

Curso: Tecnologia em Data Science, Big Data, BI & Data Engineering

Projeto: CHALLENGE (012024)

Turmas do 1o ano: 1TSCPV

Sprint 3

**DataStorm**

Ana Beatriz Azevedo, RM557420

Heloiza Oliveira, RM558881

Isabelle Nahas, RM557405

Matheus Madrid, RM555799

Sara Sitta, RM555113

**DISCIPLINA: BIG DATA ARCHITECTURE & DATA INTEGRATION**

**Professores: Rogerio Mariana Galazi, Milton Goya**

São Paulo, Setembro de 2024

## Documentação do Pipeline de Dados

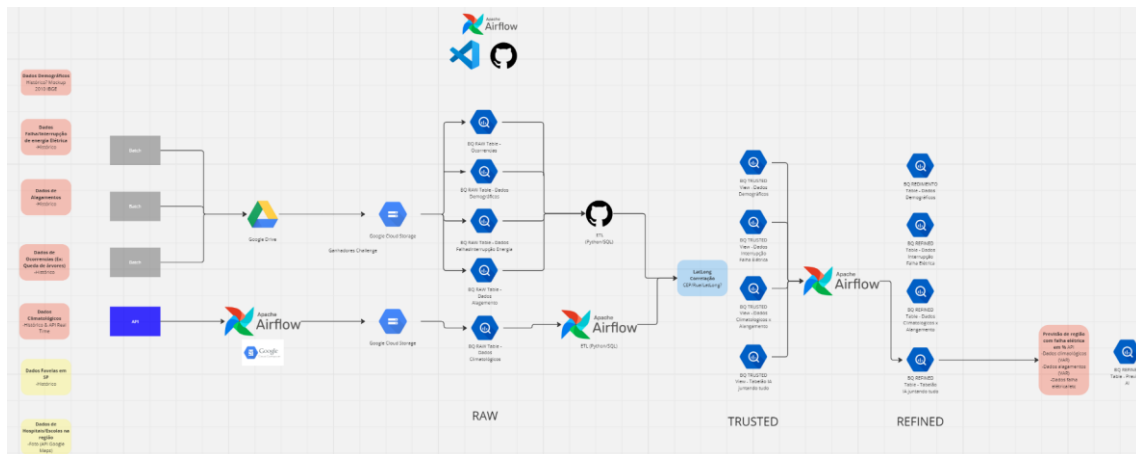
### Visão Geral

Este documento descreve o pipeline de dados implementado para ingestão, transformação, armazenamento e consumo de dados climáticos coletados de uma API, usando Apache Airflow no Google Cloud Composer, Google Cloud Storage como Data Lake, e BigQuery para armazenamento e transformação dos dados. O pipeline é composto por três DAGs principais, cada uma responsável por uma etapa específica do fluxo de dados.

### 1. Ferramenta de ETL/Ingestão

Para atender aos requisitos de coleta, ingestão, armazenamento, transformação e orquestração, foram utilizados Apache Airflow no Google Cloud Composer para orquestração, Google Cloud Storage para armazenamento temporário, e BigQuery para transformação e armazenamento dos dados estruturados.

Arquitetura atualizada:



Link: [https://miro.com/app/board/uXjVKmDD4gU/?share\\_link\\_id=695605839813](https://miro.com/app/board/uXjVKmDD4gU=?share_link_id=695605839813)

### Google Cloud Composer

console.cloud.google.com/composer/environments/detail/us-central1/weather-prod-us-central1/dags?project=elated-drive-432523-s4

Free trial status: R\$1,814.32 credit and 65 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

DISMISSACTIVATE

Google Cloud

Data Storm

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

Search

Navigation menu

Environment details

OPEN AIRFLOW UI

OPEN DAGS FOLDER

SAVE SNAPSHOT

LOAD SNAPSHOT

REFRESH

DELETE

LEARN

weather-prod-us-central1

This environment is running

MONITORINGLOGSDAGS

ENVIRONMENT CONFIGURATION

AIRFLOW CONFIGURATION OVERRIDES

ENVIRONMENT VARIABLES

LABELS

PYPI PACKAGES

FilterFilter DAGs

1 hour

6 hours

12 hours

1 day

2 days

4 days

7 days

14 days

30 days

DAG id	State	Description	Schedule interval	Last completed run	Active runs	Successful runs (1h)	Failed runs (1h)
<a href="#">airflow_monitoring</a>	Active	liveness monitoring dag	* / 10 * * * *	2 minutes ago	0	6	0
<a href="#">etl_raw_to_trusted_weather</a>	Active	Pipeline to process WEATHER and SUBPREFECTURES tables.	@hourly	2 minutes ago	0	1	0
<a href="#">etl_trusted_to_refined_weather</a>	Active	Pipeline to move WEATHER data from TRUSTED to REFINED.	@hourly	2 minutes ago	0	3	0
<a href="#">weather_data_pipeline</a>	Paused	DAG para coletar dados de clima e enviar para GCS e BigQuery	@hourly	4 hours ago	0	0	0

## Apache Airflow:

Airflow DAGs Cluster Activity Datasets Browse Admin Docs Composer

02:03 UTC

## weather-prod-us-central1

All Active Paused

Running Failed

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
airflow_monitoring	airflow		@hourly	2024-09-09, 01:50:00	2024-09-09, 02:00:00			...
etl_raw_to_trusted_weather	airflow		@hourly	2024-09-09, 01:00:00	2024-09-09, 02:00:00			...
etl_trusted_to_refined_weather	airflow		@hourly	2024-09-09, 01:35:10	2024-09-09, 02:00:00			...
weather_data_pipeline	airflow		@hourly	2024-09-08, 21:00:00	2024-09-09, 01:00:00			...

Showing 1 of 4 DAGs

## CloudStorage – Landing Zone

The screenshot shows the Google Cloud Storage console. The top navigation bar includes the Google Cloud logo, a search bar, and a 'GO TO PATH' button. The left sidebar shows the 'Cloud Storage' menu with options for 'Buckets', 'Monitoring', and 'Settings'. The main content area displays the details for the bucket 'us-central1-datastorm-prod-us-a4c8a156-landingzone'. The bucket is located in the 'us' region, has a 'Standard' storage class, and is 'Not public'. The 'Folder browser' view shows a folder named 'LANDING\_ZONE\_WEATHER/' containing a file named 'historical\_weather\_data.json'.

