

Deploying a FastAPI Chatbot to Cloud Run with Gemini Integration

Ananda Dwi Rahmawati

Google Developer Expert - Cloud





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>

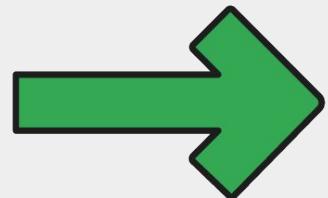


{ Build with AI }

Chapter One

Deploying a FastAPI Chatbot to Cloud Run with Gemini Integration

Overview



“

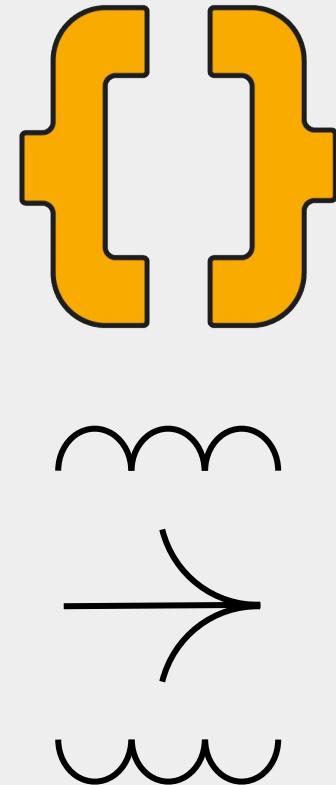
“FastAPI handles your endpoints, Gemini powers your responses, and Cloud Run makes it all effortlessly scalable.”

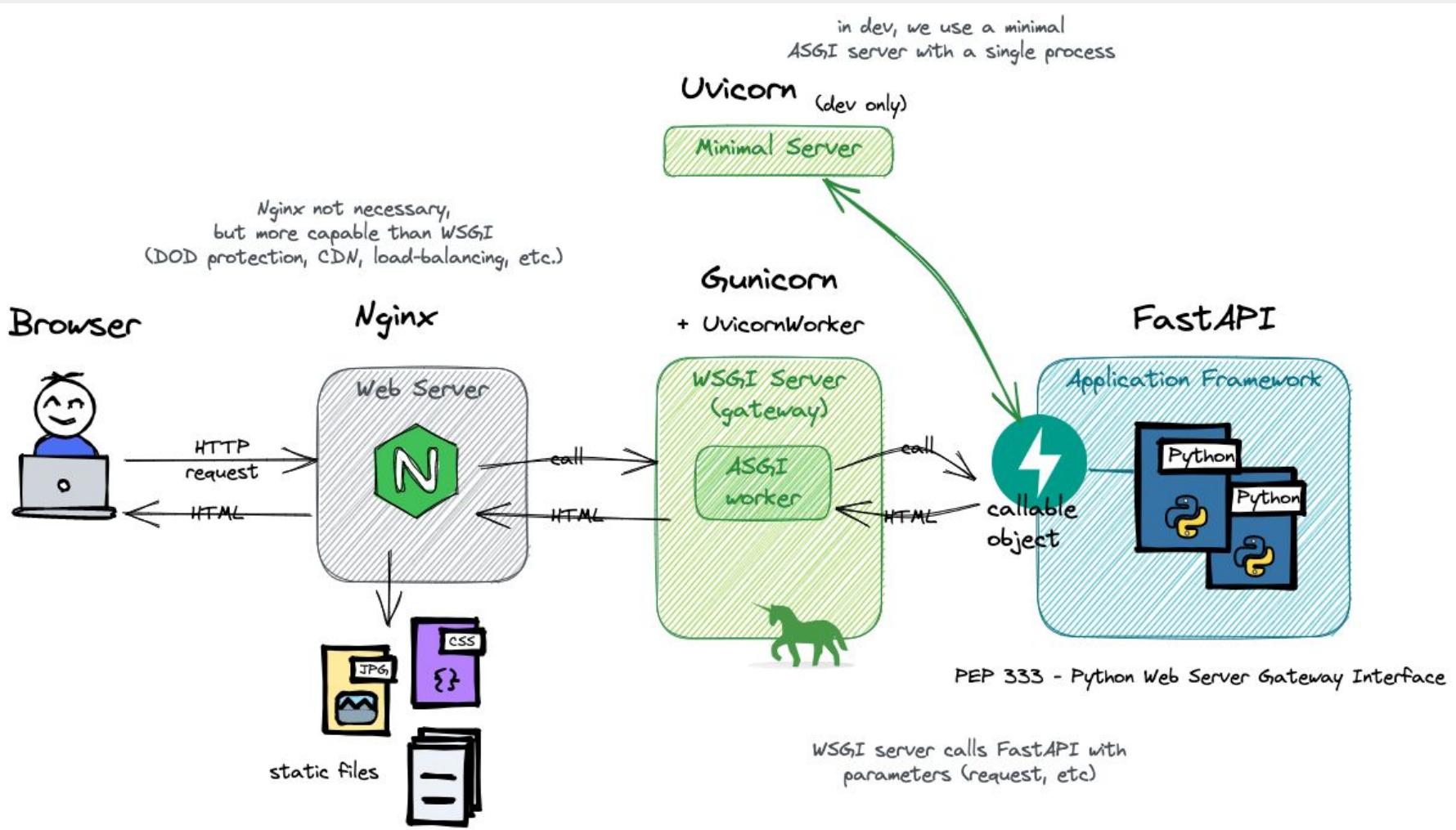
”

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.



