



08

Deploying an Internal Generative AI App

The information in this presentation is classified:

Google confidential & proprietary

⚠ This presentation is shared with you under NDA.

- Do **not** record or take screenshots of this presentation.
- Do **not** share or otherwise distribute the information in this presentation with anyone **inside** or **outside** of your organization.

Thank you!

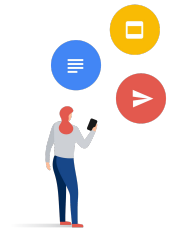


In this module, you learn to ...

- 01 Containerize a basic Generative AI app
- 02 Deploy the app to Cloud Run
- 03 Protect your app behind a load balancer
- 04 Grant access to your app through Identity-Aware Proxy



Case Study: Deploy an internal generative AI tool



You can create

- Create a tool that can summarize PDFs or ask a question that can be answered by its text
- Make the tool accessible to non-technical team members
- Control access to the tool
- Put a CI/CD process in place to easily deploy updates to the tool

What are some options for adding a UI?



Bootstrap Flask

Your engineers know [Flask](#) for Python web apps. [Bootstrap](#) is an opinionated frontend toolkit. [Bootstrap-Flask](#) helps implement Bootstrap within Flask.



Streamlit

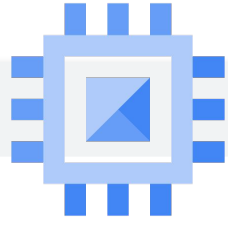
Streamlit allows you to create apps in pure Python, rendering a frontend for you.



Gradio

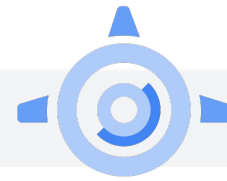
Gradio is a tool that helps you build a dashboard to control an ML model. You can even share a link to access a model that is being served from your local machine.

What are some options for serving your application?



Compute Engine

You could deploy a persistent VM if you expect few users.



App Engine

You could deploy a scalable app that can scale to zero instances.



Cloud Run

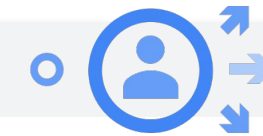
Cloud Run is the latest iteration of Google's serverless application deployments. It allows you to deploy a service with CI/CD built in.

How can you control authentication?



Identity Platform (Firebase Authentication)

You can grant access using a custom UI with passwords unique to the app or SSO with major auth providers.



Identity-Aware Proxy

You can block people from using the app at all unless they authenticate with their Google Identity which will need to have been pre-authorized.

For this app, you'll use the following:



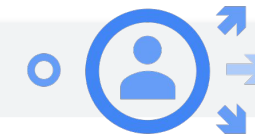
Bootstrap Flask

Your engineers already know Flask and can customize as much as they want.



Cloud Run

You like working from containers and the CI/CD deployment.



Identity-Aware Proxy

You don't want to mess with Firebase.

This is the frontend you will build

TLDR* Everything

*TLDR stands for "Too Long; Didn't Read" and is used on forums to let the reader know that what follows is a summary of a longer post.

Upload a text document in PDF form to get a summary. Or replace the request for a summary in the text field below with a question that might be answered by the document.

