# **Automação de Cluster Utilizando Comandos CLI**

## visão global

Neste laboratório, você criará um cluster usando comandos da CLI e aprenderá sobre o fluxo de trabalho do Dataproc-GCP e a automação do fluxo de trabalho.

## Objetivos

Neste laboratório, você executará as seguintes tarefas:

* Criar um cluster personalizado do Dataproc usando o Cloud Shell

#### 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.

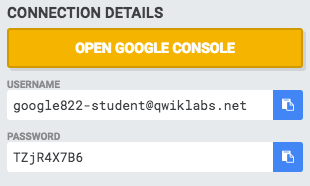
#### Comece seu laboratório

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.

Este laboratório desafia você a criar um cluster Dataproc personalizado usando apenas comandos CLI.

## Tarefa 1: preparação

Você estará criando o cluster Dataproc a partir da VM de treinamento. Existem vários valores que serão exigidos pelos comandos. Você criará variáveis ​​de ambiente para esses valores.

### Abra o terminal SSH e conecte-se à VM de treinamento

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Compute Engine** > **instâncias de VM** .
2. Localize a linha com a instância chamada **training\_vm** .
3. Na extrema direita, em 'conectar', clique em **SSH** para abrir uma janela de terminal.
4. Neste laboratório, você entrará comandos CLI no **training\_vm** .

### Verificar o intervalo a ser usado para o armazenamento temporário do Dataproc

O Dataproc pode usar um intervalo do Cloud Storage para organizar seus arquivos durante a inicialização. Você pode usar esse intervalo para preparar programas de aplicativos ou dados para uso pelo Dataproc. O bucket também pode hospedar scripts e saída de inicialização do Dataproc. O nome do intervalo deve ser globalmente exclusivo. A Qwiklabs já criou um intervalo para você que tem o mesmo nome da ID do projeto, que já é globalmente exclusiva.

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** . Verifique se o balde existe. Observe a classe de armazenamento padrão e a localização (região) desse intervalo. Você estará usando as informações desta região a seguir.

No terminal SSH **training\_vm** , configure o BUCKET.

BUCKET=<bucket-name>

Você pode usar $ BUCKET nos comandos da CLI. E se você precisar inserir o nome do intervalo <your-bucket> em um campo de texto no Console, poderá recuperar rapidamente o nome com **echo $ BUCKET** .

### Identifique uma região e zona

Você estará criando um cluster do Dataproc em uma região específica. O cluster do Dataproc e o bloco que ele usará para o armazenamento temporário devem estar na mesma região. Como o bloco que você está usando já existe, você precisará corresponder a variável de ambiente **$ MYREGION** à região do bucket.

1. Você pode usar a região usada por Qwiklabs na guia Qwiklabs em Detalhes da conexão, chamada **Região QL** .

A zona deve estar na mesma região. **$ MYZONE** conterá este valor.

1. Você pode encontrar a zona usada pelo Qwiklabs na guia Qwiklabs em Detalhes da conexão, chamada **Zona QL** .

No terminal SSH **training\_vm** , defina a REGIÃO e a ZONA.

MYREGION=<region>

MYZONE=<zone>

### Identifique um projeto

Uma variável de ambiente que você definirá é **$ PROJECT\_ID,** que contém o ID do projeto do Google Cloud necessário para acessar recursos faturáveis.

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Página inicial** . No painel com Informações do **projeto** , o **ID** do **projeto** é listado. Você também pode encontrar essas informações na guia Qwiklabs, em Detalhes da conexão, onde ela é rotulada como **ID do projeto do GCP** .

No terminal SSH **training\_vm** , defina o PROJECT\_ID.

PROJECT\_ID=<project ID>

### Identifique o endereço IP do navegador

Você usará o endereço IP do navegador para permitir que seu navegador local alcance o cluster do Dataproc.