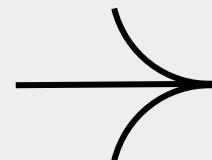
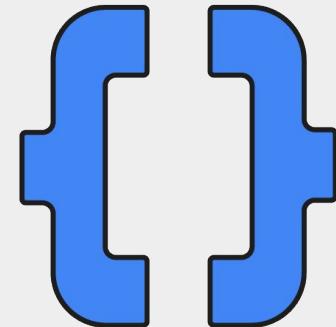


What is Gemini Chatbot?

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

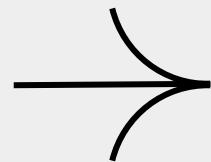
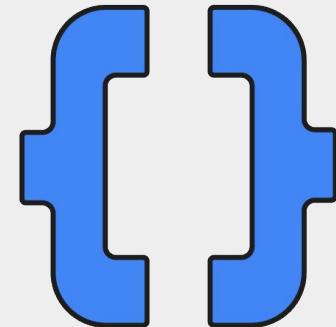


Gemini Chatbot?

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.

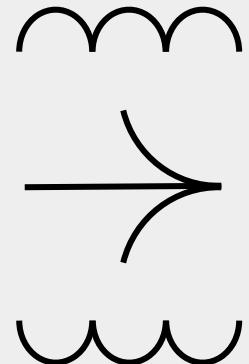
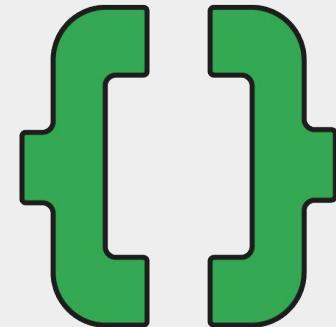
Gemini

