Discogs Album Database- More than 16 million releases Using Hadoop

Sereyoudom Eab, Daniel Garrido, Uriel Guijarro, Michael Miranda, Giovanni Munoz, Yana Polshyna
Department of Information Systems, California State University Los Angeles
Tel. 323-343-2916, Fax. 323-343—5209
CIS4560-01 Introduction to Big Data
E-mail: seab@calstatela.edu, dgarrid3@calstatela.edu, uguijar@calstatela.edu, mmiran64@calstatela.edu, gmunoz58@calstatela.edu, ypolshy@calstatela.edu

Abstract: Discogs, a prominent online music database and marketplace, offers a wealth of data for analyzing music trends across diverse genres, styles, and locations. This analysis delves into the intricacies of Discogs' extensive collection to uncover insights into popular music preferences and industry dynamics. By examining the frequency of releases and user engagement, we can identify the most popular genres and styles of music across different countries and release dates. This analysis can reveal regional music preferences and the evolution of musical trends over time. In addition to that, analysis of this data is conducted using Excel, depicting visuals such as maps, timeline and charts on trends of music in different regions of the world.

1. Introduction

This project uses Hadoop and Hive to keep and process Discogs offers a wealth of data that can be harnessed to analyze music trends across diverse genres, styles, and locations. By exploring the popularity of genres, tracking release patterns, visualizing data geographically, determining popular distribution formats, examining industry trends, and identifying influential record labels, we can gain a deeper understanding of the ever-evolving landscape of the music industry and the factors that shape it.

We have chosen this dataset because it is a valuable resource for studying music trends. It is comprehensive, accurate, granular, and accessible, and it is relevant to a wide range of research questions.

2. Related Work

A related piece of work we found was about MySpace genres that musicians used to describe themselves. There were 122 categories, the top three being hip hop, rap, and R&B. The bottom three were samba, tango, and Italian pop. This dataset was the largest available to ensure a hypothesis could be made from the data. This is similar to ours, in that we also chose a large dataset to gather good data and to be able to make observations from it. The analytical part of it is different from ours in that they extracted data from the website from the users profiles. They also used an algorithm to identify genre communities. (Silver, 2016)

Another related piece of work went over features that were often found in different music genres. For example, when looking at classical and country, the loudness level of the songs were low. Then, when looking at pop and rap music, the tempo, being the feature, was high. It is similar to ours by looking at similar attributes between different genres. The paper is different from ours because they used theory to calculate the different features of the music and the density as to how common it was in a genre. (Stetler, 2022)

A third paper relating to our work had to do with recurring themes in music genres. For example, in hip hop songs the theme of drugs was seen a lot, whereas in Latin music there was hardly any. In the theme of love, Latin music was the leader of the genres while hip hop was the lowest one. This relates to our work in that it looks at trends between the themes of genre and what is popular in each genre. This piece of work is different to ours in that they also decided to generalize the themes into positive, negative, and neutral. Determining if a genre has overall positive, negative, or neutral themes. This demonstrates that the goal of this particular paper was to demonstrate common themes in music genres and its different to ours in that we didn't associate a particular trend with something else. It would get very generalized. (Kwon, 2021) These related works demonstrate the significance of completing this project and why its important that this research be done. Music has become incredibly mainstream and the talk at many social gatherings. It is also a growing industry. Artists should understand this sort of data to see what would work in their respective genre.

3. Specification

The database is mostly to track and share music collections, provide information about music recordings, and facilitate the buying and selling of music recordings. Since the database is too large and several of the fields appear to be unneeded or unsuitable for our research, we chose to include the 12 most significant columns, which include: Release Date: The release year of the music, Release ID: This is the unique identifier for this particular music release, Status: Indicates that it has been accepted, showing that this item has a specific status, Title: The title of the music release, Artist ID: Artist ID of the album artist, Artist Name: Artist name of the album artist, Label Name: This section specifies which music label released the music, Label ID: The label's catalog number, Format: This section details in which format the music is available, Genre: Indicates the genre of the music, Country: The country of origin for this music release, Company Name: This section includes information about companies involved in the production, and Style: Specifies the style or genre of the music.

