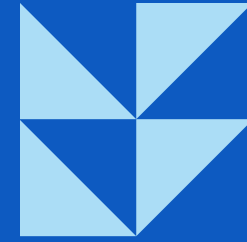


# Global Terrorism Analysis

Data streaming

Presentado por  
**Martín García**  
**David Melo**  
**Juan Andrés Ruiz**

# Sobre nosotros

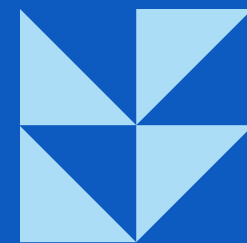


Global Terrorism Database

Utilizamos un conjunto de datos que incluye información sobre **atentados terroristas en todo el mundo** desde 1970 hasta 2017 (excepto 1993).



# Sobre nosotros



ACLED API

Es la fuente de datos que usamos para realizar un merge entre ambas fuentes de información.

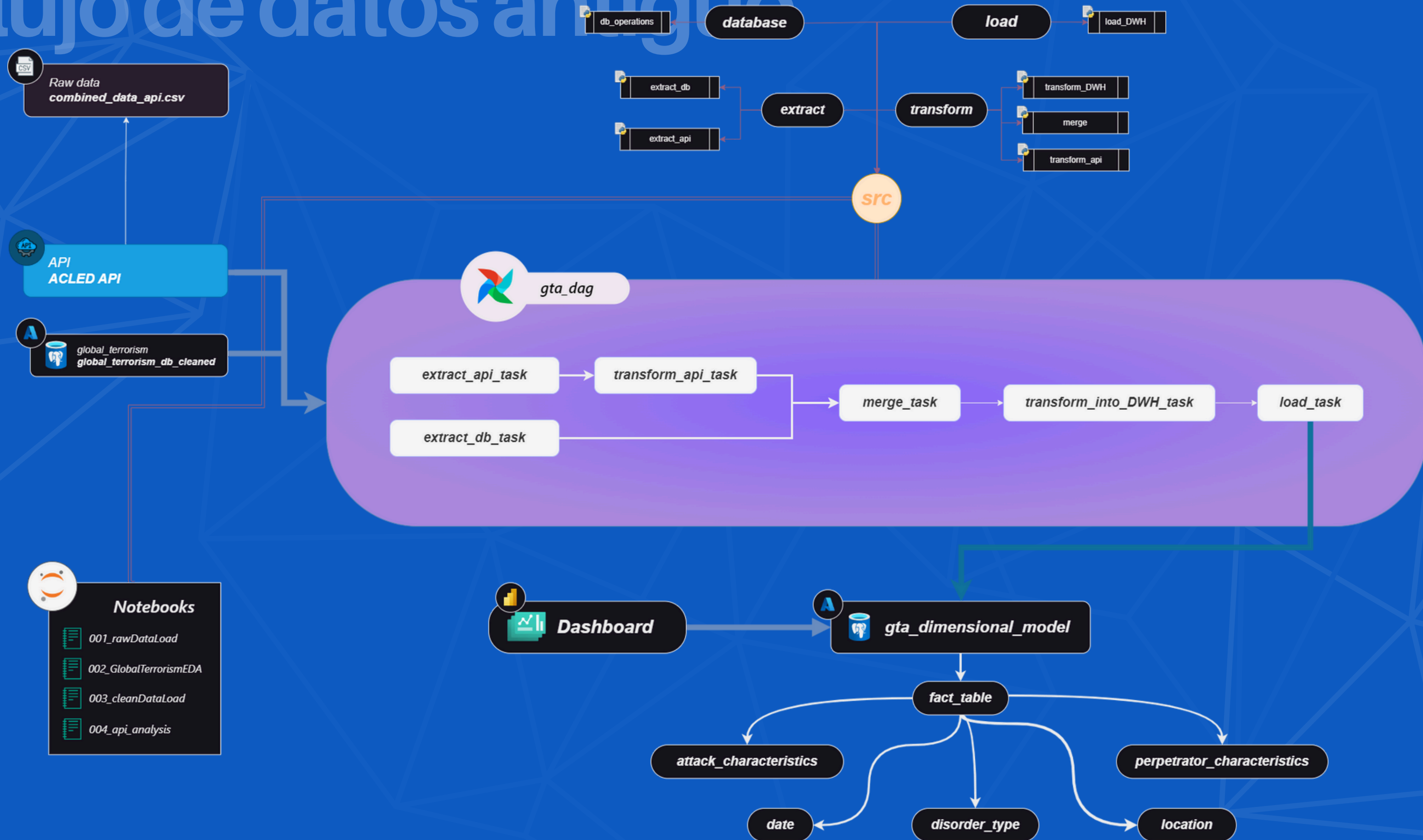




## ACLED API

```
def extracting_api_data():  
    try:  
