DWI LAB – Mini-Project I

MULTIDIMENSIONAL PROJECT DESIGN

1. Source description:

a. Brief description concerning the content of the database – general structure and its operational usage:

We chose the crash/accidents data from the **City of Chicago** having 1.1M+ records split to 3 datasets (*Crashes, Vehicles, People*). The crash dataset has 48 features, Vehicles has 71, and People has 29. The data contains information about people, vehicles and each traffic crash on city streets involved in a crash and if any injuries were sustained. The vehicle dataset contains information about vehicles (or units as they are identified in crash reports) involved in a traffic crash. “Vehicle” information includes motor vehicle and non-motor vehicle modes of transportation, such as bicycles and pedestrians. Each mode of transportation involved in a crash is a “unit” and get one entry here.

Each vehicle/pedestrian/motorcyclist/bicyclist is considered an independent unit that can have a trajectory separate from the other units. However, people inside a vehicle including the driver do not have a trajectory separate from the vehicle in which they are travelling and hence only the vehicle they are travelling in get any entry here. This type of identification of “units” is needed to determine how each movement affected the crash.

Data for occupants who do not make up an independent unit, typically drivers and passengers, are available in the People table. Many of the fields are coded to denote the type and location of damage on the vehicle. Vehicle information can be linked back to Crash data using the “RD NO” field. Since this dataset is a combination of vehicles, pedestrians, and pedal cyclists not all columns are applicable to each record. In the data of people, each record corresponds to an occupant in a vehicle listed in the Crash dataset.

Some people involved in a crash may not have been an occupant in a motor vehicle, but may have been a pedestrian, bicyclist, or using another non-motor vehicle mode of transportation. Many of the crash parameters, including street condition data, weather condition, and posted speed limits, are recorded. A vehicle can have multiple occupants and hence have a one to many relationship between Vehicle and Person dataset. However, a pedestrian is a “unit” by itself and have a one to one relationship between the Vehicle and Person table.

Data Transformation Description-

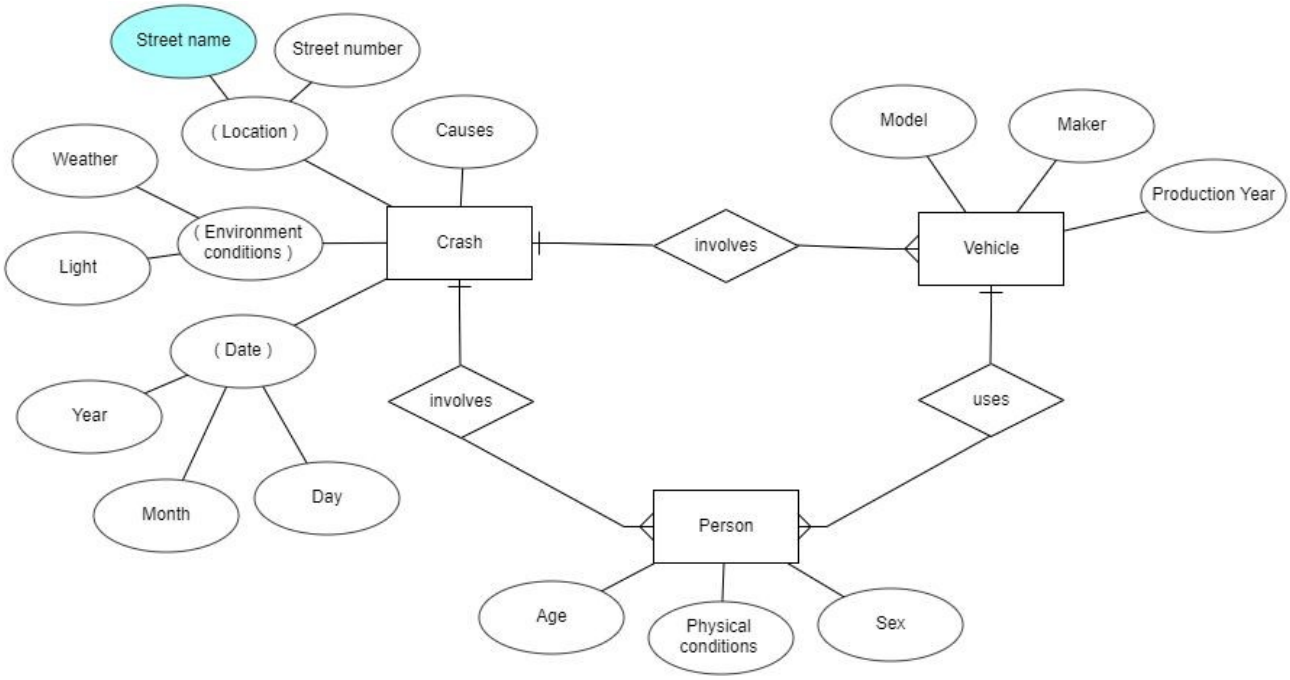
We split these 3 tables into 1 Fact, 5 Dimension tables and a Bridge table. The Crash table acts as a fact table and has Foreign Keys crash_key, date_id and location_id and measures like num_units, injuries_total, etc(see ER diag. for details). There was a many to many relationship twice between a fact(crash) and 2 dimensions, people and vehicle which is in violation of the dimensional modeling rules, so we decided to introduce a Bridge table between these 3 tables(see ER diagram for details)

These tables feature a column called **RD_NO** , which is the report number of the crash. For each individual, the report number referred to the crash they were involved in, so multiple individuals may show up with the same report number. To make our data linkage between

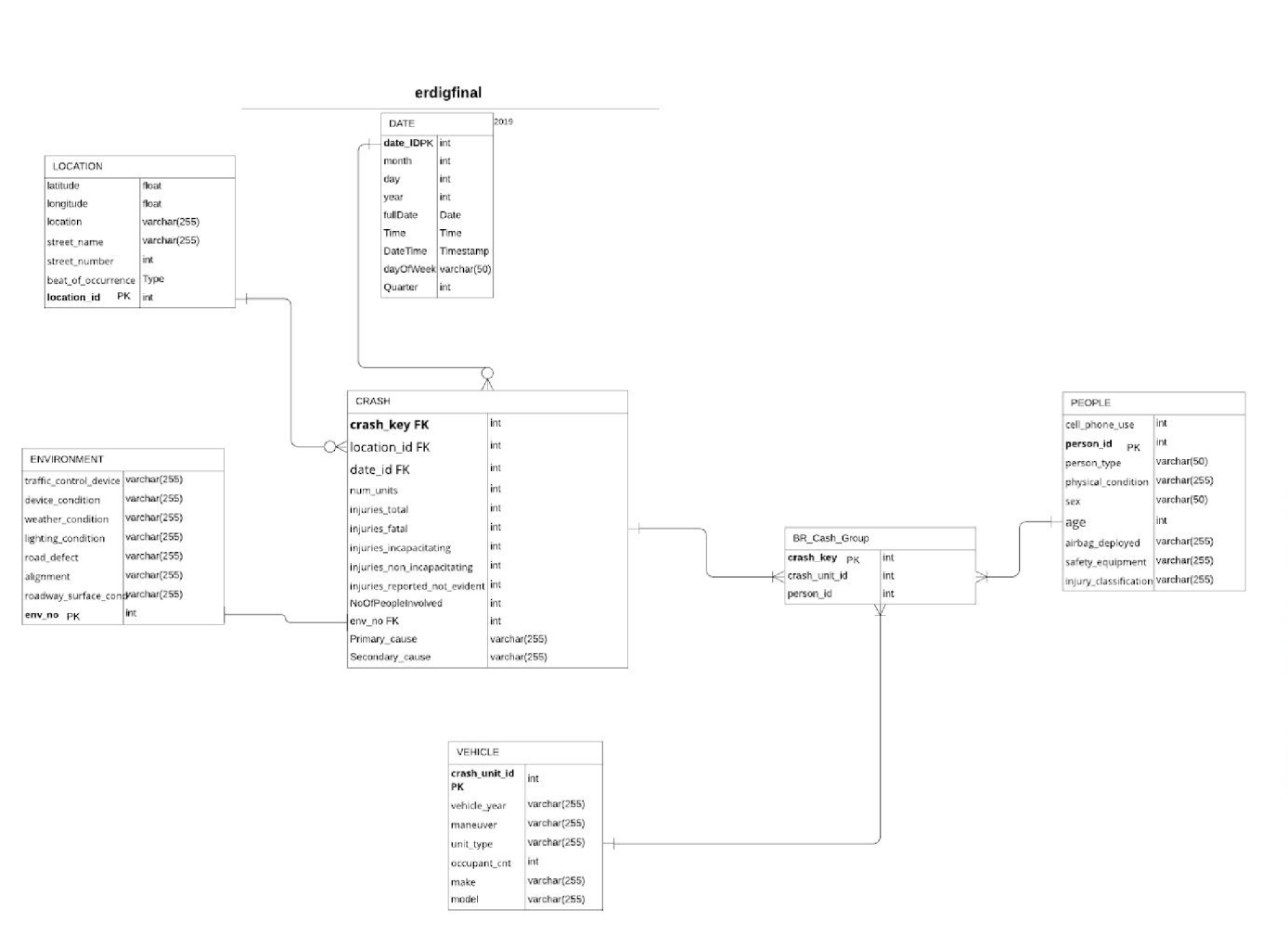
dimensions and fact, Foreign Keys and Primary keys are defined. PK date_id for the Date database will be a column having unique identifier based on day-month-year combination of the crash. Similarly, PK location_id for Location database will be a unique identifier based on latitude and longitude values where the accident has occurred.

b. Reengineer conceptual ER diagram of available data source(s):

ER diagram of raw Data Sources:



Proposed diagram of transformed Data Sources:



2. User requirements specification – identifying user information needs:

a. Please identify basic users (types) of the analytical platform (designed data mart).Try to name them and briefly characterise their interest (information-wise).