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

OBJECTS	CONFIGURATION	PERMISSIONS	PROTECTION	LIFECYCLE	OBSERVABILITY	INVENTORY REPORTS	OPERATIONS
<p>Folder browser</p> <ul style="list-style-type: none"> <li>us-central1-datastorm-prod-us-a4c8a156-landingzone <ul style="list-style-type: none"> <li>LANDING_ZONE_WEATHER/ <ul style="list-style-type: none"> <li>historical_weather_data.json</li> </ul> </li> </ul> </li> </ul>							

Name	Size	Type	Created	Storage class	Last modified	Public access
historical_weather_data.json	303.6 KB	application/octet-stream	Sep 8, 2024, 11:04:29 PM	Standard	Sep 8, 2024, 11:04:29 PM	Not public

## Bigquery GCP

Free trial status: \$21,810.37 credit and 65 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

Google Cloud Data Store

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

Explorer

elated-drive-432523-s4

Queries

Shared queries

ExploratoryDataAnalysis-We...

Notebooks

Data canvases

Data preparations

External connections

RAW

SUBPREFECTURES

WEATHER

WEATHER\_HISTORY

intermpcao2024

REFINED

WEATHER

TRUSTED

SUBPREFECTURES

WEATHER

WEATHER\_API

WEATHER

QUERY

SHARE

COPY

SNAPSHOT

DELETE

EXPORT

SCHEMA

DETAILS

PREVIEW

TABLE EXPLORER

PREVIEW

INSIGHTS

LINEAGE

DATA PROFILE

DATA QUALITY

Filter Enter property name or value

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
<input type="checkbox"/> SubprefectureName	STRING	NULLABLE	-	-	-	-	Standardized name of the subprefecture from the Open Weather API
<input type="checkbox"/> Latitude	FLOAT	NULLABLE	-	-	-	-	Subprefecture Latitude
<input type="checkbox"/> Longitude	FLOAT	NULLABLE	-	-	-	-	Subprefecture Longitude
<input type="checkbox"/> ObservationTime	TIMESTAMP	NULLABLE	-	-	-	-	Datetime of the weather conditions observation
<input type="checkbox"/> Temperature	INTEGER	NULLABLE	-	-	-	-	Recorded temperature in Kelvin
<input type="checkbox"/> FeelsLike	INTEGER	NULLABLE	-	-	-	-	Apparent temperature (feels like) in Kelvin
<input type="checkbox"/> Pressure	INTEGER	NULLABLE	-	-	-	-	Atmospheric pressure in hectopascals (hPa)
<input type="checkbox"/> Humidity	INTEGER	NULLABLE	-	-	-	-	Relative humidity percentage
<input type="checkbox"/> MinTemperature	INTEGER	NULLABLE	-	-	-	-	Minimum temperature recorded in Kelvin
<input type="checkbox"/> MaxTemperature	INTEGER	NULLABLE	-	-	-	-	Maximum temperature recorded in Kelvin
<input type="checkbox"/> WindSpeed	INTEGER	NULLABLE	-	-	-	-	Wind speed in meters per second
<input type="checkbox"/> WindDirection	INTEGER	NULLABLE	-	-	-	-	Wind direction in degrees
<input type="checkbox"/> WindGust	INTEGER	NULLABLE	-	-	-	-	Wind gust speed in meters per second
<input type="checkbox"/> CloudCoverage	INTEGER	NULLABLE	-	-	-	-	Percentage of cloud cover
<input type="checkbox"/> WeatherID	INTEGER	NULLABLE	-	-	-	-	Identifier for the weather type
<input type="checkbox"/> WeatherID	STRING	NULLABLE	-	-	-	-	Main category of the weather (e.g., Clouds, Rain)

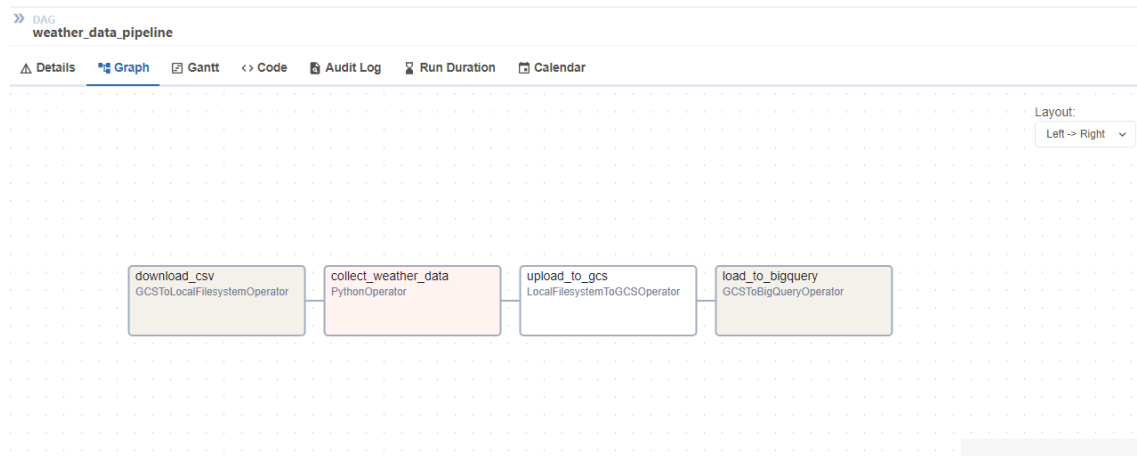
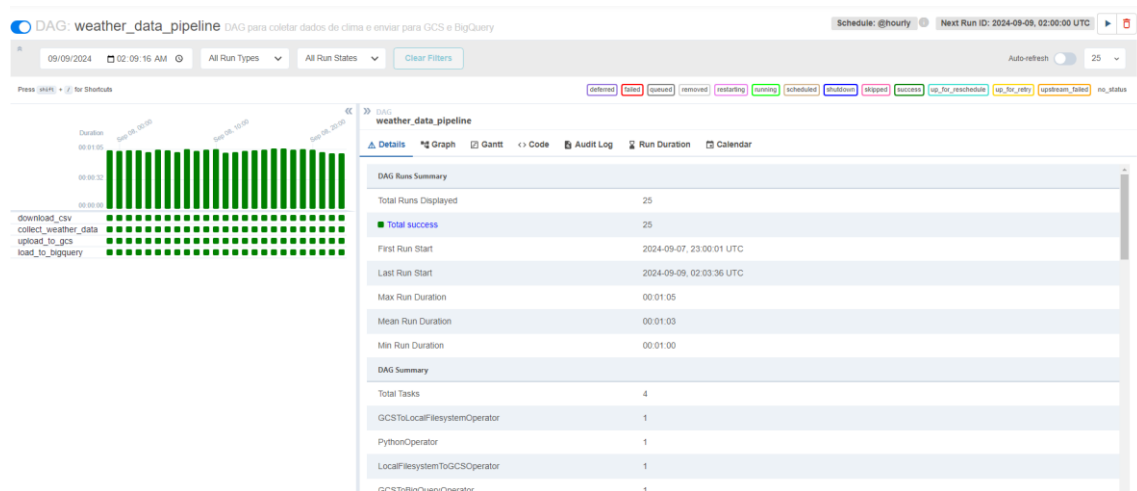
## 1.1 Coleta e Ingestão de Dados para o Data Lake e BigQuery (DAG: weather\_data\_pipeline\_dag)

