

# Project Assignment - Part 1

Roberto Pellungrini, Anna Monreale, Cristiano Landi

October 17, 2023

## Introduction

In **Part 1** of the project, you are required to create and populate a database starting from different files and perform some operations on it. In the following, you can find a set of incremental assignments, each one with a brief description of what you are required to produce and what tools you can use for the task.

## Build the datawarehouse

**Police.csv** contains the main body of data: a table with data about gun violence incidents between January 2013 and March 2018. The same table also includes information about the victims, the guns, and the locations. The file **dates.xml** maps each `date_id` from the **Police.csv** file to a real date.

Finally, the files **participant\_age.json** ( $F_1$ ), **participant\_type.json** ( $F_2$ ), and **participant\_status.json** ( $F_3$ ) are three dictionaries containing the data necessary to compute the *crime\_gravity* attribute. Given an instance  $x$ , you can compute the *crime\_gravity* using the following equation:

$$crime\_gravity(x) = F_1(x.participant\_age) * F_2(x.participant\_type) * F_3(x.participant\_status) \quad (1)$$

You have to split and integrate the main file to reproduce the schema in Figure 1.

The following assignments aim to build and deploy the schema on server `lds.di.unipi.it`. You should consider that there may be missing values, useless information, and/or the need to integrate additional data from elsewhere.

### *Assignment 0*

Create the database schema in Figure 1 using SQL Server Management Studio in server `lds.di.unipi.it`. The name of the database must be *GroupID\_DB* (example: `Group_01_DB`).

### *Assignment 1*

Write a Python program that splits the content of **Police.csv** and **dates.xml** into six separate tables: custody, gun, participant, date, incident, and geography. You will also have to write several functions to perform integration of the main data body. In particular:

- You will have to generate some missing ids, like `participant_id` and `geo_id`. Use the data that you have available in a suitable way to infer or generate these ids.
- the **crime\_gravity** attribute is the main measure of the data warehouse. You can compute its values using Eq. 1 and the additional files **participant\_age.json** ( $F_1$ ), **participant\_type.json** ( $F_2$ ), and **participant\_status.json**
- Retrieve the city and state where the incident occurred from the geographical information available in **Police.csv**. You can use additional external data to complete this task.

All the above operations must be done WITHOUT using the pandas library.

### *Assignment 2*

Write a Python program that populates the database *GroupID\_DB* with all the data you prepared in Assignment 1, establishing schema relations as appropriate.

When you want to deliver your first project, compress the folder and create a single .zip file, named LDS\_GroupID.zip. Then send an email to all teachers with the subject: LDS PART1 Group\_Id.

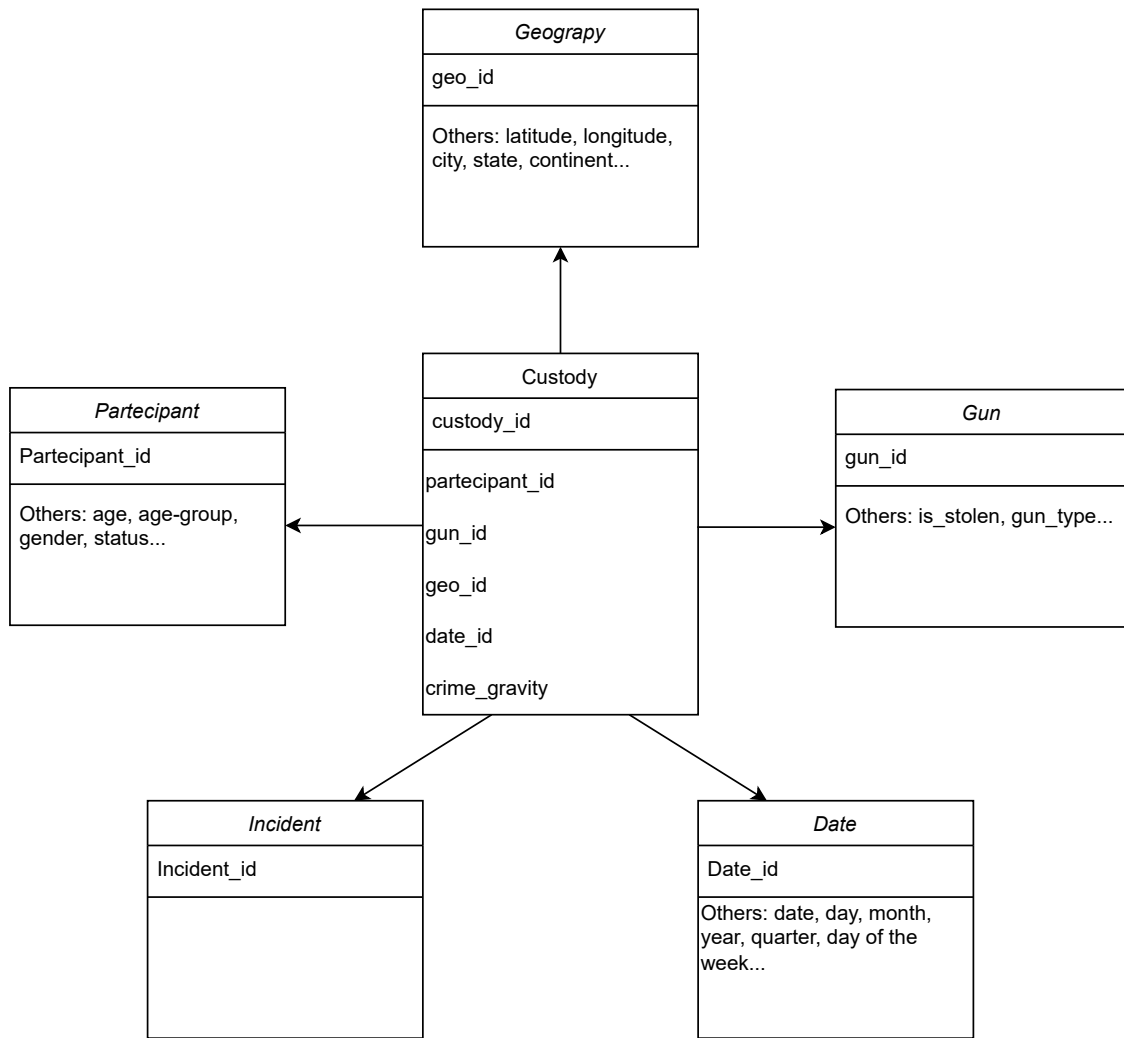


Figure 1: Datawarehouse schema of reference.