Architecture Diagram

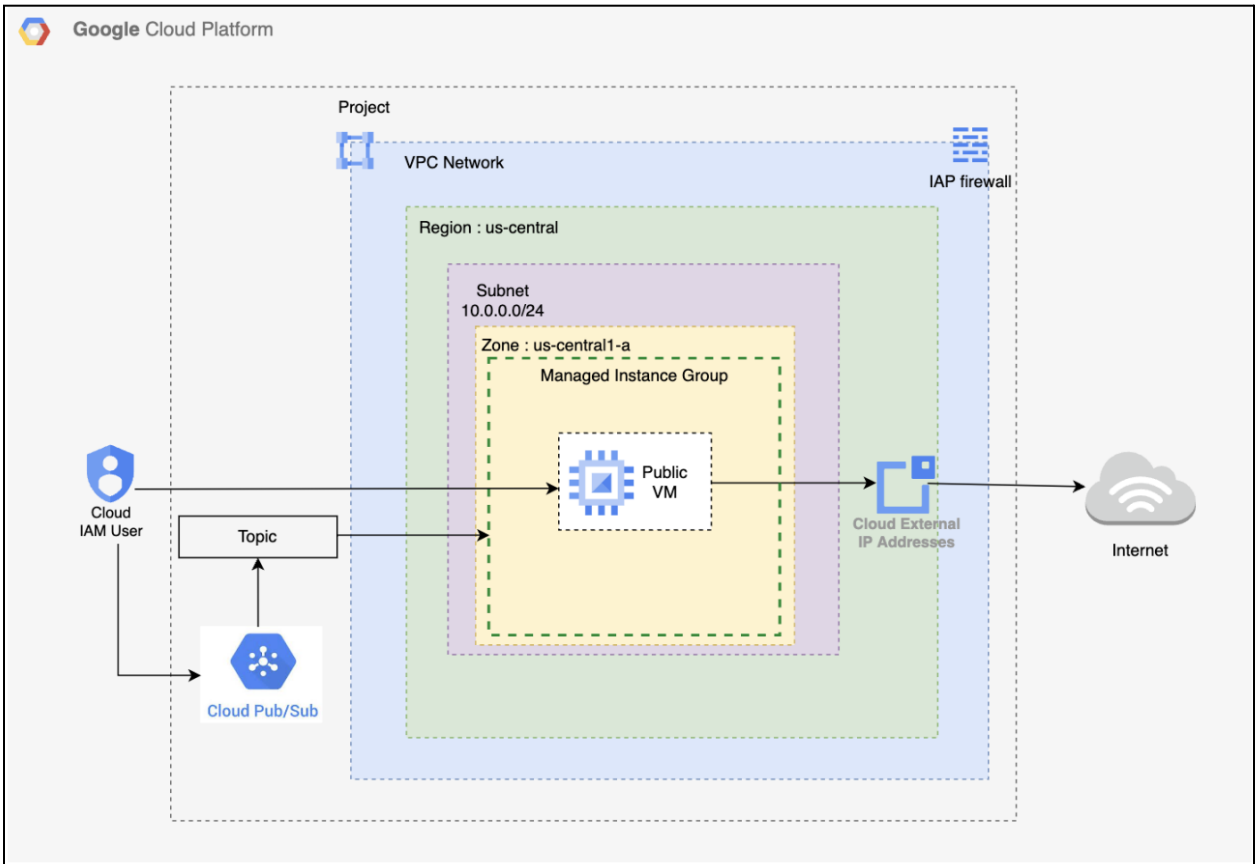


Figure: Lab_5 - Managed Instance Group Autoscaling

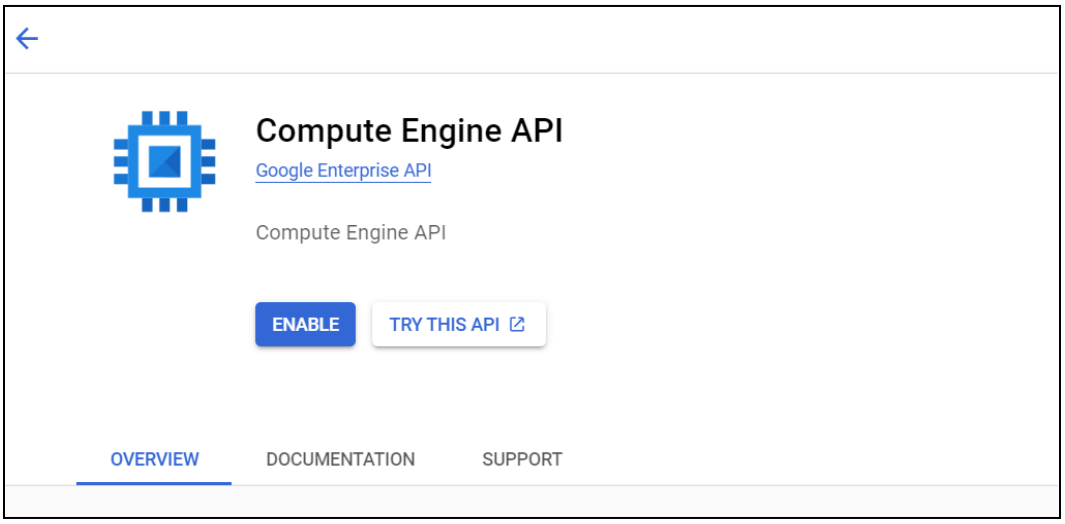
Autoscaling VM based CPU Utilization & pub/sub triggers to simulate load and autoscale the nodes

Prerequisites

Create and manage Google Cloud resources and services directly on the command line.

To create a private virtual machine in GCP following resources such as custom VPC, custom subnets, NAT gateway and IAP firewall rule is required to be provisioned.

Enable the Compute Engine API - [Compute Engine API]



Note: If you have followed Lab 1 for creation of VPC, subnets, NAT gateway, Firewall and Service account, Below prerequisites 4 steps are not required as you may have the resources created previously.

For someone who has not followed the previous lab follow all the prerequisites steps

1. Create a **Virtual Private Cloud (VPC) & Subnet**

   VPC provides networking for your cloud-based resources and services that are global, scalable, and flexible.

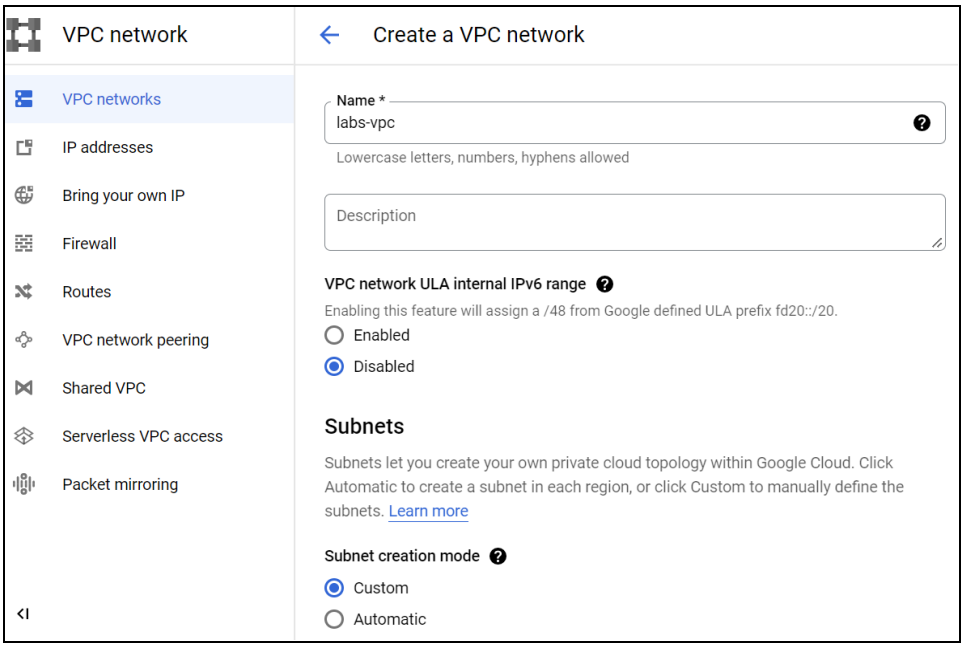
Deploying Virtual Private Connect & Subnet using Google Cloud Console

Metadata - VPC & Subnet	**Naming Convention**
Name	labs-vpc
Subnet Creation Mode	Custom
New Subnet Name Region IP Stack type IPv4 Range Dynamic Routing	 labs-subnet us-central1 IPv4 10.0.0.0/24 Regional

