NILTON KAZUYUKI UEDA

DATA ANALYTICS

BANCOS DE DADOS PARA BIG DATA

AULA 01

SUMÁRIO

O QUE VEM POR AÍ?

O Google BigQuery é uma poderosa ferramenta de análise de dados em escala empresarial, que permite explorar e extrair insights valiosos de grandes conjuntos de dados. Se você está interessado(a) em aprofundar os seus conhecimentos em consultas SQL e deseja aprender a como aplicar essas habilidades ao BigQuery, esta aula é ideal para você.

Iremos mergulhar no mundo do SQL (Structured Query Language) e explorar como ele pode ser utilizado para realizar consultas avançadas em grandes volumes de dados no Google BigQuery. Desde a criação de tabelas e carregamento de dados até a execução de consultas complexas, você aprenderá as melhores práticas e estratégias para otimizar suas análises.

Ao longo deste material, você será guiado(a) passo a passo, desde os conceitos básicos até técnicas mais avançadas de consulta SQL. Você descobrirá como usar o BigQuery de forma eficiente para obter informações valiosas e tomar decisões informadas com base nos dados.

Este material de aula foi projetado para atender tanto a iniciantes em SQL quanto a profissionais experientes que desejam expandir suas habilidades no uso do BigQuery. Independentemente do seu nível de conhecimento, você encontrará exemplos práticos, exercícios desafiadores e dicas úteis para aprimorar suas habilidades e se tornar especialista em consultas SQL no Google BigQuery.

Prepare-se para mergulhar em uma jornada de aprendizado emocionante e comece a dominar o poder das consultas!

HANDS ON

Para encontrar os materiais utilizados no nosso Hands On, acesse o github da disciplina.

Foque em entender tudo o que os códigos transmitem, replique-os em sua máquina local, e teste com muita dedicação para que o aprendizado seja definitivo!

SAIBA MAIS

O Google BigQuery foi projetado como um data warehouse “nativo da nuvem”. Ele foi criado para atender às necessidades de organizações orientadas a dados em um primeiro mundo em nuvem.

O BigQuery é o armazenamento de dados em nuvem sem servidor, altamente escalonável e econômico do Google Cloud Platform. Ele permite consultas super-rápidas em escala de petabytes usando o poder de processamento da infraestrutura do Google. Como não há infraestrutura para os clientes gerenciarem, eles podem se concentrar em descobrir insights significativos usando SQL familiar sem a necessidade de um administrador de banco de dados. Também é econômico porque eles pagam apenas pelo processamento e armazenamento que usam.

O Google BigQuery faz parte de uma suíte de soluções em nuvem que é ofertada pela empresa Google.

