



Proyecto Gestión de Datos - Spotify

Juanita Caballero Villalobos^{a,c}, Daniel Hamilton-Smith Santa Cruz 2^{a,c}

Fabian Camilo Peña^{b,c}

^a Estudiante de Maestría en inteligencia artificial

^b Profesor, Gestión de datos, Departamento de Ingeniería Industrial

^c Pontificia Universidad Javeriana, Bogotá, Colombia

Secciones del documento

1. Descripción general de cada dataset

a. Artists_mod.csv

b. tracks_mod.csv

2. Análisis de calidad de los datos y procesos de limpieza implementados

3. Análisis exploratorio de datos

4. Modelo de datos relacional

5. Esquema general de la bodega datos

6. Arquitectura del ETL

7. Arquitectura general de la solución

8. Descripción del sistema de recomendación

9. Descripción del dashboard

1. Descripción general de cada dataset

Incluya aspectos como dimensiones y diccionario de datos de las variables que considera más representativas. No olvide incluir el tipo de variable identificado.

a. Artists_mod.csv

NOMBRE	TIPO	RANGO
id	string	
followers	int64	entero positivo
genres	string	contiene un string que representa una lista de géneros
name	string	

Tabla 1. Descripción de variables, su tipo y rango en el dataset de artist

b. tracks_mod.csv

NOMBRE	TIPO	RANGO
id	string	
name	string	
popularity	int64	entero: 0 .. 100
duration_ms	int64	entero: positivo

explicit	int64	bool: 0 o 1
artists	object	string que contiene una lista
id_artists	object	string que contiene una lista
release_date	date	incluye año, opcionalmente puede incluir mes y día
danceability	float64	float: 0 .. 1
energy	float64	float: 0 .. 1
key	float64	entero: -1 .. 11
loudness	float64	float: -60 .. 0
mode	float64	bool: 0 o 1
speechiness	float64	float: 0 .. 1
acousticness	float64	float: 0 .. 1
instrumentalness	float64	float: 0 .. 1
liveness	float64	float: 0 .. 1
valence	float64	float: 0 .. 1
tempo	float64	float positivo
time_signature	float64	entero: 3 .. 7

Tabla 2. Descripción de variables, su tipo y rango en el dataset de tracks

2. Análisis de calidad de los datos y procesos de limpieza implementados

Limpieza general:

- Eliminación de todos los registros que tuvieran más de la mitad de las columnas en null.
- Para las columnas numéricas, los valores en null fueron reemplazados por la media en cada una de las columnas en las que se este fuera encontrado.
- A cada columna se revisó que cumpliera con los tipos de datos y rangos esperados.

Para los **artistas**, se agregaron dos columnas adicionales: **genres_list** y **flat_genres_list**.

- **genres_list** transforma el texto antes encontrado en la columna **genres** y lo convierte a una lista de strings, los cuales representan distintos géneros o instrumentos relacionados al artista. Ej: ['classical harp', 'harp'].

- **flat_genres_list** genera una nueva lista a partir de **genres_list**, separando las palabras encontradas. Esta nueva lista elimina duplicados. Ej: ['classical', 'harp']

Para los **tracks**, se agregaron las nuevas columnas: **artists_list**, **mode_type**, **track_type**, **instrumental_type**, **live_type** y **valence_type**.

Adicionalmente se hizo limpieza de la columnas **release_date**, pasando el valor de string a datetime. Ya que en algunos casos se tenía el mes y el día, se usaron estos datos sólo en caso de tenerse.

- **artists_list** une los datos de las columnas **artists** y **id_artists**, después de convertir los valores de las columnas previas en listas.
- **mode_type** utiliza el valor de la columna **mode** para identificar si la modalidad del track es ‘mayor’ o ‘minor’.
- **track_type** utiliza el valor de la columna **speechiness** para identificar si el track es ‘music’, ‘mixed’ o ‘speech’.
- **instrumental_type** utiliza el valor de la columna
- **instrumentalness** para identificar si el track es ‘instrumental’ o ‘vocal’.
- **live_type** usa el valor de la columna **liveness** para identificar si fue una cancion en vivo (‘live’) o grabado (‘recorded’).
- **valence_type** usa el valor de la columna **valence** para identificar si es un track alegre (‘positive’) o triste (‘negative’).

