UNIVERSITÀ
DELLA CALABRIA

Data Warehouse

Paola Guarasci mat. 231847

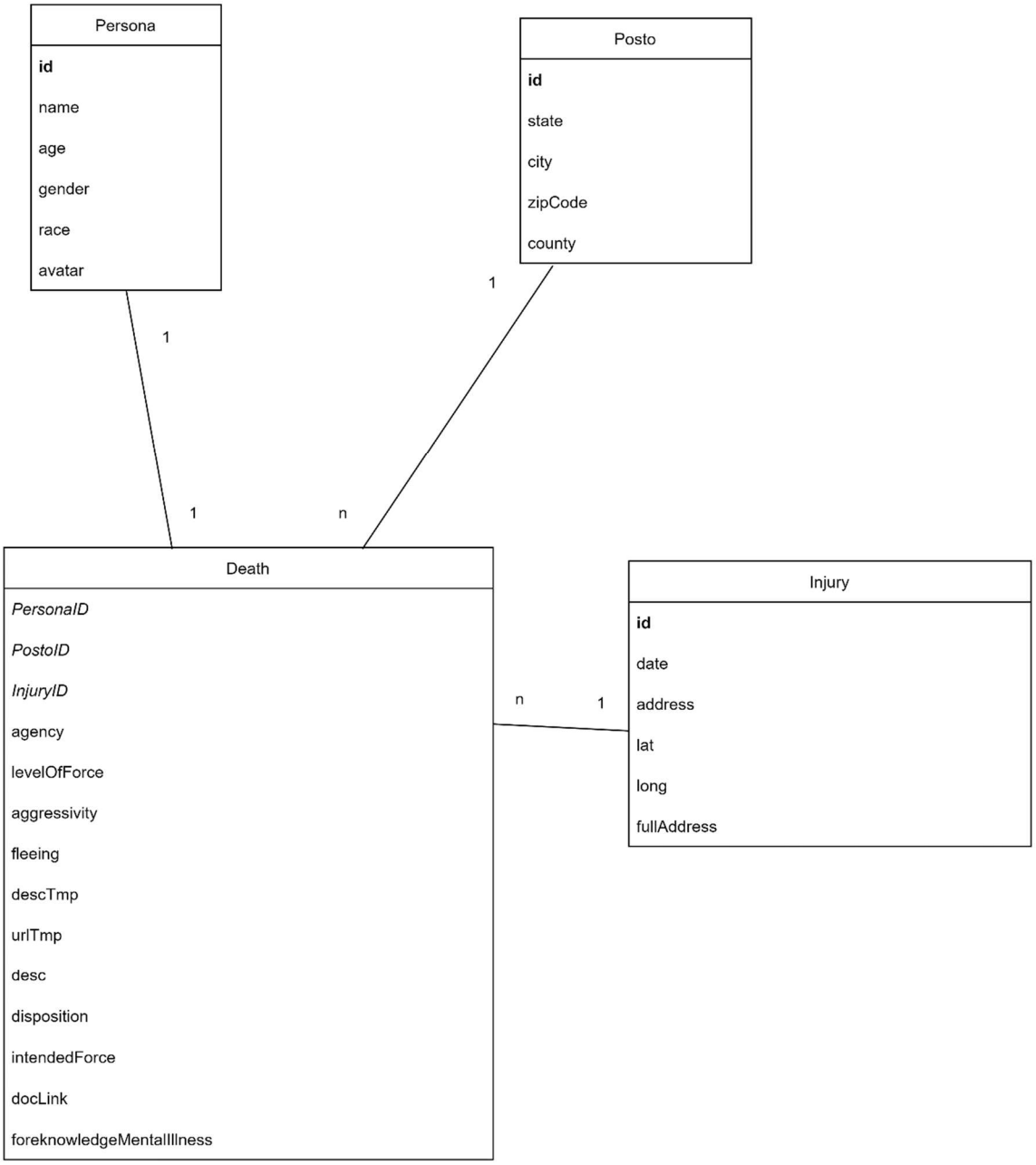
Descrizione del dataset

Il dataset utilizzato riguarda le morti innocenti per omicidio per mano della polizia (o organi di polizia) negli USA in un periodo storico che va dal 2000 al 2021. Ogni riga rappresenta un singolo evento di morte, riconducibile ad un'unica vittima.

