

NILTON KAZUYUKI UEDA

POSTECH

DATA ANALYTICS

BANCOS DE DADOS PARA BIG DATA

AULA 04

SUMÁRIO

O QUE VEM POR AÍ?	3
HANDS ON	4
SAIBA MAIS	5
O QUE VOCÊ VIU NESTA AULA?	10
REFERÊNCIAS.....	11

EMANSP

O QUE VEM POR AÍ?

Chegamos ao material de aula sobre a integração do Google BigQuery com o Google Colab! Nesta disciplina, vamos explorar como essas duas poderosas ferramentas podem trabalhar em conjunto para análise de dados em larga escala e colaboração eficiente.

O Google BigQuery é um serviço de armazenamento e análise de dados em nuvem altamente escalável, que permite executar consultas rápidas e complexas em conjuntos de dados de qualquer tamanho. Com ele é possível armazenar, compartilhar e consultar grandes volumes de dados de forma eficiente, sem a necessidade de configuração de infraestrutura complexa.

O Google Colab, por sua vez, é um ambiente de notebook colaborativo baseado na nuvem, que permite escrever e executar código Python diretamente no seu navegador. Com o Colab, você pode criar, compartilhar e colaborar em notebooks interativos, que incluem código, texto explicativo, visualizações de dados e muito mais.

Neste curso, vamos explorar como integrar o poder do Google BigQuery com a facilidade de uso do Google Colab. Você aprenderá a como conectar o Colab ao BigQuery, realizar consultas SQL em grandes conjuntos de dados e processar e visualizar os resultados. Além disso, também abordaremos técnicas avançadas, como o uso de bibliotecas Python para análise de dados e a criação de pipelines automatizados.

Prepare-se para mergulhar em um mundo de análise de dados escalável e colaboração eficiente! Ao final desta disciplina, você terá as habilidades necessárias para explorar grandes volumes de dados usando o Google BigQuery e o Google Colab, o que te capacita a realizar análises profundas e tomar decisões informadas baseadas em dados. Vamos começar!

HANDS ON

Agora que você já entende ao assunto, vamos observar as possibilidades que o Google Colab pode trazer para potencializar o uso de grandes volumes de dados que estão dentro do Google BigQuery.

EMEND

SAIBA MAIS

CONHECENDO MELHOR O GOOGLE COLAB

O Google Colab, ou Colaboratory, é uma plataforma baseada na nuvem desenvolvida pelo Google onde é possível escrever, executar e colaborar em código Python. É uma ferramenta popular para aprendizado de máquina, análise de dados e desenvolvimento de projetos baseados em Python.

Aqui estão os principais aspectos do Google Colab:

1. Ambiente de notebook: o Google Colab oferece um ambiente de notebook interativo que permite aos usuários escrever e executar códigos Python. Os notebooks são compostos por células, onde cada célula pode conter código, texto ou visualizações. Nele, é possível criar novas células, editar as existentes e executá-las individualmente.
2. Baseado em Jupyter: o Google Colab é baseado no projeto Jupyter, o que significa que ele oferece uma experiência semelhante ao Jupyter Notebook local. Isso permite que as pessoas escrevam códigos, visualizem os resultados e documentem seus projetos em um único ambiente interativo.