        logging.info("Starting API data extraction.")  
  
        url = 'https://api.acleddata.com/acled/read.json'  
        params = {  
            'key': f'{api_key}',  
            'email': f'{api_email}',  
            'fields': 'event_date|country|disorder_type|actor1',  
            'year': '1989|2017',  
            'year_where': 'BETWEEN',  
            'limit': 5000,  
            'page': 1  
        }  
    }
```

# Flujo de datos antiguo



# Merge



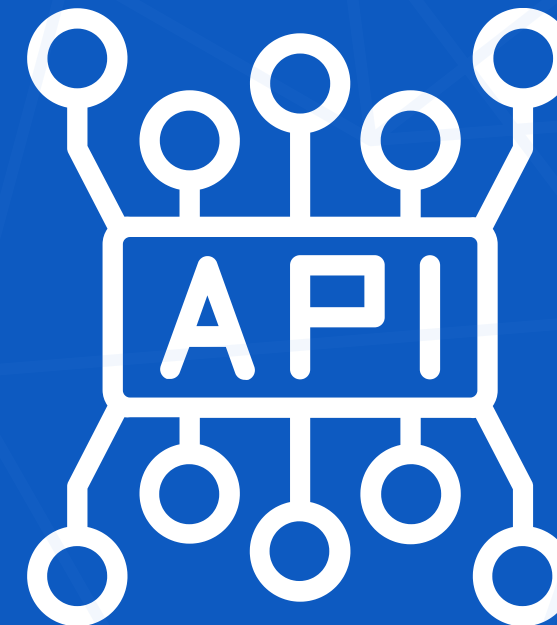
Devuelve todas las filas de la tabla izquierda, incluso si no hay coincidencias en la tabla derecha. Si no hay coincidencia, los valores de la tabla derecha serán NULL.

