University of Tours

B.D.M.A

Big Data Management Analytics

Report Phase 1 Decisional Project

Author:
Akaichi, Ines
Cissé, Ismaila
de Saint Ceran, Louis
F. Nascimento Filho, Jessé

October 17, 2018

Contents

1 Introduction

Not long ago, there was a strong belief that the internet was killing the music industry. For years now, the industry has failed to keep up with the rapid pace of technological advancement and barely had any understanding of their audience, who was buying their CDs, or cassettes.[7]

With streaming services taking over, these companies found ways to track the listening habits of users and have access to detailed information such as when, how, where, and who is listening to what.

The aim of the industry, now, is to use these customer behavior insights together with knowledge of the music itself, which is made possible only with Big Data. The raw music that is produced is essentially like unstructured data. In the digital era, this raw music can be easily digitized and analyzed.

As part of our master degree ‘s program , we choose to work on the design and implementation of a business intelligence system[8] supporting the analysis of songs. Our objective is to analyze the current music streaming industry and more precisely the songs and the artists’ popularity.

2 Management Methodology

Our team is a small multicultural and autonomous group, for this reason we found a common equation to process with ours tasks project life cycle. The result from this equation become our management methodology that is based in a lean solid and agile solution called Scrum. For this purpose we decided to follow the best practices of Scrum combined with a Kanban approach.

2.1 Scrum

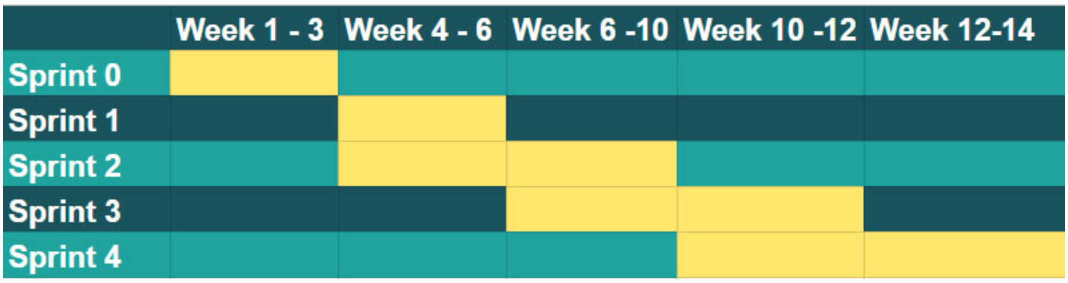
The Scrum arise as a process framework to manage complex projects since the early 1990s. The essence of Scrum is to provide effective iterations and incremental knowledge transfer to the success of a project [9].

Our Project will be composed of 4 sprints during this semester:

- Sprint 0 — Study : The list of requirements and project planning;
- Sprint 1 — Model : Conception & modeling of data warehouse;

- Sprint 2 — ETL : Data Extraction, Transformation and Loading;
- Sprint 3 — DEMO : B.I system demonstration of an initial version;
- Sprint 4 — DEFENSE : Oral project presentation;

2.1.1 Schedule



	Week 1 - 3	Week 4 - 6	Week 6 -10	Week 10 -12	Week 12-14
Sprint 0	🟨				
Sprint 1		🟨			
Sprint 2		🟨	🟨		
Sprint 3			🟨	🟨	
Sprint 4				🟨	🟨

[1]

2.1.2 Sprint 0

The name SPRINT 0 has been learned to describe the preparation phase which precedes the launching of the project. The term SPRINT 0 is being simpler to use than the preparation or inception phase, it is increasingly used in SCRUM projects. Sprint 0 does not diminish the flexibility of our project. On the contrary, it will allow us to anticipate certain actions and have an overview that will facilitate the management of changes that will emerge at the following sprints. [6]

In this Sprint we will be able to :

1. Share a clear vision of the project;
2. Identify users need;
3. Identify the preliminary workload resulting of the users need;
4. Prepare the project management plan;

2.1.3 Sprint 1

The preliminary specification of the workload in sprint 0 will help us in this sprint in modeling our data warehouse,thus the formalization of the entire workload. In addition, in this phase it is essential to maintain an active technological watch to choose our Essential BI tools used in next sprints.

2.1.4 Sprint 2

In this sprint we will be able to define our data warehouse 's architecture, assess the data quality and implement the designed ETL system.

