

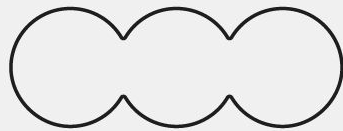



Google Developer Group
Singapore University of Technology and
Design (SUTD)

Deploying a FastAPI Chatbot to Cloud Run with Gemini Integration

Ananda Dwi Rahmawati

Google Developer Expert - Cloud



{ Build  with AI }



Google Developer Group
Singapore University of Technology and
Design (SUTD)

Ananda Dwi Rahmawati

- ❑ Cloud & DevOps Engineer, Singapore
- ❑ Google Developer Expert Cloud - Modern Architecture
- ❑ Master of Computer Science - University of Texas at Austin
- ❑ <https://linktr.ee/misskecupbung>

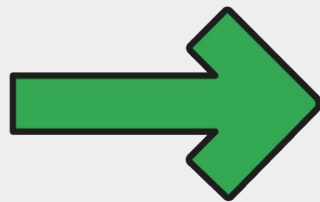


{ Build  with AI }

Chapter One

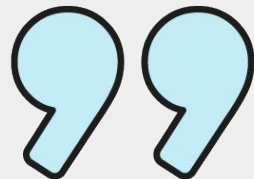
Deploying a FastAPI Chatbot to Cloud Run with Gemini Integration

Overview





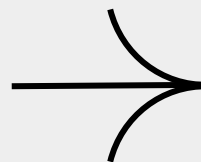
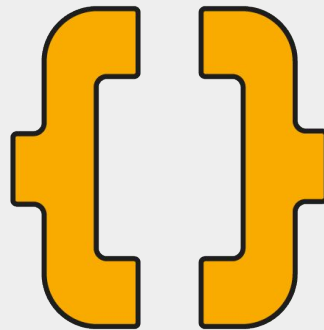
“FastAPI handles your endpoints, Gemini powers your responses, and Cloud Run makes it all effortlessly scalable.”

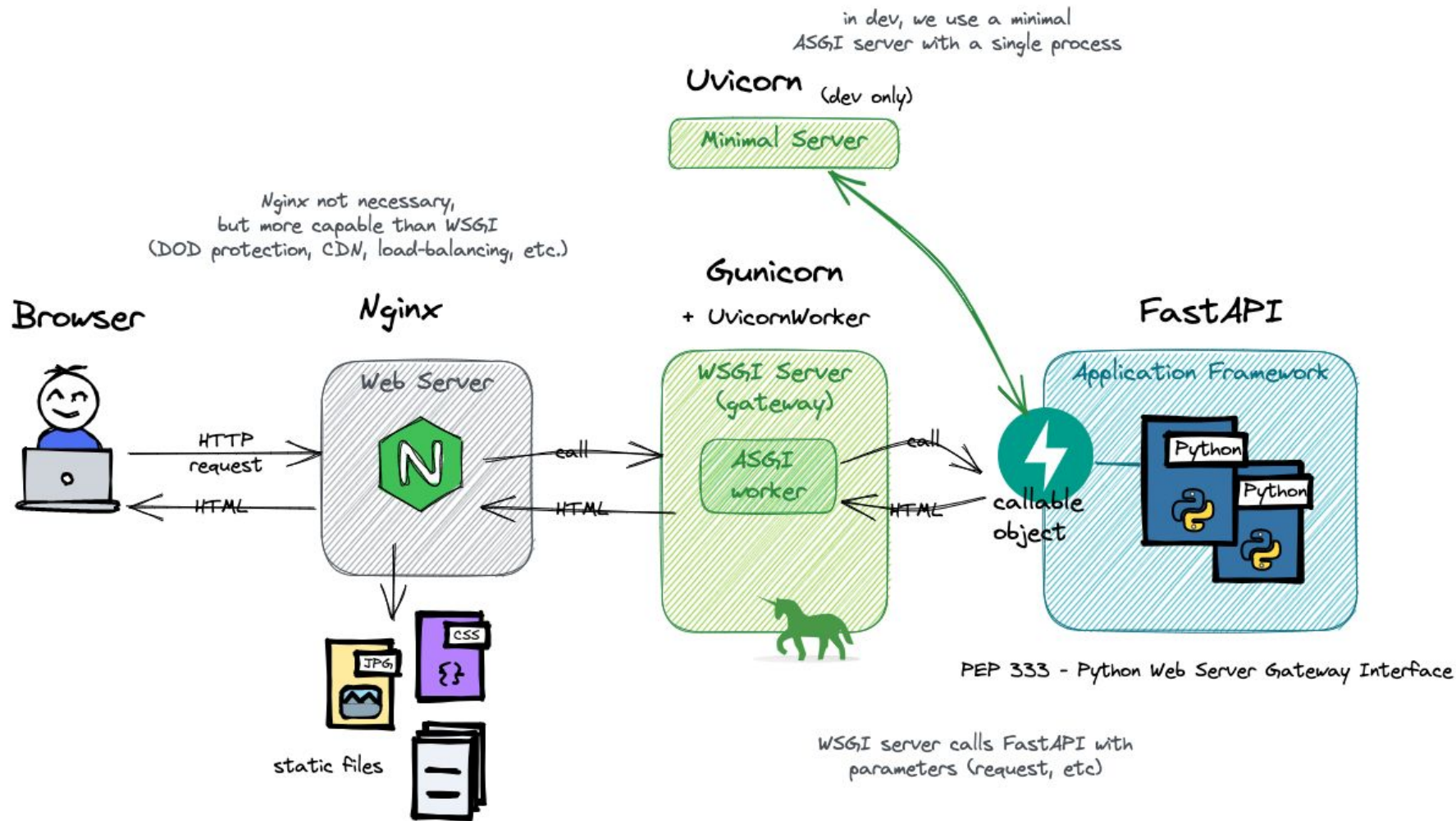


What is FastAPI

FastAPI is a modern, high-performance web framework for building APIs with Python 3.7+.

