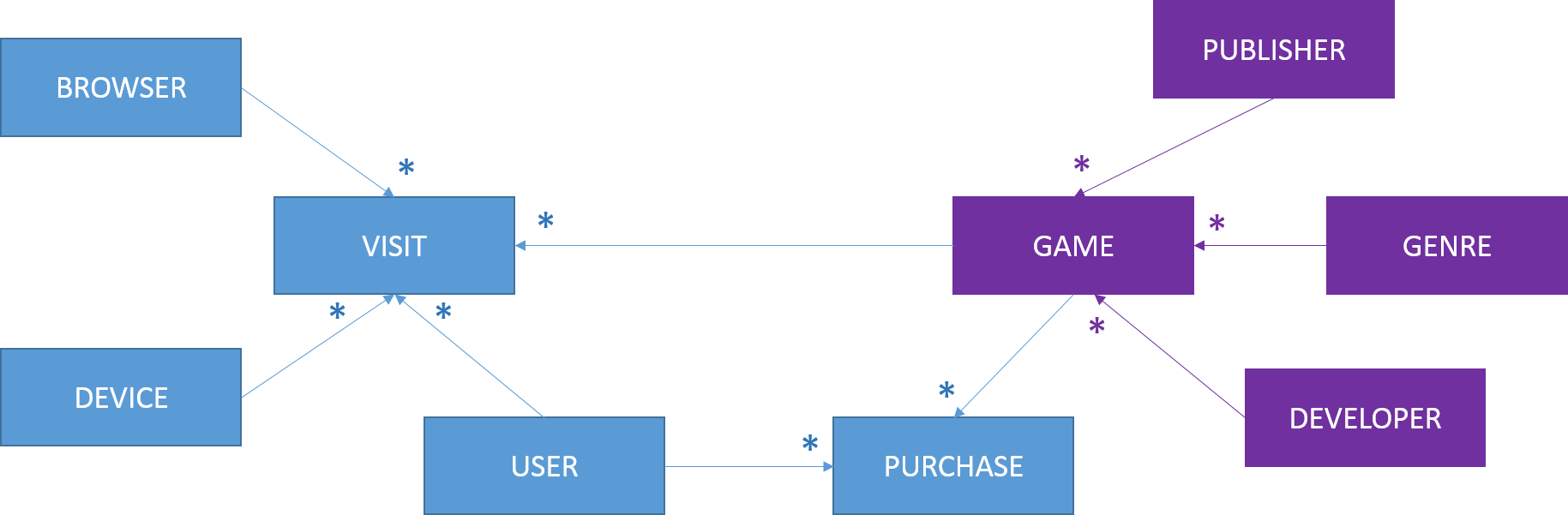
# Video Game Web Retailer

This database is for a web retailer who sells video game titles. Users visit various game detail pages and on occasion purchase a particular game. The schema is as under –