2.1.5 Sprint 3

In this sprint we will be able to visualize our data using the BI restitution tools .

2.1.6 Sprint 4

In this final sprint , We will be able to prepare an oral presentation where we summarize all the steps that we have gone through when developing our project 's data warehouse and present our work to our professors.

2.2 Kanban

In addition of Scrum methodology we choose to use Kanban approach, that means “visual card” in Japanese, to help us simplify the sprints workload. We going to make a visual work-flow using Trello for create and manage all cards with micro tasks, it will result in each sprints deliveries milestones.

2.2.1 Trello

Trello[5] is a project management software that utilizes the concept of boards to represent projects and within boards, cards to represent tasks. Trello supports Team Collaboration enabling members to discuss a project in real-time. It keeps everybody informed through task assignments, activity log, and e-mail notifications.[2]

3 Preliminary Workload

3.1 Data Sources

The main services of stream nowadays are responsible for creating an efficient way to capture data and generate real information about the customers with it, but to obtain the success in this new world where data is like gold, they need to provide a service that is capable to get users loyalty. For this reason companies and communities of independent artists are building continuously distinct forms of technologies to present not only songs, but music with value add on it, such as, variety, quality and shareability. As a result of this frequent process innovation, today It's possible for any person to have access to huge an open-sources databases about music subjects. Hence,to make decisional projects become more simple with the follow assets MusicBrainz encyclopedia[3] and Spotify Web API. We choose to use into our project these datasets and, if necessary, others data sources could be attached as a asset during this project.

MusicBrainz Database[3] includes information about artists, release groups, releases, recordings, works, and labels, as well as the many relationships between them [table 1]. It is a community-maintained open source encyclopedia of music information. This means that anyone can help contribute to the project by adding information about your favorite artists and their related works.The entire dataset is 1.8 GB.

Attributes	type	granularity
released songs	text	by title
location	text	by region, country
genre	text	by binary
date	timestamp	by month,year
artists	text	by name, alias
origins	text	by country
cover	text	by name

Table 1: A brief description of MusicBrainz Database 's attributes.

Spotify Web API [4] Based on simple REST principles, the Web API endpoints return metadata about music artists, albums, and tracks[table 8], directly from the Spotify Data Catalogue.

Attributes	type	granularity
popularity	number	0-100
energy	float	0.0 to 1.0
valence	float	0.0 to 1.0
genre	text	by binary
danceability	float	0.0 to 1.0
duration	number	by by milliseconds
loudness	float	by decibels (dB)
mode	int	by major, minor
tempo	float	by beats per minute (BPM)

Table 2: A brief description of Spotify Web API 's attributes.

3.2 User Needs

Statistics about songs and artists:

- Number and Average of released songs disaggregating by genre, year , location and artist .
- The biggest/ lowest number of released songs by location , genre , year and artist .
- Number of artists disaggregating by origins.
- Number of artists appeared every year in the music industry.
- Number of songs or artists that achieved a certain popularity.
- The average popularity of the songs where artists participated to analyze artist's performance.
- What is the less popular songs by country and find out why ?
- What is the impact of cover art on success of an album? Number of recorded covers disaggregating by artist and song .
- The most covered songs by artist and song .
- Artists that are most engaged in the last years.
- What makes a top performer based on songs 's technical features?

Statistics about song musical features:

- Average duration, average tempo by artist and/or location and/or year.

- What makes a tube based on culture, market, political time, features of the song or the category of the song.

4 Data Warehouse Modelling

4.1 conceptual Design

Conceptual modeling is widely recognized to be the necessary foundation for building a database that is well-documented and fully satisfies the user requirements. In particular, from the designer point of view the availability of a conceptual model provides a higher level of abstraction in describing the warehousing process and its architecture in all its aspects.

For modelling our data warehouse schema we used The Dimensional Fact Model or DFM which is a graphical conceptual model, specifically devised for multidimensional design. In subsection 4.2.1 , we present our conceptual models corresponding to our proposed fact schemas .

4.1.1 conceptual schemas

In order to represent our DFM models , we used the requirement-based design approach. It is a bottom-up technique that allowed us to do the mapping between our analyzed requirements in subsection 3.2 onto our available data source so that we get to locate the conceptual objects at sources and make sure that the data verifying our user needs effectively exists.

