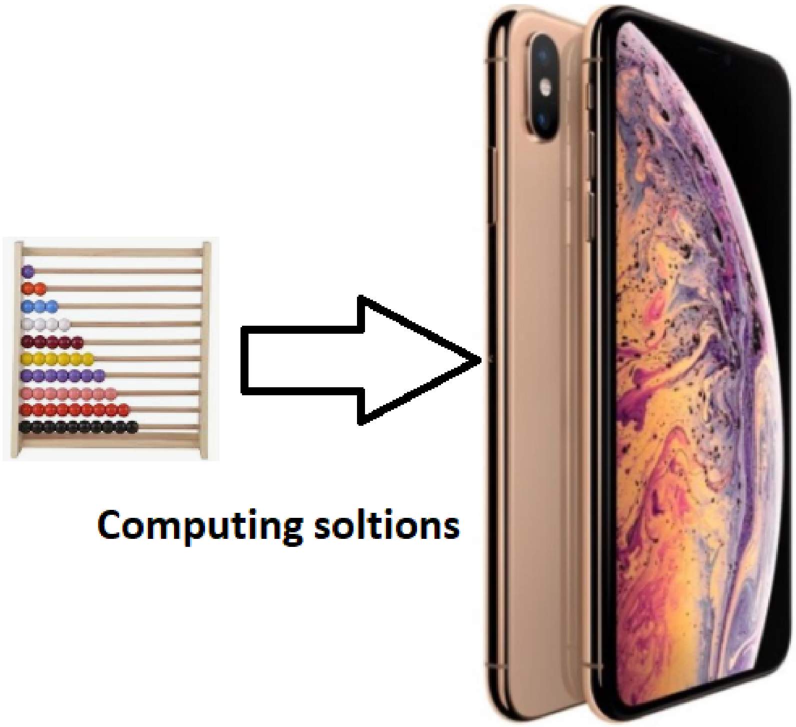
CP Certification Series: Section 3: Deploying and implementing a cloud solution: 3.1 Deploying and implementing Compute Engine resources

Prashanta Paudel

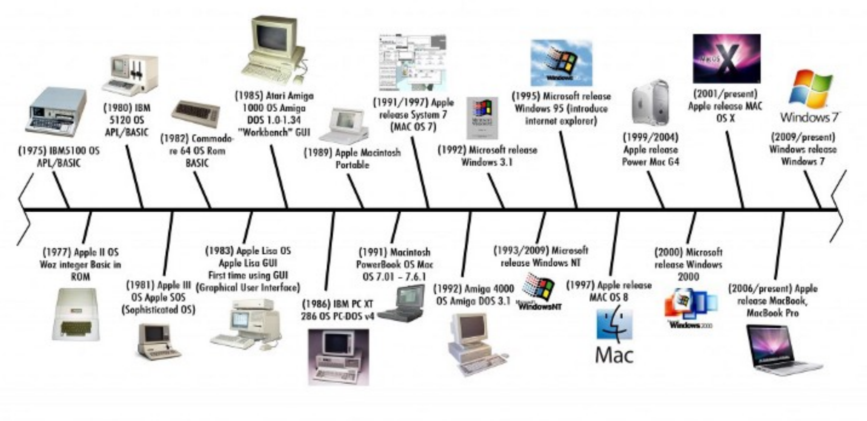
Oct 25, 2018 · 36 min read

Computers were developed primarily for serious calculation problems like population census which used to take 6–7 years to publish due to manual calculations. From primitive computing devices to today's high-end computers its shape, size, and the cost has been drastically changed and now we can have computers that fit in our palm. From abacus to iPhone XS MAX computing technology has a very important role in the transformation of how we live, work and entertain.



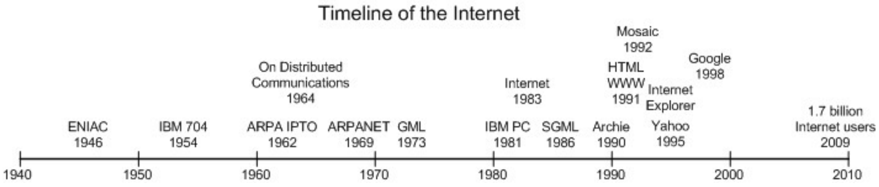
computing revolution

If we take a holistic view of general purpose computers in terms of how we do computing it can be represented in terms of where the actual computing happened.


(1975) IBM5100 OS APL/BASIC
(1980) IBM 5120 OS APL/BASIC
(1982) Commodore 64 OS Rom BASIC
(1985) Atari Amiga 1000 OS Amiga DOS 1.0-1.34 "Workbench" GUI
(1989) Apple Macintosh Portable
(1991/1997) Apple release System 7 (MAC OS 7)
(1992) Microsoft release Windows 3.1
(1995) Microsoft release Windows 95 (introduce internet explorer)
(1999/2004) Apple release Power Mac G4
(2001/present) Apple release MAC OS X
Windows 7
(2009/present) Windows release Windows 7
(1977) Apple II OS Woz integer Basic in ROM
(1981) Apple III OS Apple SOS (Sophisticated OS)
(1983) Apple Lisa OS Apple Lisa GUI First time using GUI (Graphical User Interface)
(1986) IBM PC XT 286 OS PC-DOS v4
(1991) Macintosh PowerBook OS Mac OS 7.01 - 7.6.1
(1992) Amiga 4000 OS Amiga DOS 3.1
(1993/2009) Microsoft release Windows NT
WindowsNT
(1997) Apple release MAC OS 8
Mac
(2000) Microsoft release Windows 2000
Windows 2000
(2006/present) Apple release MacBook, MacBook Pro

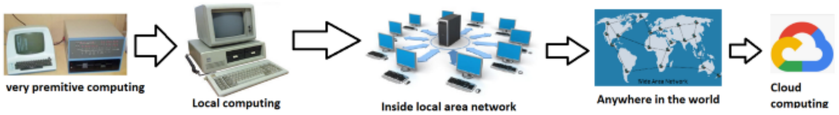
References: http://christinapierre.weebly.com/uploads/5/6/0/9/56091077/4308810_orig.jpg

In early Personal computers, almost all computing and tasks are done in PC itself and the network was merely a concept stage.



timeline

As the networking technology developed from co-axial, ethernet, fiber to wireless the computing place also changed from the same PC to LAN and then to internet and now to the cloud.



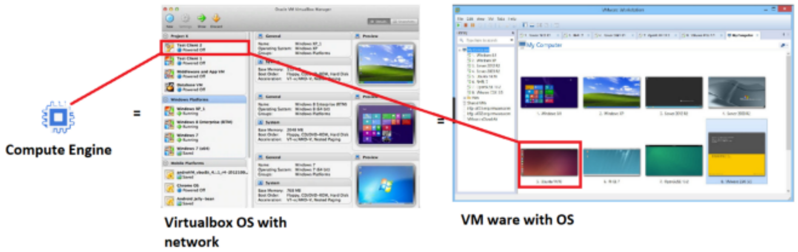
trends in computing

In cloud computing one of the most important aspects is that you can spin up any server in a matter of minutes which was not possible earlier. You can set up your whole lab, computers for all your employees and servers just with few clicks and with whatever configuration you require. All of these are done in the compute engine.

whenever we talk about having computing power like PC or server in the cloud it comes under the compute engine.

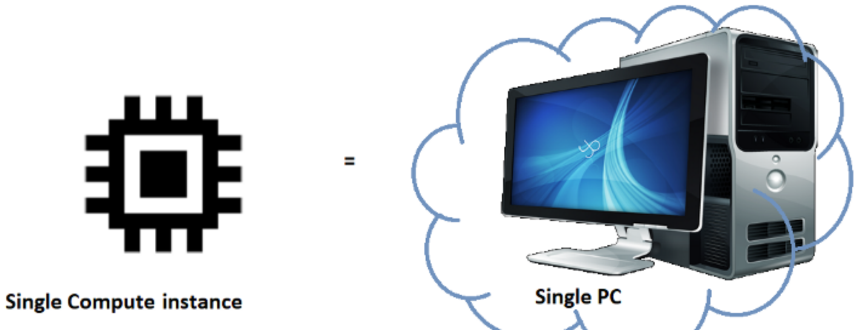
So, what is a compute engine?

As illustrated above a compute engine is the device which can perform computing for us. It can be considered as equivalent to a single computer or server with network and hard disk.



compute engine compared to other solutions

A single compute engine with persistent disk and network configured with firewalls is equivalent to having a personal computer with a network and firewall as well as having a virtual OS with disk space and network configured in it.



single compute instance

The only difference is you have to access it remotely and you own only the computing power.

Ok, now let's go to practical part.

In google cloud platform compute engine consists of the following parts.

1. **Virtual Machine(VM) Instances**

- Virtual machines
- Machine Types
- Images
- Preemptible VM Instances
- Shielded VM- Beta

2. **Storage Options**

- Compute Engine Storage Options

3. **Networking and Firewalls**

- Compute Engine Networks and Firewalls
- IP Addresses

4**. Load Balancing and Scaling**

- Compute Engine Load Balancing and Scaling

5. **Region and Zones**

- viewing available region and zones
- Global, Regional and Zonal resources

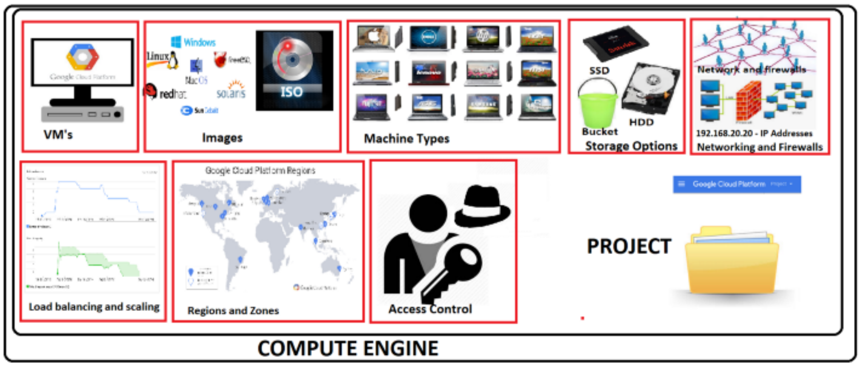
6. **Google Cloud Platform Console projects**

- Google Compute Engine console project

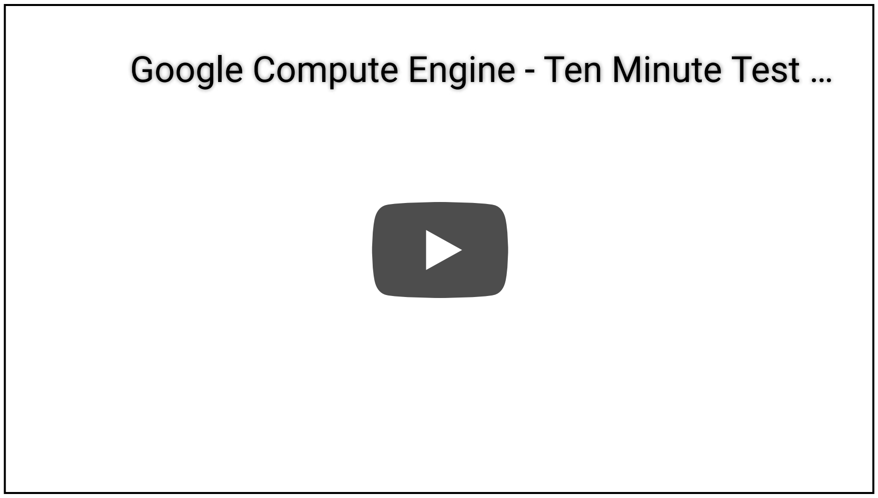
7. **Access Control**

- Google compute engine access control options
- Project team members

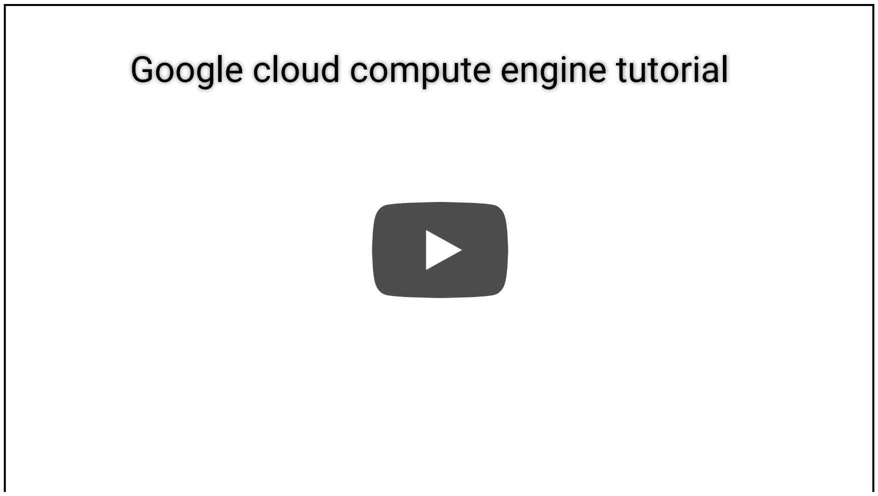
- Identity and access management roles(IAM)- Beta
- Service Accounts



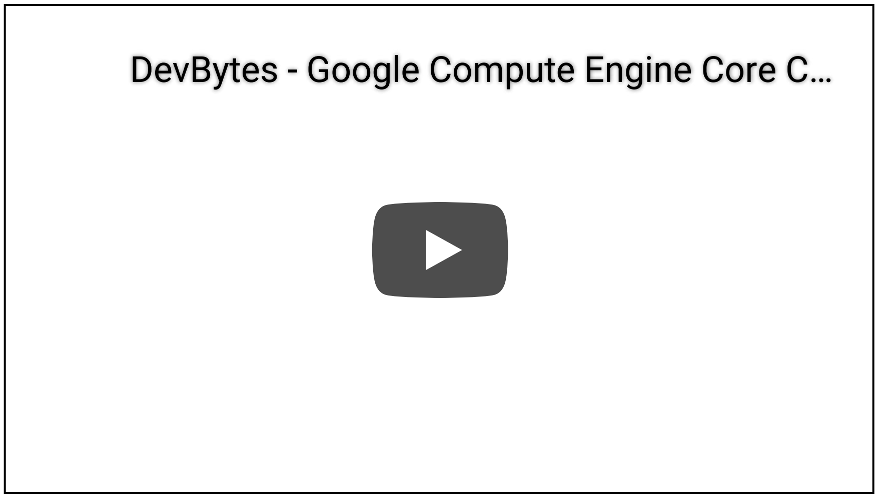
Compute Engine Component



Google Compute Engine - Ten Minute Test ...



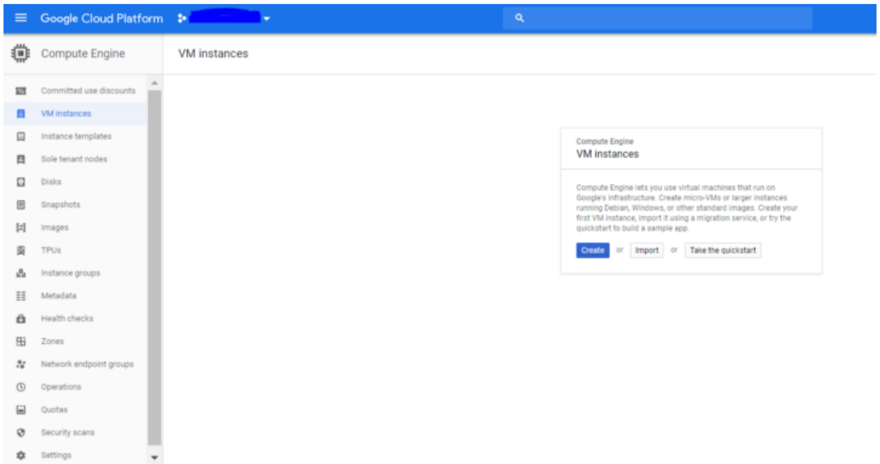
Google cloud compute engine tutorial



DevBytes - Google Compute Engine Core C…

MicroNuggets: Google Compute Engine Ex…

The screenshot of the compute Engine in the console



Compute Engine lets you use virtual machines that run on Google's infrastructure. Create micro-VMs or larger instances running Debian, Windows, or other standard images. Create your first VM instance, import it using a migration service, or try the quickstart to build a sample app.

Compute Engine Dashboard

Virtual Machine Instances

An *instance* is a virtual machine (VM) hosted on Google's infrastructure. You can create an instance by using the Google Cloud Platform Console or the command-line tool. Compute engine instances can run the public images for Linux and windows that Google explicitly provide on their platform as well as you can create or import from your existing systems. You can also deploy Docker containers, which are automatically launched on instances running the Container-Optimized OS public image. During creation, you can select no of virtual CPUs and memory by using a set of predefined machine types or by creating your own custom machine type.

