Google Associate Cloud Engineer *Cheat Sheets*

Good luck on your exam!

Google Associate Cloud Engineer CheatSheet

Exam Pro

Cloud Computing - The delivery of a shared pool of on-demand computing services over the public internet, that can be rapidly provisioned and released with minimal management effort or service provider interaction.

The 5 Characteristics of Cloud

1. On-demand Self Service - Provision resources automatically without requiring human interaction

2. Broad Network Access - Available over the network

3. Resource Pooling - Pooled resources to support a multi-tenant model allowing multiple customers to share the same applications or the same physical infrastructure

4. Rapid Elasticity - Rapidly provision and de-provision any of the cloud computing resources

5. Measured Service - Resource usage can be monitored, controlled and reported using metering capabilities

Benefits of Cloud

- **Agility -** Flexibility for provisioning resources, Innovate faster
- **Cost -** Pay as you go, Trade capital expenditure for variable expense
- **Speed -** Resources on demand, Scriptable infrastructure
- **Global -** Global data centres, Disaster recovery becomes easier, High availability
- **Security** - Always up-to-date, Physical security, Encryption at rest and in transit, Compliance

Cloud Deployment Models

- Public Cloud - 1 public cloud
- Multi-Cloud - 2 or more public clouds
- Private Cloud - on-premise cloud
- Hybrid Cloud - private cloud + public cloud

Cloud Service Models

- **Hybrid Environment** - on-premise data center + public cloud
- **IaaS:** IaaS businesses offer services such as pay-as-you-go storage, networking, and virtualization. IaaS gives users cloud-based alternatives to on-premise infrastructure, so businesses can avoid investing in expensive on-site resources.
- **PaaS:** A PaaS vendor provides hardware and software tools over the internet, and people use these tools to develop applications. PaaS users tend to be developers.

Google Associate Cloud Engineer CheatSheet

Exam Pro

- **SaaS:** SaaS platforms make software available to users over the internet, usually for a monthly subscription fee.
- **On-premise:** software that's installed in the same building as your business.

Geography and Regions

Zone

- A **zone** is a deployment area for Google Cloud resources within a region.
- The smallest entity in Google's global network.
- A single failure domain within a region
- Deploy closer to users for optimal latency

Region

- **Regions** are independent geographic areas that are sub-divided into zones
- For fault tolerance and high availability
- Intercommunication <5ms between zones within a region

Multi-Region

- Multi-Regions are large geographic areas, that contain two or more regions
- Allows Google services to maximize redundancy and distribution within and across regions
- High availability (geo-redundant)

Compute Service Options

Compute Engine

- Virtual Machines (VMs) called instances, Choose region and zone to deploy , You decide the operating system and the software you decide to put on it
- Use public or private images to create instances
- Pre-configured images and software packages available in Google Cloud Marketplace
- Manage multiple instances using instance groups
- Add/remove capacity using autoscaling with instance groups, Attach/detach disks as needed, Can be used with Google Cloud Storage, Use SSH to connect directly
- Considered to be IaaS

Google Associate Cloud Engineer CheatSheet

Exam Pro

Google Kubernetes Engine (GKE)

- Container-orchestration system for automating deploying, scaling, and managing containers
- Built on open-source Kubernetes
- Flexibility to integrate with on-premise Kubernetes
- Uses Compute Engine instances as nodes in a cluster.
- A cluster is a group of nodes or Compute Engine instances
- Considered Container as a Service (CaaS)

App Engine

- Fully managed, serverless platform for developing and hosting web applications at scale (PaaS)
- Provisions servers and scales your app instances based on demand
- Build your app in Go, Java, .NET, Node.js, PHP, Python, or Ruby
- Connect with other Google services seamlessly
- Integrates with Web Security Scanner to identify threats

Cloud Functions

- Serverless execution environment for building and connecting cloud services
- Simple, single-purpose functions that are attached to events
- Triggered when an event being watched is fired
- Your code executes in a fully managed environment
- No need to provision any infrastructure
- Cloud Functions can be written using JavaScript, Python 3, Go, or Java runtimes

Cloud Run

- Fully managed compute platform for deploying and scaling containerized applications quickly and securely
- Built upon an open standard Knative
- Abstracts away all infrastructure management
- Known as serverless for containers
- Any language, any library, any binary, Considered Function as a service(FaaS)

Google Associate Cloud Engineer CheatSheet

Exam Pro

Storage Options

Cloud Storage

- Consistent, scalable, large-capacity, highly durable object storage
- 11 9's Durability (99.999999999%)
- Unlimited storage with no minimum object size
- Use Cloud Storage for content delivery, data lakes, and backup
- Available in different storage classes and availability

Storage Classes

Standard

- Maximum availability and no limitations

Nearline

- Low-cost archival storage
- Accessed <1/month

Coldline

- Even lower-cost archival storage
- Accessed <1/quarter

Archive

- Lowest-cost archival storage
- Accessed <1/year

Availability

Region

- Single Region

Dual-region

- Pair of regions

Multi-region

- Large geographic area

Google Associate Cloud Engineer CheatSheet

Filestore

- Fully managed NFS file server
- NFSv3 compliant
- Store data from running applications
- Use with VM instances and Kubernetes clusters

Persistent Disks

Durable block storage for instances

Standard – Regular standard storage at a reasonable price

Solid State (SSD) - Lower latency/higher IOPS

Both options are available in zonal and regional options

Cloud SQL

- Fully managed database service
- PostgreSQL, MySQL, and SQL Server
- High availability across zones

Cloud Spanner

- Scalable relational database service
- Support transactions, strong consistency and synchronous replication
- High availability across regions and globally

Bigtable

- Fully managed, scalable NoSQL database
- High throughput with low latency
- Cluster resizing without downtime

Datastore

- Fast, fully managed, serverless, NoSQL document database
- For mobile, web and IoT apps
- Multi-region replication and ACID transactions

Google Associate Cloud Engineer CheatSheet

Firestore
- NoSQL, realtime database
- Optimized for offline use
- Cluster resizing without downtime

Memorystore
- Highly available in-memory service for Redis and Memcached
- Fully Managed

Virtual Private Cloud (VPC)
- Virtualized network within Google Cloud
- Core networking service
- Global resource
- Each VPC contains a default network
- Additional networks can be created in your project, but networks cannot be shared between projects.

Firewall Rules
- Govern traffic coming into instances on a network
- Default network has a default set of firewall rules
- Custom rules can be created

Routes
- Advanced networking functions for your instances
- Specifies how packets leaving an instance should be directed

Load Balancing
- Distributing Workloads across multiple instances

HTTP(S) Load Balancing
- Distribute traffic across regions to ensure that requests are routed to the closest region or, in the event of a failure or over-capacity, to a healthy instance in the next closest region.
- Distribute traffic based on content type

Google Associate Cloud Engineer CheatSheet

Network Load Balancing
- Distribute traffic among server instances in the same region based on incoming IP protocol data, such as address, port, and protocol

Google Cloud DNS
- Publish and maintain DNS records by using the same infrastructure that Google uses.
- Work with managed zones and DNS records through the CLI, API, or SDK

Cloud VPN
- Connect your existing network to your VPC through an IPsec connection.

Direct Interconnect
- Connect an existing network to your VPC using a highly available, low-latency, enterprise-grade connection.

