# MapReduce no Dataflow (Python)

## visão global

Neste laboratório, você identificará as operações Mapear e Reduzir, executar o pipeline, usar os parâmetros da linha de comando.

## Objetivo

* Identificar operações de mapeamento e redução
* Execute o pipeline
* Use os parâmetros da linha de comando

## Configuração

#### O que você precisará

Para concluir este laboratório, você precisará de:

* Acesso a um navegador de internet padrão (navegador Chrome recomendado).
* Tempo. Observe o tempo de **conclusão** do laboratório no Qwiklabs. Essa é uma estimativa do tempo necessário para concluir todas as etapas. Planeje sua programação para que você tenha tempo de concluir o laboratório. Depois de iniciar o laboratório, você não poderá pausar e retornar mais tarde (começará na etapa 1 toda vez que iniciar um laboratório).
* O tempo de **acesso** do laboratório é o tempo que seus recursos de laboratório estarão disponíveis. Se você concluir seu laboratório com o tempo de acesso ainda disponível, poderá explorar o Google Cloud Platform ou trabalhar em qualquer seção do laboratório marcada "se tiver tempo". Quando o tempo de acesso acabar, seu laboratório terminará e todos os recursos serão encerrados.
* Você **NÃO** precisa de uma conta ou projeto do Google Cloud Platform. Uma conta, projeto e recursos associados são fornecidos a você como parte deste laboratório.
* Se você já tiver sua própria conta do GCP, não a use para este laboratório.
* Se o seu laboratório solicitar que você faça login no console, **use apenas a conta do aluno fornecida pelo laboratório** . Isso impede que você incorra em cobranças por atividades de laboratório em sua conta pessoal do GCP.

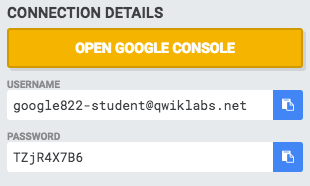
#### Comece seu laboratório

Quando estiver pronto, clique em **Iniciar Laboratório** . Você pode acompanhar o progresso do seu laboratório com a barra de status na parte superior da tela.

**Importante** O que está acontecendo durante esse tempo? Seu laboratório está gerando recursos do GCP para você nos bastidores, incluindo uma conta, um projeto, recursos dentro do projeto e permissão para você controlar os recursos necessários para executar o laboratório. Isso significa que, em vez de gastar tempo manualmente configurando um projeto e construindo recursos do zero como parte de seu laboratório, você pode começar a aprender mais rapidamente.

#### Encontre o nome de usuário e a senha do GCP do seu laboratório

Para acessar os recursos e o console deste laboratório, localize o painel Detalhes da Conexão no Qwiklabs. Aqui você encontrará o ID da conta e a senha da conta que você usará para fazer login no Google Cloud Platform:



Se o seu laboratório fornecer outros identificadores de recursos ou informações relacionadas à conexão, ele também aparecerá nesse painel.

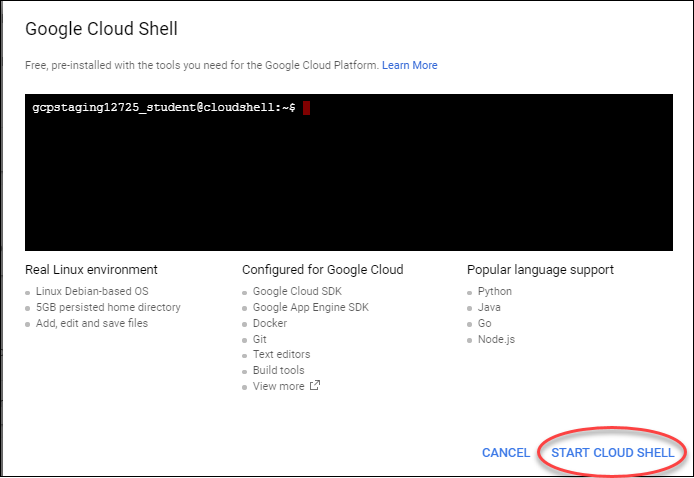
### Ativar o Google Cloud Shell

O Google Cloud Shell fornece acesso via linha de comando aos seus recursos do GCP.