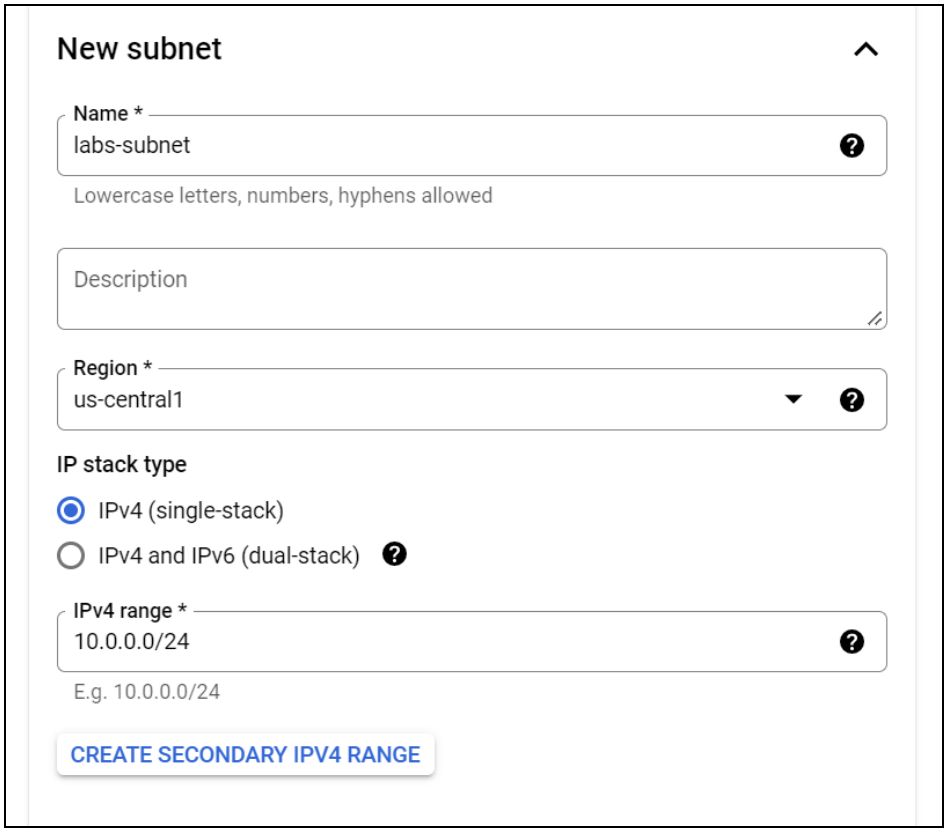
- In the Cloud Console, on the **Navigation menu** (☰), click **VPC network** > **VPC networks**.
  This may take a minute to initialize for the first time.
- To create a new VPC, click **CREATE VPC NETWORK**.

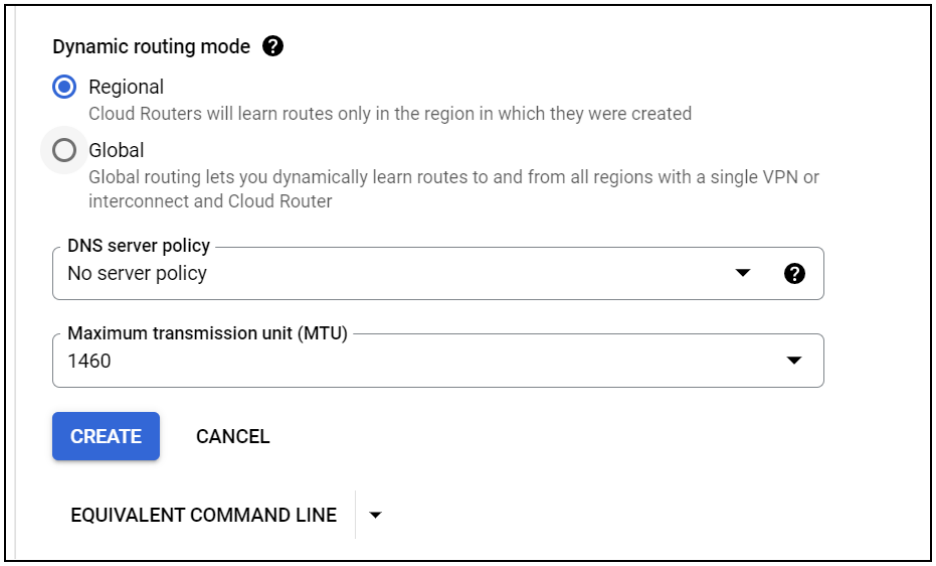
  There are many parameters you can configure when creating a **new VPC**.

- Enter VPC **Name** as it is a mandatory field.
- Description can be kept blank as it is an optional field.
- Select **Custom** checkbox for Subnet to create custom subnet.



- Click on **Add Subnet**.
- Enter Subnet **Name** as it is a mandatory field.
- Description can be kept blank as it is an optional field.
- Region here selected is us-central1. Region can be chosen upon end user's requirements to obtain low latency of resources.
- For more information about regions, see the Compute Engine guide, Regions and Zones.
- IP Stack type can be left default with IPv4 checkbox selected.
- Enter **IPv4 range** for Subnet depending upon requirement.
- Rest of the fields can be left with default values.
- Click the **CREATE** button to create a new VPC.





- Following resources are created
- VPC named **labs-vpc** and subnet named **labs-subnet**.

2. Create **NAT Gateway**

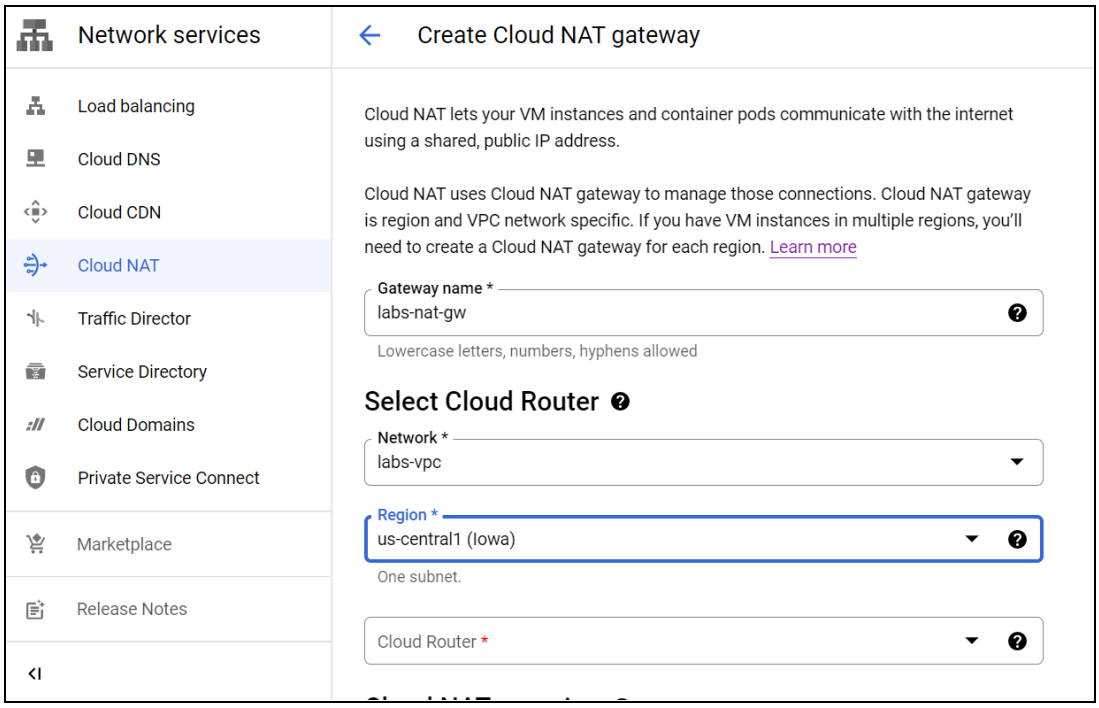
   Cloud NAT (network address translation) lets certain resources without external IP addresses create outbound connections to the internet.

