

Explore a BigQuery Public Dataset

Overview

Armazenar e consultar conjuntos de dados em massa pode ser demorado e caro sem o hardware e a infraestrutura corretos. O Google BigQuery é um data warehouse corporativo que resolve esse problema permitindo consultas SQL super rápidas usando o poder de processamento da infraestrutura do Google. Basta mover seus dados para o BigQuery e permitir que realizemos o trabalho pesado. Você pode controlar o acesso ao projeto e aos seus dados com base em suas necessidades de negócios, como permitir que outras pessoas visualizem ou consultem seus dados.

Você acessa o BigQuery por meio do console do GCP, da ferramenta de linha de comando ou fazendo chamadas para a API REST do BigQuery usando uma variedade de bibliotecas de clientes, como Java, .NET ou Python. Há também várias ferramentas de terceiros que você pode usar para interagir com o BigQuery, como visualizar os dados ou carregar os dados. Neste laboratório, você acessa o BigQuery usando a interface do usuário da web.

Você pode usar a IU da web do BigQuery no console do GCP como uma interface visual para concluir tarefas como executar consultas, carregar dados e exportar dados. Este laboratório prático mostra como consultar tabelas em um conjunto de dados público e como carregar dados de amostra no BigQuery por meio do Console do GCP.

O que você vai fazer

Neste laboratório você:

Consultar um conjunto de dados público

Crie uma tabela personalizada

Carregar dados em uma tabela

Consultar uma tabela

Configuração e requisitos

Configuração Qwiklabs

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 tem sua própria conta do GCP, não a use para este laboratório.

Se o seu laboratório solicitar que você faça o 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 andamento 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:

Abra o Google Console

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

Console do Google Cloud Platform

Faça login no Google Cloud Console

Usando a guia / janela do navegador Qwiklabs ou o navegador separado que você está usando para a sessão Qwiklabs, copie o Nome de usuário no painel Detalhes da conexão e clique no botão Abrir o Google Console.

Você será solicitado a escolher uma conta. Clique em Usar outra conta. Google\_choose\_Account

Cole o nome de usuário e, em seguida, a senha, conforme solicitado:

Faça login para continuar no Google Cloud Platform

Aceite os termos e condições.

Como esta é uma conta temporária, que você só terá que acessar para este laboratório:

Não adicione opções de recuperação

Não se inscreva para testes gratuitos

Observação: você pode visualizar a lista de serviços clicando no botão do menu de navegação do GCP na parte superior esquerda, ao lado de "Google Cloud Platform". Menu do Cloud Console

Abra o BigQuery Console

No Google Cloud Console, selecione o menu Navegação> BigQuery:

BigQuery\_menu.png

A caixa de mensagem Bem-vindo ao BigQuery na Cloud Console é aberta. Essa caixa de mensagem fornece um link para o guia de início rápido e lista as atualizações da interface do usuário.

Clique em Concluído.

Consultar um conjunto de dados público

Nesta seção, você carrega um conjunto de dados público, nomes dos EUA, no BigQuery e consulta o conjunto de dados para determinar os nomes mais comuns nos EUA entre 1910 e 2013.

Carregar conjunto de dados do Nome dos EUA

No painel esquerdo, clique em ADICIONAR DADOS> Explorar conjuntos de dados públicos.

add-dataset.png

A janela Datasets é aberta.

Na caixa de pesquisa, digite EUA, em seguida, digite.

Clique no bloco Nomes dos EUA que você vê nos resultados da pesquisa.

nomes dos EUA

Clique em VER DATASET.

O BigQuery é aberto em uma nova guia do navegador. O projeto bigquery-public-data é adicionado aos seus recursos e você vê o dataset usa\_names listado no painel esquerdo na sua árvore Resources.

usa-names-dataset.png

Consultar o conjunto de dados do nome dos EUA

Consulte bigquery-public-data.usa\_names.usa\_1910\_2013 para o nome e gênero dos bebês neste conjunto de dados e, em seguida, liste os 10 principais nomes em ordem decrescente.

Copie e cole a seguinte consulta na área de texto do editor de consultas:



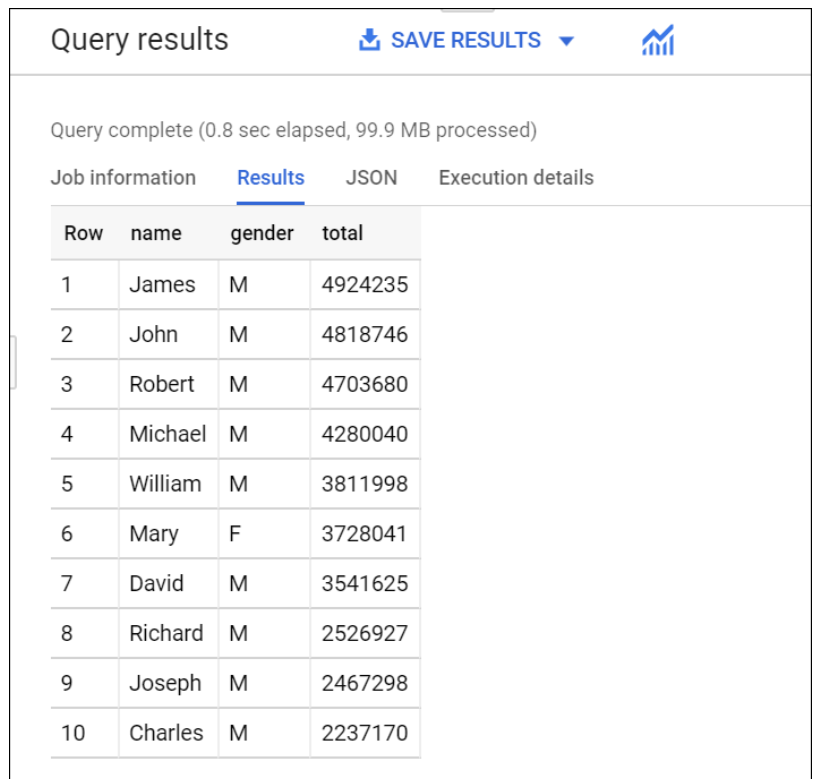
No canto inferior direito da janela, visualize o validador de consulta.