- **Fast:** Built on Starlette and Pydantic—optimized for speed.
- **Easy to Use:** Automatic validation, docs generation (Swagger, ReDoc).
- **Asynchronous:** Supports async/await for high concurrency.
- Great for ML/AI services, microservices, and REST APIs.

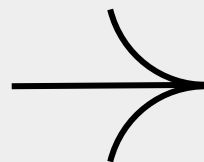
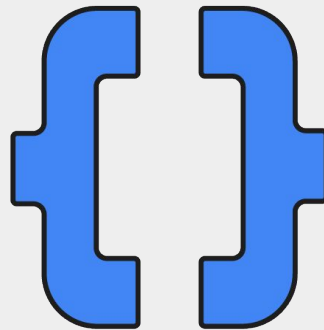




What is Gemini Chatbot?

Gemini is Google's family of advanced generative AI models (like GPT).

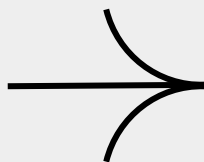
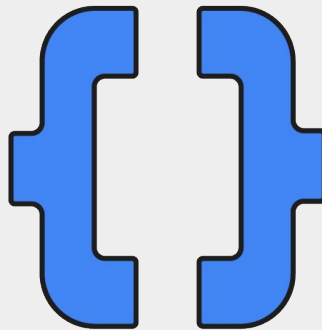
- **Built by Google DeepMind**, integrated into Google Cloud AI products.
- **Gemini Chatbot** uses these models to understand and generate human-like responses.
- **Supports multi-turn conversations**, code understanding, and contextual reasoning.
- Access via Google's Generative AI SDK or Vertex AI API.



Gemini Chatbot?

Gemini is Google's family of advanced generative AI models (like GPT).

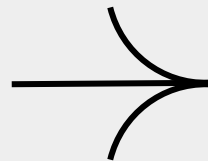
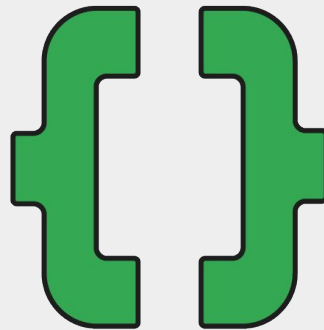
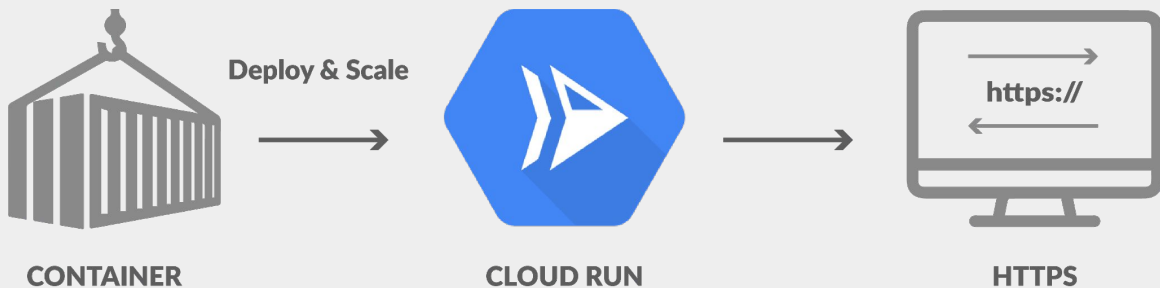
- **Built by Google DeepMind**, integrated into Google Cloud AI products.
- **Gemini Chatbot** uses these models to understand and generate human-like responses.
- **Supports multi-turn conversations**, code understanding, and contextual reasoning.
- Access via Google's Generative AI SDK or Vertex AI API.

