Università degli Studi di Milano-Bicocca

Data management and visualization

Artist investment research with Spotify

ad opera di:
Davide Banfi - 806539
Giacomo Cesareo - 805716
Jacopo Nicosia - 861812

Introduction

As companies grow, exploring new industries is more than ever a necessity for big companies which plan to diversify from their core business.
Think of Apple, initially a manufacturer of smartphones and PCs, which now offers payment services with Apple Pay, entertainment with Apple TV or in the music field through Apple Music.
For this reason, in consultancy firms, feasibility studies and market analyses on behalf of large companies are now routine.

The following elaboration aims to simulate a consulting project in the form of an analysis of the musical market, for a company interested in assessing the possibility of investing in the sector.
The process has been structured in two main parts: a first descriptive part of the macro characteristics of the musical market, and a second part of methodic research and skimming of the artists as means of investment, following a *data driven* approach.

Given these objectives, the technical approach is as follows.

It has been decided to pursue the creation of a database that would guarantee **volume** and **variety** at the expense of **velocity**, as it was necessary to have available data that could provide as much descriptive capacity as possible.

The data was scraped across the data sources Spotify, GeniusLyrics and Wikipedia using their respective APIs, and eventually stored in BSON form into MongoDB. The technical details of this process are described below.

As a database inclusion criteria the measure "*popularity*" was used, that is a measure of synthesis based on a proprietary Spotify's algorithm having a domain between 0 and 100.
This parameter has been chosen because it embodies many aspects to determine its overall value, mainly the chronological proximity of the artist's production and the amount of recent streams.
This allowed to skim "outdated" artists with greater speed without penalizing the representativeness of smaller emerging artists, also speeding up the data collection phase.
Only the artists who had a popularity greater than 60 in the last week of June 2020 have been taken into account.

Research questions

Being a "consulting project", the questions that had to be answered were those of a potential client whose knowledge of the musical market is scarce.
After an extensive exploratory analysis, the questions to which an attempt was made were:

With the aim of finding effective investors:

- **What are the high potential artists to invest in?**

With the aim of finding markets that are not yet saturated and understanding their size:

- **Where and what kind of music is played?**

Capturing and describing data

Data collection process involved the development of a *scraping* program in Python language (see "Operating Guide") and the use of a *distributed database management system*, MongoDB, to store the collected data.

A composite binary integration logic was used for data integration and enrichment.
The starting node is a JSON document powered by the Spotify API which contains the *ID keys*, *artists names*, *popularity*, *number of followers* and *url* of the artist's Spotify page as well as the associated genre.

In a secondary phase, data for album titles recorded by each artist throughout his career was downloaded. These have been used later as the base of a cycle with the aim to download the songs contained in each album.
Lyrics were then integrated thanks to Genius' API into the key songs contained in each album as well as the relative release date. At the same time the song metrics calculated with Spotify's proprietary algorithm were added to the *variables* key.

In the final phase data has been enriched using the API of Wikipedia Italia as source.
Using the key *artist*, additional demographic information such as *sex*, *age*, *city of birth* and *nationality* were added.
The ranking of the top 50 cities where each artist met the largest number of *streams* has been included as well. Each date has been assigned a unique key.

This procedure was meant to meet the requirement for data **variety**.

As for the management of the **volume**, it was decided to use a non-relational document-based solution, specifically the DBMS MongoDB.

The criteria that motivated this choice were as follows:

- **Query flexibility**: MongoDB supports searches by field, intervals and regular expression.
- **High reliability**: MongoDB provides high availability and high load management through replica sets distributed within the available servers.

- **Sharding and balancing**: MongoDB scales horizontally thanks to sharding.
  The user must choose a sharding key which determines how the data in a collection will be partitioned across the various nodes. Furthermore, Mongo includes its own data balancing mechanism in order to maintain data balance between shards.
- **Friendly user interface**: MongoDB supports Compass as a very intuitive and user friendly UI.

This way most of the problems related to volume management was solved thanks to the use of the paid version "Cluster M20" offered by Mongo (4GB of RAM, 20 GB of disk).
In particular, the free rental time offered to university students made it possible to use a distributed storage server located at our own preference. In order to reduce latency spikes the closest server, Frankfurt, was chosen.

This solution allowed to automatically manage the data distribution as well as the sharding of the database. The outcome has been a big time save channeled to the analytics phase.

The result consists of about 6500 artists covering a total of 2.6 Gigabytes, which Mongo compresses into 2 Gigabytes.