3. Análisis exploratorio de datos

El análisis exploratorio de los datos se realizó de manera independiente para cada dataset según el tipo de de los mismos se aplicaron medidas de representación univariadas y bivariadas de forma gráfica y no gráfica que se presentan a continuación:

Utilizando la librería de pandas profiling se identificó las alertas de las variables del dataset.

id: has a high cardinality: 1162095 distinct values	High cardinality
genres: has a high cardinality: 49155 distinct values	High cardinality
name: has a high cardinality: 1134430 distinct values	High cardinality
followers: is highly overall correlated with popularity	High correlation
popularity: is highly overall correlated with followers	High correlation
genres: is highly imbalanced (71.7%)	Imbalance
followers: is highly skewed (v1 = 113.8255684)	Skewed
id: is uniformly distributed	Uniform
name: is uniformly distributed	Uniform
id: has unique values	Unique
genres_list: is an unsupported type, check if it needs cleaning or further analysis	Unsupported
flat_genres_list: is an unsupported type, check if it needs cleaning or further analysis	Unsupported
followers: has 70924 (6.1%) zeros	Zeros
popularity: has 500791 (43.1%) zeros	Zeros

Fig 1. Alertas generadas con pandas profiling

La imagen anterior evidencia variables con alta cardinalidad, una alto porcentaje de presencia de ceros en las variables de seguidores y popularidad, así como una alta correlación lineal entre las mismas variables, lo cual se ratifica con los resultados del mapa de calor y la tabla de correlación (Fig 2)

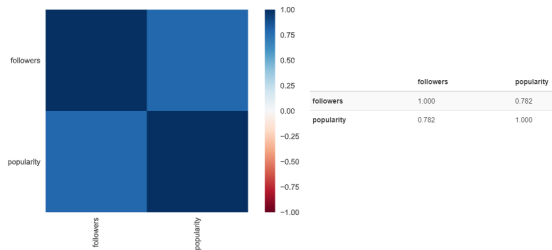


Fig 2. Mapa de correlación y tabla de correlación de las variables del dataset

Adicionalmente se realizó un análisis individual de las variables, la Fig. 3 muestra el conteo de registros, el % de registros únicos y los insights de cada una. Adicionalmente, la Fig. 4 expone las medidas de tendencia central, posición y forma para las variables cuantitativas de followers y popularity.

Quantile statistics		Descriptive statistics		Quantile statistics		Descriptive statistics	
Minimum	0	Standard deviation	13.557767	Minimum	0	Standard deviation	254.386.29
5-th percentile	0	Coefficient of variation (CV)	1.5413626	5-th percentile	0	Coefficient of variation (CV)	24.88072
Q1	0	Kurtosis	3.2728625	Q1	10	Kurtosis	20486.251
Median	2	Mean	8.7959607	Median	57	Mean	10220.808
Q3	13	Median Absolute Deviation (MAD)	2	Q3	417	Median Absolute Deviation (MAD)	55
95-th percentile	39	Skewness	1.8767052	95-th percentile	10516	Skewness	113.82527
Maximum	100	Sum	10221742	Maximum	78800234	Sum	1.9872957 + 10 ¹⁰
Range	100	Variance	183.81305	Range	78800234	Variance	6.471848 + 10 ¹⁰
Interquartile range (IQR)	13	Monotonicity	Not monotonic	Interquartile range (IQR)	407	Monotonicity	Not monotonic

Fig 4. medidas de distribución para las variables cuantitativas de followers y popularity.

Para el dataset de Tracks se realizó el mismo análisis con pandas profiling. Las alertas que generó el programa evidencian una alta cardinalidad para las variables de id, id_artists y artists, la primera presenta una distribución uniforme. Por otra parte, la variable de energy está fuertemente correlacionada con loudness y acousticness; las variables creadas que se listaron en el apartado anterior -como era de esperarse- presentan una alta correlación con la variable a partir de la cual fueron generadas. Pese a los procesos de limpieza y transformación de datos que se ejecutaron anteriormente, popularity, key, loudness e instrumentaltiness tiene un alto porcentaje de valores en ceros (Fig 5).

El análisis individual de las variables (Fig.7 hasta Fig.19) determinó las medidas de tendencia central, posición y forma para las variables del dataset, para las variables cuantitativas se adiciona el histograma de cada una de ellas para observar la distribución.

En adición, a lo anterior se generaron una serie de gráficos para responder preguntas adicionales sobre el dataset de tracks, estas incluyen el comportamiento de algunas variables en relación a la popularidad, características de sonido e interacción entre variables a lo largo del tiempo.