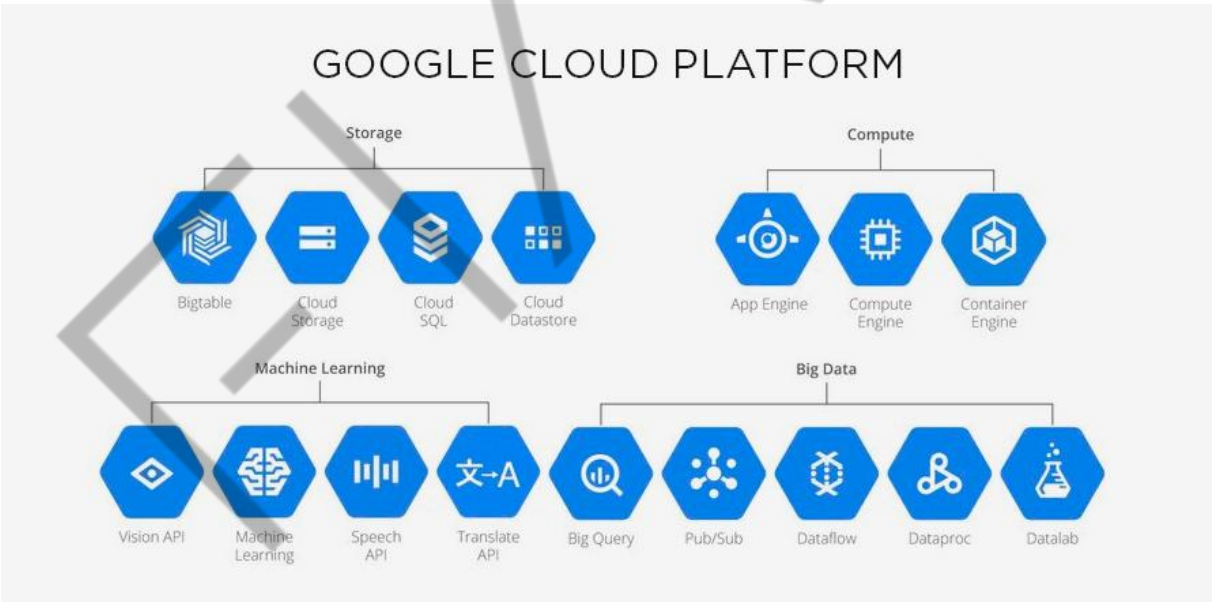

Figura 1 – Google Cloud Platform
Fonte: Dynatrace (2018), adaptada por FIAP (2023)

O BigQuery está dentro de um grupo de soluções com foco em Big Data, é possível notar que dentro do portfólio de soluções existem diversos produtos para finalidades específicas.

ENTENDENDO MELHOR O GOOGLE BIGQUERY E SUAS INTEGRAÇÕES

O BigQuery é compatível com várias maneiras de receber dados no armazenamento gerenciado. O método de ingestão específico depende da origem dos dados. Por exemplo, algumas fontes de dados no GCP, como Cloud Logging e Google Analytics, oferecem suporte a exportações diretas para o BigQuery. Basta ativar o recurso, fornecer o acesso e os dados começam a ser enviados automaticamente para o Google BigQuery na frequência desejada.

Já o BigQuery Data Transfer Service permite a transferência de dados para o BigQuery de aplicativos Google SaaS (por exemplo Google Ads e Cloud Storage), Amazon S3 e outros data warehouses (Teradata, Redshift).

Os dados de streaming, como registros ou dados de dispositivos IoT, podem ser gravados no BigQuery usando pipelines do Cloud Dataflow, jobs do Cloud DataProc ou diretamente usando a API de ingestão de stream do BigQuery.

ACESSANDO A INTERFACE DO GOOGLE BIGQUERY

Para obter acesso à interface do Google BigQuery, será necessário realizar configuração de cadastro de usuário (caso ainda não tenha acesso) e ativar sua conta.

É importante ressaltar neste momento de primeiro cadastro, que atualmente o Google Cloud possui uma oferta específica de experimentação de serviços e fornece um valor aproximado de USD 300 dólares para experimentação dentro do período de 6 meses.

Para realizar o cadastro de um usuário pela primeira vez na Google Cloud Platform (GCP), siga as etapas abaixo:

1. Acesse o site oficial da Google Cloud Platform e clique no botão "Get Started for Free" (Comece de graça) ou "Try Free" (Experimente gratuitamente).
2. Será solicitado que você faça login com a sua conta do Google. Se você não tiver uma conta do Google, clique no link "Create account" (Criar conta) para criar uma nova conta.

3. Após fazer login, você será direcionado para a página de boas-vindas da GCP. Nessa página, você precisará fornecer algumas informações, como nome, país de residência e detalhes de pagamento. O cadastro inicial geralmente inclui um período de teste gratuito e, em seguida, é necessário fornecer informações de pagamento para continuar usando a plataforma após esse período.
4. Após fornecer as informações necessárias, clique em "Start my free trial" (Iniciar minha avaliação gratuita) ou em um botão semelhante, dependendo da opção disponível.
5. Você será redirecionado(a) para o console da GCP, onde poderá criar seus projetos, configurar recursos e acessar os serviços da plataforma.
6. Antes de começar a usar a GCP, é importante revisar as configurações de faturamento e limites de uso para garantir que você esteja ciente de quaisquer encargos ou restrições. Você pode acessar essas configurações por meio do painel do console da GCP.

Depois que seguir o processo de cadastro, você deve navegar até o recurso do BigQuery ou acesse diretamente pelo link do site do Google Cloud.