Using this approach , We first start by defining the set of facts , measures and finally dimensions. Figure 1 shows the fact schema for the analysis of songs and artists musical features in which the fact consists of a released song with all its features in the month ,city of release ,the song 's genre and the group artist that released the song .

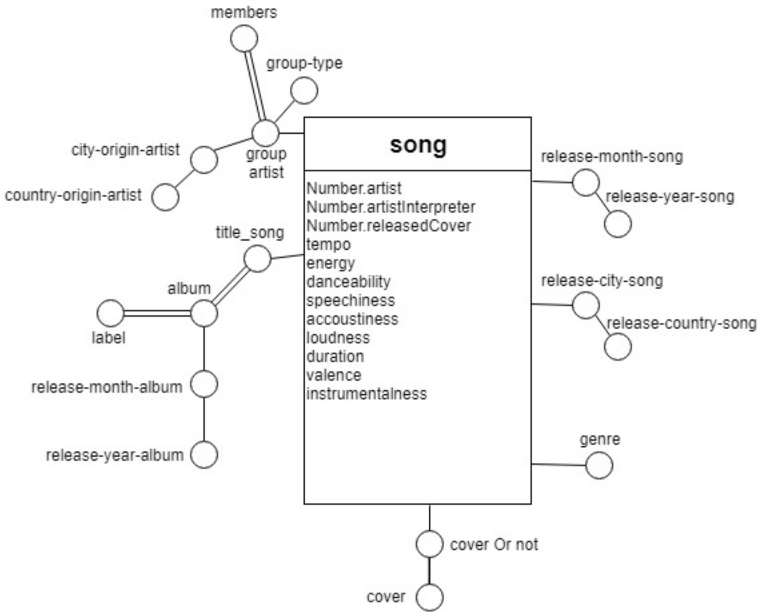


Figure 1: The Dimension Fact model for song.

Figure 2 and 3 shows the fact schema for the analysis of songs and artists popularity in the which the fact consists of song's popularity (figure 2) and the artist 's popularity (figure3) in every month .

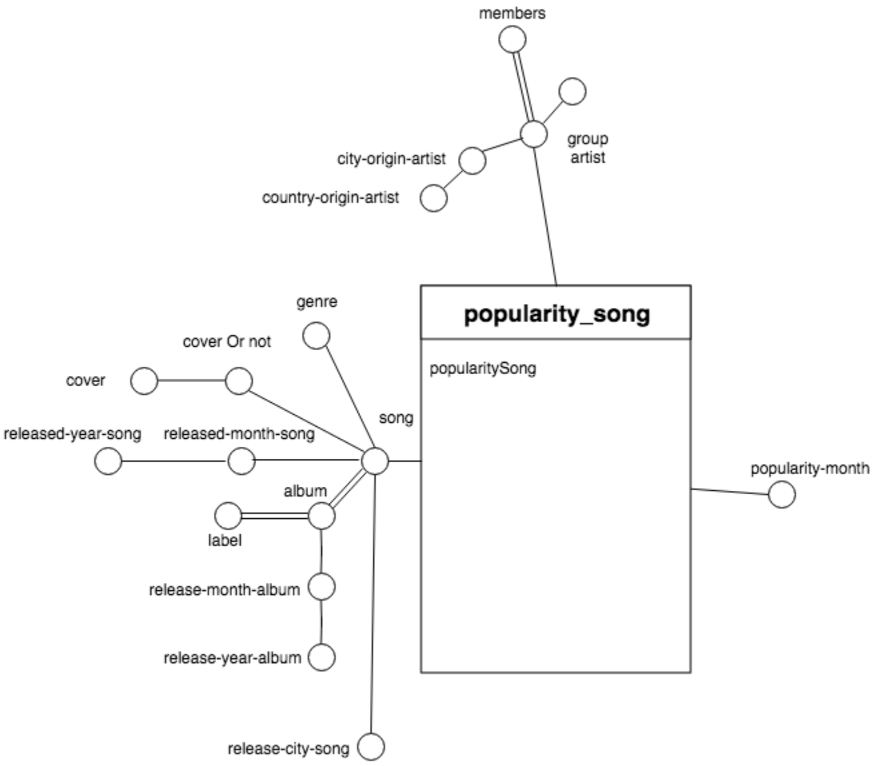


Figure 2: The Dimension Fact model for song popularity.