- **Nome da DAG:** weather\_data\_pipeline\_dag
- **Ferramenta Utilizada:** Apache Airflow com Google Cloud Composer para orquestração; Google Cloud Storage para armazenamento dos dados e BigQuery para estruturação do Lake
- **Processo:**
  - **Coleta de Dados:** A DAG coleta dados climáticos da API do OpenWeatherMap em intervalos regulares (a cada 1 hora).
  - **Armazenamento Temporário:** Os dados coletados são salvos no Google Cloud Storage, que atua como o Data Lake.
  - **Carga para BigQuery:** Após salvar no Cloud Storage, os dados são carregados na camada RAW do BigQuery, utilizando operadores específicos do Airflow para transferência de arquivos do Cloud Storage para o BigQuery.

```
weather_data_pipeline_dag.py > ...
1 from airflow import DAG
2 from airflow.operators.python import PythonOperator
3 from airflow.providers.google.cloud.transfers.local_to_gcs import LocalFileSystemToGCSOperator
4 from airflow.providers.google.cloud.transfers.gcs_to_bigquery import GCSToBigQueryOperator
5 from airflow.providers.google.cloud.transfers.gcs_to_local import GCSToLocalFileSystemOperator
6 from airflow.utils.dates import days_ago
7 import pandas as pd
8 import requests
9 import time
10 import json
11 from datetime import datetime
12
13 # Função para coletar dados das subprefeituras e salvar em formato adequado para o BigQuery
14 def collect_weather_data(**kwargs):
15     # Baixar o CSV das subprefeituras diretamente do Google Cloud Storage
16     subprefeituras = pd.read_csv('/tmp/subprefeituras-sp.csv', sep=';', header=0)
17     subprefeituras.columns = subprefeituras.columns.str.strip().str.lower()
18
19     API_KEY = '---' # Substitua pela sua chave da API
20     BASE_URL = 'https://history.openweathermap.org/data/2.5/history/city'
21
22     results = []
23     for index, row in subprefeituras.iterrows():
24         lat = round(row['latitude'], 2)
25         lon = round(row['longitude'], 2)
26         subprefeitura_nome = row['subprefeitura']
27         url = f'{BASE_URL}?lat={lat:.2f}&lon={lon:.2f}&appid={API_KEY}'
28
29         response = requests.get(url)
30         if response.status_code == 200:
31             data = response.json()
32             if 'list' in data:
33                 for entry in data['list']:
34                     result = {
35                         'subprefeitura': subprefeitura_nome,
36                         'latitude': lat,
37                         'longitude': lon,
38                         'timestamp': datetime.utcfromtimestamp(entry['dt']).isoformat(), # Converter para timestamp ISO
39                         'temperature': entry['main'].get('temp'),
40                         'feels_like': entry['main'].get('feels_like'),
41                         'pressure': entry['main'].get('pressure'),
42                         'humidity': entry['main'].get('humidity'),
43                         'temp_min': entry['main'].get('temp_min'),
44                         'temp_max': entry['main'].get('temp_max'),
45                         'wind_speed': entry['wind'].get('speed'),
46                         'wind_deg': entry['wind'].get('deg')
```

Arquivo anexado - weather\_data\_pipeline\_dag.py

weather\_data\_pipeline\_dag. DAG no Airflow



DAG weather\_data\_pipeline

Details Graph Gantt Code Audit Log Run Duration Calendar

Parsed at: 2024-09-09, 02:07:40 UTC

```
96     task_id='collect_weather_data',
97     python_callable=collect_weather_data,
98     provide_context=True,
99 )
100
101 # Task 3: Enviar JSON para o Cloud Storage na pasta especificada
102 upload_to_gcs = LocalFilesystemToGCSOperator(
103     task_id='upload_to_gcs',
104     src='/tmp/historical_weather_data.json',
105     dst='LANDING_ZONE_WEATHER/historical_weather_data.json',
106     bucket='us-central1-datastorm-prod-us-a4c8a156-landingzone',
107 )
108
109 # Task 4: Carregar dados do GCS para BigQuery
110 load_to_bigquery = GCSToBigQueryOperator(
111     task_id='load_to_bigquery',
112     bucket='us-central1-datastorm-prod-us-a4c8a156-landingzone',
113     source_objects=['LANDING_ZONE_WEATHER/historical_weather_data.json'],
114     destination_project_dataset_table='elated-drive-432523-s4:RAW.WEATHER', # Substitua pelo seu projeto e dataset
115     source_format='NEWLINE_DELIMITED_JSON',
116     write_disposition='WRITE_APPEND',
117 )
118
119 # Definir a sequência das tasks
120 download_csv >> collect_data_task >> upload_to_gcs >> load_to_bigquery
121
```

LANDING ZONE no Google Cloud Storage

Bucket details

us-central1-datastorm-prod-us-a4c8a156-landingzone

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

us-central1-datastorm-prod-us-a4c8a156-landingzone

LANDING\_ZONE\_WEATHER/

historical\_weather\_data.json

Name	Size	Type	Created	Storage class	Last modified	Public access
historical_weather_data.json	303.6 KB	application/octet-stream	Sep 8, 2024, 11:04:29 PM	Standard	Sep 8, 2024, 11:04:29 PM	Not public

## Camada RAW.WEATHER

Explorer

Search BigQuery resources

Viewing resources. SHOW STARRED ONLY

elated-drive-432523-s4

Queries

Shared queries

ExploratoryDataAnalysis-We...