A figura 2 – “Tela inicial”, representa o que você verá quando entrar no recurso.

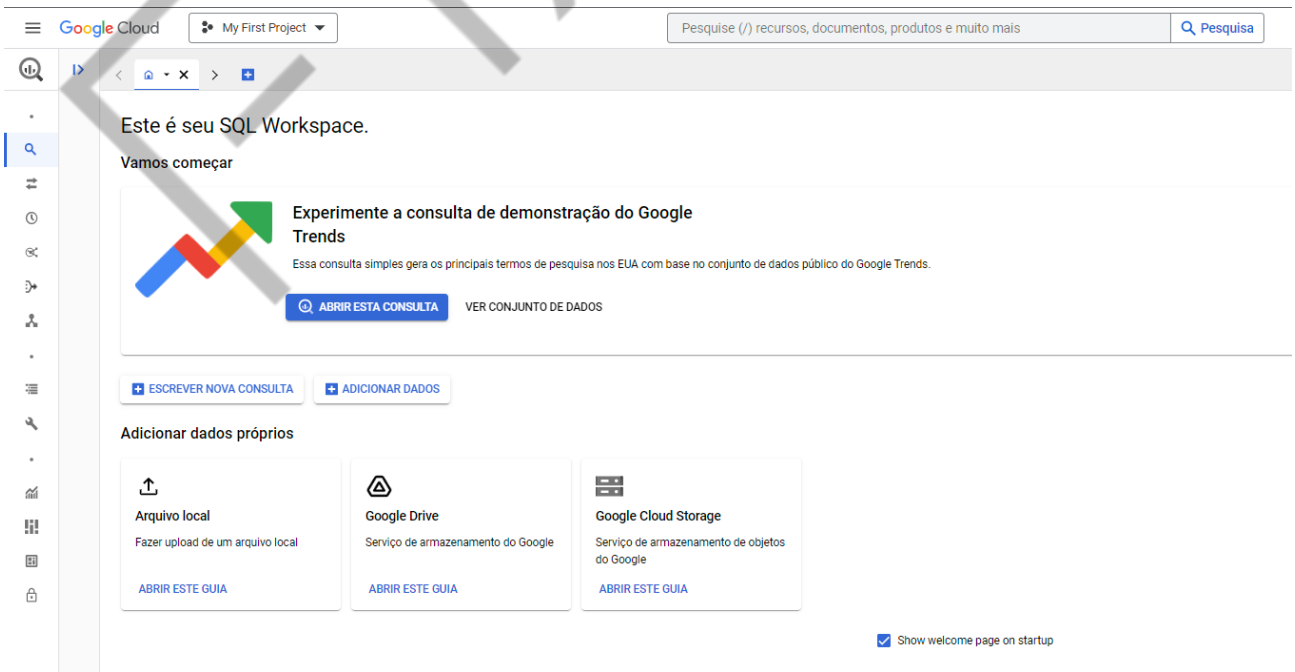

Figura 2 – Tela inicial
Fonte: elaborada pelo autor (2023)

Observe que, já na tela inicial, você tem a opção de visualizar alguns conjuntos de dados disponíveis por padrão e também a opção de levar seus dados para o BigQuery através dos importadores de Arquivo Local, Google Drive ou Google Cloud Storage.

Em breve iremos explorar estas possibilidades.

EXPLORANDO A INTERFACE DO GOOGLE BIGQUERY

Na figura 3 – “Interface do Google BigQuery”, vamos entender melhor como funciona a interface do BigQuery. Após criar o projeto, ele aparecerá para você no canto superior esquerdo **(1)**. Logo abaixo, terá uma lista de projeto fixos do BigQuery, entre esses, o “bigquery-public-data” **(2)**. A seta à esquerda do nome “bigquery-public-data” nos permite expandir a lista de todas as bases disponíveis na logo abaixo.