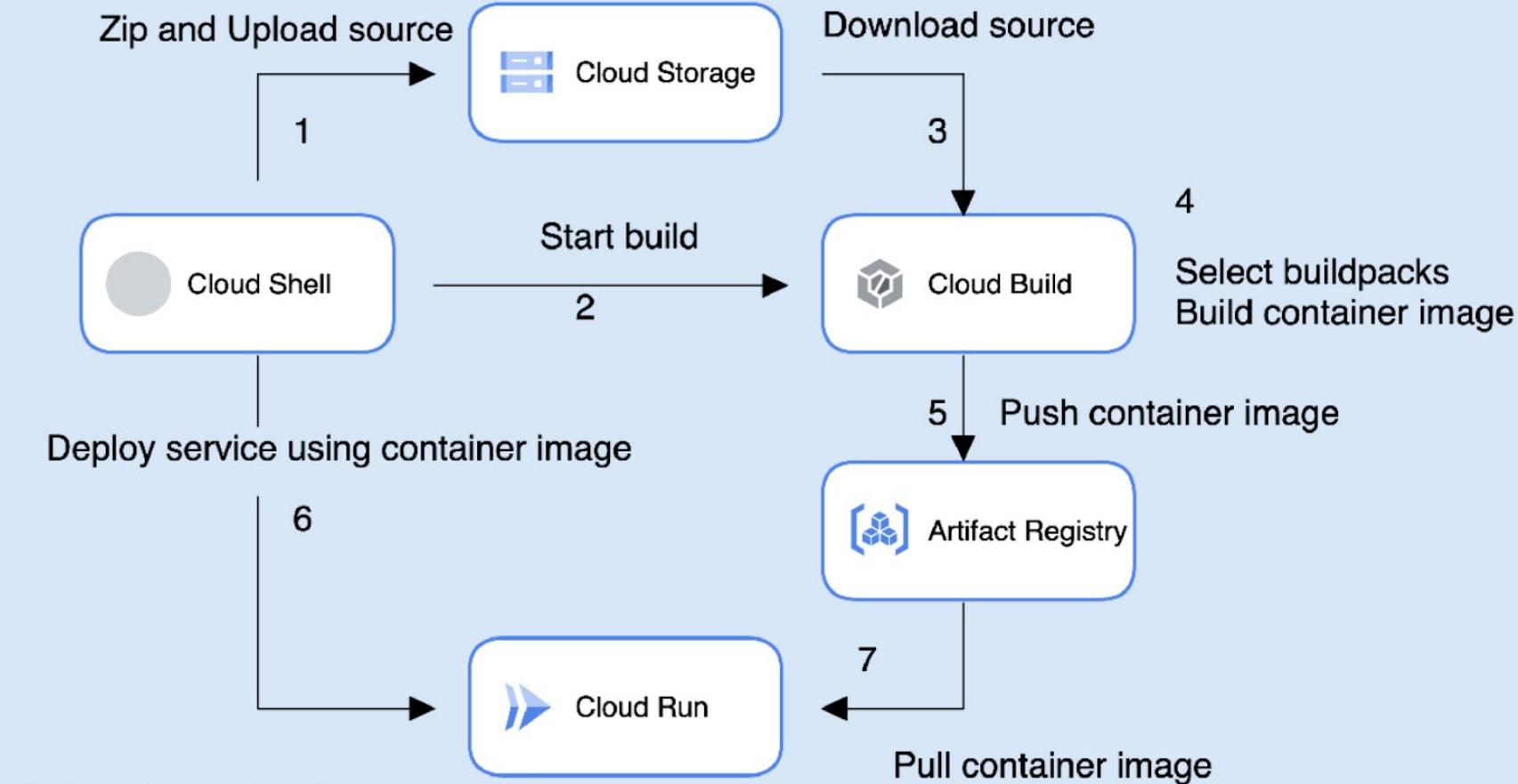


What is Cloud Run?

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.

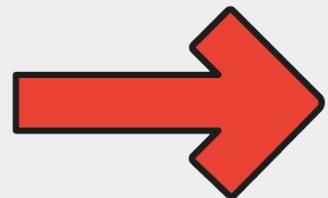


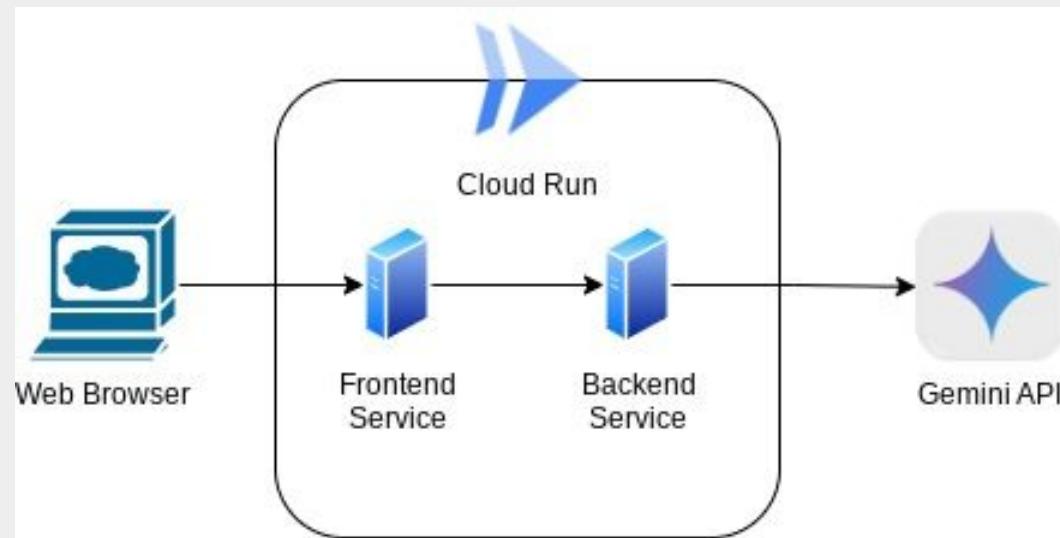


Chapter Two

Let's Demo!

Are you ready?

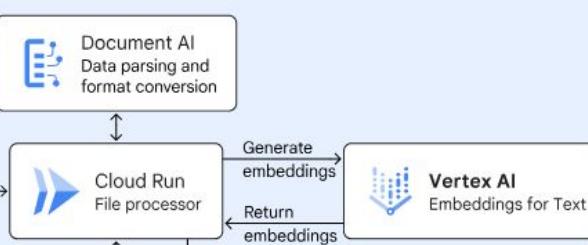




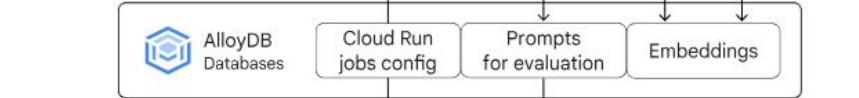
Google Cloud

Region AA

Data ingestion subsystem (for RAG)



Data upload
(RAG data,
evaluation
prompts, etc.)