Notebooks

Data canvases

Data preparations

External connections

RAW

SUBPREFECTURES

WEATHER

WEATHER\_HISTORY

WEATHER

QUERY

SHARE

COPY

SNAPSHOT

DELETE

EXPOR

SCHEMA

DETAILS

PREVIEW

TABLE EXPLORER

PREVIEW

INSIGHTS

LINEAGE

DA

Filter

Enter property name or value

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags
weather_main	STRING	NULLABLE	-	-	-	-
weather_id	INTEGER	NULLABLE	-	-	-	-
clouds_all	INTEGER	NULLABLE	-	-	-	-
wind_deg	INTEGER	NULLABLE	-	-	-	-
pressure	INTEGER	NULLABLE	-	-	-	-
temp_min	FLOAT	NULLABLE	-	-	-	-
wind_speed	FLOAT	NULLABLE	-	-	-	-
humidity	INTEGER	NULLABLE	-	-	-	-
feels_like	FLOAT	NULLABLE	-	-	-	-

## 1.2 Estruturação e Preparação dos Dados para a Camada Estruturada (DAG: etl\_raw\_to\_trusted\_weather)

- **Nome da DAG:** etl\_raw\_to\_trusted\_weather
- **Ferramenta Utilizada:** Apache Airflow com Google Cloud Composer para orquestração; Google Cloud Storage para armazenamento dos dados e BigQuery para transformação dos dados.
- **Processo:**
  - Esta DAG transforma e estrutura os dados movendo-os da camada RAW para a camada TRUSTED no BigQuery.
  - **Transformações Incluídas:**
    - Ajuste de precisão de latitude e longitude para garantir a consistência entre os conjuntos de dados.
    - Conversão de CEP de string para inteiro, removendo caracteres indesejados.
    - Criação de views para padronizar os nomes das colunas para o padrão de Data Lake (letra maiúscula, sem underscores).

Task 4 da DAG weather\_data\_pipeline\_dag

```
>> DAG
weather_data_pipeline

Details Graph Gantt <> Code Audit Log Run Duration Calendar

Parsed at: 2024-09-09, 02:07:40 UTC

96     task_id='collect_weather_data',
97     python_callable=collect_weather_data,
98     provide_context=True,
99 )
100
101 # Task 3: Enviar JSON para o Cloud Storage na pasta especificada
102 upload_to_gcs = LocalFilesystemToGCSOperator(
103     task_id='upload_to_gcs',
104     src='/tmp/historical_weather_data.json',
105     dst='LANDING_ZONE_WEATHER/historical_weather_data.json',
106     bucket='us-central1-datastorm-prod-us-a4c8a156-landingzone',
107 )
108
109 # Task 4: Carregar dados do GCS para BigQuery
110 load_to_bigquery = GCSToBigQueryOperator(
111     task_id='load_to_bigquery',
112     bucket='us-central1-datastorm-prod-us-a4c8a156-landingzone',
113     source_objects=['LANDING_ZONE_WEATHER/historical_weather_data.json'],
114     destination_project_dataset_table='elated-drive-432523-s4:RAW.WEATHER', # Substitua pelo seu projeto e dataset
115     source_format='NEWLINE_DELIMITED_JSON',
116     write_disposition='WRITE_APPEND',
117 )
118
119 # Definir a sequência das tasks
120 download_csv >> collect_data_task >> upload_to_gcs >> load_to_bigquery
121
```

## Camada RAW.WEATHER

Explorer + ADD <

Search BigQuery resources

Viewing resources. SHOW STARRED ONLY

- elated-drive-432523-s4
  - Queries
    - Shared queries
    - ExploratoryDataAnalysis-We...
  - Notebooks
  - Data canvases
  - Data preparations
  - External connections
  - RAW
    - SUBPREFECTURES
    - WEATHER**
    - WEATHER\_HISTORY

WEATHER QUERY SHARE COPY SNAPSHOT DELETE EXPORT

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DA

Filter Enter property name or value

<input type="checkbox"/>	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags
<input type="checkbox"/>	weather_main	STRING	NULLABLE	-	-	-	-
<input type="checkbox"/>	weather_id	INTEGER	NULLABLE	-	-	-	-
<input type="checkbox"/>	clouds_all	INTEGER	NULLABLE	-	-	-	-
<input type="checkbox"/>	wind_deg	INTEGER	NULLABLE	-	-	-	-
<input type="checkbox"/>	pressure	INTEGER	NULLABLE	-	-	-	-
<input type="checkbox"/>	temp_min	FLOAT	NULLABLE	-	-	-	-
<input type="checkbox"/>	wind_speed	FLOAT	NULLABLE	-	-	-	-
<input type="checkbox"/>	humidity	INTEGER	NULLABLE	-	-	-	-
<input type="checkbox"/>	feels_like	FLOAT	NULLABLE	-	-	-	-

## etl\_raw\_to\_trusted\_weather DAG para refinamento

Airflow DAGs Cluster Activity Datasets Browse Admin Docs Composer

02:13 UTC

DAG: etl\_raw\_to\_trusted\_weather Pipeline to process WEATHER and SUBPREFECTURES tables from RAW to TRUSTED layer Schedule: @hourly Next Run ID: 2024-09-09, 02:00:00 UTC

09/09/2024 02:13:15 AM All Run Types All Run States Clear Filters Auto-refresh 25