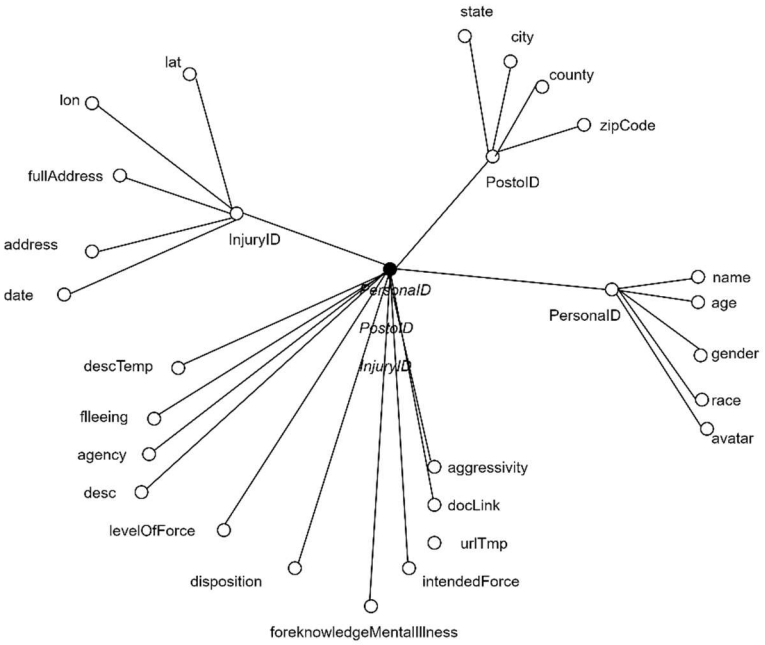
Per una questione di praticità si è deciso di applicare un mapping dei nomi con delle versioni più maneggiabili.

Nome originale	Nome Rimappapo
Unique ID	Id
Name	name
Age	age
Gender	gender
Race	race
URL of image	avatar
Date of injury resulting in death (month/day/year)	date
Location of injury (address)	address
Location of death (city)	city
State	state
Location of death (zip code)	zipCode
Location of death (county)	county
Full Address	fullAddress
Latitude	lat
Longitude	long
Agency or agencies involved	agency
Highest level of force	levelOfForce
Alleged weapon	weapon
Aggressive physical movement	aggressivity
Fleeing/Not fleeing	fleeing
Description Temp	descTemp
URL Temp	urlTemp
Brief description	desc
Dispositions/Exclusions INTERNAL USE, NOT FOR ANALYSIS	disposition
Intended use of force (Developing)	IntendedForce
Supporting document link	docLink
Foreknowledge of mental illness	foreknowledgeMetalIllness

