# Pós -graduação Ciência de Dados — PUC Rio Sprint 3 — Engenharia de Dados

Aluno: Frederico Araújo Soares

## **Objetivo**

Este trabalho tem como resultado, a criação de algumas tabelas referentes a dados de chuva/precipitação da cidade de Goiânia – GO.

O Pipeline irá ser executado para carregar os dados tratados em algumas tabelas no serviço da cloud do Google, serviço chamado BigQuery.

O objetivo deste trabalho é finalizar o pipeline para disponibilizar os dados em algumas consultas para responder as seguintes perguntas:

- É possível identificar eventos extremos com picos de precipitação, temperaturas ou rajadas de ventos?
- Investigar como a radiação solar influencia outras variáveis como tempertatura e umidade.

#### **Detalle dos dados**

O dados serão obtidos no sítio eletrônico do INMET (<a href="https://portal.inmet.gov.br/dadoshistoricos">https://portal.inmet.gov.br/dadoshistoricos</a>). Em seguida será realizado um tratamento para obter somente os dados de Goiânia.

#### Fluxo de execução do Pipeline

Será executado um Notebook dentro do BigQuery chamado "mvp3-engenharia-dados-coleta-dados.ipynb". Este notebook está presente nos arquivos aqui deste repositório github. Os seguintes passos serão executados:

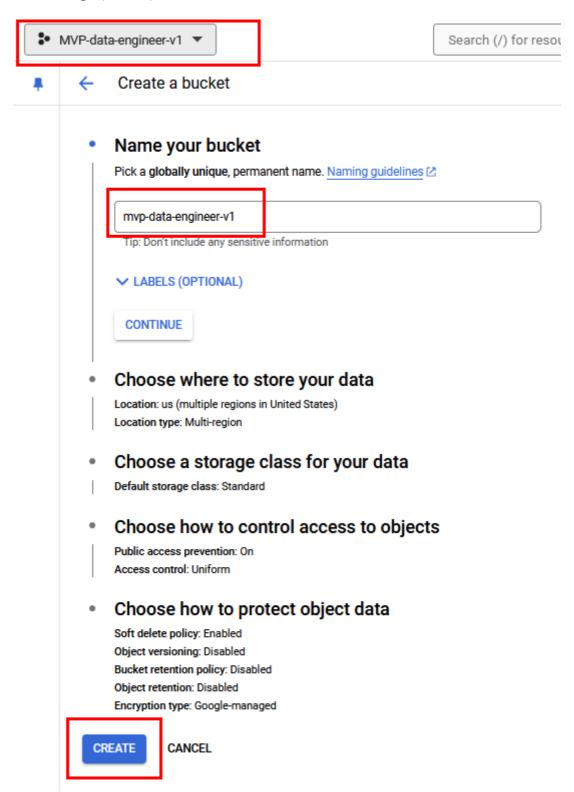
- 1. Coleta dos dados
- 2. Extração e seleção do csv de Goiânia
- 3. Tratamento para dados faltantes
- 4. Upload para o Google Cloud Storage (bucket) dos dados atuais processados e histórico dos dados executados.
- 5. Carregamento dos dados nas tabelas do BigQuery
- 6. Apresentação dos dados no BigQuery e Power BI Desktop

## Configuração da Infraestrutura

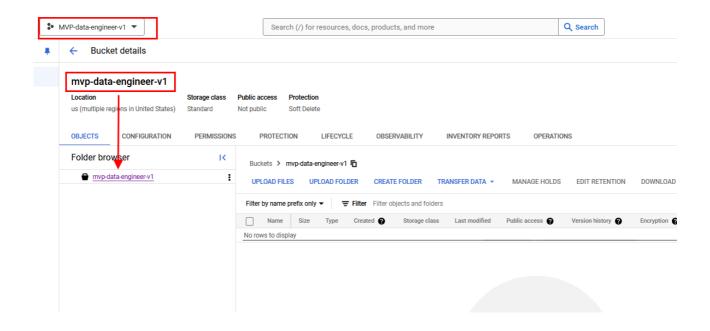
#### Criação do Projeto

Primeiramente é necessário criar o projeto no Google Cloud. No caso criei um projeto chamado MVP-data-engineer-v1.

