UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE INGENIERÍA

BASES DE DATOS NO RELACIONALES

Edición 2022

Informe de Proyecto Final

Integrantes del Grupo

Diego González Maneyro, CI: 4.139.696-9

Agustín Serra, CI: 6.441.256-2

Montevideo, Uruguay

04 de julio de 2022

1 Contenido

Contents

2 Introducción

El presente trabajo surge como presentación final de la asignatura Bases de Datos No Relacionales.

El objetivo de este trabajo es crear una base de datos de grafos en Neo4J utilizando datos obtenidos de la plataforma de streaming Spotify. Sobre esta base, se pretende realizar consultas descriptivas mediante Cypher y aplicar algoritmos de Aprendizaje Automático.

El plan de trabajo a seguir es:

- Definición y obtención de los datos a utilizar.
- Enriquecimiento de los datos: Agregar metadata a los países para ser utilizada a modo de atributo: idioma, continente.
- Diseño de la base: definir el modelo del grafo (nodos, relaciones y atributos para cada uno).
- Manipulación de los datos: mediante Python se deben realizar las transformaciones necesarias de los archivos .csv (data cruda) para contar con un input ya procesado para crear la base.
- Creación de la base en Neo4J.
- Visualización del grafo creado: haciendo uso de funcionalidades ofrecidas por AuraDB.
- Análisis descriptivo mediante Cypher: responder preguntas descriptivas utilizando dicho lenguaje de consulta.
- Aplicación de algoritmos de Aprendizaje Automático para grafos utilizando la librería de Ciencia de Datos de Neo4J: similaridad, detección de comunidades y centralidad.

3 Datos a Utilizar

Los datos a usar provienen de la empresa de servicios de multimedia llamada Spotify. La empresa ofrece una plataforma para escuchar música y podcasts, entre otras cosas. Actualmente, Spotify opera en 60 países del mundo, y en cada uno de ellos recolecta información sobre las canciones más populares en cada momento.

En base a esto, semanalmente, publica una *playlist* (lista de reproducción) llamada "Top-N", que contiene las N canciones más escuchadas de esa semana, rankeando cada canción en base a la cantidad de reproducciones dentro del país.

Para descargar la información y crear la red, se realizaron diferentes investigaciones y se verificó que la misma no se encuentra en formato de base de datos descargable ni tampoco la empresa ofrece una API abiertamente para obtenerla. Si bien existe una API de Spotify, en la misma no se ofrece la información referente al Top N de cada país.

Por lo tanto, la información de la red se obtuvo vía web-scraping con Python. Además, la web tiene ciertos controles para evitar que se pueda acceder fácilmente con librerías tales como Selenium. Debido a esto, el proceso realizado en Python fue el de simular el acceso manual y descargar uno por uno los archivos .csv de las playlist para cada país en cada semana. Se utiliza para esto la librería webbrowser de Python.

Los datos a utilizar en este caso corresponden a los de la semana comenzada el día 2021-04-09 y finalizada el día 2021-04-16. Se cuenta entonces con 60 archivos .csv distintos. En cada archivo .csv, existen 200 registros (uno por cada canción del Top 200), indicando:

- Posición en el ranking
- Nombre de canción
- Artista
- Cantidad de reproducciones semanales
- URL a la canción en Spotify

En base a estos datos, y con el enriquecimiento que se detalla en el punto siguiente, es que se procede a diseñar el esquema de la base de datos en grafos.

4 Enriquecimiento de los datos

Para poder realizar un análisis más exhaustivo en los siguientes pasos del proyecto, resulta interesante agregar información adicional a ciertos tipos de nodos.

Se decide enriquecer atributos relacionados a los nodos de países debido a que son sólo 60 en todo el grafo y buscar información relacionada a los mismos resulta más sencillo. En especial, se decide enriquecer su información con: nombre completo del país, principal idioma hablado en cada país y continente al que pertenece.

De esta manera, con esta información se pueden responder principalmente cuestiones relacionadas a la popularidad de canciones/países según continente o idioma, contando con una forma de poder realizar agrupaciones entre los países.

Por su parte, la cantidad de artistas y canciones es muy alta como para realizar un research manual de información complementaria. Además, la información contenida en el .csv de nombre de artista o de canción podría no ser suficiente como para realizar un enriquecimiento de estas clases, ya que puede haber, por ejemplo, 2 canciones distintas con el mismo nombre.

5 Diseño de la base de datos

Para el diseño de la base de datos se tienen en cuenta dos factores importantes. En primer lugar, los datos que se obtuvieron de la plataforma y segundo el tipo de análisis que se busca llevar a cabo. Se sugiere de esta manera la siguiente configuración de los datos en nodos y relaciones.

5.1 Nodos

Se proponen los siguientes tipos de nodos:

- ***Song:*** este nodo es el que contiene el nombre de la canción y la url a la canción en Spotify.
- ***Artist:*** este nodo contiene el nombre del artista.
- ***Country:*** este nodo contiene el nombre del país e información adicional que se agrega en la etapa de enriquecimiento de los datos.

5.2 Relaciones

Se proponen los siguientes tipos de relaciones entre los nodos:

- ***PERFORMS:*** Esta relación une al artista y la canción. La dirección es desde un nodo artista a un nodo canción.
- ***INCLUDES_IN_TOP:*** Esta relación vincula la canción con el país donde figura en el ranking. A su vez, la relación contiene los atributos de posición en el ranking y cantidad de reproducciones. La dirección es desde el país a la canción.

