



**Desafio Técnico - Data Engineer**

**Random User Extract**

**v.1.0.0**

**Nome:** Robton Rodrigues Brangaitis

**Data:** 02/06/2022

## Versão

Nome do Projeto	Versão
Random User Extract	1.0.0

## Dicionário de Termos:

Sigla/Abrev.	Nome	Descrição
GCS	Google Cloud Storage	Serviço de armazenamento de objetos
BigQuery	Google BigQuery	Datawarehouse para armazenamento de dados
Composer	Google Cloud Composer	Ferramenta de Orquestração de Pipelines
lib	Libraries	Libraries (Bibliotecas) do Python
Bucket	Bucket	Espaço para armazenamento de arquivos dentro do Google Cloud Storage
CSV	Comma-separated values	Arquivos de texto separado por vírgulas
JSON	JavaScript Object Notation	Arquivo para troca de informações
DAG	Directed Acyclic Graph	Tarefas a serem executadas

## Objetivo do Projeto

O projeto Random User Extract tem por objetivo extrair dados da API Pública Random User e disponibilizar os dados em uma tabela dentro do Google BigQuery (SQL).

Para tal, foram utilizadas as seguintes ferramentas:

- Google Cloud Composer: Orquestração do pipeline
- Google Cloud Storage: Salvamento de arquivos utilizados durante a execução;
- Google BigQuery: Estruturação dos dados extraídos.

Detalhes referente a arquitetura e o desenvolvimento dos pipelines de dados serão descritos nos próximos tópicos.

## Perguntas de Negócios

### Pergunta 01:

Quantos homens e mulheres – gêneros: masculino e feminino, possuem ao total de forma porcentual?

### Pergunta 02:

Quantas pessoas possuem o mesmo nome no mesmo país?

### Pergunta 03:

Qual distribuição das pessoas por gênero e país?

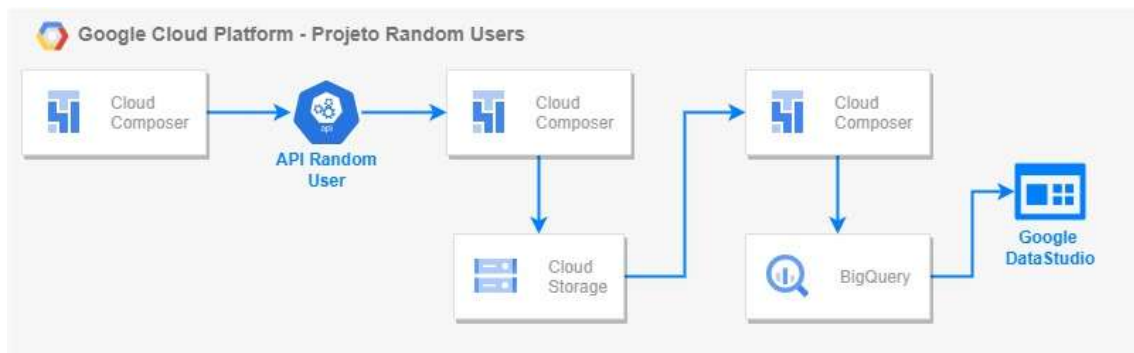
### Pergunta 04:

Quantas pessoas possuem mais de 50 anos distribuídos por país e gênero?

Os gráficos com as respostas estão no Google Data Studio, link para acessar encontra-se na seção: Informações Adicionais.

## Arquitetura

A arquitetura está construída da seguinte forma:



O Cloud Composer faz uma requisição HTTP para a API, o resultado da requisição retorna para o Composer e salva em um arquivo Json dentro do Storage.

O Composer busca o arquivo Json dentro do Storage, converte para um arquivo CSV e salva no BigQuery. O Google Data Studio consulta os dados no BigQuery.

## Desenvolvimento

Foram criados dois scripts para a execução do projeto. O primeiro: `extract_random_users.py` para extração de dados da API e disponibilização dos dados em ambiente de **Bronze**.