User: Police Department of Chicago - their interest is to help prevent road accidents. By looking at specific, prepared information, which gives conclusions about what action to take.

User: Car companies - their interest is to improve the safety in a certain car, based on cars reaction(airbag opened/ did not open) , quality of seat belts.

b. Please specify 8-10 OLAP user query types. These should not be specific/detailed queries, but rather general ones, that support decision making(actionable knowledge)– e.g.in order to decide which product groups should be removed from our offer, we analyse the quantity of ordered products from all product groups based on different locations and different time periods.

- find location where we should increase number of police patrol .in order to do that we need identify locations with higher than average number of accidents;
- paying special attention to drivers of a certain age group (with higher than avg nr of accidents for that age group);
- increasing number of police patrol at the day of the week when number of accidents is higher than average;
- identify correlation between no control devices and number of accidents;
- identify correlation between weather conditions and number of injuries;
- show the top 10 causes of crash in a location, where accident number is higher than average;
- show top 10 number of total injuries, when the person was under the influence of alcohol;
- show top 10 number of injuries, different types for different car brands.

c. Perform basic analysis of these query types, try to infer more general user requirements.
i.Identify“events”that the user is interested in–specify the snapshot, identify the type of the snapshot and perspectives.

Event is crash, snapshot is per event. User can be interested in the following perspectives:

1. location (street name and street number);
2. age group;
3. date: day of week;
4. control devices;
5. weather and lighting condition; 6. alcohol consumption;
7. car models.

3. Design dimensional Modelling – use dot modelling technique:

a. Business Process - Select and define analysed business process

- The *business process* concerns traffic crashes in the city of Chicago. The general purpose is to study these accidents in order to understand the major causes and find ways of improving safety.
- *Snapshots* are taken as a result of a crash, either self-reported at the police district by the driver(s) involved or recorded at the scene by the police officer responding to the crash.

b. Grain - Determine the grain of the business process

- The *grain* of the model is represented by a traffic crash, identified by the number of report by Chicago Police Department.

c. Dimensions - Choose the dimensions describing the business process

i. Identify needed attributes ii. Define hierarchies iii.

Define retrospection type and identify change frequency

Dot modelling entities and segments		
Entity name	**Retrospection**	**Frequency**
Date	Permanent	n/a
Metadata The date dimension contains the details of the dates and time used to specify the date and time of a traffic crash.		

Attribute name: date-id	PK Y

Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The identifier for Date. The id is generated by the system.		
Source	**Transformation**	**Data type**
n/a	None	Number
Attribute name: month		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The month associated to the date row.		
Source	**Transformation**	**Data type**
n/a	None	Number
Attribute name: year		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None

Metadata The year associated to the date row.		
Source	**Transformation**	**Data type**
n/a	None	Number
Attribute name: dayOfWeek		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The day associated to the date row.		
Source	**Transformation**	**Data type**
n/a	None	Number
Attribute name: Quarter		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The quarter of the year associated to the date row.		
Source	**Transformation**	**Data type**

n/a	None	Number
Attribute name: fullDate		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The full date associated to the row.		
Source	**Transformation**	**Data type**
n/a	None	Date
Attribute name: Time		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The time associated to the date row.		
Source	**Transformation**	**Data type**
n/a	None	Time
Attribute name: DateTime		PK N
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None

Metadata The full date associated to the row, with both time and date information.		
Source	**Transformation**	**Data type**
n/a	None	Date & Time

Dot modelling entities and segments		
Entity name	**Retrospection**	**Frequency**
Location	False	Monthly
Metadata The location dimension identifies the location of a traffic crash, specifying details about the street and the exact coordinates of the location.		
Attribute name: latitude		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The latitude associated to the exact position of the crash.		

Source	**Transformation**	**Data type**
Crashes table	None	Number
Attribute name: longitude		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The longitude associated to the exact position of the crash.		
Source	**Transformation**	**Data type**
Crashes table	None	Number
Attribute name: location		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The exact position of the crash as a point on a map.		
Source	**Transformation**	**Data type**
Crashes table	None	Point on a map
Attribute name: street_name		PK N

Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The name of the street where the crash occurred.		
Source	**Transformation**	**Data type**
Crashes table	None	Text
Attribute name: street_number		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The numerical address number of the street where the crash occurred.		
Source	**Transformation**	**Data type**
Crashes table	None	Number
Attribute name: beat_of_occurrence		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None

Metadata The Chicago Police Department Beat ID associated to the street address where the crash occurred.		
Source	**Transformation**	**Data type**
Crashes table	None	Number
Attribute name: location_id		PK Y
Retrospection	**Frequency**	**Dependency**
Permanent	n/a	None
Metadata The numerical address number of the street where the crash occurred.		
Source	**Transformation**	**Data type**
Crashes table	None	Number

Dot modelling entities and segments		
Entity name	**Retrospection**	**Frequency**
Environment	False	Monthly

Metadata The environment dimension specifies the environmental condition for a crash. It contains information about the condition of the weather, lighting, road and the presence and condition of any traffic control device.		
Attribute name: env_no		PK Y
Retrospection	**Frequency**	**Dependency**
Permanent	Monthly	None
Metadata The identifier for the rows of the table.		
Source	**Transformation**	**Data type**
Crashes table	None	Number
Attribute name: traffic_control_device		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The traffic control device present at crash location, i.e. a stop sign, traffic light…		
Source	**Transformation**	**Data type**
Crashes table	None	Text

Attribute name: device_condition	PK N

Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The condition of the traffic control device present at crash location at the time of the crash.		
Source	**Transformation**	**Data type**
Crashes table	None	Text

Attribute name: weather_condition		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The weather condition at the time of the crash.		
Source	**Transformation**	**Data type**
Crashes table	None	Text

Attribute name: lighting_condition		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None

Metadata The lighting condition at the time of the crash.		
Source	**Transformation**	**Data type**
Crashes table	None	Text
Attribute name: road_defect		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata Information on condition of the road at the time of the crash, i.e. if the road was damaged or presented any defect.		
Source	**Transformation**	**Data type**
Crashes table	None	Text
Attribute name: alignment		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The street alignment at the location of the crash.		
Source	**Transformation**	**Data type**

Crashes table	None	Text
Attribute name: roadway_surface_cond		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The surface condition of the road at the time of the crash, i.e. if the road was dry, wet or in any other condition that could affect drivers.		
Source	**Transformation**	**Data type**
Crashes table	None	Text

Dot modelling entities and segments		
Entity name	**Retrospection**	**Frequency**
Vehicle	True	Monthly
Metadata The vehicle dimension contains information about the units involved in traffic crashes. “Vehicles” can be motor vehicles and non-motor vehicles modes of transportation, such as bicycles and pedestrians.		
Attribute name: crash_unit_id		PK Y
Retrospection	**Frequency**	**Dependency**

Permanent	Monthly	None
Metadata The identifier of the vehicle involved in the crash.		
Source	**Transformation**	**Data type**
Vehicles table	None	Number
Attribute name: maneuver		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The action the unit was taking prior to the crash.		
Source	**Transformation**	**Data type**
Vehicles table	None	Text
Attribute name: unit_type		PK N

Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The type of unit. Can be a driven vehicle, a parked vehicle, a pedestrian, bicycle or others.		