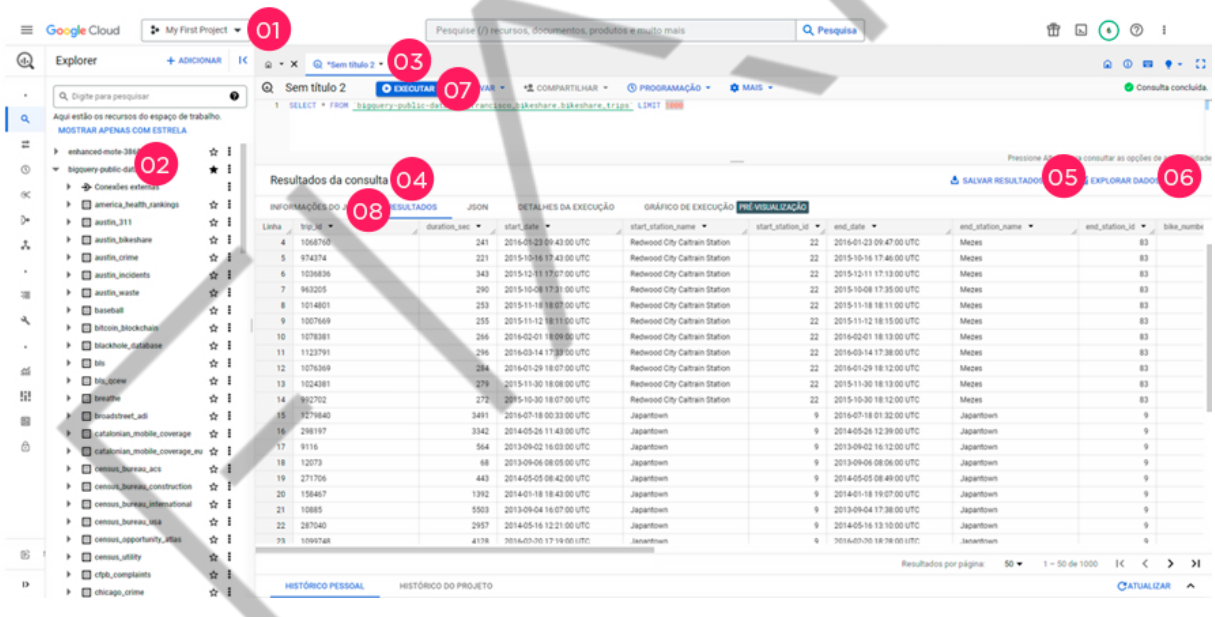

Figura 3 – Interface do Google BigQuery
Fonte: elaborada pelo autor (2023), adaptada por FIAP (2023)

Na imagem, cada botão possui uma finalidade específica, então temos:

1. Navegar pela lista de projetos disponíveis na sua conta.
2. Visualizar a lista de conjuntos de dados disponíveis.
3. Local onde você escreve sua consulta SQL, para escrever e executar consultas.

4. Onde você consegue visualizar seus resultados de consultas executadas.
5. Salvar seus resultados de execução (CSV, JSON, Google Sheets).
6. Exploração de dados com Google Sheets, Looker Studio, Geo Viz e Google Colab.
7. Executar a consulta SQL que você escreveu no console.
8. Visualizar as informações técnicas de execução do resultado da consulta.

EXPLORANDO CONJUNTOS DE DADOS PÚBLICOS NO GOOGLE BIGQUERY

Agora que você acessou o Google BigQuery, vamos entender melhor como funcionam os objetos disponíveis e seus conjuntos de dados públicos.

Os conjuntos de dados públicos são basicamente bases de dados disponibilizados pela equipe do Google Cloud para fins de testes e experimentação. Dessa forma, novos analistas de dados conseguem ter um ambiente para estudo logo após a realização do cadastro.