Direct Peering
- Exchange internet traffic between your business network and Google at one of Google's broad-reaching edge network locations

Carrier Peering
- Connect your infrastructure to Google's network edge through highly available, lower-latency connections by using service providers

Service-level resources
- Compute Instance VM's
- Cloud Storage buckets
- Cloud SQL databases

Account-level resources
- Organization
- Folders
- Projects

Resource Hierarchy
- Configure and grant access to the various resources

Resource Hierarchy Structure
- Resources are organized hierarchically using a parent/child relationship
- Designed to map organizational structure to Google Cloud

- Better management of permissions and access control
- Policies controlled by IAM
- Access control policies and configuration settings on a parent resource are inherited by the child
- Each child object has exactly one parent.

Domain (cloud level)
Organization (root node)
Folders - Grouping mechanism and isolation boundary
Projects - Core organizational component
Resources - Any service-level resource
Labels - Categorize resources
Committed Use Discounts (CUD's)

- Discounted prices when you commit to using a minimum level of resource for a specified term
- 1- or 3-year Commitment

Commitment Types - The commitment fee is billed monthly
Spend-based commitment

- Discount for a commitment to spend a minimum amount for a service (hours) in a particular region
- 25% discount for 1 year – 52% discount on a 3 year
- Available for Cloud SQL database instances and Google Cloud VMWare Engine
- Applies only to CPU and memory usage

Resource-based commitment

- Discount for commitment to spend a **minimum amount for Compute Engine** resource in a particular region.
- Available for vCPU, Memory, GPU and Local SSD
- 57% discount for most resources
- 70% for memory-optimized machine types
- For use across Projects

Google Associate Cloud Engineer CheatSheet

Sustained-use discounts

- **Automatic discounts** of running Compute Engine resources a significant portion of the billing month
- **Applies to VCPUs and memory** for most Compute Engine instance types
- Includes VM's created by **GKE**
- **Does not apply** to App Engine flexible, Dataflow and E2 machine types

GCP Pricing Calculator – Quick estimate of what your usage will cost on Google Cloud

Cloud Billing Budgets

- Enables you to track your actual Google Cloud spend against your planned spend
- Budget alert threshold rules that are used to trigger email notifications to help you stay informed about your spending

Billing Export

- Billing export enables granular billing data (such as usage, cost details, and pricing data) to be exported automatically to BigQuery for detailed analysis
- Not retroactive
- Daily cost detail data
- Pricing data

Cloud SDK

Set of command line tools that allow you to manage resources through the terminal

- gcloud
- gsutil
- bq
- Kubectl

A **user account** is a Google account that allows end users to authenticate directly to your application. For most common use cases on a single machine, using a user account is best practice.

A **service account** is a Google account associated with your GCP project and not a specific user. A service account can be used by providing a service account key to your application and is recommended to script Cloud SDK tools for use on multiple machines.

Google Associate Cloud Engineer CheatSheet

Gcloud Init - Authorizes access and performs other common Cloud SDK setup steps.
gcloud auth login - Authorize access for gcloud with Google user credentials.
Gcloud config - Allows you to configure accounts and projects.
gcloud components - Allow you to install, update and delete the components of the sdk
Principle of Least Privilege
A user, program, or process should have only the bare minimum privileges necessary to perform its function
Identity and Access Management (IAM)
You manage access control by defining who (identity) has what access (role) for which resource. This also includes organizations, folders, and projects.
A **policy** is a collection of bindings, audit configuration, and metadata.
A **binding** specifies how access should be granted on resources. It binds one or more members with a single role and any context-specific conditions that change how and when the role is granted.
The **metadata** includes additional information about the policy, such as an etag and version to facilitate policy management.
The **AuditConfig** field specifies the configuration data for how access attempts should be audited.
Google Account - Any email address that's associated with a Google Account, including gmail.com or other domains.
Service Account - An account for an application instead of an individual end user.
Google Groups - A named collection of Google Accounts and service accounts
G Suite Domain - Google Accounts that have been created in an organization's G Suite account
Cloud Identity Domain - Google Accounts in an organization that are not tied to any G Suite applications or features
AllAuthenticatedUsers - A special identifier that represents all service accounts and all users on the internet who have authenticated with a Google Account
AllUsers - A special identifier that represents anyone who is on the internet, including authenticated and unauthenticated users
Roles

- This is a named collection of permissions that grant access to perform actions on Google Cloud resources.
- You cannot grant a permission to the user directly
- You grant a role to a user and all the permissions that the role contains.

Google Associate Cloud Engineer CheatSheet

Permissions

- Determines what operations are allowed on a resource
- Correspond one-to-one with REST API methods
- Not granted to users directly

E.g., compute.instances.list

Primitive - Roles historically available in the Google Cloud

- Owner
- Editor
- Viewer

Avoid using these roles if possible

Predefined - Finer-grained access control than the primitive roles

Custom - Tailor permissions to the needs of your organization

Conditions

- Used to define and enforce conditional, attribute-based access control for Google Cloud resources.
- Conditions allow you to choose granting resource access to identities only if configured conditions are met
- When a condition exists, the access request is only granted if the condition expression = true

Metadata

To help prevent a race condition when updating the policy, IAM supports concurrency control through the use of an etag field in the policy

Audit Config

Determines which permission types are logged, and what identities, if any, are exempted from logging

Policy Limitations

- 1 policy per resource (including organizations, folders, projects)
- 1500 members or 250 Google groups per policy
- Up to 7 minutes for policy changes to fully propagate across GCP
- Limit of 100 conditional role bindings per policy

Google Associate Cloud Engineer CheatSheet

Conditions - Condition attributes are either based on resource or based on details about the request (timestamp, originating/destination IP address)

Condition Limitations

- Limited to specific services
- Primitive roles are unsupported
- Members cannot be allUsers or allAuthenticatedUsers
- Limit of 100 conditional role bindings per policy
- 20 role bindings for same role and same member

AuditConfig Logs

Specifies the audit configuration for a service. The configuration determines which permission types are logged, and what identities, if any, are exempted from logging. An AuditConfig must have one or more AuditLogConfigs.

A **service account** is a special kind of account used by an application or a virtual machine (VM) instance, not a person.

An application uses the service account to authenticate between the application and GCP services so that the users aren't directly involved

A special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs.

Service Account types

- User-managed, User created, You choose the name

Default

- Using some GCP services create user-managed service accounts
- Automatically granted the Editor role for the project

Google-managed

- Managed by Google, and they are used by Google services
- Some are visible, some hidden
- Name ends with "Service Agent" or "Service Account”

Service Account Keys

Key Management – None, All handled by Google

User managed

Key Management - Key storage, Key distribution, Key revocation, Key rotation, Protecting the keys from unauthorized users, Key recovery

Google Associate Cloud Engineer CheatSheet

Access scopes

- Service Account scopes are the legacy method of specifying permissions for your instance
- And they are used in substitution of IAM roles
- These are used specifically for default
- Or automatically created service accounts
- Based on enabled API's

Cloud Identity is an Identity as a Service (IDaaS) solution that centrally manages users and groups. This would be the sole system for authentication and that provides a single sign-on experience for all employees of an organization to be used for all your internal and external applications.

Device management - lets people in any organization access their work accounts from mobile devices while keeping the organization's data more secure.