5.3 Representación

La estructura del modelo sugerida se puede visualizar en la siguiente figura.

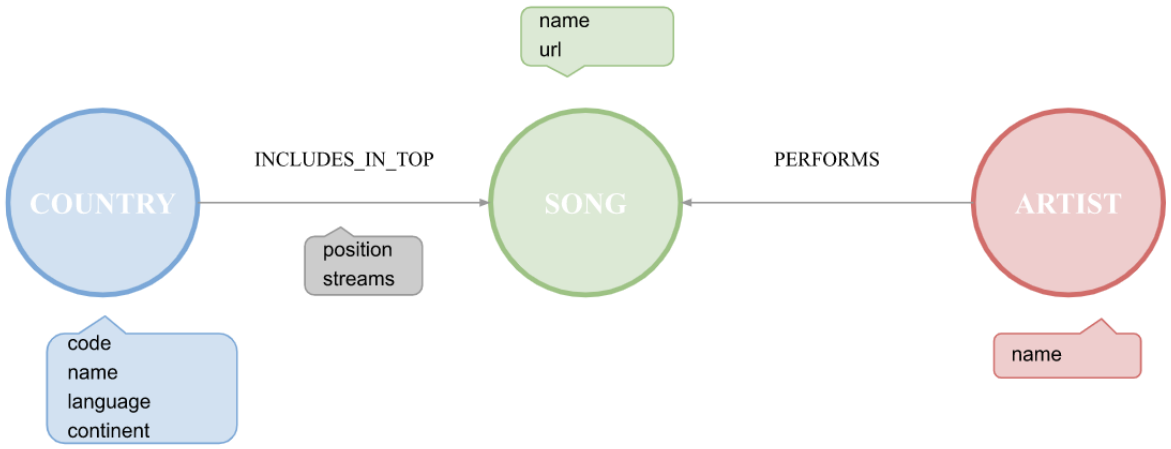

Figure 2: Estructura de los nodos y relaciones sugeridas para el desarrollo del proyecto

Pese a que no está en el alcance de este proyecto, nótese que la estructura sugerida permite fácilmente agregar canciones, artistas y países para poder contemplar los cambios de rankings semanales se dan naturalmente en el rubro de la música.

6 Manipulación de datos

En esta etapa se busca editar los archivos que se tiene inicialmente de forma que se pueda crear la base en Neo4j de manera más práctica. Para esto se realiza una primera instancia exploratoria, donde se busca entender cómo debe ser la estructura de los archivos .csv para poder llevar adelante la arquitectura diseñada en el paso anterior.

6.1 Datos iniciales

A continuación se describen los archivos .csv iniciales con los que se trabajó.

countries_data.csv: archivo resultante del enriquecimiento mencionado previamente en el plan de trabajo. Contiene los campos:

- code: código de país
- name: nombre de país

- language: lengua principal hablada en el país
- continent: continente del país

  A continuación se dejan las primeras filas de ese archivo:

codigo	name	language	continent
ar	Argentina	Spanish	South America
at	Austria	German	Europe
au	Australia	English	Oceania
be	Belgian	Dutch	Europe
bg	Bulgaria	Bulgarian	Europe

Table 1: Primeras líneas de country_data.csv

regional-[countryCode]-weekly-2021-04-09–202104-16.csv: son 60 archivos iniciales de cada país para el top de la semana escogida. Cada archivo contiene los siguientes campos:

- Position: posición de la canción en el Top de Spotify para esa semana
- Track Name: nombre de la canción
- Artist: artista o banda que interpreta la canción
- Streams: cantidad de reproducciones de esa semana
- URL: Link de Spotify en donde se puede escuchar la canción en cuestión

A continuación se dejan de ejemplo las primeras 5 filas del archivo correspondiente a Uruguay (uy):

Position	Track Name	Artist	Streams	URL
1	WACHA	KHEA	134870	`https://open.spotify.com/track/5RmSunLqeDqNWww8LWhAiK`
2	L-Gante: Bzrp Music Sessions, Vol.38	Bizarrap	134321	`https://open.spotify.com/track/1Crj1zkRMpsEjb9NOR6Zof`
3	Fiel	Los Legendarios	121450	`https://open.spotify.com/track/7Bk0uXKk1uPT0XuQbpFzvs`
4	Además de Mí - Remix	Rusherking	113140	`https://open.spotify.com/track/7I8L3vYCLThw2FDrE6LuzE`
5	Bandido	Myke Towers	106854	`https://open.spotify.com/track/1xK1Gg9SxG8fy2Ya373oqb`

Table 2: Primeras líneas de archivo de canciones en Uruguay

6.2 Datos manipulados

Se analizaron diferentes caminos de manipulación de los datos iniciales para el momento de la carga de datos en Neo4J. En especial, la principal preocupación a considerar fue el hecho de tener canciones y artistas repetidos para los distintos países.

