Alfred Custodio
Diana Melendez
Jon Arnold
Jorge Serrano

September 13th, 2021

ETL PROJECT 2 – GROUP 1

INTRODUCTION

ETL or Extract, Transform and Load has roots in the 1970s and the rise of centralized data repositories. But it wasn't until the late 1980s and early 1990s, when data warehouses took center stage, that we saw the creation of purpose-built tools to help load data into these new warehouses. Early adopters needed a way to "extract" data from siloed systems, "transform" it into the destination format, and "load" it. The first ETL tools were primitive, but they got the job done. Granted, the amount of data they handled was modest by today's standards.

For this week's ETL Assignment, Team 1 dove into the music charts in which we took a look at three of the top music websites (music.apple.com, billboard.com & rollingstone.com) in order to analyze specific elements in their Top 100 Songs Charts including:

- Rank
- Song
- Artist's Name

DESCRIPTION OF DATA

The ETL process was utilized in order to combine structured and unstructured data retrieved from the sources so it could be manipulated into table structures optimized for reporting.

EXTRACT

Data extracted from 3 sources includes Billboard.com, Rollingstone.com and Apple Music.com. Tables generated include a main table with ID, rank, song, artist_id and source_id. Lookup tables include the artist_table, and source_table. Serial Primary keys are ID, artist_id and source_id.

Extraction from the three sources included “test-find” functions to identify the information needed from the corresponding class. After the class and data was pulled, a “for-loop” was used to pull the rank, title and artist. Using this information a data frame was created with Rank, Song and Artist. The extraction for the Rollingstone.com site included adding a function to execute the “load more” button in order to retrieve the total top 100 songs from their list.

- **ROLLING STONE**

  Rolling Stone’s Top 100 Songs - Albums are ranked by Album Units, a number that combines audio streams and both digital and physical purchases, using a custom weighting system. Songs are ranked by Song Units, a number that combines audio streams and song sales, also using a custom weighting system. Album and Song Units are holistic, multilevel metrics that attempt to present the most reflective picture possible of what's happening in music.

- **BILLBOARD**

  On Billboard the charts reflect weekly sales and streams on a Friday-to-Thursday cycle. However, the mixed data charts, such as the Billboard Hot 100, use a radio airplay cycle of Monday to Sunday. Charts are refreshed every Tuesday on billboard.com and billboard.biz and the issue date is four days later on Saturday.

  For example: Charts posted on Tuesday the 24th of the month (with an issue date on Saturday the 27th) is based on the previous Friday-to Thursday cycle (12th-18th).

- **APPLE MUSIC**

  Apple Music is a music and video streaming service developed by Apple Inc. Users select music to stream to their device on-demand, or they can listen to existing playlists. The service also includes the Internet radio stations Apple Music 1, Apple Music Hits, and Apple Music Country, which broadcast live to over 200 countries 24 hours a day. The service was announced on June 8, 2015, and launched on June 30, 2015. New subscribers get a six-month free trial period before the service requires a monthly subscription.

TRANSFORM

Extracted data is moved to a staging area where transformations occur prior to loading the data into the warehouse. A source column containing the website was added to the three extracted dataframes. Afterwards, a dataframe containing the list of websites was created to be used as a lookup table for the load process. Then, the dataframes from the three websites were concatenated together to form a 300 row dataframe. Setting column names to lowercase was needed in order to be accepted when loading. After the artist and source lookup dataframes were loaded to the database, they were read back into python now containing their serial IDs that were created when loaded. Those lookup dataframes were merged with the master dataframe to give it artist and source IDs which allowed us to drop the artist and source columns to reduce data redundancy.

LOAD

A PostgreSQL database was created on GCP, google cloud, and pgAdmin was used to connect and read the cloud database. As mentioned in the transform step, the artist and source dataframes were loaded to auto-create serial IDs. After the master dataframe was merged and cleaned, it was loaded into the database successfully. All these load steps used ‘df.to_sql()’ to complete the process. A sanity check was performed using ‘pd.read_sql_query()’ to make sure the load was processed as expected. It is important to note that the dataframe columns needed to be lowercase to match the lowercase column names in the database tables.

LIMITATIONS

One limitation encountered was the time frame which showed bias toward a certain artist due to their album release the week before the data extraction. For example, in Rolling Stone's Top 100 songs, the artist "Drake" occupied 20 of the top 21 places. Incidentally, Drake's new album: Certified Lover Boy includes 21 songs and it happened to be released on Labor Day weekend or September 3rd. The only song in the album that did not make it in the top 20 was "Pipe Down"

FUTURE WORK

As a team, we concluded that if time and money were no objects, we would have liked to explore other key elements included in the datasets. Items such as: Music Label, Top 100 Songs in different Music Genres, Number of Weeks at Number 1 and the number of Weeks in the ratings. Another element that we would have liked to work on was the visualization and comparison of the top songs charts depending on the weekly and monthly trends.

FINAL CONCLUSION

In our final conclusion, we determined that music streaming services heavily influence modern top 100 charts. More than half the songs from Drake's new album are topping the charts, and everyone can access this music through their preferred streaming service. There are still artists that release singles and make it into the top 10; however, an album release has the potential to overtake the flavor of the week and move those current number one songs out of the top 10.

SOURCES

- **Rolling Stone**
- **Billboard**
- **Apple Music**