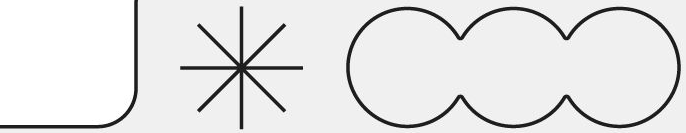
Google Developer Group
GDG Cloud Bandung

Deploying a FastAPI Gemini Chatbot with Memory to Cloud Run

Ananda Dwi Rahmawati

Google Developer Expert - Cloud



Build with AI

Google Developer Group
GDG Cloud Bandung

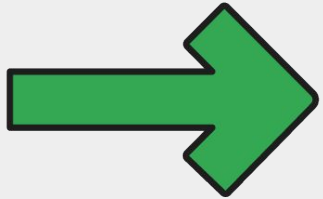
Ananda Dwi Rahmawati

- 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

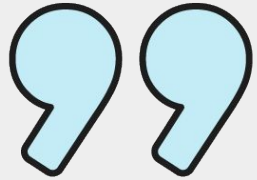
Chapter One

Deploying a FastAPI Gemini Chatbot with Memory to Cloud Run



Overview

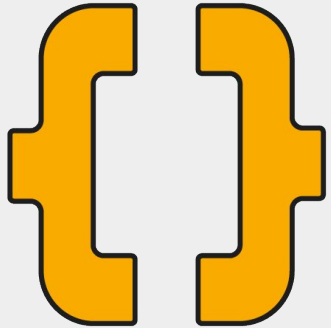
“FastAPI handles your endpoints, Gemini powers your responses, and Cloud Run makes it all effortlessly scalable.”

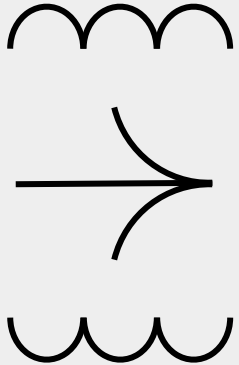


What is FastAPI

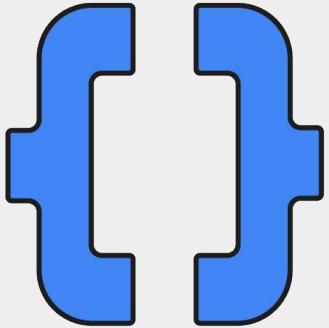
FastAPI is a modern, high-performance web framework for building APIs with Python 3.7+.

- **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.



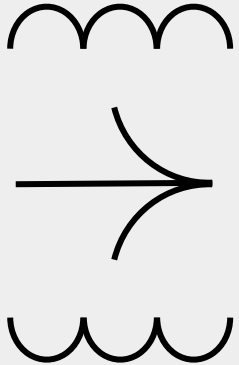


What is Gemini Chatbot?



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

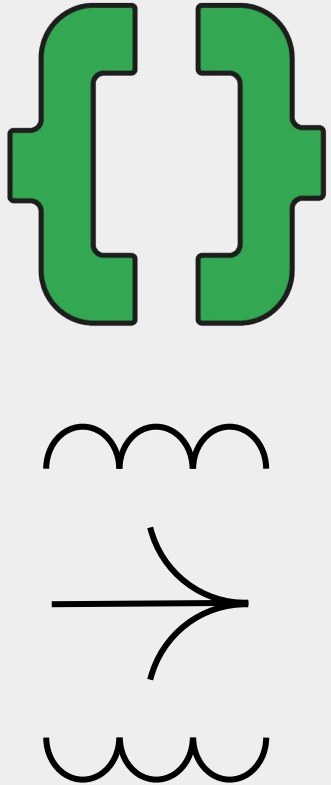
- **Built by Google DeepMind**, integrated into Google Cloud AI products.
- **Gemini Chatbot** uses these models to understand and generate human-like responses.
- **Supports multi-turn conversations**, code understanding, and contextual reasoning.
- Access via Google's Generative AI SDK or Vertex AI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



What is Cloud Run?

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

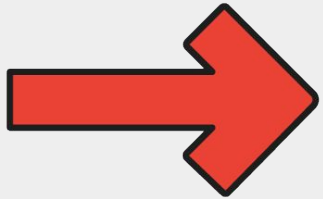
- **Deploy any Docker container**—no server management needed.
- **Scales automatically:** from zero to thousands of instances.
- Pay only when in use.
- **Ideal for microservices**, APIs, or event-driven apps.