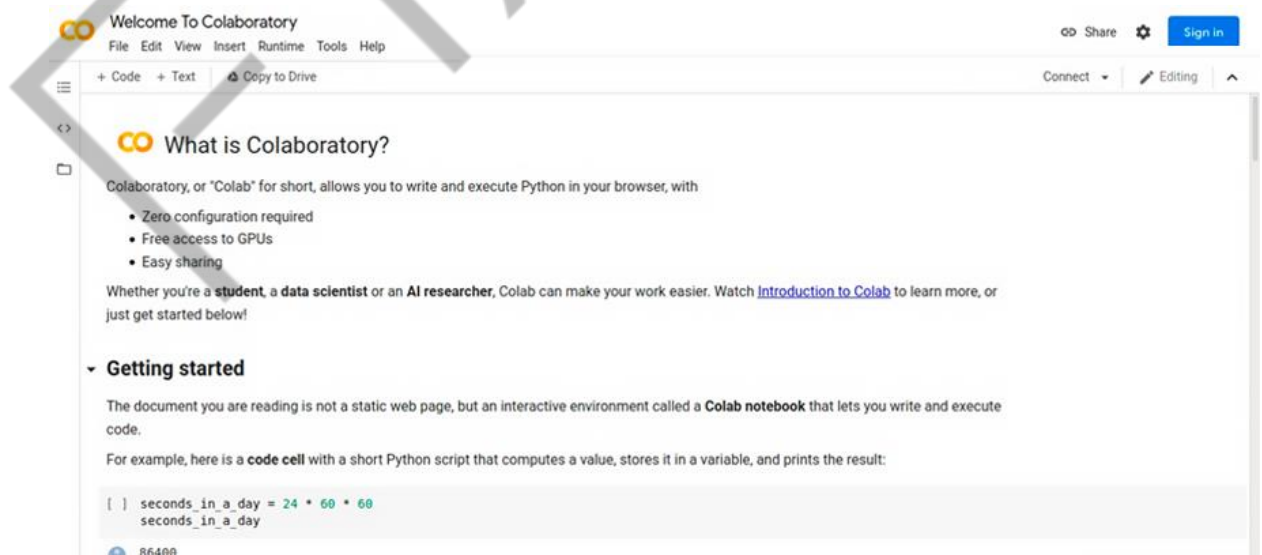


Figura 1 – Colab 1
Fonte: elaborada pelo autor (2023)

3. Acesso à GPU e TPU: uma das principais vantagens do Google Colab é a capacidade de acessar recursos de processamento de alto desempenho, como Unidades de Processamento Gráfico (GPUs) e Unidades de Processamento Tensorial (TPUs). Isso é particularmente útil para tarefas de aprendizado de máquina e processamento de dados intensivos.
4. Armazenamento e importação de dados: o Colab permite importar dados de várias fontes, como o Google Drive, GitHub e até mesmo diretamente do computador local do usuário. Os dados podem ser armazenados temporariamente no ambiente de execução do Colab ou carregados para a memória RAM para análise e processamento.

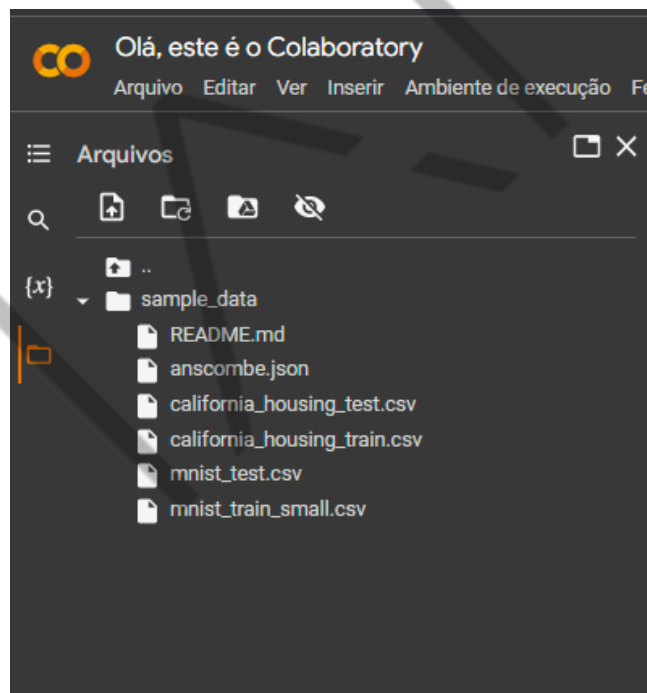


Figura 2 – Colab
Fonte: elaborada pelo autor (2023)

5. Bibliotecas e pacotes pré-instalados: o Colab já vem pré-instalado com muitas bibliotecas e pacotes populares, como NumPy, Pandas, Matplotlib e TensorFlow, o que facilita a análise de dados e o desenvolvimento de modelos de aprendizado de máquina. Além disso, é possível instalar facilmente outras bibliotecas usando o gerenciador de pacotes Python, o pip.
6. Colaboração em tempo real: o Google Colab permite que várias pessoas colaborem em um único notebook ao mesmo tempo. Você pode compartilhar

seus notebooks com outras pessoas por meio de um link e trabalhar em conjunto em projetos de maneira síncrona. Isso facilita a colaboração em equipes de desenvolvimento e o compartilhamento de conhecimento.

7. Integração com outros serviços do Google: o Colab é integrado ao ecossistema de serviços do Google. Isso significa que é possível acessar facilmente o armazenamento do Google Drive, importar conjuntos de dados do Google Sheets, visualizar gráficos usando o Google Charts e muito mais.

O Google Colab oferece uma maneira conveniente e acessível de desenvolver e executar projetos baseados em Python na nuvem. Sua natureza colaborativa, recursos de hardware poderosos e integração com outros serviços do Google o tornam uma escolha popular entre estudantes, pessoas que realizam pesquisas e especialistas em desenvolvimento.

COMO CONECTAR O GOOGLE BIGQUERY COM SEU AMBIENTE NO GOOGLE COLAB

Para conectar o BigQuery ao Google Colab, existem duas maneiras: através da interface do BigQuery e via código Python. Vamos explicar cada uma delas em detalhes.

