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.

Detalhe dos dados

O dados serão obtidos no sítio eletrônico do INMET (https://portal.inmet.gov.br/dadoshistoricos). 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:

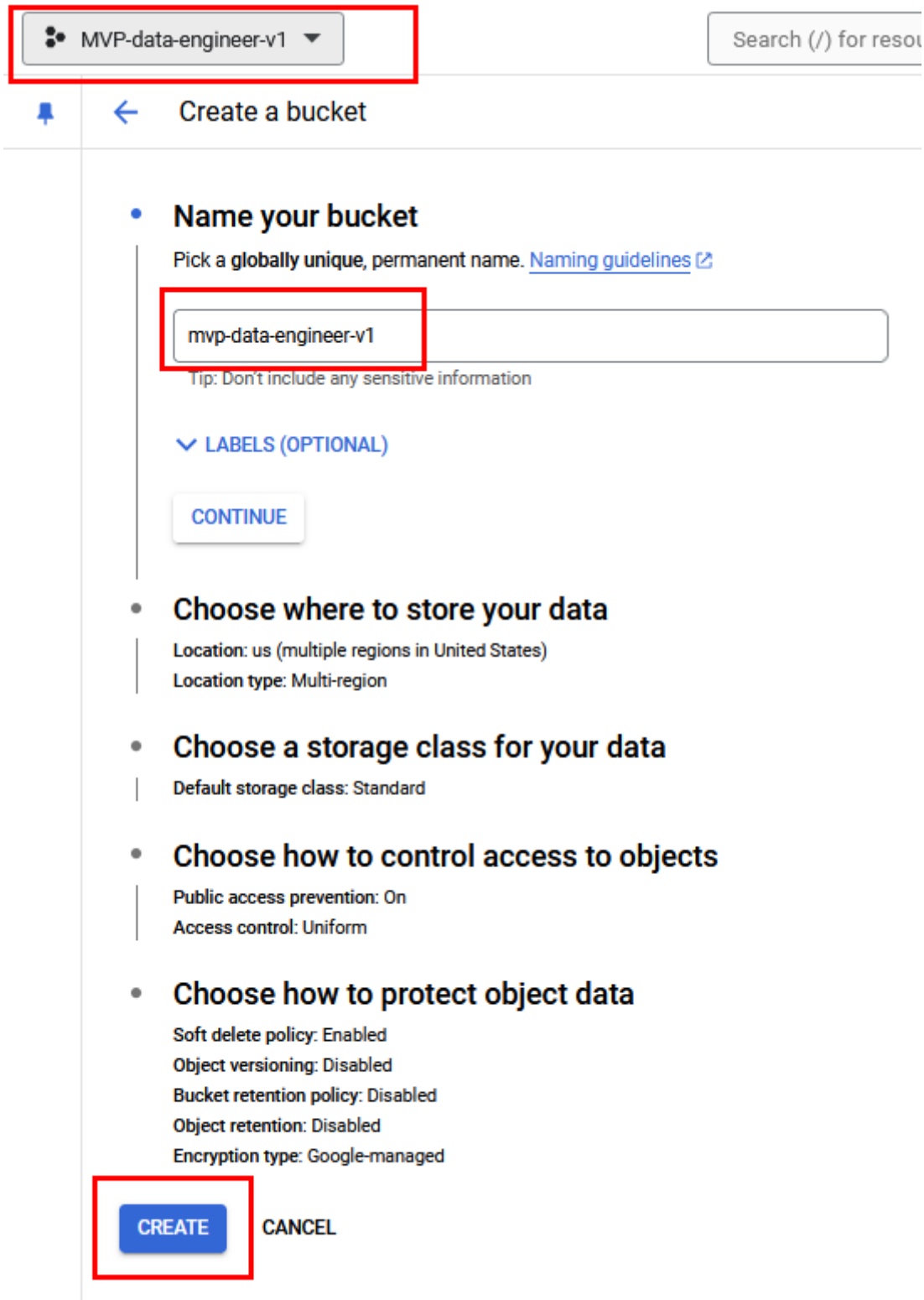
1. Coleta dos dados
2. Extração e seleção do csv de Goiânia
3. Tratamento para dados faltantes
4. Upload para o Google Cloud Storage (bucket) dos dados atuais processados e histórico dos dados executados.
5. Carregamento dos dados nas tabelas do BigQuery
6. Apresentação dos dados no BigQuery e Power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

