SMU Data Science Bootcamp September 13th, 2021

A Case Study of Extract, Transform and Load: Spotify

Steven Green

Greyson Moore

Saatvi Rajgopal

Lucas da Silva

Background and Inspiration:

Spotify is a cloud-based music streaming service that allows users to listen to millions of songs directly from their phone, computer or internet-connected device. This type of music streaming has become the new normal for music fans and Spotify is the most popular music streaming platform in the world, servicing over 248 million users. Some of the advantages that this type of music platform offers are the ability to present the users with curated music charts based on genre, similar artists and popular songs near the location of the listener. It also has charts to showcase new and potentially popular songs, such as the “Viral 50” list, that is determined by compiling streaming data along from Spotify’s platform with analytics of trending music and artists from social media sites and blogs.

Proposal Topic:

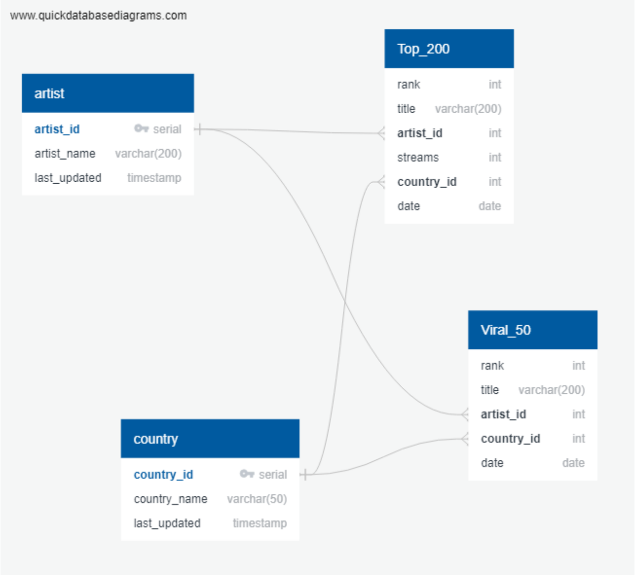
We will be creating a database on Google Cloud to store song information from Spotify’s most popular charts: the “Top 200” and the “Viral 50” . We will be analyzing the data from these charts for the most populous countries in the world where Spotify offers their services (United States, Thailand, India, Brazil, Mexico, Japan, Philippines, Egypt, Vietnam, and Turkey) and we will be comparing it with the overall global data that Spotify provides for these two charts. This data is compiled weekly and the timeframe we’ll be studying is from July 29, 2021, to September 1, 2021.

Process:

In order to create the database, the group will be web scraping song information from the Spotify charts website. The website contains lists of the “Top 200” streamed songs and the “Viral 50” songs in a week and it provides the ranking, song title, artist and streaming numbers for those songs. Using the Python library “BeautifulSoup”, the group will gather the information for each song over the selected time frame and will organize it into data frames using the Python library “pandas.” This data will be cleaned, transformed and uploaded to a PostgreSQL database. The group will then upload this database to the Google Cloud.

ERD:

We will be creating 4 tables: Artist, Country, Top 200 and Viral 50. The artist and country tables will have the names of the countries/artist and their primary keys (artist_id and country_id). These primary keys will feed into the Top 200 and the Viral 50 tables that will already have the ranking, title of the song, stream count, date and last updated information. Looking like this:



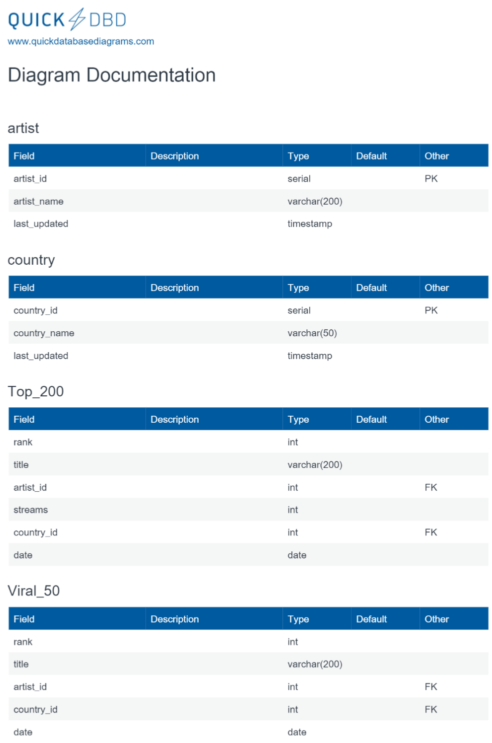


Diagram Documentation

artist

Field	Description	Type	Default	Other
artist_id		serial		PK
artist_name		varchar(200)		
last_updated		timestamp		

country

Field	Description	Type	Default	Other
country_id		serial		PK
country_name		varchar(50)		
last_updated		timestamp		

Top_200

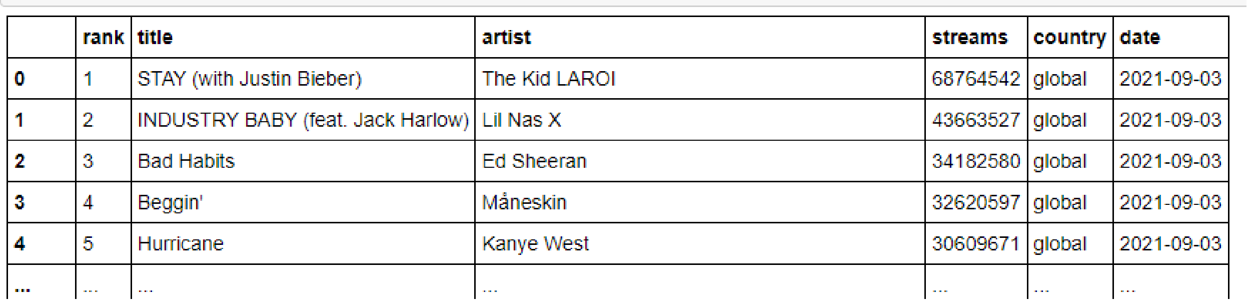
Field	Description	Type	Default	Other
rank		int		
title		varchar(200)		
artist_id		int		FK
streams		int		
country_id		int		FK
date		date		