Deploying NAT Gateway using Google Cloud Console

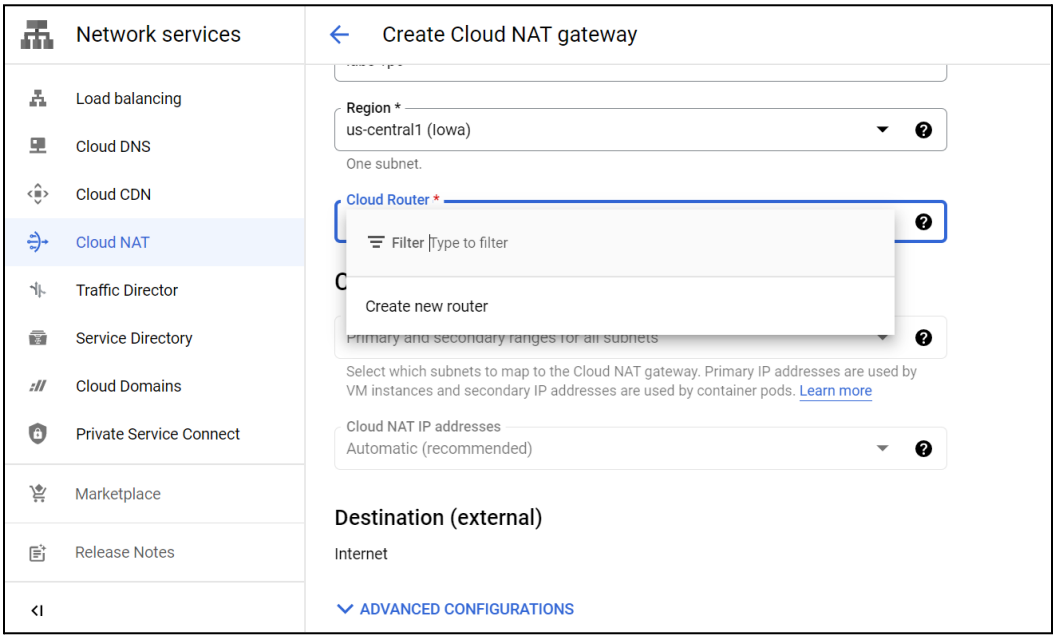
Metadata - NAT Gateway	**Naming Convention**
Name	labs-nat-gw
Select Cloud Router Network Region	 labs-vpc us-central1
Create a Router Name	 labs-nat-router
Cloud NAT Mapping Resource (Internal)	 Primary and Secondary Ranges for all subnets

- In the Cloud Console, on the **Navigation menu** (☰), click **Network services** > **Cloud NAT**.
- To create a new NAT gateway, click **CREATE CLOUD NAT GATEWAY**.

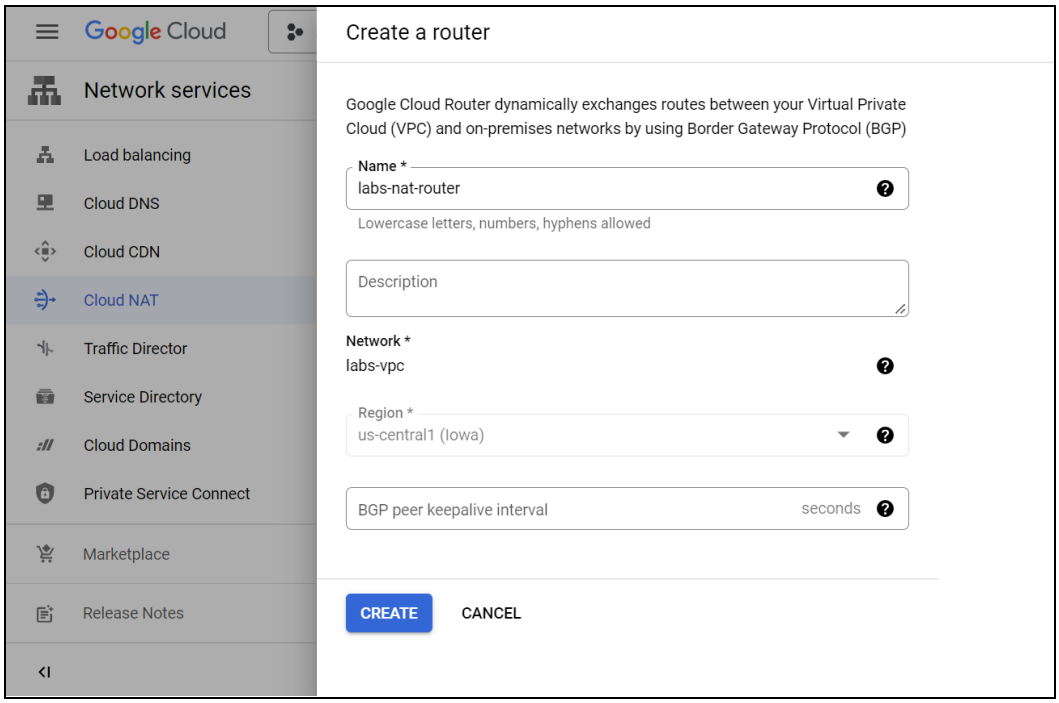
  There are many parameters you can configure when creating a new NAT gateway
- Enter NAT **Gateway name** as it is a mandatory field.
- In Select Cloud Router, select the VPC Network created previously i.e. labs-vpc
- Select the region associated with the subnet.



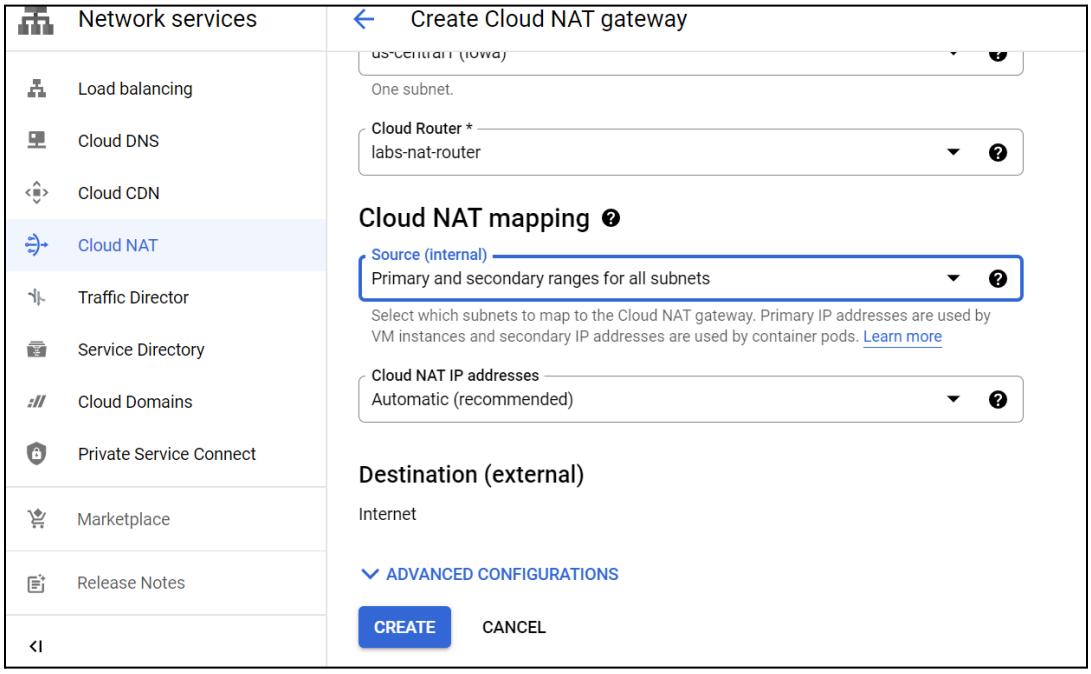
- In Cloud Router, click on create new router.



- Enter the Cloud Router name as it is a mandatory field. Rest of the fields can be left with default values.
- Click on **CREATE** to create a Cloud router.



- The router's name is populated in the Cloud Router field.
- Rest of the fields can be left with default values.



- Cloud NAT gateway named **labs-nat-gw** is created.

3. Create **IAP Firewall Rule**.

   To allow IAP to connect to your VM instances, create a firewall rule that:

   - applies to all VM instances that you want to be accessible by using IAP.
   - allows ingress traffic from the IP range **35.235.240.0/20**. This range contains all IP addresses that IAP uses for TCP forwarding.
   - allows connections to all ports that you want to be accessible by using IAP TCP forwarding, for example, port 22 for SSH and port 3389 for RDP.