Source	Transformation	Data type
Vehicles table	None	Text
Attribute name: occupant_cnt		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The number of people associated to the unit. For example, the number of passengers on a car at the moment of the crash.		
Source	**Transformation**	**Data type**
Vehicles table	None	Number
Attribute name: make		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The brand of the vehicle, if applicable.		
Source	**Transformation**	**Data type**
Vehicles table	None	Text
Attribute name: model		PK N

Retrospection	Frequency	Dependency
False	Monthly	None
Metadata The model of the vehicle, if applicable.		
Source	**Transformation**	**Data type**
Vehicles table	None	Text
Attribute name: vehicle_year		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The model year of the vehicle, if applicable.		
Source	**Transformation**	**Data type**
Vehicles table	None	Number

Dot modelling entities and segments		
Entity name	**Retrospection**	**Frequency**
People	True	Monthly

Metadata The People dimension contains information about people involved in a crash and if any injuries were sustained.		
Attribute name: person_id		PK Y
Retrospection	**Frequency**	**Dependency**
Permanent	Monthly	None
Metadata The identifier of the single person involved in the crash.		
Source	**Transformation**	**Data type**
People table	None	Number
Attribute name: person_type		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata Type of roadway user involved in the crash (driver, passenger, pedestrian…).		
Source	**Transformation**	**Data type**
People table	None	Text
Attribute name: sex		PK N

Retrospection	Frequency	Dependency

False	Monthly	None
Metadata The gender of the person.		
Source	**Transformation**	**Data type**
People table	None	Text
Attribute name: age		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata The age of the person at the time of the crash.		
Source	**Transformation**	**Data type**
People table	None	Number
Attribute name: airbag_deployed		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None

Metadata Whether vehicle occupant airbag deployed as a result of the crash.		
Source	**Transformation**	**Data type**
People table	None	Text
Attribute name: safety_equipment		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None

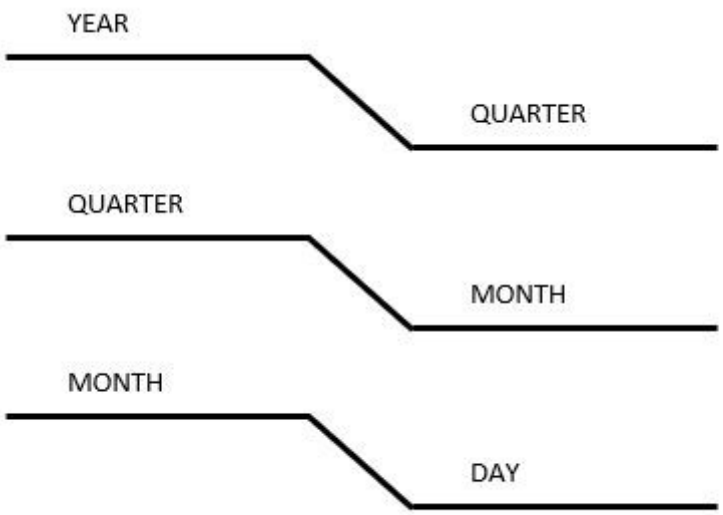
Metadata Safety equipment used by vehicle occupant in crash, if any.		
Source	**Transformation**	**Data type**
People table	None	Text
Attribute name: injury_classification		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata Severity of injury person sustained in the crash. It goes from no injury to fatal injury.		
Source	**Transformation**	**Data type**

People table	None	Text
Attribute name: physical_condition		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata Driver's apparent physical condition at the time of the crash.		
Source	**Transformation**	**Data type**
People table	None	Text
Attribute name: cell_phone_use		PK N
Retrospection	**Frequency**	**Dependency**
False	Monthly	None
Metadata Whether person was/was not using a cellphone at the time of the crash.		
Source	**Transformation**	**Data type**
People table	None	Boolean

Dot modelling hierarchies and groupings

Model name: Traffic crashes in Chicago

Dimension name: Date

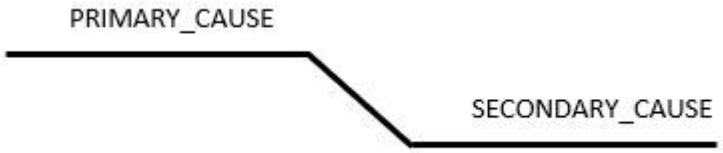


Metadata
Time hierarchy: year is composed of 4 quarters, which are composed of 3 months each, which are composed of 28-31 days.

Dot modelling hierarchies and groupings

Model name: Traffic crashes in Chicago

Dimension name: Crash (junk dimension)

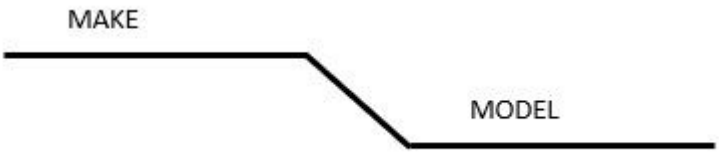


Metadata
Each primary cause for a traffic crash can have a secondary cause.

Dot modelling hierarchies and groupings

Model name: Traffic crashes in Chicago

Dimension name: Vehicle



Metadata
Make identifies the brand of a car. Models of a car are associated to a unique brand.

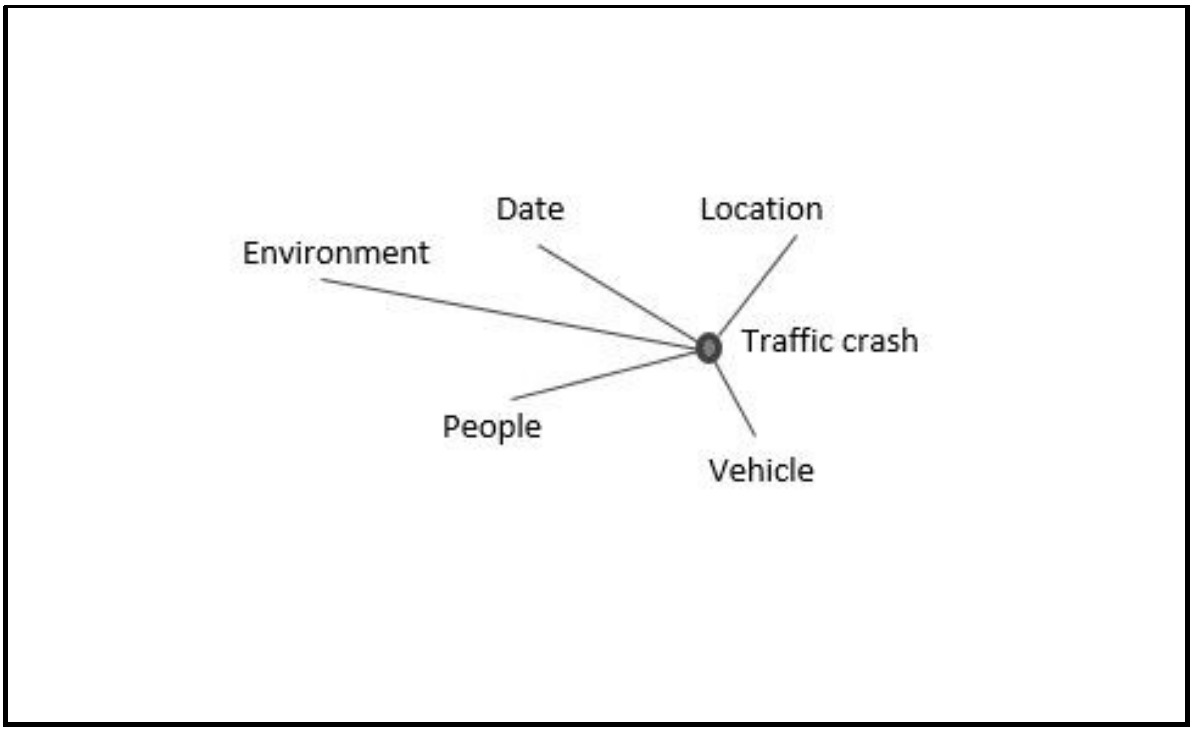
d. Facts

i. Identify the numeric measures for the facts ii.

Define aggregate functions for the measures

Dot modelling data model

Model name: Traffic crashes in Chicago	
Measurable facts	
Fact name	**Metadata**
Num_units	The total number of units involved in the traffic crash
Injuries_total	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries
Injuries_fatal	Total persons sustaining fatal injuries in the crash
Injuries_incapacitating	Total persons sustaining incapacitating/serious injuries in the crash. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they could perform before the injury occurred
Injuries_non_incapacitating	Total persons sustaining non-incapacitating injuries in the crash. Any injury, other than fatal or incapacitating injury, which is evident to observers at the scene of the crash
Injuries_reported_non_evident	Total persons sustaining possible injuries in the crash
No_of_people_involved	The total number of people involved in the traffic crash, both injured and non-injured ones
Diagram	



Dot modelling fact usage					
Model name: Traffic crashes in Chicago					
Fact name: Num_units					
Dimension	**Sum**	**Count**	**Avg**	**Min**	**Max**
Date	V	X	V	V	V
Location	V	X	V	V	V
People	V	X	V	V	V
Environment	V	X	V	V	V
Vehicle	V	X	V	V	V

Dot modelling fact usage
Model name: Traffic crashes in Chicago
Fact name: Injuries_total

Dimension	Sum	Count	Avg	Min	Max
Date	V	V	V	V	V
Location	V	V	V	V	V
People	V	V	V	V	V
Environment	V	V	V	V	V
Vehicle	V	V	V	V	V

Dot modelling fact usage					
Model name: Traffic crashes in Chicago					
Fact name: Injuries_fatal					
Dimension	**Sum**	**Count**	**Avg**	**Min**	**Max**
Date	V	V	V	V	V
Location	V	V	V	V	V
People	V	V	V	V	V
Environment	V	V	V	V	V
Vehicle	V	V	V	V	V

Dot modelling fact usage					
Model name: Traffic crashes in Chicago					
Fact name: Injuries_ incapacitating					
Dimension	**Sum**	**Count**	**Avg**	**Min**	**Max**
Date	V	V	V	V	V
Location	V	V	V	V	V
People	V	V	V	V	V