Configuração da Infraestrutura

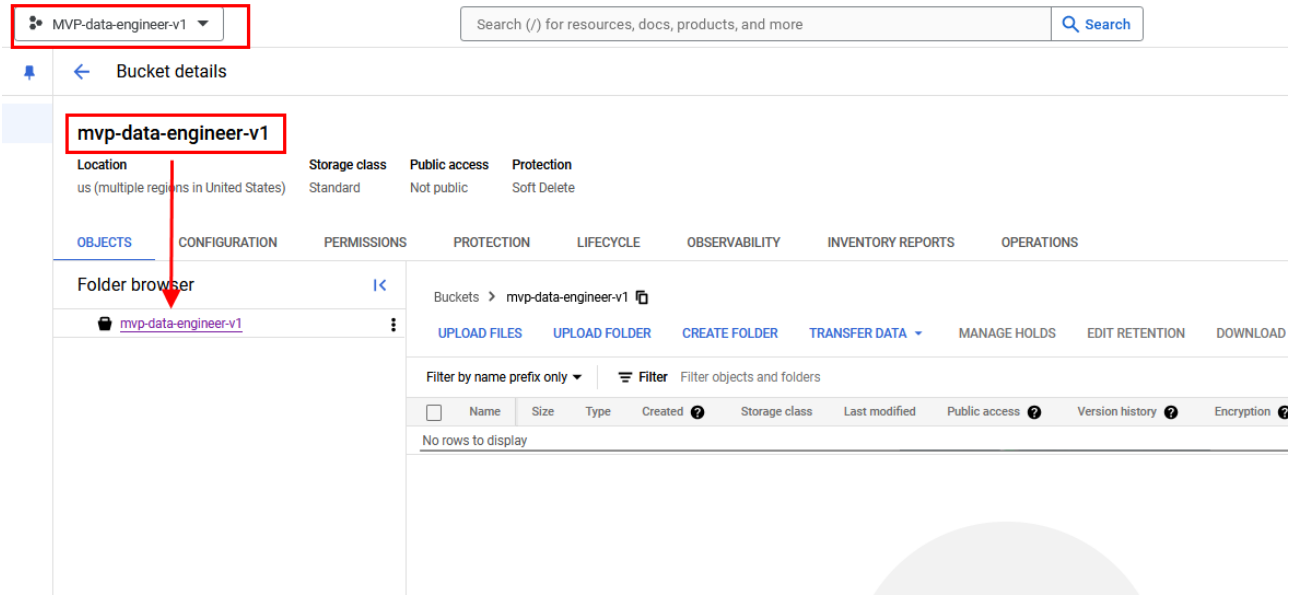
Criação do Projeto

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

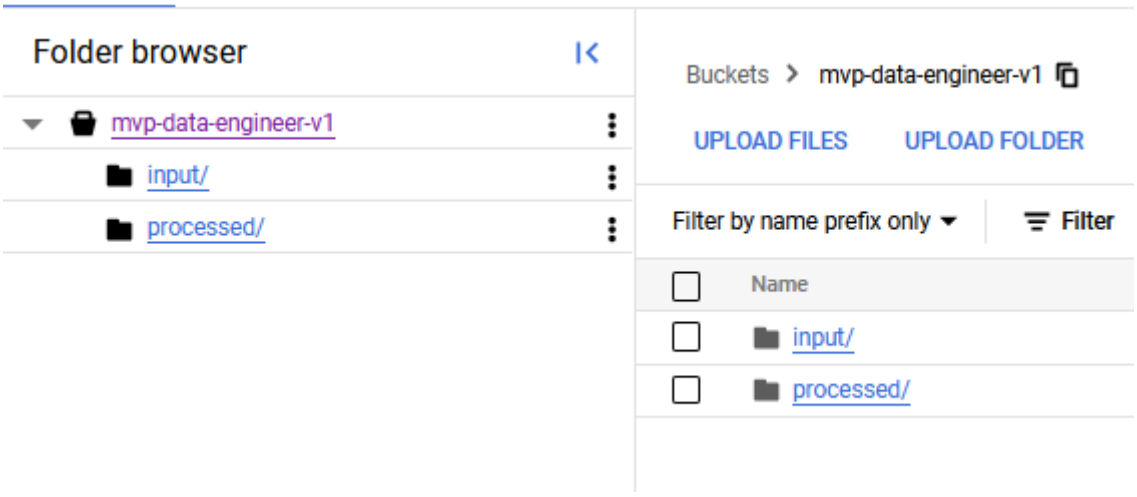
Criação do storage (bucket):



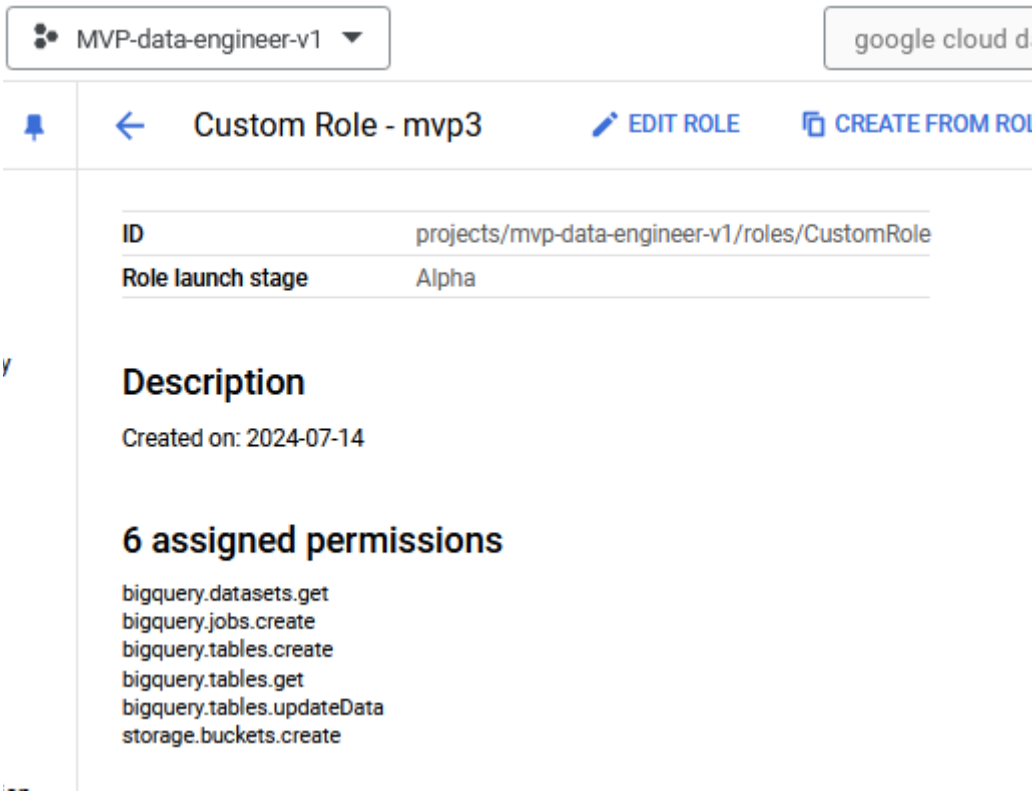
Detalhes do bucket:



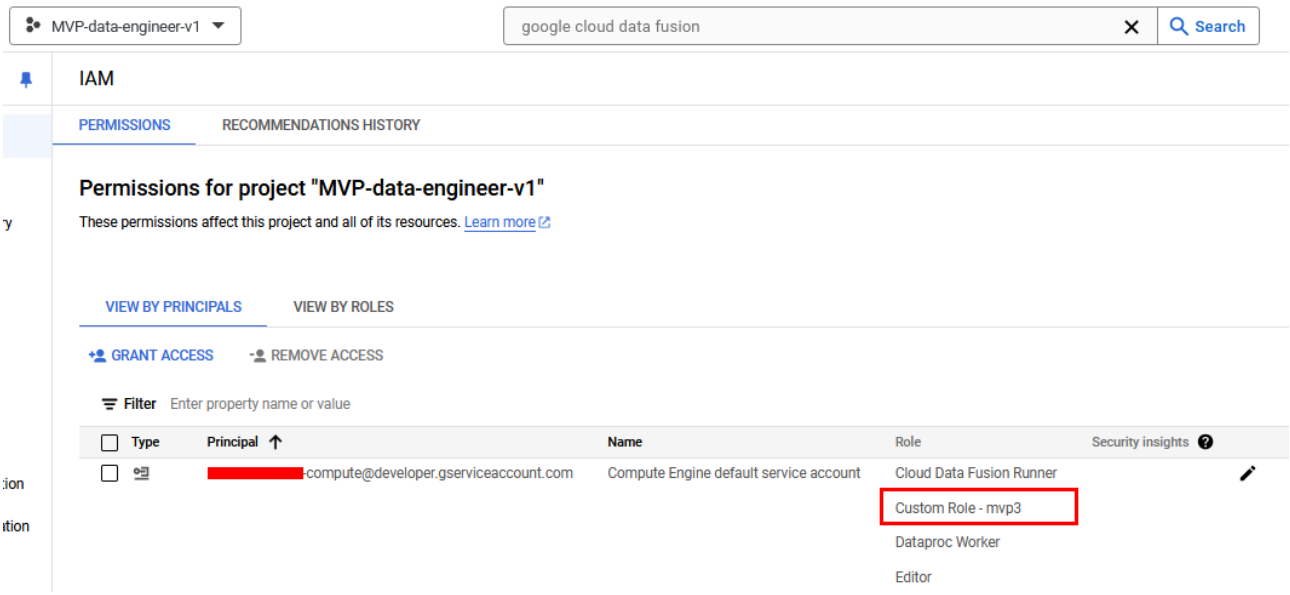
Apresentação dos diretórios



Criado um Custom Role com as seguintes permissões:



