

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 temperatura 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):

MVP-data-engineer-v1

Search (/) for resources

Create a bucket

- Name your bucket**
Pick a globally unique, permanent name. [Naming guidelines](#)
mvp-data-engineer-v1
Tip: Don't include any sensitive information
LABELS (OPTIONAL)
CONTINUE
- Choose where to store your data**
Location: us (multiple regions in United States)
Location type: Multi-region
- Choose a storage class for your data**
Default storage class: Standard
- Choose how to control access to objects**
Public access prevention: On
Access control: Uniform
- Choose how to protect object data**
Soft delete policy: Enabled
Object versioning: Disabled
Bucket retention policy: Disabled
Object retention: Disabled
Encryption type: Google-managed

CREATE CANCEL

Detalhes do bucket:

MVP-data-engineer-v1

Search (/) for resources, docs, products, and more

Search

Bucket details

mvp-data-engineer-v1

Location

us (multiple regions in United States)

Storage class

Standard

Public access

Not public

Protection

Soft Delete

OBJECTS

CONFIGURATION

PERMISSIONS

PROTECTION

LIFECYCLE

OBSERVABILITY

INVENTORY REPORTS

OPERATIONS

Folder browser

mvp-data-engineer-v1

Buckets > mvp-data-engineer-v1

UPLOAD FILES

UPLOAD FOLDER

CREATE FOLDER

TRANSFER DATA

MANAGE HOLDS

EDIT RETENTION

DOWNLOAD

Filter by name prefix only

Filter

Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption
No rows to display									

Apresentação dos diretórios

Folder browser

mvp-data-engineer-v1

input/

processed/

Buckets > mvp-data-engineer-v1

UPLOAD FILES

UPLOAD FOLDER

Filter by name prefix only

Filter

<input type="checkbox"/>	Name
<input type="checkbox"/>	input/
<input type="checkbox"/>	processed/

Criado um Custom Role com as seguintes permissões:

MVP-data-engineer-v1

google cloud d

Custom Role - mvp3

EDIT ROLE

CREATE FROM ROI

ID

projects/mvp-data-engineer-v1/roles/CustomRole

Role launch stage

Alpha

Description

Created on: 2024-07-14

6 assigned permissions

bigquery.datasets.get
bigquery.jobs.create
bigquery.tables.create
bigquery.tables.get
bigquery.tables.updateData
storage.buckets.create

No usuário do projeto, adicionei o Role criado:

MVP-data-engineer-v1

google cloud data fusion

Search

IAM

PERMISSIONS

RECOMMENDATIONS HISTORY

Permissions for project "MVP-data-engineer-v1"

These permissions affect this project and all of its resources. [Learn more](#)

