



TSMC Career Hack 2025 User Guide (ICSD)

Date: 2025-2-4

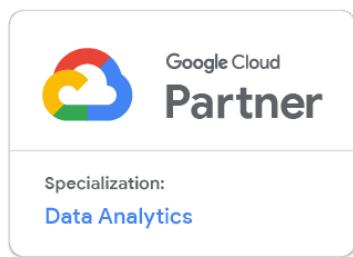
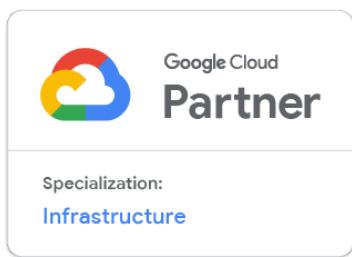


Table of contents

Table of contents.....	2
1. User Accounts.....	3
2. 使用時數限制.....	3
3. Login GCP Console.....	3
4. Select Project.....	3
5. Virtual Machine.....	4
○ Cloud Shell.....	4
○ 虛擬主機 (Google Compute Engine, GCE).....	5
6. Service Account.....	7
7. Cloud Storage.....	10
8. Gemini API.....	11
○ Vertex AI Studio UI 介面.....	11
○ curl.....	12
○ Python SDK.....	15
9. Vertex AI Embedding API model.....	16
○ curl.....	16
10. Speech-to-Text V2 API.....	18
● Curl.....	18
○ Python SDK.....	20
11. Artifact Registry.....	21
12. Cloud Run.....	22
13. Cloud SQL.....	25
14. GKE.....	26
15. Monitoring.....	28
16. Alerting.....	30

1. User Accounts

每位學生有一個 account，由題目組提供。

範例：AAID 第一組學生

AAID User 01 TSMC aaid-user01.tsmc@hackathon.cloudmile.vip
 AAID User 02 TSMC aaid-user02.tsmc@hackathon.cloudmile.vip
 AAID User 03 TSMC aaid-user03.tsmc@hackathon.cloudmile.vip
 AAID User 04 TSMC aaid-user04.tsmc@hackathon.cloudmile.vip
 以此類推

2. 使用時數限制

- 2/7 - 2/15 每項資源共可使用 60 小時
- 請注意 dashboard [CareerHack Resource Timing Dashboard \(Student View\)](#)

機器不使用時，請記得關機。以下所列資源每組上限使用 60 小時。

AAID

project_id	GCE	Workbench-0	Workbench-1	Workbench-2
aaid-test	0:00:00	0:16:35	5:41:00	0:00:00
aaid-1	0:00:00	0:00:00	0:00:00	0:00:00
aaid-2	0:00:00	0:00:00	0:00:00	0:00:00
aaid-3	0:00:00	0:00:00	0:00:00	0:00:00
aaid-4	0:00:00	0:00:00	0:00:00	0:00:00
aaid-5	0:00:00	0:00:00	0:00:00	0:00:00
aaid-6	0:00:00	0:00:00	0:00:00	0:00:00

ICSD

project_id	GCE	SQL-vector	GKE
icsd-test	0:00:00	0:31:12	2:01:57
icsd-1	0:00:00	0:00:00	0:00:00
icsd-2	0:00:00	0:00:00	0:00:00
icsd-3	0:00:00	0:00:00	0:00:00
icsd-4	0:00:00	0:00:00	0:00:00
icsd-5	0:00:00	0:00:00	0:00:00
icsd-6	0:00:00	0:00:00	0:00:00

BSID

project_id	GCE	SQL-relational	SQL-vector	Vector Search
bsid-test	0:00:00	0:00:00	0:00:00	6:00:16
bsid-1	0:00:00	0:00:00	0:00:00	0:00:00
bsid-2	0:00:00	0:00:00	0:00:00	0:00:00
bsid-3	0:00:00	0:00:00	0:00:00	0:00:00
bsid-4	0:00:00	0:00:00	0:00:00	0:00:00
bsid-5	0:00:00	0:00:00	0:00:00	0:00:00
bsid-6	0:00:00	0:00:00	0:00:00	0:00:00

TSID

project_id	GCE	GKE
tsid-test	0:01:15	0:30:51
tsid-1	0:00:00	0:00:00
tsid-2	0:00:00	0:00:00
tsid-3	0:00:00	0:00:00
tsid-4	0:00:00	0:00:00
tsid-5	0:00:00	0:00:00
tsid-6	0:00:00	0:00:00

3. Login GCP Console

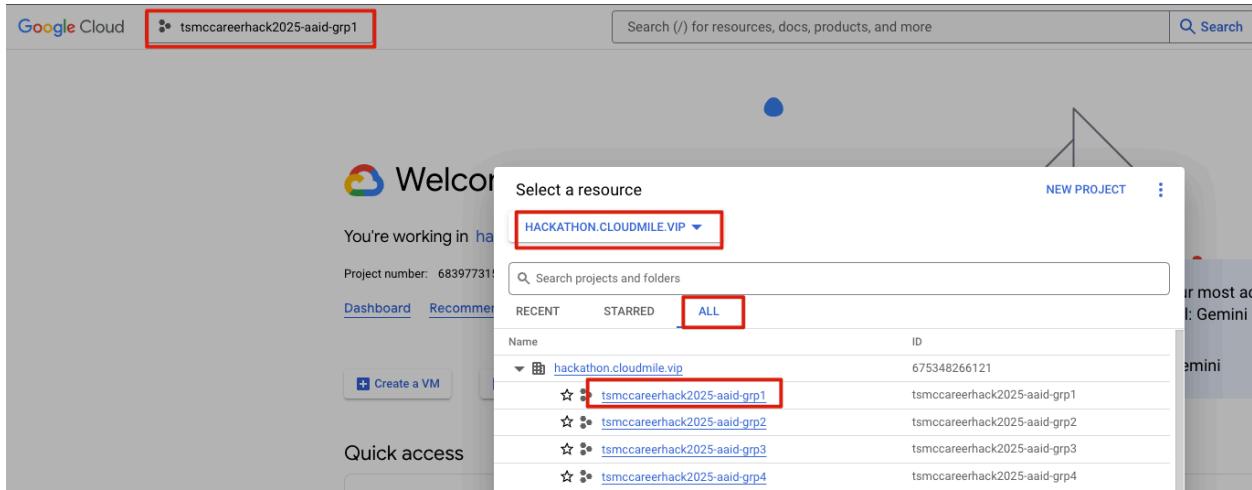
<https://console.cloud.google.com>

(Need to reset password at the first login)