Virtual Machines cannot exist alone but is always under some project. VM requires mentioning OS, zone, region and machine type before it can be deployed.

By default, VM has a small disk for booting an OS but you can add more depending on requirement.

Each VM should belong to some VPC network. VM's on the same network can communicate via LAN whereas in different VPC via the internet.

You can access VM via RDP or SSH or in a browser itself which is in the console.

Machine Types

A **machine type** specifies a particular collection of virtualized hardware resources available to a virtual machine (VM) instance, including the system memory size, virtual CPU (vCPU) count, and maximum persistent disk capability.

There are basically 2 types of Machine

Predefined machine Types

- Standard
- High memory
- High CPU

- Shared core
- Memory-optimized

Custom machine types

Images

There are basically two types of images

1. Public Images: provided by Google, open-source, default access to all users.
2. Custom Images: available only to your project, upload from your computer.

Preemptible VM

Those instances that can be terminated anytime by compute engine are called preemptive VM instances. These VM's can be implemented at a lower price than normal instances since compute engine has the right to kill it anytime.

These kinds of instances are best for fault tolerant application such as batch processing that doesn't affect the system completely even if it is down or may continue later.

Preemptible instances function like normal instances, but have the following limitations:

- Compute Engine might terminate preemptible instances at any time due to system events. The probability that Compute Engine will terminate a preemptible instance for a system event is generally low but might vary from day to day and from zone to zone depending on current conditions.
- Compute Engine always terminates preemptible instances after they run for 24 hours.
- Preemptible instances are finite Compute Engine resources, so they might not always be available.
- Preemptible instances cannot live to migrate to a regular VM instance or be set to automatically restart when there is a maintenance event.

- Due to the above limitations, preemptible instances are not covered by any Service Level Agreement (and, for clarity, are excluded from the Google Compute Engine SLA).

Instance Template

An instance template is an API 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, zone, labels, and other instance properties. You can then use an instance template to create a managed instance group or to create individual VM instances. Instance templates are a convenient way to save a VM instance's configuration so you can use it later to create new VM instances or groups of VM instances.

Instance templates are designed to create instances with identical configurations. So it is not possible to update an existing instance template or change an instance template after it has been created. If an instance template goes out of date, or you need to make changes to the configuration, create a new instance template. You can also override instance template fields when creating a VM instance from an instance template.

Instance Group

You can create and manage groups of virtual machine (VM) instances so that you don't have to individually control each instance in your project. Compute Engine offers two different types of instance groups: **managed** and **unmanaged** instance groups.

Managed instance groups

A managed instance group uses an instance template to create a group of identical instances. You control a managed instance group as a single entity. If you wanted to make changes to instances that are part of a managed instance group, you would make the change to the whole instance group

Unmanaged instance groups

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

Shielded VM

Shielded VM offers verifiable integrity of your Compute Engine VM instances, so you can be confident your instances haven't been compromised by boot- or kernel-level malware or rootkits. Shielded VM's verifiable integrity is achieved through the use of Secure Boot, virtual trusted platform module (vTPM)-enabled Measured Boot, and integrity monitoring.

Shielded VM is the first offering in the Shielded Cloud initiative. The Shielded Cloud initiative is meant to provide an even more secure foundation for all of Google Cloud Platform (GCP) by providing verifiable integrity and offering features, like vTPM shielding or sealing, that help prevent data exfiltration.

Storage Options

Compute engine instance or VM's can use following types of storage solutions

1. Zonal standard persistent disk and Zonal SSD persistent disk
2. Regional persistent disk and regional SSD persistent disk
3. Local SSD
4. Cloud storage buckets

Zonal disks can be used only within Zone

Regional disks can be used within the region

Local SSD is temporary fast storage for VM's which will be deleted when VM is deleted.

Cloud storage buckets can be used in VM's.

If you are not sure which fits you most use a persistent disk as it is most commonly used storage device.

In addition to the storage options that Google Cloud Platform provides, you can deploy alternative storage solutions on your instances.

- Create a file server or distributed file system on Compute Engine to use as a network file system with NFSv3 and SMB3 capabilities.
- Mount a RAM disk within instance memory to create a block storage volume with high throughput and low latency.

Networking and Firewall

A VPC network or just “network” is a virtual version of the physical network such as your in-house servers and workstations network. It provides connectivity to your VM, Kubernetes engine cluster, App engine, storage devices etc.

VPC networks have the following properties:

- VPC networks, including their associated routes and firewall rules, are global resources. They are *not* associated with any particular region or zone.
- *Subnets* are regional resources. Each subnet defines a range of IP addresses. For more information about networks and subnets, see networks and subnets.
- Traffic to and from instances can be controlled with network firewall rules.
- Resources within a VPC network can communicate with one another using internal (private) IPv4 addresses, subject to applicable network firewall rules. For more information, see communication within the network.
- Instances with internal IP addresses can communicate with Google APIs and services. For more information, see Private Access Options.
- Network administration can be secured using Identity and Access Management (IAM) roles.

- An organization can use Shared VPC to keep a VPC network in a common host project. Authorized IAM members from other projects in the same organization can create resources that use subnets of the Shared VPC network.
- VPC networks can be connected to other VPC networks in different projects or organizations by using VPC Network Peering.
- VPC networks can be securely connected in hybrid environments using Cloud VPN or Cloud Interconnect.
- VPC networks only support IPv4 unicast traffic. They do **not** support broadcast, multicast, or IPv6 traffic *within* the network. However, IPv6 can be used to *reach* resources in the network. For example, IPv6 addresses can be assigned to a global load balancer, and the App Engine standard environment supports IPv6.

Types of VPC networks

There are two types of VPC networks:

- When an **auto mode** network is created, one subnet from each region is automatically created within it. These automatically created subnets use a set of predefined IP ranges which fit within the `10.128.0.0/9` CIDR block. As new GCP regions become available, new subnets in those regions are automatically added to auto mode networks using an IP range from that block. In addition to the automatically created subnets, you can add more subnets manually to auto mode networks, in regions you choose, using IP ranges outside of `10.128.0.0/9` .
- When a **custom mode** network is created, no subnets are automatically created. This type of network provides you with complete control over its subnets and IP ranges. You decide which subnets to create, in regions you choose, and using IP ranges you specify.

Each project starts with a `default` auto mode network.

You can switch a network from auto mode to custom mode. This conversion is one-way; custom mode networks cannot be changed to auto mode networks. Carefully review the considerations for auto mode networks to help you decide which type of network meets your needs.

Load balancing and scaling

Google Cloud Platform (GCP) offers load balancing and autoscaling for groups of instances. If you have created many instances of the similar type and would like to split the traffic between them this could be achieved using GCP.

Load Balancing

GCP offers server-side load balancing so you can distribute incoming traffic across multiple virtual machine instances. Load balancing provides the following benefits:

- ***Scale your application***
- ***Support heavy traffic***
- ***Detect and automatically remove unhealthy virtual machine instances using health checks. Instances that become healthy again are automatically re-added.***
- ***Route traffic to the closest virtual machine***

Horizontal scaling means that you scale by adding more machines to your pool of resources whereas **Vertical scaling means** that you scale by adding more power (CPU, RAM) to an existing machine.

GCP load balancing uses forwarding rule resources to match certain types of traffic and forward it to a load balancer. For example, a forwarding rule can match TCP traffic destined to port 80 on IP address `192.0.2.1` , then forward it to a load balancer, which then directs it to healthy virtual machine instances.

GCP load balancing is a managed service, which means its components are redundant and highly available. If a load balancing component fails, it is restarted or replaced automatically and immediately.

GCP offers several different types of load balancing that differ in capabilities, usage scenarios, and how you configure them. See Load balancing for descriptions.

Autoscaling

Compute Engine offers autoscaling to automatically add or remove virtual machines from an instance group based on increases or decreases in load. This allows your applications to gracefully handle increases in traffic and reduces cost when the need for resources is lower. You just define the autoscaling policy and the autoscaler performs automatic scaling based on the measured load.

Policies

Choose from a variety of policies that an autoscaler can use to scale your virtual machines. When you create an autoscaler, you must specify at least one policy. If you use multiple policies, the autoscaler will scale an instance group based on the policy that provides the largest number of virtual machines in the group.

The following sections discuss the autoscaling policies in general; for more information about how to set up a specific autoscaling policy, see the respective policy documentation.

CPU utilization

CPU utilization is the most basic autoscaling that you can perform. This policy tells the autoscaler to watch the average CPU utilization of a group of virtual machines and add or remove virtual machines from the group to maintain your desired utilization. This is useful for configurations that are CPU-intensive but might fluctuate in CPU usage.

For more information, see Scaling Based on CPU utilization.

Load balancing serving capacity

Set up an autoscaler to scale based on load balancing serving capacity and the autoscaler will watch the serving capacity of an instance group, and scale if the virtual machines are over or under capacity.

The serving capacity of an instance can be defined in the load balancer’s backend service and can be based on either utilization or requests per second.

For more information, see Scaling Based on HTTP(S) load balancing.

Stackdriver Monitoring metrics

If you export or use Stackdriver Monitoring metrics, you can set up autoscaling to collect data of a specific metric and perform scaling based on your desired utilization level. It is possible to scale based on standard metrics provided by Stackdriver Monitoring, or using any custom metrics you create as well.

For more information, see Scaling Based on Stackdriver Monitoring Metrics.

Regions and Zones

When developing your application in GCP it is very important to understand regions and zones.

Resources are also regional and zonal so you must also have an idea about which resource is what before going in detail.

A region is a geographical location that is sub-divided into zones.

While few of the resources in GCP are global, others may be restricted by region or zone.

Regional resources can be used anywhere within the same region, while zonal resources can be used anywhere within the same zone. Some examples of this are:

Global Resources:

- Images
- Snapshots
- VPC Network
- Firewalls
- Routes

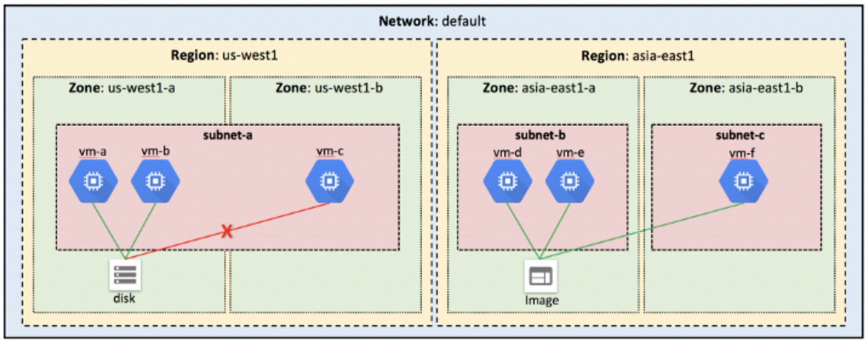
Regional Resources:

- Static external IP addresses
- Subnets

Zonal Resources:

- Instances (VMs)
- Persistent Disks

For example, I can attach a disk from one instance to another within the same zone, but I cannot do this across zones. However, since images and snapshots are Global Resources, I can use these across zones in the same region.



reference: https://www.networkmanagementsoftware.com/google-cloud-platform-gcp-networking-fundamentals/

Choosing a region and zone

You choose which region or zone hosts your resources, which controls where your data is stored and used. Choosing a region and zone is important for several reasons:

Handling failures

Distribute your resources across multiple zones and regions to tolerate outages. Google designs zones to be independent of each other: a zone usually has power, cooling, networking, and control planes that are isolated from other zones, and most single failure events will affect only a single zone. Thus, if a zone becomes unavailable, you can transfer traffic to another zone in the same region to keep your services running. Similarly, if a region experiences any disturbances, you should have backup services running in a different region. For more information about distributing your resources and designing a robust system, see Designing Robust Systems.

Decreased network latency

To decrease network latency, you might want to choose a region or zone that is close to your point of service. For example, if you mostly have customers on the East Coast of the US, then you might want to choose a primary region and zone that is close to that area and a backup region and zone that is also close by.

Identifying a region or zone

Each region in Compute Engine contains a number of zones. Each zone name contains two parts that describe each zone in detail. The first part of the zone name is the **region** and the second part of the name describes the **zone** in the region:

Region

- Regions are collections of zones. Zones have high-bandwidth, low-latency network connections to other zones in the same region. In order to deploy fault-tolerant applications that have high availability, Google recommends deploying applications across multiple zones and multiple regions. This helps protect against unexpected failures of components, up to and including a single zone or region.
- Choose regions that make sense for your scenario. For example, if you only have customers in the US, or if you have specific needs that require your data to live in the US, it makes sense to store your resources in zones in the us-central1 region or zones in the us-east1 region.

Zone

- A zone is an isolated location within a region. The fully-qualified name for a zone is made up of `<region>-<zone>`. For example, the fully-qualified name for the zone `a` in region `us-central1` is `us-central1-a`.

- Depending on how widely you want to distribute your resources, create instances across multiple zones in multiple regions for redundancy.

Google Cloud Platform Console Project

Google Cloud Platform (GCP) projects form the basis for creating, enabling, and using all GCP services including managing APIs, enabling billing, adding and removing collaborators, and managing permissions for GCP resources.

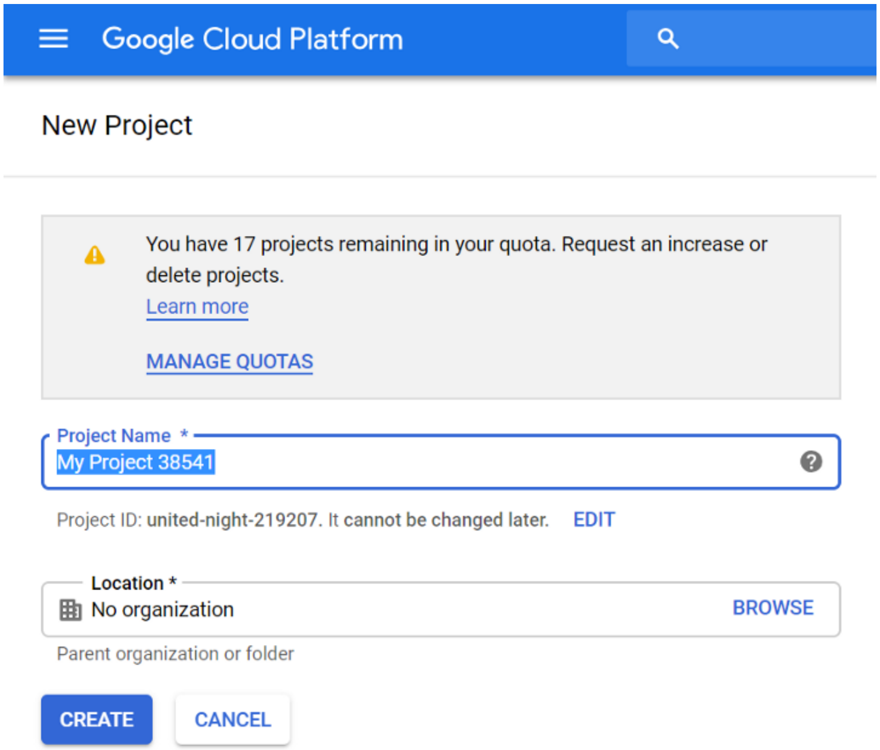
Creating the project

The most basic thing to do in GCP is creating the project. Creating the project in itself doesn’t complete anything but it will start the process so that you can add entities in the project and build your own network, create a database, write code, build server etc.

The project number and project ID are unique across the Google Cloud Platform. If another user owns a project ID for their project, you won’t be able to use the same project ID. Also, the name such as Google or SSL cannot be used for the project name.

All the instances are attached to the project. So, you got the idea everything is inside the project.

To create the project, first, you need to have access to the cloud console.



You have 17 projects remaining in your quota. Request an increase or delete projects.
Learn more
MANAGE QUOTAS

Project Name *
My Project 38541

Project ID: united-night-219207. It cannot be changed later. EDIT

Location *
No organization BROWSE
Parent organization or folder

create project

If you are running a free account you should select "individual" while creating the account itself. Free users get limited quota so if the number of projects that can be created is less than 30 you will get the message as shown above.

The project name is a human-readable name for simplicity while all the processing will be done using the project ID mentioned just below the project name.

You may be working in the console for more than one project so, it is always a good idea to check the project name while performing tasks.

To create a new project, use the `gcloud projects create` command:

```
gcloud projects create PROJECT_ID
```

Eg

```
prashantagcppaudel@cloudshell:~ (webproject-217416)$ gcloud
projects create testdemotrial123
Create in progress for
```

```
[https://cloudresourcemanager.googleapis.com/v1/projects/tes
tdemotrial123].
Waiting for [operations/cp.6134727994789518289] to
finish...done.
```

Where PROJECT_ID is the ID for the project you want to create. A project ID must start with a lowercase letter, and can contain only ASCII letters, digits, and hyphens, and must be between 6 and 30 characters.