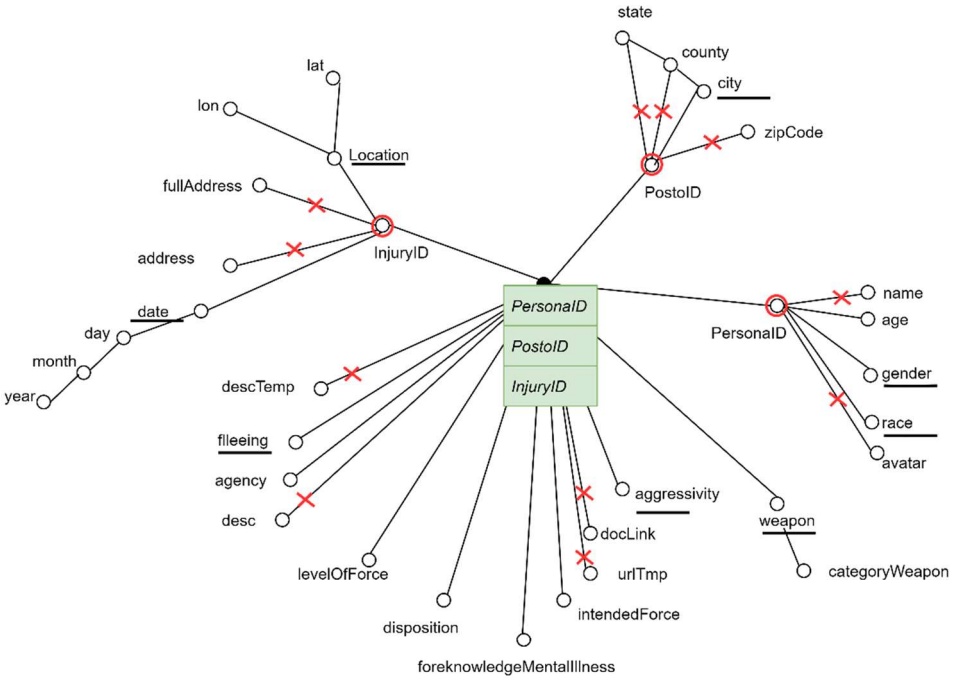
Schema ER

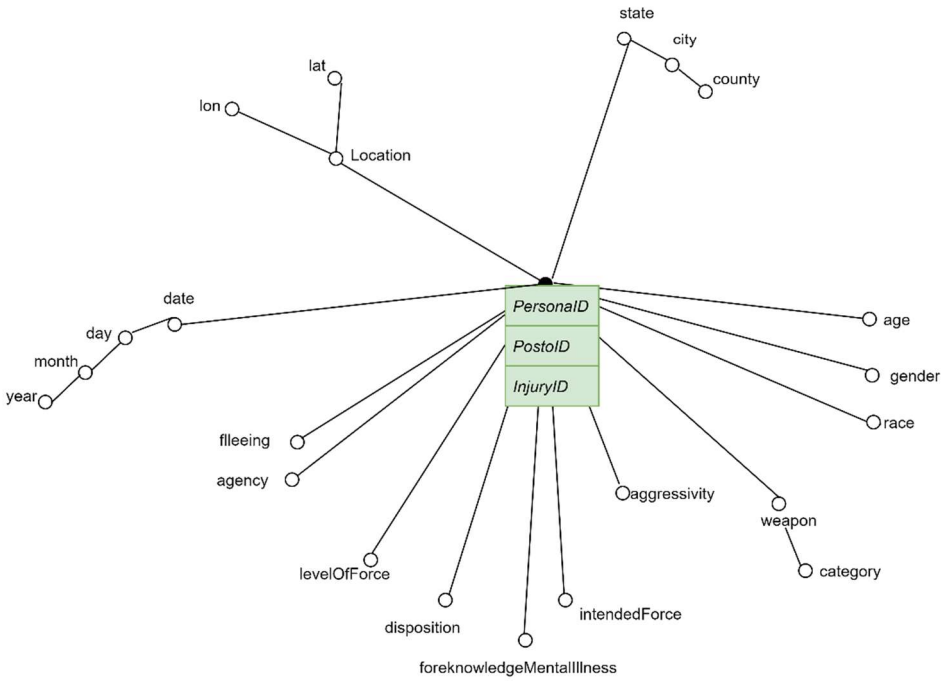

Persona
id
name
age
gender
race
avatar
Posto
id
state
city
zipCode
county
1
1
1
n
Death
PersonaID
PostoID
InjuryID
agency
levelOfForce
aggressivity
fleeing
descTmp
urlTmp
desc
disposition
intendedForce
docLink
foreknowledgeMentalIllness
n
1
Injury
id
date
address
lat
long
fullAddress