O BigQuery exibe um ícone de marca de seleção verde se a consulta for válida. Se a consulta for inválida, um ícone de ponto de exclamação vermelho será exibido. Quando a consulta é válida, o validador também mostra a quantidade de dados que a consulta processa quando você a executa. Isso ajuda a determinar o custo de execução da consulta.

Clique em Executar.

Os resultados da consulta são abertos abaixo do editor de consultas. Na parte superior da seção Resultados da consulta, o BigQuery exibe o tempo decorrido e os dados processados ​​pela consulta. Abaixo do tempo, está a tabela que exibe os resultados da consulta. A linha de cabeçalho contém o nome da coluna conforme especificado em GROUP BY na consulta.



Crie uma tabela personalizada

Nesta seção, você cria uma tabela personalizada, carrega dados nela e executa uma consulta na tabela.

Baixe os dados para o seu computador local

O arquivo que você está baixando contém aproximadamente 7 MB de dados sobre nomes populares de bebês, e é fornecido pela Administração do Seguro Social dos EUA.

Faça o download do arquivo zip do nome do bebê para o seu computador local.

Descompacte o arquivo no seu computador.

O arquivo zip contém um arquivo NationalReadMe.pdf que descreve o conjunto de dados. Saiba mais sobre o conjunto de dados.

Abra o arquivo chamado yob2014.txt para ver como são os dados. O arquivo é um arquivo de valores separados por vírgula (CSV) com as três colunas a seguir: nome, sexo (M ou F) e número de filhos com esse nome. O arquivo não possui linha de cabeçalho.

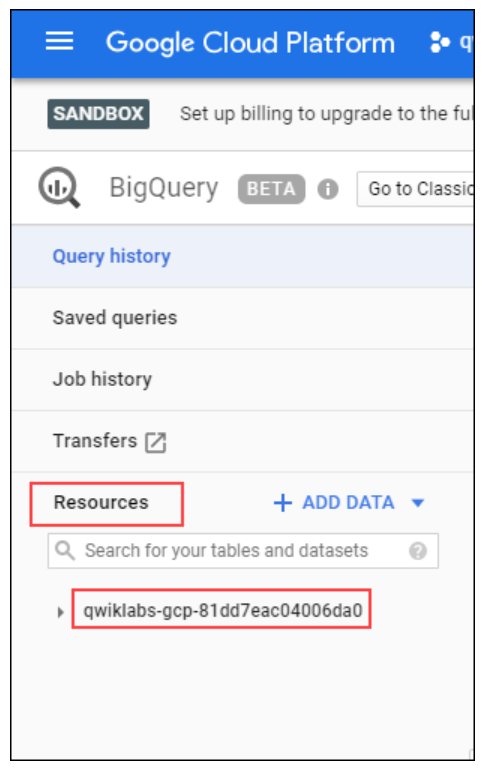
Observe a localização do arquivo yob2014.txt para que você possa encontrá-lo mais tarde.

Crie um conjunto de dados

Nesta seção, você cria um conjunto de dados para manter sua tabela, adicionar dados ao seu projeto e, em seguida, criar a tabela de dados com a qual você fará uma consulta.

Os conjuntos de dados ajudam você a controlar o acesso a tabelas e visualizações em um projeto. Este laboratório usa apenas uma tabela, mas você ainda precisa de um conjunto de dados para manter a tabela.

De volta ao console, no painel esquerdo, na seção Recursos, clique no ID do projeto do GCP (ele começará com qwiklabs).



Seu projeto é aberto no editor de consultas.

No lado direito da seção do projeto, clique em CREATE DATASET.

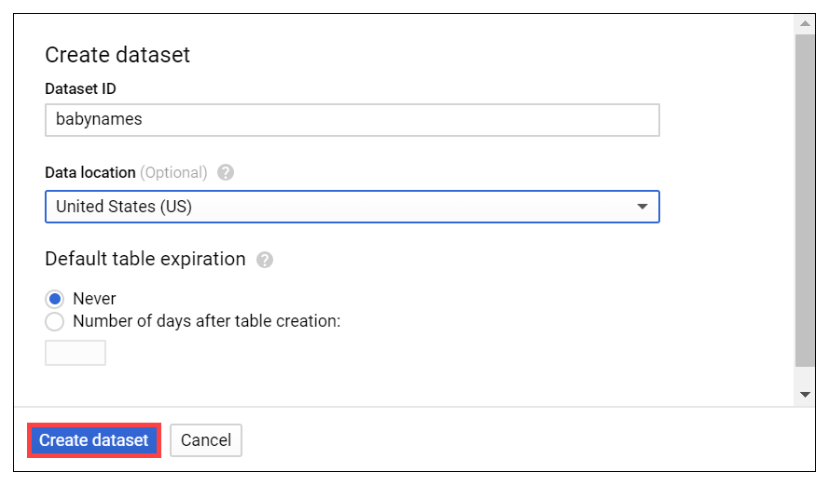
Na página Criar conjunto de dados:

Para ID do conjunto de dados, insira babynames.