Environment	V	V	V	V	V
Vehicle	V	V	V	V	V

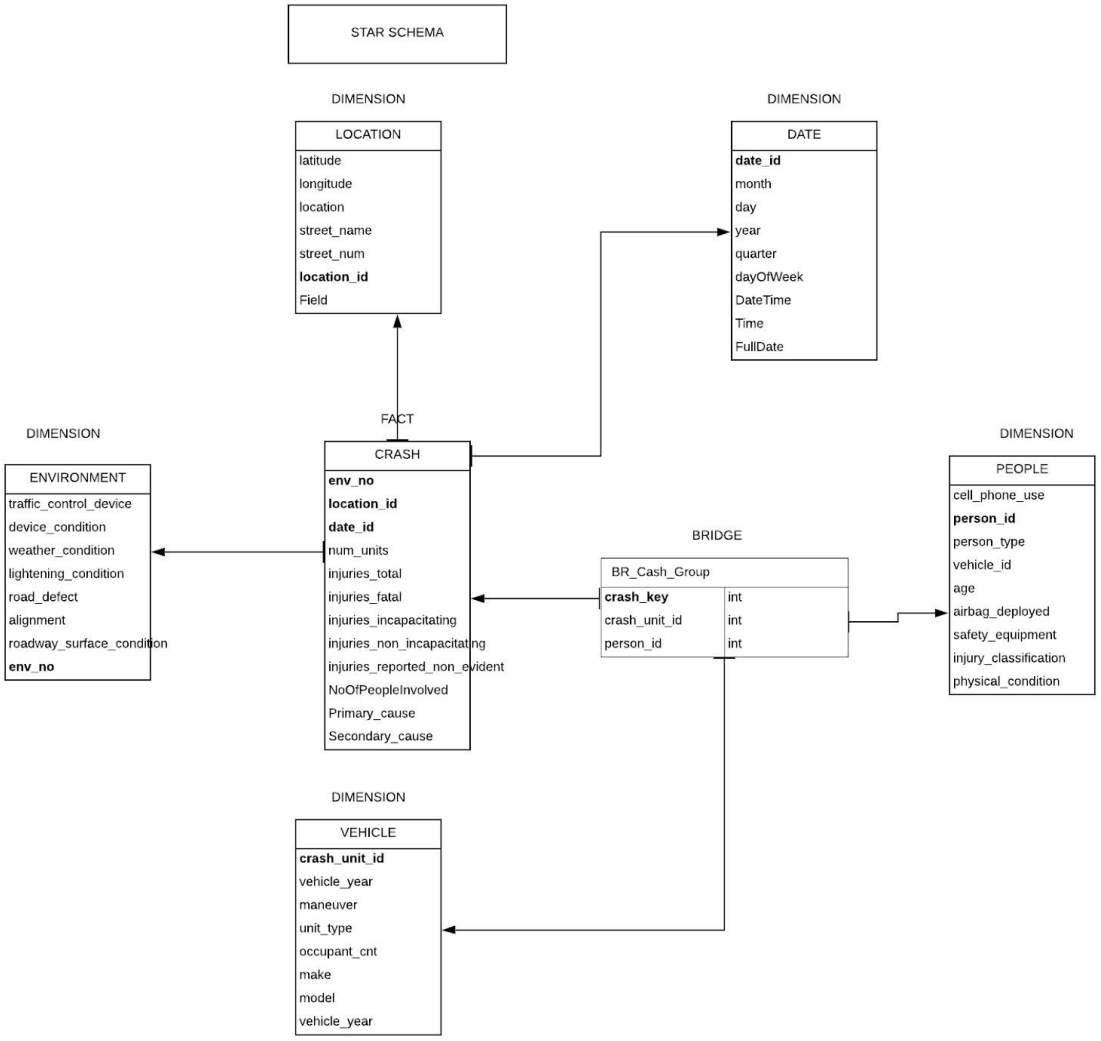
Dot modelling fact usage					
Model name: Traffic crashes in Chicago					
Fact name: Injuries_non_ incapacitating					
Dimension	**Sum**	**Count**	**Avg**	**Min**	**Max**
Date	V	V	V	V	V
Location	V	V	V	V	V
People	V	V	V	V	V
Environment	V	V	V	V	V
Vehicle	V	V	V	V	V

Dot modelling fact usage					
Model name: Traffic crashes in Chicago					
Fact name: Injuries_reported_non_evident					
Dimension	**Sum**	**Count**	**Avg**	**Min**	**Max**
Date	V	V	V	V	V
Location	V	V	V	V	V
People	V	V	V	V	V
Environment	V	V	V	V	V
Vehicle	V	V	V	V	V

Dot modelling fact usage

Model name: Traffic crashes in Chicago					
Fact name: No_of_people_involved					
Dimension	**Sum**	**Count**	**Avg**	**Min**	**Max**
Date	V	X	V	V	V
Location	V	X	V	V	V
People	V	X	V	V	V
Environment	V	X	V	V	V
Vehicle	V	X	V	V	V

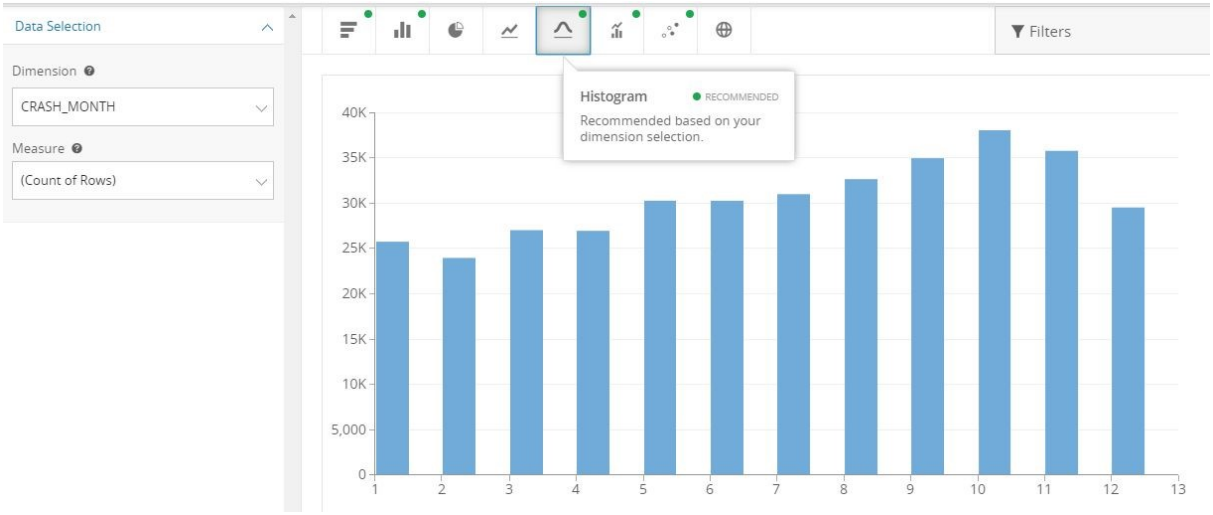
e. Proposed design schema for the multi-dimensional model (star, snowflake or fact constellation)