## *Left join*

Data set original



API

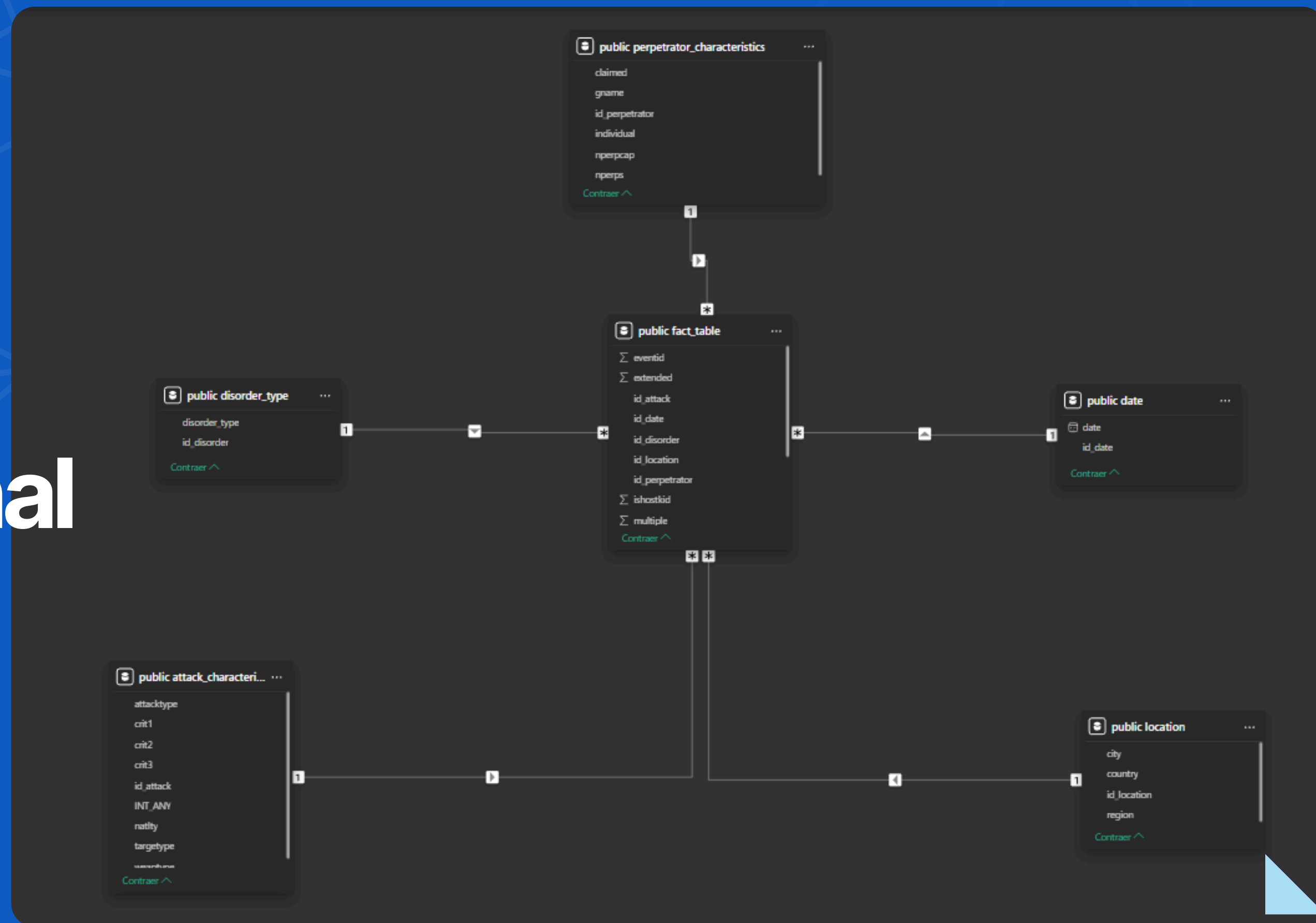






# Modelo dimensional

*arquitectura estrella*

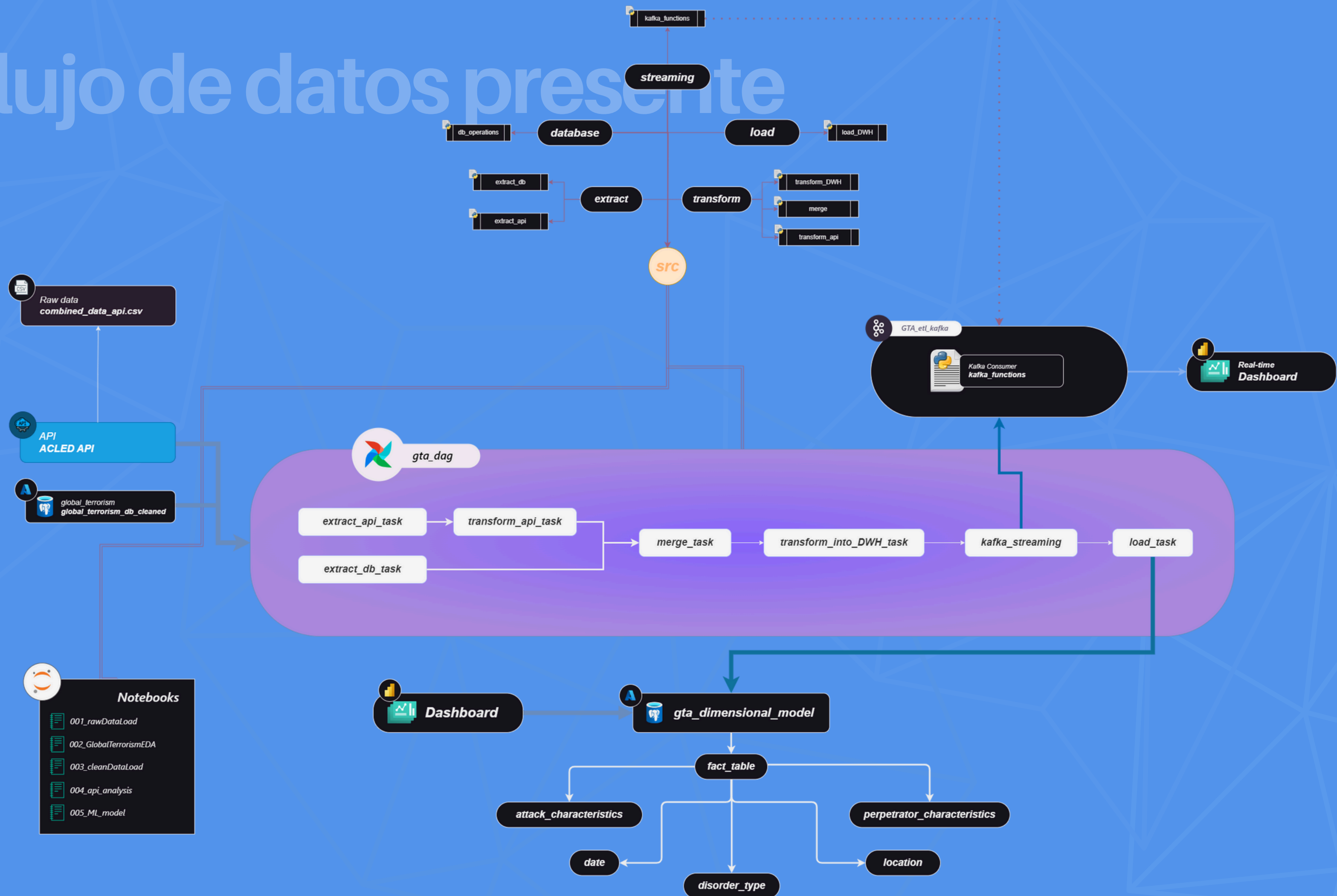


# Dashboard

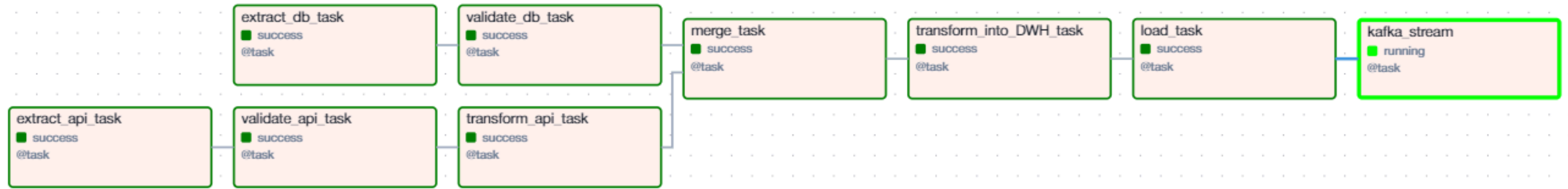




# Flujo de datos presente



# Flujo de datos presente





# Great Expectations





```
def db_validation(df_gtd):
    gtd_expectations = [
        gxe.ExpectTableColumnsToMatchOrderedList(
            column_list=[
                'eventid', 'iyear', 'imonth', 'iday', 'extended', 'country_txt',
                'country', 'region_txt', 'region', 'city', 'latitude', 'longitude',
                'vicinity', 'crit1', 'crit2', 'crit3', 'doubterr', 'multiple',
                'success', 'suicide', 'attacktype1_txt', 'attacktype1', 'targettype1_txt',
                'targettype1', 'natity1_txt', 'natity1', 'gsame', 'guncertain1',
                'individual', 'nperps', 'nperpcap', 'claimed', 'weaptype1_txt',
                'weaptype1', 'nkill', 'mound', 'property', 'ishostkid', 'INT_MW',
                'data', 'date_country_actor'
            ]
        ),
        gxe.ExpectColumnValuesToBeBetween(column='iyear', min_value=1978, max_value=2017),
        gxe.ExpectColumnValuesToBeBetween(column='imonth', min_value=1, max_value=12),
        gxe.ExpectColumnValuesToBeBetween(column='iday', min_value=1, max_value=31),
        gxe.ExpectColumnValuesToBeBetween(column='country', min_value=4, max_value=1804),
        gxe.ExpectColumnValuesToBeBetween(column='region', min_value=1, max_value=12),
        gxe.ExpectColumnValuesToBeBetween(column='latitude', min_value=-42.358458, max_value=99.8),
        gxe.ExpectColumnValuesToBeBetween(column='longitude', min_value=-157.858333, max_value=173.36),
        gxe.ExpectColumnValuesToBeBetween(column='attacktype1', min_value=1, max_value=8),
        gxe.ExpectColumnValuesToBeBetween(column='targettype1', min_value=1, max_value=22),
        gxe.ExpectColumnValuesToBeBetween(column='natity1', min_value=4, max_value=1804),