Para localização de dados, escolha Estados Unidos (EUA).

Para expiração de tabela padrão, deixe o valor padrão.

Atualmente, os conjuntos de dados públicos são armazenados no local de várias regiões dos EUA. Para simplificar, coloque seu conjunto de dados no mesmo local.



Clique em Criar conjunto de dados na parte inferior do painel.

Carrega os dados em uma nova tabela

Nesta seção, você carrega dados na tabela que você criou.

Clique em babynames encontrados no painel esquerdo na seção Recursos e, em seguida, clique em Criar tabela.

Use os valores padrão para todas as configurações, a menos que seja indicado de outra forma.

Na página Criar tabela:

Para Origem, selecione Carregar no menu suspenso Criar tabela de :.

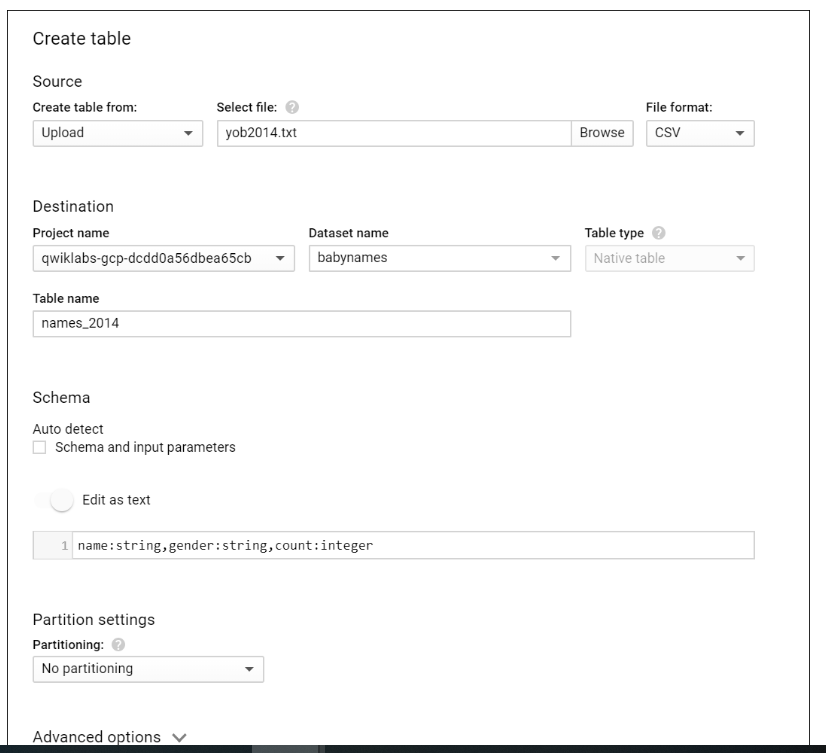
Para o arquivo Select, clique em Browse, navegue até o arquivo yob2014.txt e clique em Open.

Para formato de arquivo, escolha CSV no menu suspenso.

Para o nome da tabela, insira names\_2014.

Na seção Esquema, clique na opção Editar como texto para alternar e cole a seguinte definição de esquema na caixa de texto.

name:string,gender:string,count:integer



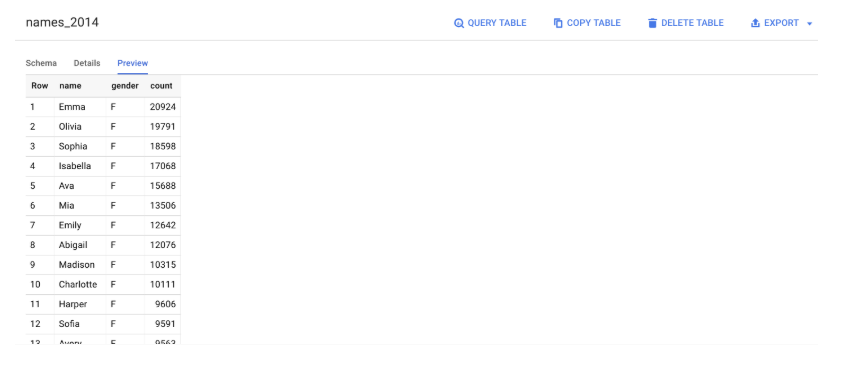
Clique em Criar tabela (na parte inferior da janela).

Aguarde o BigQuery criar a tabela e carregar os dados. Enquanto o BigQuery carrega os dados, uma sequência (1 em execução) é exibida ao lado do histórico de tarefas no painel esquerdo. A string desaparece depois que os dados são carregados.

Visualizar a tabela

No painel esquerdo, selecione babynames> names\_2014 no painel de navegação.

No painel de detalhes, clique na guia Visualizar.



