AULA 3: FAZENDO DEPLOY NA GOOGLE CLOUD PLATFORM (GCP)

O objetivo dessa aula é colocar nosso projeto no cloud da Google para que qualquer pessoa possa ter acesso através de um link, ou seja, realizar o deploy da aplicação.

O que é Deploy?

Deploy significa "colocar algo em produção". É o momento em que levamos nosso projeto (API, site, app) para um ambiente online, onde outras pessoas podem acessá-lo, seja em um servidor, nuvem ou serviço web.

Introdução à Google Cloud Platform (GCP)

A GCP é a plataforma de nuvem da Google. Nela, podemos:

- Criar servidores
- Armazenar imagens Docker
- Hospedar APIs e aplicações
- Monitorar desempenho e uso

Para isso, usamos o **Artifact Registry** e o **Cloud Run**.

O que é o Artifact Registry?

O **Artifact Registry** é o "repositório de imagens" do Google Cloud. Serve para armazenar suas imagens Docker com segurança e controle, e usá-las em diferentes projetos.

O que é o Cloud Run?

O **Cloud Run** é um serviço **serverless** da Google. Isso significa que você não precisa configurar servidores ou infraestrutura (o Google cuida disso automaticamente).

> Você só envia sua imagem Docker, define a porta, e ele faz tudo: deploy, escalabilidade, segurança, link público, etc.

Comparando com um ambiente tradicional

Ambiente tradicional	Serverless
Você escolhe um servidor (VPS)	O provedor escolhe automaticamente
Precisa configurar o sistema	Tudo já vem pronto e configurado
Paga mesmo sem usar	Só paga quando o código está rodando
Escalabilidade manual	Escala automaticamente conforme o uso

Exemplo prático com Cloud Run

Você criou uma API FastAPI. Quando você faz deploy com o **Cloud Run** (serverless):

- O Google provisiona um container só quando alguém acessa a API
- Ele escala automaticamente: se 1 pessoa acessar, roda 1 instância; se
 1.000 pessoas acessarem, ele escala pra aguentar
- E você só paga pelos segundos em que o código está ativo

Passos da aula 2:

- Fazer login na plataforma do Google Cloud
- Criar um projeto (nome dado: alura-imersao) e entrar nele
- Instalar a CLI gcloud: https://cloud.google.com/sdk/docs/install?hl=pt-br
- Reiniciar o VSCODE

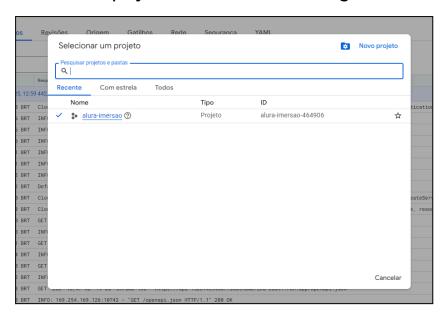
- Iniciar a sequência de comandos no terminal:
 - 1. fazer o login:

gcloud auth login

2. escolher o projeto:

gcloud config set project PROJECT_ID

para saber o ID do projeto é só clicar nele no Google Cloud:



Portanto:

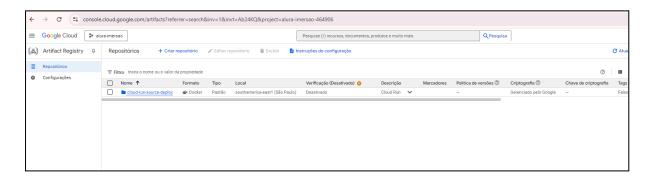
gcloud config set project alura-imersao-464906

3. Usar o cloud run: faz o deploy da aplicação, num ambiente serverless

gcloud run deploy --port=8000

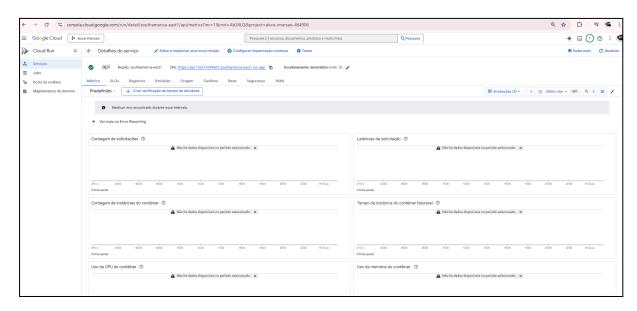
ocorre uma sequência de passos no terminal, a primeira é escolher a pasta para o deploy(se for a mesma sugerida é só dar enter, se não, colocar o caminho da pasta). O segundo é escolher um nome para o serviço (foi usado api) e em seguida dar "yes" e enter. Em seguida escolher a região onde existe Google Cloud, para ter menor latência, buscar o que estiver mais próximo, no nosso caso é usado o 32 (southamerica-east1, localizado em São Paulo). É só digitar 32 no terminal e dar enter.

Em seguida, ele pergunta se pode criar uma imagem no repositório da Google, o Artifact Registry, dar "yes". Adiante, dar "yes" para dar permissão a acesso público da aplicação. O processo inicia e é possível ver a construção do deploy no terminal ou no próprio google cloud, basta dar um search lá por "artifact registry":



É possível ver ali o tamanho da imagem e quando foi criada, o histórico fica gravado. Também é interessante colocar no search "cloud run" para ver várias métricas pertinentes disponíveis para análise:





Para acessar a aplicação, basta acessar o link fornecido no terminal e no caso do nosso projeto:

api-152614394607.southamerica-east1.run.app/docs