Por este motivo, se exploraron comandos en Neo4J que eviten realizar esta duplicación y se encontraron alternativas que se explicarán en la siguiente sección de este trabajo. Debido a esto, sólo fue necesario concatenar los 60 archivos de los países y agregar el código de cada país en las relaciones. Es decir, los archivos finales con los que se realizó la carga fueron:

- countries_data.csv: sin modificaciones
- relations_data.csv: concatena los archivos de los 60 países, agregando para cada país el código del mismo para relacionar con countries_data.csv. Se deja a continuación las primeras 5 filas de este archivo:

Position	Track Name	Artist	Streams	URL	country_code
1	L-Gante: Bzrp Music Sessions, Vol.38	Bizarrap	2380129	`https://open.spotify.com/track/1Crj1zkRMpsEjb9NOR6Zof`	ar
2	WACHA	KHEA	2374445	`https://open.spotify.com/track/5RmSunLqeDqNWww8LWhAiK`	ar
3	Además de Mí - Remix	Rusherking	2292324	`https://open.spotify.com/track/7I8L3vYCLThw2FDrE6LuzE`	ar
4	Fiel	Los Legendarios	2010410	`https://open.spotify.com/track/7Bk0uXKk1uPT0XuQbpFzvs`	ar
5	Ella No Es Tuya - Remix	Rochy RD	1640336	`https://open.spotify.com/track/5YYW3yRktprLRr47WK219Y`	ar

Table 3: Primeras líneas de archivo manipulado

Es importante destacar el hecho que para cada país sólo se cargaron sus 50 canciones más populares, para evitar relaciones con clásicos históricos (por ejemplo: Let it be de The beatles) que ocasionarían que el grafo resultante sea demasiado denso y no representante de los principales relaciones contemporáneas en cada país.

Por esta última razón, el archivos relations_data.csv cuenta entonces con 3000 filas (50 canciones por cada uno de los 60 países). Por su parte, countries_data.csv cuenta con 60 filas (una para cada país).

Los datos tanto originales como procesados se encuentran en la siguiente carpeta compartida de Google Drive.

Los notebooks y la base .dump se encuentran en la siguiente carpeta compartida de Google Drive.

El notebook de manipulación de los datos se llama *1.data_ manipulation.ipynb*

7 Creación de la base en Neo4j

Para la creación de la base en Neo4j se cuenta con las siguientes tres opciones:

- Instalar localmente en el pc utilizando Neo4j Desktop
- Utilizar en la nube Neo4j Sandbox
- Utilizar en la nube Neo4j AuraDB

En nuestro caso optamos por utilizar la tercera opción Neo4j AuraDB debido a que la herramienta fue utilizada previamente para un entregable anterior de la materia.

Dentro de esa opción, existen dos alternativas para cargar datos de un .csv y crear la base: mediante Cypher con la opción LOAD CSV o bien con el Neo4J Data Importer (una aplicación visual lanzada desde la consola).

Para continuar desarrollando el trabajo utilizando Cypher, se opta por la primera opción. Más información sobre ambas alternativas puede encontrarse en la documentación oficial de Neo4J.

Los pasos a seguir para crear la base son:

- Crear una nueva instancia en AuraDB (gratis).

Para esto, se reutilizan los datos de usuario ya gestionados para las entregas previas. Con ellos, se crea una nueva instancia.

- Disponibilizar los archivos .CSV

Debido a razones de seguridad, Neo4J no permite cargar archivos .csv locales. Según su documentación oficial, los archivos a cargar deben estar accesibles públicamente en servidores HTTP o HTTPS como GitHub, Google Drive o Dropbox. Otra alternativa es que cargar los mismos a un almacenamiento en la nube del tipo bucket (como pueden ser Google Cloud Storage - GCP - o Amazon S3) y luego configurar dicho bucket para que se pueda acceder al mismo (como un sitio web estático).

En este trabajo, se opta por dejar disponibles los archivos .csv en un Google Drive, ya que el mismo fue utilizado para entregas de tareas previas. Se deja a continuación el link al Google Drive en donde se encuentran los archivos. El mismo, por las razones explicadas en el párrafo anterior, está configurado para ser de acceso público.

- Creación de la base mediante Cypher

La creación de la base se desarrolla en un notebook titulado *2.database_ creation.ipynb*.
La librería principal utilizada en este punto es "*graphdatascience*". Esta librería se utiliza en varios ejemplos provistos por Neo4J en su documentación oficial. Se deja el link al repositorio oficial de graphdatasciente.

Los pasos seguidos para la creación se describen a continuación:

- **Conexión a la instancia.** creada en el primer punto.
- **Creación de *constraints*.** Antes de la carga de los archivos en sí, es necesario crear ciertas restricciones (constraints). Las restricciones en este caso son de unicidad para los campos de países y canciones.

En otras palabras, no se quiere permitir que se repitan dos nodos que representan al mismo país o canción. Resulta interesante resaltar el hecho que para estos dos elementos, se tiene un código de identificación único: el country_code para el país y la URL de Spotify para la canción.

No sería correcto establecer que el ID de la canción fuera su nombre, debido a que podrían existir dos canciones con el mismo nombre.

Se deja a continuación el código con el cual se realiza esta creación de constraints para ambos tipos de nodos, respectivamente:

```
# Código único para países
gds.run_cypher("""
CREATE CONSTRAINT countryCodeContraint FOR (country:Country) REQUIRE
country.code IS UNIQUE
""")
```

```
# URL única para canciones
gds.run_cypher("""
CREATE CONSTRAINT songUrlConstraint FOR (song:Song) REQUIRE song.url IS
UNIQUE
""")
```

Es importante destacar que este paso se realiza antes de poblar la base con los archivos en sí.

- **Carga de archivos**: Se realizan dos ejecuciones distintas, una para cada uno de los archivos .csv que resultaron de la manipulación de datos: *countries_data.csv* y *relations_data.csv*

La primera ejecución, destinada a cargar los países en la base de datos, es más sencilla que la segunda, debido a que la base de datos aún está vacía (por lo tanto no se pueden encontrar nodos ya existentes con los cuales haya conflicto), y además que el archivo .csv de los países no tiene elementos repetidos. Esto permite entonces hacer uso del comando CREATE de Cypher.