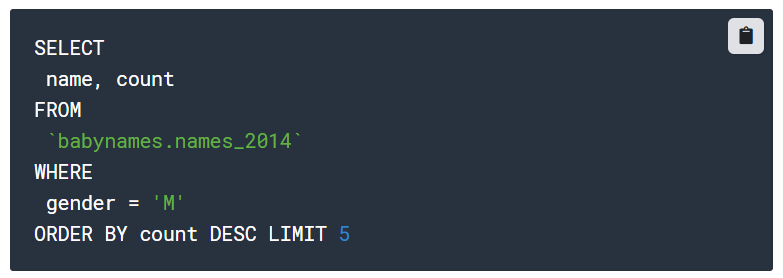
Consulta a mesa

Agora que você carregou dados na sua tabela, pode executar consultas sobre ela. O processo é idêntico ao exemplo anterior, exceto que, desta vez, você está consultando sua tabela em vez de uma tabela pública.

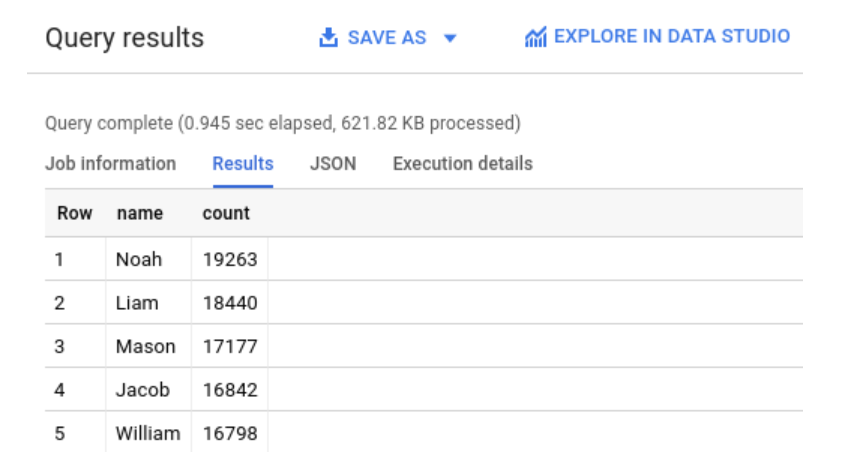
No editor de consultas, clique em Compor nova consulta.

Copie e cole a consulta a seguir no editor de consultas. Esta consulta recupera os 5 principais nomes de bebês para homens dos EUA em 2014.

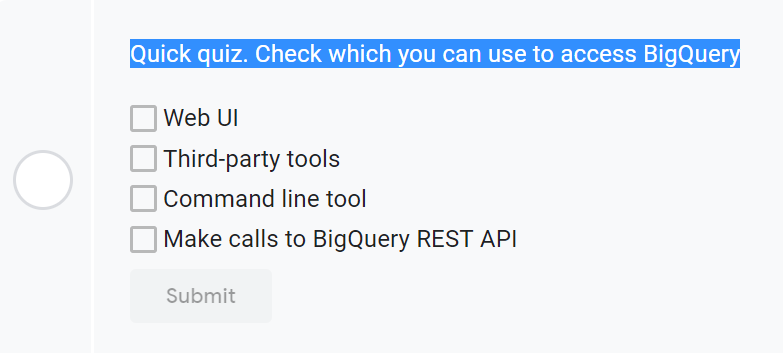
SELECIONAR



Clique em Executar. Os resultados são exibidos abaixo da janela de consulta.



Questionário rápido. Verifique qual você pode usar para acessar o BigQuery



Parabéns!

Você consultou um conjunto de dados público, criou uma tabela personalizada, carregou dados nela e executou uma consulta nessa tabela.