1. Encontre o endereço IP do navegador do seu computador abrindo uma janela do navegador e visualizando <http://ip4.me/> Copie o endereço IP.
2. Crie uma variável de ambiente denominada **BROWSER\_IP** .

BROWSER\_IP=<your-browser-ip>

### Copie os arquivos de amostra para o diretório inicial training\_vm

Os arquivos de amostra que você precisa já foram arquivados em training\_vm. Você precisará copiá-los em seu diretório de usuários com o seguinte comando.

1. Na janela do terminal **training\_vm** SSH.

cd

cp -r /training/training-data-analyst .

ls

1. Agora você deve ter o diretório **/ training** em seu diretório inicial. E deve ter arquivos dentro dele.

## Tarefa 2: Personalizar a ação de inicialização do Dataproc

### Revise o script de inicialização do cluster

1. Revise o script de customização do cluster.

cd ~/training-data-analyst/courses/unstructured/

cat init-script.sh

Este script instala a API do cliente Python do Google em todas as máquinas do cluster.

1. Personalize a inicialização para adicionar instruções para clonar o repositório do curso no nó mestre.

Use **nano** ou **vi** para editar o arquivo **init-script.sh** .

#!/bin/bash

# install Google Python client on all nodes

apt-get update

apt-get install -y python-pip

pip install --upgrade google-api-python-client

ROLE=$(/usr/share/google/get\_metadata\_value attributes/dataproc-role)

if [[ "${ROLE}" == 'Master' ]]; then

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

fi

Você pode criar suas próprias ações de inicialização de cluster usando esse script como exemplo. Ações de inicialização são comumente usadas para instalar software nos nós do cluster. E esse script mostra como distinguir o nó mestre dos nós do trabalhador.

## Tarefa 3: Criar o cluster do Dataproc

1. Verifique se o intervalo do Cloud Storage existe e se as variáveis ​​de ambiente $ BUCKET e $ MYZONE ainda estão definidas. O bucket será usado pelo cluster do Dataproc para preparar os arquivos à medida que o cluster for inicializado.

echo $BUCKET $MYREGION $MYZONE

echo $PROJECT\_ID

Permitir que o comando padrão para o seu projeto de conta e localização padrão.

1. Copie o script de personalização do cluster para o bloco.

gsutil cp init-script.sh gs://$BUCKET

O Cloud Storage é um serviço de dados distribuído e resiliente muito sofisticado que suporta os RDDs do Spark. Ele é conectado ao Dataproc por uma rede de largura de banda de bisseção de petabit, permitindo que os dados sejam processados ​​a partir de onde ele está localizado, em vez de precisar ser copiado. Então você pode usar o Cloud Storage em vez do HDFS.

Como os dados no Cloud Storage sobrevivem ao encerramento do cluster, se você os usou em vez do HDFS, poderá finalizar os clusters quando eles não estiverem sendo usados ​​para reduzir as despesas. Você pode agendar o cluster para finalizar após ficar ocioso por um período (quando as tarefas estiverem concluídas).

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/scheduled-deletion>

Você também pode usar o Cloud Storage para preparar aplicativos e armazenar arquivos de saída.

Um cluster pode usar um único endpoint global ou endpoints regionais. Você pode selecionar a zona onde o cluster será criado ou permitir que a zona seja escolhida para você.

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone>

1. Além do script de inicialização personalizada, você pode usar scripts de inicialização que foram predefinidos. O script localizado em: **gs: //dataproc-initialization-actions/datalab/datalab.sh** instala o Datalab no nó mestre. O Datalab é um ambiente de desenvolvimento baseado em notebooks baseado em notebooks Jupyter.
2. Observe que esse cluster inclui dois nós de trabalho preemptivos.
3. Crie o cluster customizado

gcloud dataproc clusters create cluster-custom \

--bucket $BUCKET \

--subnet default \

--zone $MYZONE \

--region $MYREGION \

--master-machine-type n1-standard-2 \

--master-boot-disk-size 100 \

--num-workers 2 \

--worker-machine-type n1-standard-1 \

--worker-boot-disk-size 50 \

--num-preemptible-workers 2 \

--image-version 1.2 \

--scopes 'https://www.googleapis.com/auth/cloud-platform' \

--tags customaccess \

--project $PROJECT\_ID \

--initialization-actions 'gs://'$BUCKET'/init-script.sh','gs://dataproc-initialization-actions/datalab/datalab.sh'