Un factor importante a destacar es que para que se puedan leer los archivos de Google Drive se debió obtener el link de descarga de los mismos (y no el link de compartir los mismos, como se podría suponer a priori). De hecho se realizaron varias pruebas fallidas de carga hasta dar con la solución en una consulta de StackOverflow sobre carga de archivos de Google Drive mediante Cypher.

El código utilizado para la carga de los países se deja a continuación:

```
# Carga de csv de países
gds.run_cypher("""
LOAD CSV WITH HEADERS FROM
"https://drive.google.com/u/0/uc?id=1ui0s2
KAS146vgcZn5Pb2H_ukknjKtkZH&export=download"
AS csvLine
CREATE (c:Country {code: csvLine.code, name: csvLine.name, language: csv-
Line.language, continent: csvLine.continent})
""")
```

Se puede observar cómo con el comando LOAD CSV se carga el archivo deseado y luego se recorre cada línea del mismo (csvLine), de la cual se crea un nodo del tipo Country con atributos code, name, language y continent, sin realizar transformación alguna sobre los nombres o campos originales en la carga.

En segundo lugar, se deben cargar los nodos de canciones, artistas, relaciones entre sí y con los países. Como se explicó previamente, se debe tener consideración para ciertos campos que pueden repetirse en las diferentes líneas del archivo. Por ejemplo, observando las filas de los archivos .csv provistas en la sección de Manipulación de Datos,: se debe cargar sólo una vez la relación que indica que la canción “Fiel" es interpretada por “Los Legendarios", sin importar que esta canción aparezca en muchos países a la vez, y por ende en muchas líneas del archivo relations_data.csv.

Debido a esto es que el comando CREATE es reemplazado por el comando MERGE de Cypher. El mismo lo que hace es verificar si no existe el nodo o relación ya previamente en la base. En caso de existir, no crea nada. Caso contrario, crea el nodo o relación.

Finalmente, se debe realizar un MATCH con los nodos de países para relacionar los países del archivo csv que se está cargando en esta ejecución y los nodos ya existentes en la base por la carga del otro archivo. Luego, se crea la relación entre canción y país. En este punto, no hace falta utilizar MERGE debido a que se sabe que cada relación de canción-país aparece una sóla vez en la base.

Adicionalmente, se realizan en esta carga ciertas transformaciones sobre campos como por ejemplo cambiar el nombre de “Track Name" del archivo original para las canciones por “name" dentro de la base e Neo4J, o bien aplicar la función “toInteger" de Cypher para los campos Streams y Position. Esto último es necesario debido a que Cypher, con el comando LOAD CSV aquí empleado, carga todos los datos como texto a menos que se indique lo contrario.

Se deja a continuación el código utilizado en donde se pueden observar las consideraciones antes aclaradas:

```
# Carga de csv de relaciones
gds.run_cypher("""
LOAD CSV WITH HEADERS FROM "https://drive.google.com/u/0/uc?id=1-
1n
Pqo9aP7r4h0KH8jS3oCK-a2KFRqM8&export=download"
AS csvLine
MERGE (artist:Artist {name: csvLine.Artist})
MERGE (song:Song {url: csvLine.URL, name: csvLine.‘Track Name‘})
MERGE (artist)-[:PERFORMS]->(song)
WITH song, artist, csvLine
MATCH (song:Song {url: csvLine.URL}), (country:Country {code:
csvLine.country_code})
CREATE (country)-[:INCLUDES_IN_TOP {streams: toInte-
ger(csvLine.Streams), position: toInteger(csvLine.Position)}]->(song)
""")
```

- Validación de carga: Para validar la carga, se realiza una consulta descriptiva aleatoria de prueba. Una buena práctica es inspeccionar todos los nodos y relaciones aplicando el siguiente comando sencillo:

```
# Inspeccion general
gds.run_cypher("""
MATCH (n)-[r]->(m) RETURN n, r, m
LIMIT 100
""")
```

Sin embargo, resulta más útil e interesante realizar una primera inspección de la base con la interfaz provista por Neo4J AuraDB.

- Descarga de la base a nivel local: AuraDB ofrece la posibilidad de descargar el archivo .dump de la base creada. El mismo se entrega acompañando el presente informe con el nombre *spotify_graph_neo4j.dump*.

8 Visualización del grafo creado mediante AuraDB

Mediante la funcionalidad de Inspeccionar una instancia que ofrece AuraDB, se puede realizar una visualización del grafo creado en el inciso anterior. La misma permite aplicar filtros y es de uso muy sencillo para contar con una primera idea de la base con la que se está trabajando.

En primer lugar, con esta herramienta, se puede observar que la base está compuesta por:

- 1888 nodos
  - 60 países
  - 749 artistas

  - 1079 canciones

- 4079 relaciones
  - 3000 de país-canción
  - 1079 de canción-artista

Se dejan 2 capturas de pantalla de dicha interfaz:

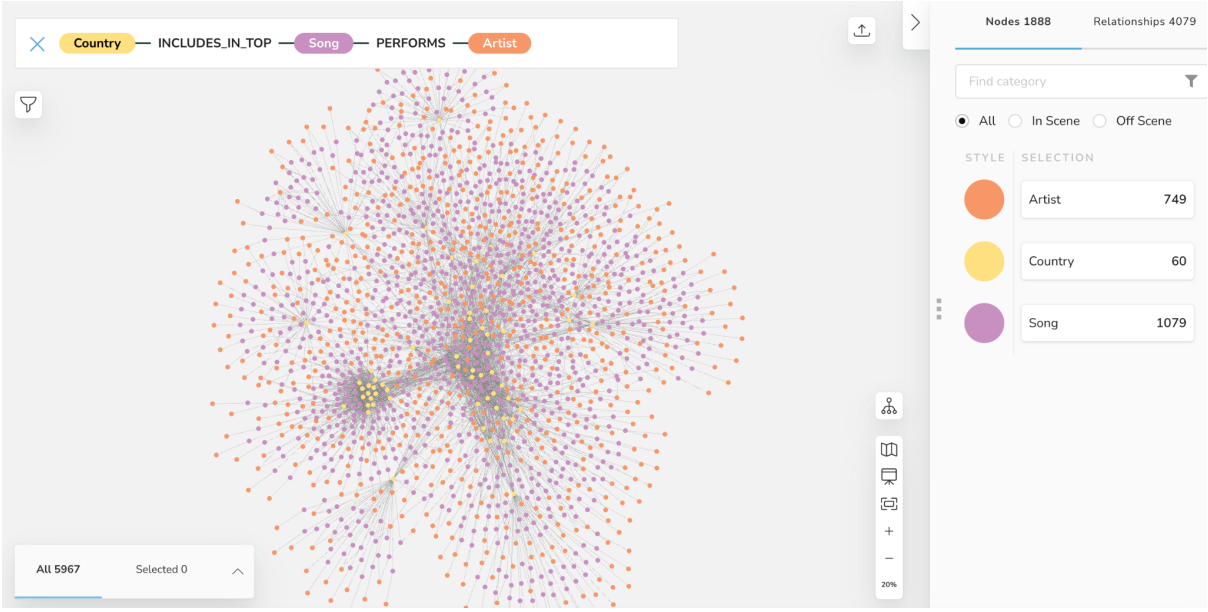

Figure 3: Red general

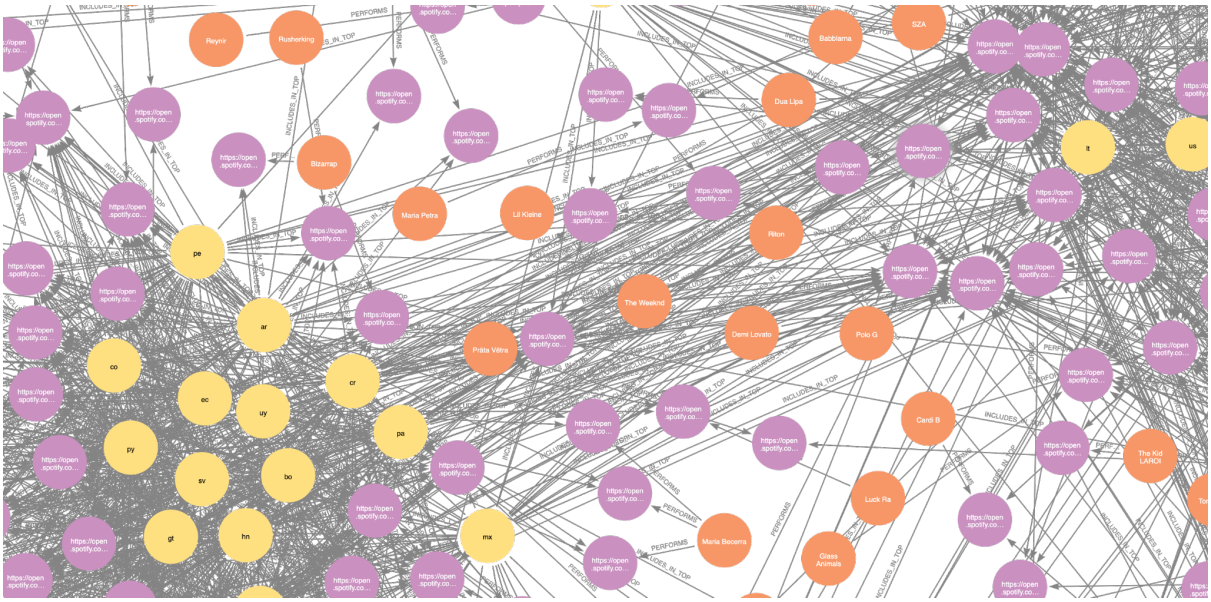

Figure 4: Fragmento de parte de red

9 Análisis descriptivo

Se realiza un análisis descriptivo para obtener información a partir de los datos almacenados. El desarrollo de este análisis se lleva a cabo en el notebook *3.descriptive_ analysis.ipynb*, donde se realiza la conexión a la base y las consultas correspondientes a las distintas preguntas que se quieren responder.

9.0.1 Artista más escuchado

Es interesante conocer en un dataset de reproducciones de canciones quién es el artista más escuchado. Se busca enriquecer ese dato consultando quién es el artista más escuchado por país, continente y por idioma. Se interpretan los ranking obtenidos en busca de darle un sentido a los datos, así como descubrir particularidades.
En la siguiente tabla se muestra el top 10 de artistas con más reproducciones, donde se puede ver que la artista Taylor Swift lidera el ranking.

Ranking	Artista	Reproducciones
1	Taylor Swift	76784978
2	Justin Bieber	66527313
3	Lil Nas X	52384697
4	Olivia Rodrigo	33681250
5	Polo G	33340577
6	The Weeknd	32939973
7	Bad Bunny	28677459
8	Doja Cat	25395382
9	Masked Wolf	22215968
10	Dua Lipa	20762111

Table 4: Top 10 artistas con más canciones

Lo siguiente que nos interesa ver es qué artistas predominan las reproducciones por país, limitando el resultado a los 10 con más reproducciones ordenados de manera descendente.