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

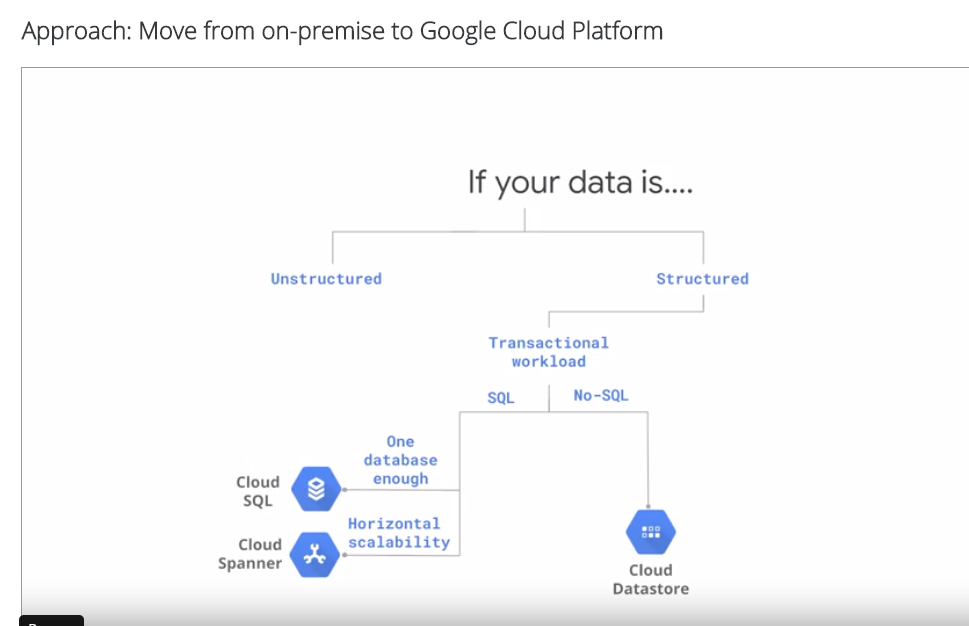
4 estrelas = Satisfeito

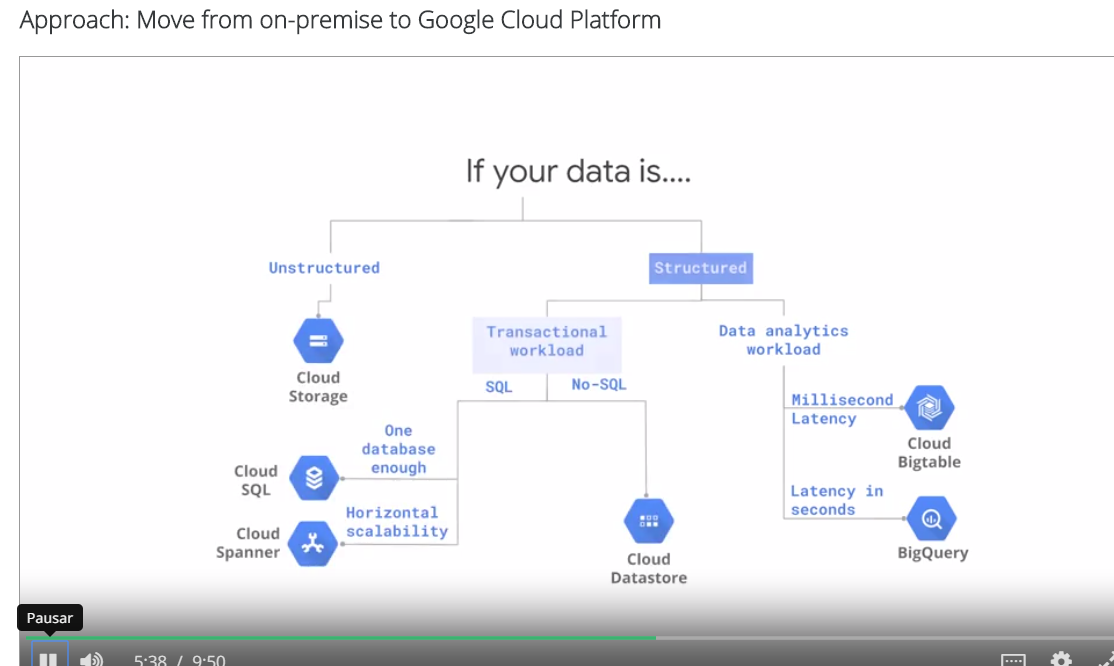
5 estrelas = muito satisfeito

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

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

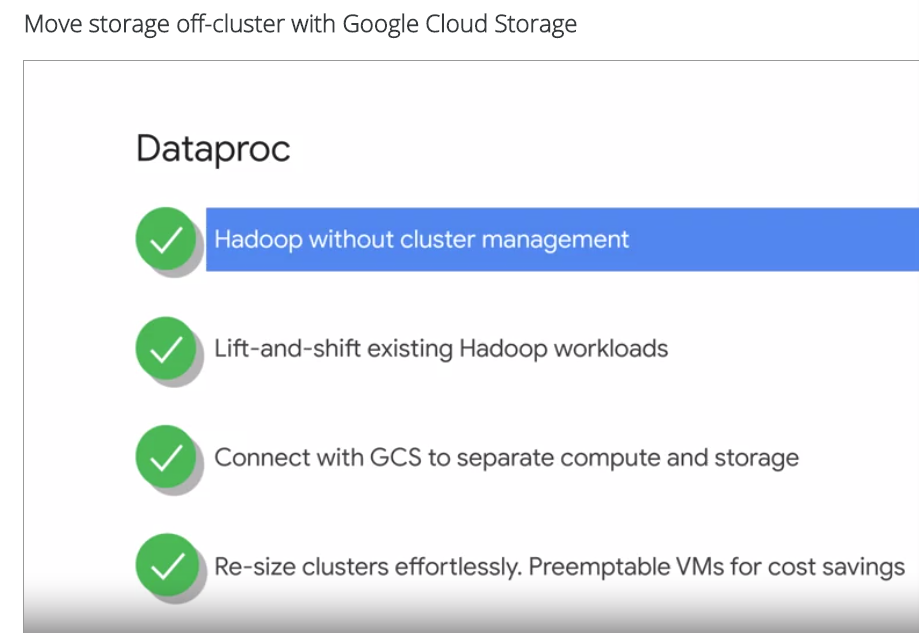
© 2019 Google LLC Todos os direitos reservados. Google e o logotipo do Google são marcas registradas do Google LLC. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas com as quais estão associados.





de entrada e saída no armazenamento do Google Cloud. Como os dados de entrada e saída estão fora do cluster, o cluster pode ser criado para um único trabalho ou tipo de carga de trabalho e pode ser desligado quando não estiver em uso. Alterar o código que funciona no local para que funcione com os dados no armazenamento em nuvem é fácil. Basta substituir o HDFS em seu Spark ou Big Cloud por GS. Então pegue as URLs do HDFS, HDFS: // e substitua por GS: // Isso fará com que o Spark de grande trabalho leia ou grave no armazenamento em nuvem, é isso.