Ele faz uma conexão com a API *Random Users* usando a *lib Requests*. O retorno da requisição é convertido em json para armazenamento na pasta: `raw_data` dentro do bucket.

Logo após utilizando a *lib Pandas*, converte para CSV e salva dentro da pasta: `stage` do bucket do projeto. Depois busca o arquivo no GCS, lê os dados e escreve em uma tabela do *BigQuery*.

O segundo script: `load_silver_random_users.py`, foi criado para carregar os dados do ambiente de **Bronze** para o **Silver**.

Nesse script, foi utilizado operadores do Airflow (Composer) específico para *BigQuery*.

Ambos os scripts podem ser consultados no GitHub, através do link:

<https://github.com/robtbrang/desafiosdedados/tree/main/airflow/dags>

Para fins de testes e execução foram adicionados dez mil (10k) registros. Sendo que cada uma das execuções, buscava um total de mil (1k) registros.

## Visualização do Ambiente

Pasta onde estão salvas as DAG:

Cloud Storage

Navegador

Monitoramento

Configurações

Marketplace

←

Detalhes do bucket

ATUALIZAR

SAIBA MAIS

us-central1-random-users-cb1d8d93-bucket

Local

Classe de armazenamento

Acesso público

Proteção

us-central1 (Iowa)

Standard

Sujeito a ACLs de objeto

Nenhum

OBJETOS

CONFIGURAÇÃO

PERMISSÕES

PROTEÇÃO

CICLO DE VIDA

Intervalos

>

us-central1-random-users-cb1d8d93-bucket

>

dags

FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR PASTA

GERENCIAR RETENÇÕES

FAZER O DOWNLOAD

EXCLUIR




Filtrar apenas pelo prefixo do nome

▼

Filtro

Filtrar objetos e pastas

Mostrar dados excluídos

	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso público
<input type="checkbox"/>	 airflow_monitoring.py	729 B	text/x-python	1 de jun...	Standard	1 de jun. de 202...	Não público
<input checked="" type="checkbox"/>	 extract_random_users.py	10,3 KB	application/octet-stream	2 de jun...	Standard	2 de jun. de 202...	Não público
<input type="checkbox"/>	 load_silver_random_users.py	4,2 KB	application/octet-stream	2 de jun...	Standard	2 de jun. de 202...	Não público

## Tela inicial do Airflow (Composer):

The screenshot shows the Airflow (Composer) dashboard. At the top, there's a navigation bar with the Airflow logo and links for DAGs, Browse, Admin, and Docs. The main heading is "DAGs". Below it, there are filters for "All" (3), "Active" (3), and "Paused" (0). A search bar labeled "Filter DAGs by tag" and "Search DAGs" is present. The main table lists DAGs with columns: DAG, Owner, Runs, Schedule, Last Run, Next Run, and Recent Tasks. Three DAGs are listed: "airflow\_monitoring" (owner: airflow, runs: 253, schedule: None, last run: 2022-06-02, 16:16:23), "etl\_extract\_users" (owner: Robton R Brangaitis, runs: 10, schedule: 1 day, 0:00:00, last run: 2022-06-02, 13:14:15, next run: 2022-06-01, 18:15:17), and "load\_silver\_random\_users" (owner: Robton R Brangaitis, runs: 2, schedule: 1 day, 0:00:00, last run: 2022-06-02, 14:21:38, next run: 2022-06-02, 14:16:07). At the bottom, there's a pagination bar showing page 1 of 1.

## Resultados das Execuções

### Bucket com os arquivos Json:

The screenshot shows the AWS S3 console interface for a bucket named "us-central1-random-users-cb1d8d93-bucket". The bucket is located in the "us-central1 (Iowa)" region. The storage class is "Standard", and the access is "Public". The protection is "None". The console shows the "OBJETOS" (Objects) tab, which lists the contents of the bucket. The objects are listed in a table with columns: Nome, Tamanho, Tipo, Criado, Classe de armazenamento, Última modificação, and Acesso público. There are 6 objects listed, all of which are JSON files created on June 2, 2022, at 00:28, 00:29, 00:31, 00:32, 00:33, and 00:34. Each object is 1.9 MB in size and has a storage class of "Standard". The access is "Not public".

Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso público
2022-06-02_00_28_random_user...	1,9 MB	application/json	1 de jun...	Standard	1 de jun. de 202...	Não público
2022-06-02_00_29_random_user...	1,9 MB	application/json	1 de jun...	Standard	1 de jun. de 202...	Não público
2022-06-02_00_31_random_user...	1,9 MB	application/json	1 de jun...	Standard	1 de jun. de 202...	Não público
2022-06-02_00_32_random_user...	1,9 MB	application/json	1 de jun...	Standard	1 de jun. de 202...	Não público
2022-06-02_00_33_random_user...	1,9 MB	application/json	1 de jun...	Standard	1 de jun. de 202...	Não público
2022-06-02_00_34_random_user...	1,9 MB	application/json	1 de jun...	Standard	1 de jun. de 202...	Não público

## Buckets com os arquivos csv:

**us-central1-random-users-cb1d8d93-bucket**

Local: us-central1 (Iowa) | Classe de armazenamento: Standard | Acesso público: Sujeito a ACLs de objeto | Proteção: Nenhum

OBJETOS | CONFIGURAÇÃO | PERMISSÕES | PROTEÇÃO | CICLO DE VIDA

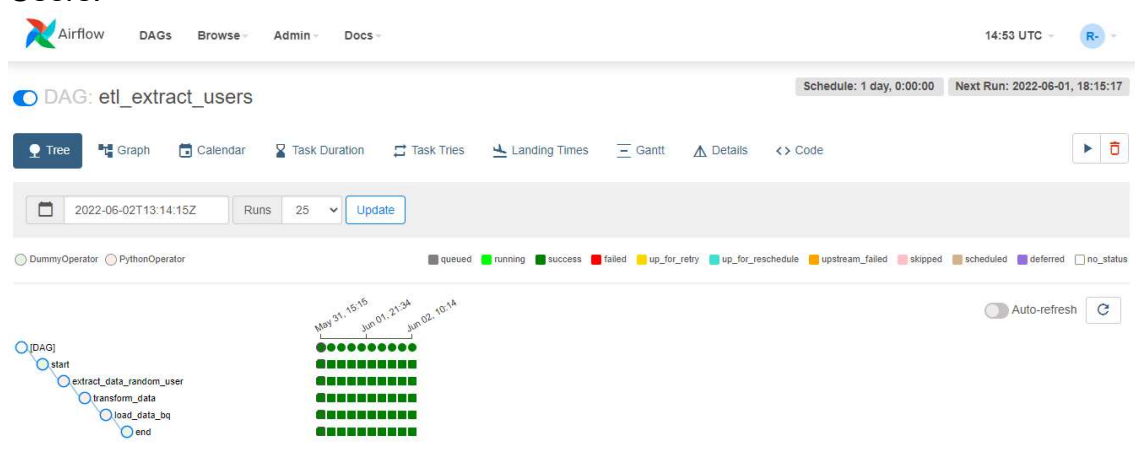
Intervalos > us-central1-random-users-cb1d8d93-bucket > stage