Evaluation
trigger

Quality evaluation subsystem

Application users

Request Response

Generative AI application

Frontend (example: chatbot)

Backend

Responsible AI
checks

Construct
prompts

Inference
request

Analytics

Observability

LLM
response

Serving subsystem

Vertex AI

LLM (Foundation or custom)



BigQuery



Cloud Logging



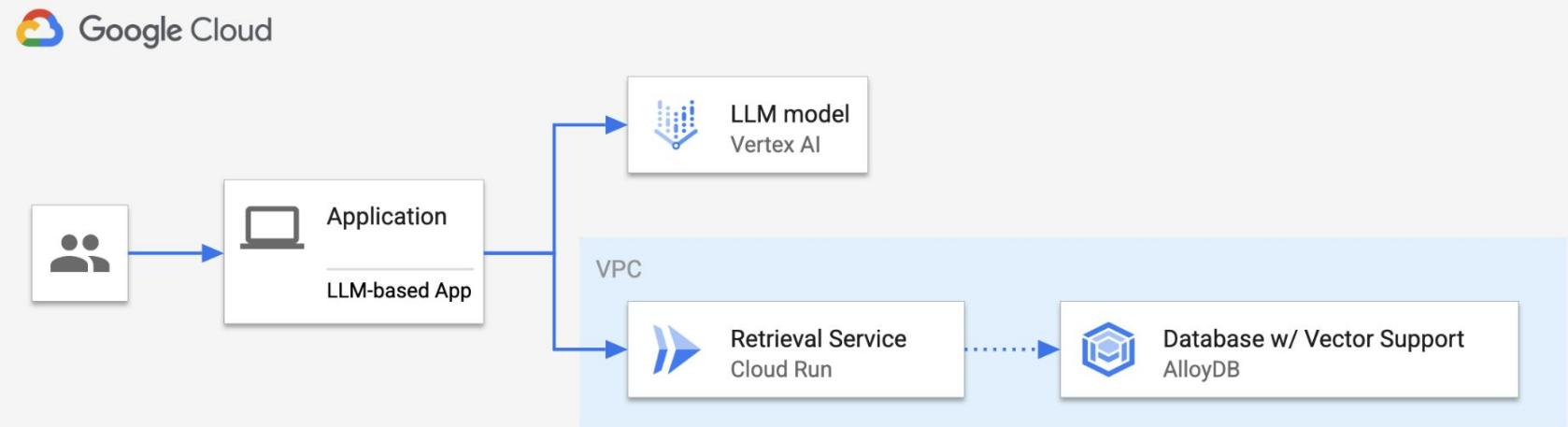
Cloud Monitoring

LLM inference stack

Vertex AI

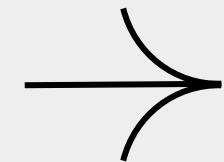
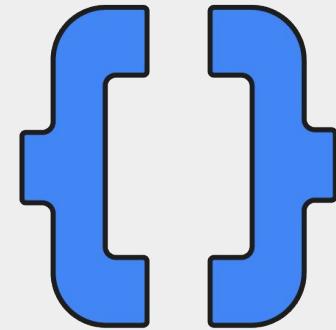
LLM (Foundation or custom)

GenAI Database Retrieval



Prerequisites

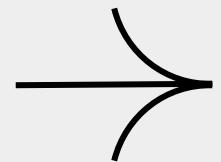
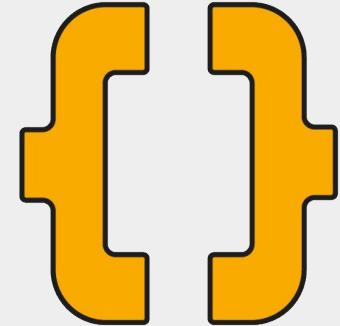
- A basic understanding of Generative AI, Gemini on Google Cloud
- A basic understanding of Cloud Run and Vertex AI Concepts



What you will learn

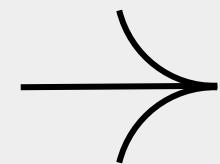
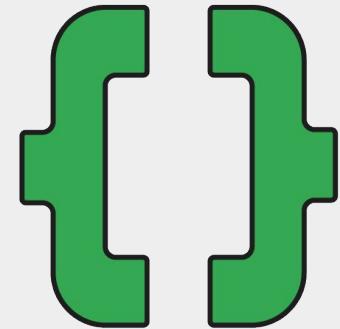
In this lab, you'll learn:

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



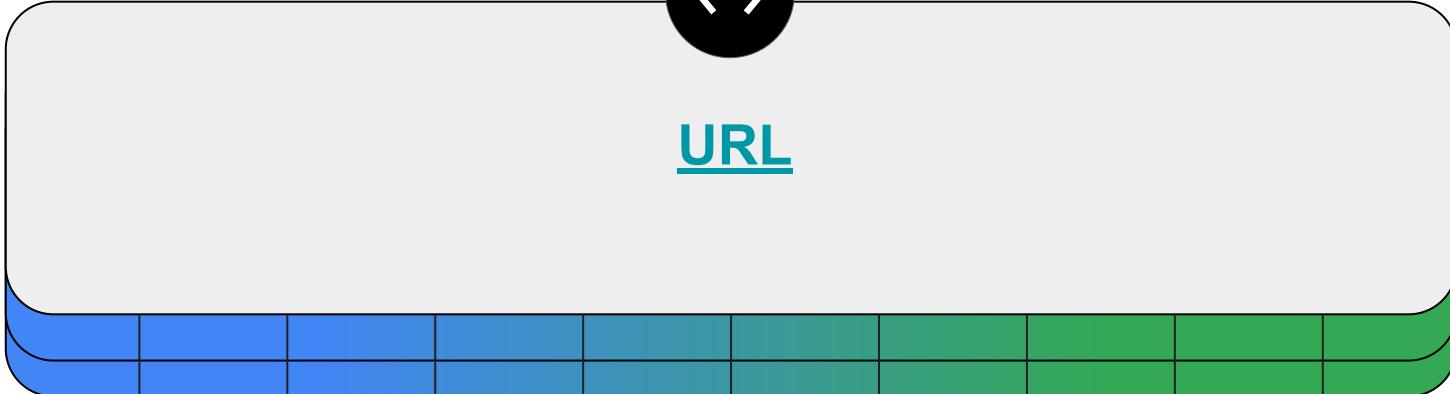
What you will need

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





URL



Open this URL



Google Developer Group

You will need to be signed in

Click on “Sign in with Google” button

Google Cloud



Sign in with Google



Please sign in

Awesome Event

You will need to be signed in

You should see a similar page. Click on this button



[Sign out](#)

Hi, welcome Cloud Developer [\(cloud-dev@gmail.com\)](mailto:cloud-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
JKT-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