A modificação do código não funciona localmente para funcionar com os dados armazenados. Basta substituir o HDFS no seu Spark ou Big Cloud por GS. Então pegue as URLs do HDFS, HDFS: // e substitua por GS: // Isso faz com que o faísca seja de grande importância. Há também uma base H



( LABORATORIO 2 )

Agora é hora do seu laboratório.

0:03

Você está encarregado de migrar a carga de trabalho de aprendizado de máquina existente da sua empresa para recomendações de armazenamento de seu cluster do Hadoop no local para a nuvem. Sua organização está satisfeita com o modelo atual, mas a infraestrutura básica no local está fazendo com que as dores de cabeça sejam ajustadas e utilizadas de maneira eficiente.

0:24

Seu CTO quer o menor atrito possível da sua infraestrutura existente no Hadoop, mas já ouviu falar das vantagens que as soluções em nuvem oferecem para escalonamento automático e gerenciamento sem servidor.

0:38

Então, aqui estão os seus objetivos de laboratório.

0:41

Crie uma instância do Cloud SQL e preencha as tabelas.

0:47

Explore os dados de aluguel usando instruções SQL usando o Cloud Shell.

0:52

Inicie o Dataproc.

0:54

Treine e aplique um modelo de aprendizado de máquina escrito no PySpark para criar recomendações de produtos.

1:02

Explore as linhas inseridas no Cloud SQL.

1:06

Aqui está como será a configuração. Na primeira etapa, configuramos o Google Cloud Storage e o Cloud SQL. Em seguida, importe os registros do GCS para o Cloud SQL. E agora que suas classificações estão no Cloud SQL, na etapa dois, você executará uma tarefa de treinamento de aprendizado de máquina no Cloud Dataproc para ler essas classificações e treinar o modelo de aprendizado de máquina. Na etapa três, você executará o modelo no Cloud Dataproc para criar recomendações. E salve as cinco principais recomendações para cada usuário no Cloud SQL.

1:43

Por fim, na quarta etapa, suas avaliações podem ser enviadas de volta aos usuários por meio do App Engine. Você não executará as etapas um e quatro neste laboratório porque essas etapas lidam principalmente com programação da web. Neste laboratório, você se concentra em fazer os passos dois e três. Experimente o laboratório e tenha em mente que você tem várias tentativas para cada laboratório e pode sempre voltar e praticar mais.

Agora é hora do seu laboratório.

0:03

Você está encarregado de migrar a carga de trabalho de aprendizado de máquina existente da sua empresa para recomendações de armazenamento de seu cluster do Hadoop no local para a nuvem. Sua organização está satisfeita com o modelo atual, mas a infraestrutura básica no local está fazendo com que as dores de cabeça sejam ajustadas e utilizadas de maneira eficiente.

0:24

Seu CTO quer o menor atrito possível da sua infraestrutura existente no Hadoop, mas já ouviu falar das vantagens que as soluções em nuvem oferecem para escalonamento automático e gerenciamento sem servidor.

0:38

Então, aqui estão os seus objetivos de laboratório.

0:41

Crie uma instância do Cloud SQL e preencha as tabelas.

0:47

Explore os dados de aluguel usando instruções SQL usando o Cloud Shell.

0:52

Inicie o Dataproc.

0:54

Treine e aplique um modelo de aprendizado de máquina escrito no PySpark para criar recomendações de produtos.

1:02

Explore as linhas inseridas no Cloud SQL.

1:06

Aqui está como será a configuração. Na primeira etapa, configuramos o Google Cloud Storage e o Cloud SQL. Em seguida, importe os registros do GCS para o Cloud SQL. E agora que suas classificações estão no Cloud SQL, na etapa dois, você executará uma tarefa de treinamento de aprendizado de máquina no Cloud Dataproc para ler essas classificações e treinar o modelo de aprendizado de máquina. Na etapa três, você executará o modelo no Cloud Dataproc para criar recomendações. E salve as cinco principais recomendações para cada usuário no Cloud SQL.

1:43

Por fim, na quarta etapa, suas avaliações podem ser enviadas de volta aos usuários por meio do App Engine. Você não executará as etapas um e quatro neste laboratório porque essas etapas lidam principalmente com programação da web. Neste laboratório, você se concentra em fazer os passos dois e três. Experimente o laboratório e tenha em mente que você tem várias tentativas para cada laboratório e pode sempre voltar e praticar mais.

# Recommend Products using ML with Cloud SQL and Dataproc

## Overview

# Neste laboratório, você preenche os dados de aluguéis no Cloud SQL para usar o mecanismo de recomendação de aluguéis. O que você aprende Neste laboratório, você irá: Criar instância do Cloud SQL Crie tabelas de banco de dados importando arquivos .sql do Cloud Storage Preencha as tabelas importando arquivos .csv do Cloud Storage Permitir acesso ao Cloud SQL Explore os dados de aluguel usando instruções SQL do CloudShell Configuração e requisitos Configuração Qwiklabs 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 tem sua própria conta do GCP, não a use para este laboratório. Se o seu laboratório solicitar que você faça o 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 andamento 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.

Console do Google Cloud Platform

Faça login no Google Cloud Console

Usando a guia / janela do navegador Qwiklabs ou o navegador separado que você está usando para a sessão Qwiklabs, copie o Nome de usuário no painel Detalhes da conexão e clique no botão Abrir o Google Console.