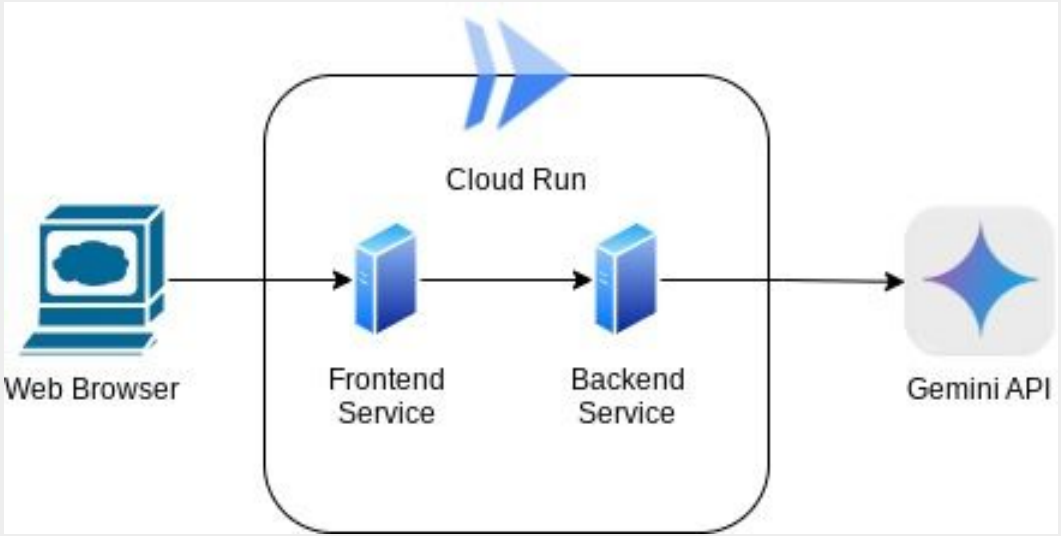


Chapter Two

Let’s Demo!

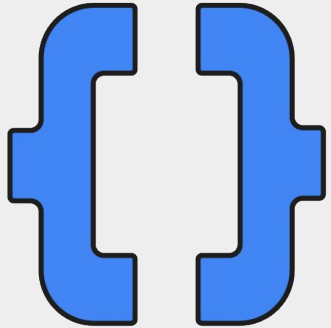
Are you ready?

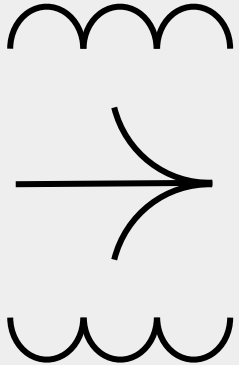



Cloud Run
Web Browser
Frontend Service
Backend Service
Gemini API

Prerequisites

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

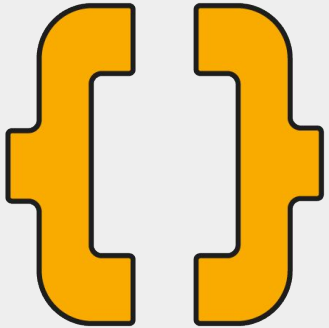


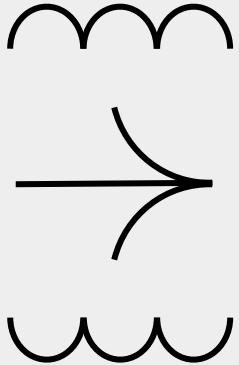


What you will learn

In this lab, you'll learn:

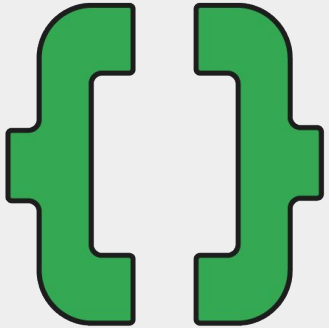
- How to deploy FastAPI to Cloud Run
- Prompt Gemini from Cloud Run in python using a Google client library