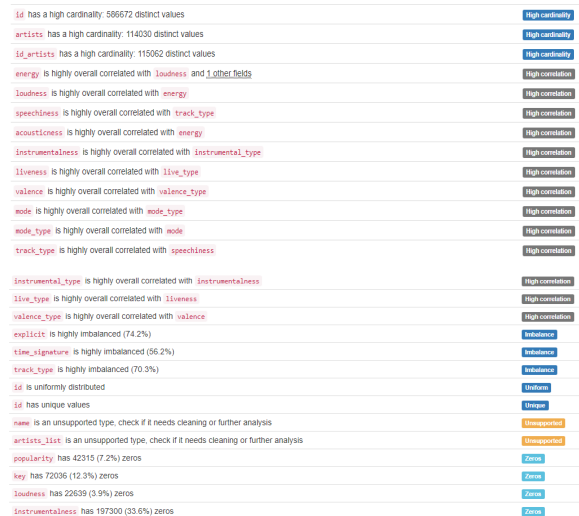


Fig 5. Alertas del dataset de tracks

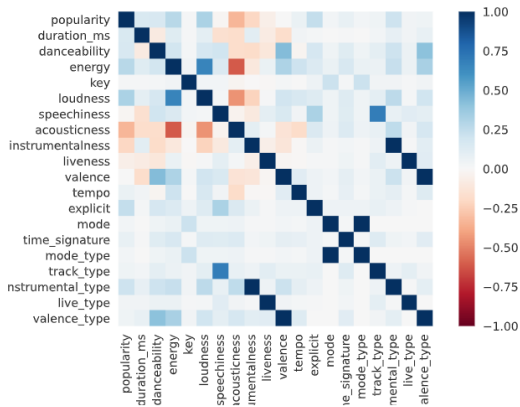


Fig 6. Mapa de correlación de las variables del dataset de tracks

● Análisis individual de las variables

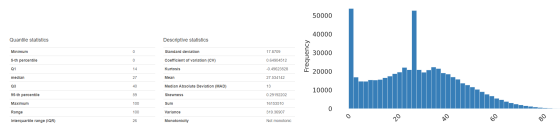


Fig 7. Análisis de la variable popularity

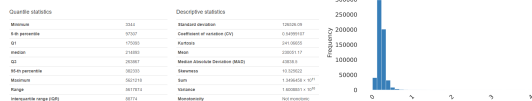


Fig 8. Análisis de la variable duration_dms

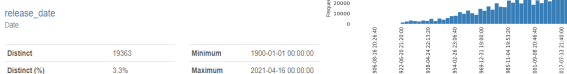


Fig 9. Análisis de la variable release date

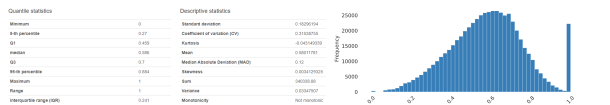


Fig 10. Análisis de la variable danceability

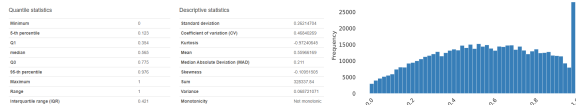


Fig 11. Análisis de la variable energy

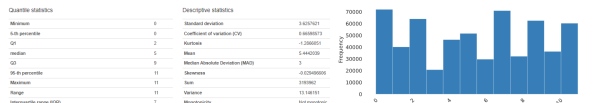


Fig 12. Análisis de la variable key

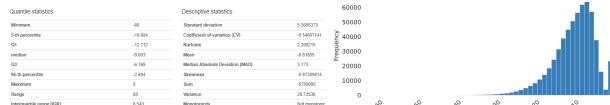


Fig 13. Análisis de la variable loudness

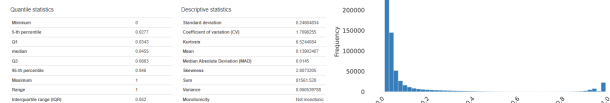


Fig 14. Análisis de la variable speechiness

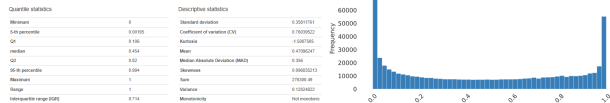


Fig 15. Análisis de la variable acousticness

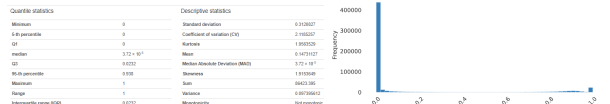


Fig 16. Análisis de la variable instrumentales

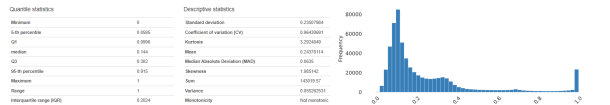


Fig 17. Análisis de la variable liveness

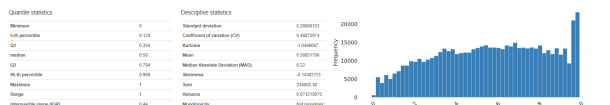


Fig 18 Análisis de la variable valence

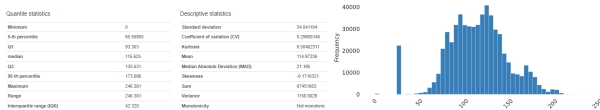


Fig 19. Análisis de la variable tempo

Los gráficos adicionales que se generaron permiten determinar:

- A. Comportamiento del lanzamiento de canciones a lo largo de las décadas (Fig. 20). Las décadas con mayor lanzamiento de canciones fueron 1990’s y 2010’s hasta el momento.

