplicaSets…" - doc
    - Inside the deployment Yaml, you can specify the container type under *spec.template.spec*
      - ⭐ A Deployment's rollout is triggered if and only if the *Deployment's Pod template* (that is, *.spec.template*) is changed - doc
  - Services
    - Types - doc
      - ⭐ ClusterIP
        Exposes the Service on a cluster-internal IP

        - LoadBalancer
          Exposes the Service externally using a cloud provider's load balancer

- time to live **(TTL)
  - mechanism to limit the lifetime of resource objects that have finished execution.
    TTL controller only handles Jobs for now - doc
- **CLI**
  - The gcloud command is still in *beta* (may 2021) - so use *beta* on the CLI,
    e.g. *gcloud beta container cluster create ...*
- Zone / Region
  - ⭐ You can select at creation time *Zonal* or *Regional* Location type
    - With Zonal type, you can always specify multiple zones of the same region
- ⭐ Install *kubectl* using *gcloud*

```
$ gcloud components install  kubectl
# |------base-----| |-what-| |-who-|

# Note that usually gcloud have different format:
$ gcloud compute instances  list
# |------base--| |--who--| |-what-|
```

- **Private cluster**
  - Makes your master inaccessible from the public internet
  - nodes do not have public IP addresses
  - ⭐ Nodes and masters communicate with each other using VPC peering.
  - Creation
    - you must specify a /28 CIDR range for the VMs that run the Kubernetes master components and you need to enable IP aliases
  - *privateIPGoogleAccess*
    - enables your cluster hosts, which have only private IP addresses (in private cluster), to communicate with Google APIs and services.
  - You can access to the master allowing your IP:

```
$ gcloud container clusters update private-cluster \
    --enable-master-authorized-networks \
    --master-authorized-networks [MY_IP/32]
```

- **VPC-native cluster**
  - A cluster that uses *alias IP address ranges* is called a VPC-native cluster. - doc
  - Other choiche:
    - A cluster that uses custom static routes in a VPC network is called a routes-based cluster.
- Monitoring - doc
  - HW Metrics collected: CPU, Memory, Disk
- Autoscaling - doc
  - You can enable autoscaling from console with *Enable autoscaling* checkbox
- Good resource: kubernetes-basicLearning

IAM

"create and manage permissions for Google Cloud resources" - doc

- defining who (identity) has what access (role) for which resource
  - `gcloud projects get-iam-policy my-project`
- resource isn't granted directly to the end user
  - **permissions** are grouped into **roles**, roles are granted to authenticated **members**
- What is a **member**:
  - Google Account (for end users)
    - could be gmail.com or other domains
  - Service account
  - Google group - Google Workspace
  - Cloud Identity
    - Identity as a Service (IDaaS)
    - Used to work with other identity providers (e.g. Active Directory) - doc
- What is a **Policy**
  - binds one or more members to a role - doc
- several kinds of roles in IAM
  - Basic roles
    - Roles historically available in GCP: Owner, Editor, and Viewer.
    - Try to avoid those roles.
  - Predefined roles
    - give finer-grained access control than the basic roles
  - Custom roles
    - tailor permissions you made

Hierarchy

- Google Cloud resources are organized hierarchically - doc
- Organization > Folders > Projects > Resources
- ⭐ You can set an IAM policy at any level in the resource hierarchy:
  Resources inherit the policies of all of their parent resources and overwrite or merge those policies - doc

Projects

- To create a project, you must have the *resourcemanager.projects.create* permission - doc
  - Permission included into *roles/resourcemanager.projectCreator*
- ⭐ By default, all users can create projects - doc
- ⭐ The max number of projects you can create is a quota traded with google

Stackdriver

Monitor, troubleshoot, and improve application performance on your Google Cloud environment.