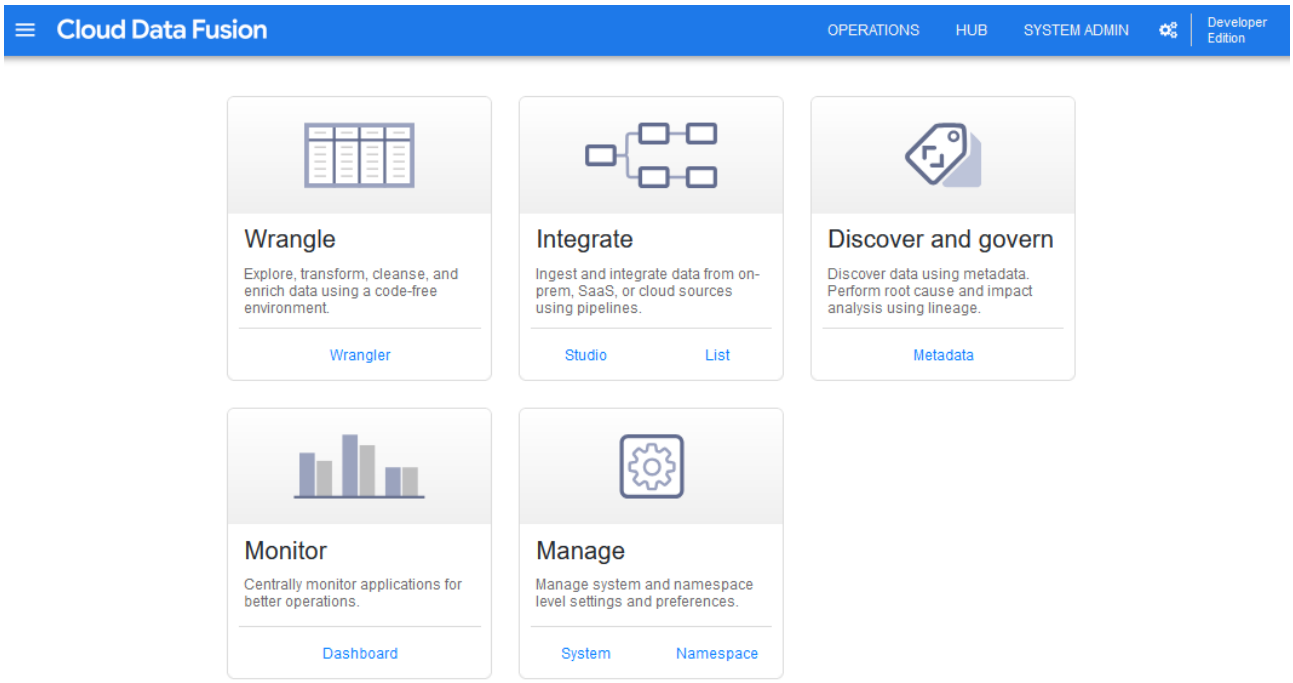
No usuário do projeto, adicionei o Role criado:



Pipeline Google Cloud Data Fusion

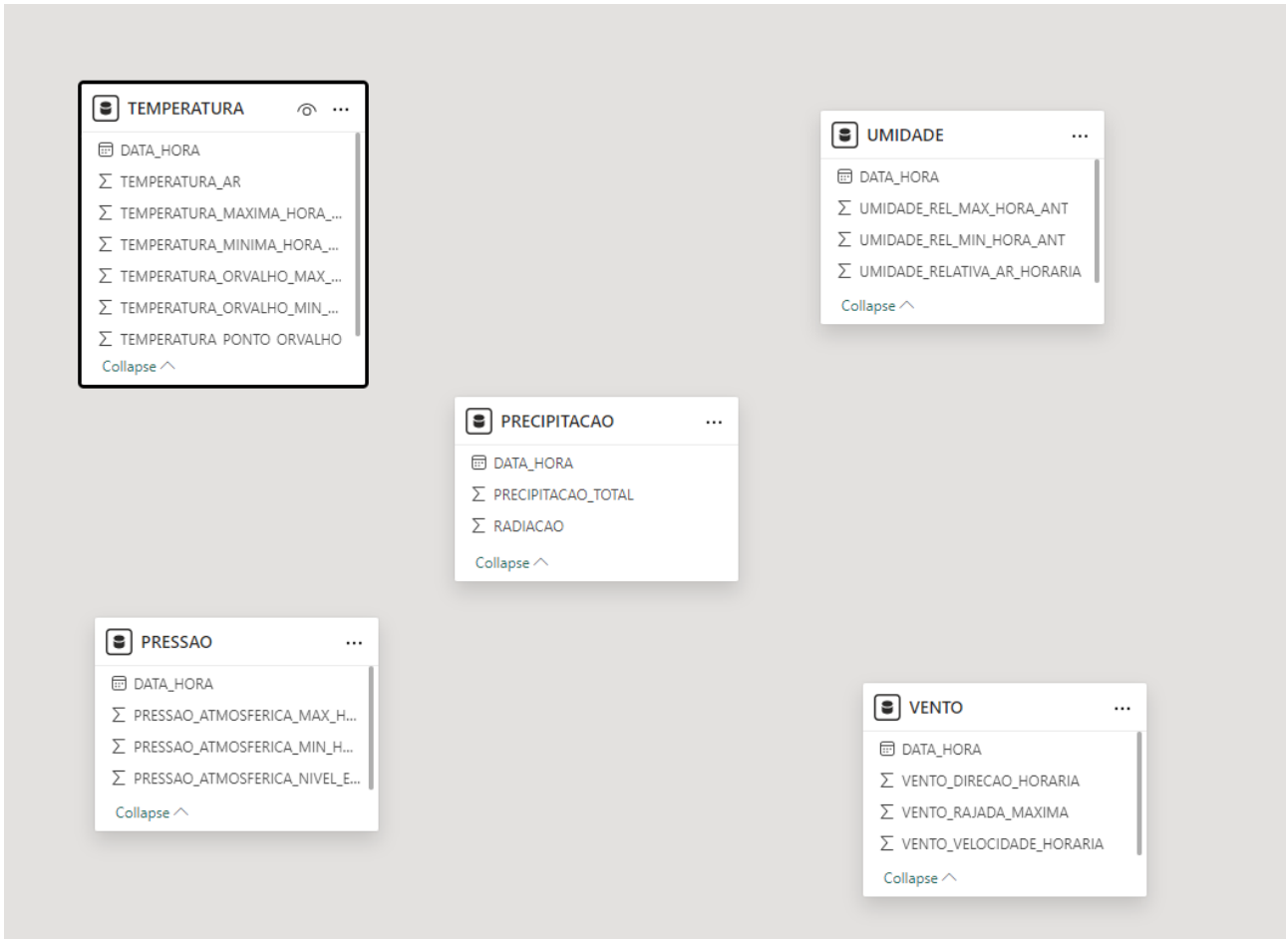
Foi criado inicialmente um pipeline com o Fusion, porém não consegui identificar a tempo um erro de transformação dos dados do CSV para a tabela do BigQuery.