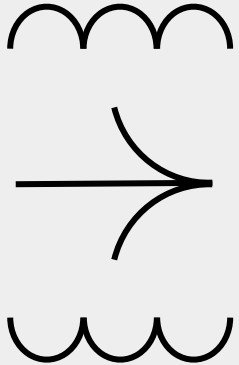


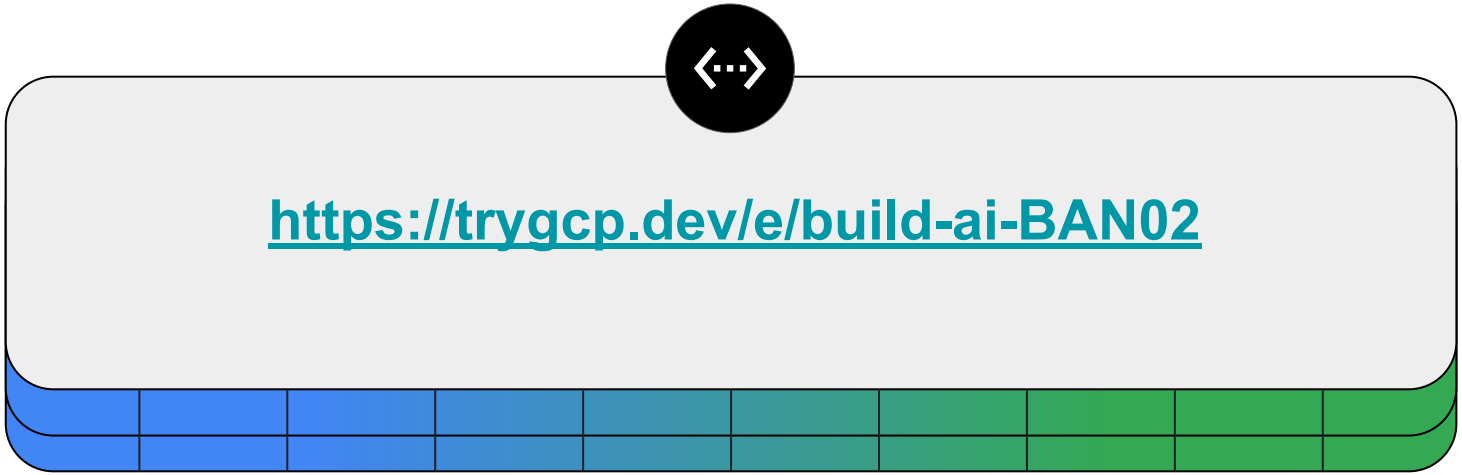


What you will need

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