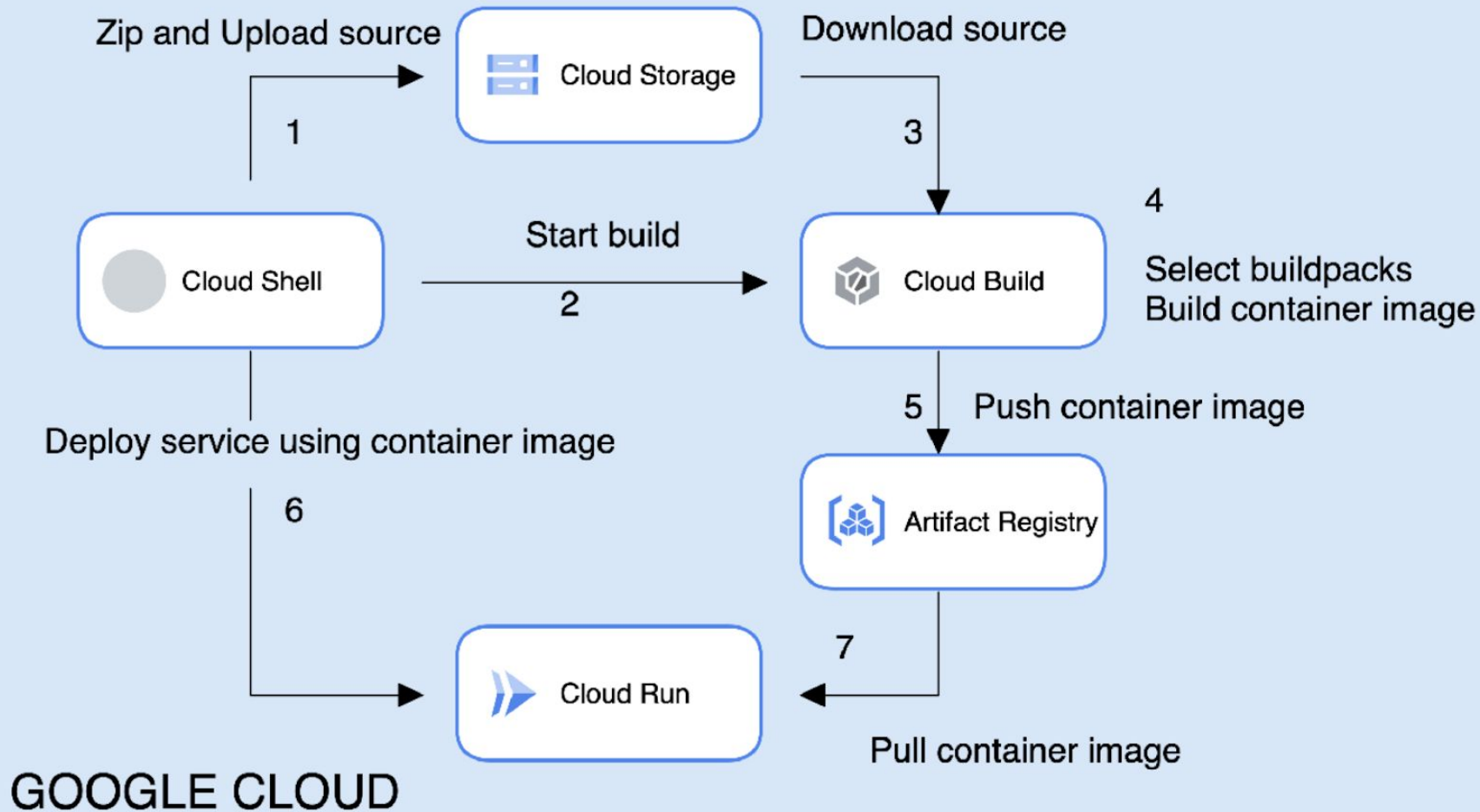


What is Cloud Run?

Cloud Run is a fully managed compute platform by Google Cloud for running containerized apps.

- **Deploy any Docker container**—no server management needed.
- **Scales automatically:** from zero to thousands of instances.
- Pay only when in use.
- **Ideal for microservices**, APIs, or event-driven apps.

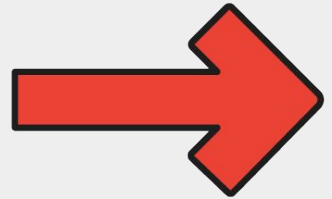


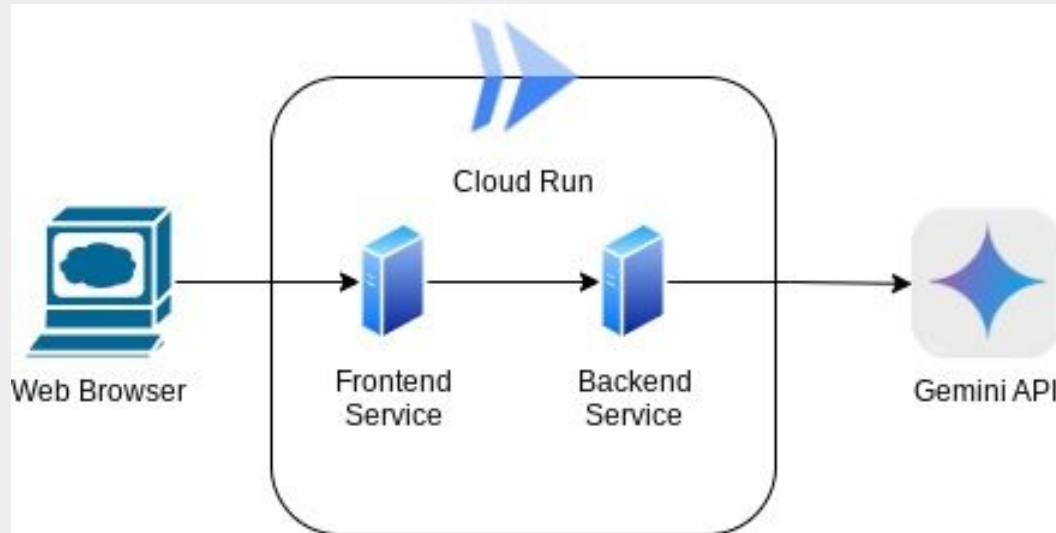


Chapter Two

Let's Demo!

Are you ready?

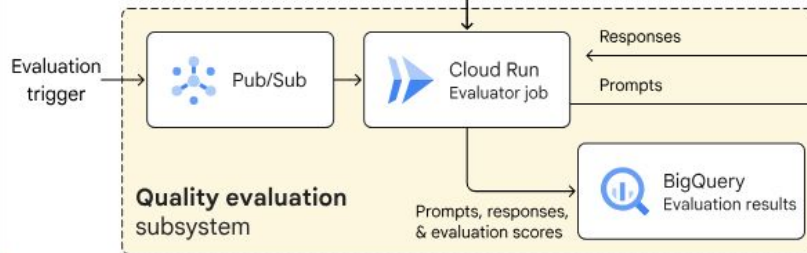
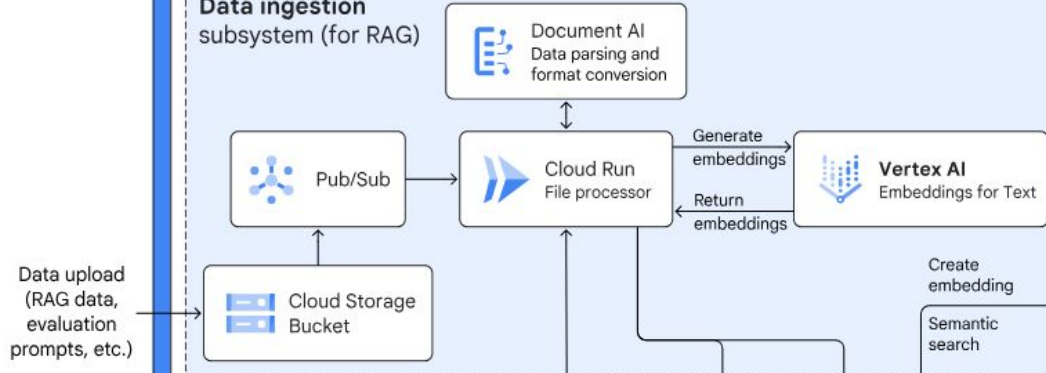