Select a PDF

Choose File No file chosen

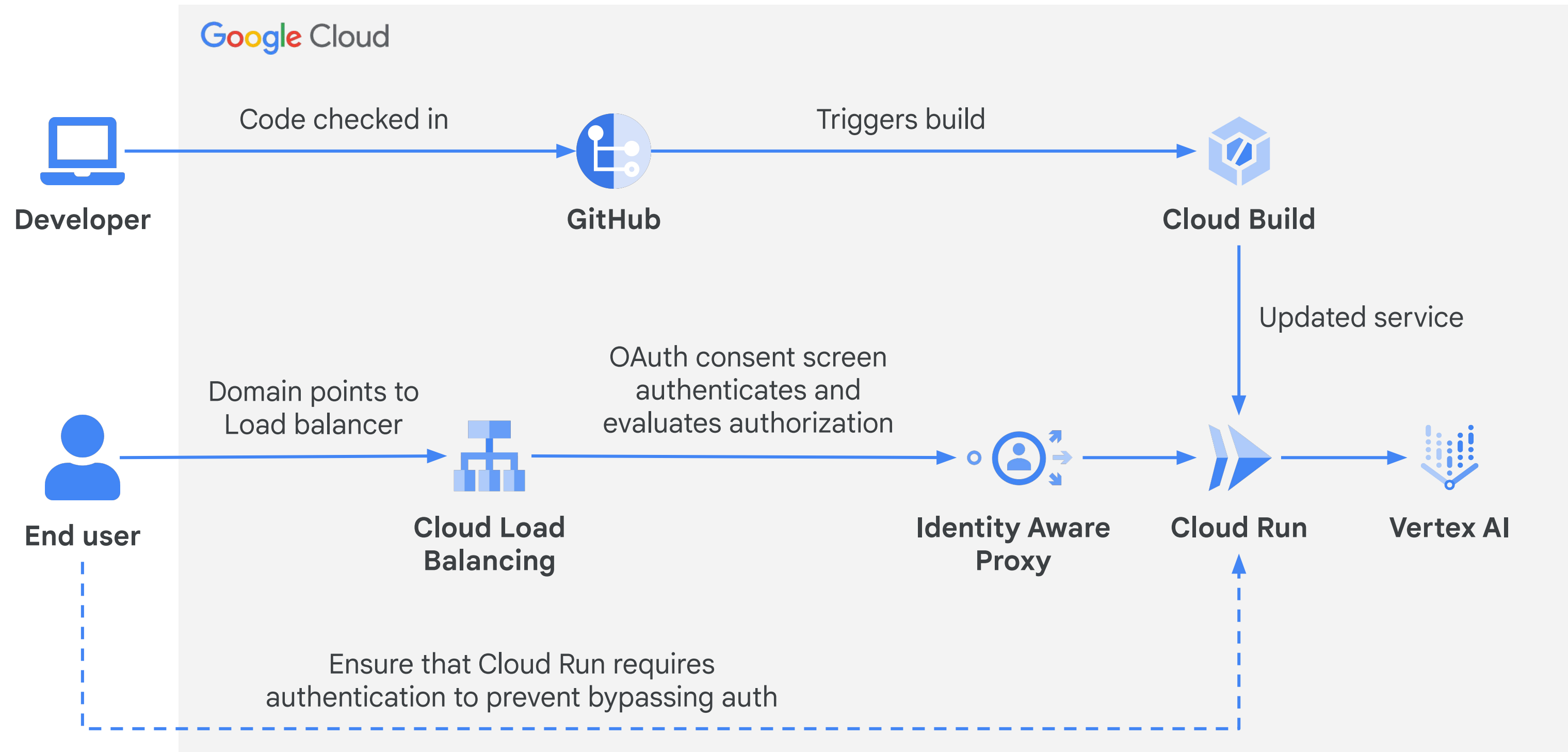
Instructions

Summarize the PDF.

Submit

© 2023 [Google](#)

The architecture allows for CI/CD updates and authenticating with Google Identities



The OAuth Consent screen is what users will be shown when granting the app access to their Google Identity

Edit app registration

1 OAuth consent screen — 2 Scopes — 3 Test users — 4 Summary

App information

This shows in the consent screen, and helps end users know who you are and contact you

App name *

TLDR Everything

The name of the app asking for consent

User support email *

For users to contact you with questions about their consent. [Learn more](#)

App logo

This is your logo. It helps people recognize your app and is displayed on the OAuth consent screen.

After you upload a logo, you will need to submit your app for verification unless the app is configured for internal use only or has a publishing status of "Testing". [Learn more](#)

Logo file to upload

BROWSE

Upload an image, not larger than 1MB on the consent screen that will help users recognize your app. Allowed image formats are JPG, PNG, and BMP. Logos should be square and 120px by 120px for the best results.

App domain

To protect you and your users, Google only allows apps using OAuth to use Authorized Domains. The following information will be shown to your users on the consent

Learn >

How is this info presented to users?

This is the consent screen that users see

1

Sign in with Google

[App Name] wants access to your Google Account

2

Select what [App Name] can access

3

Make sure you trust [App Name]

Cancel

Allow

Google Cloud

Users or whole email domains (i.e. example.com) can be granted access to IAP as IAP-Secured Web App Users

Identity-Aware Proxy

HIDE INFO PANEL

APPLICATIONSSSH AND TCP RESOURCESCONNECTORS

Identity-Aware Proxy (IAP) lets you manage who has access to services hosted on App Engine, Compute Engine, or an HTTPS Load Balancer. [Learn more](#)

To get started with IAP, add an [App Engine app](#), a [Compute Engine instance](#) or configure an [HTTPS Load Balancer](#).

CONNECT NEW APPLICATION

Premium

Filter

Enter property name or value

Resource	IAP	Method	Connection	Published	Status
<input type="checkbox"/> All Web Services					
<input type="checkbox"/> Backend Services					
<input checked="" type="checkbox"/> summary-backend	<input checked="" type="checkbox"/>	IAM	App Connector	Global HTTP(S) Load Balancer: gen-ai-lb	<input checked="" type="checkbox"/> OK

summary-backend

Use external identities for authorization

START

To grant access to the application, click "Add Principal" and select the *IAP-secured Web App User* role. [Learn more](#)

Edit or delete permissions below, or select "Add Principal" to grant new access.