País	Artista	Reproducciones
United States of America	Taylor Swift	55.726.424
Italy	Guè Pequeno	20.495.149
Philippines	Taylor Swift	13.081.530
Brazil	Os Barões Da Pisadinha	11.762.722
Japan	YOASOBI	9.204.358
México	Bad Bunny	8.742.607
Indonesia	Pamungkas	6.031.483
Germany	Luciano	5.067.940
Canada	Justin Bieber	4.457.907
France	SCH	4.425.091

Table 5: Artistas con más reproducciones por país, limitado a 10 países.

Es interesante ver que la diferencia entre reproducciones entre Estados Unidos y los demás países es considerable, se ve claramente dado que tiene casi tres veces más que el segundo en el ranking. Un dato interesante que se obtiene de estas dos consultas es que el artista que predomine las reproducciones en Estados Unidos (en este caso Taylor Swift), es sin dudas el que predomine el ranking total, mostrando que la plataforma Spotify tiene una gran adopción en ese país, sumado a su gran población.
La siguiente tabla muestra el resultado del ranking por continente

Continente	Artista	Reproducciones
North America	Taylor Swift	59.102.230
Europe	Justin Bieber	25.048.101
Asia	Taylor Swift	14.029.388
South America	Os Barões Da Pisadinha	11.762.722
Oceanía	Justin Bieber	4.565.893
Central America	Bad Bunny	2.639.379

Table 6: Top artistas por continente

El resultado nos muestra que tanto en América del Norte como en Asia, Taylor Swift lidera el ranking. Tiene sentido que sea la artista con más reproducciones en total, dado que lidera dos de los tres continentes con más reproducciones.

Por otro lado, se observa que Justin Bieber lidera el ranking en dos continentes también y como dato interesante es que en Europa es el artista con más reproducciones, pero no lidera el ranking en ninguno de los tres países con más reproducciones: Italia, Alemania y Francia. Esto deja ver que sus reproducciones están bastante distribuidas por el mundo y no están concentradas en un país en particular.

Finalmente, es interesante ver que todas las reproducciones de América del Sur de Os Baroes Da Pisadinha fueron en Brasil, dejando entrever que el artista tiene un gran alcance nacional pero no internacional. Al igual que en el caso de las mayores reproducciones en Estados Unidos, es importante tener en cuenta la influencia que podría tener la población del país respecto a los otros del continente.

Cómo un último desglose de la información de reproducciones por artista, se busca obtener información del ranking de artistas, pero agrupando por idioma.

Idioma	Artista	Reproducciones
English	Taylor Swift	62.755.590
Spanish	Bad Bunny	23.671.092
Italian	Guè Pequeno	20.495.149
Tagalog	Taylor Swift	13.081.530
Portuguese	Os Barões Da Pisadinha	11.762.722

Table 7: Top artistas por idioma

Se puede ver en el ranking que el inglés y el español son los idiomas de los países con más reproducciones. Esto tiene sentido ya que son los dos lenguajes más hablados entre los países de la base con 17 países hispanoparlantes y 6 países angloparlantes.

9.0.2 Canción más escuchada

Teniendo un panorama general de los artistas más escuchados y su influencia en los distintos países y continentes, ahora se buscará información sobre qué canciones son las más escuchadas y qué impacto tienen estas para el artista. También es interesante saber qué alcance tienen las canciones, viendo por ejemplo qué canciones están en el top 5 de más de un país.

Canciones con más reproducciones

Ranking	Artista	Canción	Reproducciones
1	Lil Nas X	MONTERO (Call Me By Your Name)	52.033.699
2	Justin Bieber	Peaches (feat. Daniel Caesar & Giveon)	46.907.779
3	Polo G	RAPSTAR	29.112.118
4	Masked Wolf	Astronaut In The Ocean	22.215.968
5	Bruno Mars	Leave The Door Open	19.560.222
6	The Weeknd	Save Your Tears	19.494.111
7	Los Legendarios	Fiel	18.777.480
8	Olivia Rodrigo	drivers license	18.217.291
9	Doja Cat	Kiss Me More (feat. SZA)	17.487.707
10	Giveon	Heartbreak Anniversary	16.804.207

Table 8: Canciones más escuchadas

Viendo el resultado de esta consulta e interpretando la información de los artistas más escuchados, se destacan dos cosas. En primer lugar, que no figura ninguna canción de Taylor Swift en el ranking, siendo esta la artista con más reproducciones totales. Profundizando en sus reproducciones por canción se puede ver que tiene 17 canciones distintas en 10 países y que la canción con más reproducciones es Mr.Perfectly Fine con 11.7M.

Por el otro lado, si se toma al artista Lil Nas X de la canción más escuchada con 52M, se puede ver que tiene centradas sus reproducciones en la canción Montero (99.3% de las reproducciones del artista) en 59 de los 60 países. Similar es el caso de Justin Bieber, que tiene 6 canciones reproducidas en 58 países, siendo la más reproducida Peaches con 47M.

Se desprende de este resultado que algunos artistas tienen pocas canciones con gran impacto a nivel global y otros muchas canciones con menor impacto pero que en su conjunto generan muchas reproducciones.

Surge de esto analizar el alcance de las canciones y su impacto, para esto se realiza la consulta de canciones en el top 5 de al menos 10 países.