        gxe.ExpectColumnValuesToBeBetween(column='nperps', min_value=0, max_value=25000),
        gxe.ExpectColumnValuesToBeBetween(column='nperpcap', min_value=0, max_value=900),
        gxe.ExpectColumnValuesToBeBetween(column='weaptype1', min_value=1, max_value=13),
        gxe.ExpectColumnValuesToBeBetween(column='nkill', min_value=0, max_value=1384),
        gxe.ExpectColumnValuesToBeBetween(column='mound', min_value=0, max_value=8191),
        gxe.ExpectColumnValuesToBeInSet(column='extended', value_set=[0, 1]),
        gxe.ExpectColumnValuesToBeInSet(column='vicinity', value_set=[0, 1, 999]),
        gxe.ExpectColumnValuesToBeInSet(column='crit1', value_set=[1, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='crit2', value_set=[1, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='crit3', value_set=[1, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='doubterr', value_set=[0, 1]),
        gxe.ExpectColumnValuesToBeInSet(column='multiple', value_set=[0, 0, 1, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='success', value_set=[1, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='suicide', value_set=[0, 1]),
        gxe.ExpectColumnValuesToBeInSet(column='guncertain1', value_set=[0, 0, 1, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='individual', value_set=[0, 1]),
        gxe.ExpectColumnValuesToBeInSet(column='claimed', value_set=[0, 0, 1, 0, 999, 8]),
        gxe.ExpectColumnValuesToBeInSet(column='property', value_set=[0, 1, 999]),
        gxe.ExpectColumnValuesToBeInSet(column='ishostkid', value_set=[0, 0, 1, 0, 999, 0]),
        gxe.ExpectColumnValuesToBeInSet(column='INT_MW', value_set=[999, 0, 1]),
        gxe.ExpectColumnValuesToNotBeNull(column='eventid'),
        gxe.ExpectColumnValuesToNotBeNull(column='iyear'),
        gxe.ExpectColumnValuesToNotBeNull(column='imonth'),
        gxe.ExpectColumnValuesToNotBeNull(column='iday'),
        gxe.ExpectColumnValuesToNotBeNull(column='extended'),
        gxe.ExpectColumnValuesToNotBeNull(column='country_txt'),