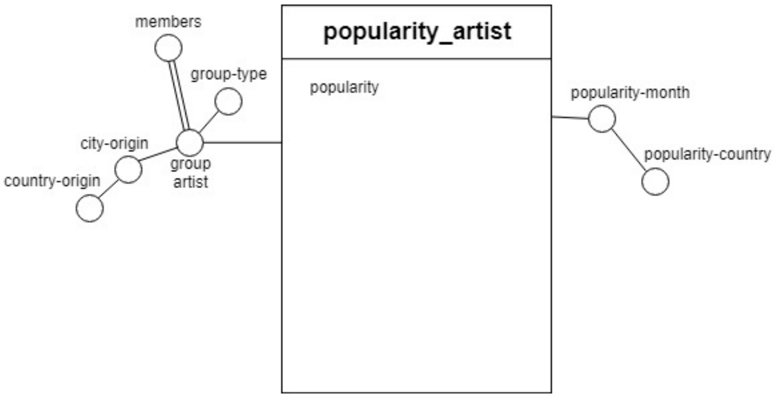


Figure 3: The Dimension Fact model for artist popularity.

4.1.2 Additivity Matrix

In the additivity matrix we identify all the aggregation functions applied on measures and grouped by dimensions .

Table 3 shows the equivalent additivity matrix for the fact schema in figure 1.

	Group-artist	Release-year-song	Release-city-song	genre	City-origin-Artist
Number.artist		sum		sum	sum
Number.releasedCover	Sum Max				
Number.releasedSong	Sum Avg Max Min	Sum Avg Max Min	Sum Avg Max Min	Sum Avg Max Min	
tempo	Avg	Avg	Avg	Avg	
energy	Avg	Avg	Avg	Avg	
danceability	Avg	Avg	Avg	Avg	
speechiness	Avg	Avg	Avg	Avg	
accoustiness	Avg	Avg	Avg	Avg	
loudness	Avg	Avg	Avg	Avg	
duration	Avg	Avg	Avg	Avg	
valence	Avg	Avg	Avg	Avg	
instrumentalness	Avg	Avg	Avg	Avg	

Table 3: Additivity matrix for Fact song.

Table 4 shows the equivalent additivity matrix for the fact schema in figure 2.

	Title-song	Popularity-month	group-artist
popularity	Min Max	Min Max	Min Max

Table 4: Additivity matrix for Fact Popularity-Song .

Table 5 shows the equivalent additivity matrix for the fact schema in figure 3.

	Group-artist	Popularity-month
popularity	Min Max	Min Max

Table 5: Additivity matrix for Fact Popularity-Artist .

4.1.3 Data dictionary

An important part of any software project is to be able to provide information in a clear and accessible way. For this purpose, to have, previously, a list with all of data variable names and description allow an efficient approach to have access to the definition of each metadata. It makes the action of manage different terminologies, formats and contents less painful for the team and users.

Attributes	Description	M	D	O
release-month-song	month of the first song release.	no	yes	no
release-year-song	year of the first song release.	no	yes	no
artist-group	the artist(s), group or lineup that the release is primarily credited.	no	yes	no
group-type	type of group artists. or artist	no	yes	no
members	members of the group.	no	yes	yes
city-origin	city of artist.	no	yes	no
country-origin	country of artist.	no	yes	no
Number.artist	a measure to count the artists or group of the album.	yes	no	no
Number.artistInterpreter	a measure to count interpreters	yes	no	no

Table 6: Data dictionary for dimensions, attributes ans measures.

Attributes	Description	M	D	O
title-song	music name or a title of a release.	no	yes	no
album	album name	no	yes	yes
label	The label which issued the release. There may be more than one; imprint, record company, music group, others	no	yes	yes
release-month-album	month of the first release of the album	no	yes	no
release-year-album	year of the first release of the album.	no	yes	no
cover	a new performance or recording by someone other.	no	yes	no
coverOrnot	a record whether a song has or not has a related cover	no	yes	no
Number.releasedCover	a measure to count the released cover	yes	no	no
release-city-song	release city location of a song	no	yes	no
release-country-song	release country location of a song	no	yes	no
genre	conventional category that identifies the music style	no	yes	no

Table 7: Data dictionary for dimensions, attributes ans measures.