VIEW BY PRINCIPALS

VIEW BY ROLES

GRANT ACCESS

REMOVE ACCESS

Filter

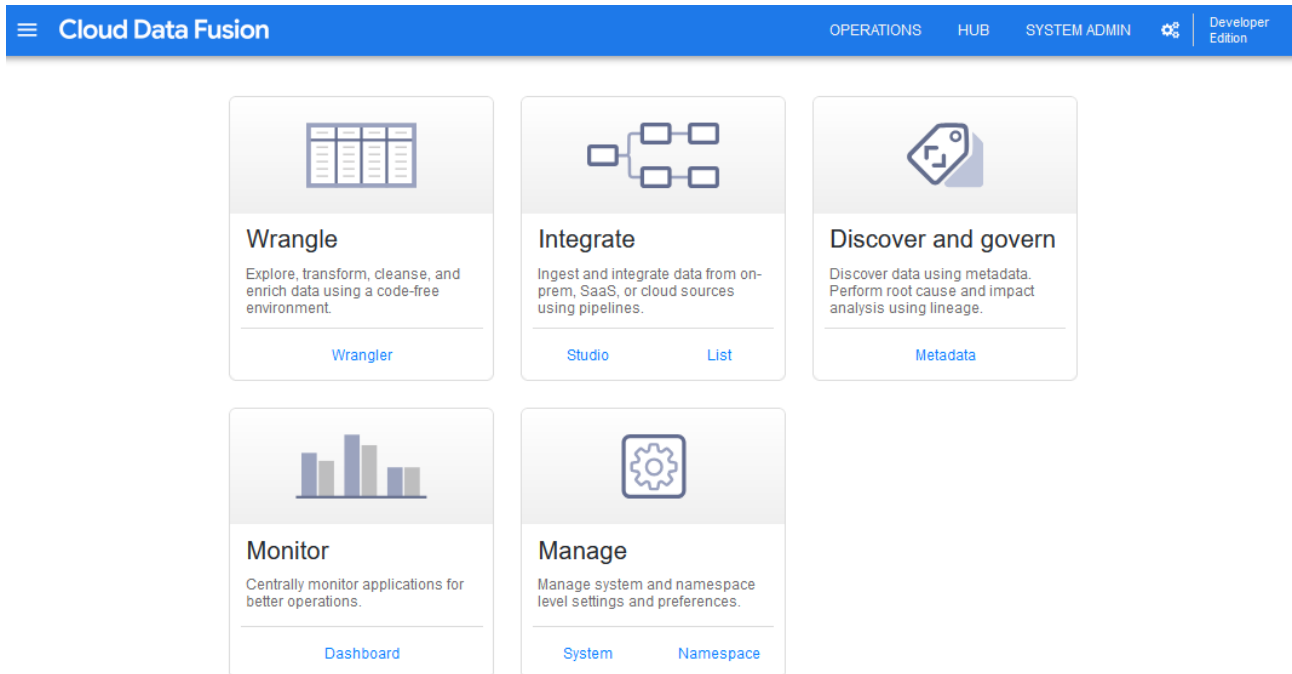
Enter property name or value

Type	Principal	Name	Role	Security insights
<input type="checkbox"/>	compute@developer.gserviceaccount.com	Compute Engine default service account	Cloud Data Fusion Runner	
<input type="checkbox"/>			Custom Role - mvp3	
			Dataproc Worker	
			Editor	

Pipeline Google Cloud Data Fusion

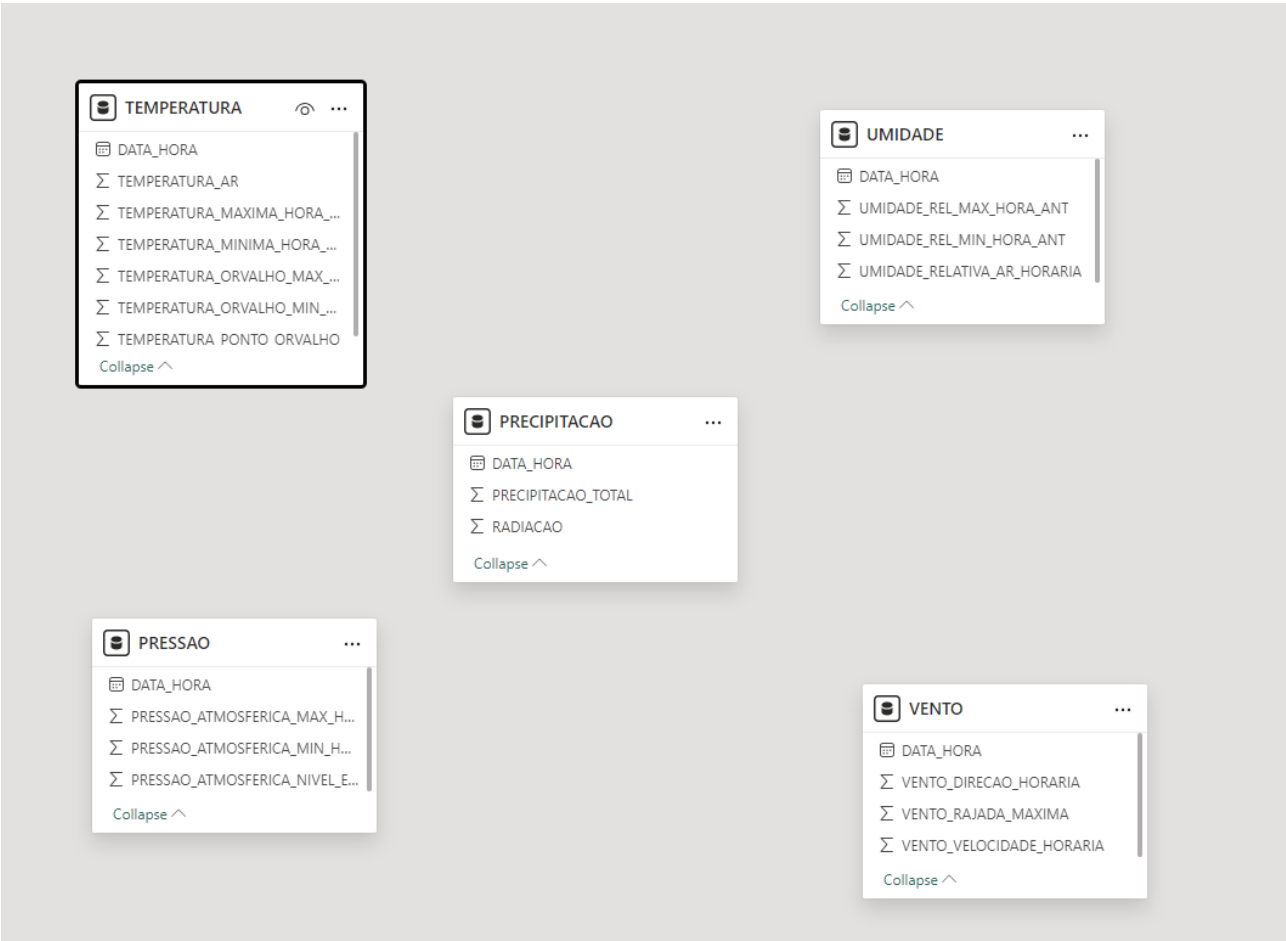
Foi criado inicialmente um pipeline com o Fusion, porém não consegui identificar a tempo um erro de transformação dos dados do CSV para a tabela do BigQuery.