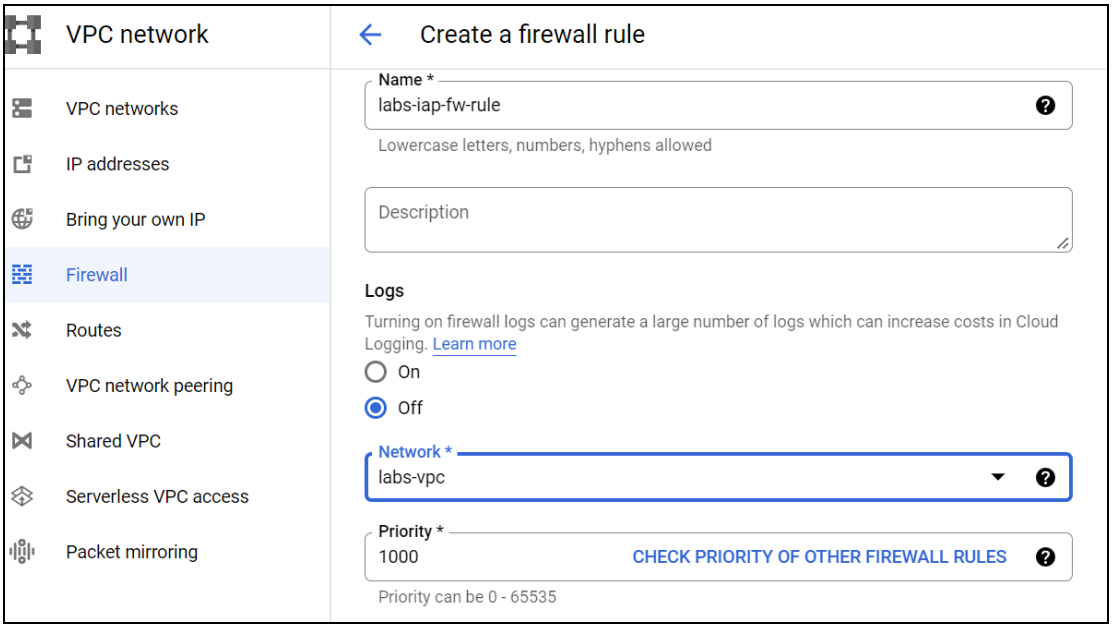
Deploying IAP Firewall using Google Cloud Console

Metadata - Firewall	**Naming Convention**
Name	labs-iap-fw-rule
Network	labs-vpc
Direction of Traffic Action on Match	Ingress Allow
Target	All Instance in Network
Source Filter	IPv4 Ranges
Source IPv4 Ranges	35.235.240.0/20
Protocol & Ports TCP	Specified Protocols & Ports 22, 3389

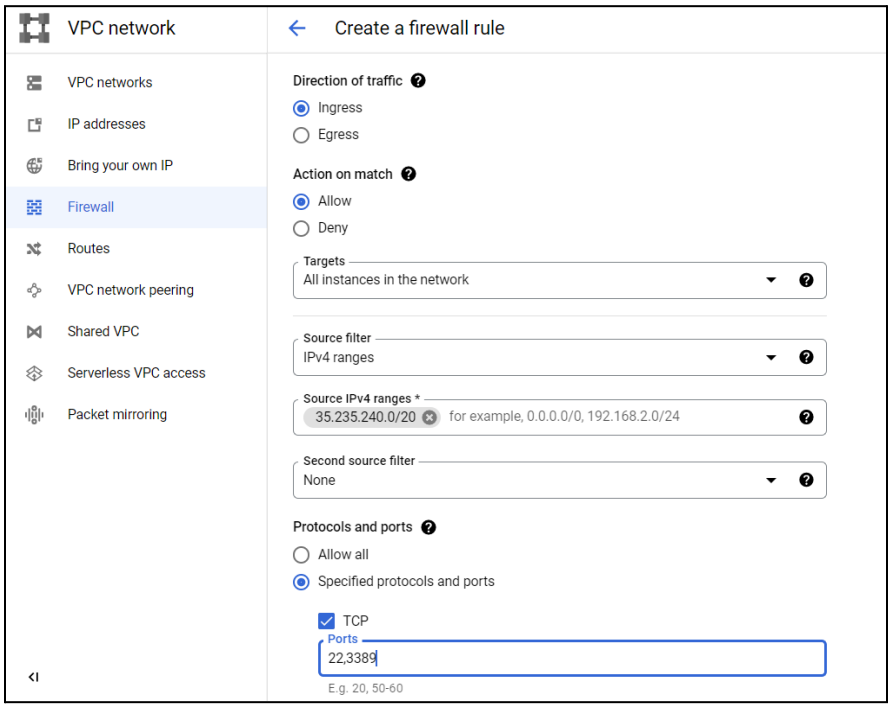
- In the Cloud Console, on the **Navigation menu** (☰), click **VPC network** > **Firewall.**
- To create a new IAP Firewall rule, click **CREATE FIREWALL RULE.**

  There are many parameters you can configure when creating a new firewall rule

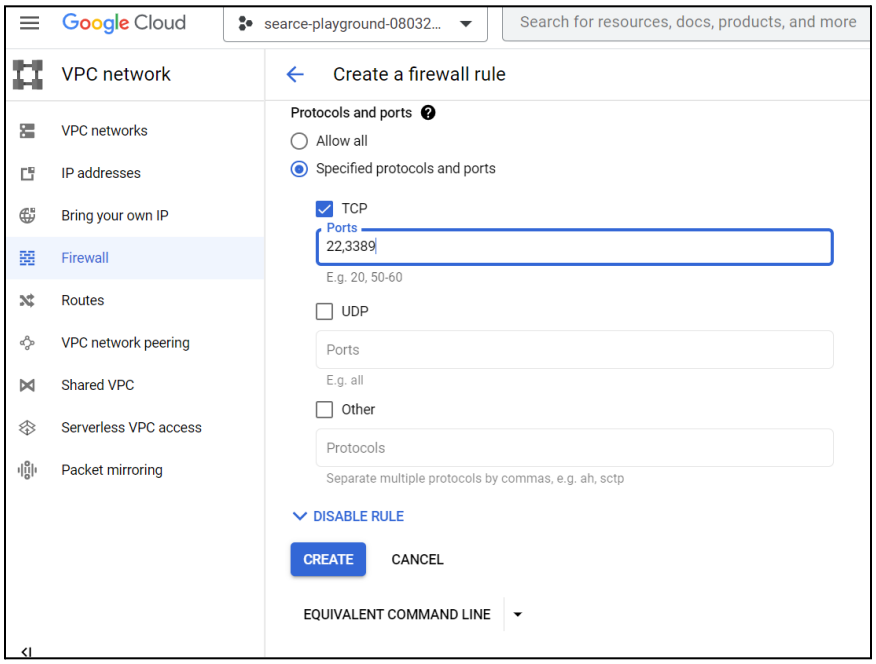
- Enter firewall **Name** as it is a mandatory field.
- Description can be left blank as it is an optional field.
- Select the VPC **Network** created previously i.e. ***labs-vpc*** to attach the firewall to that network.



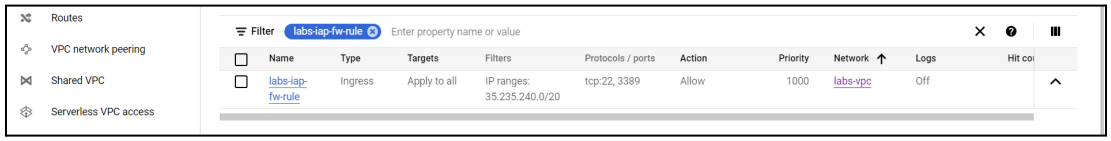
- Select Direction of traffic as **Ingress** by selecting the checkbox.
- Select Action on match as **Allow** by selecting the checkbox.
- In the Target field, select **All instances in the network**.
- In the Source filter, select IPv4 ranges.
- Enter the Source IPv4 range **35.235.240.0/20**.
- In Protocols and Ports, enable **TCP** protocol with port 22 for SSH and 3389 for RDP access.



- Click on **CREATE** to create an IAP firewall.



- The IAP firewall named **labs-iap-fw-rule** is created.



4. Create Service Account for Compute Engine

A service account is identified by its email address, which is unique to the account. Before creating the service account following the below steps

- Enable the IAM API - [IAM API]

Required roles for your IAM account.

