Database Design

This document contains 2 sections:

1. Database Description - decribes the final table for use in analysis and its origins from GDELT data.
2. Database Process & Scripts - decribes the process for deriving the final table from the GDELT data.

1. Database Description

The table on which we run our analysis has the following structure:

```
CREATE TABLE `city_day_event_counts` (
  `num_conflicts` bigint(20) DEFAULT NULL,
  `sum_num_mentions` double DEFAULT NULL,
  `sum_num_articles` double DEFAULT NULL,
  `sum_num_sources` double DEFAULT NULL,
  `sqldate` bigint(20) DEFAULT NULL,
  `feature_name` varchar(40),
  `feature_id` varchar(13),
  `state_alpha` varchar(2),
  `county_name` varchar(40),
  `prim_lat_dec` double DEFAULT NULL,
  `prim_long_dec` double DEFAULT NULL,
  PRIMARY KEY (sqldate, feature_id)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

Feature name is equivalent to city name.

num_conflicts is the dependent variable. The hypothesis is there a space-time relationship of conflict events in the United States. We are regressing `num_conflicts` in a given city on a given day on the `num_conflicts` in the same city on previous days and other cities on previous days, up to 2 weeks prior. All `num_conflicts` will be weighted by the `sum_num_*` columns. These `sum_num_*` columns measure the significance of the conflict events ocurring on that day in that city.

Origin of `city_day_event_counts` table

In order to understand the spatial-temporal relationship of conflict events in the United States from April 2013 until the recent past, the application is built on an aggregated subset of the GDELT database.

The GDELT Project collects and stores events gathered from many different news media, including broadcast, web and print. It identifies attributes of the event, such as actors, locations and significance and categorizes each into an event hierarchy.

The data is subset based on the following criteria:

- **Events ocurring on or after 1 April 2013:** We exclude any events ocurring before April 2013 for two important reasons. The first is this is when data collection by GDELT started and all events prior have been back-filled. Second, there is no SOURCEURL field. The SOURCEURL field is an important field for sanity-checking the original source of events.
- **Events within the United States:** We are limiting the scope of our analysis to events ocurring within the United States (e.g. Actor1Geo*Country and Actor2Geo*Country = 'US').
- **Events with a QuadClass of 3 or 4:** The QuadClass field is the highest level of the event type hierarchy and includes only the values 1=Verbal Cooperation, 2=Material Cooperation, 3=Verbal Conflict, 4=Material Conflict.

After subsetting the table, additional steps are required to build the conflict analysis database. These are described below in **Database Scripts**.

An image summarizing table extraction can be found below.

2. Database Process & Scripts

The process for building the conflict analysis database is as follows:

1. BGSE EC2 AMI & R: Import data from GDELT website server to RDS

`/scripts/importGdeltData.R`

2. Sqoop: Sqoop data in and out of RDS and HDFS

`/scripts/shell/sqoop-import.sh`

3. EMR Impala Build events table and subset to conflict events in the US

`/scripts/impala/events_to_usa_conflict_events.sql`

4. EMR Impala: Download and create the `gnis_features` table using `/scripts/impala/gnis_features.sql`. Subset features to cities (e.g. `FEATURE_CLASS = 'Populated Place'`).

5. EMR Impala: Join events with features to create `events_with_cities` table

`/scripts/impala/events_with_cities.sql`

6. EMR Impala: Create the city-date counts table

`/scripts/impala/city_day_event_counts.sql`

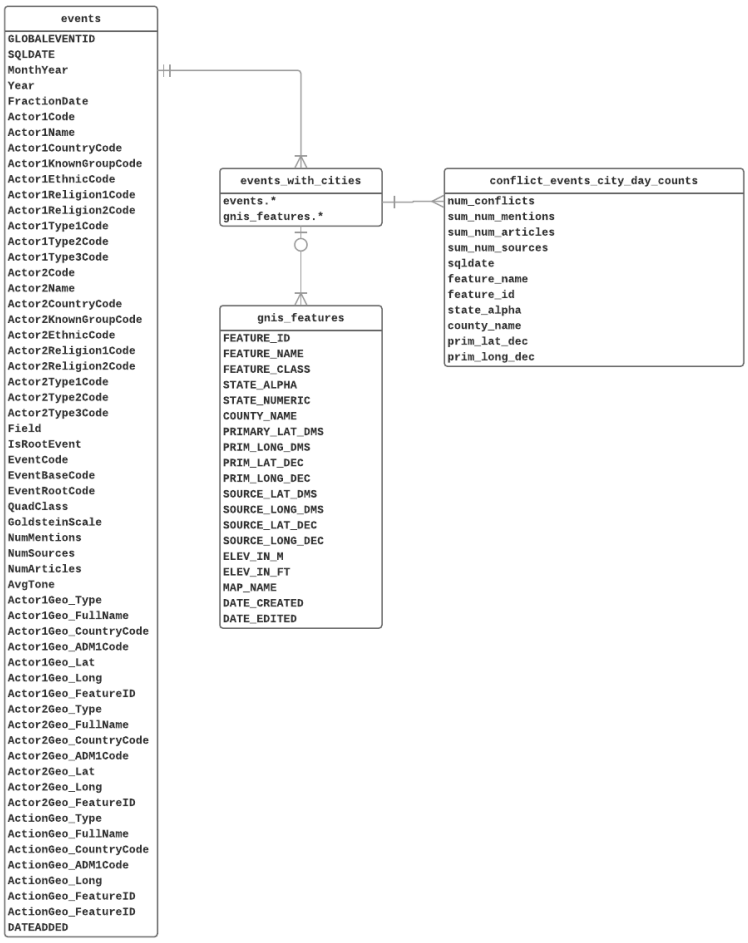

Figure 1: Data Model

Final script

The current dump file can be found on Google Drive: `city_day_event_counts.sql`.