Security - Helps by applying security best practices along with being able to deploy 2SV for the whole company along with enforcement controls and can also manage passwords to make sure they are meeting the enforced password requirements automatically.

Single Sign on - With single sign-on (SSO), users can access many applications without having to enter their username and password for each application

Reporting - This covers audit logs for logins, groups, devices and even tokens. You are even able to export these logs to BigQuery for analysis. You can then create reports from these logs that cover security, applications and activity.

Directory Management - Provides profile information for users in your organization, email and group addresses, and shared external contacts in the Directory. Using Google Cloud Directory Sync (GCDS), you can synchronize the data in your Google Account with your Microsoft Active Directory or LDAP server. GCDS doesn't migrate any content (such as email messages, calendar events, or files) to your Google Account. You use GCDS to synchronize all your users, groups, and shared contacts to match the information in your LDAP server.

Google Cloud Directory Sync is a free Google-provided tool that implements the synchronization process and can be run either on Google Cloud or in your on-premises environment. Synchronization is one-way so that Active Directory remains the source of truth.

Policy Management

- To grant access to all projects in your Organization, use an organization-level policy
- Grant roles to a Google group instead of individual users where possible
- When granting multiple roles to a particular task, create a Google group instead

Google Associate Cloud Engineer CheatSheet

iPV4 which is the original version of the internet protocol that first came on the scene in 1981
iPV6 is a newer version designed in 2017 to deal with the problem of ipv4 address exhaustion (useable ips)
Private IP addresses - Defined by standard RFC1918
Single Class A - 10.0.0.0 – 10.255.255.255 - 16,777,216 addresses
16 Class B - 172.16.0.0 – 172.31.255.255 - 1,048,576 addresses
256 Class C - 192.168.0.0 – 192.168.255.255 - 65,536 addresses
Classless Inter-Domain Routing (CIDR)
With CIDR based networks, you aren't limited to only these three classes of networks
Class A B and C have been removed for something more efficient which Will allow you to create networks in any one of those ranges.
Cider ranges are represented by it's starting IP address called a network address followed by what is called a prefix which is a / and then a number
IP - TCP/UDP
A packet is the basic unit of information in network transmission. Most networks use **TCP/IP** as the network protocol, or set of rules for communication between devices, and the rules of TCP/IP require information to be split into packets that contain a segment of data to be transferred along with the protocol and its port number, the originating address and the address of where the data is to be sent.
UDP is another protocol that is sent with IP and is used in specific applications.
Virtual Private Cloud (VPC)

- Virtualized network within Google Cloud
- A VPC is a Global resource
- Encapsulated within a Project
- VPC's do not have any IP address ranges associated with them
- Firewall rules control traffic flowing in and out of the VPC
- Resources within a VPC can communicate with one another by using internal (private) IPv4 addresses
- Support only for IPv4 addresses
- Each VPC contains a default network
- 2 Network types: **Auto Mode** or **Custom Mode**

Google Associate Cloud Engineer CheatSheet

Exam Pro

Subnets

- A subnetwork of a VPC
- Each VPC network consists of one or more subnets and each subnet is associated with a region
- The name or region of a subnet cannot be changed after you have created it
- Primary and secondary ranges for subnets cannot overlap with any allocated range

Routing

- Routes define the network traffic path from one destination to the other
- In a VPC routes consists of a single destination (CIDR) and a single next hop
- All routes are stored in the routing table for the VPC
- Each packet leaving a VM is delivered to the next hop of an applicable route based on a routing order

Routing Types

System-generated – Default, Subnet Route

Custom Routes - Static Route, Dynamic Route

Default Route

- Path to the Internet
- Path for Private Google Access
- Can be deleted only by replacing with custom route
- Lowest priority

Subnet Route

- Routes that define paths to each subnet in the VPC
- Each subnet has at least one subnet route whose destination matches the primary IP range of the subnet
- When a subnet is created, a corresponding subnet route is created for both primary and secondary IP range
- Cannot delete a subnet route unless you modify or delete the subnet

Static Route

- Can use the next hop feature
- Can be created manually

- Static routes for the remote traffic selectors are created automatically when creating Cloud VPN tunnels

Dynamic Route

- Managed by one or more Cloud Routers
- Dynamically exchange routes between a VPC and on-premises networks
- Destination IP ranges outside the VPC network
- Used with dynamically routed VPNs and Interconnect

Subnet routes are considered first because Google Cloud requires that subnet routes have the most specific destinations matching the IP address ranges of their respective subnets

VM instances that only have internal IP addresses can use **Private Google Access**. They can reach the external IP addresses of Google APIs and services

Internal IP addresses are not publicly advertised. They are used only within a network. Every VPC network or on-premises network has at least one internal IP address range. In Google Cloud you do this by defining a subnet range and Google will automatically reserve 3 IP's, as we discussed earlier.

You can assign an **external IP address** to an instance or a forwarding rule if you need to communicate with the internet, with resources in another network, or need to communicate with a public Google Cloud service

VPC firewall rules let you allow or deny connections to or from your virtual machine (VM) instances based on a configuration that you specify. And these rules apply to either incoming connections or outgoing connections, but never both at the same time.

VPC Peering

- Private connectivity across two VPC networks (RFC 1918)
- Peer across the same or different projects and organizations
- Reduces network latency
- Increases network security
- Reduces network costs

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.

VPC Flow Logs records a sample of network flows sent from and received by VM instances, including instances used as GKE nodes. These logs can be used for network monitoring, forensics, real-time security analysis, and expense optimization.

When you enable VPC Flow Logs, you enable for all VMs in a subnet.

Google Associate Cloud Engineer CheatSheet

Exam Pro

Record format - Log records contain base fields, which are the core fields of every log record, and metadata fields that add additional information. Metadata fields may be omitted to save storage costs. Base fields are always included and cannot be omitted.

Domain Name system (DNS)

A global decentralized distributed database that lets you store IP addresses and other data and look them up by name.

This system uses human readable names like www.google.com and translates it into a language that computers understand which are numeric IP addresses

DNS resource records (RR) are the basic information elements of the Domain Name System. They are entries in the DNS database which provide information about hosts. These records are physically stored in the Zone Files on the DNS server. This lesson will go through some of the most commonly used DNS records that we will be coming across throughout the course. So with that being siad, let's dive in.

Name Server (NS) - This record identifies which DNS server contains the current records for a domain.

An **A record (or Address Record)** points a domain name to an IP address

A **CNAME record,** short for Canonical Name record is a type of resource record that maps one domain name to another.

A **TXT record** (short for text record) is a type of resource record that provides text information to sources outside your domain, that can be used for a number of arbitrary purposes.

A **DNS MX record** also known as the 'mail exchange' record is the resource record that directs email to a mail server.

A **DNS pointer record** (PTR for short) provides the domain name associated with an IP address.

A **Start of authority resource record (SOA)** is created for you when you create your managed zone specifies authoritative information including global parameters about a DNS zone

Network Address Translation - NAT

- Translates local private IP(s) to public IP(s) before transferring packets
- Originally designed to deal with the scarcity of free IPv4 addresses
- IPv6 networks do not require NAT as there are no shortage of addresses
- Provides security and privacy

Types of NAT

- Static NAT - 1 private IP to 1 public IP
- Dynamic NAT - 1 private IP to 1 public IP in pool of public addresses
- Port Address Translation (PAT) - Multiple private IPs to 1 public IP

Google Associate Cloud Engineer CheatSheet

Exam Pro

- **Cloud DNS**
- Host authoritative name servers and allow authoritative DNS lookups (DNS as a Service)
- 100% SLA - Globally Resilient

Host zones through managed name servers

- Public Zone - visible to the internet
- Private Zone - visible only within your network

Virtualization is the process of running multiple operating systems on a server simultaneously.