To get the permissions that you need to manage service accounts, grant the following IAM roles on the project:

To view and create service accounts:

Create Service Accounts (roles/iam.serviceAccountCreator)

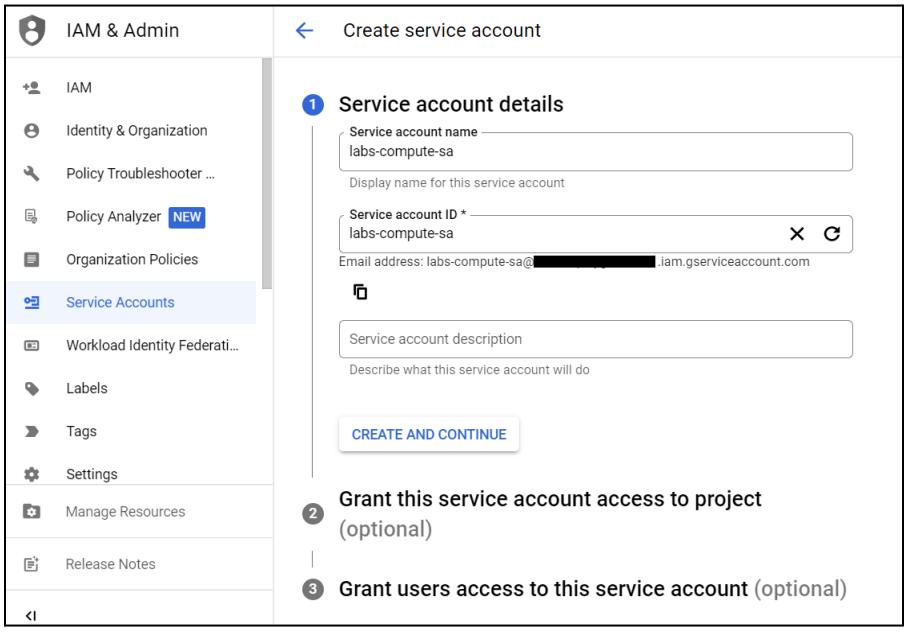
Deploying Custom Service Account using Google Cloud Console

Metadata - Service Account	**Naming Convention**
Service Account Details Service Account Name	labs-compute-sa
Add Principals New Principals	labs-compute-sa
Assign roles	Storage Admin

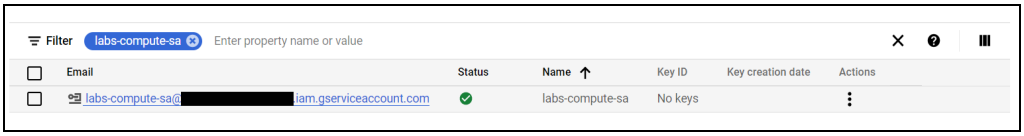
- In the Cloud Console, on the **Navigation menu** (☰), click **IAM & Admin** > **Service Account**.
- To create a new custom service account, click **CREATE SERVICE ACCOUNT**.

  There are many parameters you can configure when creating a new firewall rule

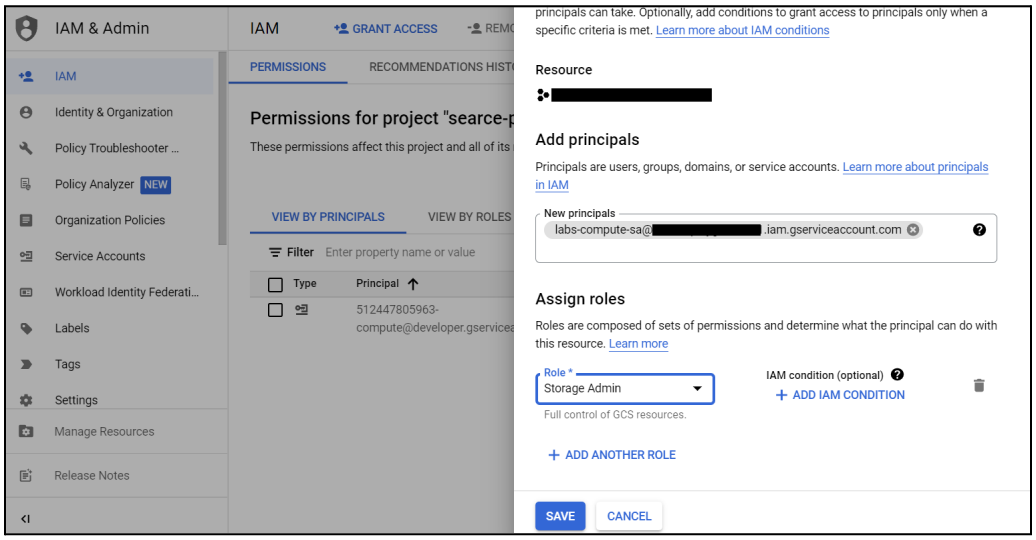
- Enter **Service Account Name** as it is a mandatory field.
- Service account ID is auto populated with service account name.
- Description can be kept blank as it is an optional field.
- Click **CREATE AND CONTINUE** to create the custom service account.

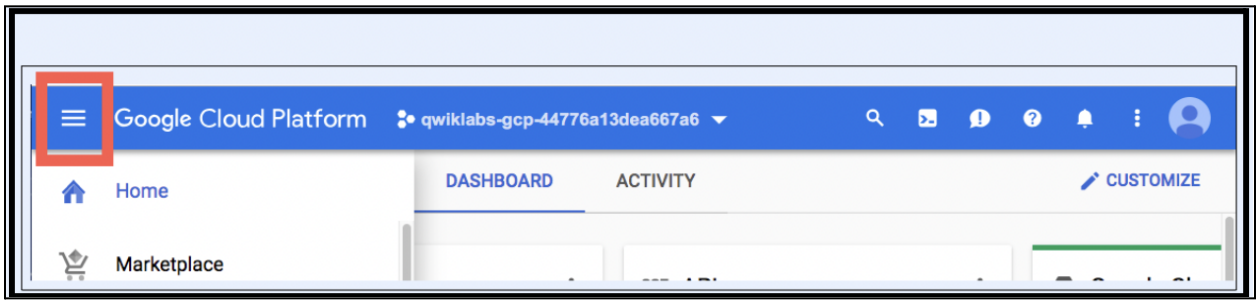


- Other fields are optional and can be skipped. The custom service account name **labs-compute-sa@project-id.iam.gserviceaccount.com** is created.



- Navigate to IAM, click **Grant ACCESS.**
- Enter the created service account name in **New Principals.**
- In Role, Select **Storage Admin** from the dropdown.

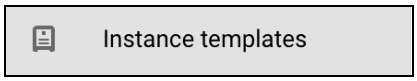




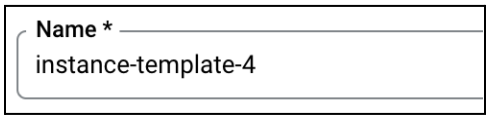
Note: You can view the menu with a list of **Google Cloud Products** and **Services** by clicking the **Navigation menu** at the top-left.

Create Instance templates for Manage instance groups.

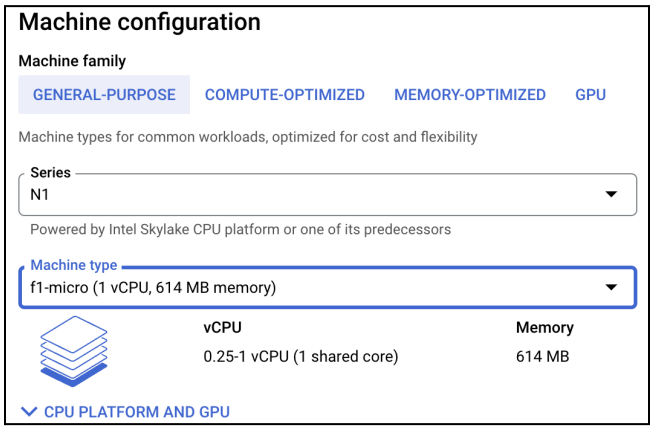
- Follow the given step to create instance template :
  - Click Compute Engine in the left navigation panel.
  - Click on instance templates.



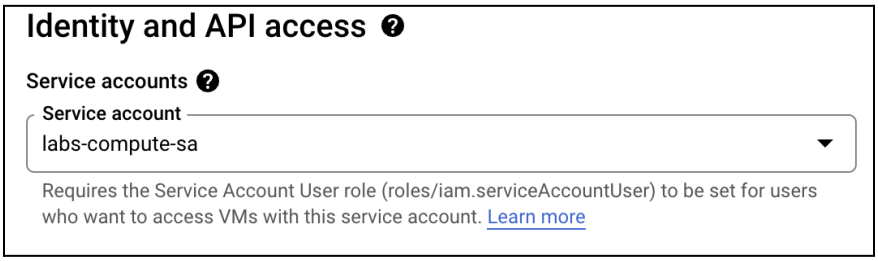
  - Click on CREATE INSTANCE TEMPLATES.
  - Enter the details > Name : <Instance template name>