- Called also “Google Cloud’s operations suite”
- To monitor the VMs, you need to install the `stackdriver-agent`
  - To do it, Google provide useful script
- Which data are collected? depend of which **agent** is installed
  - Note: agents are for both Linux and Windows - doc
  - Ops Agent - doc
    - System metrics (cpu, mem, network …)
    - Actually official GCP script install `google-cloud-ops-agent` instead `stackdriver-agent`
      - But this is due to a rebranding process (2020)
  - Logging agent - doc
    - based on *fluentd*

Quotas

From doc

- (max) Size of a log entry - 256 KB
- (max)Length of a log entry label value - 64 KB
- Retention logs *_Required* - 400 days (Not configurable)
- Retention logs *_Default* - 30 days (configurable)
- Retention logs *User-defined* - 30 days (configurable)

Cloud Source Repositories

- store, manage, and track code - doc
- Create new repo from CLI
  - `gcloud source repos create hello-world`
- Clone a repo from CLI
  - `gcloud source repos clone hello-world`

Billing

Pricing calculation

- Main resource: **Google Cloud Pricing Calculator** - web
- Total cost of ownership (TCO) - web
  - Get help from a googler to get an estimation
- ⭐ BQ query price? - docs
  - Use `bq` with `--dry_run` to estimate the number of bytes read
  - Use the G. Pricing Calculator and enter the number of bytes that are processed

Billing monitor

- Billing → Transactions page:
  - Show the GCP cost and payment history - doc

Cloud Deployment Manager

Create and manage cloud resources with simple templates.

- Like terraform, but for only GCP
  - “automates the creation and management of Google Cloud resources” - doc
- ⭐ You start with a **configuration**: a YAML file that list the *resources*
  - **Resources**
    - A configuration describes all the resources you want for a single deployment.

        - a configuration is a file written in YAML
    - Each resource must contain three components
        - name - A user-defined string to identify this resource (my-vm)
        - type - The type of the resource being deployed (compute.v1.instance)
        - properties - The parameters (zone: asia-east1-a)

- **Outputs**
    - expose key properties of your configurations or templates for other templates or users to consume
        - e.g. to get the IP of resources deployed
    - Code example:

```
mongodb.jinja
{% set MASTER = env["name"] + "-" + env["deployment"] + "-mongodb" %}
resources:
- name: {{ MASTER }}
  type: instance
  ...
outputs: # <-- [!]
- name: databaseIp
  value: $(ref.{{ MASTER }}.network[0].ip)  # Treated as a string during expansion
- name: databasePort
  value: 88
```

- You could specify **dependencies** to create a deployment timeline structure - doc
    - e.g. you need a subnet before create a VM inside it
    - Code for example:

```
resources:
- name: a-special-vm
  type: compute.v1.instances
  properties:
    ...

  metadata:
    dependsOn: # <-- [!]
    - persistent-disk-a
```

Identity-Aware Proxy

guard access to your applications and VMs - doc

- Control access to your cloud-based and on-premises applications and VMs running on Google Cloud
- ⭐ IAP lets you establish a central authorization layer for applications accessed by HTTPS, so you can use an application-level access control model instead of relying on network-level firewalls. - doc
- Implement a zero-trust access model
- Is a free services with some paid features

Cloud Logging

store, search, analyze, monitor, and alert on logging data and events from Google Cloud and Amazon Web Services. - doc

- **Access Transparency**
    - logs record the actions that Google personnel take when accessing customer content - doc
- Cloud Audit Logs - doc
    - **Admin Activity** audit logs
        - log entries for API calls or other actions that modify the configuration or metadata of resources
        - e.g. create new VM
    - **Data Access** audit logs
        - API calls that read the configuration or metadata of resources
        - ⭐ Data Access audit logs– except for BigQuery Data Access audit logs– are disabled by default because audit logs can be quite large.
    - System Event audit logs
        - log entries for Google Cloud actions that modify the configuration of resources
        - ⭐ are generated by Google systems; they are not driven by direct user action.
    - Policy Denied audit logs
        - logs when a Google Cloud service denies access to a user or service account because of a security policy violation.
        - generated by default and your Cloud project is charged for the logs storage.

Resources

- ⭐ Google developer cheat sheet - github

Sign in to comment

Styling with Markdown is supported

Sign in with GitHub