Paravirtualization (PV)

In this model a modified guest OS can speak directly to the Hypervisor. This involves having the operating system kernel to be modified and recompiled before installation into the virtual machine.

Hardware-assisted virtualization is a virtualization approach that enables efficient full virtualization using help from hardware capabilities, from the host CPU.

Kernel level Virtualization - Instead of using a hypervisor, it runs a separate version of the Linux kernel and sees the associated virtual machine as a user – space process on the physical host. This makes it easy to run multiple virtual machines on a single host. A device driver is used for communication between the main Linux kernel and the virtual machine.

Compute Engine

- Virtual machine = Instance (IaaS)
- Multiple instance sizes and types
- Per second billing
- Launched in a VPC network
- Host is available in a Zone
- Multi-tenant host or Sole-tenant node

Machine Configuration

- Many machine types - General, compute, memory
- Intel or AMD
- vCPU = single hardware hyper-thread on CPU
- Network throughput = 2Gbps per vCPU

Google Associate Cloud Engineer CheatSheet

Operating System

- Image – Linux or Windows
- Custom Image - Private Images (Snapshots/existing disk)
- Marketplace - OS + software

Storage

- Standard - Spinning Hard Drive
- Balanced - Solid State Drive (alternative to SSD)
- SSD - Solid State Drive
- Local SSD - Physically attached (swap disk)

Networking

- Auto, default, custom networks
- Many available regions and zones
- Ingress/egress firewall rules (IP ranges, tags, instances)
- Network load balancing
- Regional/global load balancing

Compute Engine Machine Types

Standard machine type - General-purpose

Standard - Balance of CPU and memory

High-memory - High memory to CPU ratio

High-CPU - High CPU to memory ratio

E2 - Day-to-day computing at a lower cost

N1 - Balanced price/performance across a wide range of VM shapes

Compute-optimised

C2 - Standard

- Ultra high performance for compute-intensive workloads

Memory-optimised
Ultra high-memory workloads
Shared-core machine types use context-switching to share a physical core between vCPUs for the purpose of multitasking. Different shared-core machine types sustain different amounts of time on a physical core.
In general, shared-core instances can be more cost-effective for running small, non-resource intensive applications than standard, high-memory or high-CPU machine types.
Custom machine types are ideal for:
Workloads that are not a good fit for the predefined machine types that are available to you.
Workloads that require more processing power or more memory, but don't need all the upgrades provided by the next larger predefined machine type.
It costs slightly more to use a custom machine type than an equivalent predefined machine type, and there are limitations in the amount of memory and vCPUs you can select.
Managing Instances
PROVISIONING - This is where Resources are being allocated for the instance. The instance is not running yet.
STAGING - After finishing the provisioning state, the lifecycle continues with the staging state.
RUNNING - Once the instance has left staging it will move onto the running state. This is where the instance is booting up or running and should allow you to login to the instance (either ssh or rdp) within a short waiting period due to any startup scripts or any boot maintenance tasks for the OS, but not immediately after it enters this state.
STOPPING - When it comes to stopping, Either a user has made a request to stop the instance or there was a failure. This is a temporary status, and the instance will move to TERMINATED.
TERMINATED - Touching on the last state is the terminated state and this is where A user either shut down the instance, or the instance encountered a failure. You can choose to restart the instance or delete it. Here you still pay for static IP's and disks, but like the suspending or stopping state, you do not pay for the CPU and memory resources allocated to the instance.
Shielded VM's offer verifiable integrity of your Compute Engine VM instances, so you can be sure your instances haven't been compromised by boot- or kernel-level malware or rootkits. This is achieved through using a 4-step process which is covered by Secure Boot, virtual trusted platform module (vTPM) running Measured Boot, and integrity monitoring.

Google Associate Cloud Engineer CheatSheet

VM Access

SSH

- Requires firewall rule allow - tcp:22
- Google Cloud console
- Cloudshell using CloudSDK
- OS Login (use 2SV)
- Manually creating SSH key pair

RDP

- Requires firewall rule allow - tcp:3389
- Connect using RDP
- Powershell terminal
- Requires setting Windows password
- RDP Chrome extension
- 3rd party RDP client

Live migration keeps your instances running during compute engine hosts that are in need of:
Regular infrastructure maintenance and upgrades, replacement of failed hardware, and system configuration changes

Compute Engine Pricing

- Each individual vCPU and each GB of memory is billed separately - resource based
- All vCPUs, GPUs, and GB of memory are charged by the second with a minimum of 1 minute
- Instance uptime - number of seconds between when you start an instance and when you stop an instance (terminated)

Reservations

Ensuring resources are available for when you need it

- Future increases in demand
- Planned or unplanned spikes
- Backup and disaster recovery
- Buffer

Google Associate Cloud Engineer CheatSheet

Include sustained use and committed use discounts
Apply only to Compute Engine, Dataproc and GKE VM's

Discount types

- Sustained use discounts
- Committed use discounts
- Preemptible VM's

Sustained use discounts are automatic discounts for running specific Compute Engine resources a significant portion of the billing month.
Compute Engine lets you purchase **committed use contracts** in return for deeply discounted prices for VM usage.
When you purchase a committed use contract, you purchase compute resource which is comprised of vCPUs, memory, GPUs, and local SSDs) at a discounted price in return for committing to paying for those resources for 1 year or 3 years.
Preemptible VMs are up to 80% cheaper than regular instances. Pricing is fixed you never have to worry about variable pricing. These prices can be found on the link to Instance pricing that I have included in the lesson text.

Storage Fundamentals

Block storage is a technology that is used to store data files on storage systems or cloud-based storage environments. Block storage is the fastest available storage type. It is also efficient, and reliable.

- Evenly sized blocks, Uniquely identifiable, Mountable, Bootable

File Storage is normally storage that is presented to users and applications as a traditional network file system.

- Network File System, Directory tree structure, Mountable, Not bootable

Object storage is a general term that refers to the way in which we organize and work with units of storage, called objects.

- Unstructured data, Infinitely scalable, Not mountable, Not bootable

IOPS - is a metric that stands for input/output operations per second. More value in the IOPS signifies the capability of executing more operations per second.

Persistent Disk Snapshots

- Backup and restore of persistent disks
- Global resources
- Support for zonal and regional PDs

- Incremental and automatically compressed
- Snapshots are stored in Cloud Storage
- Stored in regional or multi-regional location

Snapshot schedules

- Best practice for backups
- Must be in same region as pd

Managing Snapshots

- 1 snapshot = 10min
- Create regular schedules
- Eliminate excessive snapshots
- Set schedule to off-peak hours
- Windows - create VSS snapshots

Deployment Manager

Configuration

- Defines the structure of your deployment

Must contain resources section

- list of resources to create

3 Components

Name - A user-defined string to identify this resource and can be anything you choose from instance-1 my-vm, bowtie-instance

A **type** can represent a single API resource, known as a base type, or a set of resources, known as a composite type, that will be created as part of your deployment.