Google Cloud

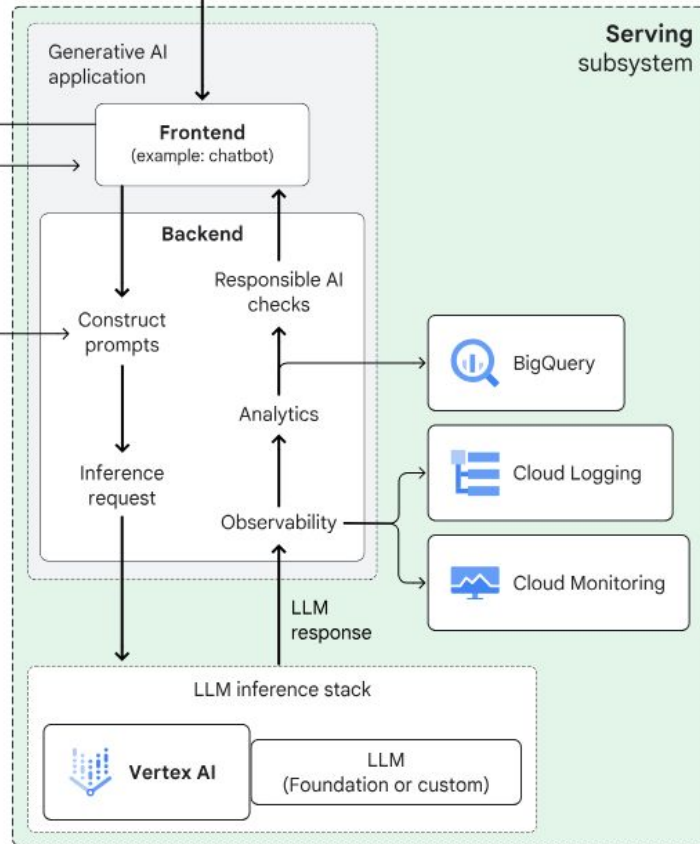
Region AA

Data ingestion subsystem (for RAG)

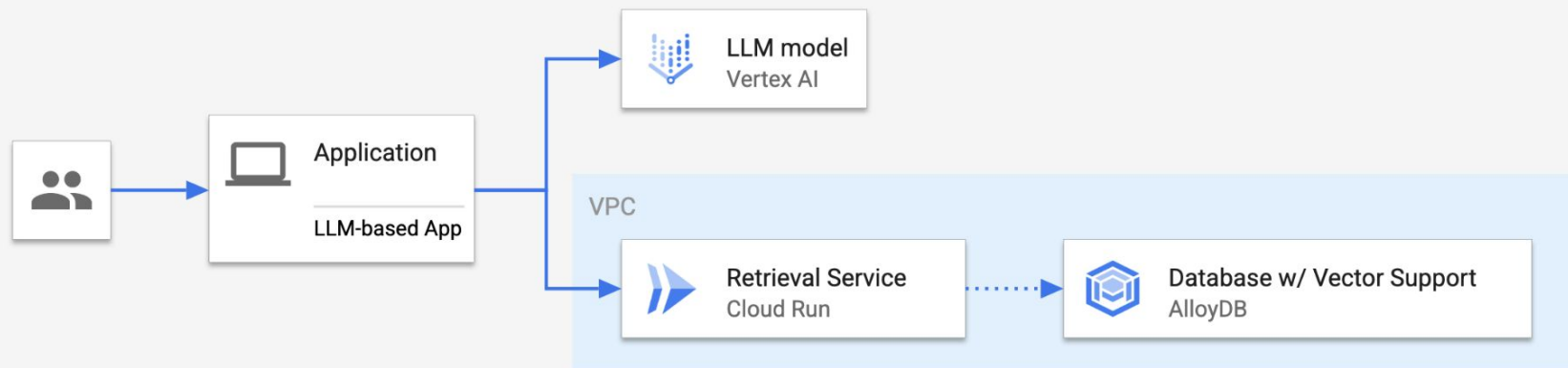


Application users

Request Response

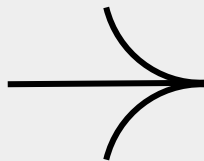
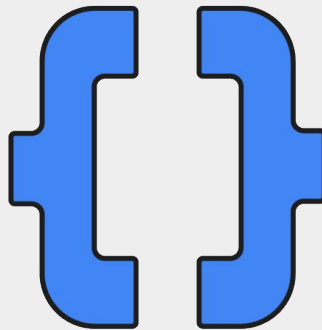


GenAI Database Retrieval



Prerequisites

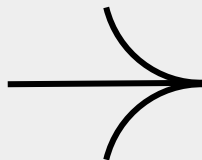
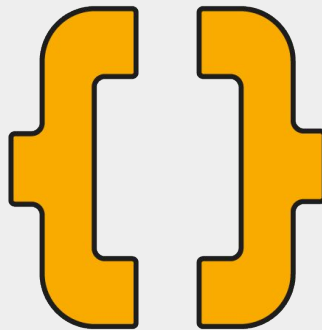
- A basic understanding of Generative AI, Gemini on Google Cloud
- A basic understanding of Cloud Run and Vertex AI Concepts



What you will learn

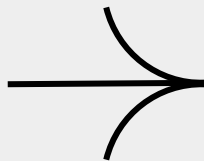
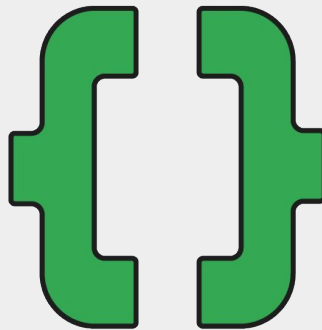
In this lab, you'll learn:

