**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
* Album Title
* Artists Name

**DESCRIPTION OF DATA**

The ELT 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. Other 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](https://en.wikipedia.org/wiki/Streaming_media) service developed by [Apple Inc.](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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

**LOAD**

The process of extracting data from multiple source systems, transforming it to suit business needs

**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….

**SOURCES**

* [**Rolling Stone**](https://www.rollingstone.com/)
* [**Billboard**](https://www.billboard.com/charts/hot-100)
* [**Apple Music**](https://music.apple.com/us/playlist/top-100-usa/pl.606afcbb70264d2eb2b51d8dbcfa6a12)