ADD PRINCIPAL

☒ Show inherited permissions

Filter

Enter property name or value

Role / Principal	Inheritance
▶ Editor (4)	
▼ IAP-secured Web App User (2)	
<div><div></div><div></div></div>	<div><div></div><div></div></div>
▶ Owner (3)	
▶ Viewer (1)	

Google Cloud

The IAP service account (which takes some time to be created) must be given the role of “Cloud Run Invoker”

service-[PROJECT-NUMBER]@gcp-sa-iap.iam.gserviceaccount.com

Add principals

Principals are users, groups, domains, or service accounts. [Learn more about principals in IAM](#)

New principals *

service-[REDACTED]@gcp-sa-iap.iam.gserviceaccount.com

?

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

Role *

Cloud Run Invoker

Can invoke a Cloud Run service.

IAM condition (optional) ?

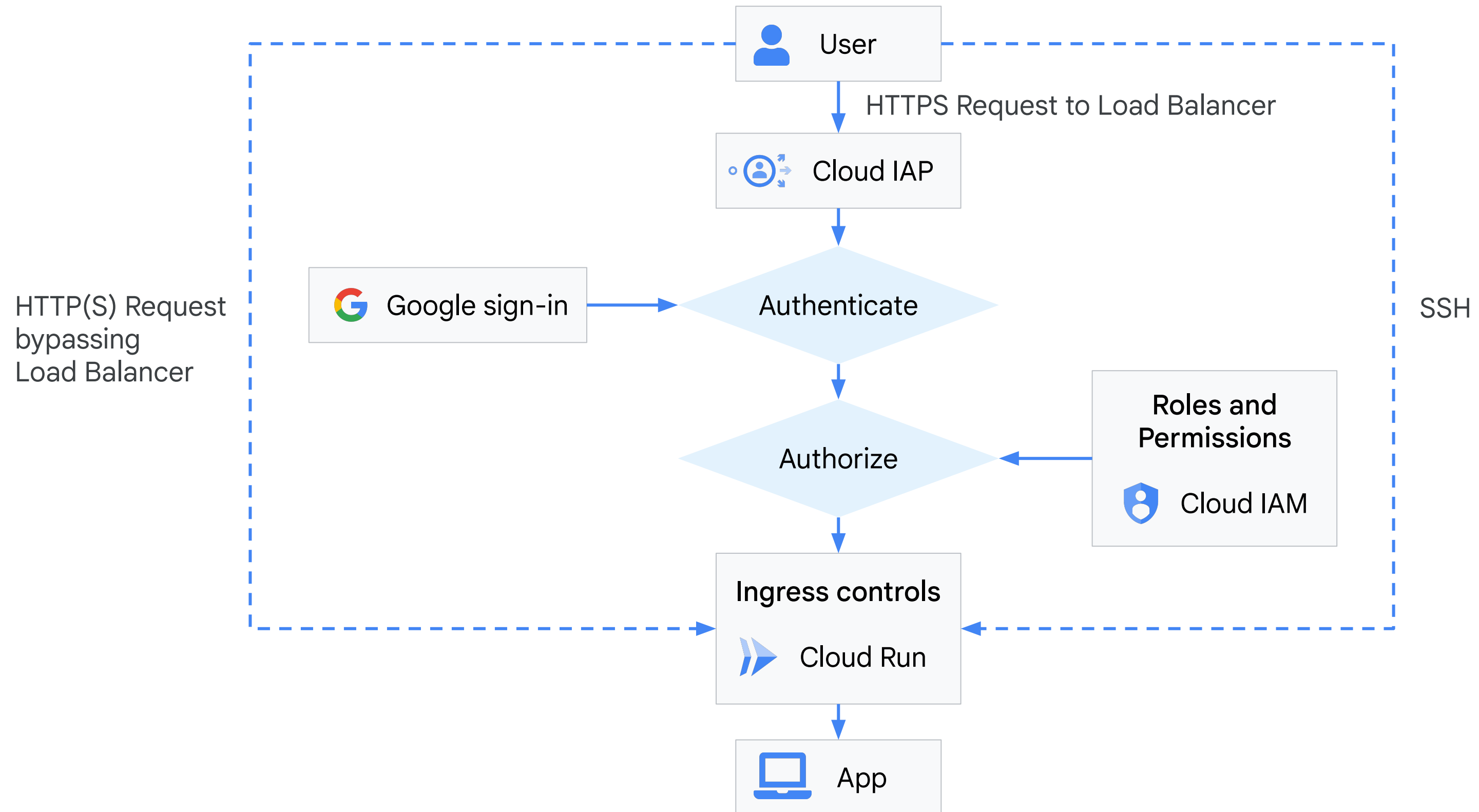
+ ADD IAM CONDITION

+ ADD ANOTHER ROLE

SAVE

CANCEL

How IAP works for a Cloud Run app



Now your users can authenticate and grant access through the OAuth Consent Screen to access the app