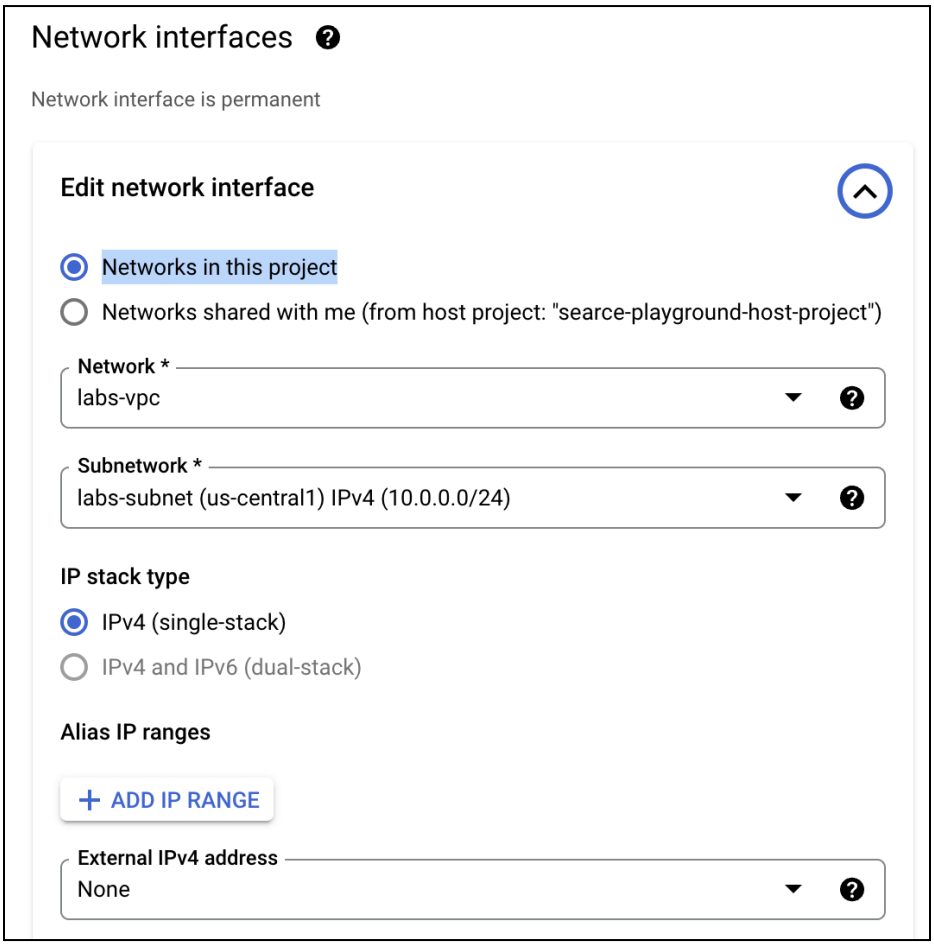
  - Machine Type >
    series : n1
    machine type : f1-micro



  - Identity and API access > Service accounts : <Add service account created in prerequisites>



- Advanced options > Network interfaces > Select : Networks in this project
  network : <select vpc name created in prerequisites>
  Subnetwork : <select subnetwork created in prerequisites>
  click on Done. External IPv4 address should be None.



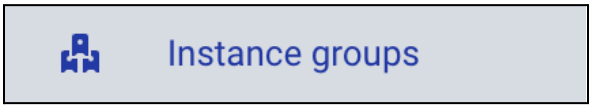
- add startup script in management > startup script :

```
#!/bin/bash
apt update
apt -y install apache2
echo "Hello world from $(hostname) $(hostname -I)" >
/var/www/html/index.html
```

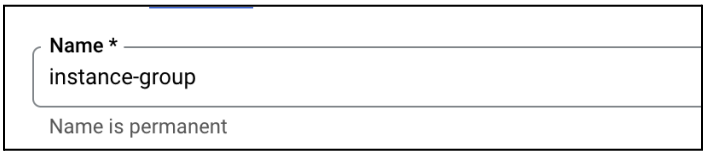
- Click on CREATE.

Configure Managed instance groups

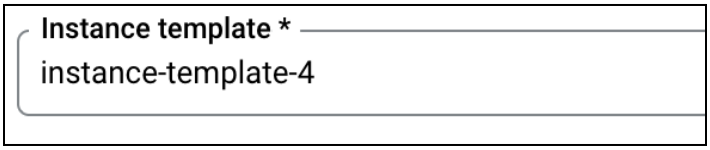
- To Configure manage instance groups these are the following steps:
    - Click Compute Engine in the left navigation panel.
    - Click on Instance groups.



    - Click on CREATE INSTANCE GROUP.
    - Select > New managed instance group (stateless)

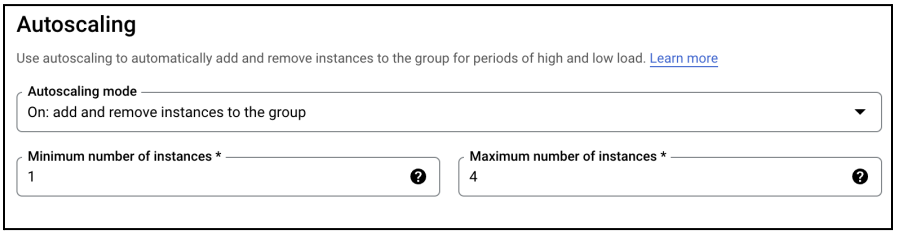


    - ENTER the following details > Name : <Enter the instance groups name>



    - Instance templates : <Select the instance template create in above step>
      Location > Select : Single Zone Region: <Select your region>
      Zone: <Select you zone>

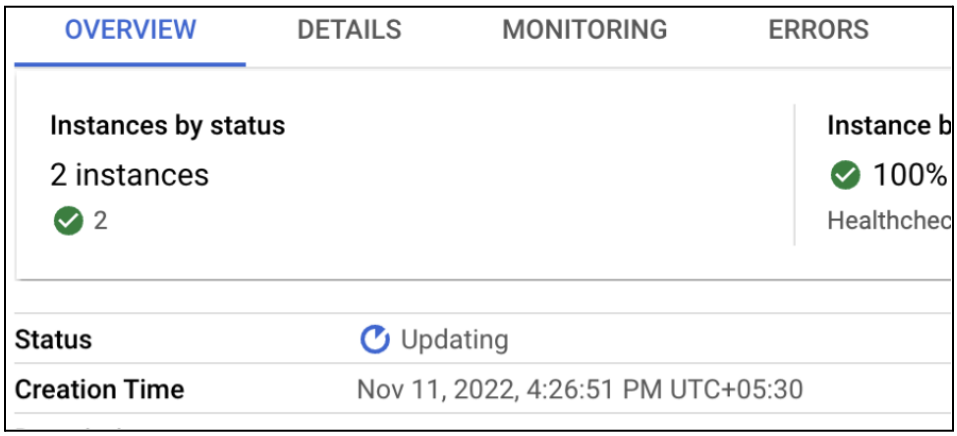
    - Autoscaling >
      Auto Scaling mode > select : On: add and remove instances to the group
      Minimum number of instances : <Enter 1 >
      Maximum number of instances : <Enter 4 >

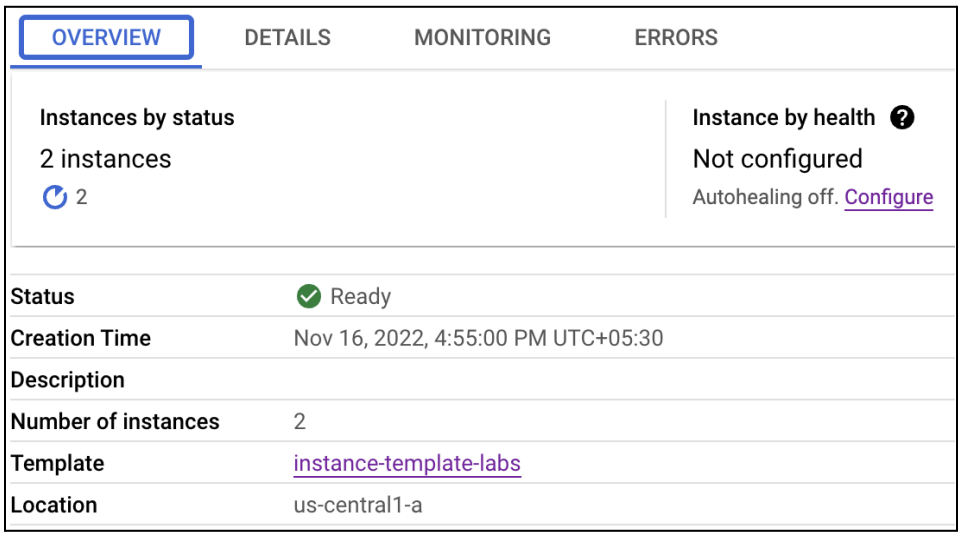


- Click on CREATE.

Result : MIG Auto scaling

Use following Command : To scale the CPU : ab -n 10000 -c 100 'http://host_ip/'
Make sure your IP Address is added in the firewall which we create in the above step.