Attributes	Description	M	D	O
tempo	it is a measure of speed or pace of a given piece and derives directly from the average beat duration.	yes	no	no
energy	is a measure energetic tracks that can help people feel fast, loud, noisy, and so on.	yes	no	no
danceability	describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.	yes	no	no
speechiness	detects the presence of spoken words in a track.	yes	no	no
accoustiness	float A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.	yes	no	no
loudness	it is the quality of a sound that is the primary psychological correlate of physical strength (amplitude).	yes	no	no
duration	The duration of the track.	yes	no	no
valence	a measure that describe the musical positiveness conveyed by a track.	yes	no	no
instrumentalness	predicts whether a track contains as little as possible no vocals.	yes	no	no
popularity-month	popularity of the song or the artist in the month	no	yes	no
popularity-year	popularity of the song or the artist in the year	no	yes	no
popularitySong	popularity of the song	yes	no	no
popularityArtist	popularity of the artist	yes	no	no

Table 8: Data dictionary for dimensions, attributes ans measures.

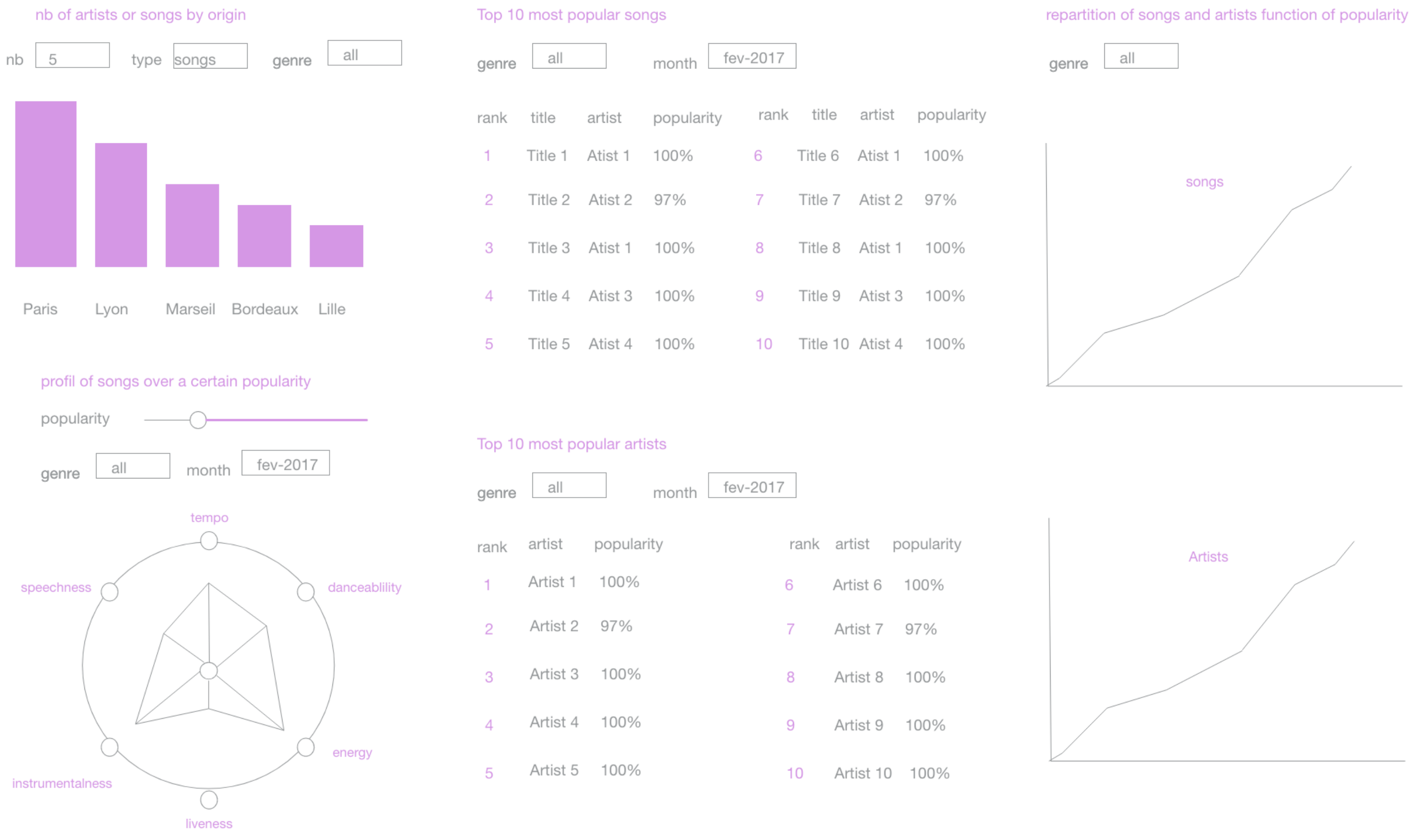
4.2 Formalization of requirements

- user need 1 :
  - SONG[city-origin].quantity
  - SONG[release-year-song].quantity
  - SONG[group artist].quantity
- user need 3 :
  - SONG[city-origin].nb-artist
  - SONG[genre].nb-artist
- user need 4 :
  - SONG[release-year-song(1);group-artist in Song[group-artist;Min(release-year-song)=realease-year-song(1)]].nb-artist
- user need 5 :
  - popularity-artist[group-artist;popularity ¿ X].quantity
  - popularity-song[title-song;popularity ¿ X].quantity
- user need 6 :
  - popularity-song[group-artist].popularity
  - popularity-song[artist].popularity
- user need 8 :
  - popularity-song[cover-or-not].popularity
  - SONG[group-artist].nb-cover-song
- user need 9 :
  - song[group-artist].max-nb-cover
- user need 10 :
  - song[artist;release-year-song = 2017].nb-songs
- user need 11 :
  - song+popularity-song[;popularity ¿ X].tempo
  - song+popularity-song[;popularity ¿ X].energy

  - ...

- aditionnal request 1 : This query calculates the difference between the popularity of artists and songs in each month in 2015.