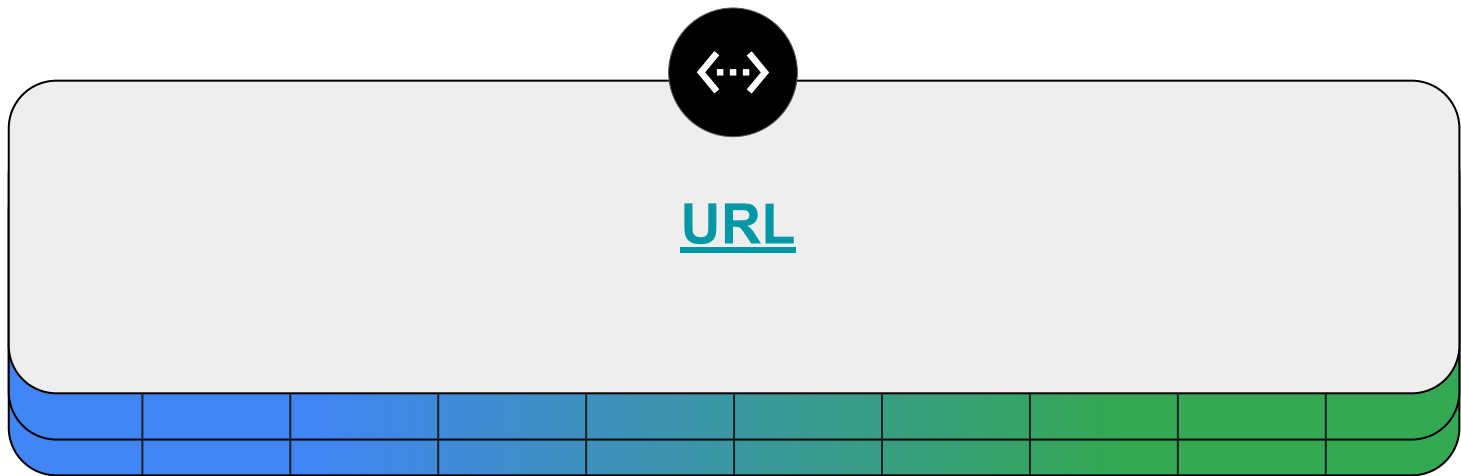
- How to deploy FastAPI to Cloud Run
- Prompt Gemini from Cloud Run in python using a Google client library



What you will need

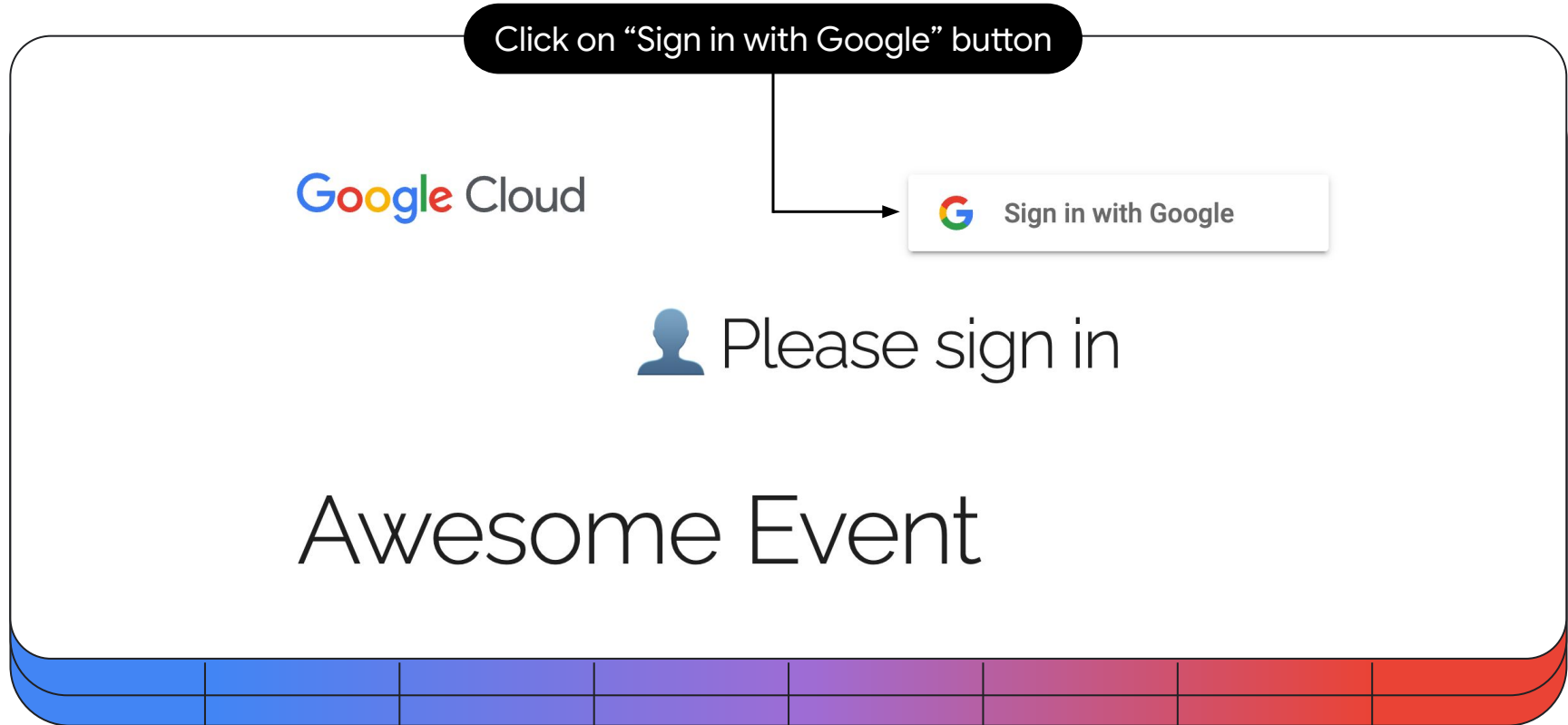
- A curious mind
- A working computer and reliable wifi
- A Google Cloud project with billing attached.





Open this URL

You will need to be signed in



You will need to be signed in

You should see a similar page. Click on this button

Google Cloud

[Sign out](#)

Hi, welcome Cloud Developer (ccloud-dev@gmail.com)

Awesome Event

Your credit will allow you to use Google Cloud [Free Tier products](#).