The below table show the specification for Oracle cluster we are using and Hadoop specification for our project.

Table 1 H/W Specifications

Number of Nodes	***3***
OCPUS	8
CPU Speed	1995.312 MHz
Memory	58 gb

4. Implementation Flowchart

Initially, the entire process yielded a plethora of data that may be used to study music patterns across various genres, styles, and regions. We used this process to obtain our desired data: Download the data from Kaggle: Log in to your Kaggle account and navigate to the dataset you want to download. Click on the "Download" button and choose the desired format (CSV, ZIP, etc.). Transfer the dataset file to the Linux server: Use a secure file transfer method, such as SCP or SFTP, to copy the downloaded data file to the Linux server. Ensure you have the necessary permissions to access the server and the destination directory. Run a bash command in order to clean up unnecessary characters within the file globally. Download the cleaned data file back to the local computer. Download and install an external app called CSViewer instead of using Excel to work with 16 millions rows of data. Excel has a difficult time handling the file because of the number of the rows and the multiple languages that are present in the dataset. Load the data into CSViewer. Change the delimiter of the whole dataset from commas to pipe. Export the data file out of CSViewer with the necessary columns. Analyze and synthesize the data with queries on Hadoop. Export the tables that were created and synthesized back out to the Linux server. Download the files on to personal computer using SCP command. Using Excel and Tableau to perform analysis and visualization of the information that derives from the queries in Hadoop. Using Excel and Tableau to perform data.

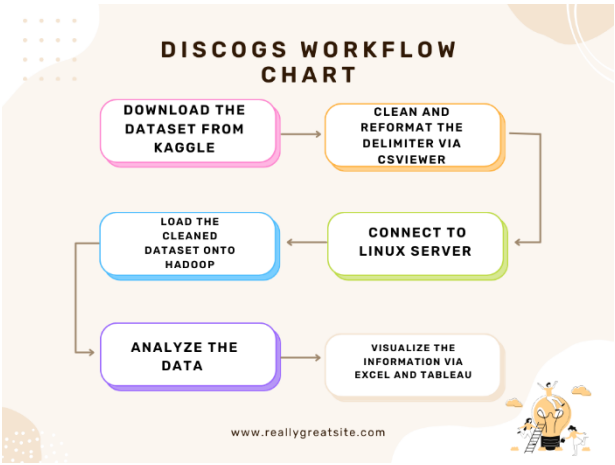

Figure 1- Discogs Flowchart

5. Data Cleaning

The data contained commas within data entries, which were causing separation issues when using a comma as the delimiter. The data analysts decided to use a pipe '|' character as the delimiter instead, as it was less frequent within data entries and deleting it would not affect the integrity of the data. The data was first uploaded to a Linux server, and all pipe characters were deleted from the file using Bash command sed -i 's/|//g'. The bash command allows us to use a steam editor to perform basic text transformations on an input stream. Sed with –i directly modified the specific file. 's/|//g', this is a substitute command that replaces any occurrences of pipe '|' with nothing, and it is globally implemented to the entirety of the file. The file was then copied back to local storage and opened in CSViewer. Unnecessary columns were removed, and the file was modified to use the pipe character as the new delimiter. Once the data was clean, the data analysts were able to create tables with consistent and accurate data. The decision to use CSViewer instead of using regular expression within Hadoop to synthesize only the data that we need is boiled down to the lack of knowledge in using regular expressions to remove commas within the string that enclosed inside of the quotation marks without impacting the integrity of the dataset as a whole; plus we shorted on time when it came to the decision to select this dataset.

Data cleaning, therefore, laid the foundation for a comprehensive and insightful analysis of Discogs' extensive music data. By ensuring the accuracy and reliability of the data, researchers were able to uncover valuable trends and patterns in the music industry.

6. Analysis and Visualization

6.1 Release Count by Year

After cleaning the data, we created a line chart that illustrates the number of music releases through Discogs per year from 1922 to 2022. The number of releases rises steadily throughout the year, reaching a high of over 200,000 in 2010. Following then, the quantity of releases reduce slightly but remains high.