Base type: [API].[VERSION].[RESOURCE]

A **composite type** contains one or more templates that are preconfigured to work together.

type: gcp-types/[PROVIDER]:[RESOURCE]

type: gcp-types/compute-v1:addresses

Properties - The parameters for this resource type.

Google Associate Cloud Engineer CheatSheet

A **configuration** can contain templates, which are essentially parts of the configuration file that has been abstracted into individual building blocks.
A t**emplate** is a separate file that is imported and used as a type in a configuration.
A **deployment** is a collection of resources that are deployed and managed together, using a configuration.
A **manifest** is a read-only property that describes all the resources in your deployment and is automatically created with each new deployment.
Manifests are not modifiable after they have been created

Load Balancing

- Distributes user traffic across multiple instances
- Single point of entry with multiple backends
- Fully distributed and software defined
- Global and Regional
- Serve content as close as possible to users
- Autoscaling with health checks

Load Balancer Types

HTTP(S) Load Balancer

Global, proxy-based Layer 7 load balancer behind a single external IP address

SSL Proxy

Reverse proxy load balancer that distributes **SSL** traffic coming from the internet to VMs

TCP Proxy

Reverse proxy load balancer that distributes **TCP** traffic coming from the internet to VMs

TCP/UDP Network Load Balancing (after this referred to as Network Load Balancing) is a regional, pass-through load balancer.
A network load balancer distributes TCP or UDP traffic among instances in the same region.

Internal Load Balancer

Internal TCP/UDP Load Balancing distributes traffic among VM instances in the same region by using an internal IP address.
An **instance group** is a collection of virtual machine (VM) instances that you can manage as a single entity.

Google Associate Cloud Engineer CheatSheet

Managed Instance Groups (MIGs) are great for **Stateless serving workloads** such as website frontends, web servers and website applications as the application does not preserve its state and saves no data to persistent storage.
All user and session data stays with the client and makes scaling up and down quick and easy
Stateless batch: high-performance or high throughput compute workloads
Stateful workloads: use stateful managed instance groups

Autohealing

- Keeps VMs in RUNNING state
- Recreate VMs not in RUNNING state
- Application-based autohealing
- Recreate VMs when app is frozen or has crashed

Regional (multi-zone) Zonal or Regional

- Regional provides higher availability
- Zonal MIGs are in one zone only
- Google recommends regional MIGs

Load Balancing

- Load balancing can use instance groups to serve traffic
- Work together to know how much traffic can be handled
- LB health checks do not send traffic to unhealthy instances

Autoscaling

- Dynamically add or remove instances from the MIG
- Scale up to meet load demands
- Shrink as the load decreases to reduce costs

Auto-updating

- Deploy new versions of software to instances
- Update deployment happens automatically
- Perform rolling updates

Google Associate Cloud Engineer CheatSheet

An **instance template** is a resource that you can use to create VM instances and managed instance groups
Instance templates define the machine type, boot disk image or container image, labels, and other instance properties
Containers are **packages of software that contain all of the necessary elements to run in any environment**. In this way, containers virtualize the operating system and run anywhere, from a private data center to the public cloud or even on a developer's personal laptop.
A **Docker image** is a collection or stack of layers that are created from sequential instructions on a docker file.
Container Registry is a single place for you to store and manage Docker images
Kubernetes is an orchestration platform for containers
Google Kubernetes Engine (GKE) - Google Cloud has since developed a managed offering for Kubernetes providing a managed environment for deploying, managing, and scaling your containerized applications using Google infrastructure.

- Cloud Load Balancing
- Node Pools
- Automatic scaling
- Automatic upgrades
- Node auto-repair
- Logging and Monitoring

Cluster Architecture

- One or more Control Planes
- One or more Nodes
- Control Plane responsible for scheduling and management
- Nodes run containerized apps and nodes responsible for Docker runtime

Control Plane - Endpoint of the cluster
API server - Point of interaction with the cluster (API calls or kubectl)
kube scheduler - Discovers and assigns newly created pods
kube controller manager - Runs all controller processes
cloud controller manager - Runs controllers specific to the cloud provider
Etcd - Key-value store that stores the state of the cluster

Google Associate Cloud Engineer CheatSheet

The **kubelet** is **the primary "node agent" that runs on each node**. It can register the node with the apiserver using one of: the hostname
kube-proxy is the component that maintains network connectivity to the pods in a cluster
The **container runtime** is the software that is responsible for running containers.
Kubernetes supports container runtimes like **Docker** and **containerd**
The **endpoint** exposes the Kubernetes API server that kubectl uses to communicate with your cluster control plane (master).
A **node pool** is a group of nodes within a cluster that all have the same configuration.
Cluster Types
Zonal clusters have a single control plane in a single zone. Depending on your availability requirements, you can choose to distribute your nodes for your zonal cluster in a single zone or in multiple zones.
A **single-zone cluster** has a single control plane running in one zone. This control plane manages workloads on nodes running in the same zone.
A **multi-zonal cluster** has a single replica of the control plane running in a single zone and has nodes running in multiple zones.
During an upgrade of the cluster or an outage of the zone where the control plane runs, workloads still run. However, the cluster, its nodes, and its workloads cannot be configured until the control plane is available.
A **regional cluster** has multiple replicas of the control plane, running in multiple zones within a given region.
Nodes also run in each zone where a replica of the control plane runs.
Cluster Version - When you create a cluster, you can choose the cluster's specific Kubernetes version, or you can make choices about its overall mix of stability and features.
Release Channel - When you enroll a new cluster in a release channel, Google automatically manages the version and upgrade cadence for the cluster and its node pools.
Rapid - Several weeks after upstream open source GA
Regular (default) - 2-3 months after releasing in Rapid
Stable - 2-3 months after releasing in Regular
Specific version
If you know that you need to use a specific supported version of Kubernetes for a given workload, you can specify it when creating the cluster.

Google Associate Cloud Engineer CheatSheet

Cluster upgrades

Control plane and nodes do not always run the same version at all times

A control plane is always upgraded before its nodes

- Zonal - Cannot launch or edit workloads during upgrade
- Regional - Each control plane is upgraded one by one

Auto-upgrade enabled by default - best practice

Manual upgrade - cannot upgrade control plane more than one minor version at a time

- Maintenance window and exclusions available

Node and Node pool upgrades

Auto-upgrade enabled by default - best practice

Manual upgrade available

- Maintenance window and exclusions available

Pods scheduled to run on another node during upgrade

Upgrade is complete only when

- All nodes have been recreated
- Cluster is in the desired state

Surge upgrades let you control the number of nodes GKE can upgrade at a time and control how disruptive upgrades are to your workloads.

Kubernetes objects are persistent entities in Kubernetes. Kubernetes uses these entities to represent the state of your cluster.

Object spec - desired state described by you

Object status - current state described by Kubernetes

Pods are the smallest, most basic deployable objects in Kubernetes. A Pod represents a single instance of a running process in your cluster.

