Data Connection Details

Participants have access to the Global Database of Events, Language, and Tone (GDELT) dataset, specifically the GDELT 1.0 Global Knowledge Graph (GKG) and Events data. This dataset provides a rich source of information on global events, news articles, and their associated metadata, allowing you to analyze and derive insights from a wide range of topics across various domains.

Time Frame

The designated time window for participants to search and download the relevant GDELT 1.0 GKG and Events data files is **[08/13/2023] to [08/13/2024].** Participants are free to use all or part of the data considered within the time window provided.

Data Access

Data **should be accessed through the RAW CSV files loaded in the GDELT webpage.** These files contain the raw, unprocessed data extracted from the GDELT database, allowing you to perform your own data preprocessing and analysis as needed.

- GDELT 1.0 GKG Data**:** The knowledge graph dataset. Each entry in the GKG dataset includes information such as event type, location, actors involved, and relevant themes.
- **GDELT 1.0 Events Data:** This dataset complements the GKG data by offering a more granular view of individual events and their characteristics.

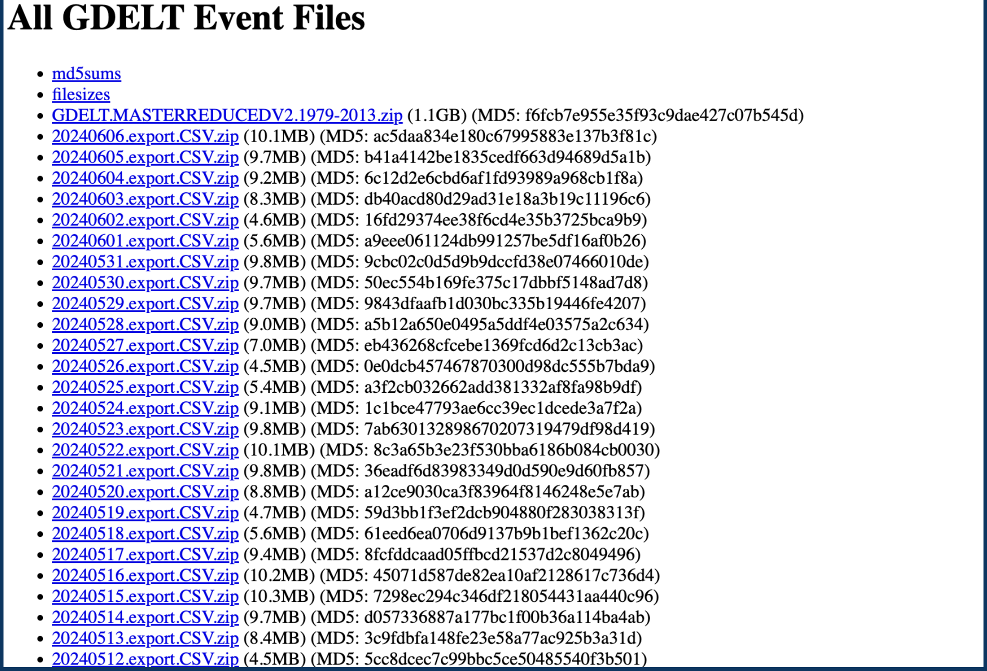
Factored

Note: In this datathon, you are not allowed to access the BigQuery version of the dataset. Use the raw files and create your own data pipelines.

Getting Started

1. Downloading the files

The dataset contains a list of raw files that can be downloaded inside an HTML page. This page can be retrieved using a GET request and then scrapped.