Some of the common operations in the project are:

1. Get an existing project
2. creating a project
3. Managing project quotas
4. listing project
5. Updating project
6. Shutting down projects
7. Restoring project

Access Control

By default, all Google Cloud Platform projects come with a single user: the original project creator. No other users have access to the project, and therefore, access to Compute Engine resources, until a user is added as a project member or is bound to a specific resource. This page describes the different ways you can add new users to your project and how to set access control for your Compute Engine resources.

In addition, if you run applications on a virtual machine (VM) instance, and the application needs access to Compute Engine or other Google Cloud Platform APIs, you can use service accounts to authenticate your applications instead of using user credentials.

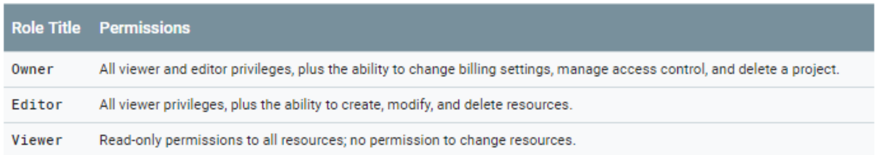
With IAM, every API method in Compute Engine requires that the identity making the API request has the appropriate permissions to use the resource. Permissions are granted by setting policies that grant roles to a user, group, or service account as a **member** of your project.

In addition to the legacy roles, owner, editor, and viewer, you can assign the following Compute Engine roles to the members of your project.

You can grant multiple roles to a project member on the same resource. For example, if your networking team manages firewall rules instead of leaving those to a separate security team, you can grant both `roles/compute.networkAdmin` and `roles/compute.securityAdmin` to the networking team's Google group.

Primitive Cloud IAM roles

Primitive Cloud IAM roles map directly to the legacy project owner, editor, and viewer roles. Generally, you should use predefined roles whenever possible; however, in some cases, where Cloud IAM is not yet supported, you might need to use a primitive role to grant the correct permissions.



Role Title	Permissions
Owner	All viewer and editor privileges, plus the ability to change billing settings, manage access control, and delete a project.
Editor	All viewer privileges, plus the ability to create, modify, and delete resources.
Viewer	Read-only permissions to all resources; no permission to change resources.

References: https://cloud.google.com/compute/docs/access/

To learn more about primitive roles, read the documentation for Primitive Roles.

If predefined or primitive roles do not meet your needs, you can create custom roles.

Service Accounts

This page describes service accounts, access scopes, and Identity and Access Management (IAM) roles that apply to service accounts. To learn how to create and use service accounts, read the Creating and Enabling Service Accounts for Instances documentation.

A service account is a special account that can be used by services and applications running on your Google Compute Engine instance to interact with other Google Cloud Platform APIs. Applications can use service account credentials to authorize themselves to a set of APIs and perform actions within the permissions granted to the service account and virtual machine instance. In addition, you can create firewall rules that allow or deny traffic to and from instances based on the service account that owns the instances.

What are service accounts?

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

You can use service accounts to create instances and other resources. If you create a resource using a service account, that resource is then owned by the creating service account. You can also change the service account of an existing instance.

An instance can have only one service account, and the service account must have been created in the same project as the instance.

Two types of service accounts are available to Compute Engine instances:

- User-managed service accounts
- Google-managed service accounts

——
— —

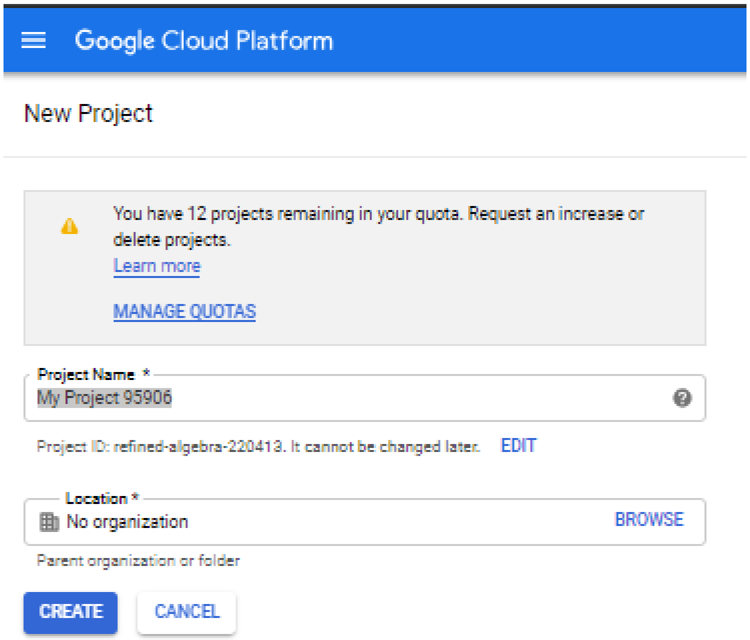
NOW let's DO POINTS IN COURSE

Launching a compute instance using Cloud Console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys).

There are many compute instance. For creating any compute instance first you need to have a project.

> *Creating a Project*

From cloud console: Goto Google cloud console (needs a google account and billing activated prior) and select New project



New Project

You have 12 projects remaining in your quota. Request an increase or delete projects.

Project Name * My Project 95906

Project ID: refined-algebra-220413. It cannot be changed later.

Location * No organization

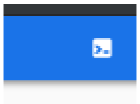
Parent organization or folder

New project

After clicking CREATE, GCP will create a new project with a single user and default policy and permissions.

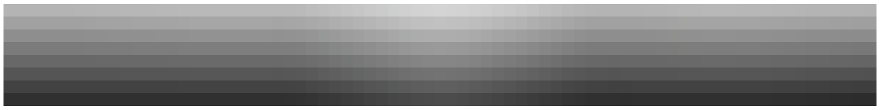
From cloud shell :

Goto Cloud Console Dashboard and click on cloud shell

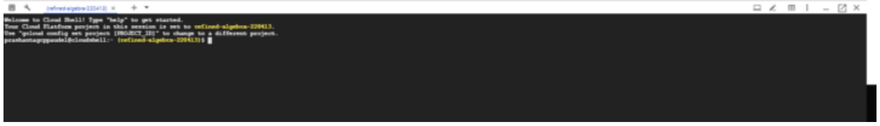


cloud Shell

It will take a while to load the cloud shell



When the SDK is loaded in temporary VM you will see the CLI interface as shown below



```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to refined-algebra-220413.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
prashantagcppaudel@cloudshell:~ (refined-algebra-220413)$
```

Cloud SDK in Cloud Shell

Now type the following command

```
prashantagcppaudel@cloudshell:~ (refined-algebra-220413)$
gcloud projects create mysmallprojecttest
Create in progress for
[https://cloudresourcemanager.googleapis.com/v1/projects/mys
mallprojecttest].
Waiting for [operations/cp.6950592308482607668] to
finish...done.
prashantagcppaudel@cloudshell:~ (refined-algebra-220413)$
```

From Cloud SDK in your computer

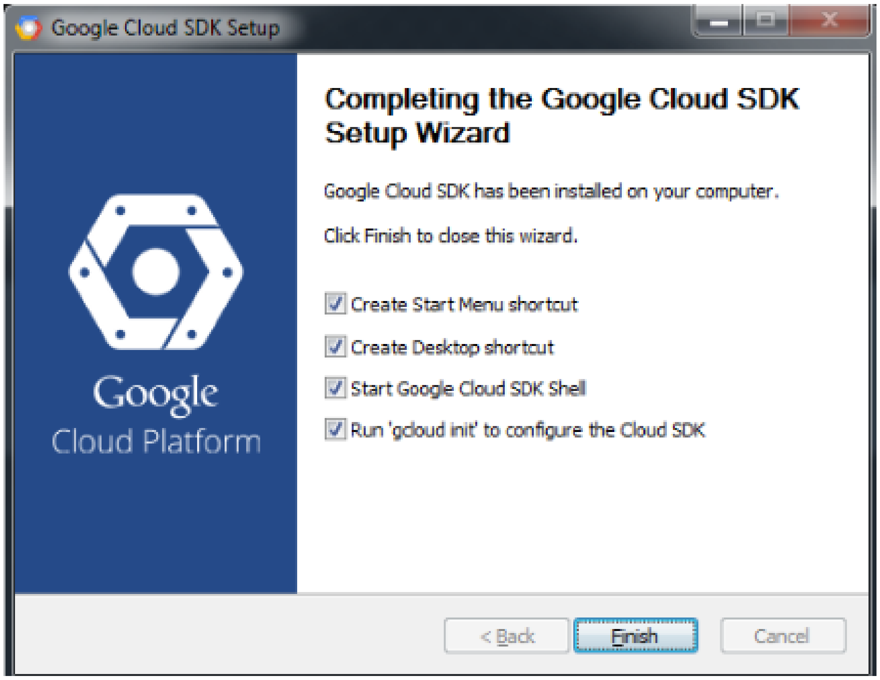
First Download SDK from Google

Cloud SDK | Cloud SDK | Google Cloud

A collection of command line tools for the Google Cloud Platform. Includes gcloud, bq, gsutil and...

cloud.google.com

Download SDK’s exe file in your desktop and install it.



Completing the Google Cloud SDK Setup Wizard

Google Cloud SDK has been installed on your computer.

Click Finish to close this wizard.

☑ Create Start Menu shortcut

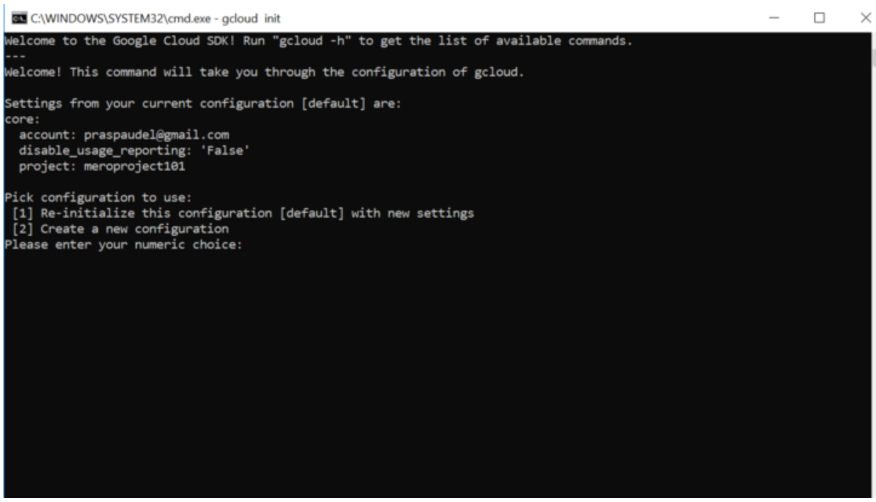
☑ Create Desktop shortcut

☑ Start Google Cloud SDK Shell

☑ Run 'gcloud init' to configure the Cloud SDK

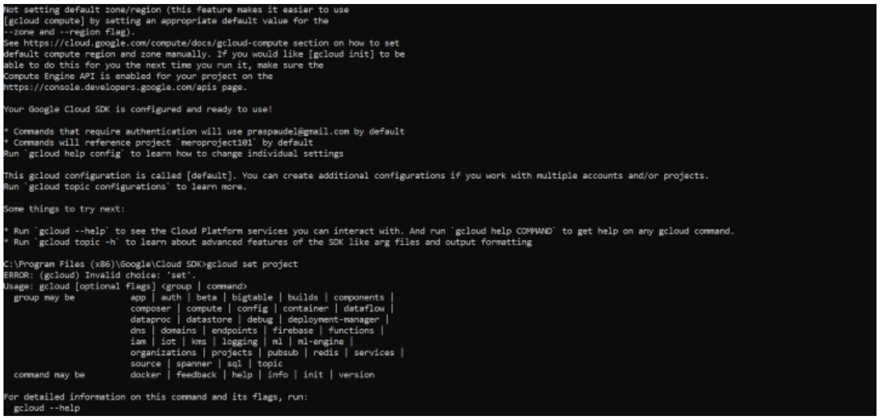
SDK

After finishing install you will get a prompt like this



```
Welcome to the Google Cloud SDK! Run "gcloud -h" to get the list of available commands.
---
Welcome! This command will take you through the configuration of gcloud.

Settings from your current configuration [default] are:
core:
  account: praspaudel@gmail.com
  disable_usage_reporting: 'False'
  project: meroproject101

Pick configuration to use:
 [1] Re-initialize this configuration [default] with new settings
 [2] Create a new configuration
Please enter your numeric choice:
```

Now follow the messages to use the credentials and once you finish you can connect and work in GCP as from Cloud shell.

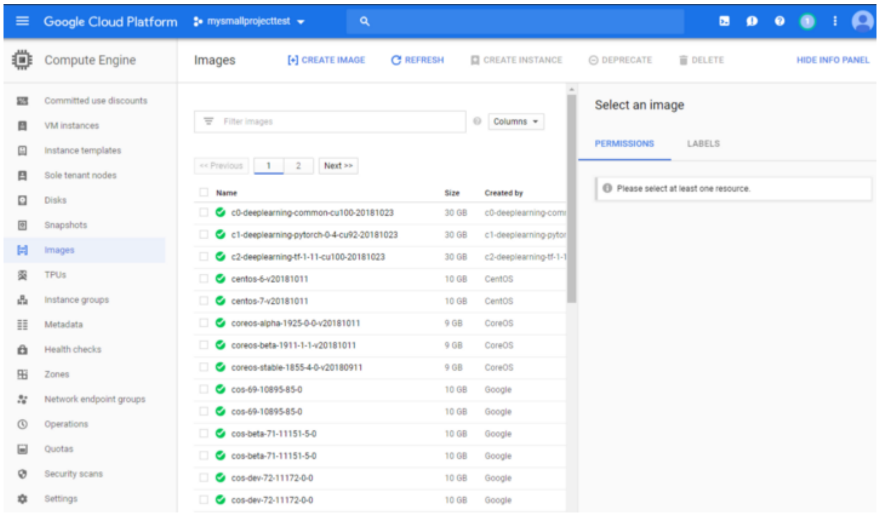


```
Not setting default zone/region (this feature makes it easier to use
[gcloud compute] by setting an appropriate default value for the
--zone and --region flag).
See https://cloud.google.com/compute/docs/gcloud-compute section on how to set
default compute region and zone manually. If you would like [gcloud init] to be
able to do this for you the next time you run it, make sure the
Compute Engine API is enabled for your project on the
https://console.developers.google.com/apis page.

Your Google Cloud SDK is configured and ready to use!

* Commands that require authentication will use praspaudel@gmail.com by default
* Commands will reference project `meroproject101` by default
Run `gcloud help config` to learn how to change individual settings

This gcloud configuration is called [default]. You can create additional configurations if you work with multiple accounts and/or projects.
Run `gcloud topic configurations` to learn more.

Some things to try next:

* Run `gcloud --help` to see the Cloud Platform services you can interact with. And run `gcloud help COMMAND` to get help on any gcloud command.
* Run `gcloud topic -h` to learn about advanced features of the SDK like arg files and output formatting

C:\Program Files (x86)\Google\Cloud SDK>gcloud set project
ERROR: (gcloud) Invalid choice: 'set'.
Usage: gcloud [optional flags] <group | command>
  group may be           app | auth | beta | bigtable | builds | components |
                         composer | compute | config | container | dataflow |
                         dataproc | datastore | debug | deployment-manager |
                         dns | domains | endpoints | firebase | functions |
                         iam | iot | kms | logging | ml | ml-engine |
                         organizations | projects | pubsub | redis | services |
                         source | spanner | sql | topic
  command may be         docker | feedback | help | info | init | version

For detailed information on this command and its flags, run:
  gcloud --help
```

Creating project from images

Another method of creating a new instance is by selecting from the already available image in GCP.

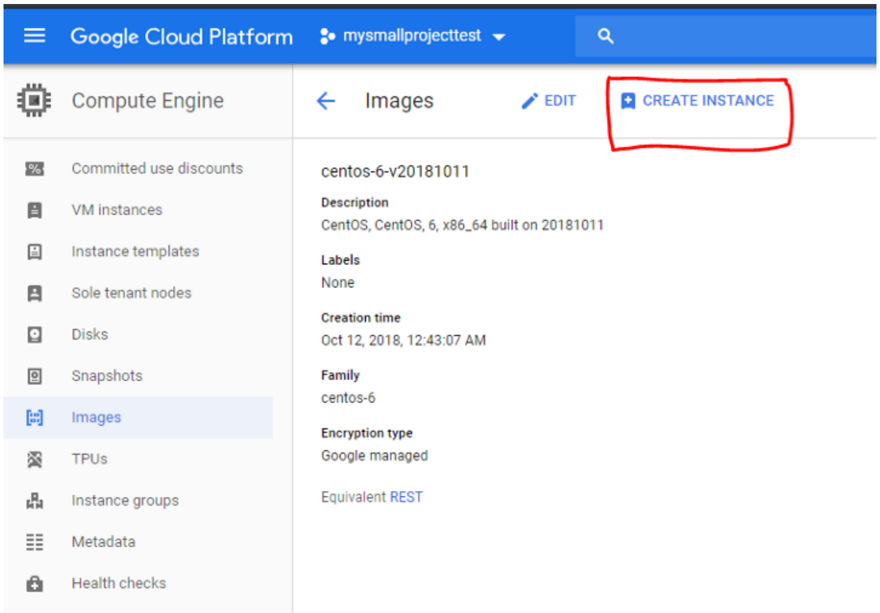
Goto Cloud Console Dashboard and select compute engine>images



Please select at least one resource.

