**SWITRS CALIFORNIA TRAFFIC ACCIDENT DATA**

## 1. Data identification

The Statewide Integrated Traffic Records System (SWITRS) is a database that collects and stores information on all traffic collisions that occur on public roads in California, in particular the data of my sample ranges from 2001 to 2020. It is maintained by the California Highway Patrol (CHP) and contains data on crashes that occur on all public roads, including freeways, highways, and local streets.

The SWITRS database includes information such as the date, time, and location of the crash, the type of collision (such as head-on or rear-end), the number of vehicles involved, the weather and lighting conditions at the time of the crash, and the types of injuries sustained by the drivers and passengers.

In addition to the basic crash data, SWITRS also includes more detailed information about the drivers involved in the collision, including their age, gender, and whether or not they were wearing a seatbelt at the time of the crash. Information about the vehicles involved is also collected, including the make, model, and year of the vehicle, as well as whether or not the vehicle was carrying passengers or cargo at the time of the crash. The analysis of such data is very useful to and policy makers to analyze traffic safety trends and develop strategies to improve roadway safety. The data can also be used by law enforcement agencies to identify high-risk areas and target enforcement efforts to reduce the number of crashes and fatalities on California's roadways.

SWITRS only includes information on collisions that are reported to law enforcement, which means that some crashes that occur may not be included in the database.

### 1.1. Datasets

There are 3 tables: COLLISIONS, PARTIES and VICTIMS

#### 1.1.1. Description

The tables belonging to the DB repository are collected by the California Highway Patrol (CHP) and its allied agencies.

#### 1.1.2. Origin (Source of data, e.g. Sensors, measurement stations, public/restricted access dataset...)

California Highway Patrol and its allied agencies.

#### 1.1.3. Collection (How data have been collected?)

The data are collected by mean of the following sources:

- Crash reports: CHP officers complete crash reports for all collisions that they investigate. The crash reports include information such as the data. Time, location, severity of the collision as well as information about the vehicles and people involved;
- Electronic data: CHP and allied agencies also collect electronic data from traffic cameras, speed radar and vehicle registration records;
- Vehicle registration records: The California Department of Motor Vehicles (DMV) maintains records of all registered vehicles in California. These records include information such as vehicle’s make, model, year, and license plate number;
- Third-party data: insurance agencies and medical providers, such information may provide a more detailed picture of the traffic collision.

The data, divided in different tables, according to the entities involved in the traffic accident, are then entered into the SWITRS database by CHP.

#### 1.1.4. Update/Timing (e.g., one single time, each hour, each week, ...)

The data is updated on a daily basis.

#### 1.1.5. Tables details

##### 1.1.5.1. Collisions

```
case_id     jurisdiction   officer_id     reporting_district    chp_shift
    population     county_city_location  county_location       special_con
dition      beat_type      chp_beat_type  city_division_lapd    chp_beat_cl
ass  beat_number    primary_road   secondary_road distance       direction
    intersection   weather_1      weather_2      state_highway_indicator
    caltrans_county       caltrans_district     state_route    route_suffi
x    postmile_prefix       postmile       location_type  ramp_intersection
    side_of_highway       tow_away       collision_severity    killed_vict
ims  injured_victims       party_count    primary_collision_factor     pcf
_violation_code    pcf_violation_category       pcf_violation  pcf_violati
on_subsection      hit_and_run    type_of_collision     motor_vehicle_invo
lved_with   pedestrian_action     road_surface   road_condition_1      roa
d_condition_2      lighting       control_device chp_road_type  pedestrian_
collision   bicycle_collision     motorcycle_collision  truck_collision
    not_private_property  alcohol_involved      statewide_vehicle_type_at
_fault      chp_vehicle_type_at_fault    severe_injury_count   other_visib
le_injury_count    complaint_of_pain_injury_count       pedestrian_killed_
count       pedestrian_injured_count     bicyclist_killed_count       bic
yclist_injured_count       motorcyclist_killed_count    motorcyclist_injur
ed_count    primary_ramp   secondary_ramp latitude       longitude      col
lision_date collision_time process_date
```