        gxe.ExpectColumnValuesToNotBeNull(column='country'),
        gxe.ExpectColumnValuesToNotBeNull(column='region_txt'),
        gxe.ExpectColumnValuesToNotBeNull(column='region'),
        gxe.ExpectColumnValuesToNotBeNull(column='city'),
        gxe.ExpectColumnValuesToNotBeNull(column='latitude'),
        gxe.ExpectColumnValuesToNotBeNull(column='longitude'),
        gxe.ExpectColumnValuesToNotBeNull(column='vicinity'),
        gxe.ExpectColumnValuesToNotBeNull(column='crit1'),
        gxe.ExpectColumnValuesToNotBeNull(column='crit2'),
        gxe.ExpectColumnValuesToNotBeNull(column='crit3'),
        gxe.ExpectColumnValuesToNotBeNull(column='doubterr'),
        gxe.ExpectColumnValuesToNotBeNull(column='multiple'),
        gxe.ExpectColumnValuesToNotBeNull(column='success'),
        gxe.ExpectColumnValuesToNotBeNull(column='suicide'),
        gxe.ExpectColumnValuesToNotBeNull(column='attacktype1_txt'),
        gxe.ExpectColumnValuesToNotBeNull(column='attacktype1'),
        gxe.ExpectColumnValuesToNotBeNull(column='targettype1_txt'),
        gxe.ExpectColumnValuesToNotBeNull(column='targettype1'),
        gxe.ExpectColumnValuesToNotBeNull(column='natity1_txt'),
        gxe.ExpectColumnValuesToNotBeNull(column='natity1'),
        gxe.ExpectColumnValuesToNotBeNull(column='gsame'),
        gxe.ExpectColumnValuesToNotBeNull(column='guncertain1'),
        gxe.ExpectColumnValuesToNotBeNull(column='individual'),
        gxe.ExpectColumnValuesToNotBeNull(column='nperps'),
        gxe.ExpectColumnValuesToNotBeNull(column='nperpcap'),
        gxe.ExpectColumnValuesToNotBeNull(column='claimed'),
        gxe.ExpectColumnValuesToNotBeNull(column='weaptype1_txt'),
        gxe.ExpectColumnValuesToNotBeNull(column='weaptype1'),
        gxe.ExpectColumnValuesToNotBeNull(column='nkill'),
        gxe.ExpectColumnValuesToNotBeNull(column='mound'),
        gxe.ExpectColumnValuesToNotBeNull(column='property'),
        gxe.ExpectColumnValuesToNotBeNull(column='ishostkid'),