Name	Size	Created by
c0-deeplearning-common-cu100-20181023	30 GB	c0-deeplearning-comi
c1-deeplearning-pytorch-0-4-cu92-20181023	30 GB	c1-deeplearning-pytor
c2-deeplearning-tf-1-11-cu100-20181023	30 GB	c2-deeplearning-tf-1-1
centos-6-v20181011	10 GB	CentOS
centos-7-v20181011	10 GB	CentOS
coreos-alpha-1925-0-0-v20181011	9 GB	CoreOS
coreos-beta-1911-1-1-v20181011	9 GB	CoreOS
coreos-stable-1855-4-0-v20180911	9 GB	CoreOS
cos-69-10895-85-0	10 GB	Google
cos-69-10895-85-0	10 GB	Google
cos-beta-71-11151-5-0	10 GB	Google
cos-beta-71-11151-5-0	10 GB	Google
cos-dev-72-11172-0-0	10 GB	Google
cos-dev-72-11172-0-0	10 GB	Google

Images



centos-6-v20181011

Description
CentOS, CentOS, 6, x86_64 built on 20181011

Labels
None

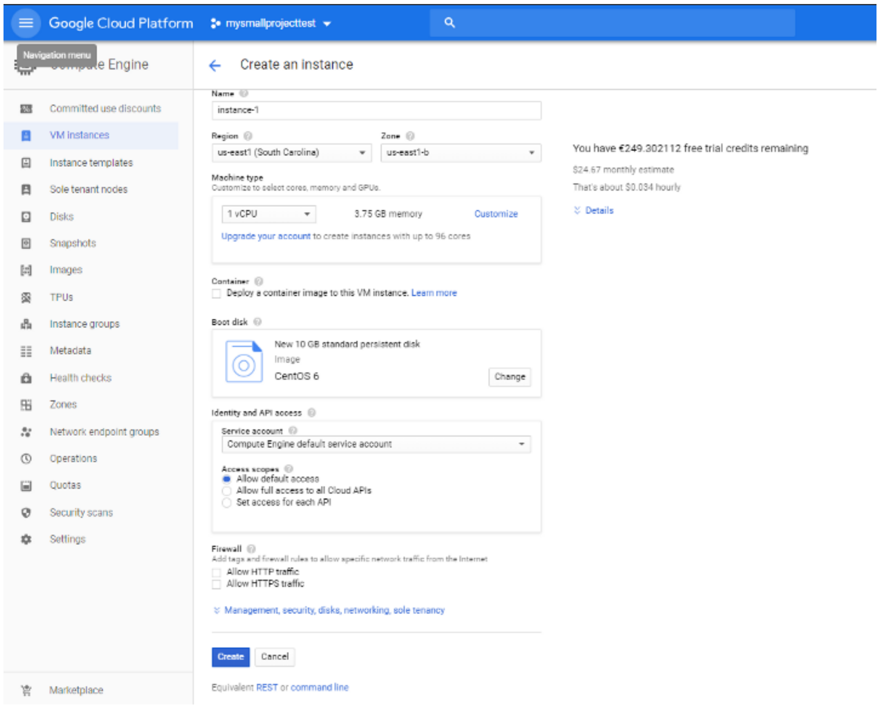
Creation time
Oct 12, 2018, 12:43:07 AM

Family
centos-6

Encryption type
Google managed

Create instance

After clicking create an instance, it will take the configuration and lands into the same page a creating a new VM.



Name
instance-1

Region
us-east1 (South Carolina)

Zone
us-east1-b

Machine type
Customize to select cores, memory and GPUs.
1 vCPU 3.75 GB memory
Upgrade your account to create instances with up to 96 cores

Container
□ Deploy a container image to this VM instance. Learn more

Boot disk
New 10 GB standard persistent disk
Image
CentOS 6

Identity and API access
Service account
Compute Engine default service account
Access scopes
● Allow default access
○ Allow full access to all Cloud APIs
○ Set access for each API

Firewall
Add tags and firewall rules to allow specific network traffic from the Internet
□ Allow HTTP traffic
□ Allow HTTPS traffic

You have €249.302112 free trial credits remaining
$24.67 monthly estimate
That's about $0.034 hourly

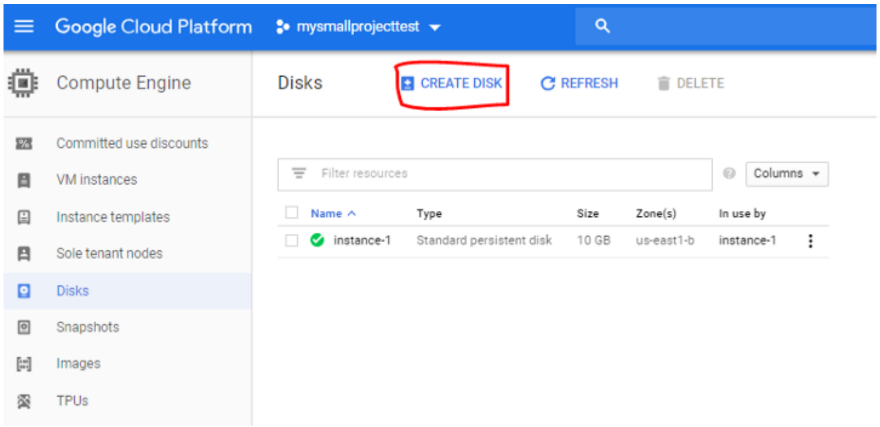
create VM

Now click on “create” to install new VM.

Assigning Disks

> *On Cloud Console*

Goto Compute engine>Disks

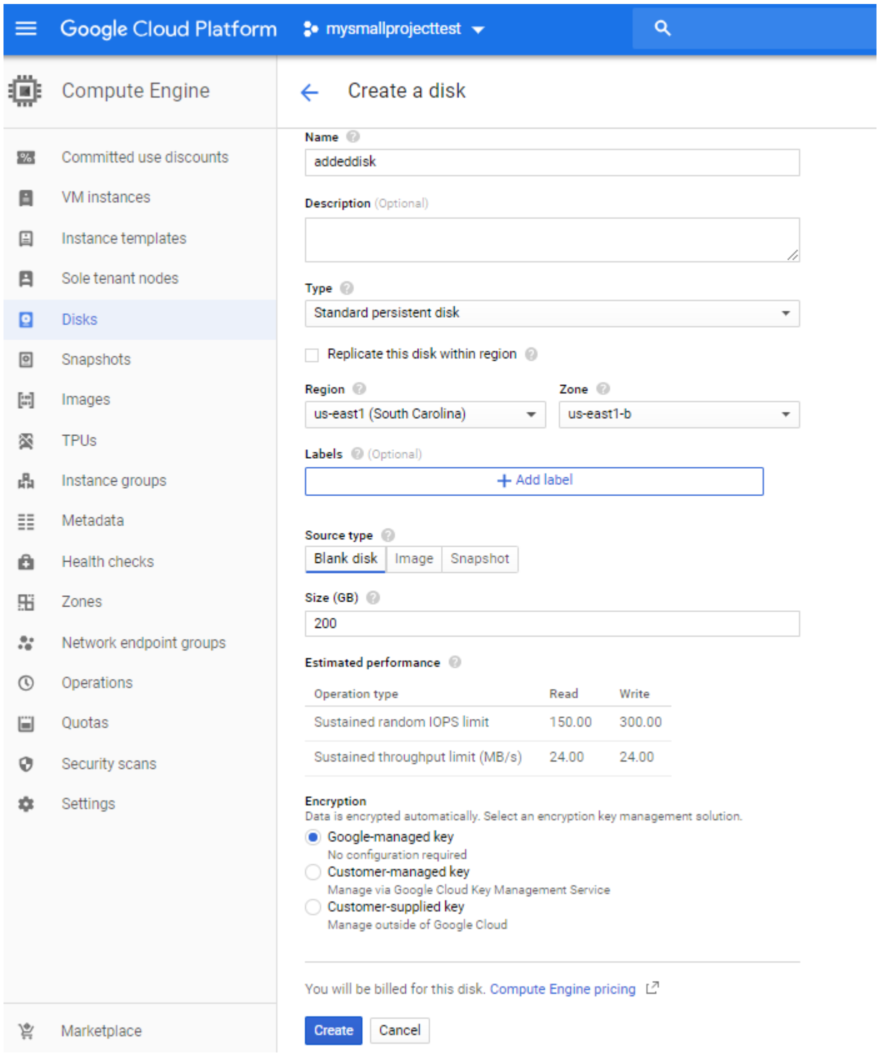


Name	Type	Size	Zone(s)	In use by
instance-1	Standard persistent disk	10 GB	us-east1-b	instance-1

disks

You will see the VM and disk used by VM.

Now, click on Create Disk at the top of the page



Create a disk

Name: addeddisk

Description (Optional)

Type: Standard persistent disk

□ Replicate this disk within region

Region: us-east1 (South Carolina) Zone: us-east1-b

Labels (Optional) + Add label

Source type: Blank disk | Image | Snapshot

Size (GB): 200

Estimated performance

Operation type	Read	Write
Sustained random IOPS limit	150.00	300.00
Sustained throughput limit (MB/s)	24.00	24.00

Encryption
Data is encrypted automatically. Select an encryption key management solution.

● Google-managed key
No configuration required
○ Customer-managed key
Manage via Google Cloud Key Management Service
○ Customer-supplied key
Manage outside of Google Cloud

You will be billed for this disk. Compute Engine pricing

Create Disk

In Name field put any name,

In type select, you want to add Standard persistent or SSD persistent disk

In region select which region you want it to be available or replicate it to other regions.

In the source, select either blank, Image or snapshot

Select Size of disk carefully for performance

lastly, select Encryption and click CREATE

> *On Cloud SDK*

```
prashantagcppaudel@cloudshell:~ (mysmallprojecttest)$ gcloud
compute disks create pappu  --size 10GB --type pd-standard
Did you mean zone [europe-west4-a] for disk: [pappu] (Y/n)?
y

WARNING: You have selected a disk size of under [200GB].
This may result in poor I/O performance. For more
information, see:
https://developers.google.com/compute/docs/disks#performance
.
Created
[https://www.googleapis.com/compute/v1/projects/mysmallproje
cttest/zones/europe-west4-a/disks/pappu].
NAME   ZONE            SIZE_GB  TYPE         STATUS
pappu  europe-west4-a  10       pd-standard  READY

New disks are unformatted. You must format and mount a disk
before it
can be used. You can find instructions on how to do this at:

https://cloud.google.com/compute/docs/disks/add-persistent-
disk#formatting

prashantagcppaudel@cloudshell:~ (mysmallprojecttest)$
```

```
prashantagcppaudel@cloudshell:~ (mysmallprojecttest)$ gcloud compute disks create pappu  --size 10GB --type pd-standard
Did you mean zone [europe-west4-a] for disk: [pappu] (Y/n)?  y

WARNING: You have selected a disk size of under [200GB]. This may result in poor I/O performance. For more information, see: https://developers.google.com/compute/docs/disks#perfor
mance.
Created [https://www.googleapis.com/compute/v1/projects/mysmallprojecttest/zones/europe-west4-a/disks/pappu].
NAME   ZONE            SIZE_GB  TYPE         STATUS
pappu  europe-west4-a  10       pd-standard  READY

New disks are unformatted. You must format and mount a disk before it
can be used. You can find instructions on how to do this at:

https://cloud.google.com/compute/docs/disks/add-persistent-disk#formatting

prashantagcppaudel@cloudshell:~ (mysmallprojecttest)$
```

Setting Instance Availability Policies

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

Choosing availability policies

A VM instance's availability policy determines how it behaves when an event occurs that requires Google to move your VM to a different host machine. For example, you can choose to keep your VM instances running while Compute Engine live migrates them to another host or you can choose to terminate your instances instead. You can update an instance's availability policy at any time to control how you want your VM instances to behave.

You can change an instance's availability policy by configuring the following two settings:

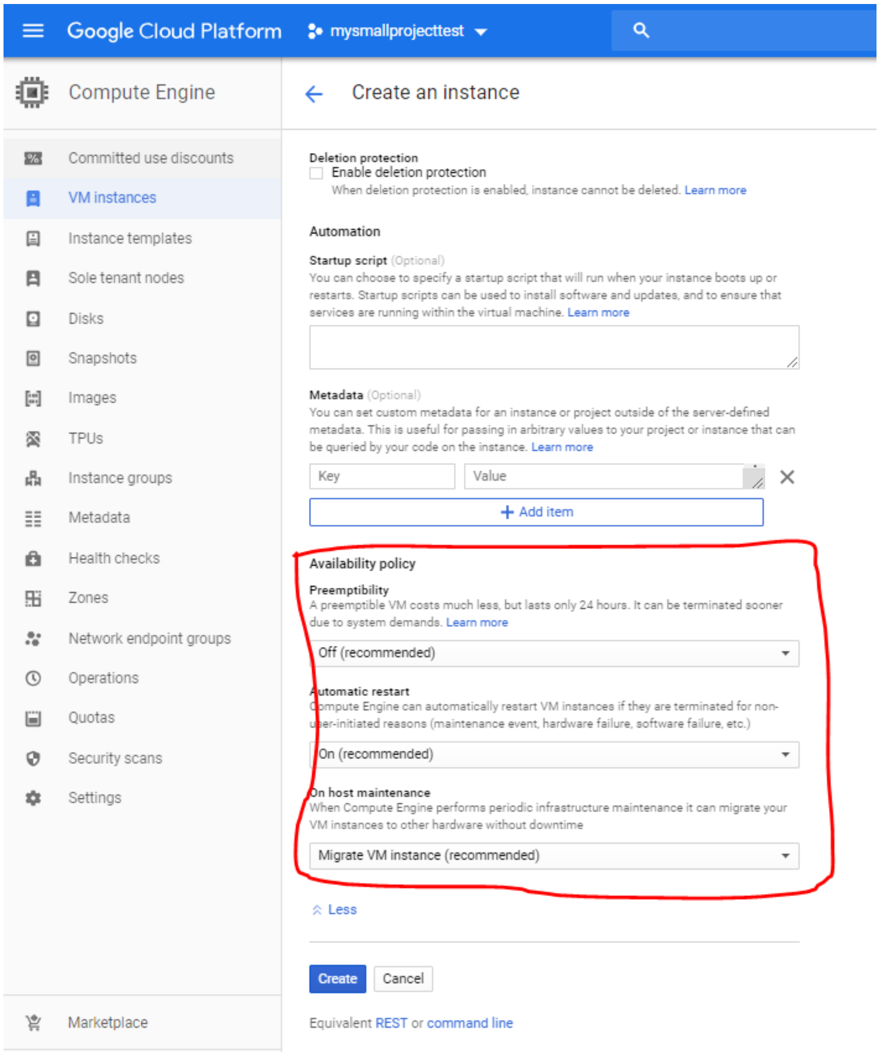
- The VM instance's **maintenance behavior**, which determines whether the instance is live migrated or terminated when there is a maintenance event.
- The instance's **restart behavior**, which determines whether the instance automatically restarts if it crashes or gets terminated.

The default maintenance behavior for instances is to live migrate, but you can change the behavior to terminate your instance during maintenance events instead.

> *On console*

Setting options during instance creation

1. In the GCP Console, go to the VM Instances page.
2. Click **Create Instance**.
3. On the **Create a new instance** page, fill in the desired properties for your instance.
4. Expand the **Management, security, disks, networking, sole tenancy** option.
5. Under **Availability policy**, set the **Automatic restart** and **On host maintenance** options.
6. Click **Create** to create the instance.



Create an instance

Deletion protection

□ Enable deletion protection

When deletion protection is enabled, instance cannot be deleted. Learn more

Automation

Startup script (Optional)

You can choose to specify a startup script that will run when your instance boots up or restarts. Startup scripts can be used to install software and updates, and to ensure that services are running within the virtual machine. Learn more

Metadata (Optional)

You can set custom metadata for an instance or project outside of the server-defined metadata. This is useful for passing in arbitrary values to your project or instance that can be queried by your code on the instance. Learn more

Key | Value

+ Add item

Availability policy

Preemptibility

A preemptible VM costs much less, but lasts only 24 hours. It can be terminated sooner due to system demands. Learn more

Off (recommended)

Automatic restart

Compute Engine can automatically restart VM instances if they are terminated for non-user-initiated reasons (maintenance event, hardware failure, software failure, etc.)