4. Select Project

Project ID: tsmccareerhack2025-icsd-grp#

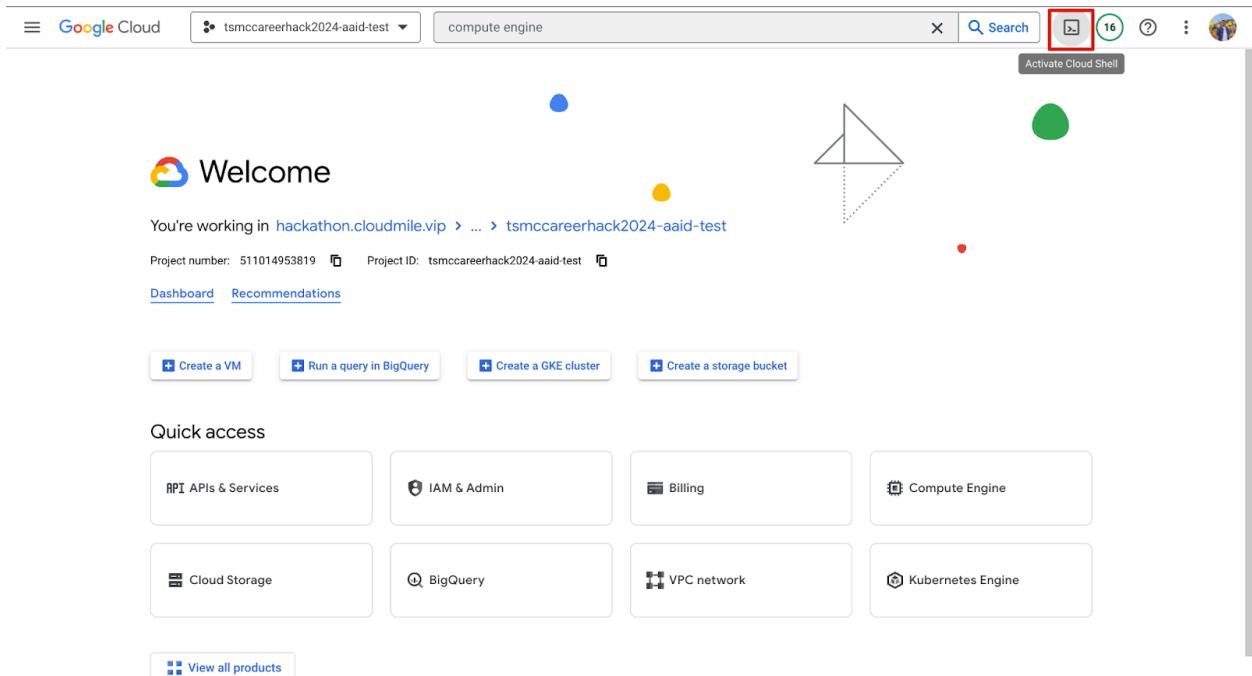
(下圖僅供參考，請選擇自己實際所屬組別的 folder 和 project。)



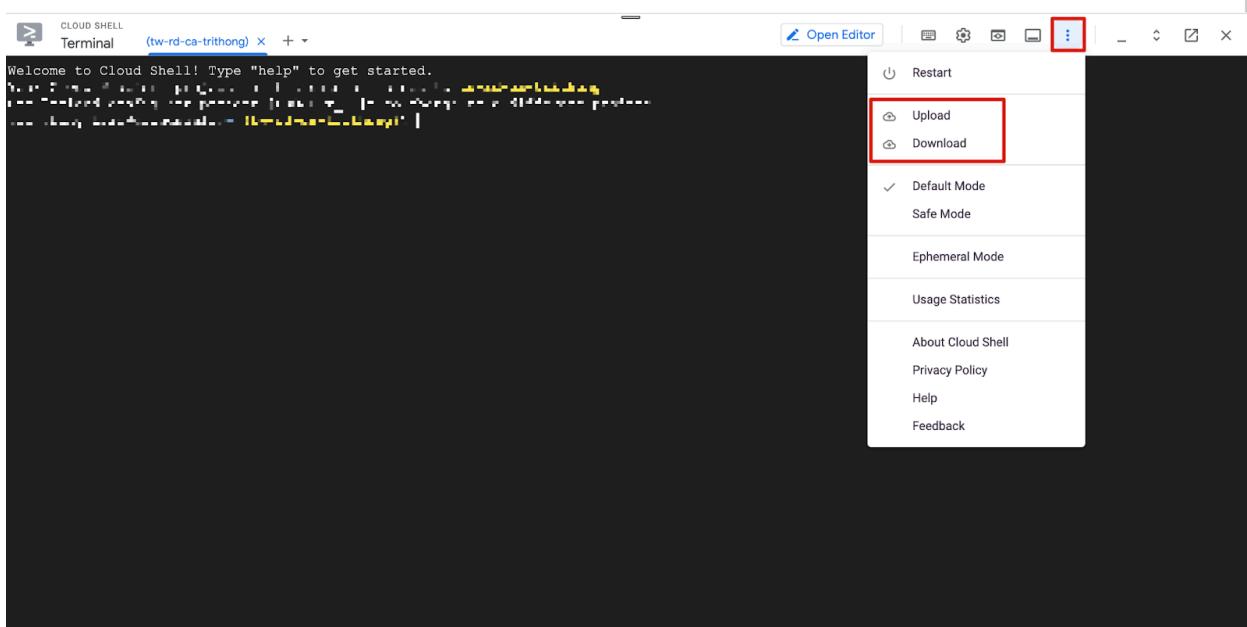
5. Virtual Machine

- Cloud Shell

Cloud Shell 供使用者在 GCP Console 上快速簡單使用 Linux 和 Google Cloud SDK 指令操作。
● 開啟 Cloud Shell :



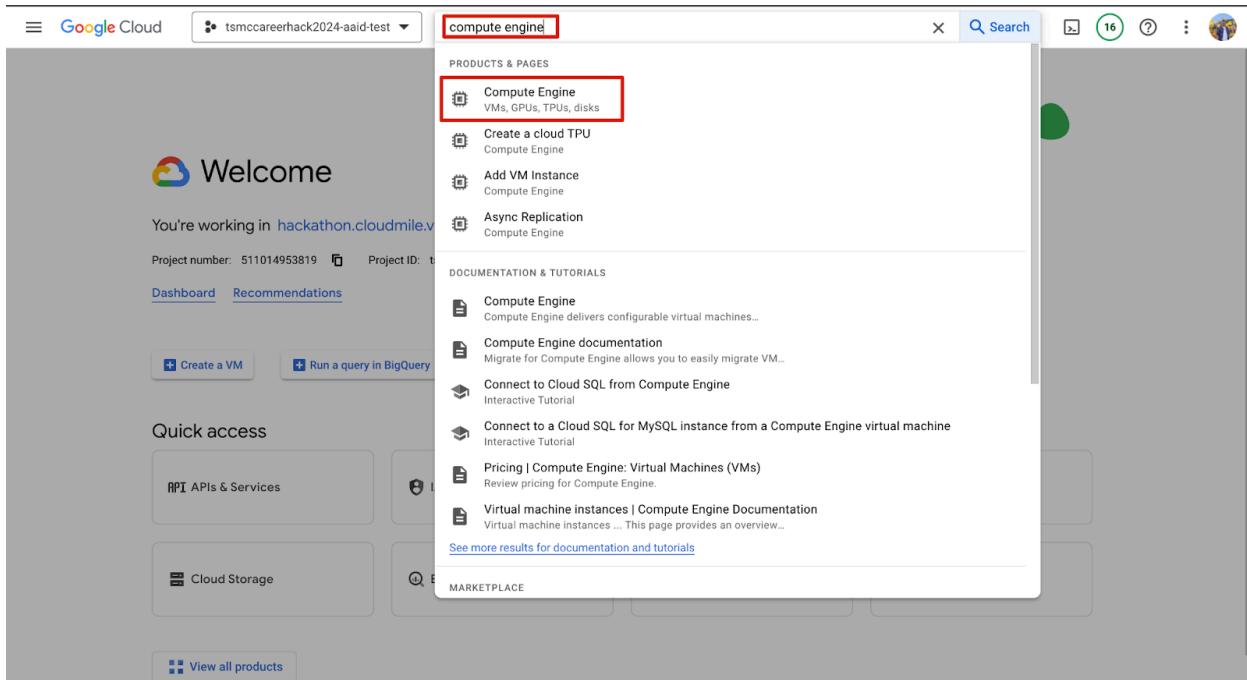
- 可透過右上角功能選單從虛擬主機下載檔案或上傳本地檔案到虛擬主機：



- 虛擬主機 (Google Compute Engine, GCE)