Espere até que o comando esteja completo no terminal antes de continuar. Pode levar vários minutos para que todas as ações de inicialização sejam concluídas.

As opções usadas neste comando incluem recursos de segurança, economia de custos e flexibilidade.

--tags: Aplica uma tag de rede para que você possa automatizar a criação de regras de firewall.

--scopes: Aplica restrições e permissões do Cloud IAM ao cluster.

--num preemptible-workers: Controla o número de nós trabalhadores de baixo custo presentes.

--inicializações: personaliza o software no cluster.

Opções para um estudo mais aprofundado:

--no-address, --network, --subnet:

As VMs possuem apenas IPs internos para maior segurança. Exige a ativação do acesso privado da API do GCP na rede, estabelecendo regras de firewall específicas e passando a sub-rede.

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network>

A criação de clusters pode ser baseada em script com base nesse comando da CLI. Portanto, a criação de cluster pode ser automatizada. Você pode criar clusters periodicamente usando um trabalho Cron. Você pode escrever um aplicativo para detectar um arquivo PySpark sendo adicionado a um intervalo no Cloud Storage e usá-lo para acionar a criação do cluster. Você também pode integrar a criação de cluster e o envio de trabalhos do Dataproc a um ambiente de Integração Contínua, como o Travis ou o Jenkins. Dessa forma, você só inicia e paga por clusters quando eles são necessários.

## Tarefa 4: verificar a personalização de cluster

Nesta tarefa, você verificará se a ação de inicialização de personalização instalou o Cloud Datalab no nó mestre.

1. Verifique se o endereço IP do seu navegador está definido em uma variável de ambiente para uso na regra de firewall.

echo $BROWSER\_IP

1. Crie uma regra de firewall.

gcloud compute \

--project=$PROJECT\_ID \

firewall-rules create allow-custom \

--direction=INGRESS \

--priority=1000 \

--network=default \

--action=ALLOW \

--rules=tcp:9870,tcp:8088,tcp:8080 \

--source-ranges=$BROWSER\_IP/32 \

--target-tags=customaccess

1. Localize o endereço IP externo do nó mestre. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Clusters** . Clique no **cluster personalizado** .
2. Clique nas **instâncias de VM** .
3. Clique em **cluster-custom-m** .
4. Na seção de interfaces de rede, localize o IP externo. Destaque e copie.
5. Abra uma nova guia ou janela do navegador. Digite **<external\_IP>: 8080** e pressione Enter.

Você deve ver o Google Cloud Datalab.

Criar o cluster personalizado é o objetivo deste laboratório. Se este foi seu ambiente de produção, seus próximos passos podem ser:

* Transforme os comandos de criação em um script para que você possa inicializar um cluster sob demanda.
* Adicione uma opção ao comando para finalizar o cluster após um período silencioso.
* Transforme a regra de firewall em um script para que você possa ativar / desativar o acesso externo (navegador) somente quando for necessário para atividades de administração.
* Desenvolva e teste sua aplicação nos notebooks do Datalab.
* Hospede o aplicativo de produção em um intervalo do Cloud Storage.
* Hospede e acesse seus dados no Cloud Storage, no BigQuery ou no Bigtable.
* Para capacidade, Edite o número de nós do trabalhador preemptivos usando o Console e o cluster em execução será adaptado.
* Encerre o cluster quando não estiver em uso ou agende o encerramento automático.

## 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** .

Última atualização: 2018-09-12

© 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.