On (recommended)

On host maintenance

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Migrate VM instance (recommended)

Less

Create Cancel

Equivalent REST or command line

Availability in console

> *On SDK*

To specify the availability policies of a new instance in `gcloud compute`, use the `--maintenance-policy` flag to specify whether the instance is `migrated` or `terminated`. By default, instances are automatically set to restart unless you provide the `--no-restart-on-failure` flag.

```
gcloud compute instances create INSTANCE .. \
     [--maintenance-policy MAINTENANCE_POLICY] \
     [--no-restart-on-failure]
```

Testing your availability policies

After you set your availability policies, you can simulate maintenance events to test the effects of these availability policies on your applications. For example, you might simulate a maintenance event on your instances in one of the following situations:

- You have instances that are configured to live migrate during maintenance events and you need to test the effects of live migration on your applications.
- You have batch jobs running on preemptible VM instances and you need to test how your applications handle preemption and shutdown of one or more instances.
- Your instances are configured to terminate and restart during maintenance events rather than live migrate, and you need to test how your applications handle this shutdown and restart process.

You can simulate a maintenance event on an instance using either the `gcloud` command-line tool or an API request.

Run the `instances simulate-maintenance-event` command to force an instance to activate its configured maintenance policy action:

```
gcloud compute instances simulate-maintenance-event
[INSTANCE_NAME] \
    --zone [ZONE]
```

where:

- `[INSTANCE_NAME]` is the name of the instance where you want to simulate the maintenance event. You can specify multiple instance names to simulate maintenance events on more than one instance in the same zone.
- `[ZONE]` is the zone where the instance is located.

————————————————————————————————
— —

Creating an autoscaled managed instance group using an instance template

When you have to build a robust and highly scalable VM instance you have to create an Instance group within a project so that you don't have to manage the instance individually.

compute engine offers two types of instance groups

- managed instance group
- unmanaged instance group

We deal now only with managed instance group.

Managed instance groups

A managed instance group uses an instance template to create a group of identical instances. You control a managed instance group as a single entity. If you wanted to make changes to instances that are part of a managed instance group, you would make the change to the whole instance group. Because managed instance groups contain identical instances, they offer the following features:

- When your applications require additional compute resources, managed instance groups can automatically scale the number of instances in the group.
- Managed instance groups work with load balancing services to distribute traffic to all of the instances in the group.
- If an instance in the group stops, crashes, or is deleted by an action other than the instance groups commands, the managed instance group automatically recreates the instance so it can resume its processing tasks. The recreated instance uses the same name and the same instance template as the previous instance, even if the group references a different instance template.
- Managed instance groups can automatically identify and recreate unhealthy instances in a group to ensure that all of the instances are running optimally.

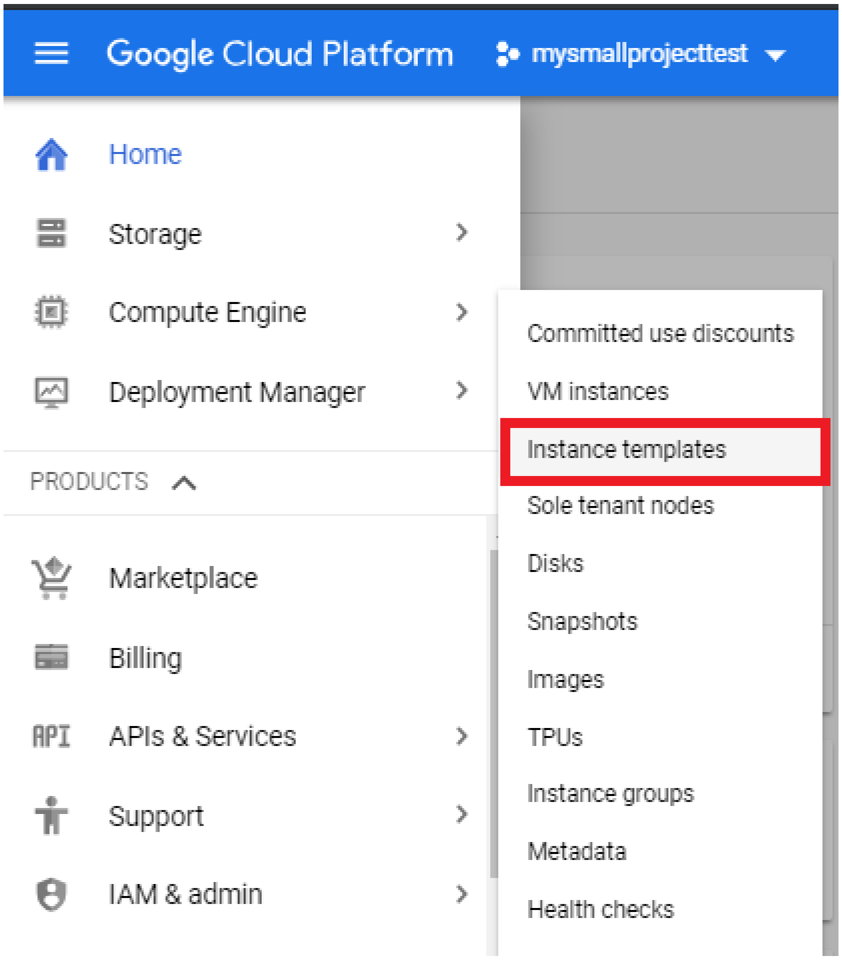
Types of managed instance groups

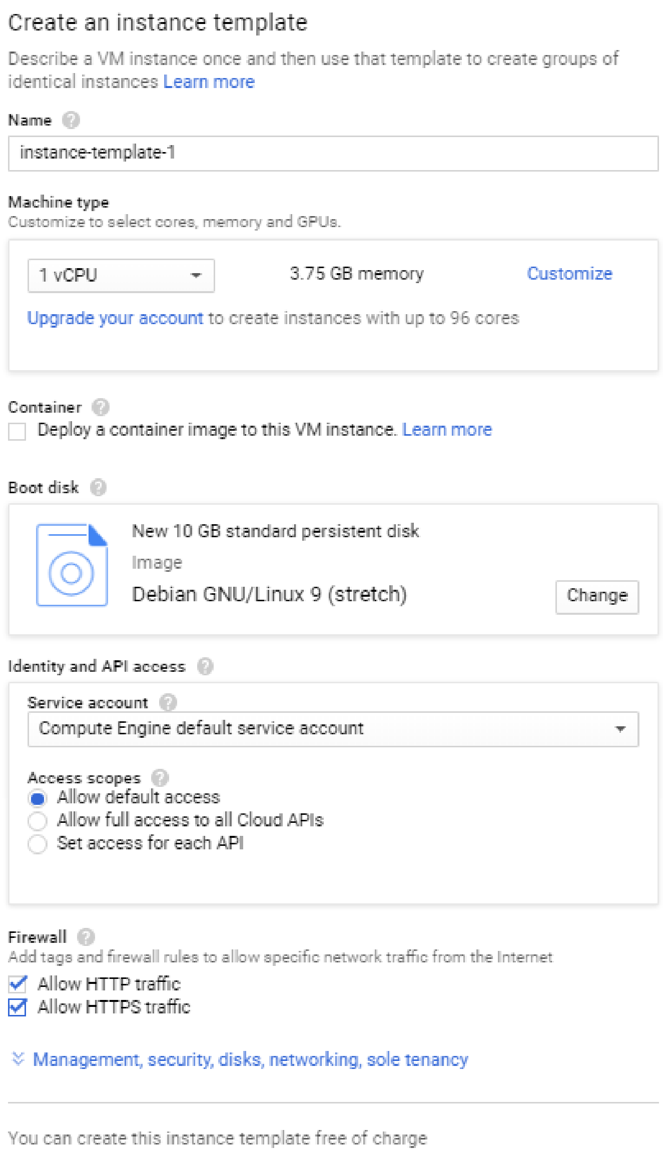
You can create two types of managed instance groups:

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

> *On Cloud Console*

First, go to Compute engine>Instance template and create a machine template that you would like to use in Instance group





Create an instance template

Describe a VM instance once and then use that template to create groups of identical instances Learn more

Name: instance-template-1

Machine type
Customize to select cores, memory and GPUs.

1 vCPU — 3.75 GB memory — Customize

Upgrade your account to create instances with up to 96 cores

Container
□ Deploy a container image to this VM instance. Learn more

Boot disk
New 10 GB standard persistent disk
Image
Debian GNU/Linux 9 (stretch)

Identity and API access
Service account: Compute Engine default service account

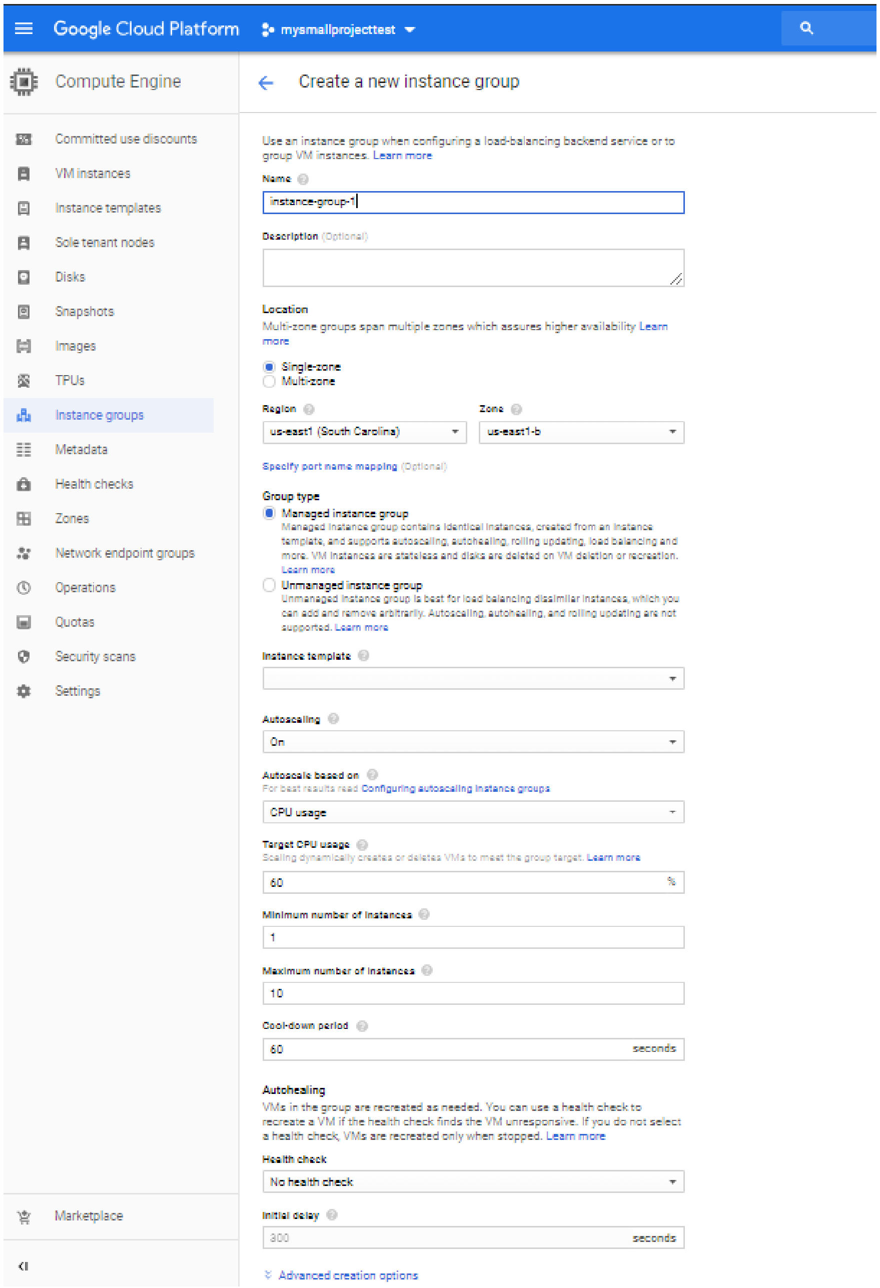
Access scopes
● Allow default access
○ Allow full access to all Cloud APIs
○ Set access for each API

Firewall
Add tags and firewall rules to allow specific network traffic from the Internet
☑ Allow HTTP traffic
☑ Allow HTTPS traffic

Management, security, disks, networking, sole tenancy

You can create this instance template free of charge

Now, save the template and goto Instance group and click on Create instance group



Create a new instance group

Use an instance group when configuring a load-balancing backend service or to group VM instances. Learn more

Name
instance-group-1

Description (Optional)

Location
Multi-zone groups span multiple zones which assures higher availability Learn more

● Single-zone
○ Multi-zone

Region: us-east1 (South Carolina)
Zone: us-east1-b

Specify port name mapping (Optional)

Group type
● Managed instance group
Managed instance group contains identical instances, created from an instance template, and supports autoscaling, autohealing, rolling updating, load balancing and more. VM instances are stateless and disks are deleted on VM deletion or recreation. Learn more
○ Unmanaged instance group
Unmanaged instance group is best for load balancing dissimilar instances, which you can add and remove arbitrarily. Autoscaling, autohealing, and rolling updating are not supported. Learn more

Instance template

Autoscaling
On

Autoscale based on
For best results read Configuring autoscaling instance groups
CPU usage

Target CPU usage
Scaling dynamically creates or deletes VMs to meet the group target. Learn more
60 %

Minimum number of instances
1

Maximum number of instances
10

Cool-down period
60 seconds

Autohealing
VMs in the group are recreated as needed. You can use a health check to recreate a VM if the health check finds the VM unresponsive. If you do not select a health check, VMs are recreated only when stopped. Learn more

Health check
No health check

Initial delay
300 seconds

Advanced creation options

You will be presented with the same dialog as creating new VM.

In place of Instance, template selects the one that you created earlier.

Select Autoscaling and other options as required and save.

> *On SDK*

First, we need to create Instance template

```
gcloud compute instance-templates create
[INSTANCE_TEMPLATE_NAME]
-------------------------------------------------------------
------


prashantagcppaudel@cloudshell:~ (mysmallprojecttest)$ gcloud
compute instance-templates create defaulttemplate
Created
[https://www.googleapis.com/compute/v1/projects/mysmallproje
cttest/global/instanceTemplates/defaulttemplate].
NAME             MACHINE_TYPE   PREEMPTIBLE
CREATION_TIMESTAMP
defaulttemplate  n1-standard-1               2018-10-
24T08:15:57.593-07:00
```

Then, use this template in the instance group

```
@cloudshell:~ (mysmallprojecttest)$ gcloud compute instance-
groups managed set-autoscaling defaultmanagedgroup --max-
num-replicas 10
Did you mean zone [europe-west4-b] for managed instance
group:
[defaultmanagedgroup] (Y/n)?  y

Created
[https://www.googleapis.com/compute/v1/projects/mysmallproje
cttest/zones/europe-west4-b/autoscalers/defaultmanagedgroup-
q6p8].
---
autoscalingPolicy:
  coolDownPeriodSec: 60
  cpuUtilization:
    utilizationTarget: 0.6
  maxNumReplicas: 10
  minNumReplicas: 2
creationTimestamp: '2018-10-24T08:32:48.171-07:00'
id: '7247587989448896112'
kind: compute#autoscaler
name: defaultmanagedgroup-q6p8
selfLink:
https://www.googleapis.com/compute/v1/projects/mysmallprojec
ttest/zones/europe-west4-b/autoscalers/defaultmanagedgroup-
q6p8
status: ACTIVE
target:
https://www.googleapis.com/compute/v1/projects/mysmallprojec
ttest/zones/europe-west4-
b/instanceGroupManagers/defaultmanagedgroup
zone:
https://www.googleapis.com/compute/v1/projects/mysmallprojec
ttest/zones/europe-west4-b
prashantagcppaudel@cloudshell:~ (mysmallprojecttest)$
```

——— -

Generating/uploading a custom SSH key for instances

If you create and manage public SSH keys yourself through the GCP Console, the `gcloud` command-line tool, or the API, you must keep track of the user keys and delete the public SSH keys for users who should not have access. For example, if a team member leaves your project, remove their public SSH keys from metadata so they cannot continue to access your instances.

Additionally, specifying your `gcloud` tool or API calls incorrectly can potentially wipe out all of the public SSH keys in your project or on your instances, which disrupts connections for your project members.

If you are not sure that you want to manage your own keys, use Compute Engine tools to connect to your instance instead.

Overview

By creating and managing SSH keys, you can allow users to access a Linux instance through third-party tools.

An SSH key consists of the following files:

- **A public SSH key file** that is applied to instance-level metadata or project-wide metadata.
- **A private SSH key file** that the user stores on their local devices.

If a user presents their private SSH key, they can use a third-party tool to connect to any instance that is configured with the matching public SSH key file, even if they are not a member of your Cloud Platform project. Therefore, you can control which instances a user can access by changing the public SSH key metadata for one or more instances.