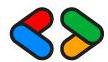
Google Developer Group

Step 1

Claim credits

Step2

Create a
Google Cloud
project

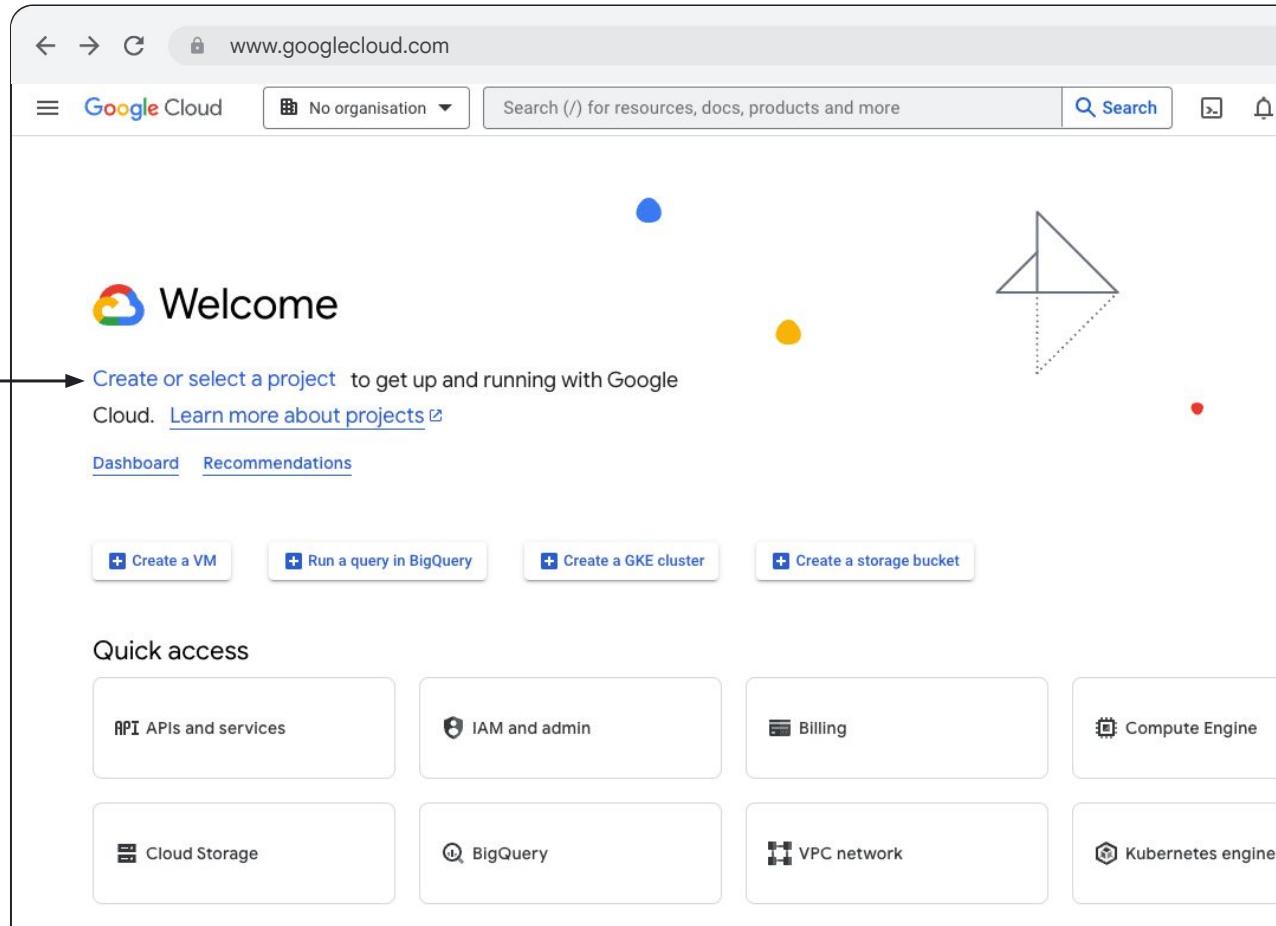


Google Developer Group

Go to:

console.cloud.google.com

Click on “Create or select a project”



The screenshot shows the Google Cloud Welcome page. At the top, there is a navigation bar with a back arrow, forward arrow, a refresh icon, and the URL www.googlecloud.com. The Google Cloud logo is on the left, followed by a dropdown menu for 'No organisation' and a search bar with placeholder text 'Search (/) for resources, docs, products and more'. Below the search bar are two small icons: a document with a checkmark and a magnifying glass.

The main heading is 'Welcome' with a colorful cloud icon to its left. Below it, a call-to-action text reads: 'Create or select a project to get up and running with Google Cloud. [Learn more about projects](#)'. Underneath this, there are two navigation links: 'Dashboard' and 'Recommendations'.

At the bottom of the main content area, there are four buttons with icons and text: '+ Create a VM', '+ Run a query in BigQuery', '+ Create a GKE cluster', and '+ Create a storage bucket'.

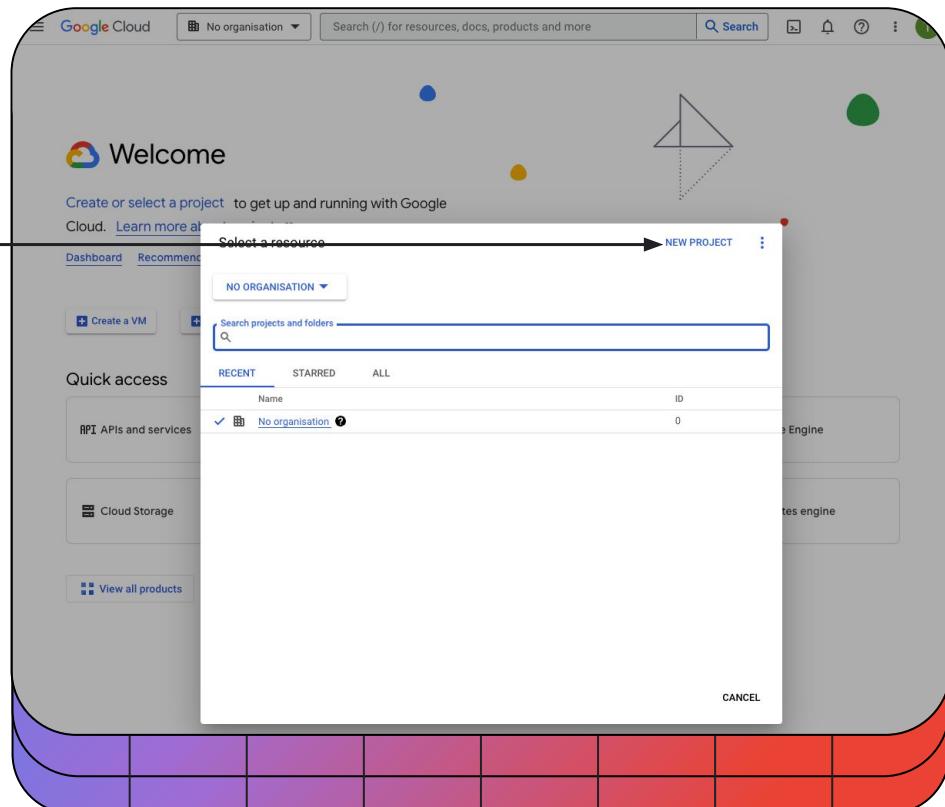
The page features a 'Quick access' section with eight cards arranged in a 2x4 grid. The cards are: 'API APIs and services', 'IAM and admin', 'Billing', 'Compute Engine' (with an icon of a computer monitor); 'Cloud Storage', 'BigQuery', 'VPC network', and 'Kubernetes engine' (with an icon of a hexagon).



Google Developer Group

Creating a new project

Click on “New Project”



Google Developer Group

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 * ?