Press <Shift> + J for Shortcuts

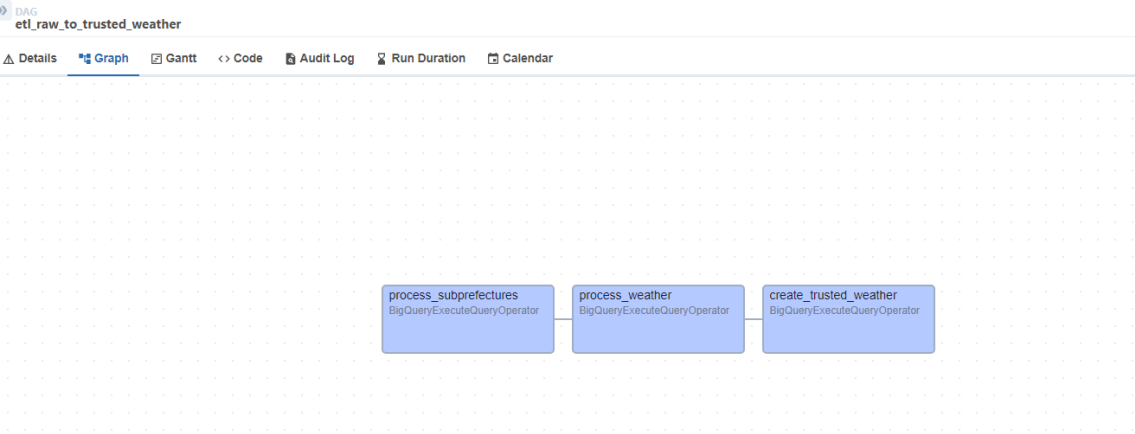
etl\_raw\_to\_trusted\_weather

Details Graph Gantt <> Code Audit Log Run Duration Calendar

DAG Run Summary

Total Runs Displayed	8
Total success	5
Total failed	3
First Run Start	2024-09-09, 00:10:48 UTC
Last Run Start	2024-09-09, 02:00:00 UTC
Max Run Duration	00:05:24
Mean Run Duration	00:02:46
Min Run Duration	00:00:12
DAG Summary	
Total Tasks	3

Tasks da DAG



DAG  
etl\_raw\_to\_trusted\_weather

Details Graph Gantt Code Audit Log Run Duration Calendar

Parsed at: 2024-09-09, 02:11:32 UTC

```
19 # Step 1: Create VIEW for SUBPREFECTURES in TRUSTED directly from RAW
20 process_subprefectures = BigQueryExecuteQueryOperator(
21     task_id='process_subprefectures',
22     sql="""
23         CREATE OR REPLACE VIEW `elated-drive-432523-s4.TRUSTED.SUBPREFECTURES` AS
24         SELECT
25             REPLACE(UPPER(SUBSTRING(Subprefeitura, 1, 1)) || LOWER(SUBSTRING(Subprefeitura, 2)), ' ', '') AS SubprefectureName,
26             ROUND(Latitude, 2) AS Latitude, -- Rounding to 2 decimal places
27             ROUND(Longitude, 2) AS Longitude, -- Rounding to 2 decimal places
28             CAST(REPLACE(CEP, '-', '') AS INT64) AS PostalCode
29         FROM `elated-drive-432523-s4.RAW.SUBPREFECTURES`
30     """,
31     use_legacy_sql=False
32 )
33
34 # Step 2: Create VIEW for WEATHER in TRUSTED directly from RAW
35 process_weather = BigQueryExecuteQueryOperator(
36     task_id='process_weather',
37     sql="""
38         CREATE OR REPLACE VIEW `elated-drive-432523-s4.TRUSTED.WEATHER_API` AS
39         SELECT
40             REPLACE(UPPER(SUBSTRING(subprefeitura, 1, 1)) || LOWER(SUBSTRING(subprefeitura, 2)), ' ', '') AS SubprefectureName,
41             ROUND(Latitude, 2) AS Latitude, -- Rounding to 2 decimal places
42             ROUND(Longitude, 2) AS Longitude, -- Rounding to 2 decimal places
43             TIMESTAMP(timestamp) AS ObservationTime,
44             temperature AS Temperature,
```

Dag anexada no zip etl\_raw\_to\_trusted\_weather.py

CAMADA TRUSTED



Explorer + ADD

Search BigQuery resources

Viewing resources. SHOW STARRED ONLY

elated-drive-432523-s4

- Queries
- Notebooks
- Data canvases
- Data preparations
- External connections
- RAW
- REFINED
- TRUSTED
  - SUBPREFECTURES
  - WEATHER
  - WEATHER\_API
  - openweathermap\_historic

WEATHER

SCHEMA DETAILS TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALITY

Filter Enter property name or value

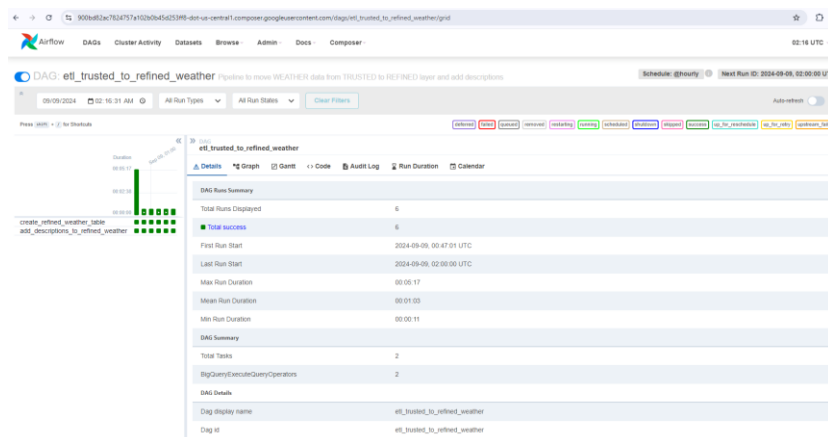
Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
SubprefectureName	STRING	NULLABLE	-	-	-	-	-
Latitude	FLOAT	NULLABLE	-	-	-	-	-
Longitude	FLOAT	NULLABLE	-	-	-	-	-
ObservationTime	TIMESTAMP	NULLABLE	-	-	-	-	-
Temperature	FLOAT	NULLABLE	-	-	-	-	-
FeelsLike	FLOAT	NULLABLE	-	-	-	-	-
Pressure	INTEGER	NULLABLE	-	-	-	-	-
Humidity	INTEGER	NULLABLE	-	-	-	-	-
MinTemperature	FLOAT	NULLABLE	-	-	-	-	-
MaxTemperature	FLOAT	NULLABLE	-	-	-	-	-
WindSpeed	FLOAT	NULLABLE	-	-	-	-	-
WindDirection	INTEGER	NULLABLE	-	-	-	-	-
WindGust	FLOAT	NULLABLE	-	-	-	-	-
CloudCoverage	INTEGER	NULLABLE	-	-	-	-	-
WeatherID	INTEGER	NULLABLE	-	-	-	-	-
WeatherMain	STRING	NULLABLE	-	-	-	-	-
WeatherDescription	STRING	NULLABLE	-	-	-	-	-
WeatherIcon	STRING	NULLABLE	-	-	-	-	-