To edit public SSH key metadata:

1. Decide which tool you will use to edit metadata:

- Edit metadata from your browser by using the Google Cloud Platform Console.
- If you prefer the command-line, use the `gcloud` command-line tool to edit metadata.
- If you are an advanced user, you can automate your public SSH key management with API methods.

2. If you need to add users to a Linux instance, prepare their public SSH keys with the following processes:

- If you need to add users who do not have SSH keys, generate a new SSH key for each new user.
- If you need to add users who have existing SSH keys, locate the public SSH key file for each user.
- Format any public SSH keys that you want to add so they will work correctly with the tool that you will use to edit metadata. Optionally, you can also format your public SSH keys to add, edit, or remove expiration times.

3. Edit public SSH key metadata to add or remove users from a Linux instance.

4. Connect to your Linux instance through a third-party tool to ensure that each public SSH key is added or removed correctly. A user can only connect to an instance if their public SSH key is available to the instance through the metadata server and if they have the matching private SSH key.

Creating a new SSH key

If you do not have an existing private SSH key file and a matching public SSH key file that you can use, generate a new SSH key. If you want to use an existing SSH key, locate the public SSH key file.

On Linux or macOS workstations, you can generate a key with the `ssh-keygen` tool.

1. Open a terminal on your workstation and use the `ssh-keygen` command to generate a new key. Specify the `-C` flag to add a

comment with your username.

- `ssh-keygen -t rsa -f ~/.ssh/[KEY_FILENAME] -C [USERNAME]`

where:

- `[KEY_FILENAME]` is the name that you want to use for your SSH key files. For example, a filename of `my-ssh-key` generates a private key file named `my-ssh-key` and a public key file named `my-ssh-key.pub`.
- `[USERNAME]` is the user for whom you will apply this SSH key.

2. This command generates a private SSH key file and a matching public SSH key with the following structure:

- ssh-rsa [KEY_VALUE] [USERNAME]

where:

- `[KEY_VALUE]` is the key value that you generated.
- `[USERNAME]` is the user that this key applies to.

1. Restrict access to your private key so that only you can read it and nobody can write to it.

- `chmod 400 ~/.ssh/[KEY_FILENAME]`

where `[KEY_FILENAME]` is the name that you used for your SSH key files.

Repeat this process for every user for who needs a new key. Then, locate the public SSH keys that you made as well as any existing public SSH keys that you want to add to a project or instance.

Locating an SSH key

There are multiple reasons why you might need to locate an SSH key. For example, if you want to add a user's public SSH key to a project or instance, you will need access to the public key file for their key. Alternatively, you might need to locate your private SSH key file in order to connect to a Linux instance.

When an SSH key is created, it is saved to a default location. The default locations and names of your public and private SSH key files depending on the tools that were used to create that key.

If you created a key on a Linux or macOS workstation by using the `ssh-keygen` tool, your key was saved to the following locations:

- Public key file: `~/.ssh/[KEY_FILENAME].pub`
- Private key file: `~/.ssh/[KEY_FILENAME]`

where `[KEY_FILENAME]` is the filename of the SSH key, which was set when the key was created.

If you need to add or remove the public SSH key from the project or instance metadata, format the public SSH key file.

To check the format of a public SSH key:

1. Locate and open the public SSH key file.
2. Check the format of the public SSH key file.

- If a public SSH key has an expiration time, then it must have the following format:
- ssh-rsa [KEY_VALUE] google-ssh {“userName”:”[USERNAME]”,”expireOn”:”[EXPIRE_TIME]”}

where:

- `[KEY_VALUE]` is the public SSH key value.
- `[USERNAME]` is the user for this SSH key, which was specified when the key was created.
- `[EXPIRE_TIME]` is a value in ISO 8601 format. For example: `2018-12-04T20:12:00+0000`.
- Otherwise, the public SSH key must have the following format:
- ssh-rsa [KEY_VALUE] [USERNAME]

where:

- `[KEY_VALUE]` is the public SSH key value.
- `[USERNAME]` is the user for this SSH key, which was specified when the key was created.

1. If your key does not match one of the above formats or if you want to add, edit, or remove an expiration time, then follow the instructions below to format your public SSH key. Otherwise, leave the file open and add the public SSH key to project or instance metadata.

To format a public SSH key for the console:

1. Make a copy of your public key file. Use the copy with Compute Engine and keep the original file to use with your other SSH configurations.
2. Open the copy of your public key file.
3. Modify the public key file so that it has the following format:

- ssh-rsa [KEY_VALUE] [USERNAME]

where:

- `[KEY_VALUE]` is the public SSH key value.
- `[USERNAME]` is the user for this SSH key, which was specified when the key was created.

1. Alternatively, if you want the public SSH key to have an expiration time, then modify the file to match the following format

- ssh-rsa [KEY_VALUE] google-ssh {“userName”:”[USERNAME]”,”expireOn”:”[EXPIRE_TIME]”}

where:

- `[KEY_VALUE]` is the public SSH key value.
- `[USERNAME]` is the user for this SSH key, which was specified when the key was created.

- `[EXPIRE_TIME]` is a value in ISO 8601 format. For example: `2018-12-04T20:12:00+0000` .

1. Save the changes that you have made, and leave the file open.

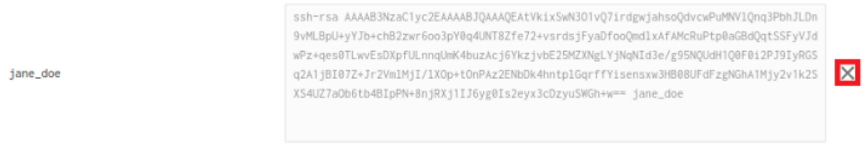
You are now ready to add the public SSH key to project or instance metadata.

Adding or removing project-wide public SSH keys

To add or remove project-wide public SSH keys from the GCP Console:

1. In the Google Cloud Platform Console, go to the metadata page for your project.
2. Under **SSH Keys**, click **Edit**.
3. Modify the project-wide public SSH keys:

- To add a public SSH key, click **Add item** at the bottom of the page. This will produce a text box. Copy the contents of your public SSH key file and paste them into the text box. Repeat this process for each public SSH key that you want to add.
- To remove a public SSH key, click the removal button next to it:



jane_doe

```
ssh-rsa AAAAB3NzaC1yc2EAAAABJQAAAQEAtVkixSwN3O1vQ7irdgwjahsoQdvcwPuMNVlQnq3PbhJLDn
9vMLBpU+yYJb+chB2zwr6oo3pY0q4UNT8Zfe72+vsrdsjFyaDfooQmdlxAfAMcRuPtp0aGBdQqtSSFyVJd
wPz+qes0TLwvEsDXpfULnnqUmK4buzAcj6YkzjvbE25MZXNgLYjNqNId3e/g95NQUdH1Q0F0i2PJ9IyRGS
q2A1jBI07Z+Jr2VmlMjI/1XOp+tOnPAz2ENbDk4hntplGqrffYisensxw3HB08UFdFzgNGhA1Mjy2v1k2S
XS4UZ7aOb6tb4BIpPN+8njRXj1IJ6yg0Is2eyx3cDzyuSWGh+w== jane_doe
```

☒

- Repeat this process for each public SSH key that you want to remove.

When you are done, click **Save** at the bottom of the page.

When you have finished, test your changes by trying to connect to your Linux instance through third-party tools.

If you encounter issues, check the metadata of the instance that you are trying to connect to. If instance-level metadata is set to block project-wide SSH keys or has a deprecated instance-only `sshKeys` value, the instance will ignore all project-wide SSH keys. To apply project-wide keys to an instance, make sure the instance allows project-wide public SSH keys and, if present, remove the deprecated instance-only `sshKeys` value from instance metadata.

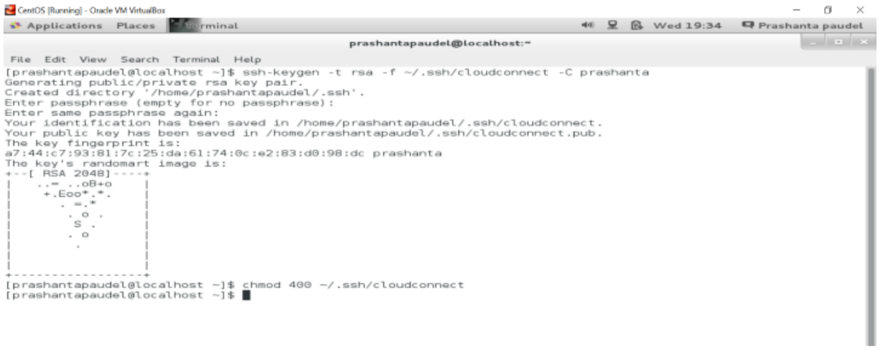
> *Connecting Linux machine to Cloud Project on GCP*

First, create an SSH key in Linux machine with the command

```
ssh-keygen -t rsa -f ~/.ssh/[KEY_FILENAME] -C [USERNAME]
```

Where

- `[KEY_FILENAME]` is the name that you want to use for your SSH key files. For example, a filename of `my-ssh-key` generates a private key file named `my-ssh-key` and a public key file named `my-ssh-key.pub`.
- `[USERNAME]` is the user for whom you will apply this SSH key.



```
[prashantapaudel@localhost ~]$ ssh-keygen -t rsa -f ~/.ssh/cloudconnect -C prashanta
Generating public/private rsa key pair.
Created directory '/home/prashantapaudel/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/prashantapaudel/.ssh/cloudconnect.
Your public key has been saved in /home/prashantapaudel/.ssh/cloudconnect.pub.
The key fingerprint is:
a7:44:c7:93:81:7c:25:da:61:74:0c:e2:83:d0:98:dc prashanta
The key's randomart image is:
+--[ RSA 2048]----+
|    ..= ..oB+o   |
|     +.Eoo*.*.   |
|       . =.*     |
|        . o .    |
|         S .     |
|        . o      |
|         .       |
|                 |
|                 |
+-----------------+
[prashantapaudel@localhost ~]$ chmod 400 ~/.ssh/cloudconnect
[prashantapaudel@localhost ~]$ 
```

Then secure the file with the command

```
chmod 400 ~/.ssh/[KEY_FILENAME]
```

Check the Key by

```
vi ~/.ssh/[KEY_FILENAME]
```

You will see random long characters

Also, check the public key file and copy the key into notepad

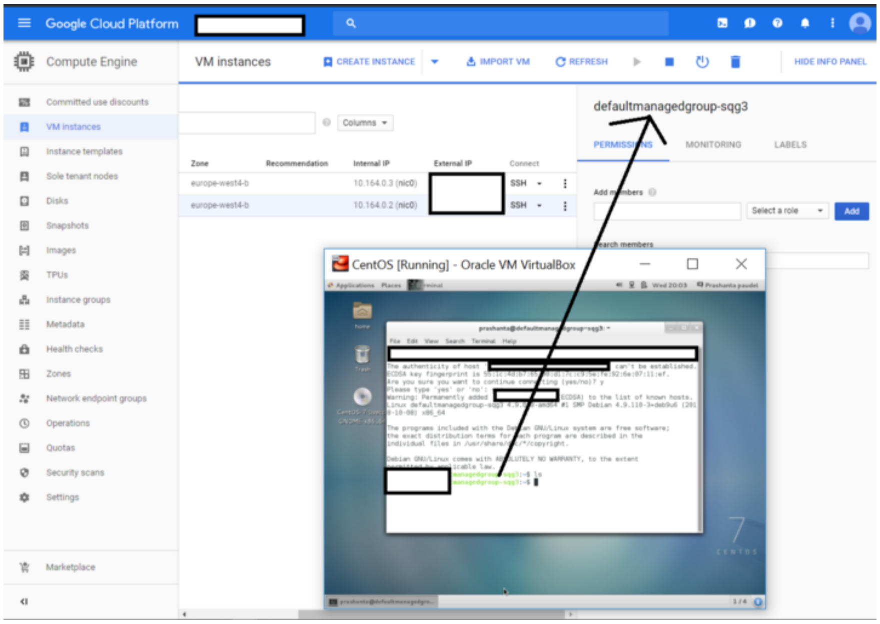
Goto COmpute Engine>Metadata>SSH Keys

paste the keys into the box and save.

Now, go to the Linux machine and use the command

SSh -i ~/.ssh/[FILENAME] [USER]@IP ADDRESS

This should connect you to the instance.



Zone	Recommendation	Internal IP	External IP	Connect
europe-west4-b		10.164.0.3 (nic0)	[redacted]	SSH
europe-west4-b		10.164.0.2 (nic0)		SSH

```
[redacted]
The authenticity of host '[redacted] can't be established.
ECDSA key fingerprint is 55:1c:4d:b7:65:00:d1:7c:c9:5e:fe:92:6e:07:11:ef.
Are you sure you want to continue connecting (yes/no)? y
Please type 'yes' or 'no':
Warning: Permanently added [redacted] (ECDSA) to the list of known hosts.
Linux defaultmanagedgroup-sqg3 4.9.0-8-amd64 #1 SMP Debian 4.9.110-3+deb9u6 (201
8-10-08) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
[redacted]managedgroup-sqg3:~$ ls
managedgroup-sqg3:~$
```

A connection between Linux and Google cloud Instance

____ -

Configuring a VM for Stackdriver monitoring and logging

A monitoring agent installed in VM is a **collectd** based daemon that gathers system metrics for virtual machine instances and sends them to stack driver monitoring.

By default, Monitoring collects disk, CPU, Network, and process metrics.

You can configure the monitoring agent to monitor third-party applications to get the full list of agent matrices.

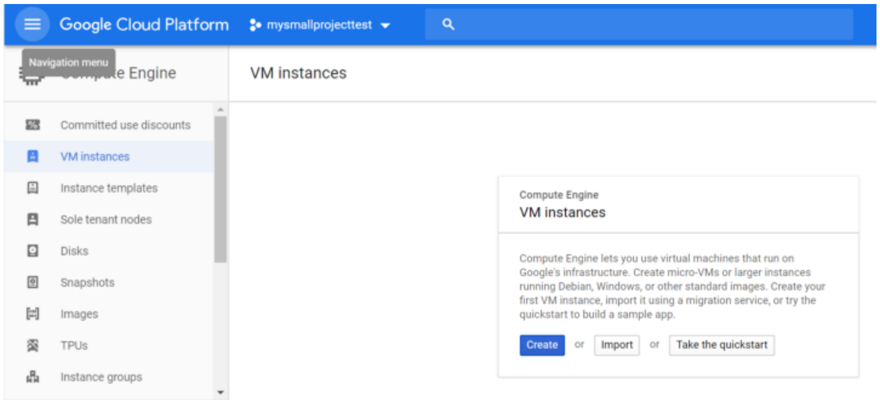
Using the Monitoring agent is optional but recommended. Monitoring can access some metrics without the Monitoring agent, including CPU utilization, some disk traffic metrics, network traffic, and uptime information. Monitoring uses the Monitoring agent to access additional system resources and application services in virtual machine (VM) instances. If you want these additional capabilities, you should install the Monitoring agent.

Installing monitoring Agent in VM

To install the agent, you should have the following:

- A supported VM instance in a GCP project or AWS account.
- A Workspace monitoring the GCP project or AWS account containing the VM instance.
- Credentials on the VM instance that authorize communication with Stackdriver. GCP VM instances generally have the credentials by default. For AWS instances, you have to install the credentials. For details, see Adding Credentials.

Step 1: Install VM

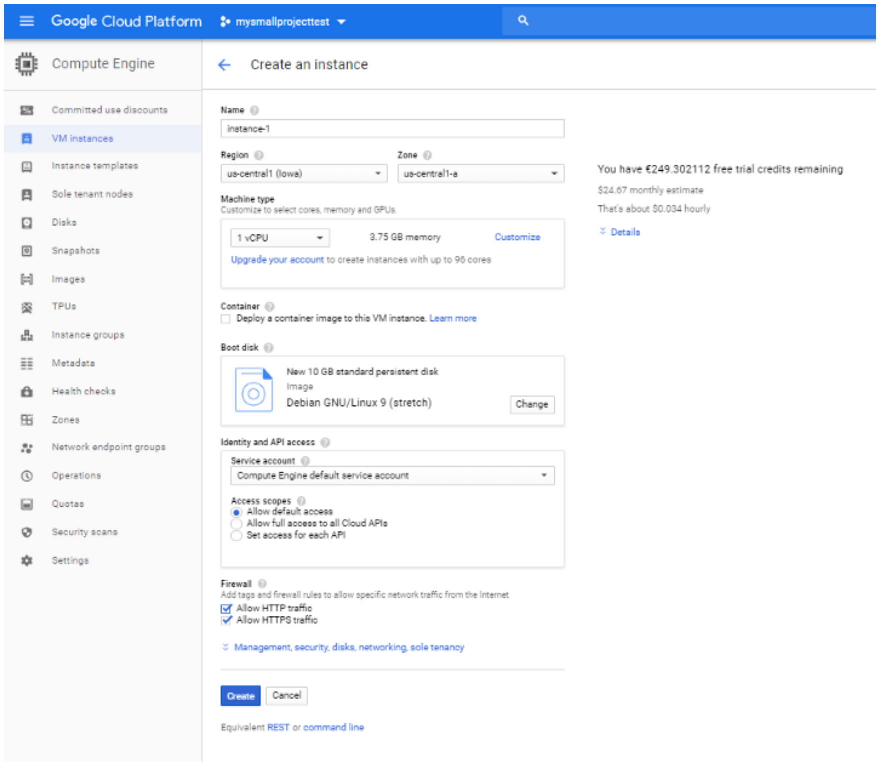


Compute Engine
VM instances

Compute Engine lets you use virtual machines that run on Google's infrastructure. Create micro-VMs or larger instances running Debian, Windows, or other standard images. Create your first VM instance, import it using a migration service, or try the quickstart to build a sample app.