##### 1.1.5.2. Parties

```
id      case_id party_number  party_type     at_fault       party_sex party_ag
e       party_sobriety party_drug_physical   direction_of_travel   party_safet
y_equipment_1  party_safety_equipment_2 financial_responsibility  hazardous_m
aterials       cellphone_in_use      cellphone_use_type    school_bus_related
oaf_violation_code    oaf_violation_category oaf_violation_section       oaf
_violation_suffix     id      case_idparty_number   victim_role    victim_sex
       victim_age     victim_degree_of_injury      victim_seating_position
       victim_safety_equipment_1    victim_safety_equipment_2    victim_ejec
ted
other_associate_factor_1     other_associate_factor_2 id  case_idparty_numbe
r       victim_role    victim_sex     victim_age     victim_degree_of_injury
       victim_seating_position       victim_safety_equipment_1    victim_safe
ty_equipment_2 victim_ejected party_number_killed  party_number_injured  mov
ement_preceding_collision    vehicle_year   vehicle_make   statewide_vehicle_
type chp_vehicle_type_towing chp_vehicle_type_towed party_race
```

##### 1.1.5.3. Victims

```
id   case_id        party_number   victim_role    victim_sex     victim_age
     victim_degree_of_injury      victim_seating_position      victim_safe
ty_equipment_1     victim_safety_equipment_2    victim_ejected
```

#### 1.1.6. Other Info

This database is a valuable resource for anyone who is interested in traffic safety. It provides comprehensive and up-to-date overview of traffic collisions in California and can be used to identify areas where safety improvements are needed.

2. **Data lifecycle** (use Arass et al. “Data lifecycles analysis: towards intelligent cycle” as a reference)

   2.1. Needed phases of DLC (What phase is mandatory? What phase could be useful?)

| | | |
|---|---|---|
| **Plan** | | |
| **Create/Receive** | 1 | |
| **Integration** | 1 | |
| **Filtering** | 1 | |
| **Anonymity** | 1 | |
| **Enrichment** | 1 | |
| **Analyze** | 1 | |
| **Visualization** | 1 | |
| **Storage** | 1 | |
| **Destruction** | | |
| **Archiving** | 1 | |
| **Total** | 9 | |

2.2. Lifecycle selection (What is the best lifecycle for your needs? Why? Look at table 2 of the previously referred paper)

Based on the results obtained and comparing with the table in references, the most appropriate data model could be a mixture of Big Data, USGS and Hindawi data models.

Moreover, it's important to point out that the most appropriate data model to the describe the database in hand would depend on the specific requirements of the California Highway Patrol, which is responsible of maintaining the database.

In the following I am giving some potential reasons about a mixture of the beforehand mentioned data models can suit the example in hand:

- Hindawi model: It could be used to model the SWITRS database because it allows for complex relationships between entities, which is important in a database that contains a large amount of interconnected data. For example, in the SWITRS database, collisions are related to multiple entities, such as drivers, passengers, and vehicles, and the Hindawi model could be used to represent these relationships;

- Big data model: The SWITRS database is a large and complex database that contains a vast amount of data. A big data model could be used to handle this volume of data, as well as to provide tools for real-time analysis and processing;

- USGS model: this data model is used to represent geological data, and it could be used to model the SWITRS database because it allows for the representation of complex geospatial relationships. For example, in the SWITRS database, collisions occur at specific locations on roads, and the USGS model could be used to represent these locations and their relationships to other entities in the database.

So, in summary, a mixture of data models, such as the Hindawi model, the Big Data model, and the USGS model, could be used to model the SWITRS database due to its complexity and size, as well as the need to represent complex relationships and geospatial data.