        gxe.ExpectColumnValuesToNotBeNull(column='INT_MW'),
        gxe.ExpectColumnValuesToNotBeNull(column='data'),
        gxe.ExpectColumnValuesToNotBeNull(column='date_country_actor'),
        gxe.ExpectColumnValuesToBeOfType(column='eventid', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='iyear', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='imonth', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='iday', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='extended', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='country_txt', type_='str'),
        gxe.ExpectColumnValuesToBeOfType(column='country', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='region_txt', type_='str'),
        gxe.ExpectColumnValuesToBeOfType(column='region', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='city', type_='str'),
        gxe.ExpectColumnValuesToBeOfType(column='latitude', type_='float64'),
        gxe.ExpectColumnValuesToBeOfType(column='longitude', type_='float64'),
        gxe.ExpectColumnValuesToBeOfType(column='vicinity', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='crit1', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='crit2', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='crit3', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='doubterr', type_='float64'),
        gxe.ExpectColumnValuesToBeOfType(column='multiple', type_='float64'),
        gxe.ExpectColumnValuesToBeOfType(column='success', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='suicide', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='attacktype1_txt', type_='str'),
        gxe.ExpectColumnValuesToBeOfType(column='attacktype1', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='targettype1_txt', type_='str'),
        gxe.ExpectColumnValuesToBeOfType(column='targettype1', type_='int64'),
        gxe.ExpectColumnValuesToBeOfType(column='natity1_txt', type_='str'),
        gxe.ExpectColumnValuesToBeOfType(column='natity1', type_='float64'),
    ]
```