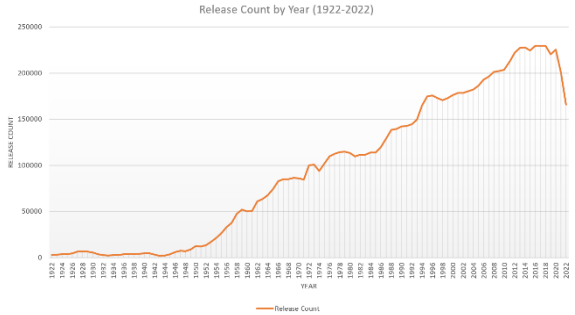

Figure 2- Release Count by Year Chart

6.2 Release Count By Country

From 1922 until 2022, this graph depicts the number of music releases by country. It is based on information from Discogs. According to the graph, the United Kingdom has continuously had the most music releases, followed by Germany and Japan. France, Italy, Canada, the Netherlands, Spain, and Australia are also significant music-releasing nations.

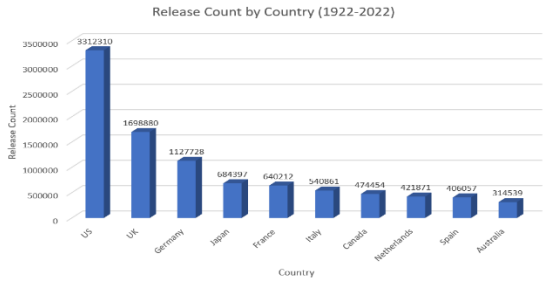

Figure 3- Release Count by Year BarChart

6.3 Release Count By Country

This GeoMap depicts the evolution of music genres from 1922 to 2022, as well as how many song genres have evolved globally over the years.

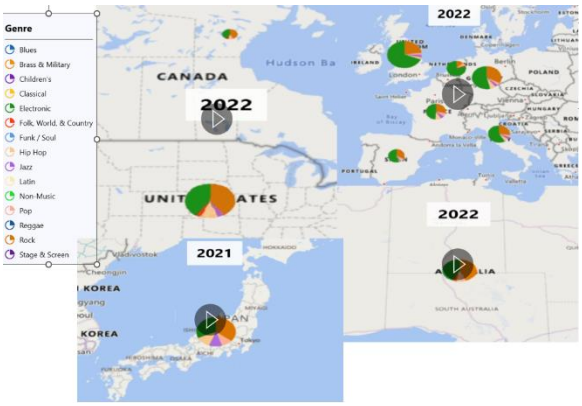

Figure 4- GeoMap Time-lapse (1922-2022)

6.4 Release Count By Genre

The graph shows the total number of music releases by genre from 1922 to 2022. The most popular genres are rock, electronic, and pop, followed by hip hop, folk, world, and country. Classical, jazz, funk/soul, and Latin are also popular genres, but have fewer releases.

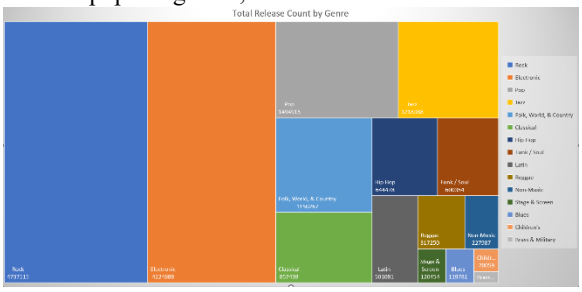

Figure 5- Heatmap Release Count by Genre

6.5 Release Formats By Genre

The distribution of music formats per rock genre is depicted in this pie chart. Vinyl is the most widely used format, followed by CD, cassette, CDr, file, DVD, 8-track cartridge, VHS, flexi-disc, and lathe cut.