Create or Import or Take the quickstart

Create VM



Create an instance

Name: instance-1

Region: us-central1 (Iowa)

Zone: us-central1-a

Machine type
Customize to select cores, memory and GPUs.

1 vCPU 3.75 GB memory Customize

Upgrade your account to create instances with up to 96 cores

Container
□ Deploy a container image to this VM instance. Learn more

Boot disk
New 10 GB standard persistent disk
Image
Debian GNU/Linux 9 (stretch) Change

Identity and API access
Service account: Compute Engine default service account

Access scopes
● Allow default access
○ Allow full access to all Cloud APIs
○ Set access for each API

Firewall
Add tags and firewall rules to allow specific network traffic from the Internet
☑ Allow HTTP traffic
☑ Allow HTTPS traffic

Management, security, disks, networking, sole tenancy

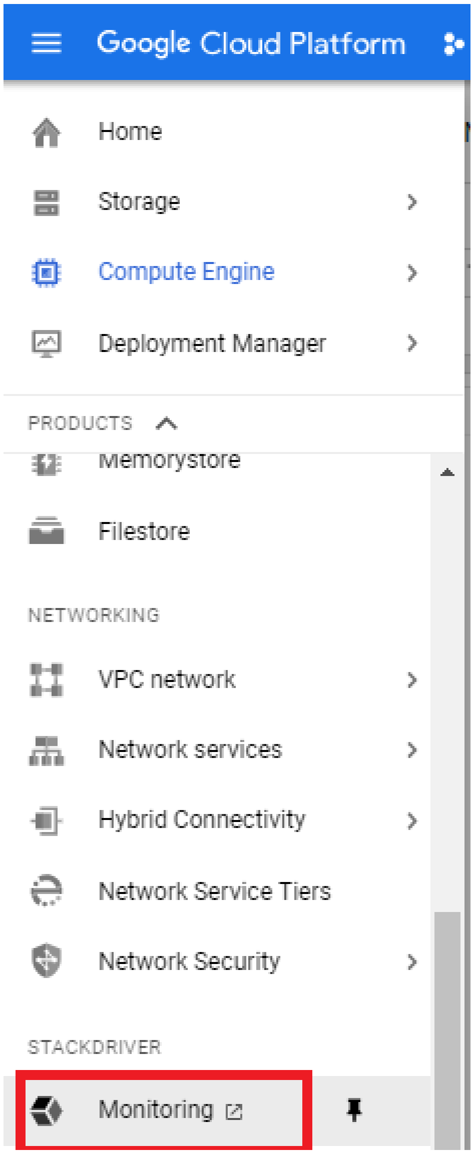
Create Cancel

Equivalent REST or command line

You have €249.302112 free trial credits remaining
$24.67 monthly estimate
That's about $0.034 hourly
Details

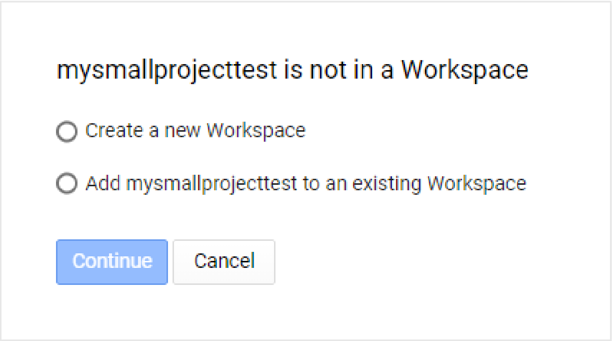
Creating a VM

Step 2: Create stack driver workspace for your project



Goto monitoring

It will redirect you to stackdriver workspace

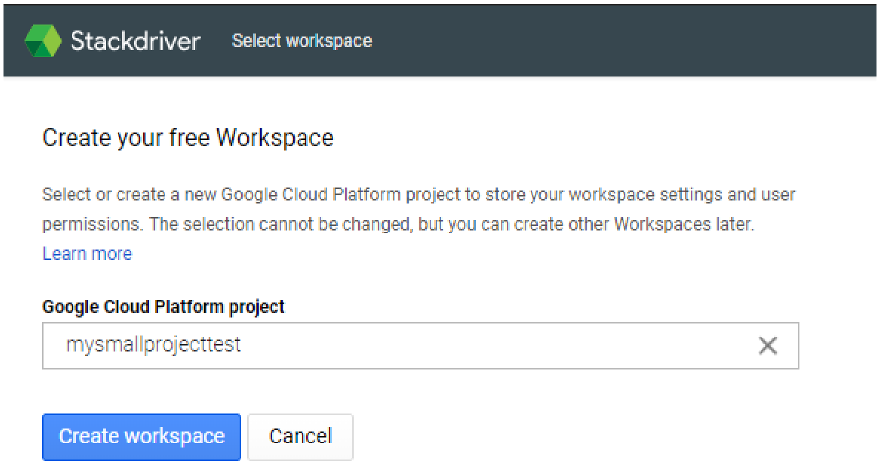


mysmallprojecttest is not in a Workspace

○ Create a new Workspace

○ Add mysmallprojecttest to an existing Workspace

If you have not created stack driver workspace yet create one.



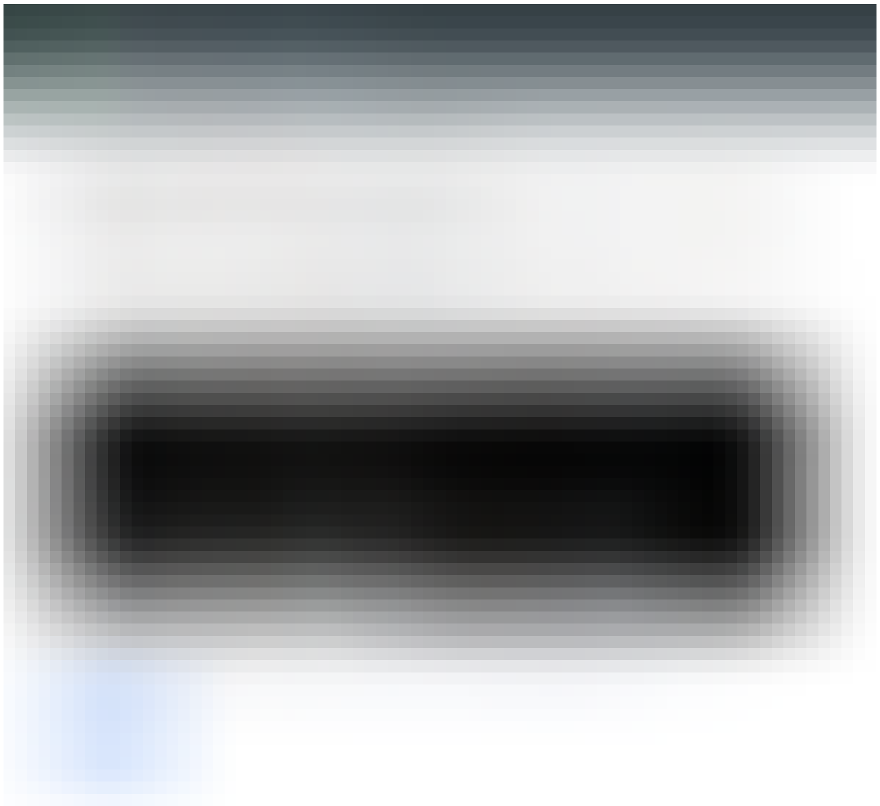
Create your free Workspace

Select or create a new Google Cloud Platform project to store your workspace settings and user permissions. The selection cannot be changed, but you can create other Workspaces later. Learn more

Google Cloud Platform project: mysmallprojecttest

select project

then it will ask to install an agent in VM



install an agent in VM

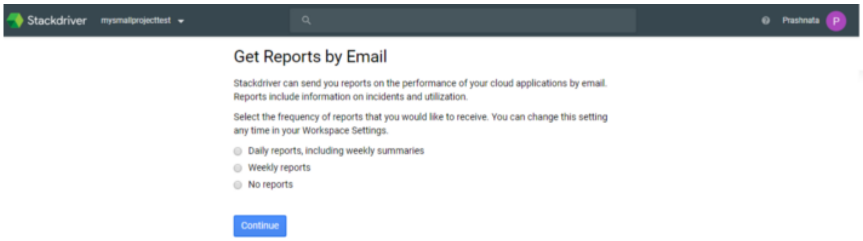
Now, I login into my VM and install agent as mentioned above and installed agents

```
Connected, host fingerprint: ssh-rsa 2048
72:9E:C7:4C:82:22:32:DD:B3:07:B4:C6:75:24:55:A5:24:4B:04:4E:
24:CF:FE:FD:33:85:EA:82:F1:48:5F:CELinux instance-1 4.9.0-8-
amd64 #1 SMP Debian 4.9.110-3+deb9u6 (2018-10-08) x86_64The
programs included with the Debian GNU/Linux system are free
software;the exact distribution terms for each program are
described in theindividual files in
/usr/share/doc/*/copyright.Debian GNU/Linux comes with
ABSOLUTELY NO WARRANTY, to the extentpermitted by applicable
law.prashantagcppaudel@instance-1:~$ curl -sSO
https://dl.google.com/cloudagents/install-monitoring-
agent.shprashantagcppaudel@instance-1:~$ sudo bash install-
monitoring-
agent.sh============================================================
==========================Starting installation of
stackdriver-
agent=========================================================
=======================Installing agent for Debian or
Ubuntu.OKReading package lists…Building dependency tree…
Reading state information…The following additional packages
will be installed: libltdl7 libpython2.7 libyajl2Suggested
packages: libmariadbclient18 libpq5 libhiredis0.13 default-
jreThe following NEW packages will be installed: libltdl7
libpython2.7 libyajl2 stackdriver-agent0 upgraded, 4 newly
installed, 0 to remove and 5 not upgraded.Need to get 3,220
kB of archives.After this operation, 9,982 kB of additional
disk space will be used.Get:1 http://security.debian.org
stretch/updates/main amd64 libpython2.7 amd64 2.7.13-
2+deb9u3 [1,071 kB]Get:2 http://deb.debian.org/debian
stretch/main amd64 libltdl7 amd64 2.4.6-2 [389 kB]Get:3
http://packages.cloud.google.com/apt google-cloud-
monitoring-stretch/main amd64 stackdriver-agent amd64 5.5.2-
382.stretch [1,736 kB]Get:4 http://deb.debian.org/debian
stretch/main amd64 libyajl2 amd64 2.1.0-2+b3 [23.2
kB]Fetched 3,220 kB in 0s (23.9 MB/s)Selecting previously
unselected package libltdl7:amd64.(Reading database … 33712
files and directories currently installed.)Preparing to
unpack …/libltdl7_2.4.6-2_amd64.deb …Unpacking
libltdl7:amd64 (2.4.6-2) …Selecting previously unselected
package libpython2.7:amd64.Preparing to unpack
…/libpython2.7_2.7.13-2+deb9u3_amd64.deb …Unpacking
libpython2.7:amd64 (2.7.13-2+deb9u3) …Selecting previously
unselected package libyajl2:amd64.Preparing to unpack
…/libyajl2_2.1.0-2+b3_amd64.deb …Unpacking libyajl2:amd64
(2.1.0-2+b3) …Selecting previously unselected package
stackdriver-agent.Preparing to unpack …/stackdriver-
agent_5.5.2-382.stretch_amd64.deb …Unpacking stackdriver-
agent (5.5.2-382.stretch) …Setting up libyajl2:amd64 (2.1.0-
2+b3) …Processing triggers for libc-bin (2.24-11+deb9u3) …
Processing triggers for systemd (232-25+deb9u4) …Setting up
libltdl7:amd64 (2.4.6-2) …Setting up libpython2.7:amd64
(2.7.13-2+deb9u3) …Setting up stackdriver-agent (5.5.2-
382.stretch) …Processing triggers for libc-bin (2.24-
11+deb9u3) …Processing triggers for systemd (232-25+deb9u4)
```

```
…
============================================================
==================Installation of stackdriver-agent-5.5.2-
382 completed successfully.Please consult the documentation
for troubleshooting advice:
https://cloud.google.com/monitoring/agent
You can monitor the monitoring agent's logfile at:
  /var/log/syslog
============================================================
==================
prashantagcppaudel@instance-1:~$ curl -sSO
https://dl.google.com/cloudagents/install-logging-agent.sh
prashantagcppaudel@instance-1:~$ sudo bash install-logging-
agent.sh
============================================================
==================
Starting installation of google-fluentd
============================================================
==================

Installing agents for Debian or Ubuntu.
OK
Selecting previously unselected package google-fluentd.
(Reading database ... 34145 files and directories currently
installed.)
Preparing to unpack .../google-fluentd_1.6.0-1_amd64.deb ...
Unpacking google-fluentd (1.6.0-1) ...
Selecting previously unselected package google-fluentd-
catch-all-config.
Preparing to unpack .../google-fluentd-catch-all-
config_0.7_all.deb ...
Unpacking google-fluentd-catch-all-config (0.7) ...
Setting up google-fluentd (1.6.0-1) ...
Adding system user `google-fluentd' (UID 108) ...
Adding new group `google-fluentd' (GID 112) ...
Adding new user `google-fluentd' (UID 108) with group
`google-fluentd' ...
Not creating home directory `/home/google-fluentd'.
Installing default conffile /etc/google-fluentd/google-
fluentd.conf ...
Setting up google-fluentd-catch-all-config (0.7) ...

============================================================
==================
Installation of google-fluentd complete.

Logs from this machine should be visible in the log viewer
at:
  https://console.cloud.google.com/logs/viewer?
project=mysmallprojecttest&resource=gce_instance/instance_id
/3368325578686794533

A test message has been sent to syslog to help verify proper
operation.

Please consult the documentation for troubleshooting advice:
  https://cloud.google.com/logging/docs/agent

You can monitor the logging agent's logfile at:
  /var/log/google-fluentd/google-fluentd.log
```

```
==============================================================
==================
prashantagcppaudel@instance-1:~$
```

Now you will be given the option to get reports



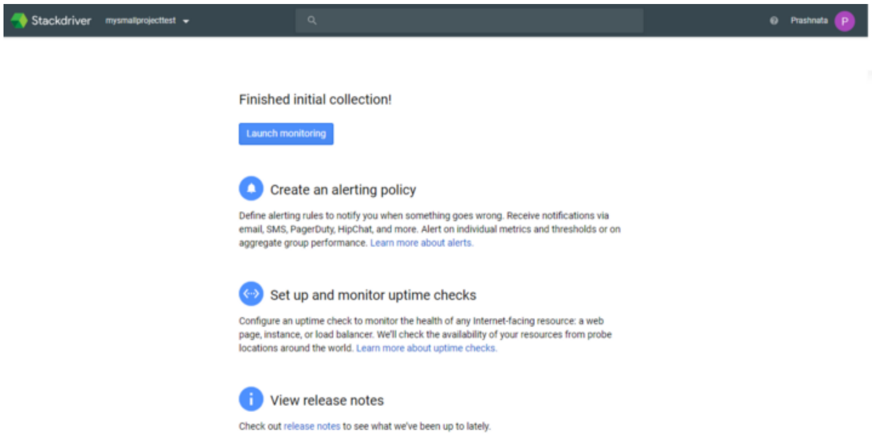
Get Reports by Email

Stackdriver can send you reports on the performance of your cloud applications by email. Reports include information on incidents and utilization.

Select the frequency of reports that you would like to receive. You can change this setting any time in your Workspace Settings.

○ Daily reports, including weekly summaries
○ Weekly reports
○ No reports

Continue



Finished initial collection!

Launch monitoring

Create an alerting policy

Define alerting rules to notify you when something goes wrong. Receive notifications via email, SMS, PagerDuty, HipChat, and more. Alert on individual metrics and thresholds or on aggregate group performance. Learn more about alerts.