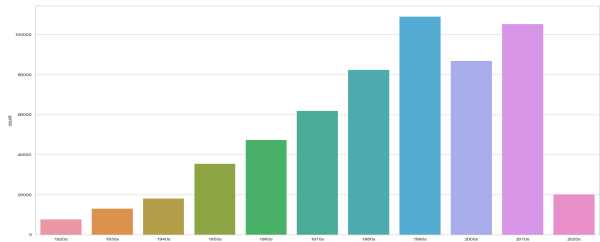


Fig 20. Recuento de lanzamiento de canciones a lo largo de las décadas

Observando el comportamiento de las variables del dataset de tracks a través del mapa de calor se observó en algunas de las representaciones un marcado comportamiento cercano a las medias de las variables. Ej: La Fig. 21 expone una tendencia de comportamiento (línea que se puede ver justo encima del 20 en el eje de popularidad) esto se debe a la decisión de utilizar la media para reemplazar valores nulos durante el proceso de análisis de calidad y limpieza de datos. Lo anterior, sugiere una oportunidad de mejora en la estrategia para reemplazar valores nulos con el fin de evitar estos comportamientos.

Por otra parte, la Fig-21 también evidencia que a mayor valor de la variable valence las canciones son más populares, lo que sugiere que las personas prefieren canciones más “alegres”.

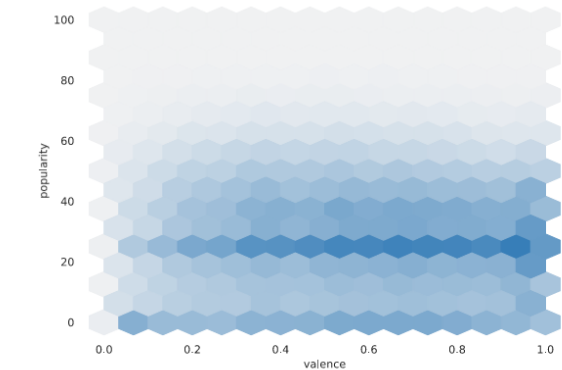


Fig 21. Mapa de interacción entre las variables popularity y valence del dataset de tracks

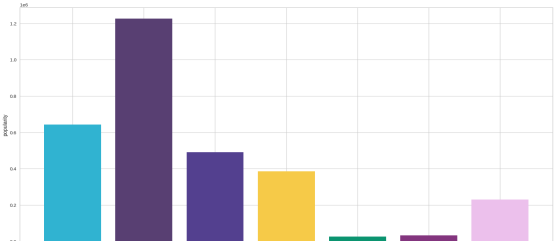


Fig 22. Popularidad de los géneros de las canciones

Adicionalmente se analizó la popularidad de la lista de géneros asociada a los 10 artistas más populares. Estos géneros fueron 'rock', 'pop', 'hip', 'rap', 'k-pop', 'reggaeton', 'trap'. La Fig. 22 evidencia que los géneros con mayor popularidad son el k-pop, el pop y el hip hop.

Para finalizar, las decisiones que se implementan en el proceso de limpieza y tratamiento de datos pueden influir en patrones de comportamiento de las variables, los analistas deben determinar las estrategias más convenientes para las características del dataset y el contexto de uso.

Así mismo, el análisis exploratorio de las variables permite observar gráficamente insights que no son evidentes en pasos previos, el uso de librerías como pandas profiling facilita el uso de diagramas interactivos con análisis integral, el valor principal desarrollo de estos gráficos con enfoque ejecutivo está en expresar de una manera clara y sencilla los patrones identificados contribuyendo a los procesos de modernización y toma de decisiones basados en datos.

4. Modelo de datos relacional

El modelo relacional para la información previamente contenida en los archivos csv de artistas y tracks se separó en 5 tablas.

Estas tablas son: Artists, Genres, ArtistGenres, Tracks, ArtistsTracks.

- Artists

Contiene únicamente la información de los artistas. Tiene las columnas:

COLUMNA	TIPO	CONTENIDO
id	varchar	id único del artista
followers	entero	cantidad de seguidores
name	varchar	nombre del artista
popularity	entero	popularidad del artista

Tabla 3. Modelo relacional de datos - Tabla Artists

- Genres

Tabla que contiene la lista completa de géneros que tienen los artistas. Tiene las columnas:

COLUMNA	TIPO	CONTENIDO
id	int	identificador único
name	varchar	nombre del género

Tabla 4. Modelo relacional de datos - Tabla Genres

- ArtistGenres

Tabla intermedia entre entre Artists y Genres

COLUMNA	TIPO	CONTENIDO
artist_id	varchar	id artista
genre_id	int	id genero

Tabla 5. Modelo relacional de datos - Tabla ArtistGenres

- Tracks:

Tabla que contiene toda la información de los tracks. Se agregaron algunas columnas con procesamiento adicional. Estas se escriben como [columna origen]:[nombre columna].