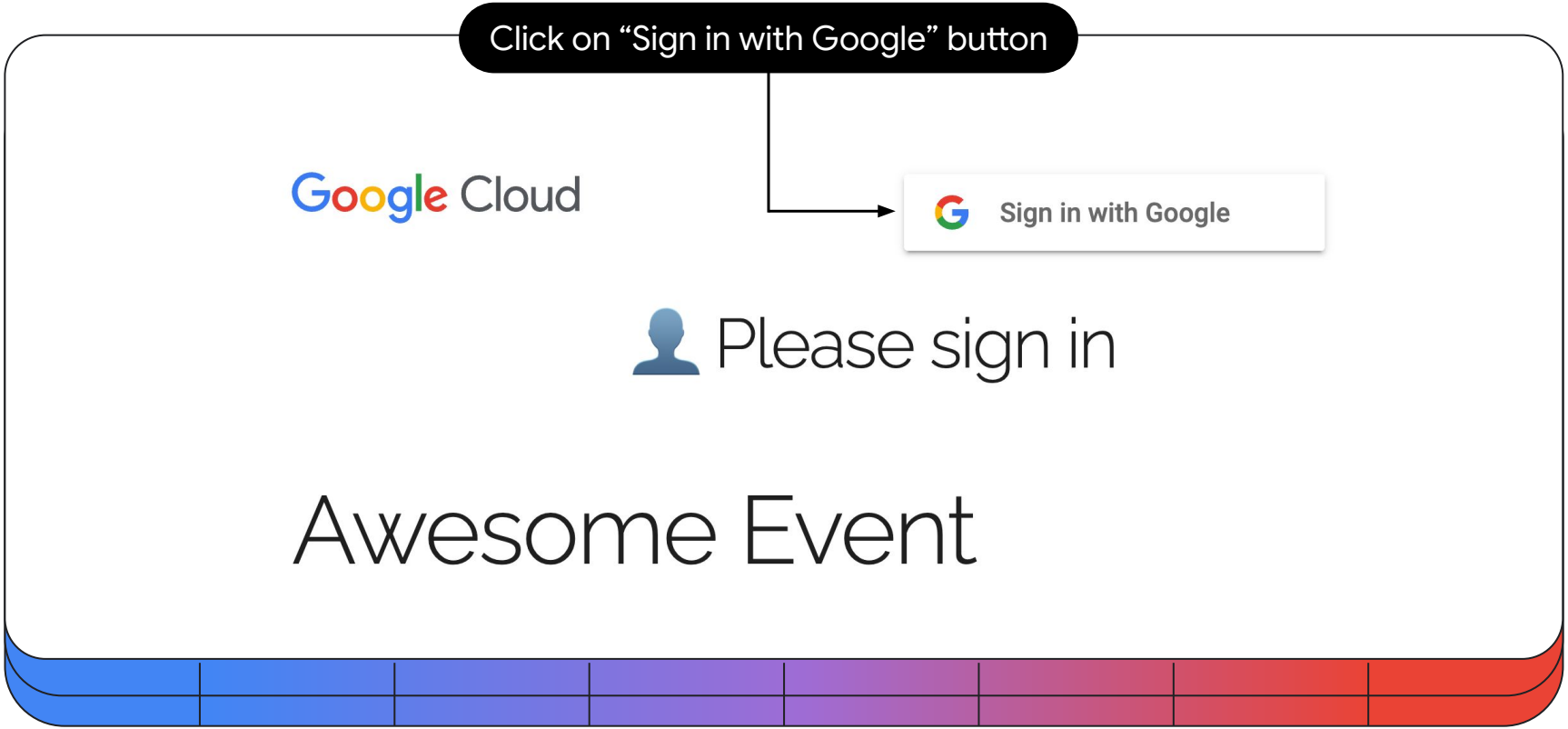




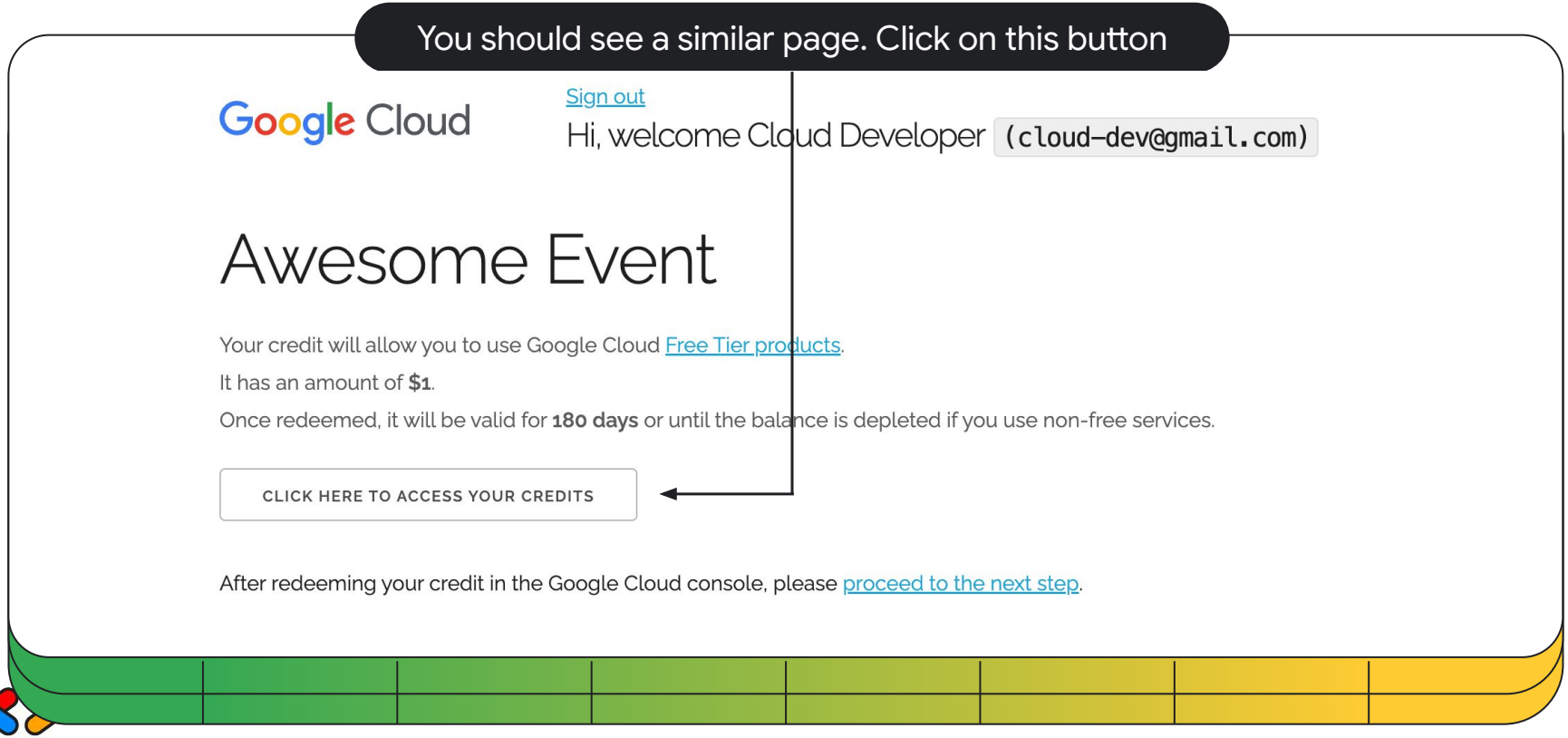


Open this URL

You will need to be signed in