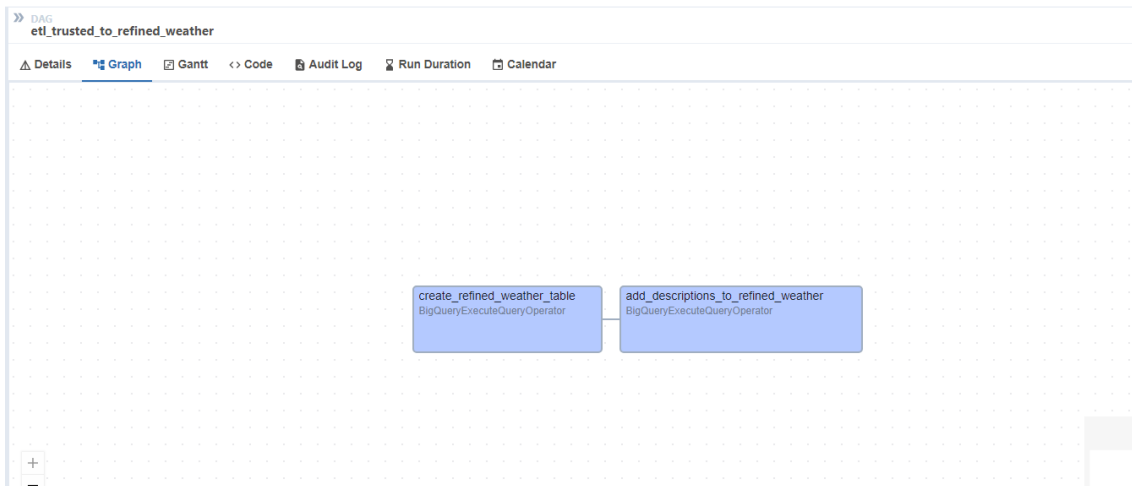
EDIT SCHEMA

### 1.3 Refinamento e Enriquecimento dos Dados (DAG: etl\_trusted\_to\_refined\_weather)

- **Nome da DAG:** etl\_trusted\_to\_refined\_weather
- **Ferramenta Utilizada:** Apache Airflow com Google Cloud Composer para orquestração; BigQuery para refinamento e enriquecimento dos dados.
- **Processo:**
  - A última DAG move os dados da camada TRUSTED para a camada REFINED no BigQuery.
  - **Refinamento Inclui:**
    - Criação de tabelas finais na camada REFINED.
    - Adição de descrições das colunas para melhorar a documentação e a compreensão dos dados.
    - Junção de dados climáticos com informações de subprefeituras para fornecer um conjunto de dados enriquecido.

DAG etl\_trusted\_to\_refined\_weather no Airflow





» DAG  
etl\_trusted\_to\_refined\_weather

Details Graph Gantt **Code** Audit Log Run Duration Calendar

Parsed at: 2024-09-09, 02:15:25 UTC

```
29 # Step 2: Consolidated ALTER TABLE to Add Descriptions to REFINED.WEATHER Table
30 add_descriptions_to_refined_weather = BigQueryExecuteQueryOperator(
31     task_id='add_descriptions_to_refined_weather',
32     sql="""
33         ALTER TABLE `elated-drive-432523-s4.REFINED.WEATHER`
34         SET OPTIONS (
35             description = 'Refined table with enriched weather and subprefecture data'
36         );
37
38         ALTER TABLE `elated-drive-432523-s4.REFINED.WEATHER`
39         ALTER COLUMN SubprefectureName SET OPTIONS (description = 'Standardized name of the subprefecture from the Open Weather API')
40         ALTER COLUMN Latitude SET OPTIONS (description = 'Subprefecture Latitude'),
41         ALTER COLUMN Longitude SET OPTIONS (description = 'Subprefecture Longitude'),
42         ALTER COLUMN ObservationTime SET OPTIONS (description = 'Datatime of the weather conditions obsevation'),
43         ALTER COLUMN Temperature SET OPTIONS (description = 'Recorded temperature in Kelvin'),
44         ALTER COLUMN Feelslike SET OPTIONS (description = 'Apparent temperature (feels like) in Kelvin'),
45         ALTER COLUMN Pressure SET OPTIONS (description = 'Atmospheric pressure in hectopascals (hPa)'),
46         ALTER COLUMN Humidity SET OPTIONS (description = 'Relative humidity percentage'),
47         ALTER COLUMN MinTemperature SET OPTIONS (description = 'Minimum temperature recorded in Kelvin'),
48         ALTER COLUMN MaxTemperature SET OPTIONS (description = 'Maximum temperature recorded in Kelvin'),
49         ALTER COLUMN WindSpeed SET OPTIONS (description = 'Wind speed in meters per second'),
50         ALTER COLUMN WindDirection SET OPTIONS (description = 'Wind direction in degrees'),
51         ALTER COLUMN WindGust SET OPTIONS (description = 'Wind gust speed in meters per second'),
52         ALTER COLUMN CloudCoverage SET OPTIONS (description = 'Percentage of cloud cover'),
53         ALTER COLUMN WeatherID SET OPTIONS (description = 'Identifier for the weather type'),
```