TLDR* Everything

*TLDR stands for "Too Long; Didn't Read" and is used on forums to let the reader know that what follows is a summary of a longer post.

Upload a text document in PDF form to get a summary. Or replace the request for a summary in the text field below with a question that might be answered by the document.

Select a PDF

Choose File No file chosen

Instructions

Summarize the PDF.

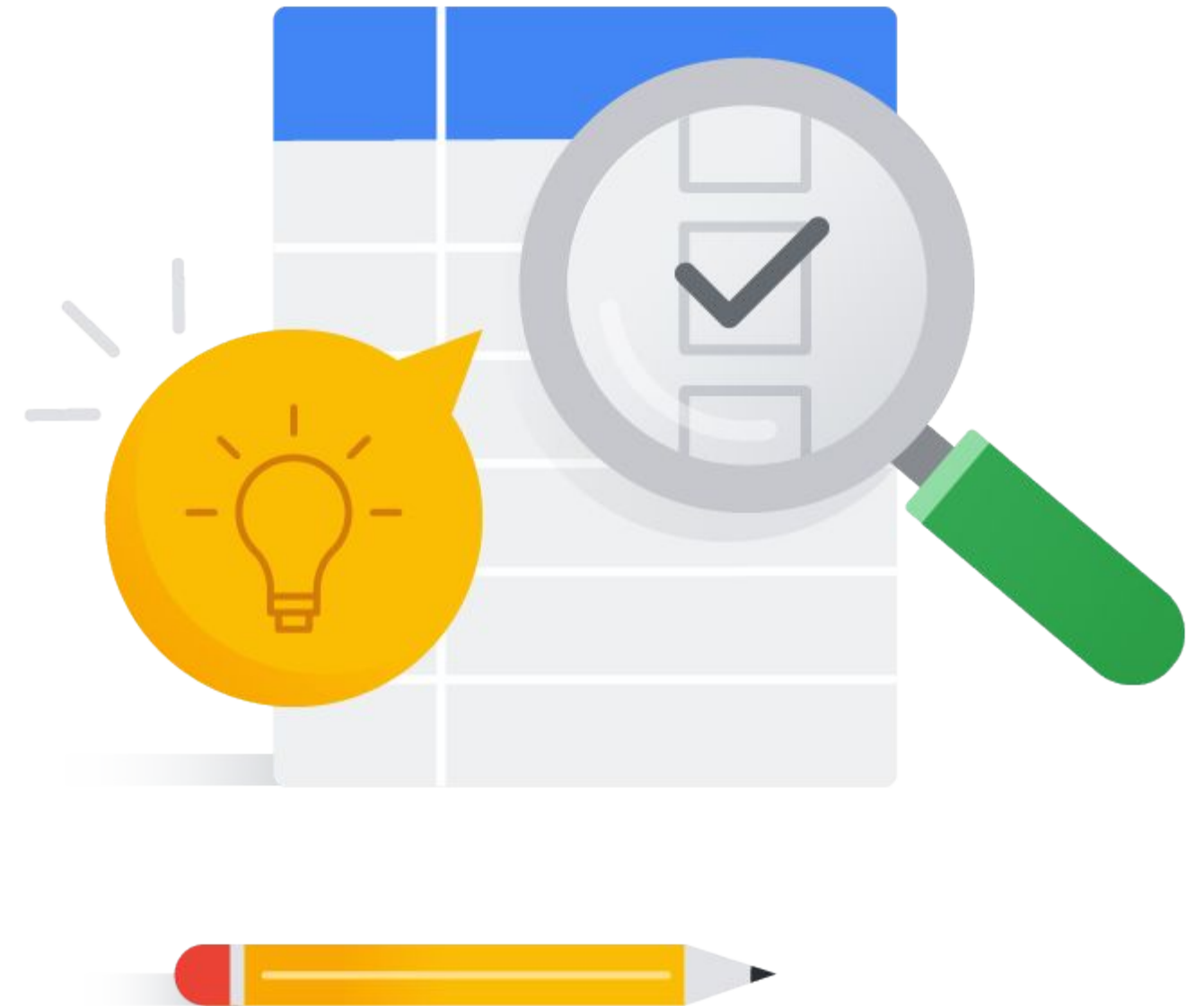
Submit

© 2023 [Google](#)