The structure of the final JSON is pictured below.

```
|_ _id                    # id dell'artista
|_ artist                 # nome dell artista
|_ albums                 # Contiene una lista di dizionari, uno per ogni album
|____ name                # Nome dell'album
|____ tracks              # Lista di dizionari inerenti alle canzoni
|_______ songname         # Nome della canzone
|_______ language         # Lingua del testo della canzone
|_______ lyrics           # Lyrics
|_______ variables        # lista di dizionari dove ogni variabile è chiave del valore effettivo.
|__________ danceability
          ...
          ...
          ...
|__________ loudness
|_ related                # Lista contenente i 4 principali artisti correlati.
|_ age                    # Età dell'artista.
|_ città                  # Città di nascita dell'artista.
|_ sex                    # Sesso dell'artista.
|_ nationality            # Nazionalità di provenienza.
|_ followers              # Numero di followers su Spotify.
|_ genres                 # Lista di generi dell'artista
|_ popularity             # Livello di popolarita su Spotify
|_ url                    # Url della pagina Spotify dell'artista
|_ streams                # Lista di dizionari dove il timestamp è chiave.
|____ 01/01/2020
     ...
     ...
     ...
|____ 31/12/2020
|_______ città 1          # Dizionario per ogni citta dove la città e chiave e
       ...                # il numero di streams è il valore.
       ...
       ...
|_______ città n
```

Description of problems encountered in exploratory analysis (Data Cleaning)

The main problems encountered mostly concerned data *quality*.
One of the major findings was the quality of the data coming from Wikipedia, which required numerous adjustments in the programming phase to eventually be converted in a form with tolerable level of errors.

The notable errors are listed below:

- The lack of matching artists in Wikipedia, mainly caused by the syntactic difference of a same name reported in Spotify and in Wikipedia. The presence of special characters within the name highly influenced the outcome.
  This last problem was solved by using the Wikipedia API as a data socket together with an edit distance algorithm to reduce matching errors.
  E.g.: the artist "Adele" is reported as "Adele_(singer)" on Wikipedia and simply as "Adele" on Spotify, so it was necessary to use an edit distance algorithm with a tolerance threshold of 9.
  The threshold was chosen by making an approximate average of the length of the words following the artist's name, such as "rapper", "musical group", "musician", ...
- The insertion of the wrong sex, caused by the variance of the introductive expression used on Wikipedia, namely "è un artista/è un'artista/sono...". The correction required a further adjustment in the scraper.
- The collection of dates of birth sparked problems during the age calculation: this is the case of groups and deceased artists.
  The relative correction consists in detecting past tense phrases like "è stato/è stata" and using them as markers to identify deceased artists.
  In addition to that, the average group age was roughly calculated considering all members as 20 years old at the moment of group foundation. This was chosen as a representative value which was encountered to be a very approximate mean for the majority of the group members (the majority were between 14 to 26 years old at the time of foundation).

Despite all the corrections, the demographic information from Wikipedia Italia generated a substantial amount of null values due to the fact that the Italian version does not have pages dedicated to internationally less famous artists.
It was therefore decided to use for the analysis only the data of the artists with all the information.

- In the *cleaning phase* most of the problems were encountered in the correction of the field "musical genre", as raw data presented many thousands of facets the majority of which being meaningless or wrong.
- Some lyrics were found to be wrong. In fact, some songs initially had abnormally sized texts probably generated by the presence of some podcasts on the platform.
  To overcome this kind of error, a filter was placed which forced the scraper to skip the lyrics in case it exceeded 9000 characters, so as to balance the possibility of very

long lyrics and avoiding the influence of podcasts.
Periodic random checks were also carried out to verify overall correctness.

- At last, the detection of daily streams is not available in a consistent manner for all artists. Eventually such artists were not taken into account.

Infographics

https://public.tableau.com/profile/davide.b#!/

In order to evaluate the data visualization process, a survey has been conducted to evaluate different parameters of the infographics.
The characteristics that were asked to evaluate were: clearness, beautiness, informativeness, intuitiveness, and eventually an overall score.
After the psychometric modules were completed, it was discovered that the least appreciated infographics were the word cloud and the quadrant matrix.
For what concerns the quadrant matrix, the bad feedback was given mainly by the population ranging from 40 to 60. They were probably the least familiar in the whole sample with basic knowledge of statistics, and indeed the visualization resulted a little bit too technical for them.
The word cloud instead was dragged down by the field of informativeness, which was lower compared to the other infographics.
Both the above infographics were kept in the final version of the project because the overall score was above the target threshold.

The geographical exploration has been the most appreciated one, with almost the maximum score achieved on each field.

Conclusions and potential improvements

The results obtained were overall satisfactory as the extracted artist, 070 Shake, is emerging and growing at a fast pace: nonetheless she is already recorded under the record label of Kanye West since 2016, one year after his debut.
The good result is therefore to be able to find without previous knowledge of the music market an emerging artist with high potential and on which an expert in the field would invest.
The counter downside concerns the timing as being already under contract reduces the investment opportunities.
Possible alternatives include the purchase of the contract from the label under which she is registered, or the organization of tours on behalf of the label.
The results obtained give hope that the search and selection process is fundamentally correct.

The findings showed substantial room for improvement in the procedure currently used.
In the first instance, the maturity of 070 Shake is already slightly more advanced than the target initially desired, and this tells us that an improvement could be to include in the analysis artists with less than 60 popularity.

In addition, having more time and resources available, the improvements that could have been made to the project would have been:

- trying to enrich the database with data from other music streaming platforms.
- collect streams data for longer periods in a way that you can get a more complete picture of the seasonality of the music market.
- implement a data update tool in the architecture (*e*.g. Kafka) so as to be able to have an appropriate tool to catch up the speed required for this kind of investment research.