A close up of a logo

Description automatically generated**A Million Rows of Music**

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |

ETL Project Proposal - By Ema, Kayti & Dinesh

**Description:**

The purpose of this project is to

* Extract the data of trending music from various data-sources,
* Collect the missing information using web-scraping & API calls.
* Cleanse the data by adding / removing the raw attributes
* And loading the final dataset to the SQL/No-SQL database for further analysis.

**Use-cases:**

The stored dataset can be used to analyze the trending musical numbers of recent past across the globe and identify the most popular genre, artist or the track.

The dataset can be drilled down to a lower granularity to find the musical tastes of people in a specific country or in a specific year.

Further-more, we can study the changing trends of world music.

**Data-sources:**

1. CSV from Kaggle

<https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking>

1. Web Scraping

<https://en.wikipedia.org/wiki/ISO_3166-1>

1. API

<http://ws.audioscrobbler.com/2.0/?method=track.getInfo&api_key=b848087a7bcf37ce7a1404dc164ed41d&artist=J%20Balvin&track=Safari&format=json>

**The strategy:**

We have chosen CSV with a Million Records as our primary source of truth, the file has all the basic information such as Track, Artist, Number of Streams and Country Codes.

However, the file has a lot of missing information such as Genre, Name of the Album, Description, Lyrics etc. So, we found an API which provides all the missing pieces.

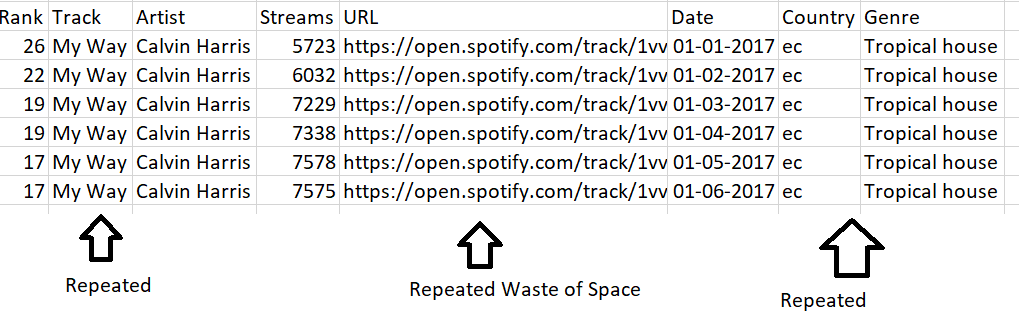
We are planning to do web-scraping to derive the country names form the country-codes.

**Loading:**

We are planning to store the data in Mongo-DB because,

When we store the data into a SQL database with the base attributes of each song such as Artist, Genre, Album, Lyrics, Country will be repeated every-time the song repeats.

**SQL Way:**

**Mongo Way:**