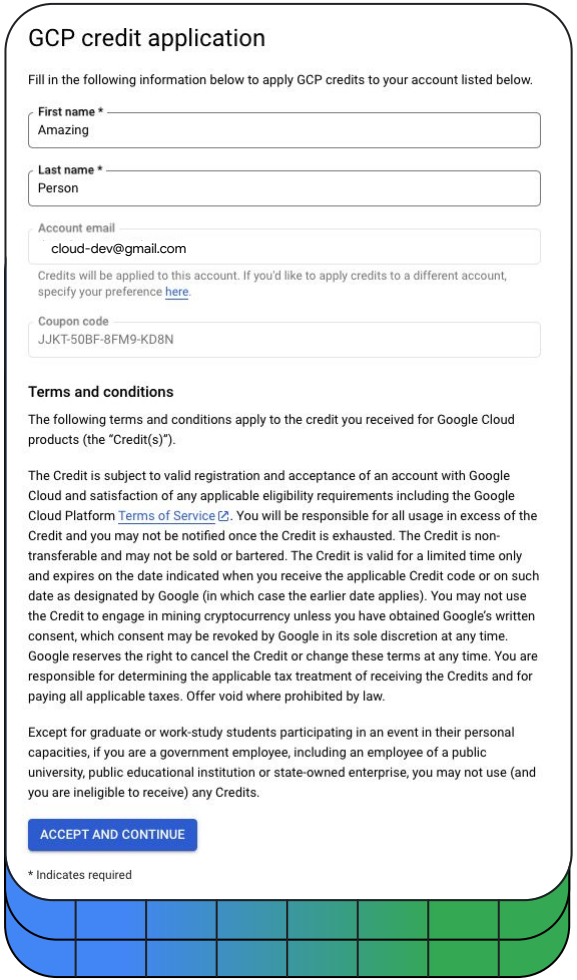


You will need to be signed in



Make sure that you are applying for the correct account

Click on “Accept and Continue” to proceed