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

Billing account * ▼

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

Location * 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 *

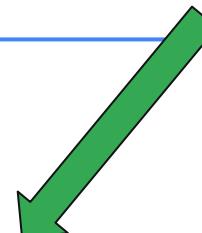
Google Cloud Platform Trial Billing Account



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

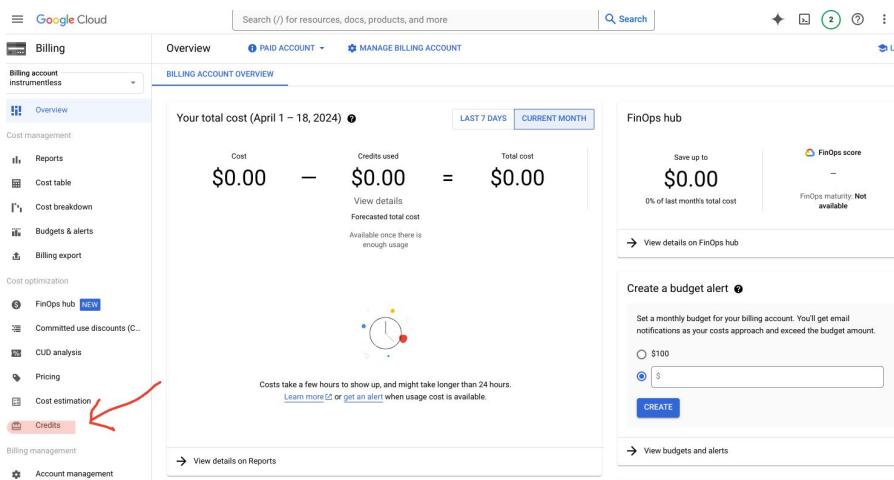
CANCEL

SET ACCOUNT



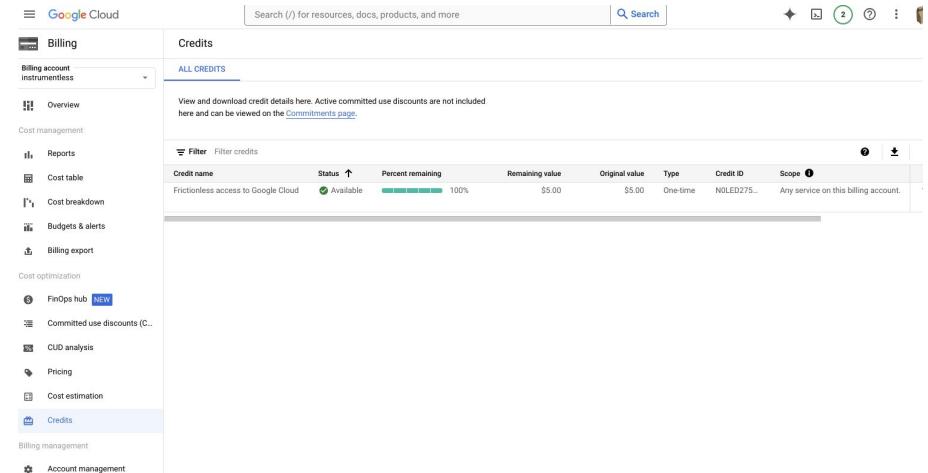
Issues When attempting to redeem credits

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



The screenshot shows the Google Cloud Billing Overview page. The left sidebar includes links for Billing account instrumentless, Overview, Reports, Cost table, Cost breakdown, Budgets & alerts, Billing export, Cost optimization, FinOps hub (NEW), Committed use discounts (CUD), CUD analysis, Pricing, Cost estimation, and Credits. The Credits link is highlighted with a red arrow. The main content area displays the Billing Account Overview with a summary of total cost (\$0.00), credits used (\$0.00), and total cost (\$0.00). It also features a FinOps hub section with a clock icon and a 'Create a budget alert' form. A note at the bottom states: 'Costs take a few hours to show up, and might take longer than 24 hours. Learn more or get an alert when usage cost is available.'