FAZER UPLOAD DE ARQUIVOS | CARREGAR PASTA | CRIAR PASTA | GERENCIAR RETENÇÕES | FAZER O DOWNLOAD | EXCLUIR

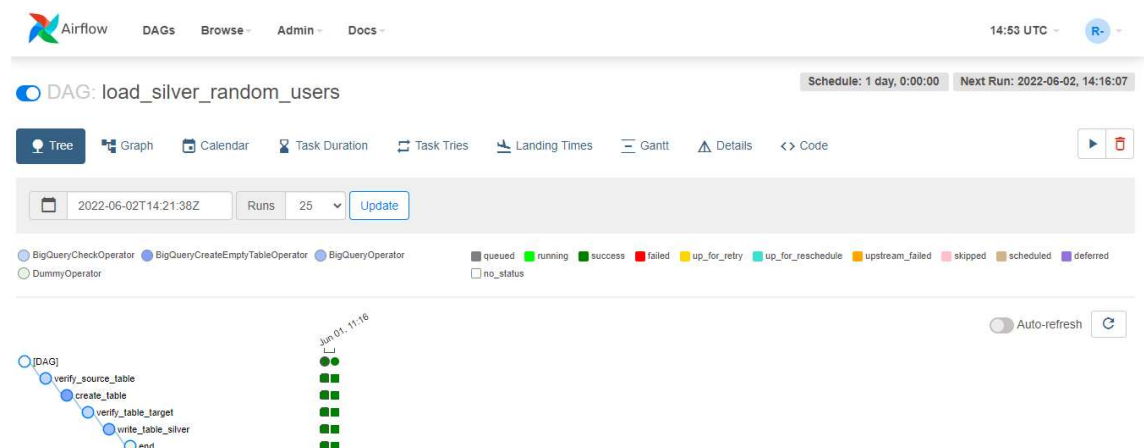
Filtrar apenas pelo prefixo do nome | Filtro: Filtrar objetos e pastas | Mostrar dados excluídos

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso público	Hist
<input type="checkbox"/>	2022-06-01_22-57_random_users...	1,1 KB	text/csv	1 de jun....	Standard	1 de jun. de 202...	Não público	— ↓ ⋮
<input type="checkbox"/>	2022-06-01_23-04_random_users...	1,1 KB	text/csv	1 de jun....	Standard	1 de jun. de 202...	Não público	— ↓ ⋮
<input type="checkbox"/>	2022-06-01_23_25_random_user...	1,1 KB	text/csv	1 de jun....	Standard	1 de jun. de 202...	Não público	— ↓ ⋮
<input type="checkbox"/>	2022-06-01_23_28_random_user...	1,1 KB	text/csv	1 de jun....	Standard	1 de jun. de 202...	Não público	— ↓ ⋮
<input type="checkbox"/>	2022-06-01_23_34_random_user...	1,1 KB	text/csv	1 de jun....	Standard	1 de jun. de 202...	Não público	— ↓ ⋮
<input type="checkbox"/>	2022-06-01_23_53_random_user...	1,1 KB	text/csv	1 de jun....	Standard	1 de jun. de 202...	Não público	— ↓ ⋮

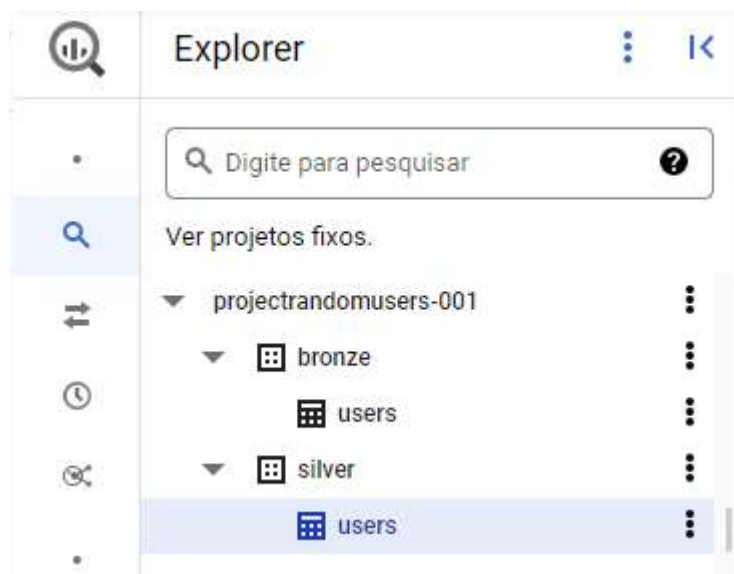
## DAG do Airflow (Composer) – Extração de Dados da API Random Users:



## DAG do Airflow (Composer) – Carregamento dos Dados em Ambiente Silver:



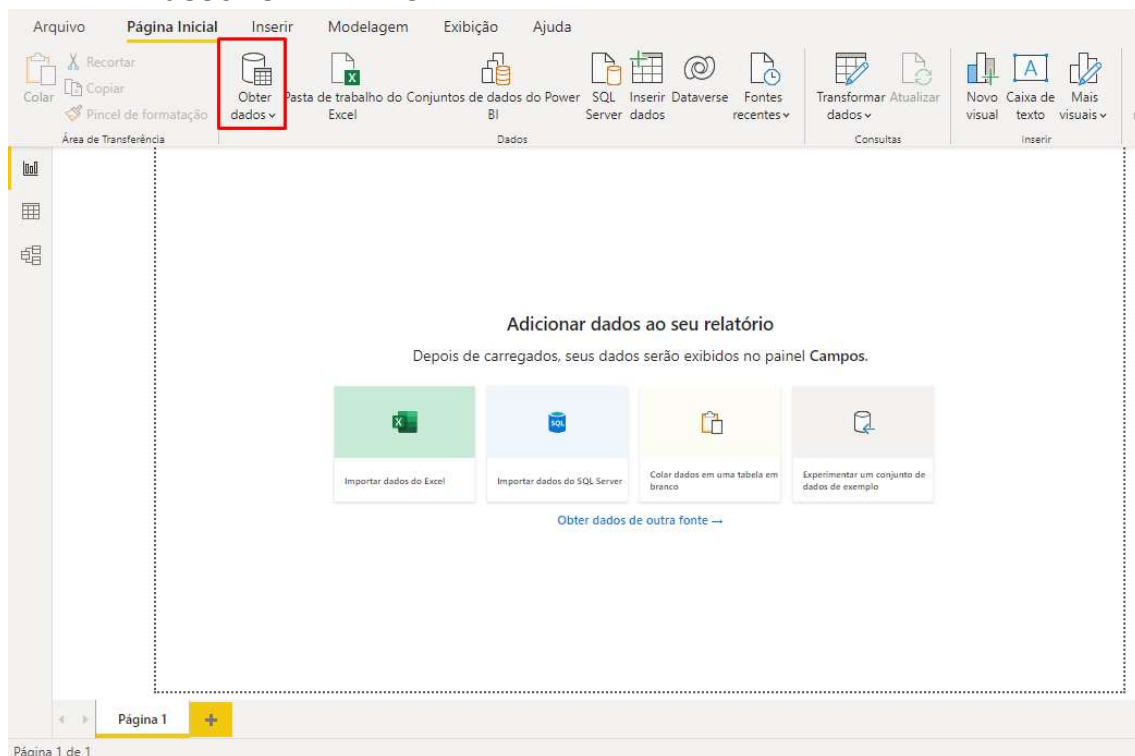
## Datasets e Tabelas no BigQuery:



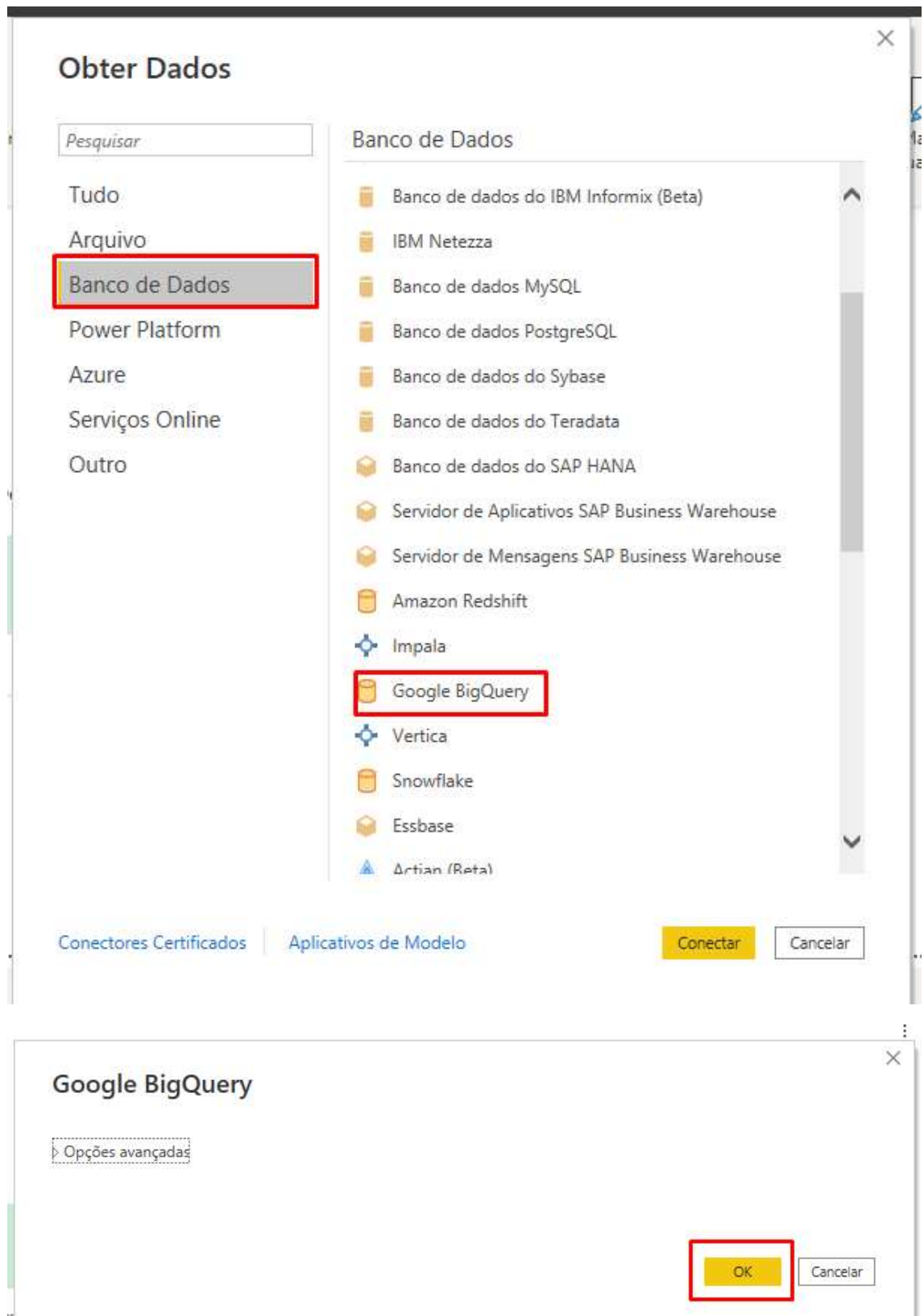
## Conectando BigQuery ao Microsoft Power BI

Foi criada uma conta de serviço para conexão ao *BigQuery*. Abaixo mostro como conectá-lo ao Microsoft Power BI Desktop.