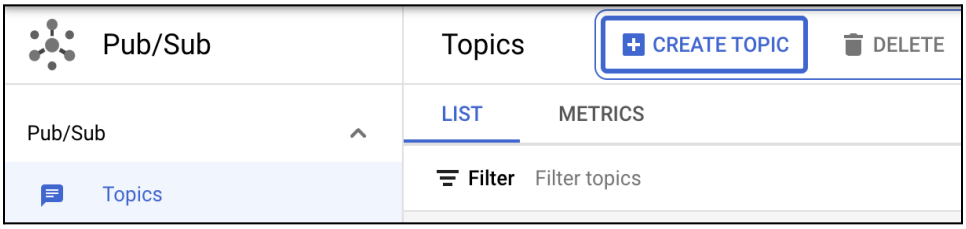
Configure the PUB/SUB

To configure PUB/SUB below are the following step :
To set up Auto Scaling via Pub/Sub. First of all, create a Topic and add that topic into Subscriptions.

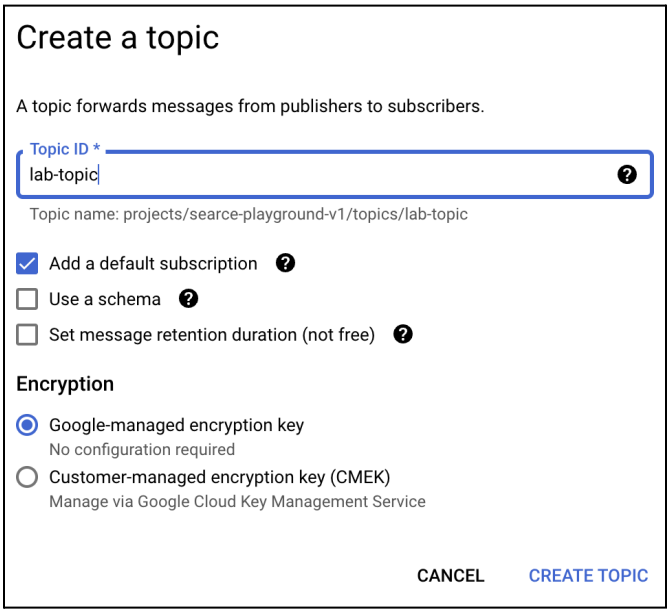
- Create TOPIC
  - In google console search pub/sub in the search tab.

  - Click on Pub/Sub.

  - Click on Pub/sub > Topics > CREATE TOPIC

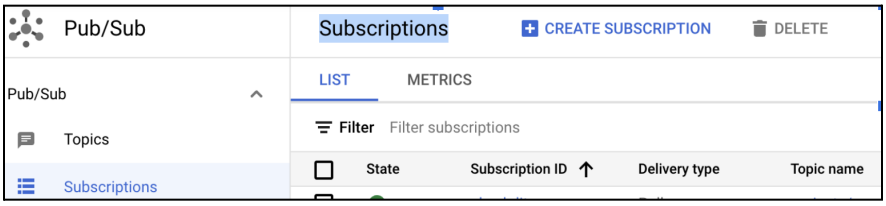


  - Enter the Topic name and leave the rest of the fields default.

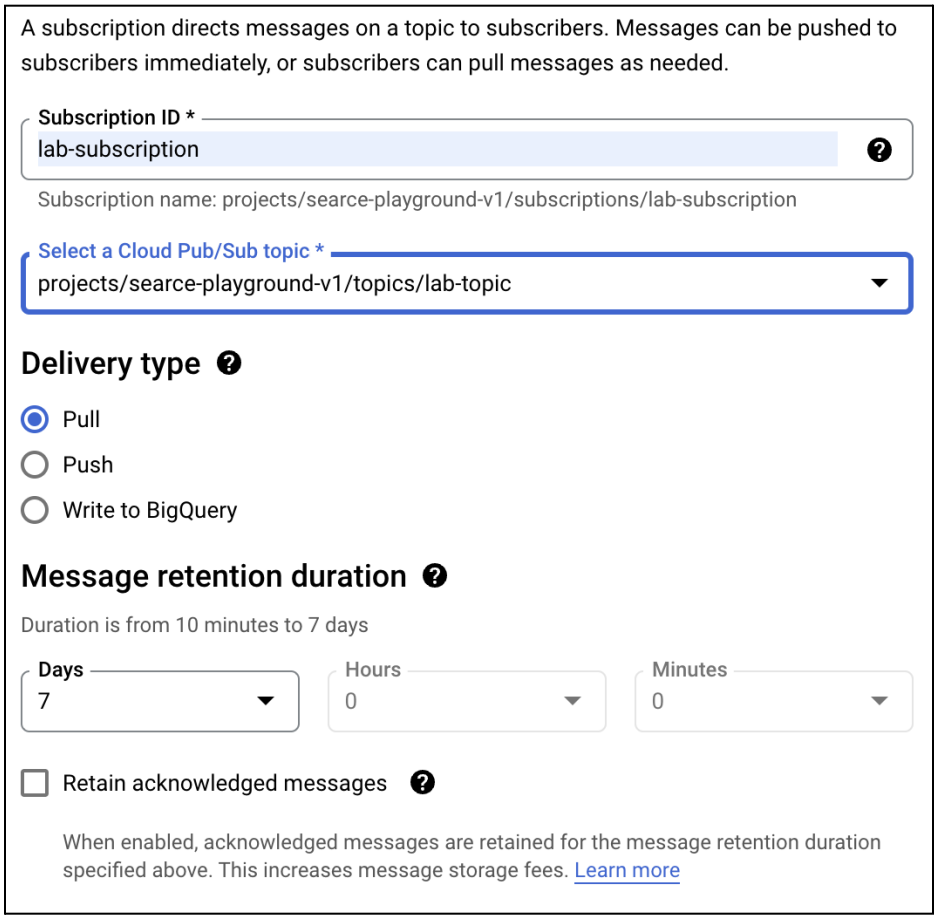


- Click on CREATE TOPIC.

- Create Subscriptions.
  - Click on the Subscriptions > CREATE SUBSCRIPTION.



  - Enter the following details >
    Subscriptions : <Enter the name of Subscription>
    Select a Cloud Pub/Sub topic : <Select the topic created in above step>



  - Leave the rest of the field default and Click on the CREATE.

To configure auto scaling based on unacknowledged messages in a Pub/Sub subscription, use the subscription/num_undelivered_messages metric provided by pubsub and filter by the subscription ID.

The subscription/num_undelivered_messages metric exports the total number of messages in the subscription, including messages that are currently being processed but that are not yet acknowledged.

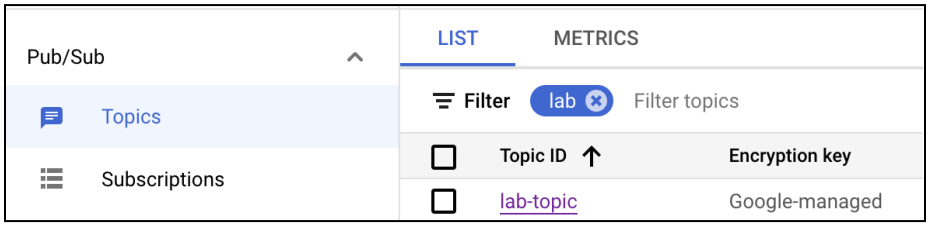
Run the below Command on Cloud shell editor.

```
gcloud compute instance-groups managed set-autoscaling MIG_NAME \
--max-num-replicas=MAX_INSTANCES \
--min-num-replicas=MIN_INSTANCES \
--update-stackdriver-metric=pubsub.googleapis.com/subscription/num_undelivered_messages \
--stackdriver-metric-filter="resource.type=\"pubsub_subscription\" AND resource.labels.subscription_id=\"SUBSCRIPTION_ID\"" \
--stackdriver-metric-single-instance-assignment=NUMBER_OF_MESSAGES_TO_ASSIGN_TO_EACH_VM
```

```
Example : gcloud compute instance-groups managed set-autoscaling instance-group-mig \
--max-num-replicas=4 \
--min-num-replicas=1 \
--update-stackdriver-metric=pubsub.googleapis.com/subscription/num_undelivered_messages \
--stackdriver-metric-filter="resource.type=\"pubsub_subscription\" AND resource.labels.subscription_id=\"lab-subscriptionl\"" \
--stackdriver-metric-single-instance-assignment=2
```

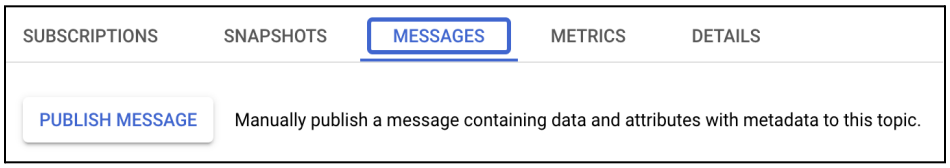
Now Setup is complete for pub/sub triggers to simulate load and autoscaling of nodes.