Você será solicitado a escolher uma conta. Clique em Usar outra conta. Google\_choose\_Account

# 

Aceite os termos e condições.

Como esta é uma conta temporária, que você só terá que acessar para este laboratório:

Não adicione opções de recuperação

Não se inscreva para testes gratuitos

# 

Introdução

Neste laboratório, você preenche os dados de aluguéis no Cloud SQL para usar o mecanismo de recomendação de aluguéis. O próprio mecanismo de recomendações será executado no Dataproc usando o Spark ML.

Criar instância do Cloud SQL

No console do GCP, clique em SQL (na seção Storage).

Clique em Criar instância.

Escolha o MySQL. Clique em Avançar, se necessário.

Para ID da Instância, digite aluguel.

ab1cdf08212ecadf.png

Role para baixo e especifique uma senha root. Antes de esquecer, anote a senha do root.

Clique em Criar para criar a instância. Levará um minuto ou mais para que sua instância do Cloud SQL seja provisionada.

Crie tabelas

Enquanto você espera que sua instância seja criada, leia o script mySQL abaixo e responda às perguntas que seguem

# 

# Quantas tabelas esse script criará?

# Quando um usuário avalia uma casa (dando quatro estrelas, por exemplo), uma entrada é adicionada à tabela \_\_\_3\_\_\_\_.

# General information about houses, such as the number of rooms they have and their average rating is stored in the \_\_\_\_\_\_\_\_\_ table.

O trabalho do mecanismo de recomendação é preencher a tabela \_\_\_\_\_\_\_\_\_\_\_ para cada usuário e casa: essa é a classificação prevista dessa casa por esse usuário.

# 

No Cloud SQL, clique em aluguéis para visualizar as informações da instância.

Conecte-se ao banco de dados

Encontre a caixa Conectar a esta instância na página e clique em conectar usando o Cloud Shell

Observação: você também pode se conectar à sua instância a partir de uma VM dedicada do Cloud Compute Engine, mas, por enquanto, o Cloud Shell criará uma micro-VM para nós e operará a partir dela.

Aguarde até que o Cloud Shell carregue

Quando o Cloud Shell for carregado, você verá o comando abaixo já digitado:

gcloud sql connect aluguéis --user = root --quiet

Pressione Enter

Aguarde até que seu endereço IP seja colocado na lista de permissões

# Whitelisting your IP for incoming connection for 5 minutes...⠹

Quando solicitado, digite sua senha e pressione Enter (nota: você não verá sua senha digitada ou mesmo \*\*\*\*)

Agora você pode executar comandos no seu banco de dados!

# 

Execute o comando abaixo



Você deve ver os bancos de dados do sistema padrão:

# 

Nota: Você sempre deve terminar seus comandos mySQL com um ponto e vírgula;

Copie e cole a instrução SQL abaixo que você analisou anteriormente e cole-a na linha de comando

# 

# Confirme que você vê recomendação\_spark como um banco de dados agora:

# 

# 

# 

Dados do palco no Google Cloud Storage

Opção 1: use a linha de comando

Abra uma nova guia do Cloud Shell (não use a guia mySQL Cloud Shell existente)

Cole no comando abaixo

# Paste in the below command

# echo "Creating bucket: gs://$DEVSHELL\_PROJECT\_ID"

# gsutil mb gs://$DEVSHELL\_PROJECT\_ID

# echo "Copying data to our storage from public dataset"

# gsutil cp gs://cloud-training/bdml/v2.0/data/accommodation.csv gs://$DEVSHELL\_PROJECT\_ID

# gsutil cp gs://cloud-training/bdml/v2.0/data/rating.csv gs://$DEVSHELL\_PROJECT\_ID

# echo "Show the files in our bucket"

# gsutil ls gs://$DEVSHELL\_PROJECT\_ID

# echo "View some sample data"

# gsutil cat gs://$DEVSHELL\_PROJECT\_ID/accommodation.csv

# Opção 2: usar a interface do usuário do console

# Ignore estas etapas se você já tiver carregado seus dados usando a linha de comando

# Navegue até Armazenamento e selecione Armazenamento> Navegador

# Clique em Criar Balde (se ainda não existir)

# Especifique o nome do seu projeto como o nome do intervalo

# Clique em Criar

# Faça o download dos arquivos abaixo localmente e faça o upload deles no seu novo intervalo:

* [accommodation.csv](https://storage.googleapis.com/cloud-training/bdml/v2.0/data/accommodation.csv)
* [rating.csv](https://storage.googleapis.com/cloud-training/bdml/v2.0/data/rating.csv)

Como carregar dados do Google Cloud Storage em tabelas do Cloud SQL

Navegue de volta para o SQL

Clique em aluguéis

Importar dados de acomodação

Clique em Importar (menu superior).

Especifique o seguinte:

Arquivo do Cloud Storage: navegue para selecionar accommodation.csv

Formato de importação: CSV

Banco de dados: selecione recomendação\_spark na lista suspensa

Tabela: copiar e colar: Alojamento

Clique em Importar

Como carregar dados do Google Cloud Storage em tabelas do Cloud SQL

Navegue de volta para o SQL

Clique em aluguéis

Importar dados de acomodação

Clique em Importar (menu superior).