Conectar o BigQuery ao Google Colab pela interface do BigQuery:

1. Abra o Google Colab no seu navegador.
2. Na barra de menu superior, clique em "Inserir" e, em seguida, em "Código".
3. Será criada uma célula de código vazia. Digite o seguinte código para instalar e importar a biblioteca necessária:

```
!pip install google-cloud-bigquery
```

```
from google.cloud import bigquery
```

4. Acesse o Console do Google Cloud [no site do console](#).
5. No canto superior direito da página, clique no ícone de seleção de projeto e selecione o projeto do Google Cloud em que você está trabalhando.

6. No painel de navegação à esquerda, clique em "BigQuery" para abrir a interface do BigQuery.
7. Na interface do BigQuery, clique no nome do projeto no canto superior esquerdo e selecione o projeto que corresponde ao projeto do Google Cloud que você selecionou anteriormente no Google Colab.
8. Agora você precisa obter as credenciais de autenticação para conectar o BigQuery ao Colab. Para isso, clique no ícone de chave inglesa no canto superior direito da página e selecione "Configurações".
9. Na guia "Credenciais", role para baixo até "Credenciais da API" e clique em "Criar credenciais".
10. Selecione "Chave da conta de serviço" na lista suspensa.
11. Na seção "Papéis", selecione "BigQuery" > "BigQuery Admin" para conceder permissões de administrador.
12. Clique em "Continuar" e, em seguida, em "Criar sem opção de fornecer o consentimento do usuário".
13. Será feito o download de um arquivo JSON contendo suas credenciais. Guarde-o em um local seguro.
14. Volte ao Google Colab e clique em "Arquivo" > "Carregar..." e carregue o arquivo JSON de credenciais que você baixou.
15. Para criar uma conexão com o BigQuery, adicione o seguinte código em uma célula de código no Colab:

```
from google.colab import auth
auth.authenticate_user()
```
16. Execute a célula de código e siga as instruções para fazer login na sua conta do Google.
17. Após autenticar, você pode usar o BigQuery no Colab usando a biblioteca ``google.cloud.bigquery``.

Conectar o BigQuery ao Google Colab via código Python:

1. Abra o Google Colab no seu navegador.
2. Na barra de menu superior, clique em "Inserir" e, em seguida, em "Código".
3. Será criada uma célula de código vazia. Digite o seguinte código para instalar e importar a biblioteca necessária:

```
!pip install google-cloud-bigquery  
  
from google.cloud import bigquery
```

4. Para autenticar a conexão com o BigQuery no Colab, adicione o seguinte código em uma célula de código:

```
from google.colab import auth  
  
auth.authenticate_user()
```

5. Execute a célula de código e siga as instruções para fazer login na sua conta do Google.

6. Depois de autenticar, você pode usar o BigQuery no Colab usando a biblioteca `google.cloud.bigquery`. Por exemplo, você pode executar consultas SQL e obter resultados diretamente no Colab.

Essas são as maneiras de conectar o BigQuery ao Google Colab. A primeira opção permite que você use a interface do BigQuery para executar consultas, enquanto a segunda opção permite que você execute consultas diretamente usando código Python no Colab.

O QUE VOCÊ VIU NESTA AULA?

Nesta aula, tivemos a oportunidade prática de conhecer mais sobre o Google Colab e como esta plataforma de ciência de dados proporciona maiores possibilidades de uso dos dados quando combinada com o poder de processamento do Google BigQuery.

O que achou do conteúdo? Conte-nos no Discord, onde estaremos te esperando para responder eventuais dúvidas e comentários!

REFERÊNCIAS

Modern Data Warehousing with BigQuery (Cloud Next '19), 2019. Disponível em: <<https://www.youtube.com/watch?v=eOQ3YJKgvHE>>. Último acesso em: 20 jun 2023.

BigQuery Documentação Oficial, [s.d.]. Disponível em: <<https://cloud.google.com/bigquery/>>. Último acesso em: 20 jun 2023.

BONNER, ANNE. **Getting started with Google Colab**, 2019. Disponível em: <<https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c>>. Último acesso em: 20 jun 2023.

PARK, Song. **Google Colab 101: Connecting to & Querying BigQuery Data**, 2021. Disponível em: <<https://songjoyce.medium.com/google-colab-101-connecting-to-bigquery-3a2481706907>>. Último acesso em: 20 jun 2023.

WAIBEL, XINRAN. **BigQuery Best Practices**, 2020. Disponível em: <<https://medium.com/google-cloud/bigquery-best-practices-9452c294c9d9>>. Último acesso em: 20 jun 2023.

PALAVRAS-CHAVE

Palavras-chave: Cloud, Data Lake, Data Warehouse, Data Lakehouse, SQL, Google Cloud, BigQuery, Pipeline, Integração de Dados, Tableau, Colab, Python.

EMSE

The background is a dark blue field filled with numerous small, light blue dots. Overlaid on this are several large, wavy, translucent lines in shades of blue and yellow. A vertical line on the left side has a small 'x' at its base. A circle containing the number '7' is positioned in the upper left. A hexagon is located in the lower right. The text 'POSTECH' is centered in the middle of the image.

POSTECH