虛擬主機供參賽組別公用操作環境，可使用 Google Cloud SDK、可部署應用程式。使用虛擬主機的操作步驟如下：

- 進入 Compute Engine 服務頁面：



- 勾選擬器，點擊 Start 開機、Stop 關機（請記得不使用時要關機！）：

VM instances

INSTANCES OBSERVABILITY INSTANCE SCHEDULES

X 1 instance selected

Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input checked="" type="checkbox"/>	gce-instance	us-central1-a			10.0.0.5 (nic0)		SSH
<input type="checkbox"/>	workbench-instance-0	us-central1-a			10.0.0.6 (nic0)	34.71.173.107 (nic0)	SSH

- 開機後點擊右端“SSH”按鈕進行 SSH 連線：

VM instances

INSTANCES OBSERVABILITY INSTANCE SCHEDULES

VM instances

Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	gce-instance	us-central1-a			10.0.0.5 (nic0)		SSH
<input type="checkbox"/>	workbench-instance-0	us-central1-a			10.0.0.6 (nic0)	34.71.173.107 (nic0)	SSH

- 可透過右上角功能選單從虛擬主機下載檔案或上傳本地檔案到虛擬主機：

SSH-in-browser

Linux bastion 5.10.0-26-cloud-amd64 #1 SMP Debian 5.10.197-1 (2023-09-29) 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.

[root@bastion ~]#

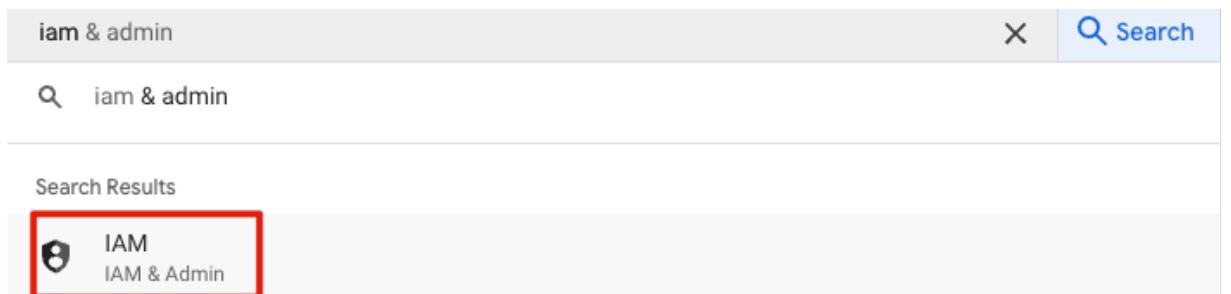
更多連線方式，請參考：<https://cloud.google.com/compute/docs/connect/ssh-using-ia>

6. Service Account

若有使用 service account key 的需求，請參考以下步驟

請勿將 key 上傳至公開 repository

進入 IAM 頁面



iam & admin

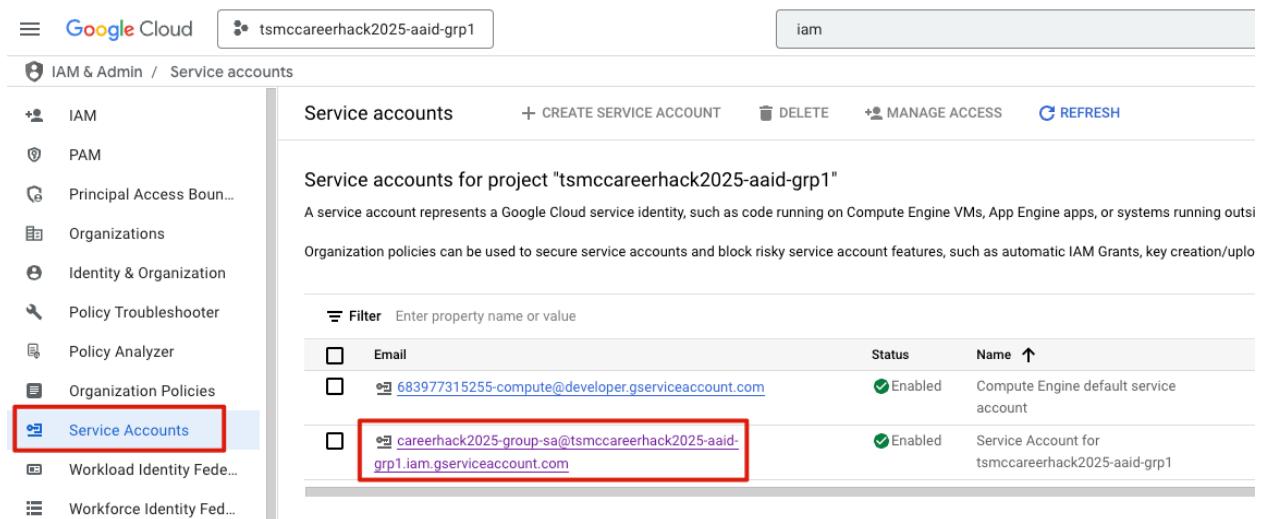
Search Results

IAM

IAM & Admin

點擊 Service Accounts，選擇小組的 service account:

careerhack2025-group-sa@tsmccareerhack2025-icsd-grp#.iam.gserviceaccount.com



Google Cloud tsmccareerhack2025-aaid-grp1 iam

IAM & Admin / Service accounts

Service accounts

CREATE SERVICE ACCOUNT DELETE MANAGE ACCESS REFRESH

Service accounts for project "tsmccareerhack2025-aaid-grp1"

A service account represents a Google Cloud service identity, such as code running on Compute Engine VMs, App Engine apps, or systems running outside Google Cloud.

Organization policies can be used to secure service accounts and block risky service account features, such as automatic IAM Grants, key creation/upload, and more.