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>

Google Cloud

MVP-data-engineer-v1

Search (/) for resources, docs, products, and more

Search

Explorer

+ ADD

Q Type to search

Viewing resources

SHOW STARRED ONLY

mvp-data-engineer-v1

Queries

Shared queries

mvp3-solucao-final

mvp3_ddl

Notebooks

Untitled notebooks

Shared notebooks

mvp3-engenharia-dados-col...

Data canvases

External connections

mvp3_dataset

PRECIPITACAO

PRESSAO

TEMPERATURA

UMIDADE

VENTO

SHOW MORE

SUMMARY

ACTIVITY

Jul 16, 2024

Saved - Fred 8:15 PM

Saved - Fred 12:34 PM

Saved - Fred 12:31 PM

Created - Fred 11:56 AM

mvp3_ddl

RUN

MORE

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

1 create or replace table 'mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO' (
2 DATA_HORA TIMESTAMP NOT NULL,
3 PRECIPITACAO_TOTAL FLOAT64 NOT NULL,
4 RADIACAO FLOAT64
5);
6 ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO'
7 ADD PRIMARY KEY (DATA_HORA) NOT ENFORCED;
8
9
10
11 create or replace table 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO' (
12 DATA_HORA TIMESTAMP,
13 PRESSAO_ATMOSFERICA_NIVEL_ESTACAO FLOAT64,
14 PRESSAO_ATMOSFERICA_MAX_HORA_ANT FLOAT64,
15 PRESSAO_ATMOSFERICA_MIN_HORA_ANT FLOAT64
16);
17 ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO'
18 ADD PRIMARY KEY (DATA_HORA) NOT ENFORCED;
19 ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO'
20 ADD CONSTRAINT DATA_HORA_FK FOREIGN KEY (DATA_HORA)
21 REFERENCES 'mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO' (DATA_HORA) NOT ENFORCED;
22
23
24 create or replace table 'mvp-data-engineer-v1.mvp3_dataset.TEMPERATURA' (
25 DATA_HORA TIMESTAMP,
26 TEMPERATURA_AR FLOAT64,
27 TEMPERATURA_PONTO_ORVALHO FLOAT64,
28 TEMPERATURA_MAXIMA_HORA_ANT FLOAT64,
29 TEMPERATURA_MINIMA_HORA_ANT FLOAT64,
30 TEMPERATURA_ORVALHO_MAX_HORA_ANT FLOAT64,
31 TEMPERATURA_ORVALHO_MIN_HORA_ANT FLOAT64
32);
33 ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.TEMPERATURA'
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