COLUMNA	TIPO	CONTENIDO
id	varchar	id track
name	varchar	nombre
popularity	int	popularidad del track
duration_ms	int	duración milisegundos
explicit	bool	si contiene lenguaje explícito
release_date	date	fecha salida
danceability	float	que tan bailable es
energy	float	que tan energética es
key	int	numero entre -1 y 11, pitch
loudness	int	entre -60 y 0, que tan ruidosa es
mode	bool	modo de la canción
speechiness	float	que tanto se habla
acousticness	float	entre 0 y 1, que tan acústica es
instrumentalness	float	entre 0 y 1, si contiene

		vocales
liveness	float	entre 0 y 1, que tan animada es
valence	float	entre 0 y 1, que tan alegre es
tempo	float	BPM (beats per minute)
time_signature	int	cantidad de beats en cada barra
mode:mode_type	varchar	“mayor” o “minor”
speeches:track_type	varchar	‘music’, ‘mixed’ o ‘speech’
instrumental:instrumental_type	varchar	‘instrumental’ o ‘vocal’
live:live_type	varchar	‘live’ o ‘recorded’
valence:valence_type	varchar	‘positive’ o ‘negative’

Tabla 6. Modelo relacional de datos - Tabla Tracks

- ArtistsTracks

Tabla intermedia entre artistas y tracks. Una fila representa la presencia de un artista en un track.

COLUMNA	TIPO	CONTENIDO
artist_id	varchar	id artista
track_id	varchar	id track

Tabla 7. Modelo relacional de datos - Tabla ArtistsTracks

5. Esquema general de la bodega de datos

Luego de implementar el ETL en Python, con los datos lo mejor preparados posible sin normalización se cargaron los dataset de artists y tracks a una Bodega de BigQuery para su posterior entrenamiento. Los dataframes se transformaron a archivos csv para facilitar su lectura, se eliminaron los índices y se asignó un encoding ‘uft-8’.

La autenticación para la subida de archivos se realizó a través OAuth2 por lo cual no se incluyó llave de logueo json. A continuación se presenta la estructura de la bodega de datos para artists y tracks.

- Artists_clean

ID de la tabla	gestiondedatos-380300.SPOTIFY_JUANITA_DANIEL.artists_clean
Creado	16 mar 2023, 16:59:14 UTC-5
Última modificación	16 mar 2023, 16:59:14 UTC-5
Vencimiento de la tabla	15 may 2023, 16:59:14 UTC-5
Ubicación de los datos	US

Fig 22. Detalles de la tabla artists_clean (Datos de artistas después del ETL)

Editar esquema


Filtro Ingresar el nombre o el valor de la propiedad					
<input type="checkbox"/>	Nombre del campo	Tipo	Modo	Intercalación	Valor predeterminado 
<input type="checkbox"/>	id	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	followers	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	genres	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	name	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	popularity	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	genres_list	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	flat_genres_list	STRING	NULLABLE		Valor predeterminado

Fig 23. Esquema de la tabla artists_clean en BigQuery

- Tracks_clean

ID de la tabla	gestiondedatos-380300.SPOTIFY_JUANITA_DANIEL.tracks_clean
Creado	16 mar 2023, 17:16:55 UTC-5
Última modificación	16 mar 2023, 17:16:55 UTC-5
Vencimiento de la tabla	15 may 2023, 17:16:55 UTC-5
Ubicación de los datos	US

Fig 24. Detalles de la tabla trackss_clean (Datos de artistas después del ETL)

Filtro Ingresar el nombre o el valor de la propiedad					
<input type="checkbox"/>	Nombre del campo	Tipo	Modo	Intercalación	Valor predeterminado 
<input type="checkbox"/>	id	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	name	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	popularity	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	duration_ms	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	explicit	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	artists	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	id_artists	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	release_date	STRING	NULLABLE		Valor predeterminado
<input type="checkbox"/>	danceability	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	energy	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	key	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	loudness	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	mode	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	speechiness	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	acousticness	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	instrumentalness	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	liveness	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	valence	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	tempo	FLOAT	NULLABLE		Valor predeterminado
<input type="checkbox"/>	time_signature	FLOAT	NULLABLE		Valor predeterminado

<input type="checkbox"/>	artists_list	STRING	NULLABLE	Valor predeterminado
<input type="checkbox"/>	mode_type	STRING	NULLABLE	Valor predeterminado
<input type="checkbox"/>	track_type	STRING	NULLABLE	Valor predeterminado
<input type="checkbox"/>	instrumental_type	STRING	NULLABLE	Valor predeterminado
<input type="checkbox"/>	live_type	STRING	NULLABLE	Valor predeterminado
<input type="checkbox"/>	valence_type	STRING	NULLABLE	Valor predeterminado

Fig 25. Esquema de la tabla tracks_clean en BigQuery

6. Arquitectura del ETL

El proceso de ETL se implementó en Python con el fin de obtener los datos lo más preparados posibles para futuros entrenamientos pero sin aplicar los procesos de normalización. La Fig.26 muestra a detalle el diagrama de arquitectura del ETL implementado el cual puede ser consultado a detalle en el Anexo del repositorio de Github.