Viral_50

Field	Description	Type	Default	Other
rank		int		
title		varchar(200)		
artist_id		int		FK
country_id		int		FK
date		date		

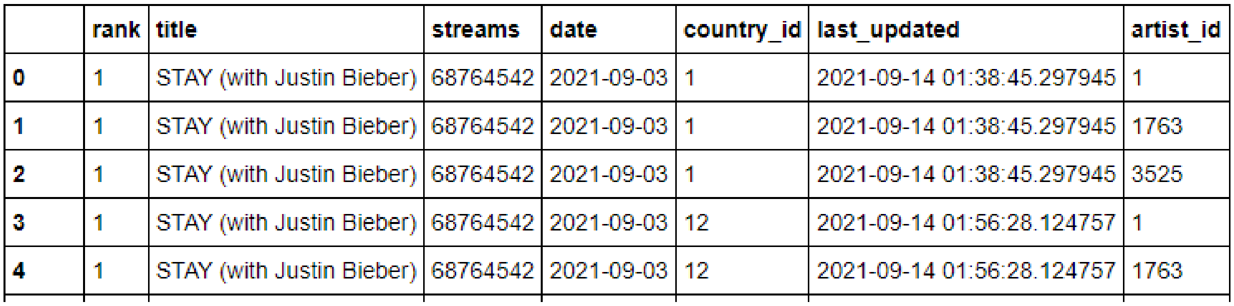
Web Scraping and Data Cleaning:

The group chose to use Python's Beautiful Soup to extract the data, and Pandas to help transform it from a disorganized array into an easy-to-read data-frame. The base URL chosen was https://spotifycharts.com/regional/{countriesList[i]}/weekly/{start_date}--{end_date}, the country list and dates are present above in the Proposal Topic section of this document. After the data is extracted, Pandas organizes it in a table with the following column headers: Rank, Title, Artist, Streams, Country and Date:



	rank	title	artist	streams	country	date
0	1	STAY (with Justin Bieber)	The Kid LAROI	68764542	global	2021-09-03
1	2	INDUSTRY BABY (feat. Jack Harlow)	Lil Nas X	43663527	global	2021-09-03
2	3	Bad Habits	Ed Sheeran	34182580	global	2021-09-03
3	4	Beggin'	Måneskin	32620597	global	2021-09-03
4	5	Hurricane	Kanye West	30609671	global	2021-09-03
...	...	...	...	...	...	...

Then, the charts extracted from Spotify are merged and transformed with the country and artist tables, so they generate a table with chart data, primary keys, foreign keys and time stamp:



	rank	title	streams	date	country_id	last_updated	artist_id
0	1	STAY (with Justin Bieber)	68764542	2021-09-03	1	2021-09-14 01:38:45.297945	1
1	1	STAY (with Justin Bieber)	68764542	2021-09-03	1	2021-09-14 01:38:45.297945	1763
2	1	STAY (with Justin Bieber)	68764542	2021-09-03	1	2021-09-14 01:38:45.297945	3525
3	1	STAY (with Justin Bieber)	68764542	2021-09-03	12	2021-09-14 01:56:28.124757	1
4	1	STAY (with Justin Bieber)	68764542	2021-09-03	12	2021-09-14 01:56:28.124757	1763

Conclusion:

After the tables were loaded, the group would have liked to run a few queries to answer a few basic questions about the data. The first query concerned artists that overlapped between the global Top 200 chart and Viral 50 chart. The purpose would be to show if viral successes could lead to mainstream successes as well and, if that was the case, how often did it happen?

The second query would concern the most popular artist in our dataset and if other artists that formed partnerships with this top artist could eventually become mainstream successes as well. This could be used to show that songs in the viral list featuring these top artists could reach global success because of the influence of these more established artists.

The last query would have been about what country has the most influence on the global lists. This could have helped determine if popular songs, by stream count in specific countries, are more likely to become global successes if they initially become popular in certain markets.

The answers to these questions are heavily influenced by the limitations of the dataset, limitations of time and constraints of this project. If time permitted, the group would have liked to include more information from datasets, covering longer periods of time, to the tables to have a better understanding of long-term trends and relationships between top artists and songs in the Top 200 chart and Viral 50 chart.

Some other interesting questions the group would have liked to explore, if it wasn't for the limitations listed above, are: based on stream counts and country data, can we predict what artist/song is going to have the most streams for the month in the global charts? Can we project which artists are going to rise and fall during the year? Is there seasonality to music? Are some artists/genres more likely to be at the top during the summer? Or during the winter?

With a greater dataset, we could conceivably answer all of these questions and have a much stronger statistical support for any hypothesis that the group would have liked to explore

Sources:

https://spotifycharts.com/regional/global/weekly/latest

https://jennifer-franklin.medium.com/how-to-scrape-the-most-popular-songs-on-spotify-using-python-8a8979fa6b06

https://www.worldometers.info/world-population/population-by-country/