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



Studio??????

Modelo dos dados via PowerBI

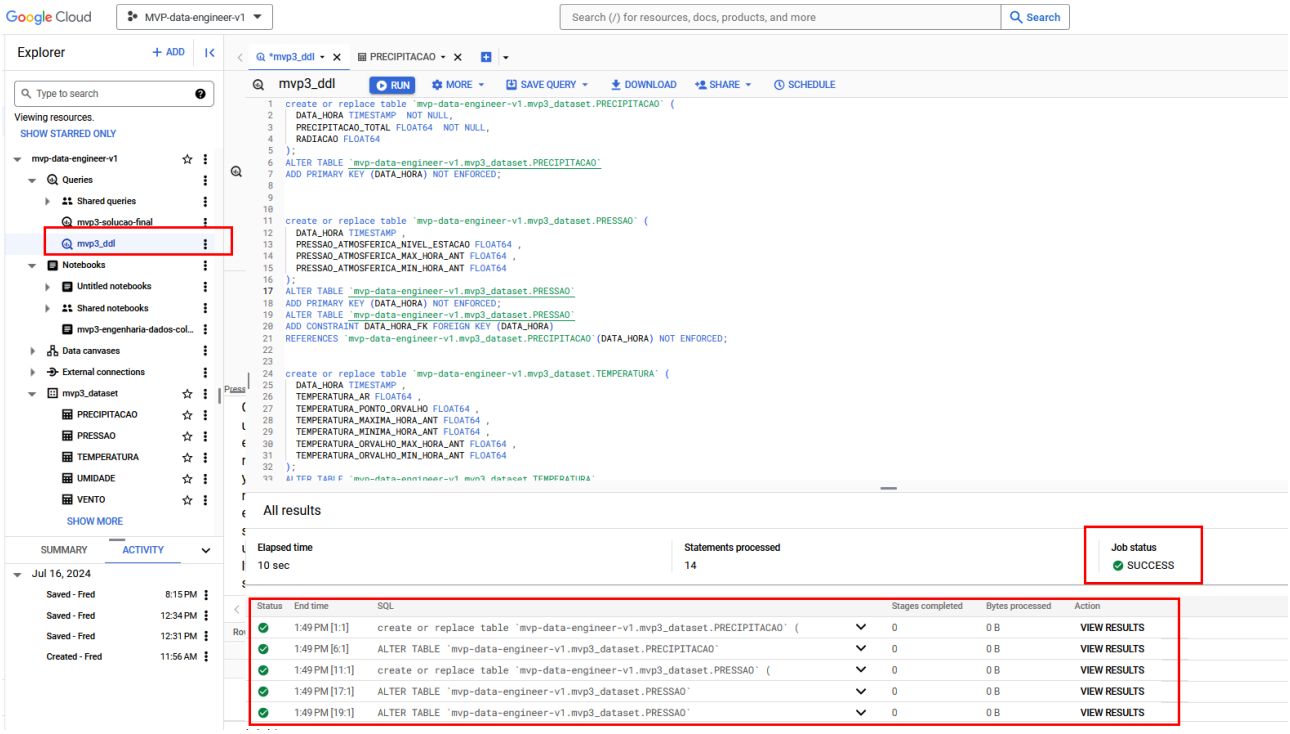


Criação no BigQuery das Tabelas

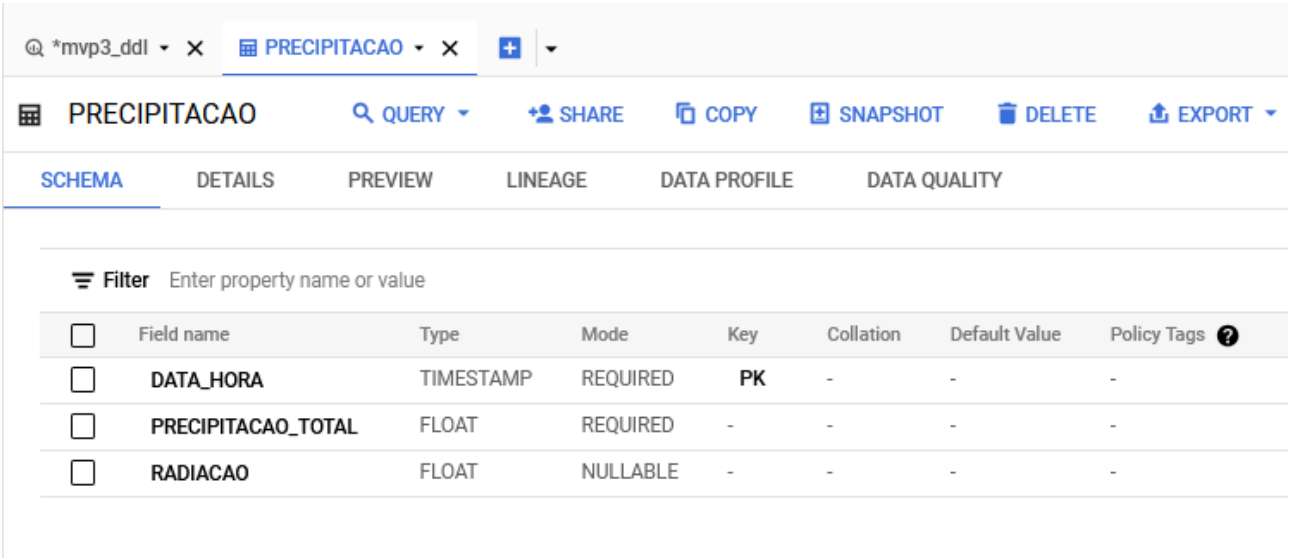
O arquivo mvp3_ddl.sql possui o sql de criação das tabelas.

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`

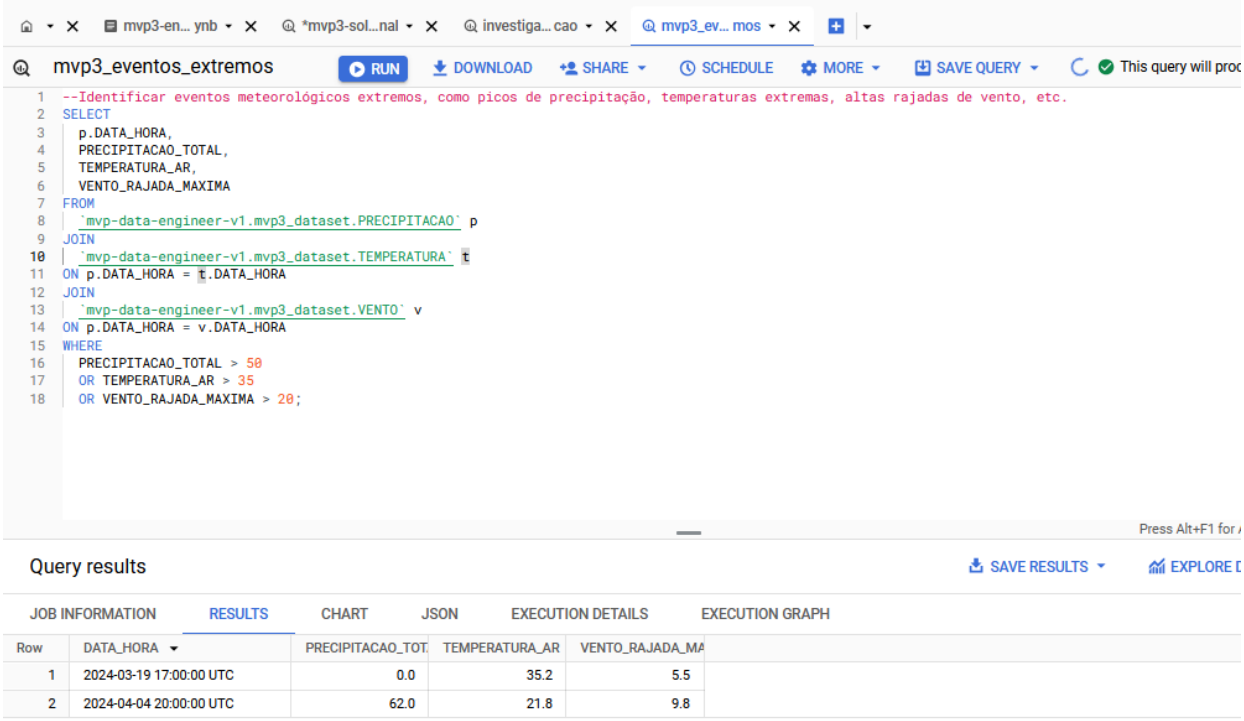


Exemplo da Tabela PRECIPITACAO:



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:



mvp3_eventos_extremos

RUN DOWNLOAD SHARE SCHEDULE MORE SAVE QUERY This query will proc

```
--Identificar eventos meteorológicos extremos, como picos de precipitação, temperaturas extremas, altas rajadas de vento, etc.
SELECT
  p.DATA_HORA,
  PRECIPITACAO_TOTAL,
  TEMPERATURA_AR,
  VENTO_RAJADA_MAXIMA
FROM
  `mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO` p
JOIN
  `mvp-data-engineer-v1.mvp3_dataset.TEMPERATURA` t
ON p.DATA_HORA = t.DATA_HORA
JOIN
  `mvp-data-engineer-v1.mvp3_dataset.VENTO` v
ON p.DATA_HORA = v.DATA_HORA
WHERE
  PRECIPITACAO_TOTAL > 50
  OR TEMPERATURA_AR > 35
  OR VENTO_RAJADA_MAXIMA > 20;
```