  - popularity-artist+popularity-song[month; year = '2015'].(popularityArtist - popularitySong)
- additionnal request 2 : This query calculates the popularity of group artists in each month in 2016 and its members are six or less.
  - Popularity-artist[month, groupArtist; year='2016' groupArtist in song[year, groupArtist; year = '2016' and Number.artist ¡= 6].groupArtist].popularityArtist

4.3 Prototypes of charts



Top 10 most popular songs

genre all month fev-2017

rank	title	artist	popularity	rank	title	artist	popularity
1	Title 1	Atist 1	100%	6	Title 6	Atist 1	100%
2	Title 2	Atist 2	97%	7	Title 7	Atist 2	97%
3	Title 3	Atist 1	100%	8	Title 8	Atist 1	100%
4	Title 4	Atist 3	100%	9	Title 9	Atist 3	100%
5	Title 5	Atist 4	100%	10	Title 10	Atist 4	100%

Top 10 most popular artists

genre all month fev-2017

rank	artist	popularity	rank	artist	popularity
1	Artist 1	100%	6	Artist 6	100%
2	Artist 2	97%	7	Artist 7	97%
3	Artist 3	100%	8	Artist 8	100%
4	Artist 4	100%	9	Artist 9	100%
5	Artist 5	100%	10	Artist 10	100%

Figure 4: prototype of charts

References

[1] Coolors pattern. https://coolors.co/1a535c-20a39e-f7fff7-ff6b6b-ffe66d. accessed: 21.09.2018.

[2] Kanban. https://www.tutorialspoint.com/kanban/kanban_tutorial.pdf. accessed: 21.09.2018.

[3] Musicbrainz database and schema. https://musicbrainz.org/doc/MusicBrainz_Database/Schema. accessed: 21.09.2018.

[4] Spotify api. https://developer.spotify.com/documentation/web-api/. accessed: 21.09.2018.

[5] Trello. https://trello.com/. accessed: 21.09.2018.

[6] Christian DESTREMAU. Méthode scrum partie 2 : définition de la méthode. https://www.supinfo.com/articles/single/6054-methode-scrum-partie-2-definition-methode. accessed: 21.09.2018.

[7] Avantika Monnappa. Predicting the next big hit - big data and the music industry. https://www.simplilearn.com/big-data-science-in-music-industry-article. accessed: 21.09.2018.

[8] L.T. Moss and S. Atre. *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*. Addison-Wesley Information Technology Series. Pearson Education, 2003.

[9] Ken Schwaber and Jeff Sutherland. The scrum guide™ : The definitive guide to scrum: The rules of the game. https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf. accessed: 21.09.2018.