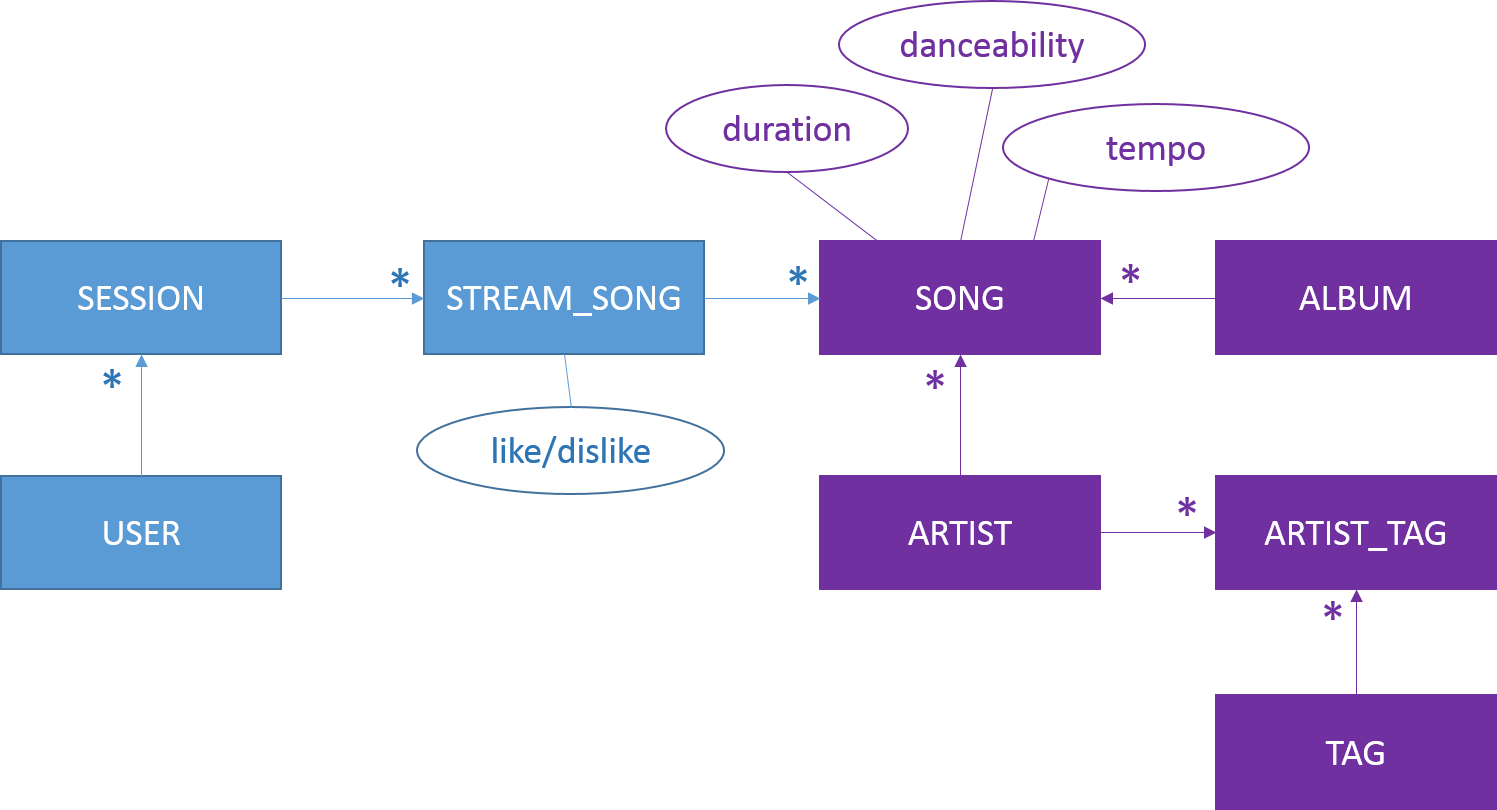
Some of our demo queries for this db can be (in increasing order of complexity) –

1. Which genres are the most popular (in terms of website traffic)?
2. Which publishers are the most profitable (in terms of number of purchases)?
3. Does the number of repeat visits to a game detail page predict its purchase by that user?
4. Do users visiting from a mobile device have a genre preference?

The raw data for this dataset will come from a [list of video games from Wikipedia](https://en.wikipedia.org/wiki/Lists_of_video_games). I’ll reshape that raw data into entities shown in purple in the picture above. I’ll also generate synthetic data for the entities in blue. We’ll have around 10,000 songs in the SONG table and 500,000 users. This will give us around a million rows in the PURCHASE table and 3-4 million rows in the VISIT table.

# Music Streaming Service

This database is for a “Pandora” like service where a stream of songs are played back for the user and the user can either like or dislike the song. The entity-relationship looks as under –



Some of our demo queries for this can be (in increasing order of complexity) –

1. What is the average number of songs streamed in a user session?
2. Do the first 3 songs predict the kind of songs the user wants to listen to in the session?
3. What kind of artist (as described by their tags) does a user like?

The raw data for this dataset will come from the [Million Song Dataset](http://labrosa.ee.columbia.edu/millionsong/). I’ll reshape that raw data into the entities shown in purple in the picture above. I’ll also generate synthetic data for the blue entities. We’ll have around a million songs in the SONG table and around half a million users in the USER table. This will give us couple of million rows in the SESSION table 5 – 6 million rows in the STREAM\_SONG table.

The gist of the [licensing terms](http://labrosa.ee.columbia.edu/millionsong/pages/what-are-licensing-terms) for the Million Song Dataset is a CC-AB license which means that all we have to do is cite them on our web page after getting prior permission from them.