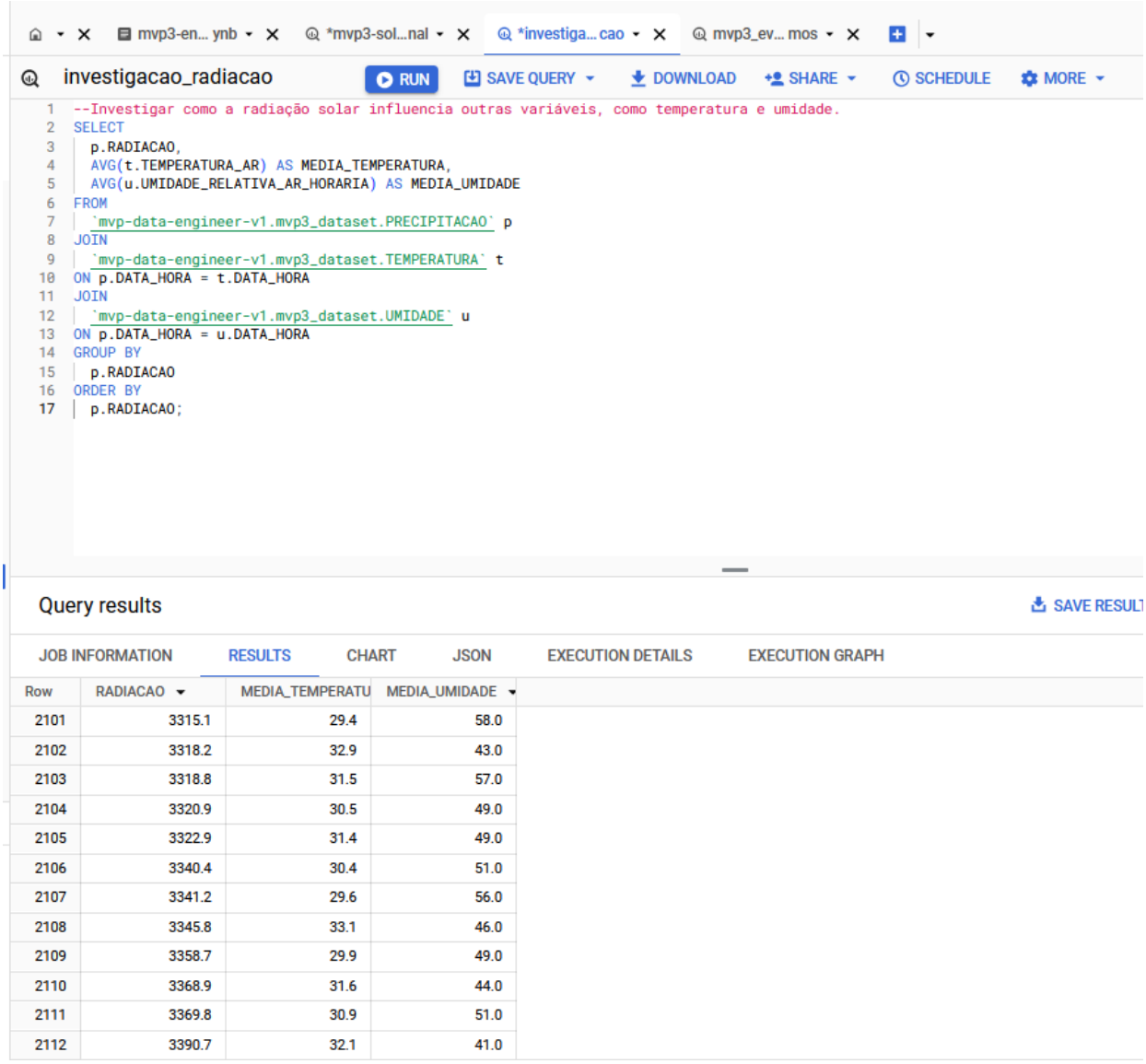
Query results

SAVE RESULTS EXPLORE D

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	DATA_HORA	PRECIPITACAO_TOT	TEMPERATURA_AR	VENTO_RAJADA_MA
1	2024-03-19 17:00:00 UTC	0.0	35.2	5.5
2	2024-04-04 20:00:00 UTC	62.0	21.8	9.8

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



investigacao_radiacao

RUN | SAVE QUERY | DOWNLOAD | SHARE | SCHEDULE | MORE

```
1  --Investigar como a radiação solar influencia outras variáveis, como temperatura e umidade.
2  SELECT
3    p.RADIACAO,
4    AVG(t.TEMPERATURA_AR) AS MEDIA_TEMPERATURA,
5    AVG(u.UMIDADE_RELATIVA_AR_HORARIA) AS MEDIA_UMIDADE
6  FROM
7    `mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO` p
8  JOIN
9    `mvp-data-engineer-v1.mvp3_dataset.TEMPERATURA` t
10 ON p.DATA_HORA = t.DATA_HORA
11 JOIN
12   `mvp-data-engineer-v1.mvp3_dataset.UMIDADE` u
13 ON p.DATA_HORA = u.DATA_HORA
14 GROUP BY
15   p.RADIACAO
16 ORDER BY
17   p.RADIACAO;
```