Artista	Canción	Qty de países
Justin Bieber	Peaches (feat. Daniel Caesar & Giveon)	35
Lil Nas X	MONTERO (Call Me By Your Name)	32
Masked Wolf	Astronaut In The Ocean	19
Los Legendarios	Fiel	17
Myke Towers	Bandido	15
Polo G	RAPSTAR	15
Bad Bunny	DÁKITI	12
Sech	911	10
Bad Bunny	LA NOCHE DE ANOCHE	10
Riton	Friday (feat. Mufasa & Hypeman) - Dopamine Re-...	10

Table 9: Canciones en el top 5 de al menos 10 países

Se puede corroborar con esta tabla el planteo del impacto masivo de canciones particulares. Como se puede ver, tanto Peaches como Montero tienen un gran impacto en muchos países, reflejándose en la posición del ranking.

10 Algoritmos de aprendizaje automático

Para aplicar los algoritmos de aprendizaje automático fue necesario crear una nueva instancia de la base, debido a que en AuraDB no se puede aplicar la librería *graphdatascience* con sus algoritmos. Para eso, se necesita tener una base creada en AuraDS, la cual actualmente tiene un costo.

Por esta razón, se creó un Neo4J Sandbox, el cual ofrece una instancia gratuita por un período de días determinados.

Los comandos de Cypher para crear la base son los mismos que los utilizados al crear la primera base en AuraDB.
El trabajo de esta sección se encuentra en el notebook llamado *4.graph_ data_ science.ipynb*

Los algoritmos aplicados son 3:

10.1 Similaridad

El algoritmo de similaridad compara un set de nodos basándose en los nodos a los cuales están conectados. Dos nodos se consideran similares si comparten muchos vecinos entre sí. El algoritmo de similaridad computa similaridades de a pares basándose en la métrica de Jaccard.

Dados dos sets A y B, la fórmula de Jaccard se computa de la siguiente manera:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

El algoritmo se aplica a los países, ponderando las relaciones por posición en el top de cada país. Los resultados son los siguientes:

- Países más similares entre sí:

1. Colombia - Ecuador (score 0.5828)
2. El Salvador - Honduras (score 0.5653)
3. Honduras - Nicaragua (score 0.5178)

- Países menos similares entre sí:

1. Japan - Peru (score 0.0007)
2. Japan - Thailand (score 0.0031)
3. Japan - Hong Kong (score 0.0039)

Es interesante destacar cómo los países de latinos (del norte de Sudamérica y Centroamérica) son los más similares, mientras que Japón parece ser el país más aislado. Esto podrá corroborarse aplicando el algoritmo de detección de comunidades.

10.2 Detección de comunidades

Para realizar la detección de comunidades, la librería graphdatascience se basa en el algoritmo de Louvian, descrito en Lu, Hao, Mahantesh Halappanavar, and Ananth Kalyanaraman "Parallel heuristics for scalable community detection".

Este algoritmo maximiza un score de modularidad para cada comunidad, en donde la modularidad cuantifica la calidad de asignación de nodos a comunidades. Esto significa evaluar qué tan conectados están los nodos dentro de una comunidad, comparados con qué tan conectados estarían en una red aleatoria.

El algoritmo tal como está implementado no permite distinguir entre distintos tipos de nodos. Debido a esto, al ejecutarlo sobre el grafo de este trabajo, dentro de cada comunidad pueden aparecer nodos de los tres tipos: países, canciones y artistas.

Se ejecuta el algoritmo y quedan agrupados los 1888 nodos iniciales en 246 comunidades distintas. La más grande de ellas contiene 246 nodos, mientras que las más pequeñas sólo están compuestas por 2 nodos.

Se analiza cómo quedan compuestas algunas comunidades en especial.

- Comunidad en donde aparece el artista “Migrantes", banda perteneciente a la música del género tropical/cumbia. Este clúster está compuesto por 238 nodos entre países, artistas y canciones.
  - Países: En este clúster quedan todos países de habla hispana: incluyendo países Latinoamericanos como Uruguay, Ecuador, Argentina, México, Honduras, entre muchos otros, así como también España. Se observa entonces que el lenguaje de un país podría tener influencia fuerte respecto al clúster final en el cual cada país es clasificado.

  - Artistas: Se observa que la mayoría de los artistas que componen la comunidad realizan canciones de género similar a la banda en cuestión, performando principalmente canciones del género cumbia, reggaetón o trap. Por destacar algunos, se encuentran Bizarrap, KHEA, Maria Becerra, entre otros.
  - Canciones: Como es esperable, las canciones también son del mismo género y principalmente en idioma español.

- Comunidad en donde aparece el país "Japón". Resulta interesante estudiar esta comunidad, ya que Japón, según los algoritmos de similaridad estudiados previamente, parece ser el país menos parecido al resto.

  - Países: Un dato interesante es que Japón es el único país que integra el clúster, reforzando la idea de que su música es muy distinta a la del resto de los países del mundo.
  - Artistas: Entre los artistas de la comunidad aparecen (cantautor japonés y youtuber), YOASOBI (dúo japonés de música pop), y Awesome City Club (banda japonesa que interpreta el estilo J-POP). A simple vista, se puede destacar cómo Japón parece ser un gran consumidor de artistas totalmente locales y que no son consumidos por otros países en el mundo.
  - Canciones: Algunas de las canciones que integran el clúster son , y . Siguiendo sus URL de Spotify se verifica que todas pertenecen al género pop japonés. El hecho de que estas canciones sean en lengua japonesa puede ser una explicación de que las mismas no se escuchen en otros países, al no estar el idioma japonés tan esparcido en el resto del mundo como otras lenguas tales como el inglés o el español.