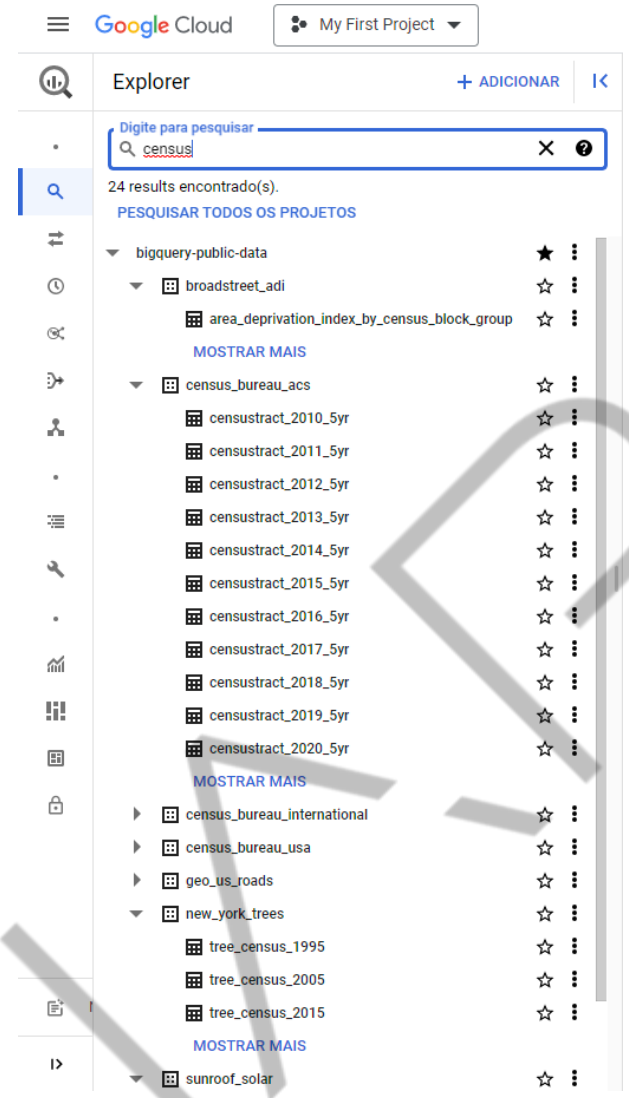

Figura 4 – Bases de dados

Fonte: elaborada pelo autor (2023)

Na figura 4 – “Bases de dados”, estamos pesquisando por bases de dados disponíveis públicas que contenham a palavra “census”. A pesquisa retornou 24 resultados encontrados e esses dados podem ser utilizados para fins de experimentação e de estudos.

ENTENDENDO CONSULTAS SQL NO GOOGLE BIGQUERY

Agora que você já entendeu o que é o Google BigQuery, como é seu funcionamento e sua estrutura principal de interface, vamos colocar a mão na massa através de consultas SQL.

O SQL é uma das linguagens mais reconhecidas no mundo todo para extração, processamento e manipulação de dados. Ela é uma importante ferramenta de trabalho e pode trazer bastante autonomia e produtividade para os trabalhos que envolvam análise de dados.

Como a grande maioria das linguagens de consumo de dados, o SQL é uma linguagem que precisa de um compilador (ou interface) para ser executado. Neste caso, estamos usando o BigQuery como camada de acesso aos dados.

Dentro de uma consulta SQL, existem várias cláusulas e comandos que podem ser usados para personalizar e refinar os resultados. Abaixo estão os principais comandos e cláusulas utilizados em uma consulta SQL:

- SELECT: o comando SELECT é usado para especificar as colunas que você deseja selecionar na consulta. Você pode selecionar colunas específicas separando-as por vírgulas ou usar o asterisco (*) para selecionar todas as colunas da tabela.
- FROM: a cláusula FROM é aplicada para especificar a tabela ou tabelas a partir das quais você deseja recuperar os dados. Ela permite que você especifique a fonte dos dados para a consulta.
- WHERE: a cláusula WHERE é utilizada para filtrar os resultados com base em uma condição específica. Você pode usar operadores de comparação (como igual a, maior que, menor que) e operadores lógicos (como AND, OR, NOT) para construir condições complexas.
- GROUP BY: a cláusula GROUP BY é usada para agrupar os resultados com base em uma ou mais colunas. Ela permite que você faça agregações, como contar, somar ou encontrar valores máximos/mínimos, dentro de cada grupo usando funções de agregação, como COUNT, SUM, MAX, MIN.
- HAVING: a cláusula HAVING é aplicada em conjunto com a cláusula GROUP BY para filtrar grupos com base em condições específicas. Ela funciona de maneira semelhante à cláusula WHERE, mas opera nos resultados após a execução da cláusula GROUP BY.
- ORDER BY: a cláusula ORDER BY é utilizada para classificar os resultados em uma determinada ordem. Você pode organizar os resultados em ordem

ascendente (ASC) ou descendente (DESC) com base em uma ou mais colunas.