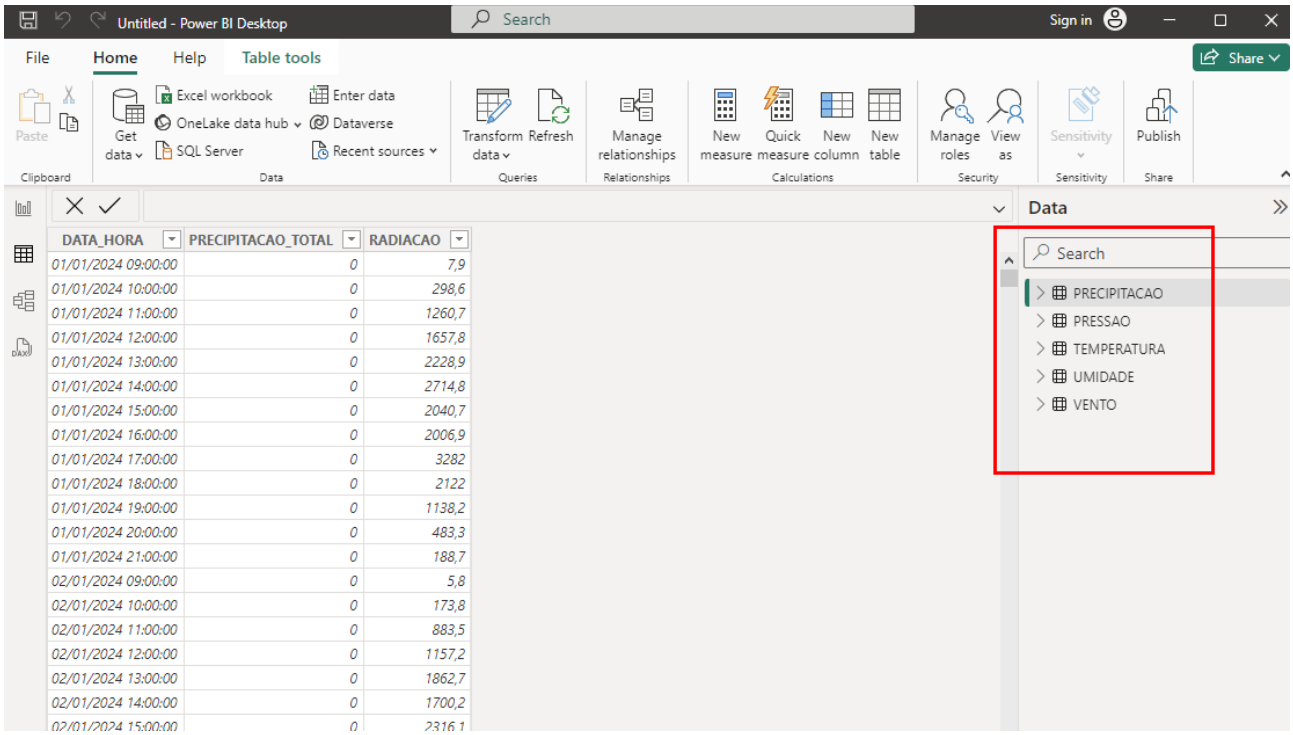
Query results

SAVE RESULT

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	RADIACAO	MEDIA_TEMPERATU	MEDIA_UMIDADE
2101	3315.1	29.4	58.0
2102	3318.2	32.9	43.0
2103	3318.8	31.5	57.0
2104	3320.9	30.5	49.0
2105	3322.9	31.4	49.0
2106	3340.4	30.4	51.0
2107	3341.2	29.6	56.0
2108	3345.8	33.1	46.0
2109	3358.7	29.9	49.0
2110	3368.9	31.6	44.0
2111	3369.8	30.9	51.0
2112	3390.7	32.1	41.0

Dados Apresentados via Power BI acessando o BigQuery



DATA_HORA	PRECIPITACAO_TOTAL	RADIACAO
01/01/2024 09:00:00	0	7,9
01/01/2024 10:00:00	0	298,6
01/01/2024 11:00:00	0	1260,7
01/01/2024 12:00:00	0	1657,8
01/01/2024 13:00:00	0	2228,9
01/01/2024 14:00:00	0	2714,8
01/01/2024 15:00:00	0	2040,7
01/01/2024 16:00:00	0	2006,9
01/01/2024 17:00:00	0	3282
01/01/2024 18:00:00	0	2122
01/01/2024 19:00:00	0	1138,2
01/01/2024 20:00:00	0	483,3
01/01/2024 21:00:00	0	188,7
02/01/2024 09:00:00	0	5,8
02/01/2024 10:00:00	0	173,8
02/01/2024 11:00:00	0	883,5
02/01/2024 12:00:00	0	1157,2
02/01/2024 13:00:00	0	1862,7
02/01/2024 14:00:00	0	1700,2
02/01/2024 15:00:00	0	2316,1