# Validaciones para

- Orden de columnas
- Rangos de valores
- Set de valores
- Nulos
- Tipo de dato





# ML model



```
Model: Linear Regression  
Accuracy: 0.1510  
Model: Lasso  
Accuracy: 0.1262  
Model: Ridge  
Accuracy: 0.1510  
Model: RidgeCV  
Accuracy: 0.1510  
Model: LassoCV  
Accuracy: 0.1257  
Model: LassoLars  
Accuracy: 0.1262  
Model: LassoLarsCV  
Accuracy: 0.1510  
Model: LassoLarsIC  
Accuracy: 0.1510  
Model: XGBRegressor  
Accuracy: 0.1571
```



```
X = df.drop(['eventid', 'id_location', 'id_date',  
'id_attack', 'id_perpetrator', 'id_disorder',  
'nkill', 'property', 'multiple'], axis=1)  
  
y = df['nkill']
```

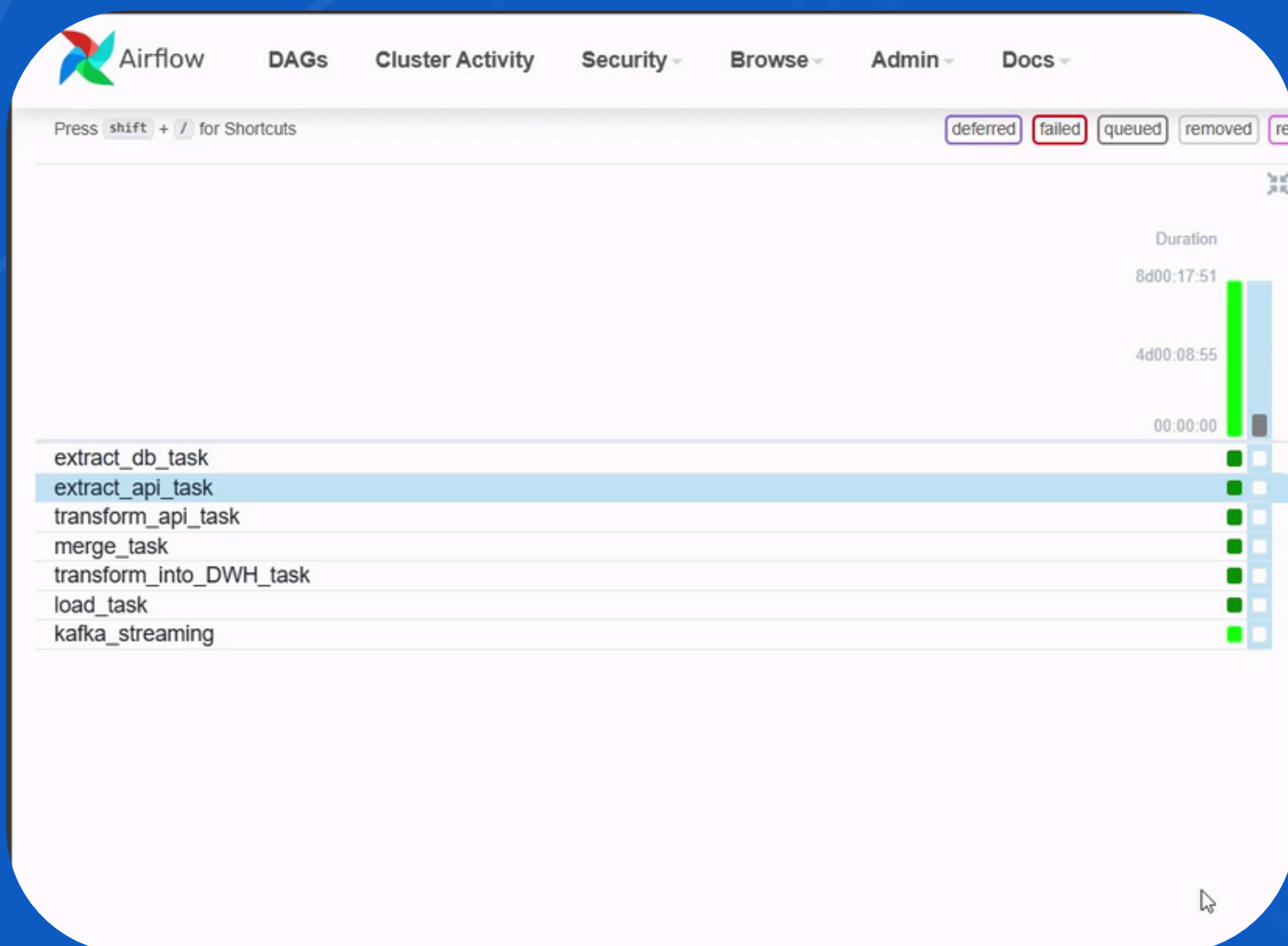




## Consumer kafka\_functions.py



# Kafka



```
[2024-10-23, 02:25:49 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708140003, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:25:53 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708140026, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:25:57 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708140039, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:01 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708150012, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:05 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708150026, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:09 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708150040, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:13 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708160014, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:17 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708160034, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:21 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708170012, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:25 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708170024, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:29 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708180003, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:33 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708220019, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:37 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708190003, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:41 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708190030, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:45 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708200003, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:49 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708200023, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:53 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708210001, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
[2024-10-23, 02:26:57 UTC] {kafka_functions.py:43} INFO - Batch sent: [{'eventid': 201708210018, 'extended': 0, 'multiple': 0, 'success': 1, 'suicide': 0, 'nkill': 0, 'property': 1, 'ishostkid': 0, 'nwound': 7.0, 'id_location': '20910Buca district', 'id_date': 19700101, 'id_attack': 33209.06111999, 'id_perpetrator': '201708310015Unknown0', 'id_disorder': 5}]
```

```
["id_location": "1679Saint Petersburg", "id_date": 19700101, "id_attack": 71167.08111999, "id_perpetrator": "201708310013Unknown0", "id_disorder": 5}, {"eventid": 201708310015, "extended": 0, "multiple": 0, "success": 1, "suicide": 0, "nkill": 0, "property": 1, "ishostkid": 0, "nwound": 7.0, "id_location": "20910Buca district", "id_date": 19700101, "id_attack": 33209.06111999, "id_perpetrator": "201708310015Unknown0", "id_disorder": 5}]
^CProcessed a total of 457 messages
user@01075cfd6e86 ~]$ ^C
```

# Real-time dashboard





**Muchas gracias  
por su atención**