No console do GCP, clique no ícone do **Cloud Shell** na barra de ferramentas superior direita:

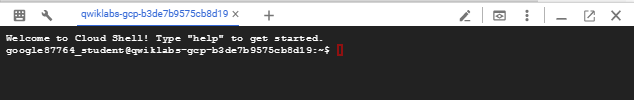


Em seguida, clique em **START CLOUD SHELL** :



Você pode clicar em **START CLOUD SHELL** imediatamente quando a caixa de diálogo aparecer, em vez de esperar na caixa de diálogo até as disposições do Cloud Shell.

Demora alguns momentos para provisionar e se conecta ao ambiente:



O Cloud Shell é uma máquina virtual carregada com todas as ferramentas de desenvolvimento de que você precisa. Ele oferece um diretório pessoal de 5 GB persistente e é executado no Google Cloud, aprimorando muito o desempenho e a autenticação da rede.

Uma vez conectado ao shell da nuvem, você verá que já está autenticado e o projeto está definido como seu PROJECT\_ID :

gcloud auth list

Saída:

Credentialed accounts:

- <myaccount>@<mydomain>.com (active)

**Observação:** gcloud é a ferramenta de linha de comando poderosa e unificada do Google Cloud Platform. A documentação completa está disponível no [Google Cloud gcloud Overview](https://cloud.google.com/sdk/gcloud) . Ele vem pré-instalado no Cloud Shell e suporta a conclusão de guias.

gcloud config list project

Saída:

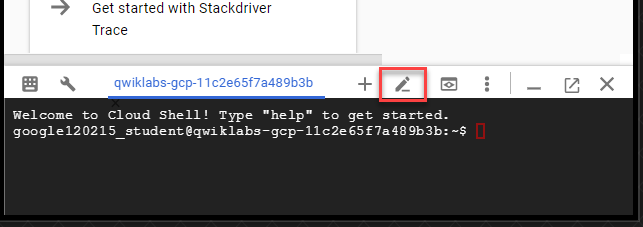
[core]

project = <PROJECT\_ID>

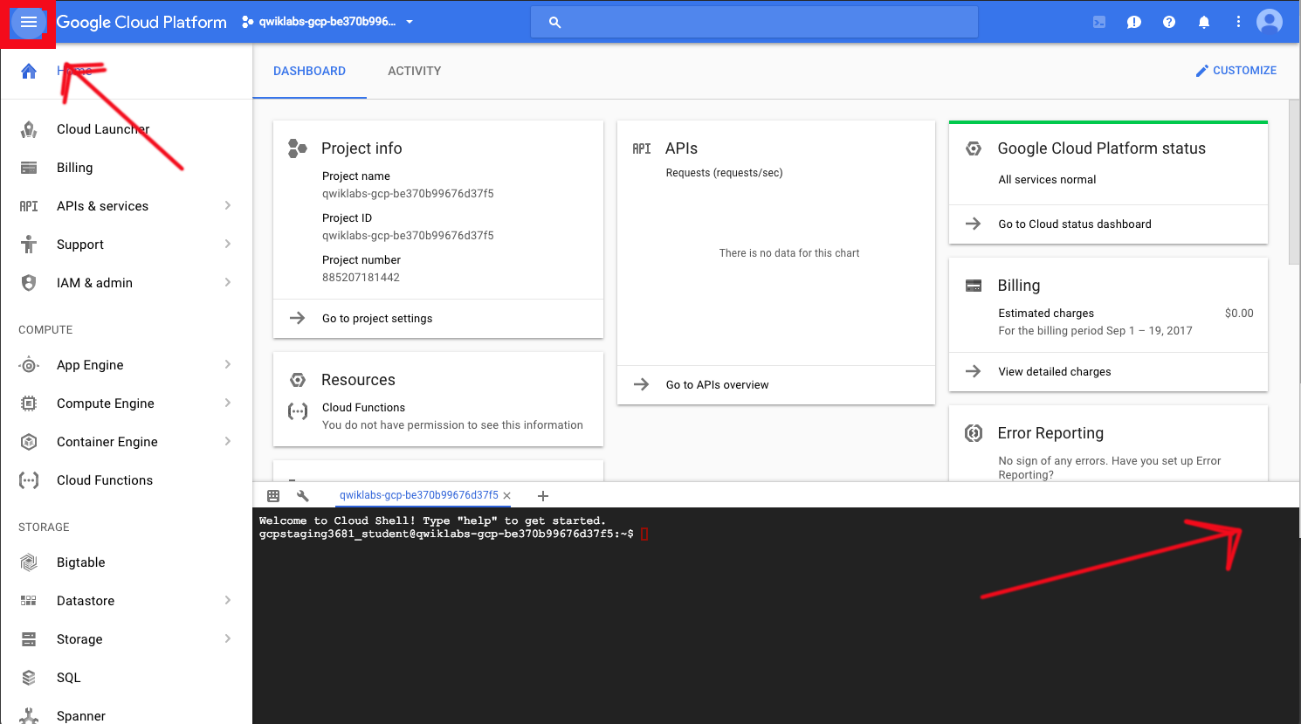
### Inicie o Editor de códigos do Google Cloud Shell