Google Developer Group

Step 1

Claim credits

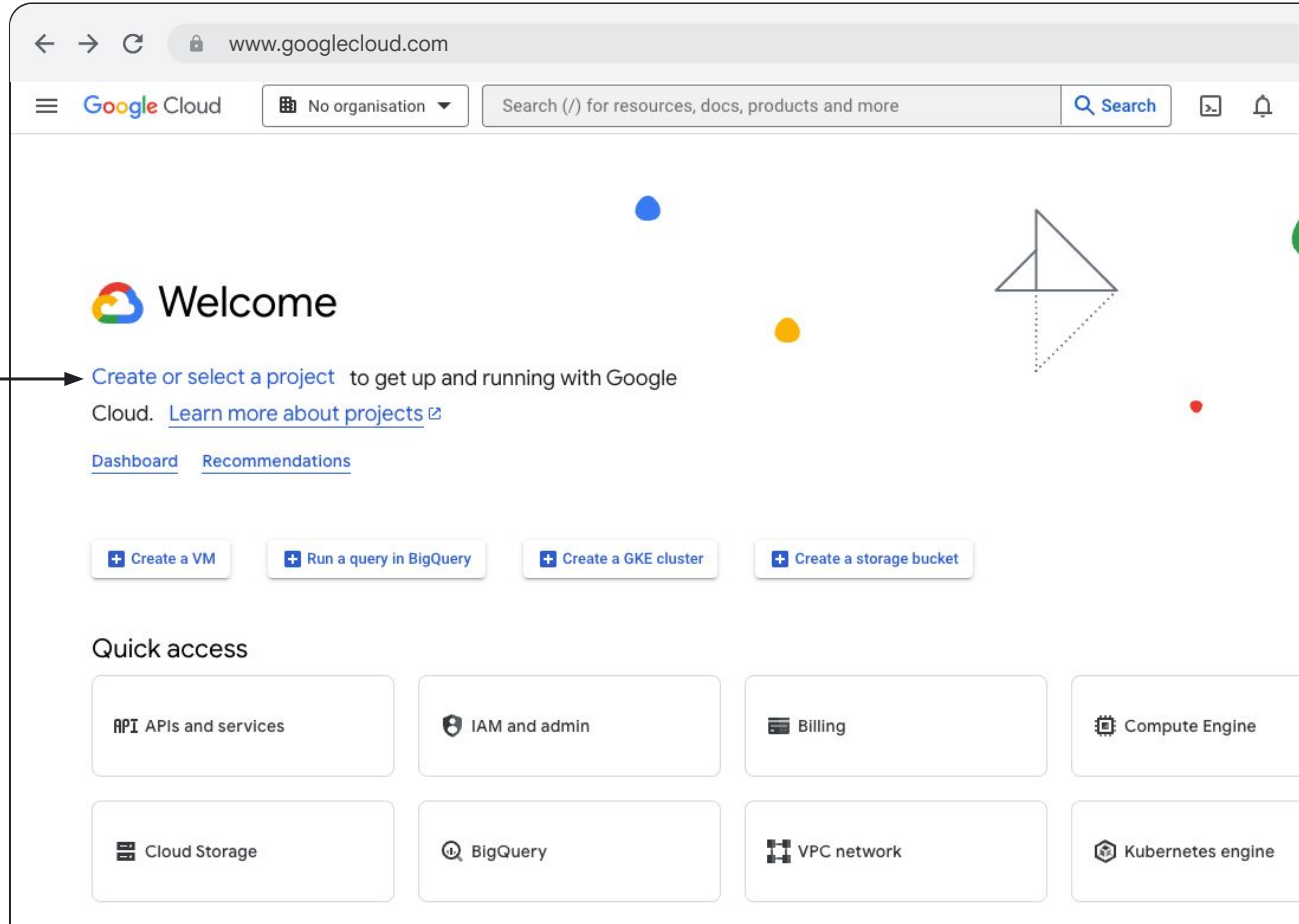
Step2

Create a Google Cloud project

Go to:

console.cloud.google.com

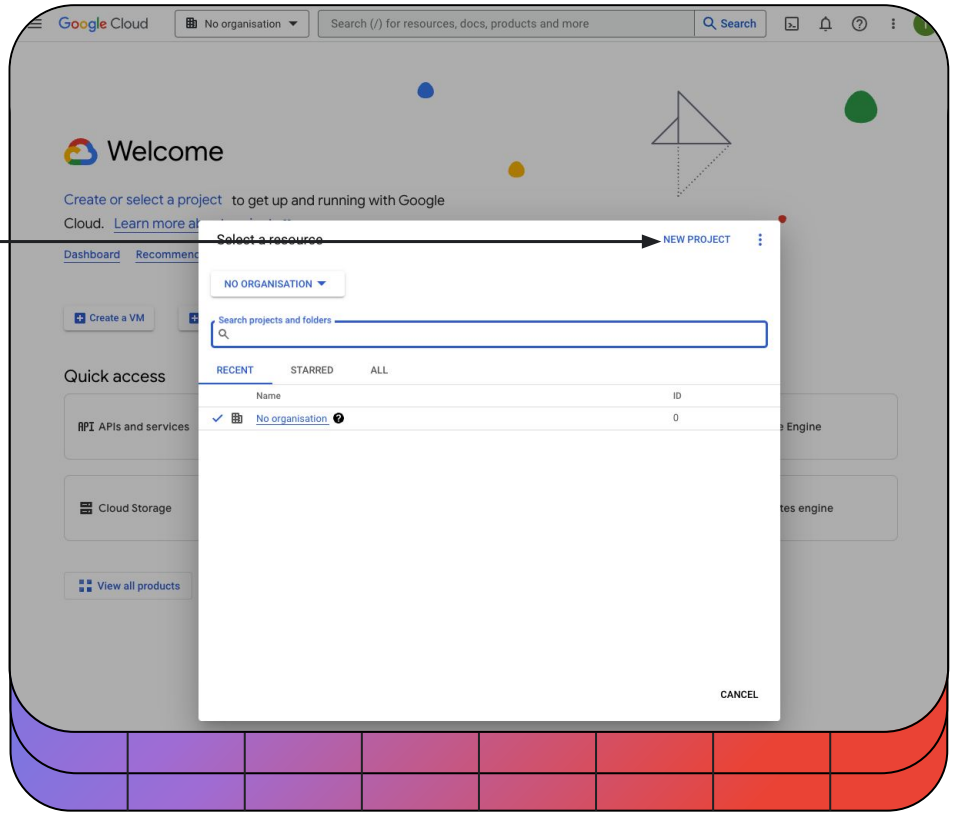
Click on “Create or select a project”



Google Developer Group

Creating a new project

Click on “New Project”



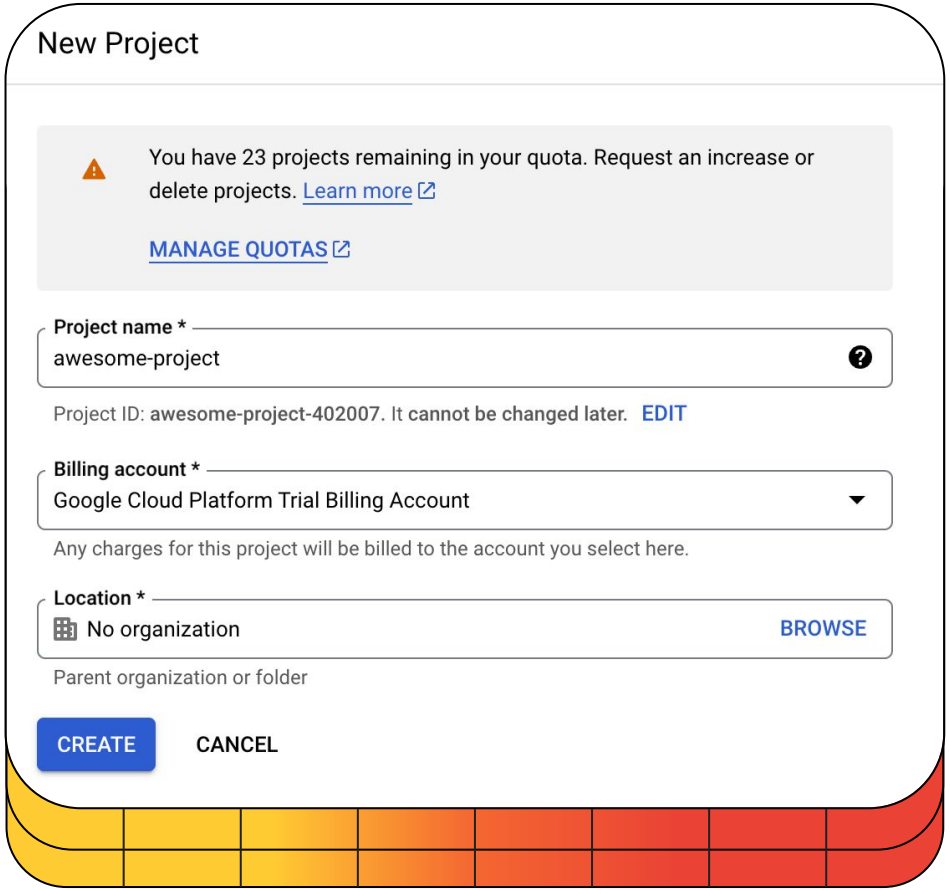
Google Developer Group

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”