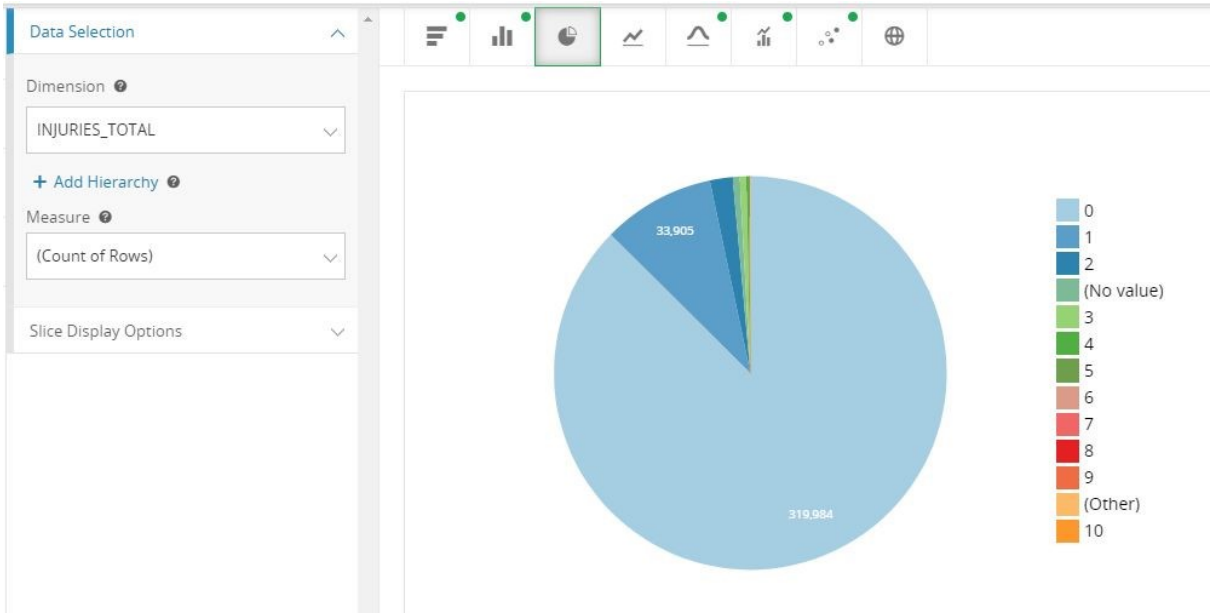


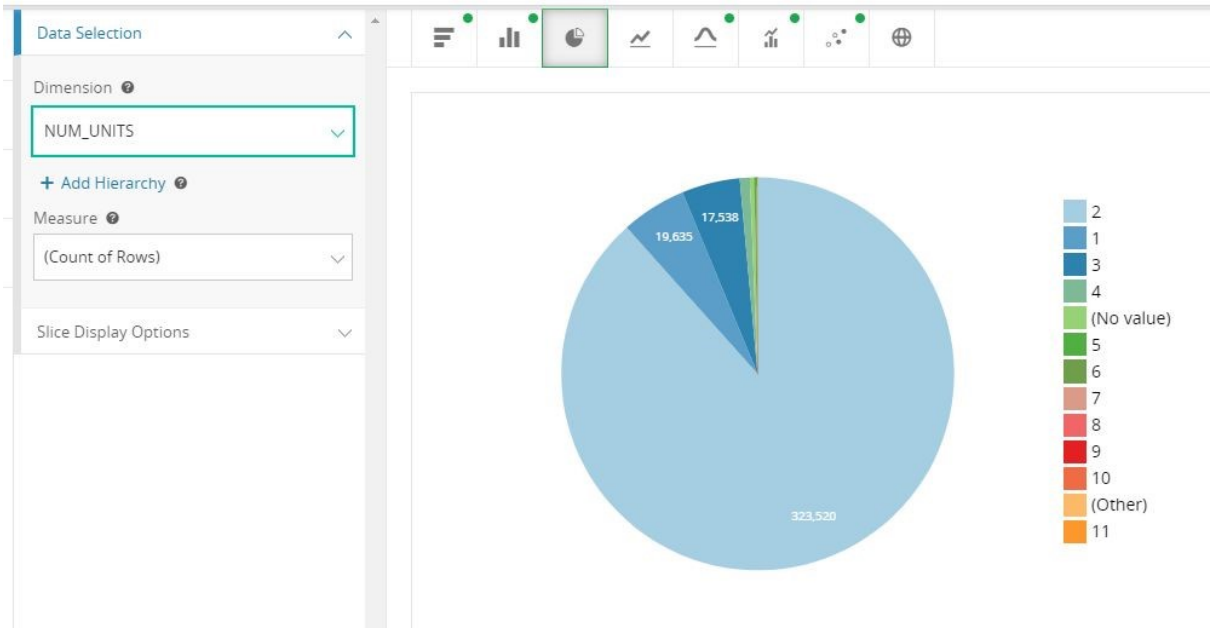
Table_Name	Column_Name	Datatype	Table_Type	SCD	Database_Name	STable_Name	SColumn_Name	SDatatype	Transformation
Location	latitude	float	Dimension	1	city of chicago	Crashes	latitude	float	-
Location	longitude	float	Dimension	1	city of chicago	Crashes	longitude	float	-
Location	location	varchar(255)	Dimension	1	city of chicago	Crashes	location	varchar(255)	-
Location	street_name	varchar(255)	Dimension	0	city of chicago	Crashes	street_name	varchar(255)	-
Location	street_num	int	Dimension	0	city of chicago	Crashes	street_num	int	-
Location	Beat_of_occurance	varchar(255)	Dimension	1	city of chicago	Crashes	Beat_of_occurance	varchar(255)	-
Location	location_id	int	Dimension	1	city of chicago	Crashes	latitude,longitude	float	Generate
Date	date_id	int	Dimension	1	city of chicago	Crashes	crash_date,hour	int	date_id AS CONVERT(CHAR(8), [crash_date], 112),
Date	month	int	Dimension	1	city of chicago	Crashes	crash_month	int	-
Date	day	int	Dimension	1	city of chicago	Crashes	crash_date	int	day(crash_date)
Date	year	int	Dimension	1	city of chicago	Crashes	year	int	-
Date	fullDate	date	Dimension	1	city of chicago	Crashes	crash_date	date	dd-mm-yyyy(crash_date)
Date	time	time	Dimension	1	city of chicago	Crashes	crash_date	time	hh-mm-ss(crash_date)
Date	datetime	timestamp	Dimension	1	city of chicago	Crashes	crash_date	timestamp	Natural Key
Date	dayOfWeek	varchar(50)	Dimension	1	city of chicago	Crashes	dayOfWeek	varchar(50)	-
Date	quarter	int	Dimension	1	city of chicago	Crashes	crash_date	int	DATEPART(QUARTER, [fullDate]),
People	person_id	int	Dimension	2	city of chicago	People	person_id	int	Primary key - People
People	person_type	varchar(50)	Dimension	1	city of chicago	People	person_type	varchar(50)	-
People	vehicle_id	int	Dimension	1	city of chicago	People	vehicle_id	int	-
People	sex	varchar(50)	Dimension	1	city of chicago	People	sex	varchar(50)	-
People	age	int	Dimension	1	city of chicago	People	age	int	-
People	airbag_deployed	varchar(255)	Dimension	1	city of chicago	People	airbag_deployed	varchar(255)	-
People	safety_equipment	varchar(255)	Dimension	1	city of chicago	People	safety_equipment	varchar(255)	-
People	injury_classification	varchar(255)	Dimension	1	city of chicago	People	injury_classification	varchar(255)	-
People	Physical_condition	varchar(255)	Dimension	1	city of chicago	People	Physical_condition	varchar(255)	-
People	cellphone_use	varchar(255)	Dimension	1	city of chicago	People	cellphone_use	varchar(255)	-
Vehicle	crash_unit_ID	int	Dimension	2	city of chicago	Vehicles	crash_unitID	int	Primary key - Vehicle
Vehicle	manevour	varchar(255)	Dimension	0	city of chicago	Vehicles	manevour	varchar(255)	-
Vehicle	unit_type	varchar(255)	Dimension	1	city of chicago	Vehicles	unit_type	varchar(255)	-
Vehicle	occupant_cnt	int	Dimension	1	city of chicago	Vehicles	occupant_cnt	int	-
Vehicle	make	varchar(255)	Dimension	0	city of chicago	Vehicles	make	varchar(255)	-
Vehicle	model	varchar(255)	Dimension	0	city of chicago	Vehicles	model	varchar(255)	-
Vehicle	vehicle_year	int	Dimension	0	city of chicago	Vehicles	vehicle_year	int	-
Environment	traffic_control_devi	varchar(255)	Dimension	1	city of chicago	Crashes	traffic_control_devic	varchar(255)	-
Environment	device_condition	varchar(255)	Dimension	1	city of chicago	Crashes	device_condition	varchar(255)	-
Environment	weather_condition	varchar(255)	Dimension	1	city of chicago	Crashes	weather_condition	varchar(255)	-
Environment	lightening_condition	varchar(255)	Dimension	1	city of chicago	Crashes	lightening_condition	varchar(255)	-
Environment	road_deffect	varchar(255)	Dimension	1	city of chicago	Crashes	road_deffect	varchar(255)	-
Environment	alignment	varchar(255)	Dimension	1	city of chicago	Crashes	alignment	varchar(255)	-
Environment	roadway_surface_c	varchar(255)	Dimension	1	city of chicago	Crashes	roadway_surface_co	varchar(255)	-
Environment	env_no	int	Dimension	1	city of chicago	-	-	-	Primary key - Environent, random generatio
Crash	env_no	int	Fact	1	city of chicago	-	-	-	Foreign Key - Environment
Crash	location_id	int	Fact	n/a	city of chicago	Crashes	location_id	int	Foriegn Key - Location
Crash	date_id	int	Fact	n/a	city of chicago	Crashes	date_id	int	Foreign Key - Date
Crash	num_units	int	Fact	n/a	city of chicago	Crashes	num_units	int	-
Crash	injuries_total	int	Fact	n/a	city of chicago	Crashes	injuries_total	int	-
Crash	injuries_fatal	int	Fact	n/a	city of chicago	Crashes	injuries_fatal	int	-
Crash	injuries_incapacitati	int	Fact	n/a	city of chicago	Crashes	injuries_incapacitatin	int	-
Crash	injuries_non_incapa	int	Fact	n/a	city of chicago	Crashes	injuries_non_incapac	int	-
Crash	injuries_reported_n	int	Fact	n/a	city of chicago	Crashes	injuries_reported_no	int	-
Crash	numberOfPeopleInv	int	Fact	n/a	city of chicago	Crashes	injuries_total, injuries	int	Combine injuries_total, injuries_fatal
Crash	primaryCause	varchar(255)	Fact	n/a	city of chicago	Crashes	primaryCause	varchar(255)	-
Crash	SecondaryCause	varchar(255)	Fact	n/a	city of chicago	Crashes	SecondaryCause	varchar(255)	-
BR_Crash_group	crash_key	int	Bridge	n/a	city of chicago	-	-	-	Surrogate Key
BR_Crash_group	crash_unit_ID	int	Bridge	n/a	city of chicago	Vehicles	crash_unit_ID	int	-
BR_Crash_group	person_id	int	Bridge	n/a	city of chicago	People	person_id	int	-