- JOIN: o comando JOIN é usado para combinar registros de duas ou mais tabelas com base em uma condição de correspondência entre elas. Existem diferentes tipos de junções, como INNER JOIN, LEFT JOIN, RIGHT JOIN e FULL JOIN, que determinam como os registros são combinados.
- DISTINCT: o comando DISTINCT é utilizado para retornar apenas valores distintos das colunas selecionadas. Ele elimina duplicatas nos resultados da consulta.

Esses são os principais comandos e cláusulas usados em uma consulta SQL. Combinando-os de maneira adequada, você pode personalizar e refinar os resultados de acordo com suas necessidades específicas. Vamos experimentá-los juntos?

ESCREVENDO A ESTRUTURA DE UMA CONSULTA SQL

Agora que você já entendeu os principais comandos SQL que existem dentro uma consulta SQL, podemos colocar a mão na massa e usar o poder do BigQuery para processar os dados de maneira massiva e otimizada.

Alguns exemplos para facilitar seu entendimento:

Consulta SQL sem agregação:

```
SELECT [nome_da_coluna_1], [nome_da_coluna_2], [nome_da_coluna_3]
FROM [nome_do_esquema].[nome_da_tabela]
```

Consulta SQL sem agregação e com condição de filtro e ordenação de resultado por campo:

```
SELECT [nome_da_coluna_1], [nome_da_coluna_2], [nome_da_coluna_3]
FROM [nome_do_esquema].[nome_da_tabela]
WHERE [condição]
ORDER BY [campo_de_ordenacao]
```

Consulta SQL com agregação de SOMA por 3 campos:

SELECT [nome_da_coluna_1], [nome_da_coluna_2], [nome_da_coluna_3], SUM([nome_do_campo] AS [nome_do_campo]

FROM [nome_do_esquema].[nome_da_tabela]

GROUP BY [nome_da_coluna_1], [nome_da_coluna_2], [nome_da_coluna_3]

Consulta SQL com agregação de SOMA por 3 campos e ordenação de resultado por campo:

SELECT [nome_da_coluna_1], [nome_da_coluna_2], [nome_da_coluna_3], SUM([nome_do_campo] AS [nome_do_campo]

FROM [nome_do_esquema].[nome_da_tabela]

GROUP BY [nome_da_coluna_1], [nome_da_coluna_2], [nome_da_coluna_3]

ORDER BY [campo_de_ordenacao]

O QUE VOCÊ VIU NESTA AULA?

O que é o Google Cloud Platform, o que é o BigQuery e como usar a infraestrutura do Google Cloud para consulta e processamento avançado de grandes volumes de dados dentro da nuvem.

Não se esqueça de participar da comunidade do Discord! Lá você pode conversar com os docentes, seus colegas e pessoa de Community Management. Explore!

REFERÊNCIAS

A brief intro to full stack performance monitoring on Google Cloud Plataform, 2018. Disponível em: <https://www.dynatrace.com/news/blog/a-brief-intro-to-full-stack-performance-monitoring-on-google-cloud-platform/>. Último acesso em: 20 jun 2023.

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

Google BigQuery (SQL) 101, 2021. Disponível em: <https://medium.com/basedosdados/bigquery-101-8b39da1ce52b>. Último acesso em: 20 jun 2023.

Modern Data Warehousing with BigQuery (Cloud Next '19), 2019. Disponível em: <https://www.youtube.com/watch?v=eOQ3YJKgvHE>. Último acesso em: 13 jun 2023.

THALLAM, Rajesh. **New Blog Series - BigQuery Explained: An Overview**, 2020. Disponível em: <https://medium.com/google-cloud/bigquery-explained-overview-357055ecfda3>. Último acesso em: 13 jun 2023.

PALAVRAS-CHAVE

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