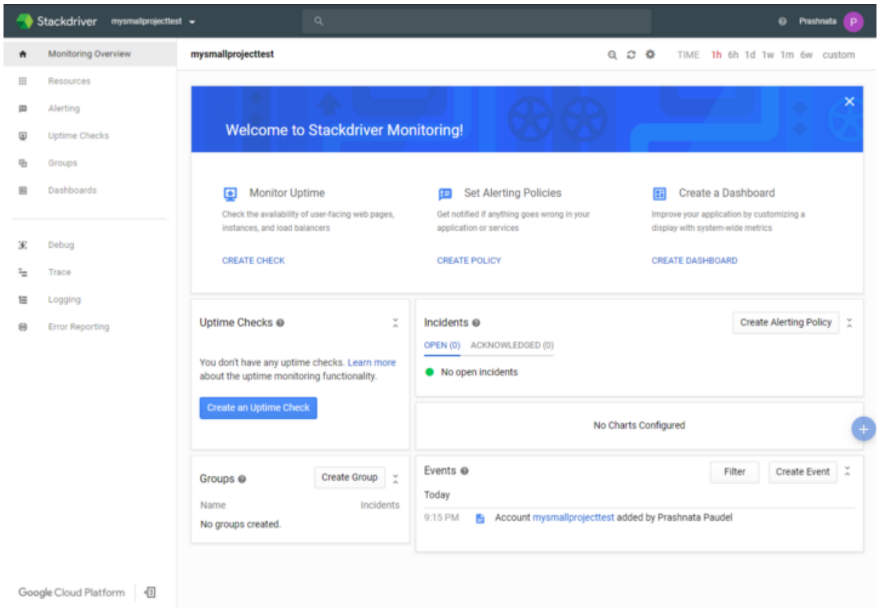
Set up and monitor uptime checks

Configure an uptime check to monitor the health of any Internet-facing resource: a web page, instance, or load balancer. We'll check the availability of your resources from probe locations around the world. Learn more about uptime checks.

View release notes

Check out release notes to see what we've been up to lately.

Now your VM is under monitoring in stackdriver



Welcome to Stackdriver Monitoring!

Check the availability of user-facing web pages, instances, and load balancers

Get notified if anything goes wrong in your application or services

Improve your application by customizing a display with system-wide metrics

You don't have any uptime checks. Learn more about the uptime monitoring functionality.

No open incidents

No Charts Configured

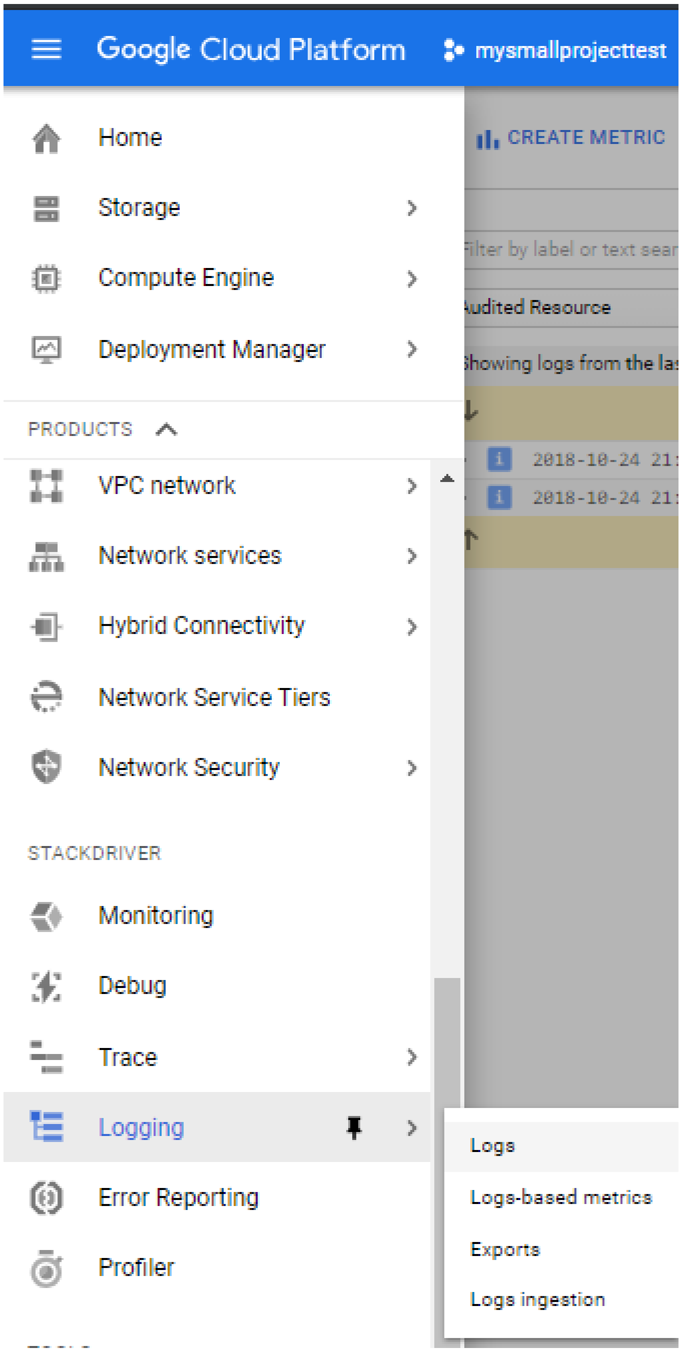
No groups created.

9:15 PM Account mysmallprojecttest added by Prashnata Paudel

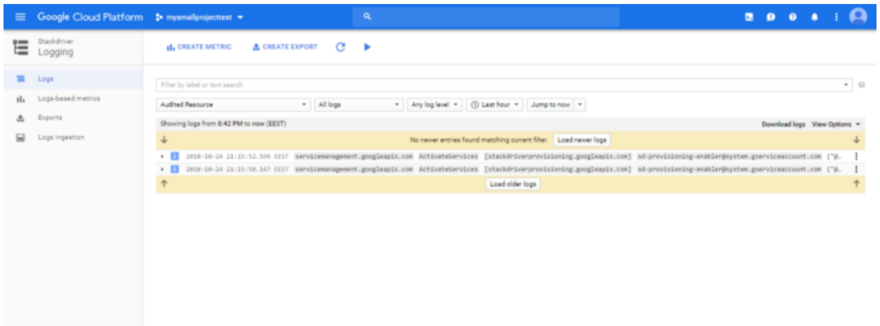
Stackdriver workspace

For logging

Goto Compute Engine>Logging



Here you will see all the logs from GCP in one place.



Showing logs from 8:42 PM to now (EEST)

No newer entries found matching current filter.

```
2018-10-24 21:15:52.596 EEST servicemanagement.googleapis.com ActivateServices [stackdriverprovisioning.googleapis.com] sd-provisioning-enabler@system.gserviceaccount.com {"@...
2018-10-24 21:15:50.147 EEST servicemanagement.googleapis.com ActivateServices [stackdriverprovisioning.googleapis.com] sd-provisioning-enabler@system.gserviceaccount.com {"@...
```

logging

———————————————————————————————————————

Assessing compute quotas and requesting increases

Resource Quotas

Compute Engine enforces quotas on resource usage for a variety of reasons. For example, quotas protect the community of Google Cloud Platform users by preventing unforeseen spikes in usage. Google Cloud Platform also offers Free trial quotas that provide limited access for projects that are just exploring Google Cloud Platform on a free trial basis.

Not all projects have the same quotas. As your use of Google Cloud Platform expands over time, your quotas may increase accordingly. If you expect a notable upcoming increase in usage, you can proactively request quota adjustments from the Quotas page in the GCP Console.

Free Trial Quotas

Compute Engine limitations

During the free trial, the following limitations apply to your Compute Engine resources:

- Your project can have no more than 8 cores (or virtual CPUs) running at the same time. To run more than 8 vCPUs at a time, you must upgrade your account.

- For more information about the types of virtual machines available and the number of cores they use, see Machine type pricing.
- You cannot add GPUs to your Compute Engine instances.
- Preemptible VM instances are not included in the free trial.
- You cannot request a quota increase. For an overview of quotas, see Resource Quotas.

Cloud TPU usage

Cloud TPUs are not part of the Always Free program.

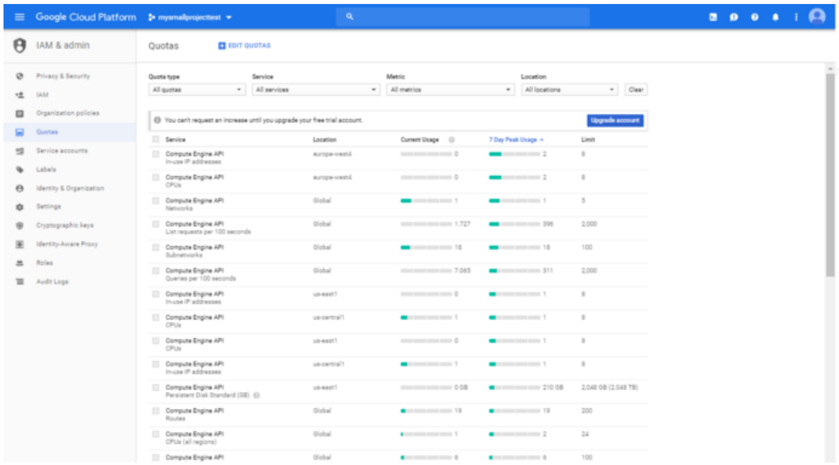
Restricted actions

Some actions are prohibited during the free trial. For example, during your free trial, you may not use Google Cloud Platform services to engage in mining cryptocurrency.

For additional restrictions, see the Free Trial Terms of Service agreement and the Terms of Service for Google Cloud Platform.

Checking Quotas

In GCP Console Goto IAM and Services > Quotas



Quota type: All quotas
Service: All services
Metric: All metrics
Location: All locations

You can't request an increase until you upgrade your free trial account.

Service	Location	Current Usage	7 Day Peak Usage	Limit
Compute Engine API In-use IP addresses	europe-west4	0	2	8
Compute Engine API CPUs	europe-west4	0	2	8
Compute Engine API Networks	Global	1	1	5
Compute Engine API List requests per 100 seconds	Global	1,727	396	2,000
Compute Engine API Subnetworks	Global	18	18	100
Compute Engine API Queries per 100 seconds	Global	7,065	311	2,000
Compute Engine API In-use IP addresses	us-east1	0	1	8
Compute Engine API CPUs	us-central1	1	1	8
Compute Engine API CPUs	us-east1	0	1	8
Compute Engine API In-use IP addresses	us-central1	1	1	8
Compute Engine API Persistent Disk Standard (GB)	us-east1	0 GB	210 GB	2,048 GB (2,048 TB)
Compute Engine API Routes	Global	19	19	200
Compute Engine API CPUs (all regions)	Global	1	2	24
Compute Engine API	Global	6	6	100

Quotas

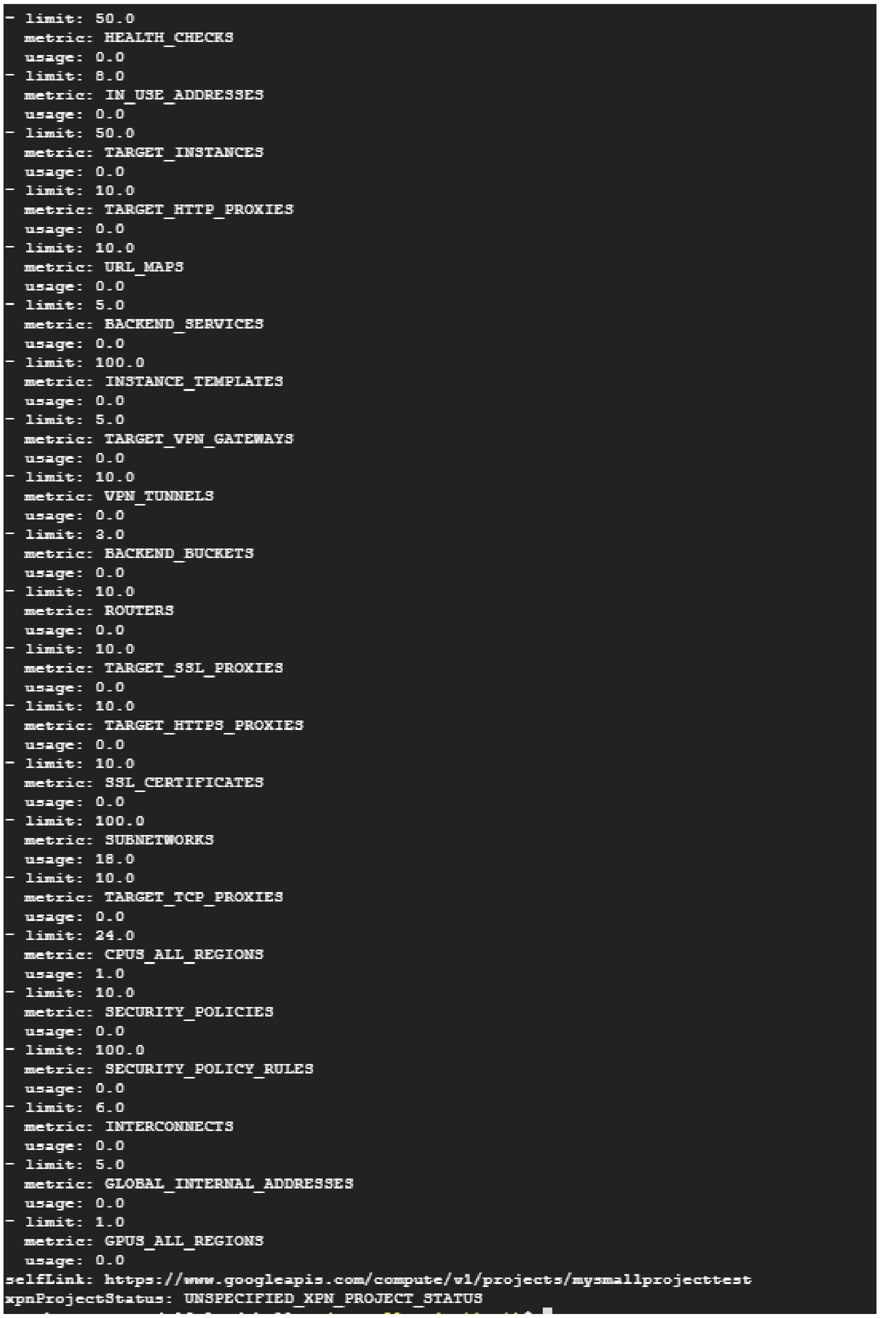
Here you can see all the quotas allocated for you in GCP

Alternatively, you can use SDK to check quotas

```
gcloud compute project-info describe --project myproject
```

To check your used quota in a region, run:

```
gcloud compute regions describe example-region
```



```
- limit: 50.0
  metric: HEALTH_CHECKS
  usage: 0.0
- limit: 8.0
  metric: IN_USE_ADDRESSES
  usage: 0.0
- limit: 50.0
  metric: TARGET_INSTANCES
  usage: 0.0
- limit: 10.0
  metric: TARGET_HTTP_PROXIES
  usage: 0.0
- limit: 10.0
  metric: URL_MAPS
  usage: 0.0
- limit: 5.0
  metric: BACKEND_SERVICES
  usage: 0.0
- limit: 100.0
  metric: INSTANCE_TEMPLATES
  usage: 0.0
- limit: 5.0
  metric: TARGET_VPN_GATEWAYS
  usage: 0.0
- limit: 10.0
  metric: VPN_TUNNELS
  usage: 0.0
- limit: 3.0
  metric: BACKEND_BUCKETS
  usage: 0.0
- limit: 10.0
  metric: ROUTERS
  usage: 0.0
- limit: 10.0
  metric: TARGET_SSL_PROXIES
  usage: 0.0
- limit: 10.0
  metric: TARGET_HTTPS_PROXIES
  usage: 0.0
- limit: 10.0
  metric: SSL_CERTIFICATES
  usage: 0.0
- limit: 100.0
  metric: SUBNETWORKS
  usage: 18.0
- limit: 10.0
  metric: TARGET_TCP_PROXIES
  usage: 0.0
- limit: 24.0
  metric: CPUS_ALL_REGIONS
  usage: 1.0
- limit: 10.0
  metric: SECURITY_POLICIES
  usage: 0.0
- limit: 100.0
  metric: SECURITY_POLICY_RULES
  usage: 0.0
- limit: 6.0
  metric: INTERCONNECTS
  usage: 0.0
- limit: 5.0
  metric: GLOBAL_INTERNAL_ADDRESSES
  usage: 0.0
- limit: 1.0
  metric: GPUS_ALL_REGIONS
  usage: 0.0
selfLink: https://www.googleapis.com/compute/v1/projects/mysmallprojecttest
xpnProjectStatus: UNSPECIFIED_XPN_PROJECT_STATUS
```

gcloud quota

If you have a paid account you will get the option to request an increase in quota.

———————————————————————————————————————

In this way, we completed all the tasks required in 3.1

Goodnight!