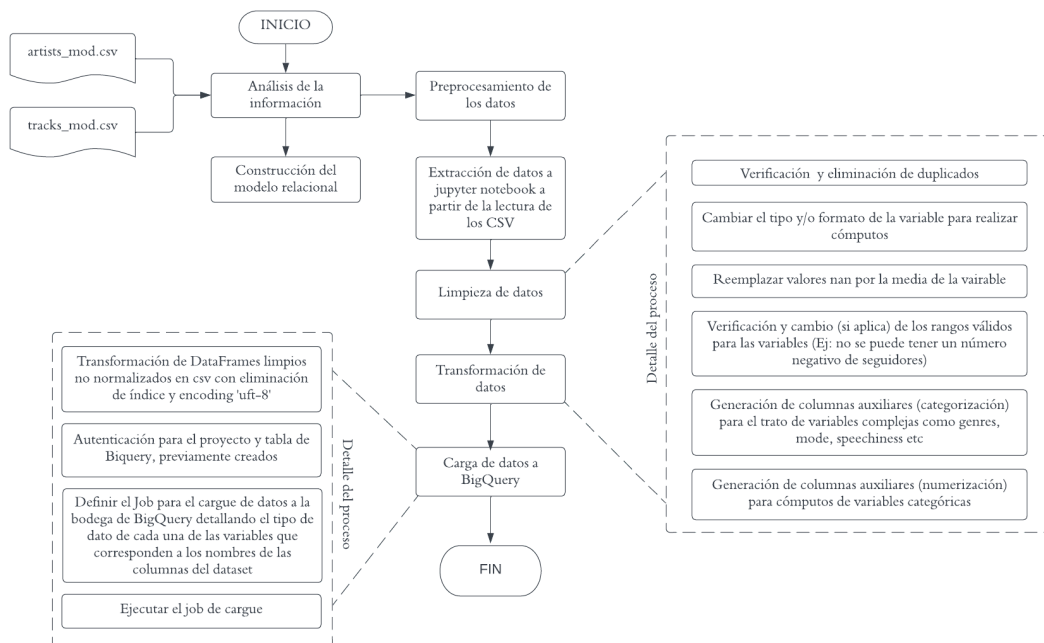


Fig 26. Arquitectura del ETL implementado proyecto gestiondedatos-380300-SPOTIFY_JUANITA_DANIEL

7. Arquitectura general de la solución

La figura 28. en la siguiente página, evidencia la arquitectura global de la solución incluyendo las secciones de ETL, el sistema de recomendación y el dashboard, dado que en otras secciones del documento se detalló el proceso del ETL y en secciones posteriores profundizaremos sobre la arquitectura del dashboard, este numeral se enfocará principalmente en la arquitectura del sistema de recomendación.

El sistema tiene dos tipos de despliegue, en uno se despliega únicamente como un REST api y en el otro despliega un front.

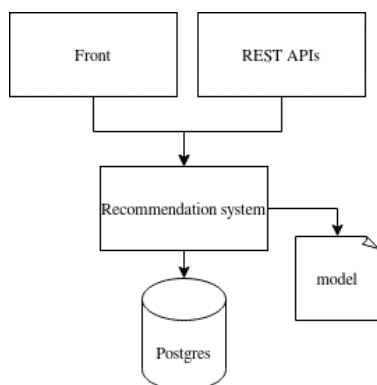


Fig 27. Diagrama Genérico de sistema de recomendación

Ambos métodos fueron desarrollados utilizando python y utilizan los mismos sistemas internos. Para el caso del rest api, se utilizaron las librerías: fastapi y uvicorn. En el caso del front se usó gradio. Una ventaja de esta librería es que también permite desplegar a una url accesible de forma pública.

El sistema realiza el almacenamiento de datos en la base de datos Postgres. La tabla más relevante para este proyecto es la de Tracks, ya que las recomendaciones se hicieron principalmente a base de esta tabla. Hipotéticamente, podrían usarse otras tablas para generar recomendaciones, como podría ser en el caso de realizar similitud entre artistas.

Al desplegar el proyecto se realiza el entrenamiento del modelo con base a las columnas que se hayan tomado como feature list. Este modelo después será utilizado para generar recomendaciones, utilizando la librería annoy.