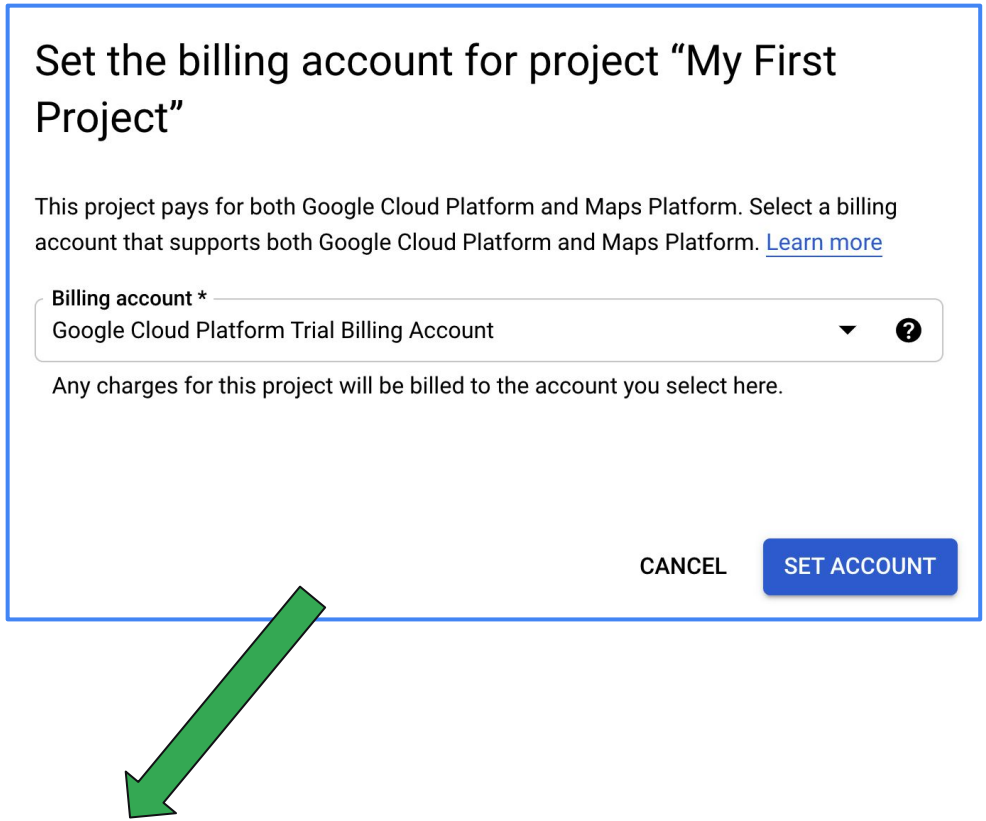


Google Developer Group

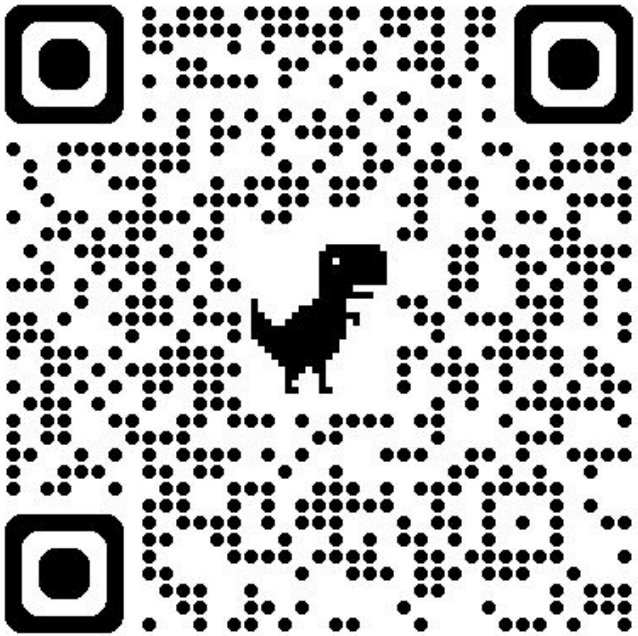
If you don't see a billing account in the previous step:

1) Go to Billing from Google Cloud Console and

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



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