Albero degli attributi Originale

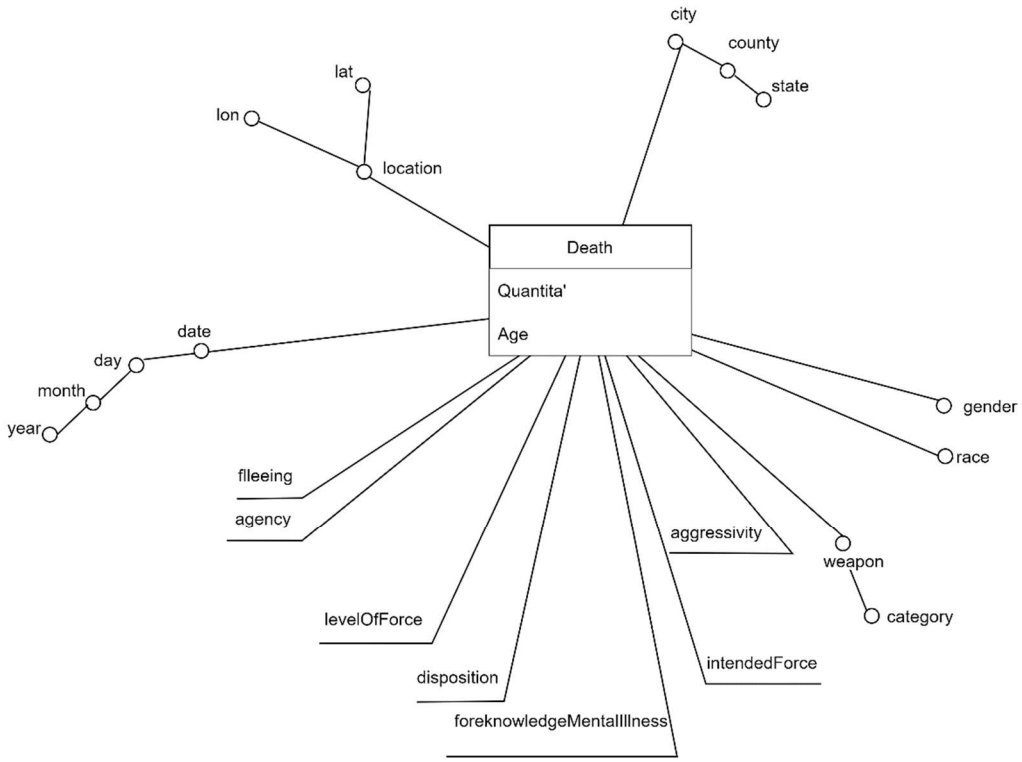


Albero degli attributi (potature e innesti)

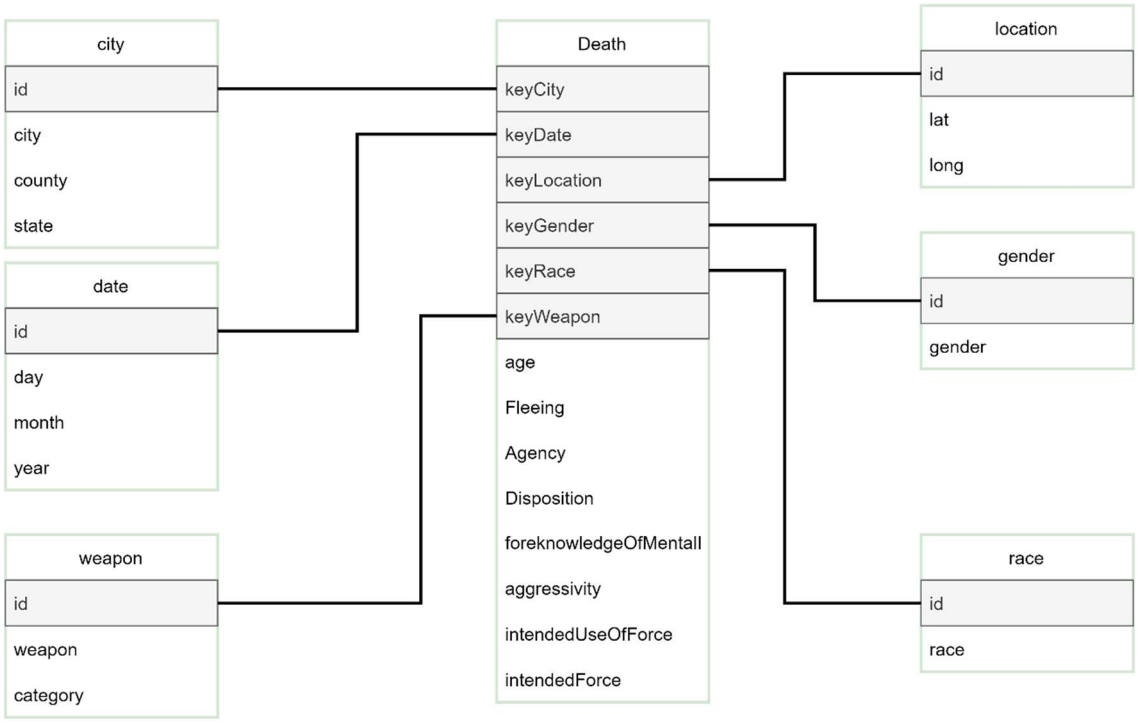




Schema di fatto



Schema STAR


city
id
city
county
state
date
id
day
month
year
weapon
id
weapon
category
Death
keyCity
keyDate
keyLocation
keyGender
keyRace
keyWeapon
age
Fleeing
Agency
Disposition
foreknowledgeOfMentall
aggressivity
intendedUseOfForce
intendedForce
location
id
lat
long
gender
id
gender
race
id
race

ETL

Per la preparazione del dataset è stato utilizzato Tableau Prep Build, il software messo a disposizione dalla stessa azienda che fornisce lo strumento OLAP Tableau, utilizzato successivamente in fase di analisi.