Filter Enter property name or value

Email	Status	Name ↑
683977315255-compute@developer.gserviceaccount.com	Enabled	Compute Engine default service account
careerhack2025-group-sa@tsmccareerhack2025-aaid-grp1.iam.gserviceaccount.com	Enabled	Service Account for tsmccareerhack2025-aaid-grp1

點擊 KEYS，按下 ADD KEY → Create new key 以下載 key 的 json 檔。

請勿將 key 上傳至公開 repository

The screenshot shows the Google Cloud IAM & Admin / Service accounts / Service account: 117025319245551209354 / Keys page. The 'KEYS' tab is selected. A red box highlights the 'Create new key' button. The page displays a warning about service account keys being a security risk if compromised, and a note that Google automatically disables service account keys detected in public repositories.

請勿將 key 上傳至公開 repository

Create private key for "Service Account for tsmccareerhack2025-aaid-grp1"

Downloads a file that contains the private key. Store the file securely because this key can't be recovered if lost.

Key type

JSON

Recommended

P12

For backward compatibility with code using the P12 format

CANCEL CREATE

7. Cloud Storage

Bucket name:

- 競賽題目的資源下載：`careerhack2025-icsd-resource-bucket`
 - 此 bucket 位於 `tsmccareerhack2025-icsd-test` project 中，學生無法從 UI 存取，需要下指令
- 供參賽者存放資料：`tsmccareerhack2025-icsd-grp#-bucket`

UI 操作：

The screenshot shows the Google Cloud Platform interface for Cloud Storage. The top navigation bar includes the CloudMile logo, the project name `tsmccareerhack2025-aaid-grp1`, and a search bar. Below the navigation is a sidebar with options: Overview, Buckets (which is highlighted with a red box), Monitoring, and Settings. The main content area is titled 'Buckets' and features a 'CREATE' button and a 'REFRESH' button. A 'Filter' section allows filtering by Name (升序). A table lists one bucket: `tsmccareerhack2025-aaid-grp1-bucket`, created on Jan 20, 2025, 4:38:43 PM, located in Region.

指令操作：

- 下列操作指令可在上述 Cloud Shell 和 GCE VM 環境執行以存取 Google Cloud Storage 的 bucket。
- 基本操作指令：
 - `gcloud storage ls`
 - `gcloud storage cp`
 - `gcloud storage mv`
 - `gcloud storage rm`
 - ...
- 完整指令操作手冊：<https://cloud.google.com/sdk/gcloud/reference/storage>
- 常用指令範例：
 - 列出目錄中的物件：

```
Unset
```

```
gcloud storage ls gs://BUCKET_NAME
```

- 下載檔案到當前本機目錄：

```
Unset
```

```
gcloud storage cp gs://BUCKET_NAME/OBJECT_NAME .
```

- 下載資料夾到當前本機目錄：

Unset

```
gcloud storage cp -r gs://BUCKET_NAME/FOLDER_NAME .
```

- 上傳檔案：

Unset

```
gcloud storage cp LOCAL_FILE_NAME gs://BUCKET_NAME
```

- 上傳資料夾：

Unset

```
gcloud storage cp -r LOCAL_DIR_NAME gs://BUCKET_NAME
```

8. Gemini API

請留意各 model 的 rate limit

Token per minute

Base model	Tokens per minute
base_model: gemini-1.5-flash (version 001)	4M (4,000,000)
base_model: gemini-1.5-pro (version 001)	4M (4,000,000)

Request per minute

Base model	Requests per minute
base_model : gemini-1.5-flash	200
base_model : gemini-1.5-pro	60

請參考：<https://cloud.google.com/vertex-ai/generative-ai/docs/quotas>

○ Vertex AI Studio UI 介面

Gemini 1.5 Pro for Text

The screenshot shows the Vertex AI Studio UI for Gemini 1.5 Pro for Text. The left sidebar includes sections for Tools (Dashboard, Model Garden, Pipelines), Notebooks (Colab Enterprise, Workbench), Vertex AI Studio (Overview, Freeform), and Build with Gen AI (Extensions). The main area displays a 'System instructions' box and a 'Prompt' box where 'Write a prompt, or create one with ⚡ Help me write' is entered. The right panel contains configuration options for the Model (set to 'gemini-1.5-pro-002'), Region (us-central1 (Iowa)), Temperature (0 to 2), Output token limit (1 to 8192), Grounding (Source: Google Search), and Safety Filter Settings.

The screenshot shows the Vertex AI Studio UI for code-gecko@002. The sidebar and main interface are similar to the Gemini 1.5 Pro version, but the right panel shows different configuration options. The Model is set to 'code-gecko@002', Region to 'us-central1 (Iowa)', Temperature to 1 (0.2), Output token limit to 64, and Safety filter threshold to 'Block few'. Other settings like Grounding and Safety Filter Settings are also present.

- curl

- 若使用 Cloud Shell 或所提供的虛擬主機操作，請跳過此步驟。若欲從使用者本地機器操作，請先下載並安裝 [Google Cloud SDK](#)。安裝完成後需要執行下列指令登入帳號取得授權：

```
Unset
```

```
gcloud auth application-default login
```

curl 指令範例：

- 範例 request.json：

```
Unset
```

```
{  
  "contents": [ {  
    "role": "user",  
    "parts": [ {  
      "text": "Tell me how to win a hackathon"  
    }]  
  }]  
}
```

- 範例 input：

```
Unset
```

```
export PROJECT_ID=YOUR_PROJECT_ID  
export MODEL_ID="gemini-1.5-pro-002"
```

```
curl \  
  -X POST \  
  -H "Authorization: Bearer $(gcloud auth application-default  
print-access-token)" \  
  -H "Content-Type: application/json" \  
  -d @request.json \  
  https://us-central1-aiplatform.googleapis.com/v1/projects/${PROJECT_ID}/locatio  
ns/us-central1/publishers/google/models/${MODEL_ID}:generateContent
```