Especifique o seguinte:

Arquivo do Cloud Storage: navegue para selecionar accommodation.csv

Formato de importação: CSV

Banco de dados: selecione recomendação\_spark na lista suspensa

Tabela: copiar e colar: Alojamento

Clique em Importar

# 

Você será redirecionado de volta para a página Visão geral. Espere um minuto para os dados serem carregados.

Importar dados de classificação do usuário

Clique em Importar (menu superior).

Especifique o seguinte:

Arquivo do Cloud Storage: navegue para selecionar rating.csv

Formato de importação: CSV

Banco de dados: selecione recomendação\_spark na lista suspensa

Tabela: copiar e colar: Classificação

Clique em Importar

Você será redirecionado de volta para a página Visão geral. Espere um minuto para os dados serem carregados.

Explore os dados do Cloud SQL

Se você fechou a conexão do Cloud Shell com mySQL, abra-a novamente localizando Conectar-se a esta instância e clicando em Conectar usando o Cloud Shell

Pressione Enter quando solicitado para efetuar login

Forneça sua senha e pressione Enter

Consultar os dados das classificações:

# USE recommendation\_spark;

# SELECT \* FROM Rating

# LIMIT 15;

# Use uma função de agregação SQL para contar o número de linhas na ta

# SELECT COUNT(\*) AS num\_ratings

# FROM Rating;bela.

# 

# Qual é a avaliação média de nossas acomodações?

# SELECT

# COUNT(userId) AS num\_ratings,

# COUNT(DISTINCT userId) AS distinct\_user\_ratings,

# MIN(rating) AS worst\_rating,

# MAX(rating) AS best\_rating,

# AVG(rating) AS avg\_rating

# FROM Rating;

# 

No aprendizado de máquina, precisaremos de um rico histórico de preferências do usuário para o modelo aprender. Execute a consulta abaixo para ver quais usuários forneceram mais avaliações

# SELECT

# userId,

# COUNT(rating) AS num\_ratings

# FROM Rating

# GROUP BY userId

# ORDER BY num\_ratings DESC;

# 

# Você pode sair do prompt do mysql digitando exit.

Gerando Recomendações de Alojamento com Aprendizado de Máquina Usando o Cloud Dataproc

Neste laboratório, você realiza recomendações de aprendizado de máquina usando o Dataproc.

O que você aprende

Neste laboratório, você irá:

Lançamento do Dataproc

Executar tarefas do SparkML usando o Dataproc

Introdução

Neste laboratório, você usa o Dataproc para treinar o modelo de aprendizado de máquina de recomendações com base nas classificações anteriores dos usuários. Em seguida, você aplica esse modelo para criar uma lista de recomendações para cada usuário no banco de dados.

Neste laboratório, você irá:

Lançamento do Dataproc

Treine e aplique o modelo ML escrito no PySpark para criar recomendações de produtos

Explorar linhas inseridas no Cloud SQL

Lançamento do Dataproc

Para iniciar o Dataproc e configurá-lo para que cada uma das máquinas no cluster possa acessar o Cloud SQL:

No Console do GCP, no menu Navegação (menu Navegação), clique em SQL e observe a região da sua instância do Cloud SQL:

# 

Na foto acima, a região é us-central1.

No Console do GCP, no menu Navegação (menu Navegação), clique em Dataproc e clique em Ativar API, se solicitado.

Uma vez ativado, clique em Criar cluster e nomeie seus aluguéis de cluster

Deixe a região como está, ou seja, global e altere a região para us-central1-a (na mesma zona da sua instância do Cloud SQL). Isso minimizará a latência de rede entre o cluster e o banco de dados.

Para o nó mestre, para o tipo de máquina, selecione 2 vCPUs (n1-standard-2).

Para nós do Worker, para o tipo Machine, selecione 2 vCPUs (n1-standard-2).

Deixe todos os outros valores com seus padrões e clique em Criar. Levará de 1 a 2 minutos para provisionar seu cluster.

Observe os nós de trabalho Nome, Zona e Total em seu cluster.

Copie e cole o script abaixo em seu Cloud Shell (opcionalmente, altere CLUSTER, ZONE, NWORKERS, se necessário, antes de executar)

# echo "Authorizing Cloud Dataproc to connect with Cloud SQL"

# CLUSTER=rentals

# CLOUDSQL=rentals

# ZONE=us-central1-a

# NWORKERS=2

# machines="$CLUSTER-m"

# for w in `seq 0 $(($NWORKERS - 1))`; do

# machines="$machines $CLUSTER-w-$w"

# done

# echo "Machines to authorize: $machines in $ZONE ... finding their IP addresses"

# ips=""

# for machine in $machines; do

# IP\_ADDRESS=$(gcloud compute instances describe $machine --zone=$ZONE --format='value(networkInterfaces.accessConfigs[].natIP)' | sed "s/\[u'//g" | sed "s/'\]//g" )/32

# echo "IP address of $machine is $IP\_ADDRESS"

# if [ -z $ips ]; then

# ips=$IP\_ADDRESS

# else

# ips="$ips,$IP\_ADDRESS"

# fi

# done

# echo "Authorizing [$ips] to access cloudsql=$CLOUDSQL"

# gcloud sql instances patch $CLOUDSQL --authorized-networks $ips

# 

# 

# 

# 

# 

# 

# 

# Teste 2 PURCHASES WITH BIGQUERY ML

# 

# 

# 

# 

# 

# TESTE 3 PRODUCCTS USING CLOUD SQL AND SPARK

# 

# 

# 

# 

# 