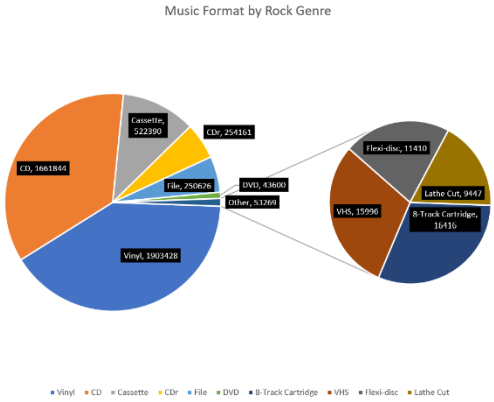

Figure 6- Music Format by Rock Genre Pie Chart

This pie chart also discusses the most purchased/used music formats for Electronic Music, with Files and Vinyl being the most popular.

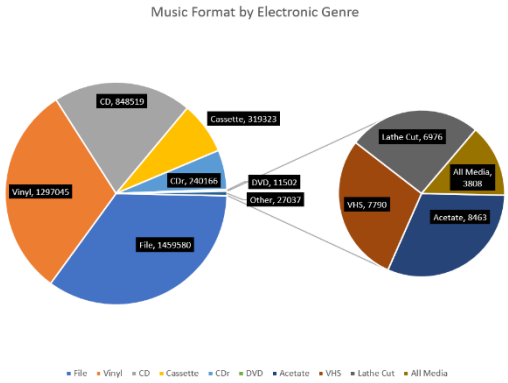

Figure 7- Music Format by Electronic Genre Pie Chart

This Pie chart also looks at the most used/bought format for the Pop Genre, which reveals that vinyl and CDs are the most popular formats for this genre.

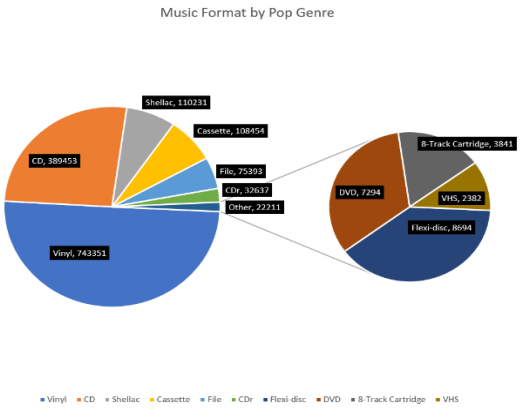

Figure 8- Music Format by Pop Genre Pie Chart

2 In Figure 3- Release Count by Year BarChart We opted to include only the top ten nations that release music in the barchart because it is crucial to represent the top ten countries that had the greatest impact on this database for us and how it could impact future data.

7. Conclusion

Finally, to summarize what we discussed in this paper, the largest peaks of music releases occurred between 2016-2018, with the United States releasing the most music, followed by the United Kingdom. Also, the most popular music formats for the Rock and Pop genres were vinyl and CD, but for electronic music, the most popular formats were vinyl and files.

We learned about music genre trends from the 16 million data points collected by Discogs users between 1922 and 2022, as well as the most popular music formats for each genre. We now understand how music genre trends fluctuate depending on place and how much they might influence the music industry.

For more information such as codes/slides visit the GitHub link provided bellow.

References

[1] Kwon, L., Medina, D., Ghattas, F., & Reyes, L. (2021). Trends in positive, negative, and neutral themes of Popular Music from 1998 to 2018: Observational study. JMIR Pediatrics and Parenting, 4(2), e26475. https://doi.org/10.2196/26475

[2] Silver, D., Lee, M., & Childress, C. (2016). Genre complexes in popular music. PLOS ONE, 11(5), e0155471. https://doi.org/10.1371/journal.pone.0155471

[3] Stetler, R., "Exploring Music Genres: A Study of Optimal Differentiation by Feature" (2022). Honors Projects. 788. https://scholarworks.bgsu.edu/honorsprojects/788

[4]GitHubLink: https://github.com/danielgarrido1/Term-Project-Abstract-Discogs

[5]DataSource: https://www.kaggle.com/datasets/ofurkancoban/discogs-releases-dataset

3. GitHub Link: https://github.com/danielgarrido1/Term-Project-Abstract-Discogs