- 範例輸出（注意：usageMetadata 回傳了該 prompt 的 input 和 output token count）：

```
Unset
{
  "candidates": [
    {
      "content": {
        "role": "model",
        "parts": [
          {
            "text": "Winning a hackathon isn't just about writing the most technically brilliant code, it's about creating a well-rounded project that solves a problem and presents well. Here's a breakdown of how to increase your chances:\n\n**Before the Hackathon:**\n* **Team Up (Optional but Recommended):** A diverse team with complementary skills (coding, design, business, marketing) is a huge advantage. Find teammates beforehand if possible.\n* **Brainstorm Ideas:** Don't wait until the hackathon starts. Having a few project ideas in your back pocket saves valuable time. Consider current events, trending technologies, and personal passions.\n* **Practice:** Familiarize yourself with the technologies you plan to use. The more fluent you are, the more time you can spend on building, not learning.\n* **Gather Resources:** Identify APIs, libraries, and frameworks you might need. Knowing where to find resources can save you hours of searching during the event.\n* **Pack Smart:** Charger, laptop stand, external monitor (if allowed), headphones, snacks, water bottle – the essentials for a long haul.\n\n**During the Hackathon:**\n* **Scope Realistically:** Ambition is good, but a finished, functional project is better than an unfinished masterpiece. Focus on a core feature and expand if time allows.\n* **Divide and Conquer:** If you're in a team, leverage everyone's strengths. Clear roles and responsibilities prevent overlap and maximize efficiency.\n* **Version Control (Git):** Essential for team projects. Use branches and merge requests to keep the codebase organized and avoid conflicts.\n* **Focus on the User:** Keep the end-user in mind throughout the process. A user-friendly interface and intuitive design are crucial.\n* **Seek Feedback:** Talk to mentors and other participants. External perspectives can highlight blind spots and spark new ideas.\n* **Document Everything:** Keep track of your progress, decisions, and challenges. This will be invaluable for your presentation and future development.\n* **Don't Neglect the Presentation:** A compelling presentation can make or break your project. Practice your pitch and highlight the problem, solution, and impact of your creation.\n* **Rest (Seriously):** Pulling an all-nighter might seem productive, but fatigue leads to errors and poor decision-making. Short breaks and power naps can boost your performance.\n\n**Key Ingredients for a Winning Project:**\n* **Originality/Innovation:** A fresh perspective or a novel application of existing technology.\n* **Impact/Usefulness:** Addresses a real-world problem"
      }
    }
  ]
}
```

```
or provides a valuable service.\n* **Technical Execution:** Clean,  
well-documented code that functions as intended.\n* **Design/User Experience:**  
An intuitive and aesthetically pleasing interface.\n* **Presentation:** A  
clear, concise, and engaging pitch that highlights the project's  
strengths.\n\n**After the Hackathon:**\n* **Network:** Connect with other  
participants, judges, and sponsors. Hackathons are excellent networking  
opportunities.\n* **Learn from Feedback:** Even if you don't win, gather  
feedback and use it to improve your skills for future events.\n* **Iterate and  
Improve:** Don't abandon your project after the hackathon. Continue  
developing and refining it – you might have a viable product on your  
hands.\n\n**Common Pitfalls to Avoid:**\n* **Feature Creep:** Trying to  
implement too many features and ending up with nothing finished.\n* **Ignoring  
the User:** Building something technically impressive but impractical or  
difficult to use.\n* **Poor Presentation:** Failing to communicate the value  
and impact of your project effectively.\n* **Lack of Teamwork:** Inefficient  
collaboration and communication within the team.\n\nWinning a hackathon is a  
combination of skill, preparation, and a bit of luck. By focusing on these  
tips, you'll significantly increase your chances of success and have a  
rewarding experience. Good luck!\n"  
        }  
    ]  
,  
    "finishReason": "STOP",  
    "avgLogprobs": -0.20640207265878652  
}  
,  
"usageMetadata": {  
    "promptTokenCount": 8,  
    "candidatesTokenCount": 847,  
    "totalTokenCount": 855  
},  
"modelVersion": "gemini-1.5-pro-002"  
}
```

更多範例請參考：[Generate content with the Gemini Enterprise API | Generative AI on Vertex AI | Google Cloud](#)

- Python SDK
 - Python 版本需求：Python 3.8 or higher
 - 安裝 Vertex AI Python SDK：

Mac/Linux:

```
Unset
pip install virtualenv
virtualenv <your-env>
source <your-env>/bin/activate
<your-env>/bin/pip install google-cloud-aiplatform
```

Windows:

```
Unset
pip install virtualenv
virtualenv <your-env>
<your-env>\Scripts\activate
<your-env>\Scripts\pip.exe install google-cloud-aiplatform
```

更多資訊請參考 [Vertex AI Python SDK](#).
[Python client library | Google Cloud](#)
[Install the Vertex AI client libraries | Google Cloud](#)

- 本地開發登入授權，本地開發需先在本地[安裝 gcloud CLI](#)：

```
Unset
gcloud auth application-default set-quota-project <PROJECT_ID>
gcloud auth application-default login
```

- 範例程式：

```
Python
import vertexai
from vertexai.generative_models import GenerativeModel

# TODO(developer): Update below line
PROJECT_ID = "PROJECT_ID"

REGION = "us-central1"
vertexai.init(project=PROJECT_ID, location=REGION)

model = GenerativeModel("gemini-1.5-pro-002")

response = model.generate_content()
```

```
"Tell me how to win a hackathon"  
)  
  
print(response.text)
```

9. Vertex AI Embedding API model

text-multilingual-embedding-002

[Text embeddings API | Generative AI on Vertex AI | Google Cloud](#)

[Embeddings for Text | Model Garden](#)

- curl

curl 指令範例：

- 範例 request.json :

```
Unset  
{  
  "instances": [  
    {  
      "task_type": "RETRIEVAL_DOCUMENT",  
      "title": "document title",  
      "content": "I would like embeddings for this text!"  
    }  
  ],  
  "parameters": {  
    "outputDimensionality": 256  
  }  
}
```

- 範例 input :

```
Unset
MODEL_ID="text-multilingual-embedding-002"
PROJECT_ID=PROJECT_ID

curl \
-X POST \
-H "Authorization: Bearer $(gcloud auth print-access-token)" \
-H "Content-Type: application/json" \
-d @request.json \
https://us-central1-aiplatform.googleapis.com/v1/projects/${PROJECT_ID}/locations/us-central1/publishers/google/models/${MODEL_ID}:predict
```

- 範例輸出：

```
Unset
{
  "predictions": [
    {
      "embeddings": {
        "statistics": {
          "token_count": 14,
          "truncated": false
        },
        "values": [
          0.039897624403238297,
          -0.011670297011733055,
          0.03811752051115036,
          0.024373488500714302,
          0.065585844218730927,
          0.013257214799523354,
          0.0053430162370204926,
          0.041858211159706116,
          ...
          ...
        ]
      }
    },
    ],
    "metadata": {
      "billableCharacterCount": 45
    }
  }
```

```
}
```

10. Speech-to-Text V2 API

詳細步驟請參考官方文件：[Transcribe speech to text by using the command line](#)

- 若使用 Cloud Shell 或所提供的虛擬主機操作，請跳過此步驟。若欲從使用者本地機器操作，請先下載並安裝 [Google Cloud SDK](#)。安裝完成後需要執行下列指令登入帳號取得授權：

```
Unset
```

```
gcloud auth application-default login
```

- Curl

- 修改檔案路徑 `/full/path/to/audio/file.wav`

```
Unset
```

```
echo \"{
  \"config\": {
    \"auto_decoding_config\": {},
    \"language_codes\": [\"en-US\"],
    \"model\": \"long\"
  },
  \"content\": \"$(base64 -w 0 /full/path/to/audio/file.wav | sed 's/+/-/g;
s///_g')\""
}" > /tmp/data.txt
```

- 範例 input :

```
Unset
```

```
PROJECT_ID=PROJECT_ID
```

```
curl -X POST -H "Content-Type: application/json; charset=utf-8" \
-H "Authorization: Bearer $(gcloud auth print-access-token)" \
-d @/tmp/data.txt \
```

```
https://speech.googleapis.com/v2/projects/${PROJECT_ID}/locations/global/recognizers/_:recognize
```

- 範例輸出：

```
Unset
{
  "results": [
    {
      "alternatives": [
        {
          "transcript": "how old is the Brooklyn Bridge",
          "confidence": 0.98267895
        }
      ]
    }
  ]
}
```