All GDELT Event Files

- md5sums
- filesizes
- GDELT.MASTERREDUCEDV2.1979-2013.zip (1.1GB) (MD5: f6fcb7e955e35f93c9dae427c07b545d)
- 20240606.export.CSV.zip (10.1MB) (MD5: ac5daa834e180c67995883e137b3f81c)
- 20240605.export.CSV.zip (9.7MB) (MD5: b41a4142be1835cedf663d94689d5a1b)
- 20240604.export.CSV.zip (9.2MB) (MD5: 6c12d2e6cbd6af1fd93989a968cb1f8a)
- 20240603.export.CSV.zip (8.3MB) (MD5: db40acd80d29ad31e18a3b19c11196c6)
- 20240602.export.CSV.zip (4.6MB) (MD5: 16fd29374ee38f6cd4e35b3725bca9b9)
- 20240601.export.CSV.zip (5.6MB) (MD5: a9eee061124db991257be5df16af0b26)
- 20240531.export.CSV.zip (9.8MB) (MD5: 9cbc02c0d5d9b9dccfd38e07466010de)
- 20240530.export.CSV.zip (9.7MB) (MD5: 50ec554b169fe375c17dbbf5148ad7d8)
- 20240529.export.CSV.zip (9.7MB) (MD5: 9843dfaafb1d030bc335b19446fe4207)
- 20240528.export.CSV.zip (9.0MB) (MD5: a5b12a650e0495a5ddf4e03575a2c634)
- 20240527.export.CSV.zip (7.0MB) (MD5: eb436268cfcebe1369fcd6d2c13cb3ac)
- 20240526.export.CSV.zip (4.5MB) (MD5: 0e0dcb457467870300d98dc555b7bda9)
- 20240525.export.CSV.zip (5.4MB) (MD5: a3f2cb032662add381332af8fa98b9df)
- 20240524.export.CSV.zip (9.1MB) (MD5: 1c1bce47793ae6cc39ec1dcede3a7f2a)
- 20240523.export.CSV.zip (9.8MB) (MD5: 7ab630132898670207319479df98d419)
- 20240522.export.CSV.zip (10.1MB) (MD5: 8c3a65b3e23f530bba6186b084cb0030)
- 20240521.export.CSV.zip (9.8MB) (MD5: 36eadf6d83983349d0d590e9d60fb857)
- 20240520.export.CSV.zip (8.8MB) (MD5: a12ce9030ca3f83964f8146248e5e7ab)
- 20240519.export.CSV.zip (4.7MB) (MD5: 59d3bb1f3ef2dcb904880f283038313f)
- 20240518.export.CSV.zip (5.6MB) (MD5: 61eed6ea0706d9137b9b1bef1362c20c)
- 20240517.export.CSV.zip (9.4MB) (MD5: 8fcfddcaad05ffbcd21537d2c8049496)
- 20240516.export.CSV.zip (10.2MB) (MD5: 45071d587de82ea10af2128617c736d4)
- 20240515.export.CSV.zip (10.3MB) (MD5: 7298ec294c346df218054431aa440c96)
- 20240514.export.CSV.zip (9.7MB) (MD5: d057336887a177bc1f00b36a114ba4ab)
- 20240513.export.CSV.zip (8.4MB) (MD5: 3c9fdbfa148fe23e58a77ac925b3a31d)
- 20240512.export.CSV.zip (4.5MB) (MD5: 5cc8dcec7c99bbc5ce50485540f3b501)

Figure 1. Example list of files from the GDELT dataset.

2. Processing initial time frame

Once the list of files has been parsed, the participants must figure out a way of downloading the files, uncompress them as they arrive as .ZIP files, and store them on a data lake. Then, use a processing engine to pick those files and transform them into tables. Furthermore, use those tables on downstream processes to feed your data model and your application.

You might want to take a look at the medallion architecture as a guide to structure data layers inside a data platform.

3. Processing Newer files

GDELT updates the list of files daily, so newer files can use a very similar process to download the latest file every day and append the contents to the events table or generate an incremental load for the Knowledge Graph table.

4. Important considerations

One important detail is that this exercise should resemble a production-ready platform. As such, the following aspects should be carefully considered:

- Modularize your data processing code. Use classes and functions to achieve that objective.
- If possible, create automated tests for your code and data.
- Consider a data processing engine that is scalable and fault-tolerant.

  One example is Apache Spark and its hosted cloud solution Databricks. A 14-day trial is available with Free Databricks credits in Microsoft Azure. **Although, the underlying compute and storage resources is not free of charge.**

- Design and implement a data quality strategy according to your application needs.

  Take a look at tools such as Soda, great expectations, or Delta Live Tables.

- Data governance, including security, access, discoverability, and documentation is very important. You want your customers to consume your data!
- Consider orchestration tools like Apache Airflow to create production-ready data and machine learning pipelines.

Factored

GDELT Dataset Documentation

The GDELT project offers comprehensive documentation about their datasets. They contain a description of the purpose of the dataset, columns involved, and an overview of how it was obtained. You can read the documentation in the following links:

- GDELT 1.0 GKG Data
- GDELT 1.0 Events Data

You can also find the list of columns, separated by tabs, for the **events dataset** in this link. There is no header file for the GKG dataset, so you must use the names in the documentations to add the columns manually.