Use o Editor de códigos do Google Cloud Shell para criar e editar facilmente diretórios e arquivos na instância do Cloud Shell.

Depois de ativar o Google Cloud Shell, clique no botão **Iniciar editor de códigos**(parece um lápis) para abrir o Cloud Shell Code Editor.

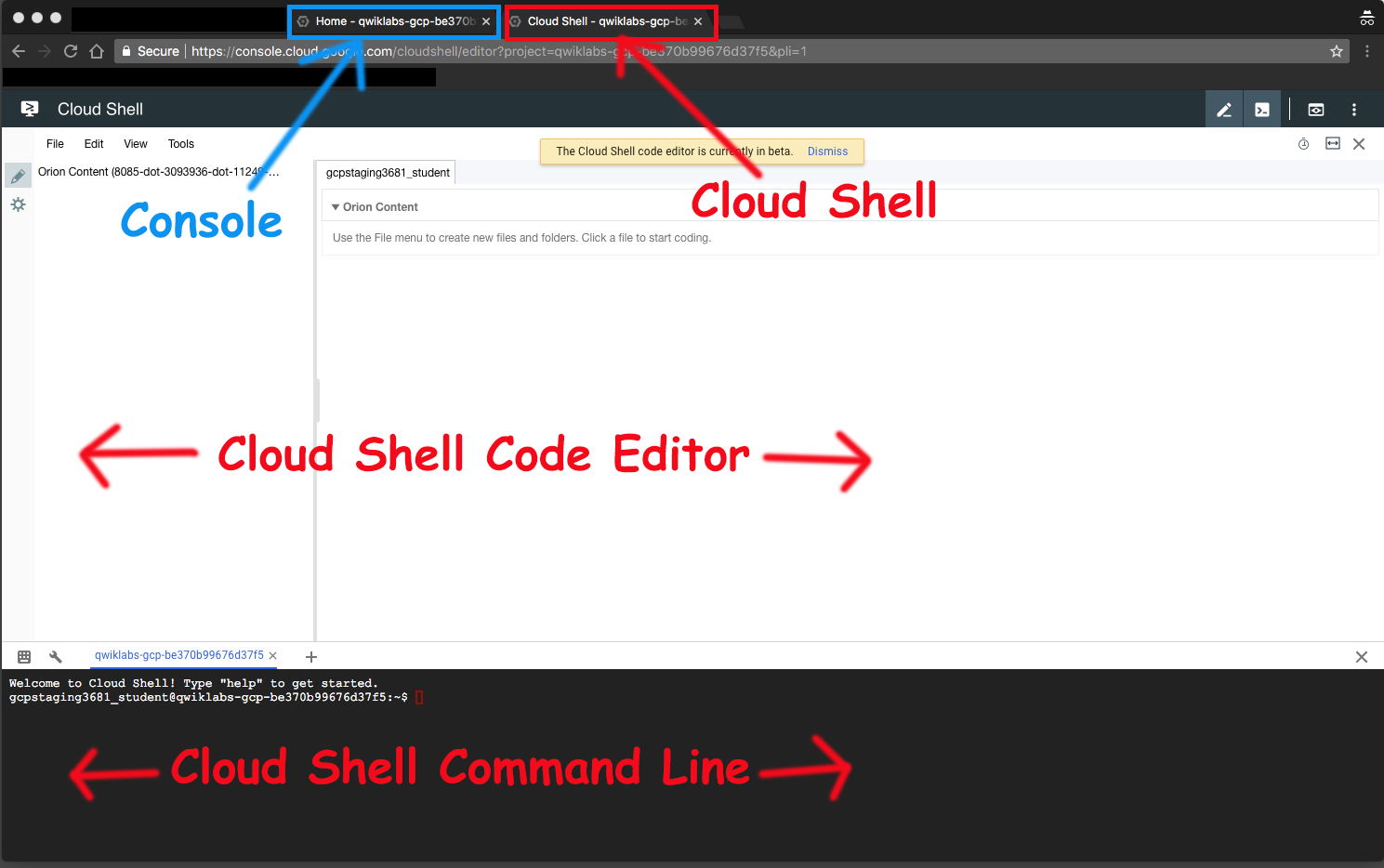


**Nota** : O botão **Launch code editor** pode estar fora da tela à direita. Pode ser necessário clicar no botão do **menu Navegação** para fechar o menu para ver os botões.