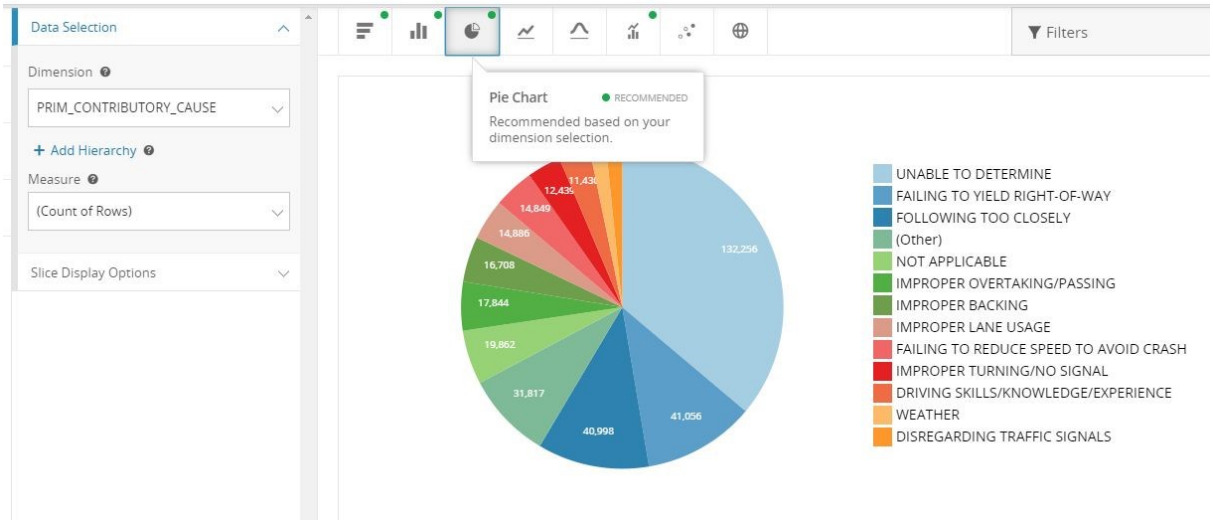
4. Create logical data map of the ETL process-

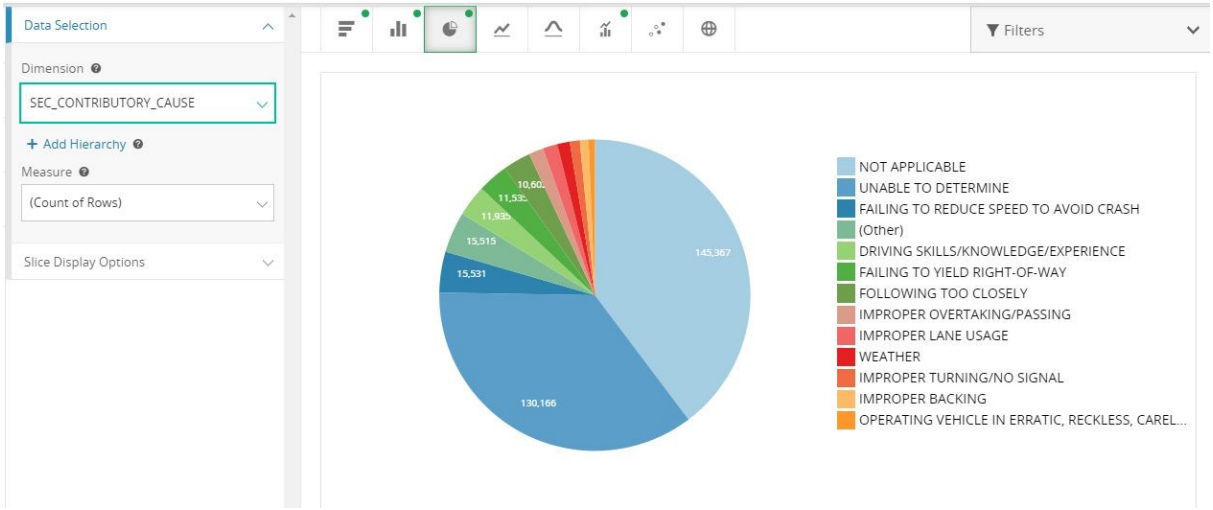
5. Perform basic data profiling of the data source