Lab

🕒 2 hours 🧑‍🔧

Lab: Deploy and Secure a Gen AI Web Application

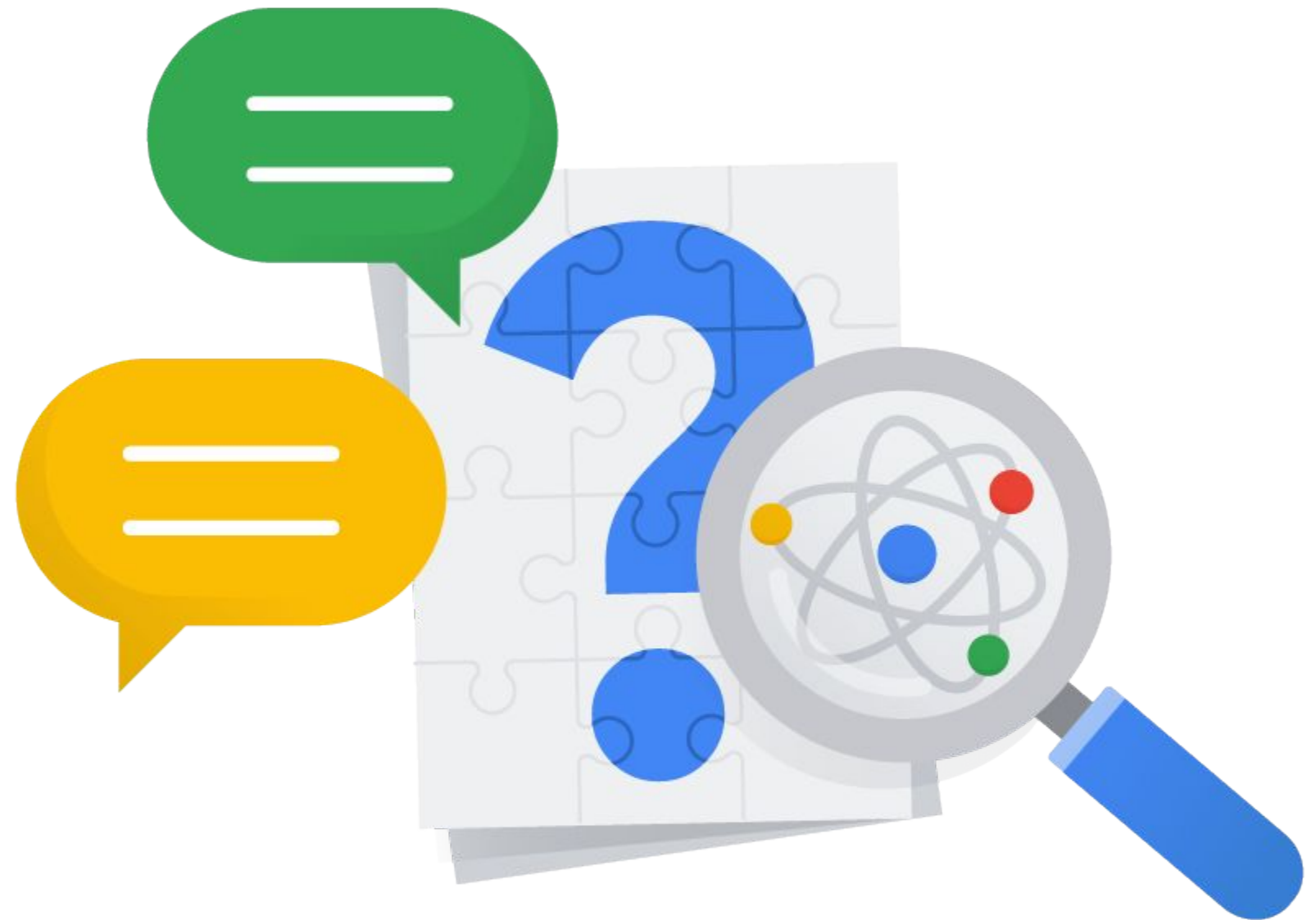


In this module, you learned to ...

- 01 Containerize a basic Generative AI app
- 02 Deploy the app to Cloud Run
- 03 Protect your app behind a load balancer
- 04 Grant access to your app through Identity-Aware Proxy



Questions and answers



Google Cloud