It has an amount of **\$1**.

Once redeemed, it will be valid for **180 days** or until the balance is depleted if you use non-free services.

CLICK HERE TO ACCESS YOUR CREDITS

After redeeming your credit in the Google Cloud console, please [proceed to the next step](#).

Make sure that you are applying for the correct account

GCP credit application

Fill in the following information below to apply GCP credits to your account listed below.

First name *

Amazing

Last name *

Person

Account email

cloud-dev@gmail.com

Credits will be applied to this account. If you'd like to apply credits to a different account, specify your preference [here](#).

Coupon code

JJKT-50BF-8FM9-KD8N

Terms and conditions

The following terms and conditions apply to the credit you received for Google Cloud products (the "Credit(s)").

The Credit is subject to valid registration and acceptance of an account with Google Cloud and satisfaction of any applicable eligibility requirements including the Google Cloud Platform [Terms of Service](#). You will be responsible for all usage in excess of the Credit and you may not be notified once the Credit is exhausted. The Credit is non-transferable and may not be sold or bartered. The Credit is valid for a limited time only and expires on the date indicated when you receive the applicable Credit code or on such date as designated by Google (in which case the earlier date applies). You may not use the Credit to engage in mining cryptocurrency unless you have obtained Google's written consent, which consent may be revoked by Google in its sole discretion at any time. Google reserves the right to cancel the Credit or change these terms at any time. You are responsible for determining the applicable tax treatment of receiving the Credits and for paying all applicable taxes. Offer void where prohibited by law.

Except for graduate or work-study students participating in an event in their personal capacities, if you are a government employee, including an employee of a public university, public educational institution or state-owned enterprise, you may not use (and you are ineligible to receive) any Credits.

ACCEPT AND CONTINUE

* Indicates required

Click on "Accept and Continue" to proceed



Google Developer Group

Step 1

Claim credits

Step2

Create a
Google Cloud
project

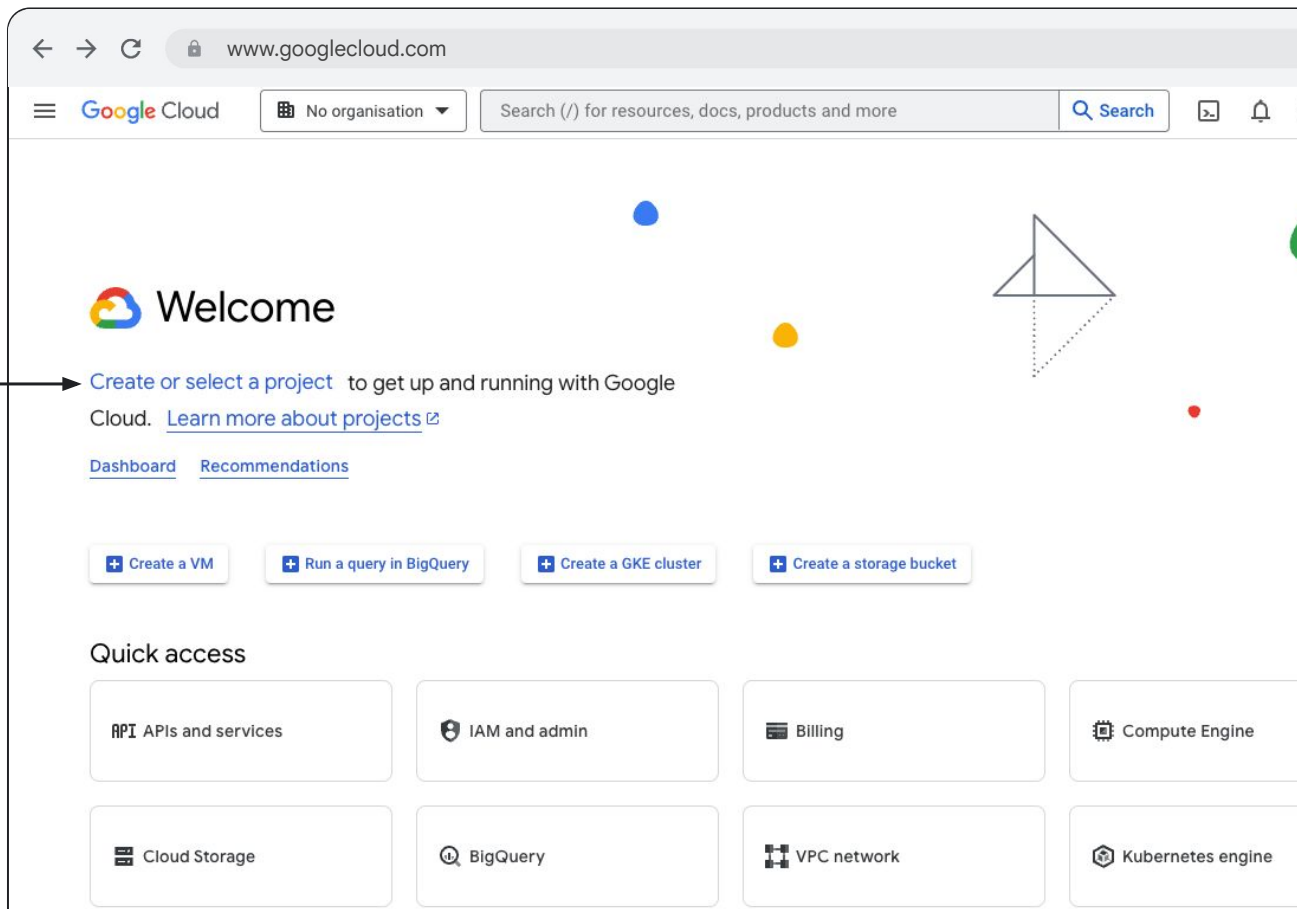


Google Developer Group

Go to:

console.cloud.google.com

Click on “Create or select a project”