8. Descripción del sistema de recomendación

De la totalidad de las columnas que se tienen en el sistema: ['id', 'name', 'popularity', 'duration_ms', 'explicit', 'danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'time_signature', 'mode_type', 'track_type', 'instrumental_type', 'live_type', 'valence_type', 'release_date']

Se seleccionaron las siguientes columnas al momento de realizar el sistema de recomendación:

['popularity', 'explicit', 'danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', 'time_signature'].

Lo anterior, porque se consideró que estas serían las más relevantes al momento de buscar tracks similares. Todas estas representan variables cuantitativas. Debido a que se

está utilizando annoy para generar el modelo, y este genera algo similar a un knn, es preferible utilizar variables cuantitativas sobre variables cualitativas. La excepción en este caso es explicit, siendo una variable de tipo cualitativo, pero al ser nominal es fácil de acoplar al sistema de recomendación. Además puede ser importante tener esta categoría dado que a una persona puede que no le gusten canciones con lenguaje explícito.

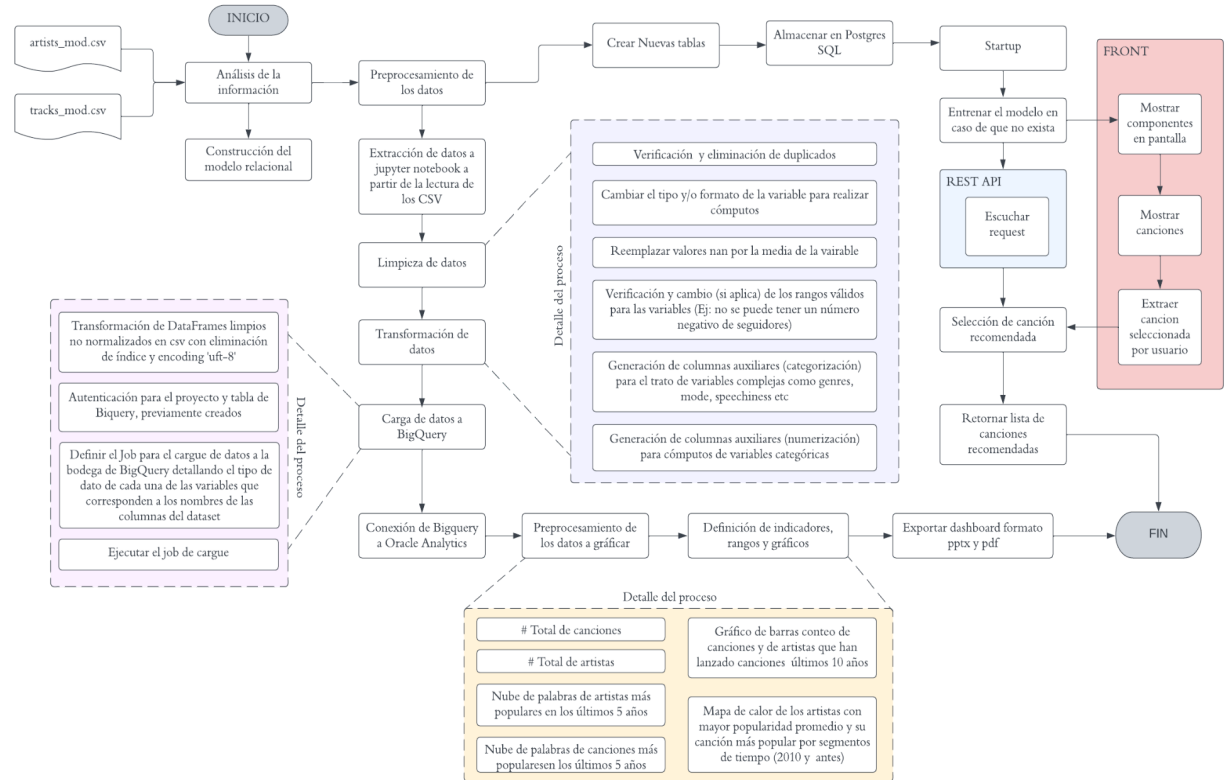


Fig 28. Diagrama de arquitectura general proyecto gestiondedatos-380300-SPOTIFY_JUANITA_DANIEL

9. Descripción del dashboard

El dashboard se realizó en Oracle Analytics Cloud, se realizó una conexión con los datos que se cargaron a Bigquery, posteriormente con la conexión establecida se establecieron los formatos que solicita la herramienta en términos de tratamiento y métodos de agregación.