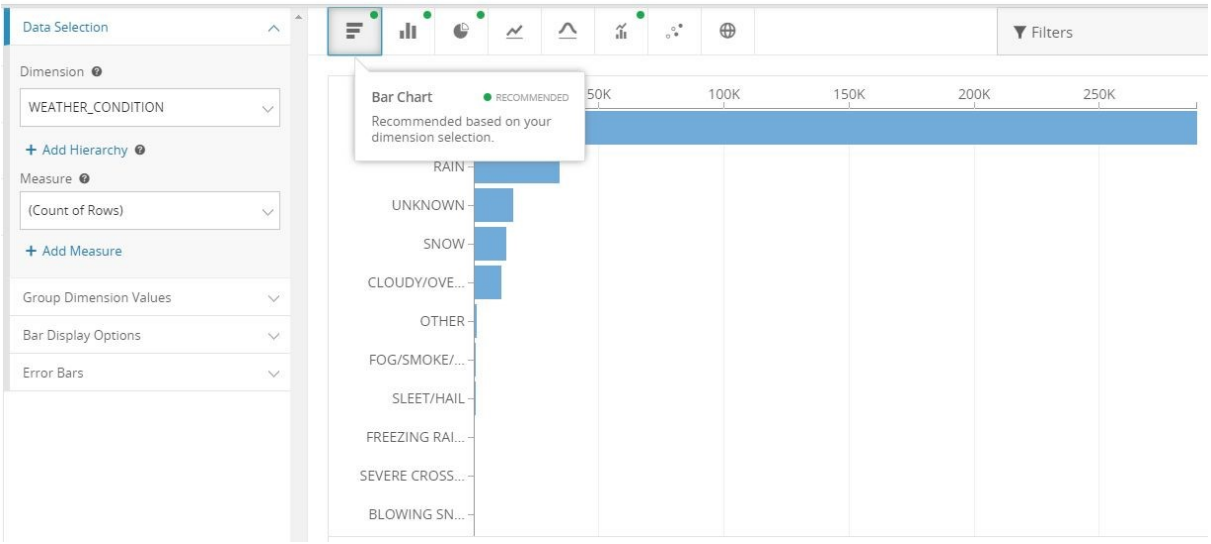


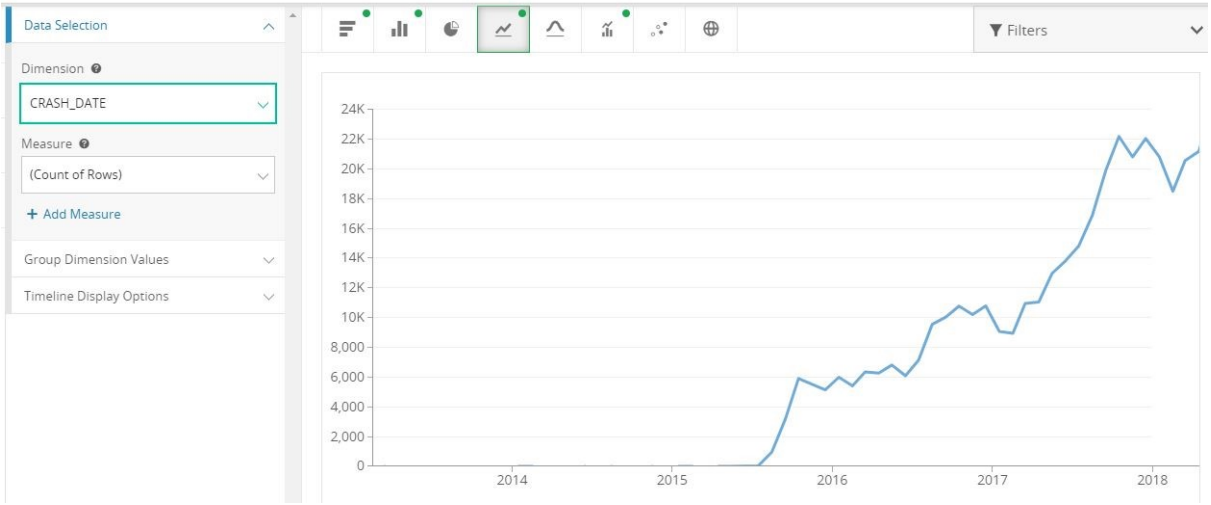


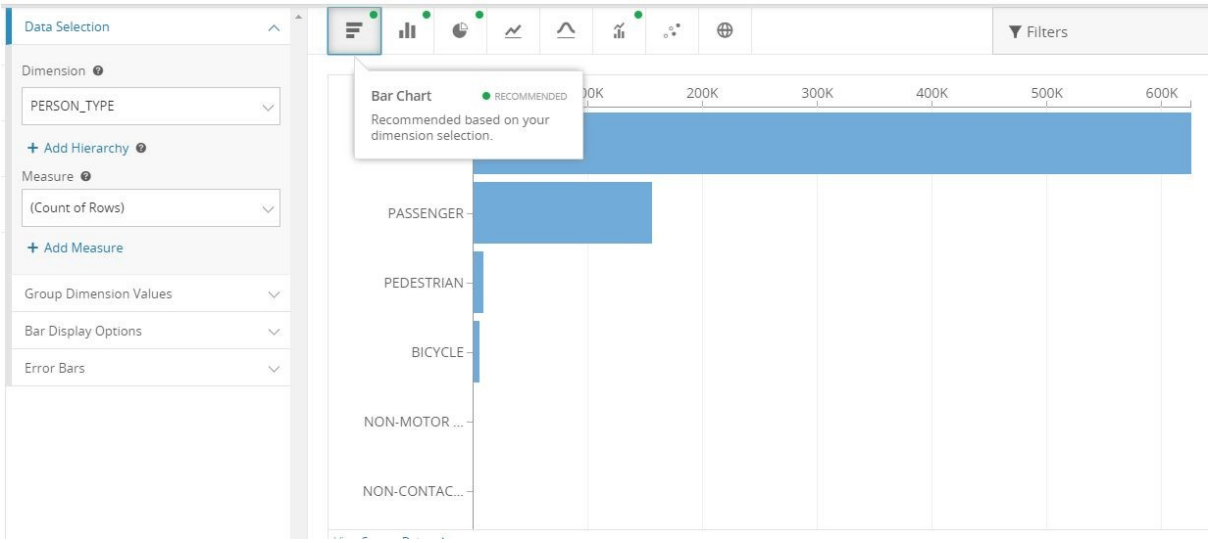

Data Selection
Dimension
NUM_UNITS
+ Add Hierarchy
Measure
(Count of Rows)
Slice Display Options
17,538
19,635
323,520
2
1
3
4
(No value)
5
6
7
8
9
10
(Other)
11

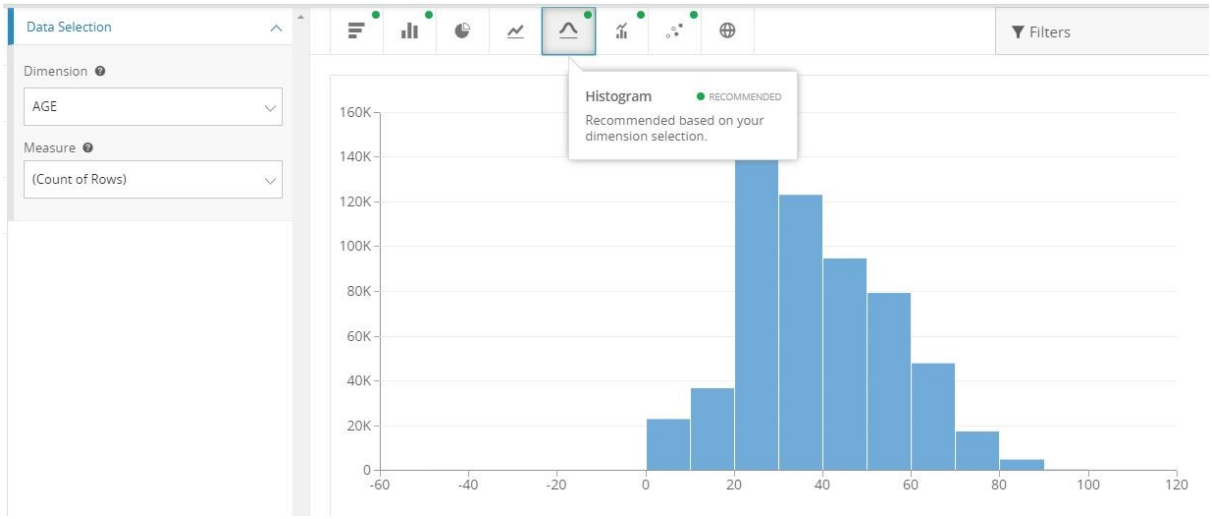

Data Selection
Dimension
PRIM_CONTRIBUTORY_CAUSE
+ Add Hierarchy
Measure
(Count of Rows)
Slice Display Options
Filters
Pie Chart
RECOMMENDED
Recommended based on your dimension selection.
132,256
41,056
40,998
31,817
19,862
17,844
16,708
14,886
14,849
12,435
UNABLE TO DETERMINE
FAILING TO YIELD RIGHT-OF-WAY
FOLLOWING TOO CLOSELY
(Other)
NOT APPLICABLE
IMPROPER OVERTAKING/PASSING
IMPROPER BACKING
IMPROPER LANE USAGE
FAILING TO REDUCE SPEED TO AVOID CRASH
IMPROPER TURNING/NO SIGNAL
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE
WEATHER
DISREGARDING TRAFFIC SIGNALS

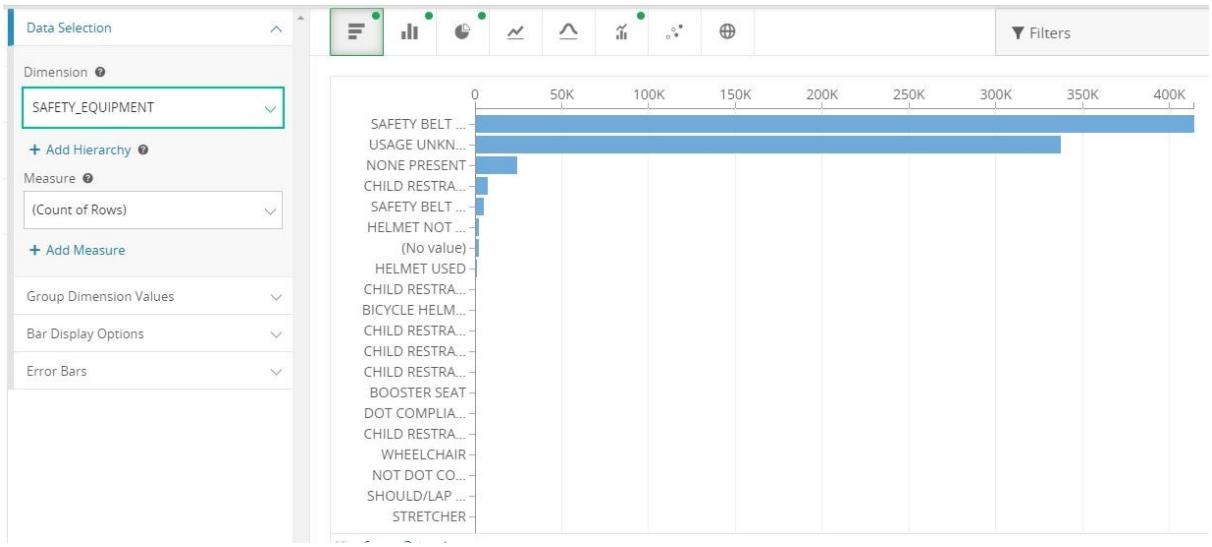

Data Selection
Dimension
SEC_CONTRIBUTORY_CAUSE
+ Add Hierarchy
Measure
(Count of Rows)
Slice Display Options
Filters
145,367
130,166
15,531
15,515
NOT APPLICABLE
UNABLE TO DETERMINE
FAILING TO REDUCE SPEED TO AVOID CRASH
(Other)
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE
FAILING TO YIELD RIGHT-OF-WAY
FOLLOWING TOO CLOSELY
IMPROPER OVERTAKING/PASSING
IMPROPER LANE USAGE
WEATHER
IMPROPER TURNING/NO SIGNAL
IMPROPER BACKING
OPERATING VEHICLE IN ERRATIC, RECKLESS, CAREL...