3. **Conceptual modelling of data** (using E-R model)

3.1. List of Entities and relationships with attributes and keys

For the collision table the PK is the case_id;

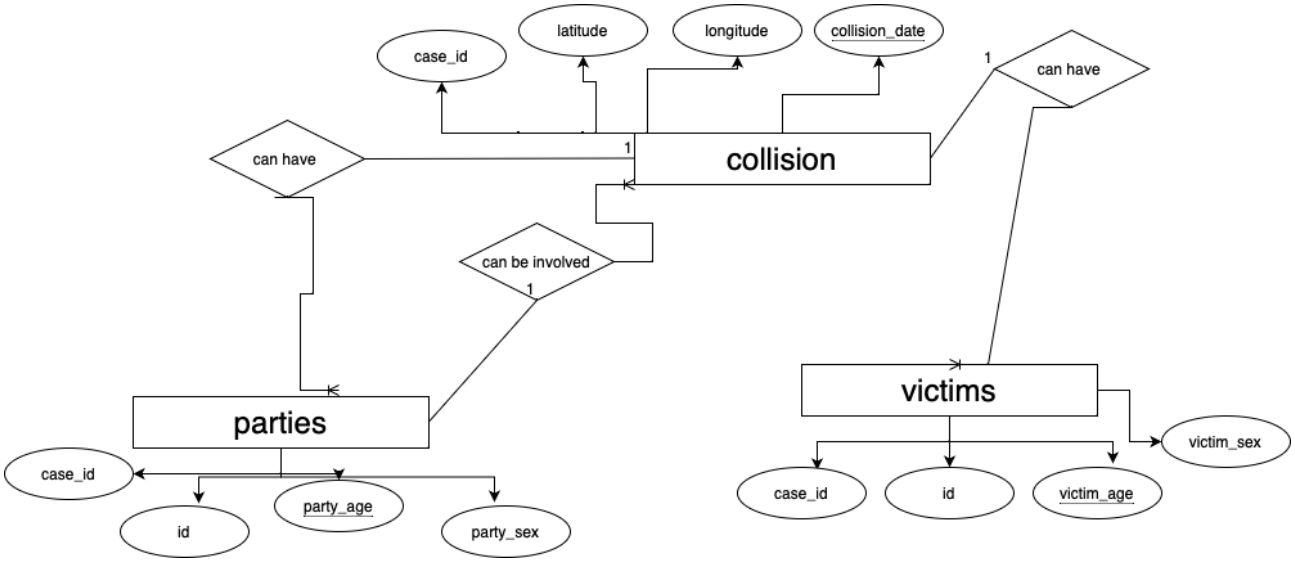
For the victims table the PK is the victim_id, the FK is the case_id;

For the parties table the PK is the party_id, the FK is the case_id;

Some relationship:

- A collision can have many parties;
- A party can be involved in many collisions;
- A collision can have many victims.