That is where they will see the credits



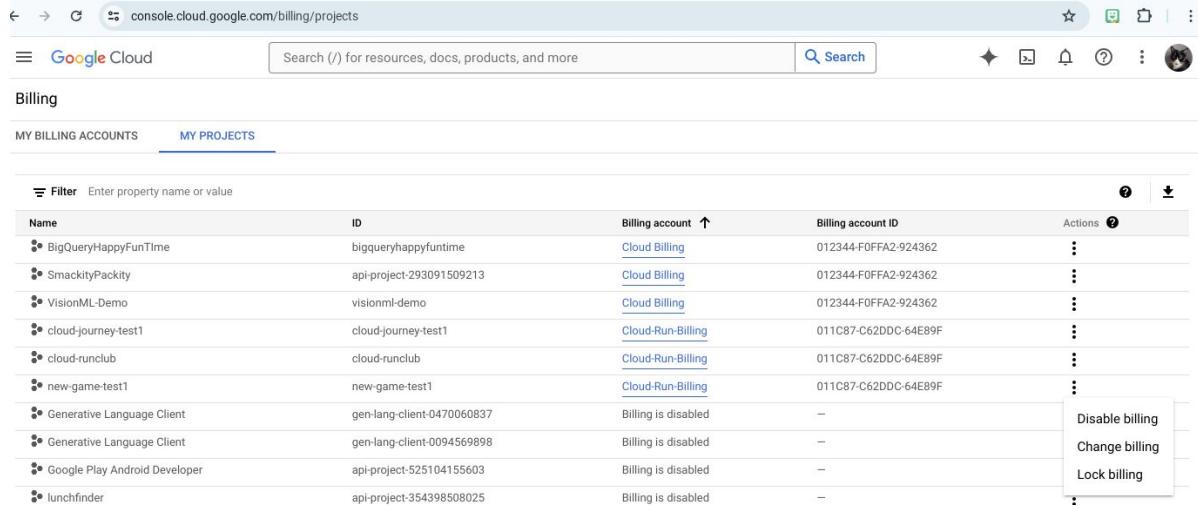
The screenshot shows the Google Cloud Credits page. The left sidebar includes links for Billing account instrumentless, Overview, Reports, Cost table, Cost breakdown, Budgets & alerts, Billing export, Cost optimization, FinOps hub (NEW), Committed use discounts (CUD), CUD analysis, Pricing, Cost estimation, and Credits. The Credits link is highlighted. The main content area displays a table with one credit entry:

Credit name	Status	Percent remaining	Remaining value	Original value	Type	Credit ID	Scope
Frictionless access to Google Cloud	Available	100%	\$5.00	\$5.00	One-time	NOLED27...	Any service on this billing account.

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

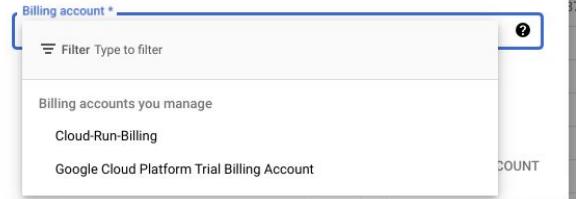


The screenshot shows the Google Cloud Platform Billing interface. At the top, there is a navigation bar with a back arrow, a forward arrow, a refresh icon, and a URL bar that reads "console.cloud.google.com/billing/projects". Below the navigation bar is a search bar with the placeholder "Search (/) for resources, docs, products, and more" and a "Search" button. To the right of the search bar are several icons: a star, a copy icon, a refresh icon, a question mark, and a three-dot menu. The main content area is titled "Billing". It has two tabs: "MY BILLING ACCOUNTS" (which is selected) and "MY PROJECTS". Below the tabs is a "Filter" button and a text input field for filtering by property name or value. The main table lists 12 projects, each with a three-dot menu icon. The columns are: Name, ID, Billing account, Billing account ID, and Actions. The "Actions" column contains three 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"

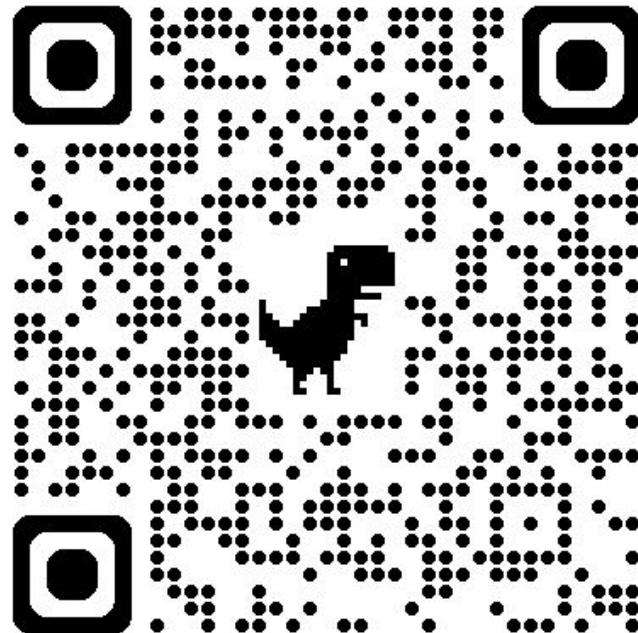
This project pays for both Google Cloud Platform and Maps Platform. Select a billing account that supports both Google Cloud Platform and Maps Platform. [Learn more](#)



The screenshot shows a "Change billing" dialog box. At the top, there is a "Billing account" dropdown menu with a blue outline, a "Filter Type to filter" button, and a question mark icon. Below the dropdown is a list titled "Billing accounts you manage". It contains two items: "Cloud-Run-Billing" and "Google Cloud Platform Trial Billing Account". To the right of the list is a "COUNT" button.

Learn more about Google
Cloud at goo.gle/clouddevs





Google Developer Group

References

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