- Comunidad que esté compuesta sólo por 2 nodos. Se elige de manera aleatoria una de las comunidades más chicas para analizar qué la compone.

  - Países: la comunidad no está compuesta por ningún país.
  - Artistas: Incluye al nodo de artista Nemazalány, el cual es un dúo femenino de origen húngaro
  - Canciones: La única canción que incluye esta comunidad es Üres szívek. Como es de esperar, esta canción es interpretada por el dúo Nemazalány. Se observa nuevamente en esta comunidad cómo las cuestiones de lengua pueden tener gran influencia sobre la conformación de los clústers.

10.3 Centralidad

La centralidad de intermediación es una manera de detectar la cantidad de información que un nodo tiene sobre todo el flujo de información del grafo. Sirve generalmente para identificar qué nodos actúan como una especie de "puente" entre una parte del grafo y la otra.

El algoritmo calcula caminos más cortos (sin peso) entre todos los pares de nodos del grafo. Cada nodo recibe un *score* dependiendo de todos los caminos más cortos de los cuales forma parte. Aquellos nodos que estén en mayor parte de aquellos caminos recibirán un *score* mayor que los que no lo hacen.

La implementación de la librería *graphdatascience* se basa en el Algoritmo de aproximación de Brandes para grafos sin peso.

El algoritmo no tolera grafos heterogéneos, por lo que la centralidad se calcula para todos los nodos como si fueran del mismo tipo. Luego de aplicar el algoritmo, se separan entonces los distintos tipos de nodos: artistas, países y canciones.

Se dejan a continuación los resultados de centralidad para los 3 tipos de nodos:

10.3.1 Artistas

- Más centrales

Nombre	Score
BTS	917.876142
Justin Bieber	263.047886
Bruno Mars	204.116162
Taylor Swift	197.368383
Bad Bunny	166.347647

Table 10: Artistas más centrales

- Menos centrales

Nombre	Score
nân	0.0
Tlinh	0.0
Hoàng Yn Chibi	0.0
Thiu Bo Trâm	0.0
hooligan.	0.0

Table 11: Artistas menos centrales

Notar cómo los artistas más centrales tienen correlato con lo analizado en la sección descriptiva de esta entrega.

10.3.2 Canciones

- Más centrales

Nombre	Score
MONTERO (Call Me By Your Name)	327430.549427
Peaches (feat. Daniel Caesar & Giveon)	262152.618111
Astronaut In The Ocean	129976.662471
Save Your Tears	118738.030218
Leave The Door Open	79507.357695

Table 12: Canciones más centrales

- Menos centrales

Nombre	Score
Si Estuviésemos Juntos	20.837726
Break My Heart	19.874588
Stay Gold	18.500000
A Tu Merced	18.226098
Si Veo a Tu Mamá	18.226098

Table 13: Canciones menos centrales

Nuevamente las canciones más centrales tienen sentido con lo analizado en la parte descriptiva de esta entrega, donde se puede observar en el ranking de canciones en el top 5 de al menos 10 países. Si bien el órden no es necesariamente el mismo, las canciones se repiten.

10.3.3 Países

- Más centrales

Nombre	Score
Turkey	140983.090433
Japan	135902.376142
Vietnam	135417.015389
Brazil	135336.176096
Italy	124378.939993

Table 14: Países más centrales

- Menos centrales

Nombre	Score
Canada	15514.658789
Bolivia, Plurinational State of	13390.868888
El Salvador	12999.033556
Guatemala	12625.131415
Ecuador	11217.922637

Table 15: Países menos centrales

Dentro de los países más centrales, se destacan los que tienen una cultura nacional muy fuerte pero que mantienen un vínculo con la música del exterior. Esto hace que de no estar esos nodos, se pierde una conexión de los demás países con la música del país “retirado".

Esto se puede ver fácilmente con el grupo Os Baroes Da Pisadinha de Brasil y con el cantante Guè Pequeno de Italia. Estos tienen una cantidad de reproducciones muy alta pero solo dentro del país y a esas reproducciones sólo se accede desde el nodo del país. También se ve reflejado en los clústers para el caso de Japón, donde todas las canciones del clúster son de procedencia japonesa.

11 Conclusiones

En este trabajo se analizaron diversas cuestiones relacionadas con la cultura musical del mundo, o al menos 60 países del mundo en donde la plataforma Spotify opera.
Según los análisis realizados, se puede concluir que:

- Existe una comunidad muy marcada respecto al idioma español: incluyendo países latinoamericanos y España. Las canciones y artistas que lo integran son de género tropical, trap, y ritmos del estilo.
- Japón es el país más distinto al resto, priorizando canciones y artistas locales.
- Taylor Swift es la artista más escuchada en el mundo durante la semana analizada, sin embargo no tiene una canción que destaque dentro de las más escuchadas del mundo.
- Artistas como Justin Bieber y The Weeknd no sólo integran el top de los artistas más escuchados, sino también los artistas más centrales para el grafo.
- Existen comunidades pequeñas de nodos totalmente alejados del resto del grafo. Por ejemplo: la banda húngara Nemazalány y una canción de su autoría integran una comunidad propia.

Los análisis aquí realizados son sólo una primera exploración del grafo, pero la base contiene suficiente información como para profundizar en muchas otras líneas de investigación, ya sea tanto descriptiva como predictiva.