- Python SDK
 - 請參考前一章節安裝 Vertex AI Python SDK

```
Python
import os
```

```
from google.cloud.speech_v2 import SpeechClient
from google.cloud.speech_v2.types import cloud_speech

PROJECT_ID = os.getenv("GOOGLE_CLOUD_PROJECT")

def quickstart_v2(audio_file: str) -> cloud_speech.RecognizeResponse:
    """Transcribe an audio file.

    Args:
        audio_file (str): Path to the local audio file to be transcribed.
```

```
    Returns:  
        cloud_speech.RecognizeResponse: The response from the recognize  
        request, containing  
            the transcription results  
        """  
  
        # Reads a file as bytes  
        with open(audio_file, "rb") as f:  
            audio_content = f.read()  
  
        # Instantiates a client  
        client = SpeechClient()  
  
        config = cloud_speech.RecognitionConfig(  
            auto_decoding_config=cloud_speech.AutoDetectDecodingConfig(),  
            language_codes=["en-US"],  
            model="long",  
        )  
  
        request = cloud_speech.RecognizeRequest(  
            recognizer=f"projects/{PROJECT_ID}/locations/global/recognizers/_",  
            config=config,  
            content=audio_content,  
        )  
  
        # Transcribes the audio into text  
        response = client.recognize(request=request)  
  
        for result in response.results:  
            print(f"Transcript: {result.alternatives[0].transcript}")  
  
    return response
```

11. Artifact Registry

- 本活動為每一組準備好一個 private docker repository，可用於存放 docker images：
`tsmccareerhack2025-icsd-grp#-repository`
- 使用說明：
 - 上傳 docker image：
 - 授權：於 Cloud Shell 或 GCE VM 執行下列指令

Unset

```
sudo gcloud auth configure-docker us-central1-docker.pkg.dev
```

■ tag 本地 image

Unset

```
sudo docker tag SOURCE-IMAGE  
us-central1-docker.pkg.dev/PROJECT-ID/REPOSITORY/IMAGE:TAG
```

■ 將 image 推到 Artifact Registry :

Unset

```
sudo docker push us-central1-docker.pkg.dev/PROJECT-ID/REPOSITORY/IMAGE:TAG
```

更多說明請參考官方文件：

<https://cloud.google.com/artifact-registry/docs/docker/pushing-and-pulling#push-tagged>

○ 從 Artifact Registry 拉 image 到本機 docker :

Unset

```
sudo docker pull  
us-central1-docker.pkg.dev/PROJECT-ID/REPOSITORY/IMAGE@IMAGE-DIGEST
```

更多說明請參考官方文件：

<https://cloud.google.com/artifact-registry/docs/docker/pushing-and-pulling#pulling>

12. Cloud Run

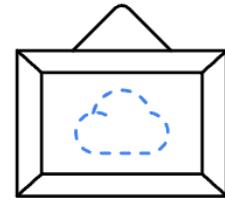
- Cloud Run 是一個完全託管的運算環境，用於部署和擴展無伺服器 HTTP 容器。
- 從選單或搜尋找到 Cloud Run，點擊 CREATE SERVICE 創建一個 Cloud Run 服務。

The screenshot shows the Google Cloud Services page for the project 'tsmccareerhack2025-bsid-test'. The 'SERVICES' tab is selected. A red box highlights the 'Service' section, which contains the text: 'Each service has a unique endpoint & autoscales deployed code.' Below it, another red box highlights the 'Job' section, which contains the text: 'Job Execute code to completion.'

Services

Filter services

Name	Deployment type	Req/sec	Region	Authentication	Ingress	Recommendation	Last deployed
No results to display							



- 選擇 container image :
 - 快速選用範例 container image：點擊 TEST WITH A SAMPLE CONTAINER
 - 選用自備的 container image：點擊 SELECT，從 Artifact Registry 中選擇欲使用的 image。

The screenshot shows the 'Create service' page for Cloud Run. In the 'Container image URL' field, the text 'TEST WITH A SAMPLE CONTAINER' is entered, and a red box highlights the 'SELECT' button next to it. To the right, a modal window titled 'Select container image' is open, showing the 'ARTIFACT REGISTRY' tab selected. It lists two items under 'Project: tsmccareerhack2025-bsid-test' (with 'CHANGE' link):

- ▶ Demo containers
- ▶ us-central1-docker.pkg.dev/tsmccareerhack2025-bsid-test/tsmccareerhack2025-bsid-test-repository

 The 'SELECT' button in the modal is also highlighted with a red box.

Configure

Service name *

Region *

Endpoint URL *

https://service-name-727146887239.us-central1.run.app

- 其他設定：根據實際需求設定參數：

CPU allocation and pricing

CPU is only allocated during request processing
You are charged per request and only when the container instance processes a request.

CPU is always allocated
You are charged for the entire lifecycle of the container instance.

Autoscaling
Minimum and maximum numbers of instances the created revision scales to.

Minimum number of instances * Maximum number of instances *

Set to 1 to reduce cold starts. [Learn more](#)

Ingress control

Internal
Allow traffic from your project, shared VPC, and VPC service controls perimeter. Traffic from another Cloud Run service must be routed through a VPC. Limitations apply. [Learn more](#)

All
Allow direct access to your service from the internet

Authentication *

Allow unauthenticated invocations
Check this if you are creating a public API or website.

Require authentication
Manage authorized users with Cloud IAM.

Container(s), Volumes, Networking, Security

CREATE CANCEL

- 點擊 CREATE 創建該服務。
- 創建成功後，點擊上方的 URL 連結以開啟所部署的服務。
- 若需要連接 Cloud SQL，請於設定中選擇 instance：

Cloud SQL connections

Cloud SQL instance 1
tsmccareerhack2025-bsid-test:us-central1:sql-instance-relational

+ ADD CONNECTION

詳細設定請參考官方文件：[Quickstart: Connect to Cloud SQL for PostgreSQL from Cloud Run | Google Cloud](#)

13. Cloud SQL

- Vector DB : sql-instance-vector
- 點擊 Stop 關機 (請記得不使用時要關機！) :

The screenshot shows the Google Cloud SQL Overview page for the instance 'sql-instance-vector'. The left sidebar lists various instance management options like Overview, Cloud SQL Studio, System insights, etc. The main area displays the instance details, including its PostgreSQL version (PostgreSQL 16). A warning message states: 'This instance's backups settings don't follow your organization policy "Resource Location Restriction"'. Below this is a chart section showing CPU utilization over the last 1 day. At the bottom, there is a link to 'Go to Query insights for more in-depth info on queries and performance'.

- 連線方式：

- 查看 instance 的 private IP

The screenshot shows the Google Cloud Instances page. It lists three instances: 'sql-instance-relational' and 'sql-instance-vector' (both selected), and 'sql-instance-vector'. The 'sql-instance-vector' row is highlighted with a red box. The 'Private IP address' column shows '10.1.0.3' and '10.1.1.3'. The 'Location' column shows 'us-central1-c' for both.