Le operazioni di pulizia svolte sul dataset sono state le seguenti:

Campo	Operazioni	Note
id	--	0% valori null
name	Eliminato (non necessario)	
age	Rimozione righe con valore null	1191 (<4%) valori null
gender	Rimozione righe con valore null	136 (<1%) valori null
race	Accorpamento valori leggermente diversi per errori di battitura	"european-American/White" accorpato con "European-American/White"
avatar	Eliminato (non necessario)	
date	--	Rimozione di una riga con data molto distante rispetto al range ottenuto dopo la pulizia degli altri campi
address	Eliminato (non necessario)	
city	Rimozione righe con valori null, Accorpamento automatico in base alla vicinanza lessicale	38 (<1%) valori null Es. Abingdon e Abington sono stata accorpate in un'unica entità.
state	Accorpamento manuale di due entità	DC e Washington accorpate
zipCode	Eliminato (non necessario)	
county	Rimozione righe con valori null, Accorpamento automatico in base alla vicinanza lessicale	15 (<1%) valori null Es. Alleghany e Allegheny sono stata accorpate in un'unica entità.
fullAddress	Eliminato (non necessario)	
lat	--	
long	Rimozione righe con valori null	1 (<1%) valori null
agency	Rimozione righe con valori null	77 (<1%) valori null
levelOfForce	Rimozione righe con valori null	3 (<1%) valori null
weapon	Rimozione righe con valori null Raggruppamento di entry diverse solo per il case	17000 (55%) valori null
aggressivity	Rimozione righe con valori null Raggruppamento valori simili (errori di battitura)	16997 (55%) valori null None e derivati tutto in None
fleeing	Rimozione righe con valori null Raggruppamento valori simili (errori di battitura)	16997 (55%) valori null Not fleeing (64%) e Not Fleeing (<1%) raggruppati tutti in Not fleeing
descTemp	Eliminato (non necessario)	
urlTemp	Eliminato (non necessario)	
desc	Eliminato (non necessario)	
disposition	Rimozione righe con valori null	1 (<1%)
IntendedForce	Rimozione righe con valori null	2 (<2%)
docLink	Eliminato (non necessario)	
foreknowledgeMetalIllness	Rimozione righe con valori null	61(<1%)

In generale la percentuale di dati null è poca e il dataset è sufficientemente ampio e consistente: si tratta di un totale di 30860 righe prima delle operazioni di pulizia. Dopo tutte le rimozioni è diventato circa un terzo, 11465 righe, in ogni caso sufficiente per le analisi svolte.

I campi utilizzati come attributi descrittivi sono stati lasciati tal quali, anche se spesso sono strutturati come categoria/sottocagoria. Per i campi utilizzati come dimensioni si è proceduto allo scorporamento della colonna utilizzando un campo calcolato ed un'espressione regolare per alimentarlo, al fine di ricostruire la corretta gerarchia.

Il dataset originale è in formato csv. Il formato CSV è un file di testo in cui si utilizza un formalismo per cui ogni riga è una riga della tabella e ogni campo è separato da un carattere speciale, spesso una virgola (,) o un punto e virgola (;). Ci si è resi conto che il parsing del file da parte di Microsoft Excel ha generato molti più dati null del previsto e soprattutto alcune righe presentavano campi *sfalsati* rispetto all'ordine reale delle colonne. Dopo una attenta analisi del file originale si è compresa la causa di questi *shift*, ovvero la presenza del delimitatore, del carattere speciale, anche all'interno di porzioni di testo e non solo come delimitatore in sé. Tutto ciò ha portato ad una serie di righe con valori spostati. La soluzione a questo inconveniente è stata importare direttamente il file csv in Tableau Prep Builder, senza quindi la conversione xlxs intermedia. Il parsing interno di Tableau si è dimostrato più efficace rispetto a quello utilizzato dal software Microsoft, riuscendo a distinguere quando il carattere speciale è in effetti usato come delimitatore e quando, al contrario, è presente in un virgolettato.