The screenshot shows the Google Cloud console homepage. At the top, there's a navigation bar with the Google Cloud logo, a dropdown for 'No organisation', a search bar, and a search button. Below the navigation bar, the main content area features a 'Welcome' message with the Google Cloud logo. A callout arrow from the text 'Click on “Create or select a project”' points to the link 'Create or select a project' in the 'Welcome' section. Below the 'Welcome' message, there are links for 'Dashboard' and 'Recommendations'. Further down, there are four buttons: 'Create a VM', 'Run a query in BigQuery', 'Create a GKE cluster', and 'Create a storage bucket'. At the bottom, there's a 'Quick access' section with eight tiles: 'API APIs and services', 'IAM and admin', 'Billing', 'Compute Engine', 'Cloud Storage', 'BigQuery', 'VPC network', and 'Kubernetes engine'.

www.googlecloud.com

Google Cloud

No organisation

Search (/) for resources, docs, products and more

Search

Welcome

Create or select a project to get up and running with Google Cloud. [Learn more about projects](#)

[Dashboard](#) [Recommendations](#)

Create a VM Run a query in BigQuery Create a GKE cluster Create a storage bucket

Quick access

API APIs and services IAM and admin Billing Compute Engine

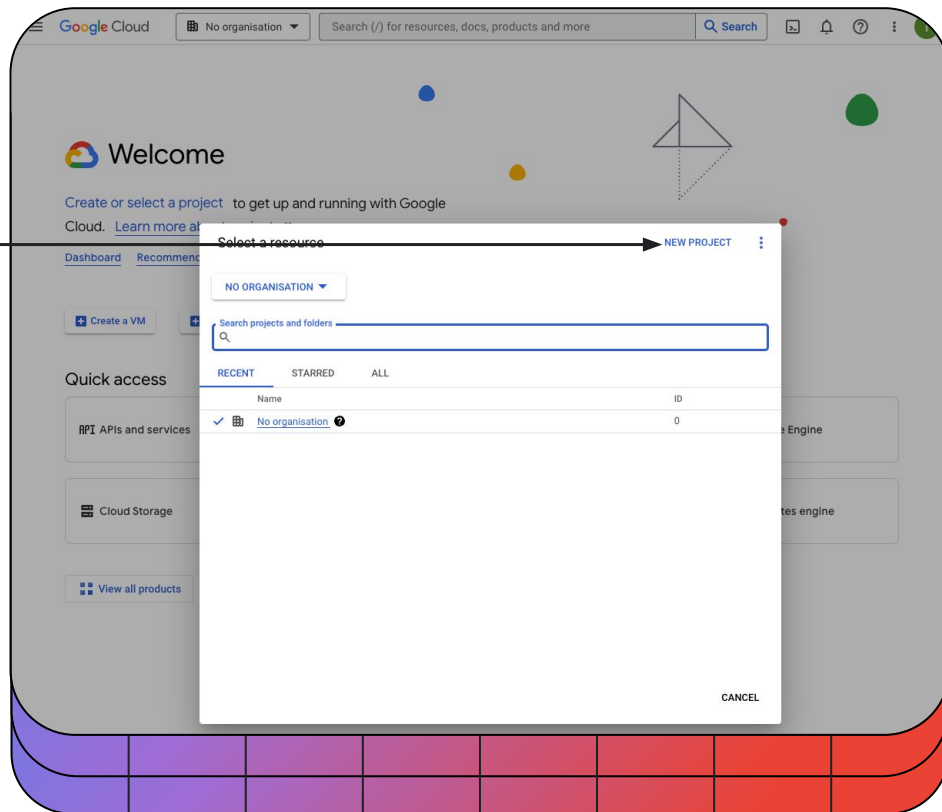
Cloud Storage BigQuery VPC network Kubernetes engine



Google Developer Group

Creating a new project

Click on “New Project”



Creating a new project

Name your project

If you see a billing account, make sure to select the "Trial Billing Account". If NOT, still create the project and go to next slide

Click "CREATE"

New Project



You have 23 projects remaining in your quota. Request an increase or delete projects. [Learn more](#)

[MANAGE QUOTAS](#)

Project name *

awesome-project



Project ID: awesome-project-402007. It cannot be changed later. [EDIT](#)

Billing account *

Google Cloud Platform Trial Billing Account



Any charges for this project will be billed to the account you select here.

Location *

No organization

[BROWSE](#)

Parent organization or folder

CREATE

CANCEL



Google Developer Group

If you don't see a billing account in the previous step:

1) Go to Billing from Google Cloud Console and

2) Set your project's billing account to **Google Cloud Platform Trial Billing Account**



Set the billing account for project "My First Project"

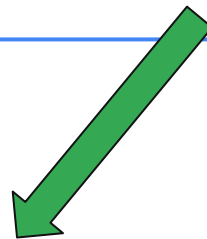
This project pays for both Google Cloud Platform and Maps Platform. Select a billing account that supports both Google Cloud Platform and Maps Platform. [Learn more](#)

Billing account *

Google Cloud Platform Trial Billing Account

Any charges for this project will be billed to the account you select here.

CANCELSET ACCOUNT



Issues When attempting to redeem credits

It may be unclear where the credits land. If your users are confused, have them navigate to the Credits tab on the page on the console