Instance ID	Issues	Cloud SQL edition	Internal connection method	Private IP address	Location
<input type="checkbox"/> sql-instance-relational		Enterprise	Private Services Access (PSA)	10.1.0.3	us-central1-c
<input checked="" type="checkbox"/> sql-instance-vector		Enterprise	Private Services Access (PSA)	10.1.1.3	us-central1-c

- 從提供的 GCE VM 中使用 postgres client

```

@gce-instance:~$ psql -h 10.1.1.3 -U postgres
Password for user postgres:
psql (15.10 (Debian 15.10-0+deb12u1), server 16.6)
WARNING: psql major version 15, server major version 16.
          Some psql features might not work.
SSL connection (protocol: TLSv1.3, cipher: TLS_AES_256_GCM_SHA384, compression: off)
Type "help" for help.

postgres=>
  
```

- Vector DB 使用方式請參考官方文件：[Work with vector embeddings | Cloud SQL for PostgreSQL](#)

- 使用 psql 連接 DB 後，需建立 EXTENSION

```
Unset
\c DB_NAME

GRANT EXECUTE ON FUNCTION embedding TO USER_NAME;
```

14. GKE

- Cluster: careerhack-cluster-icsd

The screenshot shows the Google Cloud Platform interface for managing Kubernetes clusters. On the left, there's a sidebar with 'Resource Management' expanded, showing 'Clusters' selected. The main area displays a table of clusters, with one row for 'careerhack-cluster-icsd' highlighted. At the top right, a search bar contains the text 'gke'. A dropdown menu appears, listing several options related to Kubernetes Engine, with the first item, 'Kubernetes Engine', highlighted by a red box.

This screenshot shows the same Google Cloud Platform interface as the previous one, but the search results for 'gke' are now displayed in a separate window or panel. The 'Kubernetes Engine' result from the dropdown is the top item in the list. Other results include 'GKE Clusters Dashboard', 'etcd', 'GKE overview | Google Kubernetes Engine (GKE)', 'GKE cluster architecture | Google Kubernetes Engine (GKE)', 'Java application deployment on GKE', 'GKE Multi-Cloud API', 'Google Kubernetes Engine documentation', 'WordPress', 'Apache Superset', and a note about resource results for the current project.

連線方式：

需於 GCE VM 中操作 (請參考前面章節連線至 GCE VM)

```

SSH-in-browser
Linux gce-instance 6.1.0-28-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.119-1 (2024-11-22) 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.
Creating directory '/home/_'.
@gce-instance:~$ gcloud container clusters get-credentials careerhack-cluster-tsid-1 --zone us-central1-a --project tsmccareerhack2025-tsid-test
Fetching cluster endpoint and auth data.
kubeconfig entry generated for careerhack-cluster-tsid-1.
@gce-instance:~$ kubectl get ns
NAME      STATUS   AGE
default   Active   94m
gke-managed-cim Active   94m
gke-managed-system Active   94m
gmp-public Active   94m
gmp-system Active   94m
kube-node-lease Active   94m
kube-public Active   94m
kube-system Active   95m
@gce-instance:~$ 

```

取得 gke credentials

Unset

```
gcloud container clusters get-credentials careerhack-cluster-icsd --region
us-central1 --project tsmccareerhack2025-icsd-test
```

取得 credentials 後可使用 kubectl 操作 cluster (請參考上圖範例)

Resize Node Pool:

開機 - 將 worker node resize to 1

關機 - 將 worker node resize to 0 (請記得不使用時要關機 !)

點擊 cluster: careerhack-cluster-tsid

The screenshot shows the Google Cloud Kubernetes Engine interface. The top navigation bar includes 'Google Cloud' and 'tsmccareerhack2025-icsd-grp1'. The main area has tabs for 'OVERVIEW', 'OBSERVABILITY', and 'COST OPTIMIZATION'. On the left, a sidebar lists 'All Fleets', 'Resource Management' (with 'Clusters' selected), 'Workloads', 'Teams', and 'Applications'. The 'Clusters' table lists one cluster: 'careerhack-cluster-icsd' (Status: Active, Location: us-central1-a, Tier: Standard, Number of nodes: 0, Total vCPUs: 0). A red box highlights the cluster name 'careerhack-cluster-icsd'.

Status	Name	Location	Tier	Number of nodes	Total vCPUs
<input type="checkbox"/>	<input checked="" type="checkbox"/> careerhack-cluster-icsd	us-central1-a	Standard	0	0

點擊 NODES → Node Pools → default-pool

Google Cloud tsmccareerhack2025-icsd-grp1 gke

Career Hack Cluster - careerhack-cluster-icsd

Your cluster has one or more unschedulable pods. [Autoscaling documentation](#)

VIEW DETAILS AND POSSIBLE ACTIONS

Automatic upgrades and other Google maintenance tasks may run at any time, increasing transient disruptions to your workloads. Set a maintenance window. [VIEW DETAILS](#)

Nodes

Name	Status	Version	Number of nodes	Machine type	Image type	Autoscaling	Default IPv4 Pod IP address range
default-pool	Ok	1.31.4-gke.1256000	0	g2-standard-24	Container-Optimized OS with containerd (cos_containerd)	Off	10.144.0.0/14

開機 - 將 worker node resize to 1

關機 - 將 worker node resize to 0 (請記得不使用時要關機 !)

Google Cloud tsmccareerhack2025-icsd-grp1 gke

Node pool details

default-pool

RESIZE

Node pool basics

Cluster	careerhack-cluster-icsd
Node version	1.31.4-gke.1256000
Current COS version	cos-117-18613-75-66
End of standard support	?
End of extended support	?

Size

Number of nodes * 0

Nodes

CANCEL RESIZE

15. Monitoring

- 進入 Monitoring 服務的 Dashboards，打開 Usage Monitoring Dashboard 查看 GCE VM 的使用時數以及 gemini 使用次數。

The screenshot shows the Google Cloud Monitoring Dashboards Overview page. The left sidebar has a 'Monitoring' section with various sub-options like Metrics Scope, Overview, Dashboards, Integrations, Services, etc. The 'Dashboards' option is selected and highlighted with a red box. The main content area is titled 'Dashboards Overview' and features a 'DASHBOARD LIST' tab. It displays a table of dashboards categorized into 'Categories' (Recently Viewed, Favorites, Custom, GCP, Integrations, Other) and 'All Dashboards'. A 'SAMPLE LIBRARY' link is also present. The 'Usage Monitoring Dashboard' is listed under the 'Custom' category and is also highlighted with a red box.

Categories	All Dashboards	LABELS
Filter by category	Filter Dashboards	
All	9	<input type="checkbox"/> Name
Recently Viewed	1	Autoscaler Monitoring
Favorites	0	Cloud Storage
Custom	1	Disks
GCP	8	Firewalls
Integrations	0	GCE VM Instance Monitoring
Other	0	GCE VM Lifecycle Events Monitoring
		Infrastructure Summary
		<input type="checkbox"/> Usage Monitoring Dashboard
		VM Instances

← Usage Monitoring Dashboard + Add widget ⋮ Share

Annotations (3) Group by Filter Autosave

Review usage of timed resources

Please note that the resources listed have a strict limit of 60 hours.

Remember to power off the machines when they are not in use.

[CareerHack Resource Timing Dashboard \(Student View\)](#)

GCE VM uptime total

No data is available for the selected time frame.

UTC+8 1:50PM 2:00PM 2:10PM 2:20PM 2:30PM

0 time series

Vertex AI model invocation c...