- Check the above setup using following step :
  - In google console search pub/sub in the search tab.
  - Click on Pub/sub > Topic > your topic name

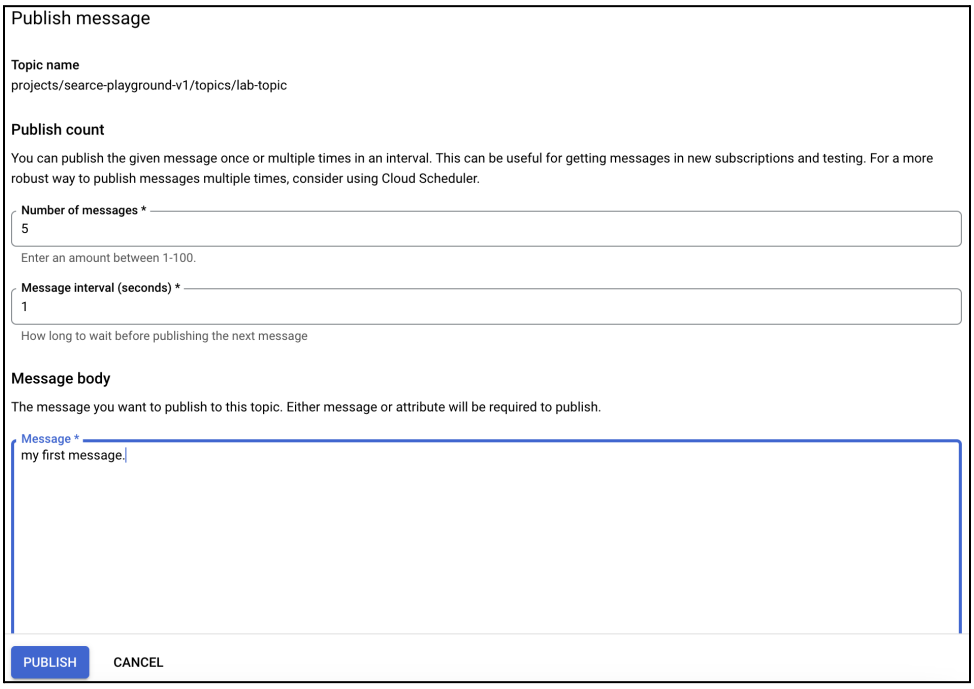


  - Click on Your topic name.

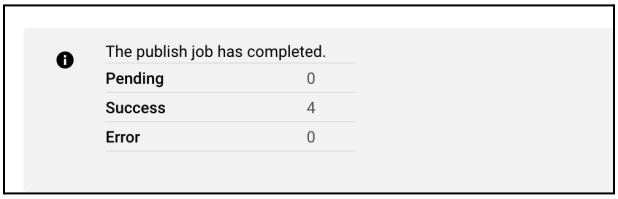
- Click on MESSAGE > PUBLISH MESSAGE.



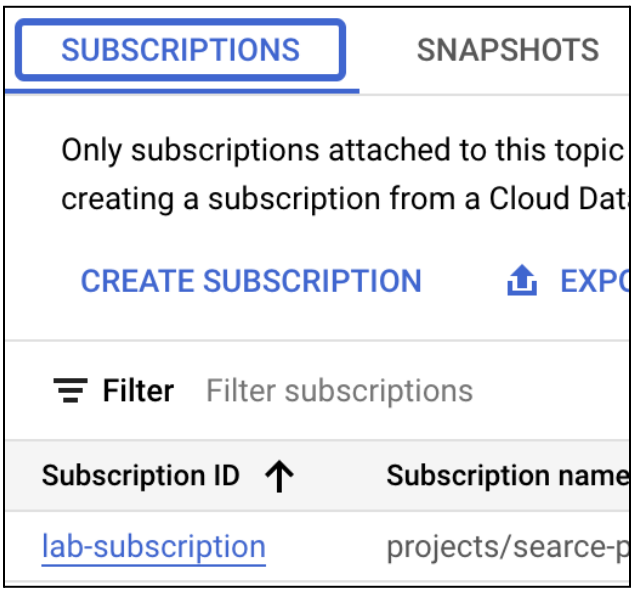
- Enter the following details :
  Number of messages : <Enter you desire number to publish your message>
  Message interval (Seconds) : <Enter you desire number So, It publish your message in given interval>
  Message body > Message : <Enter your contain that you what to publish>



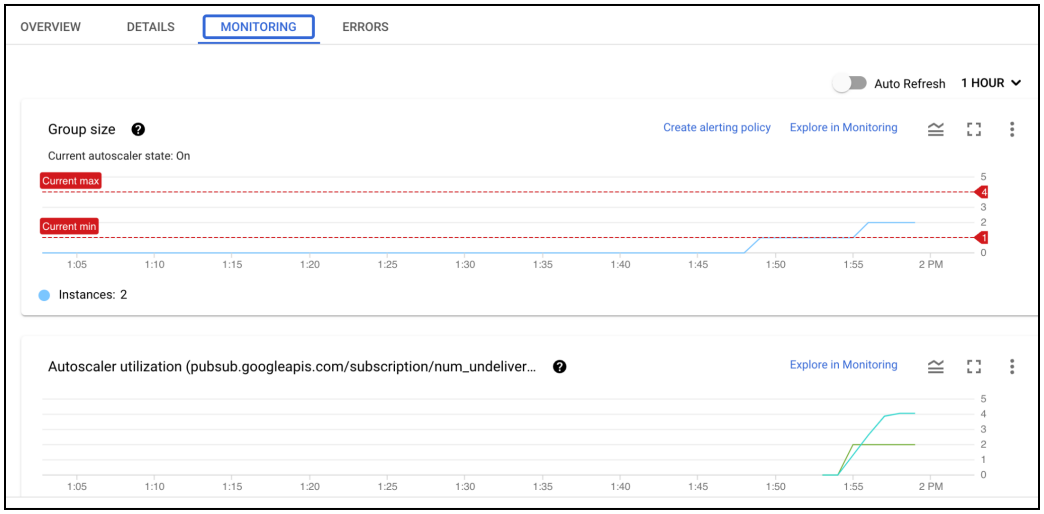
- Click on PUBLISH to publish your message.

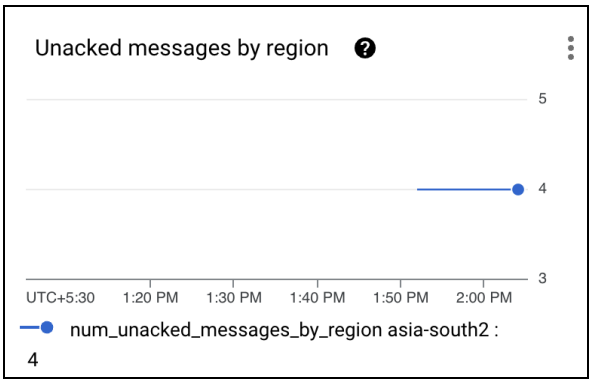


- Click on Subscription <your Subscription ID>

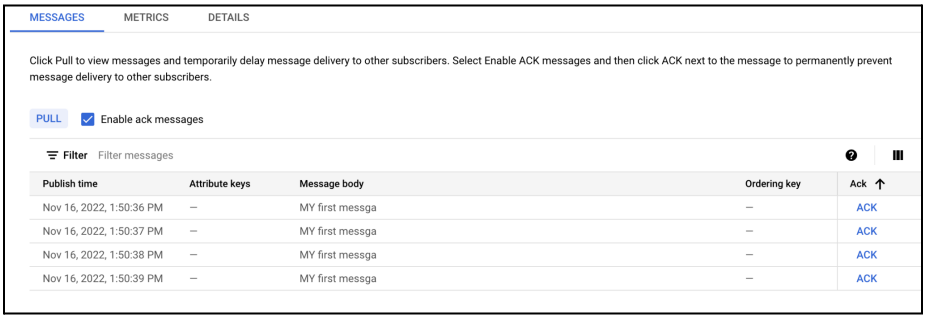


Now check the METRICS of your subscription, <Unacked messages by region>, here It shows Unacked message Four. Now check your Manage instance groups (MIG) METRICS Via click on MONITORING. In MIG graphs It shows the scaled instance and Unacked message with the threshold value line.





- To scale down the instance just click on ACK.
    - Click Pull to view messages and temporarily delay message delivery to other subscribers. Select Enable ACK messages and then click ACK next to the message to permanently prevent message delivery to other subscribers. Note : ACK will show as a deadline exceed in some second so complete the ACK step ASAP



RESULTS : It takes 5 to 10 mint for showing the Scale down result. It will look like the images below.