El panel general muestra 3 indicadores, 6 gráficos y 5 descripciones automáticas de algunas de las variables del dataset. Cada una de las gráficas cuenta con filtros específicos de tiempo y de variables de interés en cada una para resolver consultas ejecutivas y contribuir a la toma de decisiones. Adicionalmente, el dashboard en producción

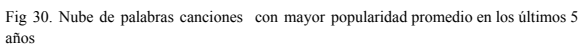
cuenta con una sección “Mis cálculos” para personalizar consultas SQL sobre la data.

Los indicadores (Fig. 29) exponen el número total de canciones, artistas y la popularidad promedio en todo el set.

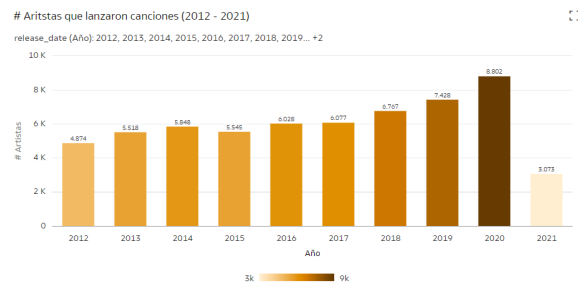
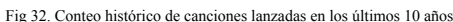


Fig 29. Indicadores del dashboard principal

Otros de los gráficos fueron las nubes de palabras, en este caso, se crearon dos (Fig. 30 y Fig. 31) para evidenciar los artistas y canciones con mayor popularidad promedio en los últimos 5 años.

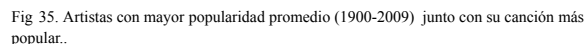
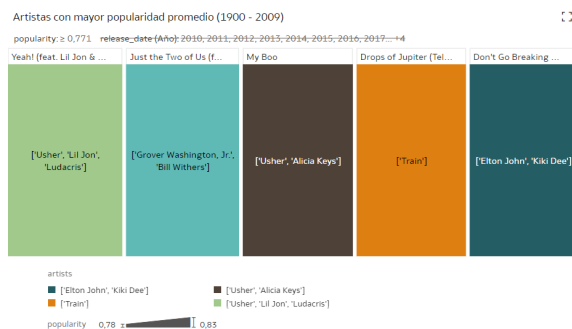
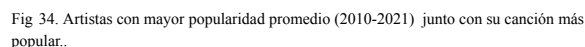
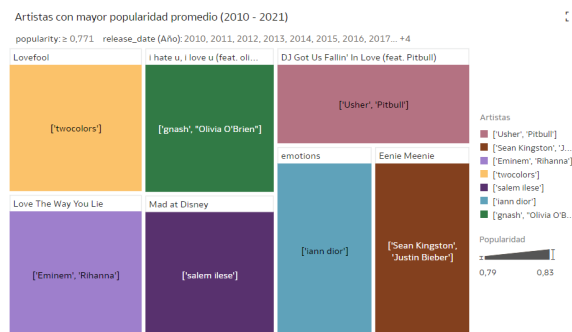


Por otra parte, los gráficos de barras nos muestran el conteo histórico del lanzamiento de canciones y la misma métrica sobre los artistas en los últimos 10 años, cada uno de los filtros como se mencionó al inicio de la sección se puede modificar de forma independiente para otorgar distintas visualizaciones de interés.



Finalizando la sección de gráfica del dashboard encontrarán dos gráficos de rectángulos (Fig. 34 y 35) los

cuales permiten determinar los artistas con mayor popularidad promedio para diferentes periodos de tiempo junto al nombre de su canción más popular.



Las gráficas e indicadores anteriores son de relevancia para el negocio porque permiten tener un balance general de las métricas más importantes (canciones, artistas y popularidad) de estos pilares principales se evidencia una evolución histórica y relación con las demás características propias de las canciones del periodo de tiempo analizado.

Para concluir la descripción del dashboard se decidió incluir una funcionalidad automática de Oracle Analytics que habilita el análisis exploratorio de algunas variables del dataset. Esta funcionalidad también puede ser utilizada en generación de reportes automáticos para perfiles ejecutivos y se puede complementar con opciones de desarrollador para automatización de procesos manuales.

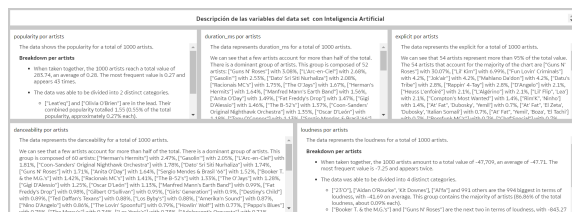


Fig 31. Descripción automática de algunas variables del dataset

Anexos Entrega Final

1. Link Repositorio de GITHUB: <https://github.com/dhasane/gestion-datos>
2. Video / Presentación Proyecto:
https://drive.google.com/file/d/1DYRuAFFX250dITBBSZkluXwK_Z47FfcO/view
<https://youtu.be/TGJzpZJ0gTc>