### 1º Passo: Clicar em Obter Dados no Power BI:

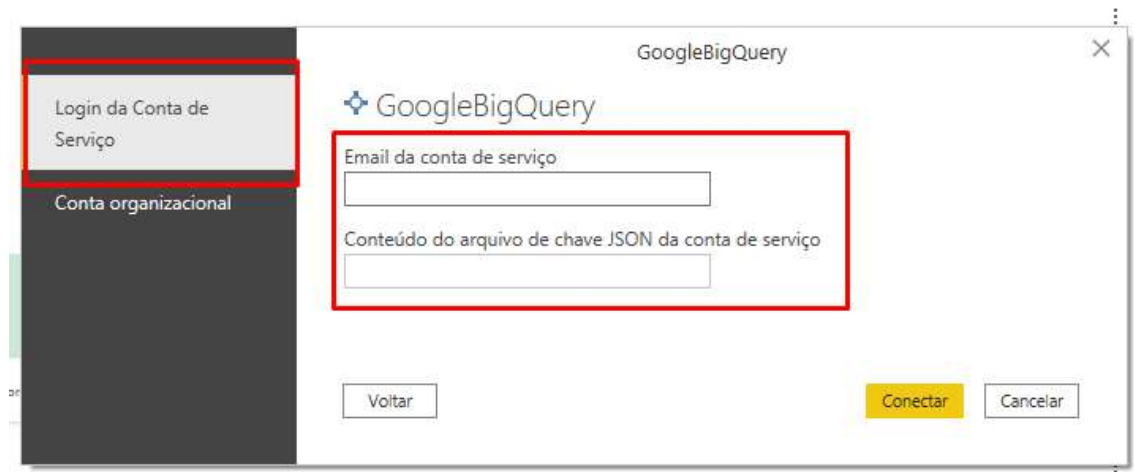


**2º Passo:** Selecionar Google BigQuery e após clicar em Conectar e após em Ok.

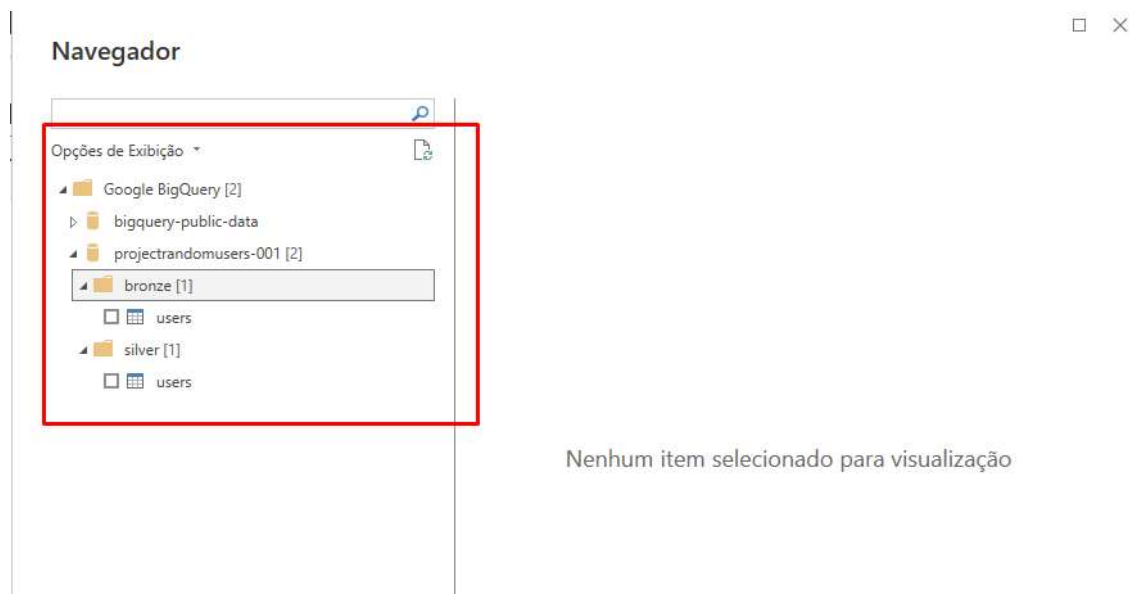




**3º Passo:** Selecionar Login da Conta de Serviço e preencher os dados solicitados referente a Chave Json (que será encaminhada juntamente com arquivos):



**4º Passo:** Clique em Ok e na nova tela que for aberta selecione o dataset e a tabela desejada.



## Informações Adicionais

- Link do Dashboard no Data Studio

<https://datastudio.google.com/reporting/750f8e54-c7bb-437d-a51f-3a35ae9cdbcf>

## Próximos Passos

Neste tópico irei abordar algumas implementações que poderão contribuir para melhoria do processo:

- Configurar Google Cloud Data Catalog para implementação de Governança de Dados nos datasets do BigQuery;
- Esteira de CI/CD para deploy de scripts de desenvolvimento;
- Conectar tasks do Airflow (Composer) a um serviço de monitoramento, onde possa ser avisado caso dê erro na execução da DAG.