Pods Remains on the node until:

- The pod's process is complete
- The pod is deleted
- The pod is evicted from the node due to lack of resources
- The node fails

Google Associate Cloud Engineer CheatSheet

Exam Pro

Kubernetes starts with four initial **namespaces**:

- **Default –** For objects with no other namespace. When creating new objects without a namespace, your object will automatically be assigned to this.
- **kube-system -** For objects created by the Kubernetes system
- **kube-public -** This namespace is created automatically and is readable by all users.
- **kube-node-lease** - For the lease objects associated with each node which improves the performance of the node heartbeats as the cluster scales.

Labels are key/value pairs that are attached to resources.

Pods are ephemeral. They are not designed to run forever, and when a Pod is terminated it cannot be brought back.

A **Deployment** runs multiple replicas of your application and automatically replaces any instances that fail or become unresponsive.

Workloads

Deployments - runs multiple replicas of your app and automatically replaces any instances that fail or become unresponsive

StatefulSets - used for apps that requires persistent storage

DaemonSets - ensures that every node in the cluster runs a copy of a pod

Jobs - used to run a finite task until completion

CronJobs - similar to jobs but runs until completion on a schedule

ConfigMaps - configuration info for any workload to reference

A **service** can be defined as a logical set of pods. It is an abstraction on the top of the pod which provides a single persistent IP address and DNS name by which pods can be accessed it allows for routing external traffic into your Kubernetes cluster and used inside your cluster for more intelligent routing.

- Persistent single IP
- Internal and external
- Load balancing
- Scaling

When a **network request** is made to the **service**, it selects all Pods in the cluster matching the **service's selector**, chooses one of them, if there more than 1 with the **same label** and forwards the **network request** to it.

Service Types

- A **ClusterIP service** is the default Kubernetes service.
- It gives you a service inside your cluster that other apps inside your cluster can access.

Google Associate Cloud Engineer CheatSheet

NodePort - When you create a Service of type NodePort, you specify a nodePort value.

- The nodeport is a static port and is chosen from a pre-configured range between 30000-32767. You can specify your own value within this range

LoadBalancer - The service is exposed as a load balancer in the cluster.

- LoadBalancer services will create an internal Kubernetes Service that is connected to a cloud-providers Load Balancer

Multi-port Services

- When using multiple ports for a Service, you must give all of your ports names and if you have multiple service ports these names must be unique.

ExternalName

- Provides an internal alias for an external DNS name.
- Internal clients make requests using the internal DNS name, and the requests are redirected to the external name.

Headless - Sometimes you don't need or want load-balancing and a single service IP. In this case, you can create "headless" services by specifying "None" for the cluster IP.

- This allows you to choose other service discovery mechanisms, without being tied to Kubernetes' implementation.
- In GKE, an **Ingress object** defines rules for routing **HTTP(S) traffic** to applications running in a cluster.

An **Ingress object** is associated with one or more Service objects, each of which is associated with a set of Pods.

- To use Ingress, you must have the HTTP load balancing add-on enabled.

Network endpoint groups (NEG)

This is a configuration object that specifies a group of backend endpoints or services.

NEGs are useful for Container native load balancing where each Container can be represented as an endpoint to the load balancer.

Health Checks

- Default and inferred parameters are used if there are no specified health check parameters
- Should be explicitly defined by using a Backend Config custom resource definition (CRD)
- Anthos Ingress controller
- >1 container
- Specific port for LB health check
- Backend service's health check
- healthCheck parameter of a BackendConfig CRD referenced by service

Google Associate Cloud Engineer CheatSheet

SSL Certificates - Three ways to provide SSL certificates to an HTTP(S) load balancer:

Google-managed

- Completely managed by Google
- Do not support wildcard domains

Self-managed

- Managed and shared with Google Cloud
- Provision your own certificates
- List the certificate in annotation for use

Self-managed as Secrets

- Provision your own certificates
- Create a secret to hold the certificate
- Refer to the secret for use

Multiple certificates: specify in Ingress manifest

GKE Storage Options

If you need to connect a **database** to your cluster, consider using **Cloud SQL, Datastore or Cloud Spanner**

Object storage – Recommended to use **Cloud Storage**

Filestore is a great option for when your application requires managed **Network Attached Storage (NAS)**

If your application requires **block storage**, the best option is to use **persistent disks**

A **Docker volume** is a directory on disk or in another container.

A **Docker container** has a writeable layer, and this is where the data is stored by default.

Three ways to mount data inside a **Docker container**

1. A Docker volume is the first way to mount data and sits inside the Docker area, within the host's filesystem and can be shared amongst other containers.
2. Bind mounting is the second way to mount data and is coming directly from the host's file system.
   - Bind mounts are great for local application development yet cannot be shared across containers.
3. using tmpfs and is stored in the host's memory.
   - This way is great for ephemeral data and increases performance as it no longer lies in the container's writeable layer.

Google Associate Cloud Engineer CheatSheet

Volumes - Basic storage unit that decouples the storage from the container and tie it to the pod

Volume created when the Pod is created. Terminated when pod is terminated or deleted

Pod spec - how directory is created storage medium used directory's initial contents

Types of volumes

emptyDir

- empty directory that containers in the Pod can read and write from

ConfigMap

- provides a way to inject configuration data into Pods

Secret

- used to make sensitive data available to applications

Downward API

- used to make Downward API data available to applications

PersistentVolumeClaim

- provision durable storage to be used by applications

PersistentVolume resources are used to manage durable storage in a cluster.

Persistent Volume Claims this is a request for and claim to a PersistentVolume resource.

PersistentVolumeClaim objects request a specific size, access mode, and StorageClass for the PersistentVolume

The default **StorageClass** is used when a PersistentVolumeClaim doesn't specify a StorageClassName and can also be replaced with one of your choosing.

Persistent Volume Access

Access Modes

ReadWriteOnce

- mounted as read-write by a single node

ReadOnlyMany

- mounted as read-only by many nodes

ReadWriteMany

- mounted as read-write by many nodes

Google Associate Cloud Engineer CheatSheet

Regional persistent disks
- Replicate between 2 zones
- Can failover workloads (HA)

Zonal persistent disks
- If no zone specified, one is chosen at random
- Pods referencing disk are scheduled in same zone

Cloud VPN
- Connects your peer network to your VPC network through an IPsec VPN connection.
- IPsec tunnel over the public internet
- Encrypted by one VPN gateway, and then decrypted by the other VPN gateway.
- Regional Service
- Site to site VPN only (no site to client)
- Allows Private Google Access for on-premises hosts
- Supports up to 3Gbps per tunnel
- Dynamic and static routing
- Supports IKEv1 and IKEv2 using Shared Secret

Types of Cloud VPN

Classic VPN
- 99.9% SLA
- Static and dynamic routing
- 1 external IP address for a single interface
- Deprecating functionality in 2021

HA VPN
- Dynamic routing only
- 2 external IPs to be configured for 2 interfaces
- New default VPN

Google Associate Cloud Engineer CheatSheet

When to use Cloud VPN

- Public internet access is needed
- Peering location is not available
- Budget constraints
- High speeds/ low latency not needed
- Outgoing traffic (egress) from GCP

Cloud Interconnect

- Low latency, highly available connection between your on-premises and Google Cloud VPC networks
- Directly accessible internal IP addresses - Private Google Access
- Does not traverse the public internet
- Dedicated connection
- Not encrypted
- Expensive

Dedicated Interconnect provides direct physical connections between your on-premises network and Google's network.