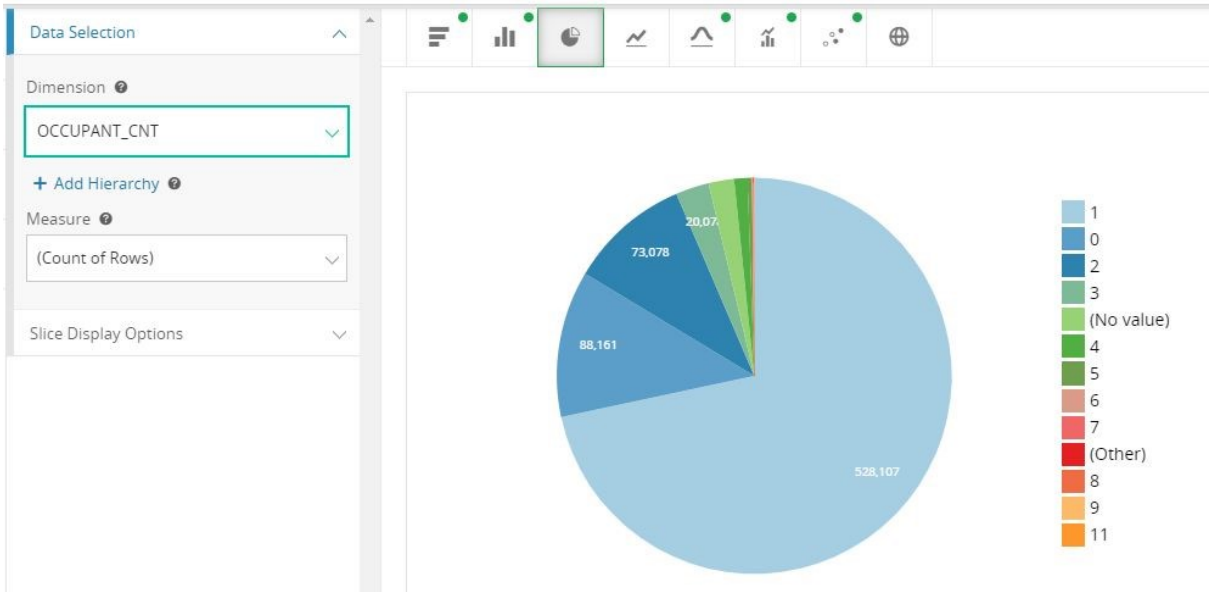

Data Selection
Dimension
WEATHER_CONDITION
+ Add Hierarchy
Measure
(Count of Rows)
+ Add Measure
Group Dimension Values
Bar Display Options
Error Bars
Filters
Bar Chart
RECOMMENDED
Recommended based on your dimension selection.
50K
100K
150K
200K
250K
RAIN
UNKNOWN
SNOW
CLOUDY/OVE...
OTHER
FOG/SMOKE/...
SLEET/HAIL
FREEZING RAI...
SEVERE CROSS...
BLOWING SN...

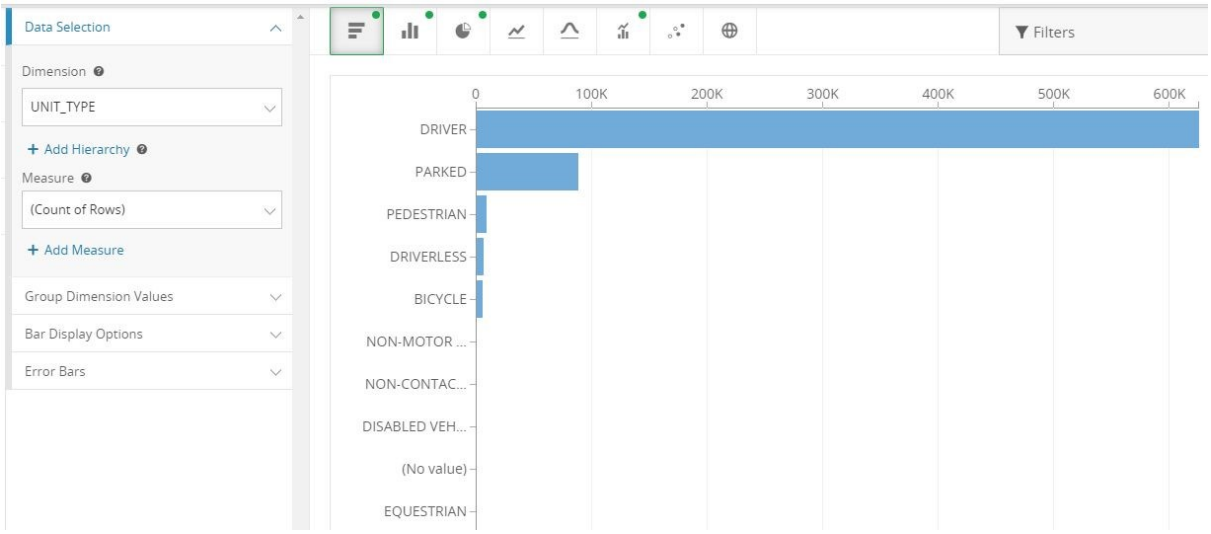

Data Selection
Dimension
CRASH_DATE
Measure
(Count of Rows)
+ Add Measure
Group Dimension Values
Timeline Display Options
Filters
24K
22K
20K
18K
16K
14K
12K
10K
8,000
6,000
4,000
2,000
0
2014
2015
2016
2017
2018

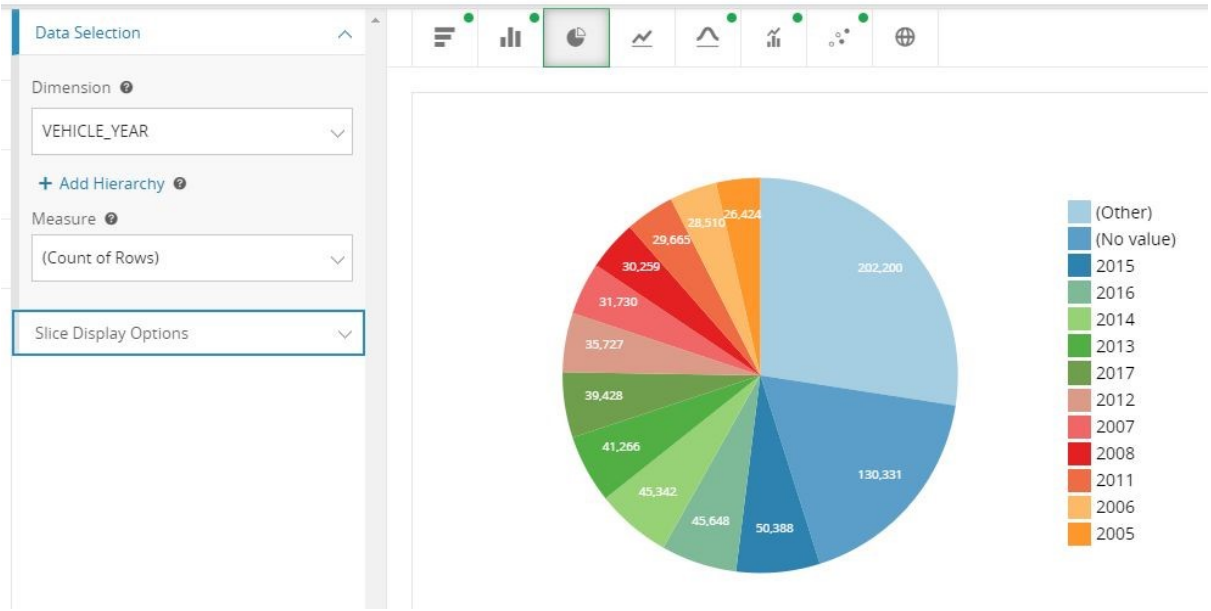

Data Selection
Dimension
PERSON_TYPE
+ Add Hierarchy
Measure
(Count of Rows)
+ Add Measure
Group Dimension Values
Bar Display Options
Error Bars
Filters
Bar Chart
RECOMMENDED
Recommended based on your dimension selection.
200K
300K
400K
500K
600K
PASSENGER
PEDESTRIAN
BICYCLE
NON-MOTOR ...
NON-CONTAC...


Data Selection
Dimension
AGE
Measure
(Count of Rows)
Filters
Histogram
RECOMMENDED
Recommended based on your dimension selection.
160K
140K
120K
100K
80K
60K
40K
20K
0
-60
-40
-20
0
20
40
60
80
100
120


Data Selection
Dimension
SAFETY_EQUIPMENT
+ Add Hierarchy
Measure
(Count of Rows)
+ Add Measure
Group Dimension Values
Bar Display Options
Error Bars
Filters
0
50K
100K
150K
200K
250K
300K
350K
400K
SAFETY BELT ...
USAGE UNKN...
NONE PRESENT
CHILD RESTRA...
SAFETY BELT ...
HELMET NOT ...
(No value)
HELMET USED
CHILD RESTRA...
BICYCLE HELM...
CHILD RESTRA...
CHILD RESTRA...
CHILD RESTRA...
BOOSTER SEAT
DOT COMPLIA...
CHILD RESTRA...
WHEELCHAIR
NOT DOT CO...
SHOULD/LAP ...
STRETCHER


Data Selection
Dimension
OCCUPANT_CNT
+ Add Hierarchy
Measure
(Count of Rows)
Slice Display Options
73,078
88,161
528,107
1
0
2
3
(No value)
4
5
6
7
(Other)
8
9
11





ADDITIONAL TASK

We are using the update-driven approach to perform further operations on our dataset. In this way, the information from multiple heterogeneous sources are integrated in advance and will be stored stored in a warehouse. This information will be available for direct querying and analysis.
According to this approach, we are firstly performing Data Extraction, Data Cleaning, and Data Transformation.

After we load our initial Data involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions, we will perform refreshing which involves updating from data sources to warehouse.

We will be updating vehicles and people dimensions as there can be new crash_unit_id and person_id as new people and vehicles can encounter with the crashes and thus be updated monthly into the respective databases.

As we update the vehicle and people databases, connections in the bridge table called BR_crash_group will also have to get updated.