All results

Elapsed time 10 sec

Statements processed 14

Job status SUCCESS

Status	End time	SQL	Stages completed	Bytes processed	Action
✓	1:49 PM [1:1]	create or replace table 'mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO' (DATA_HORA TIMESTAMP NOT NULL, PRECIPITACAO_TOTAL FLOAT64 NOT NULL, RADIACAO FLOAT64); ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO' ADD PRIMARY KEY (DATA_HORA) NOT ENFORCED;	0	0 B	VIEW RESULTS
✓	1:49 PM [6:1]	ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO'	0	0 B	VIEW RESULTS
✓	1:49 PM [11:1]	create or replace table 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO' (DATA_HORA TIMESTAMP, PRESSAO_ATMOSFERICA_NIVEL_ESTACAO FLOAT64, PRESSAO_ATMOSFERICA_MAX_HORA_ANT FLOAT64, PRESSAO_ATMOSFERICA_MIN_HORA_ANT FLOAT64); ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO' ADD PRIMARY KEY (DATA_HORA) NOT ENFORCED;	0	0 B	VIEW RESULTS
✓	1:49 PM [17:1]	ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO'	0	0 B	VIEW RESULTS
✓	1:49 PM [19:1]	ALTER TABLE 'mvp-data-engineer-v1.mvp3_dataset.PRESSAO'	0	0 B	VIEW RESULTS

Exemplo da Tabela PRECIPITACAO:

*mvp3_ddl

PRECIPITACAO

+

+

PRECIPITACAO

QUERY

SHARE

COPY

SNAPSHOT

DELETE

EXPORT

SCHEMA

DETAILS

PREVIEW

LINEAGE

DATA PROFILE

DATA QUALITY

Filter

Enter property name or value

	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags
<input type="checkbox"/>	DATA_HORA	TIMESTAMP	REQUIRED	PK	-	-	-
<input type="checkbox"/>	PRECIPITACAO_TOTAL	FLOAT	REQUIRED	-	-	-	-
<input type="checkbox"/>	RADIACAO	FLOAT	NULLABLE	-	-	-	-

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-en... ynb × *mvp3-sol...nal × investiga...cao × mvp3_ev...mos ×

mvp3_eventos_extremos RUN DOWNLOAD SHARE SCHEDULE MORE SAVE QUERY This query will proc

```
1 --Identificar eventos meteorológicos extremos, como picos de precipitação, temperaturas extremas, altas rajadas de vento, etc.
2 SELECT
3   p.DATA_HORA,
4   PRECIPITACAO_TOTAL,
5   TEMPERATURA_AR,
6   VENTO_RAJADA_MAXIMA
7 FROM
8   `mvp-data-engineer-v1.mvp3_dataset.PRECIPITACAO` p
9 JOIN
10  `mvp-data-engineer-v1.mvp3_dataset.TEMPERATURA` t
11 ON p.DATA_HORA = t.DATA_HORA
12 JOIN
13  `mvp-data-engineer-v1.mvp3_dataset.VENTO` v
14 ON p.DATA_HORA = v.DATA_HORA
15 WHERE
16   PRECIPITACAO_TOTAL > 50
17   OR TEMPERATURA_AR > 35
18   OR VENTO_RAJADA_MAXIMA > 20;
```

Press Alt+F1 for

Query results SAVE RESULTS EXPLORE I

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 temperatura 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 RESULTS

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

Untitled - Power BI Desktop

Search

Sign in

Share

FileHomeHelpTable tools

Paste

Get data

Excel workbook

OneLake data hub

SQL Server

Enter data

Dataverse

Recent sources

Transform data

Refresh data

Manage relationships

New measure

Quick measure

New measure column

New table

Manage roles

View as

Sensitivity

Sensitivity

Publish

Clipboard

Data

Queries

Relationships

Calculations

Security

Sensitivity

Share

×

✓

Search

PRECIPITACAO

PRESSAO

TEMPERATURA

UMIDADE

VENTO

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