No data is available for the selected time frame.

UTC+8 1:50PM 2:00PM 2:10PM 2:20PM 2:30PM

0 time series

GKE Instance group size [MAX]

No data is available for the selected time frame.

UTC+8 1:50PM 2:00PM 2:10PM 2:20PM 2:30PM

0

Vertex AI Endpoint - Prediction count [SUM]

Filter Enter property name or value

No rows to display.

- 若欲查看 GCE VM 的 metrics (CPU utilization, memory utilization, disk operations...)，可透過在 Console 點擊進入 Compute Engine 的 OBSERVABILITY 頁面查看。

The screenshot shows the Google Cloud Compute Engine Observability dashboard. The left sidebar is collapsed, showing the Compute Engine icon and the 'VM instances' option highlighted with a red box. The main content area has tabs for 'INSTANCES', 'OBSERVABILITY' (which is selected and highlighted with a red box), and 'INSTANCE SCHEDULES'. Below these tabs are 'RECOMMENDED ALERTS' and 'SAVE AS DASHBOARD' buttons. A time range selector shows 'RESET ZOOM' and options for '1 hour', '6 hours', '1 day', '1 week', '1 month', '6 weeks', and 'Custom'. The dashboard displays four line charts: 'CPU Utilization (Top 5 VMs)', 'Memory Utilization (Top 5 VMs)', 'Disk Utilization (Top 5 VMs)', and 'Processes by CPU Usage (Top 5)'. Each chart includes a legend, a timestamp from 'UTC+8 1:10 PM' to '1:50 PM', and a zoom control. To the right of the charts is a callout box with the text 'Visualize your entire infrastructure' and 'Get instant, visual answers to all your infrastructure related questions.' with a 'EXPLORE' button.

16. Alerting

- 若有需要獲取一些監控 metric 的告警 (alerts)，可在 Monitoring → Alerting 設定。

The screenshot shows the Google Cloud Monitoring interface. The left sidebar has a red box around the 'Monitoring' section, which is currently selected. The main content area has two red boxes: one around the '+ CREATE POLICY' button and another around the 'EDIT NOTIFICATION CHANNELS' button. The 'Alerting' tab is also highlighted in blue.

Metrics Scope
1 project

Alerting

+ CREATE POLICY

EDIT NOTIFICATION CHANNELS

Summary

Incidents firing	Incidents acknowledged
0	0

Incidents

State	Severity	Policy name	Incident summary	Opened	Closed
No rows to display					

→ See all incidents

Snoozes

+ CREATE SNOOZE

State	Name
No rows to display	

- 選定特定 metric 作為觸發告警的條件及其門檻值（請根據各自需求設定各參數，更多詳細說明請參考官方說明：<https://cloud.google.com/monitoring/alerts/using-alerting-ui>）

Google Cloud tsmccareerhack2024-bsid-test Search (/) for resources, docs, products, and more Search

Create alerting policy + ADD ALERT CONDITION DELETE ALERT CONDITION MQL

ALERT CONDITIONS

- New condition
- Configure trigger

ALERT DETAILS

- Notifications and name
- Review alert

Select a metric ?

SELECT A METRIC

Select a metric Filter by resource or metric name

Active

POPULAR RESOURCES

VM Instance 76 metrics >

ACTIVE RESOURCES

Audited Resource 2 metrics >

Consumed API 4 metrics >

Consumer Quota 5 metrics >

GCP Location 4 metrics >

Disk 10 metrics >

Firewall 2 metrics >

Instance 28 metrics >

Interface 3 metrics >

Logs-based metrics 2 metrics >

Memory 2 metrics >

ACTIVE METRIC CATEGORIES

Agent 7 metrics >

Cpu 4 metrics >

Disk 10 metrics >

Firewall 2 metrics >

Instance 28 metrics >

Interface 3 metrics >

Logs-based metrics 2 metrics >

Memory 2 metrics >

ACTIVE METRICS

Disk bytes read agent.googleapis.com/disk/read_bytes_count

Disk bytes used agent.googleapis.com/disk/bytes_used

Disk bytes written agent.googleapis.com/disk/write_bytes_count

Disk I/O time agent.googleapis.com/disk/io_time

Disk merged operations agent.googleapis.com/disk/merged_operations

Disk operation time agent.googleapis.com/disk/operation_time

Disk operations agent.googleapis.com/disk/operation_count

Disk pending operations

Selection preview VM Instance > disk

Cancel Apply

CREATE POLICY PROVIDE FEEDBACK CANCEL

Google Cloud tsmccareerhack2024-bsid-test Search (/) for resources, docs, products, and more Search

Create alerting policy + ADD ALERT CONDITION DELETE ALERT CONDITION MQL VIEW CODE

ALERT CONDITIONS

- VM Instance - CPU usage
- Configure trigger

ALERT DETAILS

- Notifications and name
- Review alert

Configure alert trigger

Condition Types

Threshold Condition triggers if a time series rises above or falls below a value for a specific duration window

Metric absence Condition triggers if any time series in the metric has no data for a specific duration window

Forecast PREVIEW Condition triggers if any timeseries in the metric is projected to cross the threshold in the near future.

Alert trigger Any time series violates

Threshold position Above threshold

Threshold value

Advanced Options

Condition name * VM Instance - CPU usage

NEXT CREATE POLICY PROVIDE FEEDBACK CANCEL

VM Instance - CPU usage

Filter Enter property name or value

Metric	Value
usage_time	0.002

● 設定接收告警的管道

The screenshot shows the 'Create alerting policy' interface in Google Cloud. On the left, a sidebar lists 'ALERT CONDITIONS' (VM Instance - CPU usage) and 'ALERT DETAILS' (Notifications and name, which is selected and highlighted with a red box). The main panel is titled 'Configure notifications and finalize alert'. It includes a 'Configure notifications' section with a 'Use notification channel' toggle (which is checked and highlighted with a red box), a 'Notification Channels' section (which displays a message: 'There are no available notification channels for this workspace.' and a 'MANAGE NOTIFICATION CHANNELS' button, both highlighted with red boxes), and a note about redundancy. Below these are sections for 'Notify on incident closure' (checkbox) and 'Incident autoclose duration' (dropdown set to '7 days'). At the bottom are 'CREATE POLICY', 'PROVIDE FEEDBACK', and 'CANCEL' buttons.

The screenshot shows the 'Notification channels' configuration interface in Google Cloud. The sidebar shows 'ALERT CONDITIONS' (VM Instance - CPU usage) and 'ALERT DETAILS' (Notifications and name). The main panel lists various notification channels: Google Chat (PREVIEW), PagerDuty Services, PagerDuty Sync (BETA), Slack (with a question mark icon), Webhooks, Email (highlighted with a red box), SMS (highlighted with a red box), and Pub/Sub. Each channel has a status message ('Monitoring now supports both user-scoped and device-scoped Cloud Console Mobile notification channels') and an 'ADD NEW' button. A 'LEARN MORE' link and a 'DISMISS' button are also present at the top right.

17. Review 後點擊 CREATE POLICY 創建該 alert policy。