Arquivo anexado no zip etl\_trusted\_to\_refined\_weather.py

CAMADA REFINED com a tabela WEATHER final com as Descrições

Explorer

Search BigQuery resources

Viewing resources

SHOW STARRED ONLY

elated-drive-432523-s4

Queries

Notebooks

Data canvases

Data preparations

External connections

RAW

REFINED

WEATHER

TRUSTED

openweathermap\_historic

WEATHER

QUERY

SHARE

COPY

SNAPSHOT

DELETE

EXPORT

SCHEMA

DETAILS

PREVIEW

TABLE EXPLORER

PREVIEW

INSIGHTS

LINEAGE

DATA PROFILE

DATA QUALITY

Filter

Enter property name or value

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
SubprefectureName	STRING	NULLABLE	-	-	-	-	Standardized name of the subprefecture from the Open Weather API
Latitude	FLOAT	NULLABLE	-	-	-	-	Subprefecture Latitude
Longitude	FLOAT	NULLABLE	-	-	-	-	Subprefecture Longitude
ObservationTime	TIMESTAMP	NULLABLE	-	-	-	-	Datetime of the weather conditions observation
Temperature	FLOAT	NULLABLE	-	-	-	-	Recorded temperature in Kelvin
FeelsLike	FLOAT	NULLABLE	-	-	-	-	Apparent temperature (feels like) in Kelvin
Pressure	INTEGER	NULLABLE	-	-	-	-	Atmospheric pressure in hectopascals (hPa)
Humidity	INTEGER	NULLABLE	-	-	-	-	Relative humidity percentage
MinTemperature	FLOAT	NULLABLE	-	-	-	-	Minimum temperature recorded in Kelvin
MaxTemperature	FLOAT	NULLABLE	-	-	-	-	Maximum temperature recorded in Kelvin
WindSpeed	FLOAT	NULLABLE	-	-	-	-	Wind speed in meters per second
WindDirection	INTEGER	NULLABLE	-	-	-	-	Wind direction in degrees
WindGust	FLOAT	NULLABLE	-	-	-	-	Wind gust speed in meters per second
CloudCoverage	INTEGER	NULLABLE	-	-	-	-	Percentage of cloud cover
WeatherID	INTEGER	NULLABLE	-	-	-	-	Identifier for the weather type
WeatherMain	STRING	NULLABLE	-	-	-	-	Main category of the weather (e.g., Clouds, Rain)
WeatherDescription	STRING	NULLABLE	-	-	-	-	Detailed description of the weather conditions
WeatherIcon	STRING	NULLABLE	-	-	-	-	Code representation the weather icon

2. Análise Exploratória no Ambiente SQL (BigQuery)

Utilizando BigQuery para análise exploratória dos dados refinados, foram realizadas as seguintes operações:

2.2a) Contagem de Registros

```
SELECT COUNT(*) AS TotalRecords
FROM `elated-drive-432523-s4.REFINED.WEATHER`;
```

ExploratoryDataAnalysis-Weather

RUNSAVE QUERYDOWNLOAD

1

-- 2.2a) Count of Records

2

-- This part counts the total number of records in the REFINED.WEATHER table.

3

SELECT

4

'TotalRecords' AS Description,

5

COUNT(\*) AS Value

6

FROM `elated-drive-432523-s4.REFINED.WEATHER`

7

;

Query results

JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAILS	E
Row	Description	Value			
1	TotalRecords	23808			

2.2b) Visualização de Amostra dos Dados (Primeiras Dez Linhas)

SELECT \*  
FROM `elated-drive-432523-s4.REFINED.WEATHER`  
LIMIT 10;

ExploratoryDataAnalysis-Weather

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

MORE

9

-- 2.2b) View a Sample of Data (First Ten Rows)

-- This part displays the first ten rows to give a sample of the data.

SELECT \*

FROM -elated-drive-432523-s4.REFINED\_WEATHER

LIMIT 10

15

16

Query results

Press Alt+F1 for Acc

SAVE RESULTS

EXPLORE DATA

JOB INFORMATION

RESULTS

CHART

JSON

EXECUTION DETAILS

EXECUTION GRAPH

Row	SubprefectureName	Latitude	Longitude	ObservationTime	Temperature	FeelsLike	Pressure	Humidity	MinTemperature	MaxTemperature
1	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
2	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
3	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
4	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
5	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
6	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
7	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
8	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
9	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53
10	Sé	-23.55	-46.63	2024-09-07 11:00:00 UTC	291.53	291.75	1020	89	290.62	291.53

2.2c) Descrição dos Campos da Tabela

SELECT  
    table\_name  
    ,column\_name  
    ,field\_path  
    ,data\_type  
    ,description  
FROM `elated-drive-432523-s4.REFINED.INFORMATION\_SCHEMA.COLUMN\_FIELD\_PATHS`  
WHERE table\_name = 'WEATHER'  
;

ExploratoryDataAnalysis-Weather

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

MORE

Query results

SAVE RESULTSEXPLORE DATA

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS		EXECUTION GRAPH				
Row	table_name	column_name	field_path	data_type	description						
1	WEATHER	SubprefectureName	SubprefectureName	STRING	Standardized name of the subprefecture from the Open Weather API						
2	WEATHER	Latitude	Latitude	FLOAT64	Subprefecture Latitude						
3	WEATHER	Longitude	Longitude	FLOAT64	Subprefecture Longitude						
4	WEATHER	ObservationTime	ObservationTime	TIMESTAMP	Datetime of the weather condit...						
5	WEATHER	Temperature	Temperature	FLOAT64	Recorded temperature in Kelvin						
6	WEATHER	FeelsLike	FeelsLike	FLOAT64	Apparent temperature (feels lik...						
7	WEATHER	Pressure	Pressure	INT64	Atmospheric pressure in hecto...						
8	WEATHER	Humidity	Humidity	INT64	Relative humidity percentage						
9	WEATHER	MinTemperature	MinTemperature	FLOAT64	Minimum temperature recorde...						
10	WEATHER	MaxTemperature	MaxTemperature	FLOAT64	Maximum temperature recorde...						
11	WEATHER	WindSpeed	WindSpeed	FLOAT64	Wind speed in meters per seco...						
12	WFATHFR	WindDirection	WindDirection	INT64	Wind direction in degrees						

2.2d) Contagem de Valores Distintos

Exemplo

SELECT COUNT(DISTINCT SubprefectureName) AS DistinctSubprefectureNames  
FROM `elated-drive-432523-s4.REFINED.WEATHER`;