Dedicated Interconnect enables you to transfer large amounts of data between your network and Google Cloud, which can be more cost-effective than purchasing additional bandwidth over the public internet.

Partner Interconnect provides connectivity between your on-premises network and your Virtual Private Cloud (VPC) network through a supported service provider.

A **Partner Interconnect connection** is useful if a Dedicated Interconnect colocation facility is physically out of reach, or your workloads don't warrant an entire 10-Gbps connection.

Direct Peering

- Direct peering connection between your on-premises network and Google's edge network
- 100 locations in 33 countries
- Direct egress pricing available
- Direct Peering connection with Google is FREE

Google Associate Cloud Engineer CheatSheet

Exam Pro

CDN Interconnect

- Enables select third-party CDN providers to establish direct peering links with Google's edge network
- Direct traffic from VPC networks to the provider's network
- Reduced pricing on egress costs

When to use Cloud Interconnect

- Prevent traffic from traversing the public internet
- Dedicated physical connection
- Extension of your VPC network
- High speed/low latency is needed - 200 Gbps
- Heavy outgoing traffic (egress) from GCP
- Private Google Access

App Engine Overview

- Fully managed, serverless platform to develop and host web apps
- PaaS service
- Code or containers - Python, Java, Node.js, Go, Ruby, PHP, or .NET
- Autoscaling based on load
- Versions - Allow for rollbacks, migrating or traffic splitting
- Support for connecting to external storage
- Standard and Flexible environments

Standard and Flexible environments

Standard

- Apps run in sandbox environment
- Specific versions of runtimes used
- Run for free or at very low cost
- Designed for sudden and extreme spikes of traffic
- Pricing based on instance hours

Google Associate Cloud Engineer CheatSheet

Flexible

- Apps run in docker containers
- Any version of runtimes used
- No free quota available
- Designed for consistent traffic
- Pricing based on VM resources
- Managed VMs

Deploying applications to App Engine is as simple as using the **gcloud app deploy** command

Managing Instances

- Automatically create and shut down instances
- Specify a number of instances to run
- Specify a scaling type

Automatic scaling

- based on metrics like request rate and response latencies

Basic scaling

- creates instances when your application receives requests

Manual scaling

- specifies the number of instances that continuously run

Traffic migration switches the request routing between the versions within a service of your application, moving traffic from one or more versions to a single new version.

You can use **traffic splitting** to specify a percentage distribution of traffic across two or more of the versions within a service.

Cloud Functions

- Serverless
- FaaS - Function as a Service
- Runtime - Python, Java, Node.js, Go, .NET core
- Event-driven

Google Associate Cloud Engineer CheatSheet

- Triggers - HTTP, Pub/Sub, Cloud Storage (Firestore, Firebase)
- Billing - time + resources provisioned (memory)
- Free Tier

Cloud Storage

- Consistent, scalable, large-capacity, highly durable object storage - not file or block
- Worldwide accessibility and worldwide storage locations
- Use for data files, text files, pictures, videos
- Excels for content delivery, big data sets and backups
- Buckets and Objects

Cloud Storage buckets

- basic container that holds your data
- Organize your data
- Access control
- Storage Classes

Hot data

Standard

- Maximum availability
- No storage duration
- Analytical workloads and transcoding
- $0.02 /GB/month

Nearline

- Low-cost for infrequently accessed data
- 30 day min. storage duration
- Data backup and data archiving
- $0.01 /GB/month
- $ Data access

Google Associate Cloud Engineer CheatSheet

Cold data

Coldline

- Very low-cost for infrequently accessed data
- 90 day min. storage duration
- Data backup and data archiving
- $0.004 /GB/month
- $$ Data access

Archive

- Lowest-cost archival storage
- 365 day min. storage duration
- Cold data storage
- Disaster recovery
- $0.0012 /GB/month
- $$$ Data access

Access Control

IAM

- standard IAM permissions
- permissions inherited hierarchically
- Recommended over ACLs
- Two levels of granularity: project or bucket level
- Roles available: Primitive, Standard, Legacy
- Legacy roles are equivalent to ACLs

Access Control List (ACL)

- defines who has access to your buckets and objects, as well as what level of access they have
- Granular permissions
- Entry = permission + scope

Google Associate Cloud Engineer CheatSheet

Signed URLs

- time-limited read/write access URL
- access the object for the duration of time you specify
- Allows users without credentials to perform specific actions on a resource

Signed Policy Documents

- specify what can be uploaded to a bucket

Objects are immutable, which means that an uploaded object cannot change throughout its storage lifetime.

Object Lifecycle Management

Rules - there are a set of rules, conditions and the action when the conditions are met. Rules are any set of conditions for any action

- Any set of conditions
- for any action

Conditions - Conditions is something an object must meet before the action defined in the rule occurs on the object

- 1 or multiple

Action - where you only have the option to delete or set storage class

- Delete
- SetStorageClass

Cloud SQL

- Fully managed, relational database service (RDBMS)
- DBaaS (Database as a Service)
- Low latency, transactional, relational db workloads
- MySQL, PostgreSQL and SQL Server - NEW
- Replication - Read Replicas
- High Availability
- On-demand and automatic backups
- Point in time recovery

- 30TB storage capacity
- Automatic storage increase
- Encryption at rest and in transit
- Billed for instance, persistent disk and egress traffic

Cloud SQL instances are not available in the same instance types as Compute Engine and are only available in the **shared-core, standard and high memory CPU types**

Storage types for Cloud SQL are only available in **HDD** and **SSDs**

You can configure it with a **Public or Private IP**, but after configuring the instance with a private IP, it cannot be changed

Connecting with a private IP is preferred when connecting from a client on a resource with access to a VPC

The **Cloud SQL proxy** allows you to authorize and secure your connections using Identity and Access Management (IAM) permissions.

In a **Cloud SQL instance**, the instance that is replicated is called the primary instance and the copies are called read replicas.

A **Cloud SQL instance** configured for High Availability (HA) is also called a regional instance and is located in a primary and secondary zone within the configured region.

If an **HA-configured instance** becomes **unresponsive**, Cloud SQL automatically switches to serving data from the standby instance. This is called a **failover**.

When the primary instance is available again, a **failback** will happen, and this is when traffic will be redirected back to the primary instance

Backups help you restore lost data to your Cloud SQL instance.

Cloud Spanner

- Fully managed relational database service that is both strongly consistent and horizontally scalable
- DBaaS (Database as a Service)
- Supports schemas, ACID transactions, and SQL queries
- Globally distributed
- Handles replicas and sharding
- Synchronous data replication
- Automatic scaling and node redundancy
- Up to 99.999% availability
- Data layer encryption, audit logging, IAM integration

- Designed for financial services, ad tech, retail and global supply chain, gaming
- Pricing: $0.90 /node/hr + $0.30/GB/mo.

To use Cloud Spanner, you must first create a Cloud Spanner instance
This instance is an allocation of resources that is used by Cloud Spanner databases created in that instance.
As Cloud Spanner gets populated with data, sharding happens which is also known as a split

Performance

- 10,000 queries QPS of reads or 2,000 QPS of writes
- 2TB of storage per node
- Add nodes to increase data throughput and QPS
- Scale nodes automatically using Cloud Monitoring metrics triggered by Cloud Functions

NoSQL Databases

Cloud Bigtable

Fully managed, wide-column NoSQL database designed for terabyte to petabyte-scale workloads that offers low latency and high throughput.