Agora você tem três interfaces disponíveis:

* O editor de código do Cloud Shell
* A linha de comando do Cloud Shell
* Console (clicando na guia). Você pode alternar entre o console e o Cloud Shell clicando na guia.



## Tarefa 1. Revisar Preparações

Estas preparações já devem ter sido feitas:

* Criar intervalo do Cloud Storage
* Clonar [repositório do](https://github.com/GoogleCloudPlatform/training-data-analyst) github para o Cloud Shell

git clone https://github.com/GoogleCloudPlatform/training-data-analyst

* Atualizar pacotes e instalar o Apache Beam

sudo apt-get install python-pip -y

pip install google-cloud-dataflow oauth2client==3.0.0

# downgrade as 1.11 breaks apitools

sudo pip install --force six==1.10

pip install -U pip

pip -V

sudo pip install apache\_beam

## Tarefa 2. Identificar Mapear e Reduzir Operações

1. No editor de códigos do Cloud Shell, navegue até o diretório /training-data-analyst/courses/data\_analysis/lab2/pythone visualize o arquivo is\_popular.pyno editor do Cloud Shell. **Não faça alterações no código.**

Alternativamente, você poderia ver o arquivo com nano. **Não faça alterações no código.**

cd ~/training-data-analyst/courses/data\_analysis/lab2/python

nano is\_popular.py

Normalmente, você desenvolveria esse código Java em um Ambiente de Desenvolvimento Integrado, como Eclipse ou Intellij (não no CloudShell).

Você pode responder a estas perguntas sobre o arquivo is\_popular.py?

* Quais argumentos personalizados são definidos?
* Qual é o prefixo de saída padrão?
* Como é a variável output\_prefix no main()set?
* Como os argumentos do pipeline são --runnerdefinidos?
* Quais são os principais passos no pipeline?
* Quais dessas etapas ocorrem em paralelo?
* Quais dessas etapas são agregações?

## Tarefa 3. Executar o pipeline

1. Execute o pipeline localmente:

cd ~/training-data-analyst/courses/data\_analysis/lab2/python

python ./is\_popular.py

Nota: se você vir um erro que diz "No handlers could be found for logger "oauth2client.contrib.multistore\_file", você pode ignorá-lo. O erro é simplesmente dizer que o log da biblioteca oauth2 irá para stderr.

1. Identifique o arquivo de saída. Deve ser **saída** <sufixo> e pode ser um arquivo particionado.

ls -al /tmp

1. Examine o arquivo de saída, substituindo '- \*' pelo sufixo apropriado.

cat /tmp/output-\*

## Tarefa 4. Usar Parâmetros da Linha de Comandos

1. Altere o prefixo de saída do valor padrão:

python ./is\_popular.py --output\_prefix=/tmp/myoutput

1. Qual será o nome do novo arquivo que está escrito?
2. Note que agora temos um novo arquivo no **diretório / tmp** :

ls -lrt /tmp/myoutput\*

## Termine seu laboratório

Quando você tiver concluído seu laboratório, clique em **Finalizar Laboratório** . O Qwiklabs remove os recursos que você usou e limpa a conta para você.

Você terá a oportunidade de avaliar a experiência do laboratório. Selecione o número de estrelas aplicável, digite um comentário e clique em **Enviar** .

O número de estrelas indica o seguinte:

* 1 estrela = muito insatisfeito
* 2 estrelas = insatisfeito
* 3 estrelas = neutra
* 4 estrelas = Satisfeito
* 5 estrelas = muito satisfeito

Você pode fechar a caixa de diálogo se não quiser fornecer feedback.

Para comentários, sugestões ou correções, use a guia **Suporte** .

Data da última atualização: 2018-11-21

Data do último teste: 2018-11-21

© 2018 Google LLC Todos os direitos reservados. Google e o logotipo do Google são marcas registradas do Google LLC. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas com as quais estão associados.