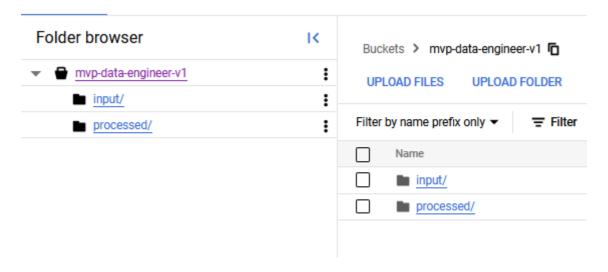
### Criação do storage (bucket):



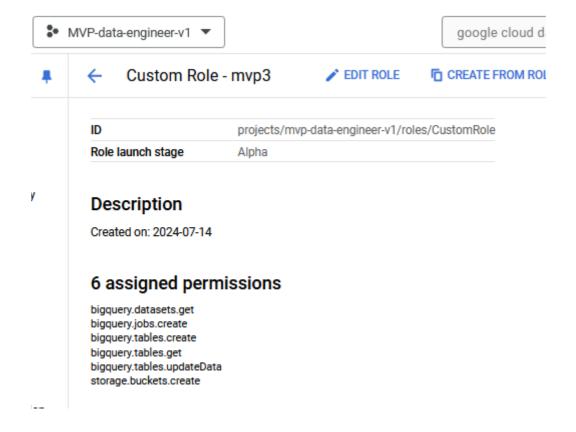
#### **Detalhes do bucket:**



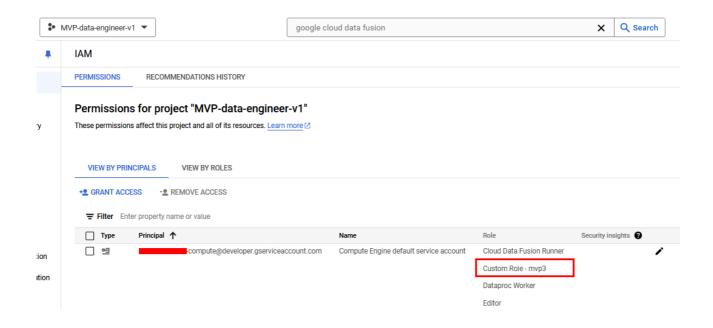
#### Apresentação dos diretórios



#### Criado um Custom Role com as seguintes permissões:



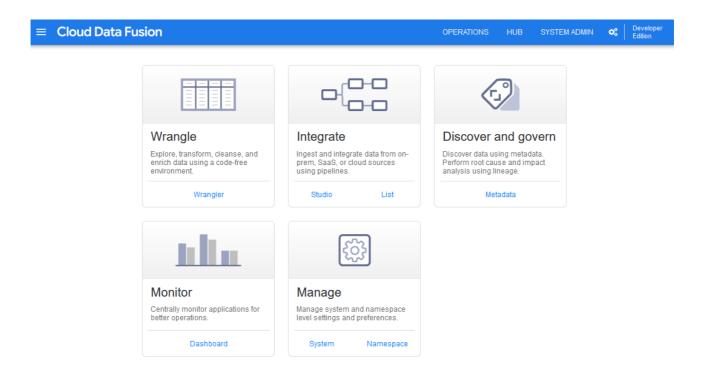
### No usuário do projeto, adicionei o Role criado:



### **Pipeline Google Cloud Data Fusion**

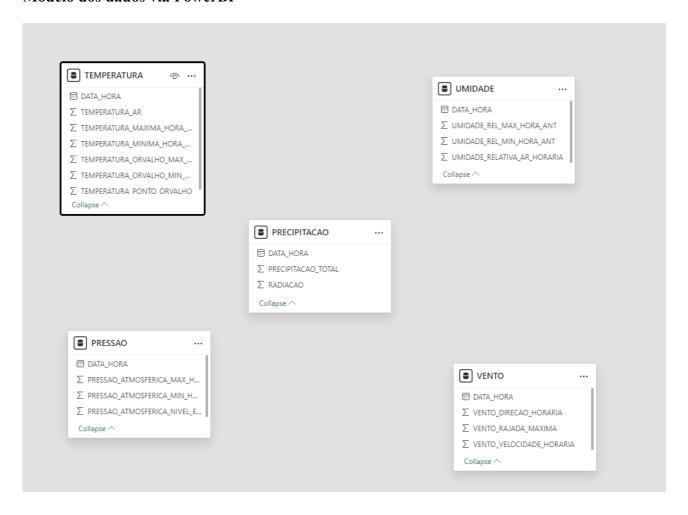
Foi criado inicialmente um pipeline com o Fusion, porém não consegui identificar a tempo um erro de transformação dos dados do CSV para a tabela do BigQuery.

Após criar a instância do Fusion, clique para visualizar a instancia:



Studio??????