- Built for real-time app serving & large-scale analytical workloads
- Regional Service
- Automated replication
- Store large amounts of single-keyed data
- Add nodes when you need them
- Cluster resizing
- Ideal data source for MapReduce operations, High-priced

Use cases

- Time-series data
- Marketing data
- Financial data
- IoT data
- Graph data

Google Associate Cloud Engineer CheatSheet

Exam Pro

Cloud Datastore

Fully managed, highly scalable NoSQL document database built for automatic scaling, high performance, and ease of application development

- High-availability of reads and writes
- Atomic transactions
- Automatic scaling
- SQL-like query language (GQL)
- Strong and eventual consistency
- Encryption at Rest
- Being retired in favour of Cloud Firestore in 2021

Use cases

- Product Catalogs
- User profiles
- Transactions based on ACID properties

Firestore for Firebase

Flexible, scalable NoSQL cloud database to store and sync data for client and server-side development

- Serverless
- Multi-region replication
- Flexibility
- Expressive querying
- Realtime updates
- Offline support
- Secure
- Realtime Database
- Simpler version of Firestore

Firebase

A mobile app development platform that provides tools and cloud services to help enable developers to develop apps faster and more easily

Google Associate Cloud Engineer CheatSheet

Exam Pro

Memorystore

Fully managed service for either Redis or Memcached in-memory data store to build application caches

- Fully managed
- High Availability
- Scale as needed
- Secure
- Always up to date

Use cases

- Caching
- Gaming (leaderboards, user profiles)
- Stream processing

What is Big Data?

Massive amounts of data that would typically be too expensive to store, manage, and analyze using traditional database systems.

Traditional databases are not cost effective

- No flexibility for storing unstructured data
- Inability to accommodate "real time" data
- Lacks support for petabyte-scale data volumes
- Apache Hadoop & NoSQL to the rescue
- Extremely complex to deploy and manage

When this data is captured, formatted, manipulated, stored and then analyzed, can help a company make better decisions (business value).

- Gain useful insight
- Increase revenue
- Get or retain customers
- Improve operations
- Better with Machine Learning

Google Associate Cloud Engineer CheatSheet

Exam Pro

Big Data Services

Big Query

Fully managed, petabyte scale, low cost analytics data warehouse

- Serverless
- Real-time analytics insertion
- Use Standard SQL for querying
- Process external data
- Dataproc, Dataflow, Cloud Storage, Big Table, Cloud SQL, Google Drive
- Parquet, ORC, Google sheets

Data Transfer Service (DTS)

- 145 Services - Teradata, Amazon S3, Azure Blob, etc.

Run open source data science workloads

- Spark, Tensorflow, Dataflow, Apache Beam, MapReduce

Automatic backups

Automatic high availability

Data Governance and security

- Geographic data control
- Data encryption at rest and in-transit

Composer - Managed workflow orchestration service, built on Apache Airflow

Dataflow - Fully managed processing service for executing Apache Beam pipelines for batch and realtime data streaming

Dataproc - Fully managed Spark and Hadoop service

- Can be used to replace on-prem Hadoop infrastructure

DataLab - An easy-to-use interactive tool for data exploration, analysis, visualization, and machine learning.

Pub/Sub - Fully-managed, real-time messaging service that allows you to send and receive messages between independent applications.

Dataprep - Serverless, intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning

Google Associate Cloud Engineer CheatSheet

What is Machine Learning?

Functionality that enables software to perform tasks without any explicit programming or rules.

- Trained to recognize patterns in collected data using algorithmic models
- Collected data includes video, images, speech or text
- Cloud is an efficient place for ML due to the use of massive computation at scale
- Better with Big Data

What can Machine Learning do?

- Categorize images such as photos, faces, or satellite imagery
- Look for keywords in text documents or emails
- Flag potentially fraudulent transactions
- Enable software to respond accurately to voice commands
- Translate languages in text or audio

Sight

Vision - Pre-trained machine learning models that allow you to assign labels to images and quickly classify them into millions of predefined categories

Video Intelligence

- Pre-trained machine learning models that automatically recognize a vast number of objects, places, and actions in stored and streaming video

Language

Natural Language - Derive insights from unstructured text using Google machine learning

Translation - Translation enables you to dynamically translate between languages using Google's pre-trained or custom machine learning models

Conversation

Dialog Flow

- Natural language understanding platform that makes it easy to design and integrate a conversational user interface into your application or device

Speech-to-Text

- Accurately convert speech into text using Google's AI technologies

Text-to-Speech

- Enables developers to synthesize natural-sounding speech with 100+ voices, available in multiple languages and variants

Google Associate Cloud Engineer CheatSheet

Cloud AutoML is a suite of machine learning products that enables developers with limited machine learning expertise to train high-quality models specific to their business needs.

Operations Suite

A suite of tools for logging, monitoring, and application diagnostics

- Available for GCP and AWS
- VM monitoring with agents
- Available for on-premises environments
- Google Cloud native integration
- Monitoring, Logging, Error Reporting, Debugger, Trace, Profiler

Cloud Monitoring

Collects measurements, or metrics, to help you understand how your applications and system services are performing

- Collects metrics to provide insights
- Dashboards and charts
- Workspaces are needed to use cloud monitoring
- Agents are recommended to monitor VMs
- Works together with cloud logging
- Support to monitor GKE and Alerting

Cloud Logging - Central repository for log data from multiple sources

- Logs Viewer only shows logs from one project
- Log Entry records a status or an event
- Logs are a named collection of log entries within a GCP resource
- Retention period how long your logs are kept

Types of Logs

- Audit Logs who did what, where, and when
- Access Transparency Logs actions taken by Google staff
- Agent Logs

- Real-time log management and analysis
- Tight integration with monitoring
- Platform, system and application logs
- Export logs to other sources

Error Reporting - Real time error monitoring and alerting

- Counts, analyzes, and aggregates all the errors in your GCP environment
- Alerts you when a new application error occurs
- Integrated into Cloud Functions and GAE Standard
- Issue tracking integration
- In beta for GCE, GKE, GAE Flexible, AWS EC2
- Go, Java, Node.js, .Net, PHP, Python, Ruby

Debugger - Inspect the state of a running application in real time, without stopping or slowing it down

- Debug a running application with no latency
- “Snapshot” the call stack in your application
- Logpoints allow you to inject logging into running services
- Can be hooked into remote Git repo - Github, GitLab, Bitbucket
- Can be installed on non-GCP environments
- Java, Go, Node.js, Python, .Net, PHP, Ruby

Trace - Collects latency data from App Engine, HTTPS load balancers and applications

- Helps to understand how long it takes your application to handle incoming requests (latency)
- Collects latency data from cloud resources and apps
- Integrated with GAE Standard
- Can be installed on GCE, GKE, and GAE
- Can be installed on non-GCP environments
- C#, Go, Java, Node.js, PHP, Python, Ruby

Google Associate Cloud Engineer CheatSheet

Profiler - Continuously gathers CPU usage and memory allocation information from your applications

- Helps discover patterns of resource consumption
- Low-profile
- Needs profiling agent to be installed
- Can be installed on GCE, GKE, GAE
- Can be installed on non-GCP environments
- Go, Java, Node.js, Python