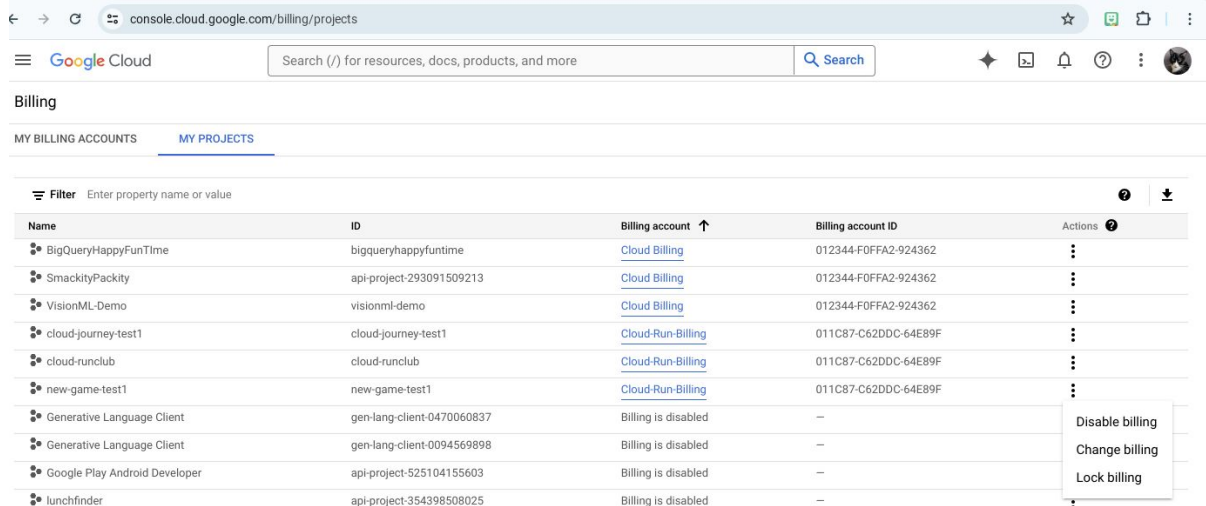
That is where they will see the credits

The screenshot shows the Google Cloud Billing console. The left sidebar contains a navigation menu with the following items: Billing, Overview, Cost management (Reports, Cost table, Cost breakdown, Budgets & alerts, Billing export), Cost optimization (FinOps hub, Committed use discounts (CUD), CUD analysis, Pricing, Cost estimation), and Billing management (Account management). The 'Credits' tab is highlighted in red in the sidebar, and a red arrow points to it. The main content area shows the 'Overview' page for the 'PAID ACCOUNT'. It displays 'Your total cost (April 1 - 18, 2024)' as \$0.00, with 'Cost' at \$0.00 and 'Credits used' at \$0.00. It also shows the 'FinOps hub' with a 'Save up to \$0.00' and a 'Create a budget alert' section.

The screenshot shows the Google Cloud Billing console with the 'Credits' page selected. The left sidebar is the same as the previous screenshot. The main content area shows the 'Credits' page with a table of credits. The table has columns: Credit name, Status, Percent remaining, Remaining value, Original value, Type, Credit ID, and Scope. The table contains one row: 'Frictionless access to Google Cloud' with a status of 'Available', 100% remaining, and a remaining value of \$5.00. The page also includes a 'Filter' section and a 'View details on Reports' link.

Issues if the user ALREADY has a existing billing account make sure they use the correct one

Click on the 3 Dot menu on the project, and select "Change Billing," from the drop down select that Google Cloud Platform Trial Billing

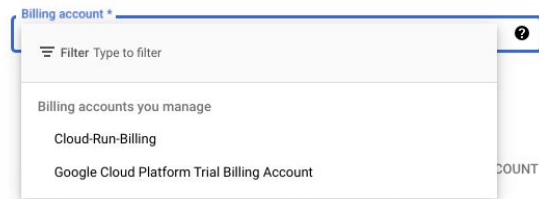


The screenshot shows the Google Cloud Billing console. The browser address bar is `console.cloud.google.com/billing/projects`. The page title is "Billing". Below the title, there are tabs for "MY BILLING ACCOUNTS" and "MY PROJECTS". The "MY PROJECTS" tab is selected. A filter bar is present with the text "Filter Enter property name or value". Below the filter bar is a table with the following columns: Name, ID, Billing account, Billing account ID, and Actions. The table lists several projects, including "BigQueryHappyFunTime", "SmackityPackity", "VisionML-Demo", "cloud-journey-test1", "cloud-runclub", "new-game-test1", "Generative Language Client", "Google Play Android Developer", and "lunchfinder". The "Billing account" column shows the account name (e.g., "Cloud Billing", "Cloud-Run-Billing") and a link to "Change billing". The "Billing account ID" column shows the account ID (e.g., "012344-F0FFA2-924362"). The "Actions" column shows a three-dot menu with options: "Disable billing", "Change billing", and "Lock billing".

Name	ID	Billing account	Billing account ID	Actions
BigQueryHappyFunTime	bigqueryhappyfuntime	Cloud Billing	012344-F0FFA2-924362	⋮
SmackityPackity	api-project-293091509213	Cloud Billing	012344-F0FFA2-924362	⋮
VisionML-Demo	visionml-demo	Cloud Billing	012344-F0FFA2-924362	⋮
cloud-journey-test1	cloud-journey-test1	Cloud-Run-Billing	011C87-C62DDC-64E89F	⋮
cloud-runclub	cloud-runclub	Cloud-Run-Billing	011C87-C62DDC-64E89F	⋮
new-game-test1	new-game-test1	Cloud-Run-Billing	011C87-C62DDC-64E89F	⋮
Generative Language Client	gen-lang-client-0470060837	Billing is disabled	—	⋮
Generative Language Client	gen-lang-client-0094569898	Billing is disabled	—	⋮
Google Play Android Developer	api-project-525104155603	Billing is disabled	—	⋮
lunchfinder	api-project-354398508025	Billing is disabled	—	⋮

Set the billing account for project "new-game-test1"

This project pays for both Google Cloud Platform and Maps Platform. Select a billing account that supports both Google Cloud Platform and Maps Platform. [Learn more](#)



The screenshot shows a dropdown menu titled "Billing account *". It has a search bar with the placeholder text "Filter Type to filter". Below the search bar, there are two options: "Cloud-Run-Billing" and "Google Cloud Platform Trial Billing Account". The "Google Cloud Platform Trial Billing Account" option is highlighted.

Learn more about Google
Cloud at goo.gle/clouddevs





References

- <https://cloud.google.com/run/docs>
- <https://codelabs.developers.google.com/cloud-run-starter-app#0>
- <https://ai.google.dev/gemini-api/docs>