### Modelo dos dados via PowerBI

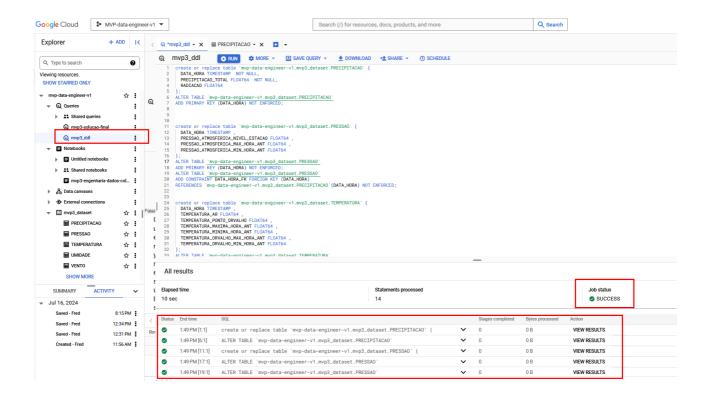


# Criação no BigQuery das Tabelas

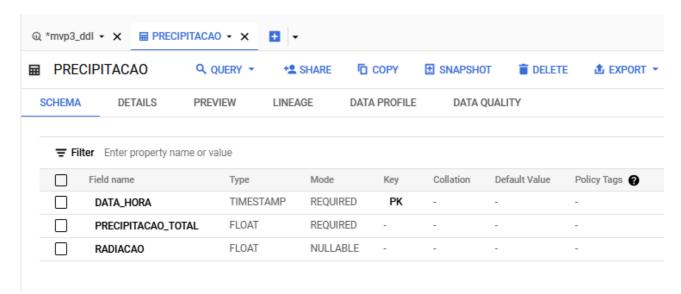
O arquivo mvp3 ddl.sql possui o sql de criação das tabelas.

Aqui é importante ressaltar as limitações do BigQuery. As chaves primárias e estrangeiras são "not enforced" de acordo com a documentação:

https://cloud.google.com/bigquery/docs/information-schema-table-constraints#limitations



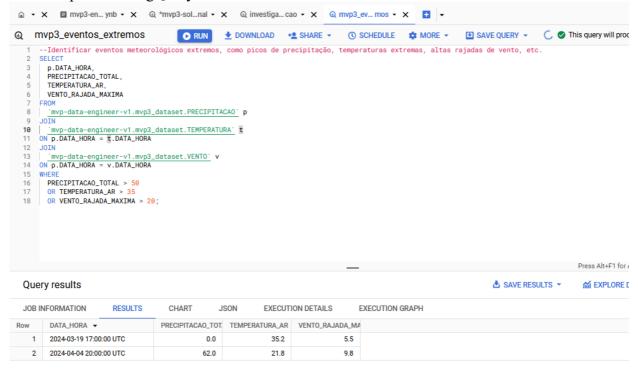
#### Exemplo da Tabela PRECIPITACAO:



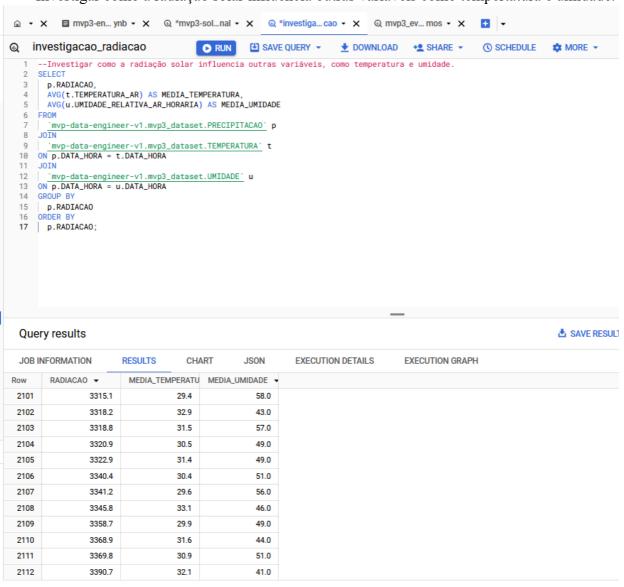
## SOLUÇÃO: Analisando se as perguntas dos objetivos foram atendidas

• É possível identificar eventos extremos com picos de precipitação, temperaturas ou rajadas de ventos?

Resposta via BigQuery:



• Investigar como a radiação solar influencia outras variáveis como tempertatura e umidade.



## Dados Apresentados via Power BI acessando o BigQuery

