UNIVERSIDAD
DE MÁLAGA

Advanced Analytics on Big Data

Task 7: Big Data Warehouse

Authors: Juan Morales Conde

Martín Bláquez Moreno

Juan Rafael Caro Romero

Module: Data Analytics

Málaga, 13^{th} January 2020

The matter of this work is to extend the previous document done in the Task 6. As explained, we supposed that we work for an insurance company and want to know the significant data of the collisions between vehicles in the city of New York and their reasons in order to adjust the price of our products. Once we know the data, is time to propose a design for the data warehouse capable of store and shown, in a simple way, all the information required.

To accomplish our objective, we need to create a multidimensional model. This model is made up of fact(s), dimension(s) and measure(s). In our case, the fact of the multidimensional model is the **Collision** due to the purpose of our study is to analyze the number and causes of the crashes in the city. As we observed in the data set, we decide to create five dimensions:

- Date: to know the exact time of the collision. This attribute include different levels of hierarchy.
- Location: to know the coordinates and the place of the accident. This attribute include different levels of hierarchy.
- Vehicle: this field store all the characteristics of the vehicles involved in the accident: type, make, model, year, damage...
- Cause: factors contributing of the collision.
- People involved: this field collect the data of all people involved in the accident: drivers, passengers, pedestrians...

The measures represent the variables involved that change in the time and they are accumulated. In our study, we are interested in the number of people killed, people injured and the number of collisions. This allow us to relation this measures with the different dimensions and attributes.

This model can be represented by a star schema and logical model.

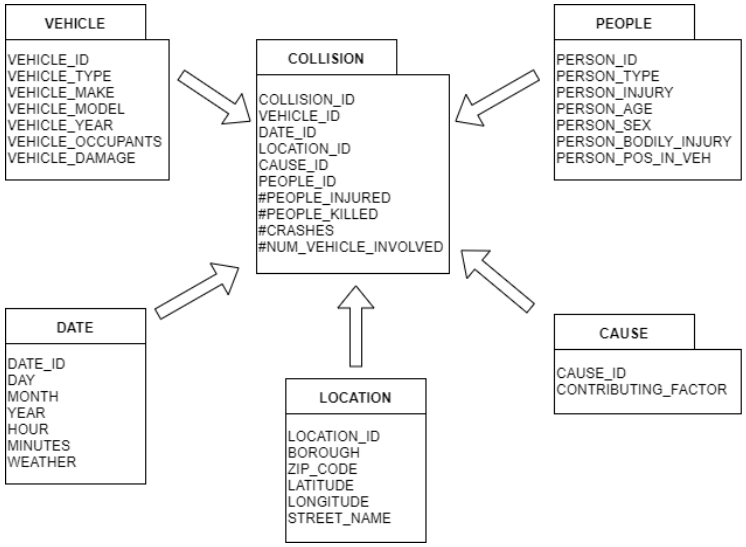

Figura 1: Star schema

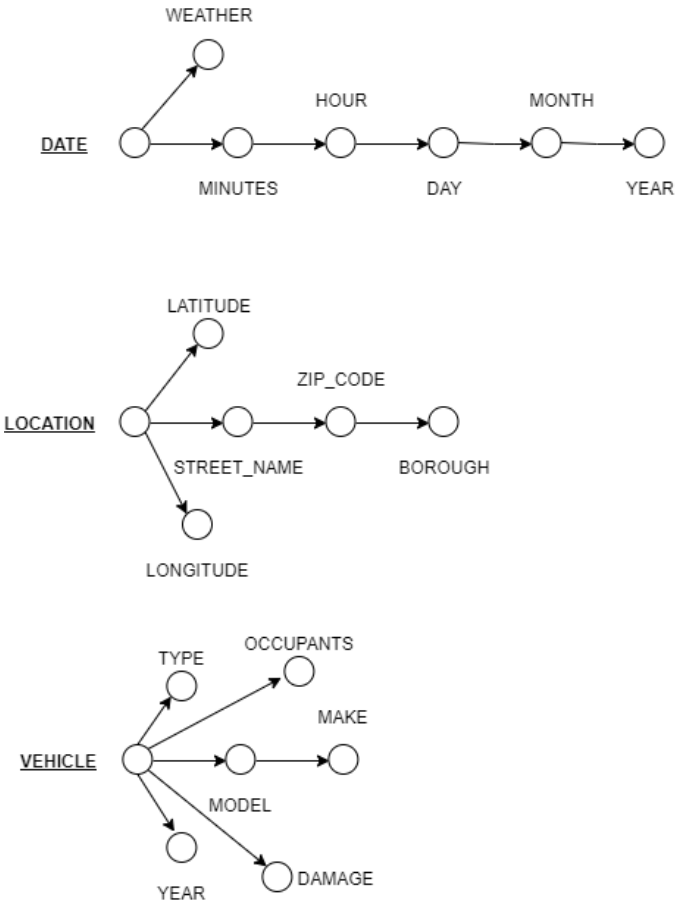

Figura 2: Logical schema

To implement this data warehouse, the first step will be to create a relational database creating the fact and the dimensions as tables with the purpose of connecting with the datawarehouse. After that, we need to modify and load the data into this tables. We set up all the dimensions as a multidimensional schema and, finally, publish the XML file to analyze in the server.