🔍 ExploratoryDataAnalysis-Weather ▶ RUN 📄 SA

```
28
29 -- 2.2d) Count of Distinct Values
30 -- This part counts the distinct values for specific fields.
31 SELECT
32   `DistinctSubprefectureNames` AS Description,
33   COUNT(DISTINCT SubprefectureName) AS Value
34 FROM `elated-drive-432523-s4.REFINED.WEATHER`
35
36 UNION ALL
37
38 SELECT
39   `DistinctTemperatures` AS Description,
40   COUNT(DISTINCT Temperature) AS Value
41 FROM `elated-drive-432523-s4.REFINED.WEATHER`
42
43 UNION ALL
44
45 SELECT
46   `DistinctWeatherMain` AS Description,
47   COUNT(DISTINCT WeatherMain) AS Value
48 FROM `elated-drive-432523-s4.REFINED.WEATHER`
49 ;
--
```

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECU
Row	Description ▾	Value ▾			
1	DistinctWeatherMain	5			
2	DistinctTemperatures	389			
3	DistinctSubprefectureNames	32			

2.2e) Estatísticas Descritivas (Média, Mínimo, Máximo)

Exemplo de cálculo para a coluna Temperature:

```
SELECT
  `TemperatureStats` AS Description,
  TO_JSON_STRING(t) AS Value
FROM (
  SELECT
    AVG(Temperature) AS AvgTemperature,
    MIN(Temperature) AS MinTemperature,
    MAX(Temperature) AS MaxTemperature
```

```

FROM `elated-drive-432523-s4.REFINED.WEATHER`
) t

UNION ALL

SELECT
  'WindSpeedStats' AS Description,
  TO_JSON_STRING(t) AS Value
FROM (
  SELECT
    AVG(WindSpeed) AS AvgWindSpeed,
    MIN(WindSpeed) AS MinWindSpeed,
    MAX(WindSpeed) AS MaxWindSpeed
  FROM `elated-drive-432523-s4.REFINED.WEATHER`
) t
;

```

ExploratoryDataAnalysis-Weather RUN SAVE QUERY

```

51 -- 2.2e) Descriptive Statistics (Mean, Min, Max)
52 -- This part calculates the mean, minimum, and maximum for numerical field
53 SELECT
54   'TemperatureStats' AS Description,
55   TO_JSON_STRING(t) AS Value
56 FROM (
57   SELECT
58     AVG(Temperature) AS AvgTemperature,
59     MIN(Temperature) AS MinTemperature,
60     MAX(Temperature) AS MaxTemperature
61   FROM `elated-drive-432523-s4.REFINED.WEATHER`
62 ) t
63
64 UNION ALL
65
66 SELECT
67   'WindSpeedStats' AS Description,
68   TO_JSON_STRING(t) AS Value
69 FROM (
70   SELECT
71     AVG(WindSpeed) AS AvgWindSpeed,
72     MIN(WindSpeed) AS MinWindSpeed,
73     MAX(WindSpeed) AS MaxWindSpeed

```

#### Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS
Row	Description	Value			
1	WindSpeedStats	{"AvgWindSpeed":1.9099172547042977,"MinWindSpeed":0,"MaxWindSpeed":6.69}			
2	TemperatureStats	{"AvgTemperature":297.14541666666759,"MinTemperature":288.81,"MaxTemperature":307.64}			

## 2.2f) Distribuição de Frequência

```

-- 2.2f) Frequency Distribution (Count of Elements for Each Field)
-- This part shows the frequency distribution of categorical fields.
SELECT
  'WeatherMainFrequency' AS Description,
  TO_JSON_STRING(t) AS Value
FROM (
  SELECT
    WeatherMain,
    COUNT(*) AS Frequency
  FROM `elated-drive-432523-s4.REFINED.WEATHER`

```

```

GROUP BY WeatherMain
ORDER BY Frequency DESC
) t

UNION ALL

SELECT
  'SubprefectureNameFrequency' AS Description,
  TO_JSON_STRING(t) AS Value
FROM (
  SELECT
    SubprefectureName,
    COUNT(*) AS Frequency
  FROM `elated-drive-432523-s4.REFINED.WEATHER`
  GROUP BY SubprefectureName
  ORDER BY Frequency DESC
) t;

```

ExploratoryDataAnalysis-Weather RUN SAVE QUERY DOWNLOAD SHARE

```

78 --2.2f) Frequency Distribution (Count of Elements for Each Field)
79 --This part shows the frequency distribution of categorical fields.
80 SELECT
81   'WeatherMainFrequency' AS Description,
82   TO_JSON_STRING(t) AS Value
83 FROM (
84   SELECT
85     WeatherMain,
86     COUNT(*) AS Frequency
87   FROM `elated-drive-432523-s4.REFINED.WEATHER`
88   GROUP BY WeatherMain
89   ORDER BY Frequency DESC
90 ) t
91
92 UNION ALL
93
94 SELECT
95   'SubprefectureNameFrequency' AS Description,
96   TO_JSON_STRING(t) AS Value
97

```

## Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	Description	Value				
1	WeatherMainFrequency	{"WeatherMain": "Clouds", "Frequency": 8420}				
2	WeatherMainFrequency	{"WeatherMain": "Clear", "Frequency": 14360}				
3	WeatherMainFrequency	{"WeatherMain": "Mist", "Frequency": 510}				
4	WeatherMainFrequency	{"WeatherMain": "Haze", "Frequency": 358}				
5	WeatherMainFrequency	{"WeatherMain": "Fog", "Frequency": 160}				
6	SubprefectureNameFrequency	{"SubprefectureName": "Sé", "Frequency": 744}				
7	SubprefectureNameFrequency	{"SubprefectureName": "Lapa", "Frequency": 744}				
8	SubprefectureNameFrequency	{"SubprefectureName": "Mooca", "Frequency": 744}				
9	SubprefectureNameFrequency	{"SubprefectureName": "Penha", "Frequency": 744}				
10	SubprefectureNameFrequency	{"SubprefectureName": "Perus", "Frequency": 744}				

## 3. Documentação e Evidências

Todas as etapas foram capturadas e documentadas para comprovar o funcionamento do pipeline.

## Conclusão