**Area:** American iTunes TOP sales 2013-2016 representing popularity of songs among Americans within specified timeframe.

**Sources of the dataset:** https://kworb.net/pop/archive/ , https://www.riaa.com/ , en.wikipedia.org.

**Description of the dataset:** provided dataset contains information about songs, their authors, sales, certifications and other parameters which can be used to analyze data and support further making decisions.

**Problems appeared while loading data:** the first stage ended up with no problems due to availability of data on the Internet (see *Sources of the dataset*). As all data were being extracted from multiple sources with different logics of storing and representation, data cleansing & transform was a way more challenging. It took much time to consolidate data and make them legible. The data was loadedvia DDL command – *INSERT*.

**Description of data model:** the whole model consists of 7 tables – 1 fact and 4 dimensions (2 out of 4 illustrated by 2 tables => 7) and represented as a ***star***:

1. *‘songs’* table contains information about each song appeared in TOP-10 within specified timeframe. It includes all aspects relevant for making analysis (*length, genre, authors, date of release, etc.*). It is commonplace for songs to be upon impact of these factors while being chosen by audience;
2. *‘labels’* – another table for the first dimension. Any song is released with support of a label. It can have a profound impact on promo, radio rotations, even just name of a label gives great support;
3. *‘performers’* – whatever information about people performed selling songs. In current project, only names are provided. Much as normalization is required, division by first name and last name would be senseless, as a name of performer is a brand to some extent.

Americans are both into national values and equality in any instance, at which point race, gender and some other attributes are considered;

1. *‘countries’* – one more table connected to ‘performers’ to build up a dimension. As said above, it is customary for those people to consider subtle aspects of human being, and origin as well;
2. *‘promo’* table illustrates events used by performers to present their songs to great quantity of people. For better usage, only huge actions (awards, festivals) can be provided. This table may well be viable for deep inspection of interdependencies. By and large, promos have indirect impact on sales, so in current project this table is empty and is not connected to fact table;
3. *‘times’* table includes information about each time point used for sales tracking. For study purposes, only information about number of week, month and half year is illustrated;
4. *‘sales’* – the only fact table in the schema. It includes link to dimension tables and some kind of additional data: numbers of copies sold within previous week, possible certifications, price (as a rule, there are some fixed values but rarely songs can be under discount), estimate income (in terms of iTunes it is not relevant, as performers get almost nothing for sales. Sold copies are crucial element of this area).

**Additional information:** certification is sort of admitting that a song has succeeded. There are several key points for getting this document. For USA they are: 500k, 1..10M. These values can vary in different countries. The most challenging issue is that number of sold copies very often does not correspond with provided certification. Because labels tend to redeem them later (or not to redeem at all).