3.2. E-R graphic representation




## 4. On-Line Analytical Processing (OLAP) modelling

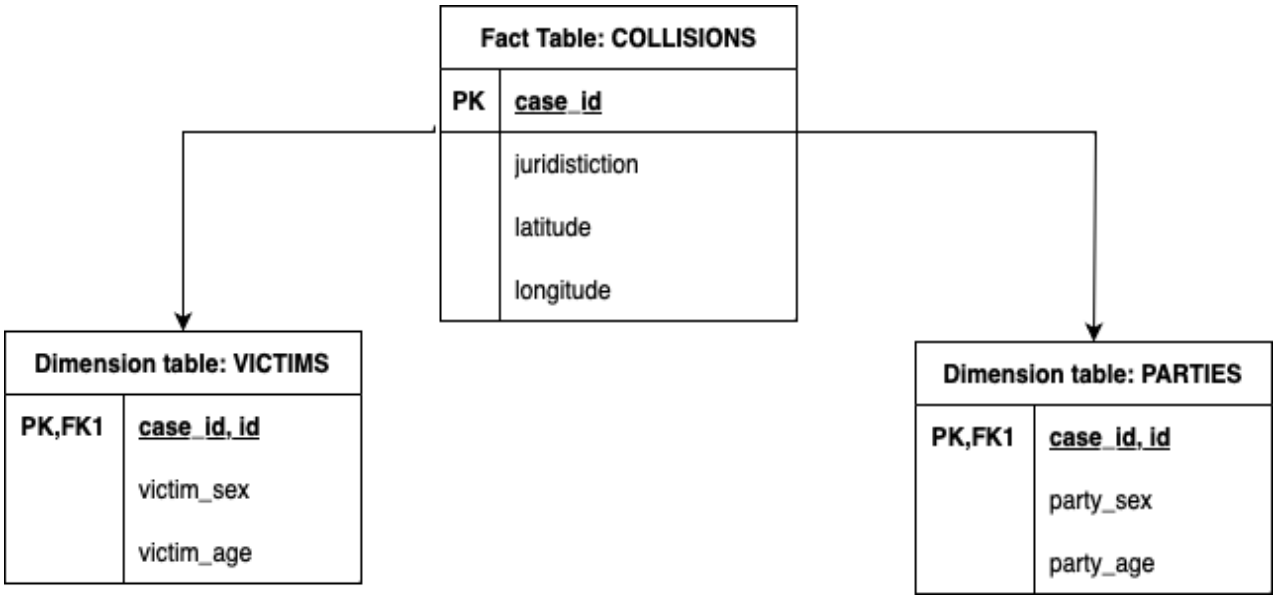
### 4.1. OLAP conceptual model of facts

The OLAP conceptual model of facts can be the following:

- Fact: Collisions
- Dimensions:
  - Jurisdiction
  - Latitude
  - Longitude
  - Collision date
  - Collision time
  - Pedestrian killed count
  - Killed victims
  - ….
  - ….
- Measures:
  - Number of collisions
  - Map of the collisions
  - Average age of involved people
  - …
  - …

### 4.2. STAR schema of facts

Being the tables so wide (many variables) in the following star schema I have only reported the fact tables with some attributes as well as dimensions tables with some attributes too.




4.3. Some sample queries

```
SELECT collision_date , 1 as {CRASH_COL} , IIF(COLLISION_SEVERITY='fatal', 1, 0) as {FATALITY_COL}
FROM collisions WHERE {DATE_COL} IS NOT NULL AND {DATE_COL} BETWEEN '2019-01-01' AND
'2020-11-30'
```

This query could be used to investigate trends in collision fatalities over time. Moreover, the number of fatal collisions has generally decreased over the specified time range (2019-01-01 to 2020-11-30); there are certain times of the year (or specific days) when fatal collisions are more likely to occur.

```
SELECT collision_date , {FACTOR_COL} , count(1) as total FROM collisions AS c LEFT JOIN parties as
p ON p.case_id = c.case_id WHERE {DATE_COL} IS NOT NULL AND {DATE_COL} BETWEEN
'{START_YEAR}-01-01' AND '{LAST_YEAR}-11-30' AND p.{FACTOR_COL} in ('{MAKE_1}', '{MAKE_2}')
GROUP BY 1, 2 ORDER BY 1, total DESC
```

This query could be used to investigate trends in collisions involving specific factors over time. In fact, there has been an increase in the number of collisions involving certain vehicle makes over the specified time range. Would be interesting to investigate if the number of collisions involving certain vehicle makes varies by geographic region within California.

## 5. Conclusions

The launched queries needed for the analyses of the SWITRS dataset provide valuable insights into the patterns and causes of traffic collisions in California. These insights can represent an example to help guide policy decisions and resource allocation towards areas with the highest need for intervention and improvement.

Appendix: Table 2 of Arass et al. paper

**Table 2 : Lifecycle score according to the retained phases**

| | Information lifecycle | Hindawi | DataONE | USGS | Big Data | IBM | DDI | CIGREF | CRUD | Enterprise | Pyramid | PII |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Plan** | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Create/Receive** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **Integration** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| **Filtering** | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Anonymity** | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Enrichment** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Analyze** | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| **Visualization** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Storage** | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| **Destruction** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| **Archiving** | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| **Total** | 5 | 8 | 5 | 7 | 6 | 5 | 4 | 4 | 5 | 4 | 3 | 3 |