Master Universitario en Bioinformática para Ciencias de la Salud
Facultad de Informática
UNIVERSIDADE DA CORUÑA

Práctica 1. Calidad e integración de datos.

INTELIGENCIA COMPUTACIONAL PARA DATOS DE ALTA DIMENSIONALIDAD.

Cristian Morillo Losada | Pedro Sánchez García

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA PARA CIENCIAS DE LA SALUD.
PROFESOR: DR. CARLOS EIRAS FRANCO.

PARTE 1: CALIDAD DE DATOS

INTRODUCCIÓN.

Los análisis de Big Data representan una versatilidad para la extracción de valor y realización de predicciones útiles a partir de conjuntos de datos. No obstante, la calidad de estos últimos conforma un requisito para lograr beneficios correspondientes, pues existen varios peligros asociados a la mala calidad, tales como los errores en toma de decisiones, incremento en costes de gestión y daños en la relación con los clientes. De esta forma, la primera parte de la práctica se centra en la implementación de la calidad de datos a través de los componentes de Talend, con el fin de justificar un perfilado completo sobre los conjuntos de datos proporcionados.

OBJETIVOS.

- Analizar la calidad de los datos sobre los *releases* de **musicbrainz** (datos erróneos, duplicados, etc.).
- Analizar la calidad del conjunto de datos resultante de cruzar (hacer un *join*) *musicbrainz_release* y *bestalbums3000* mediante el atributo *artist*.

ANÁLISIS DE CALIDAD DE DATOS EN *RELEASES* DE MUSICBRAINZ.

CONFIGURACIÓN DE LOS ANÁLISIS.

Se plantea examinar los datos proporcionados y alcanzar unas estadísticas o resúmenes informativos sobre estos que nos muestren una visión global de su calidad. Para ello, nos centramos en la existencia de posibles valores nulos y atípicos, que no corresponden con los patrones habituales del tipo de datos que se manejan. Además, cabe destacar la relevancia en la detección de duplicados, con el fin de evitar diferentes referencias a una misma representación determinada.

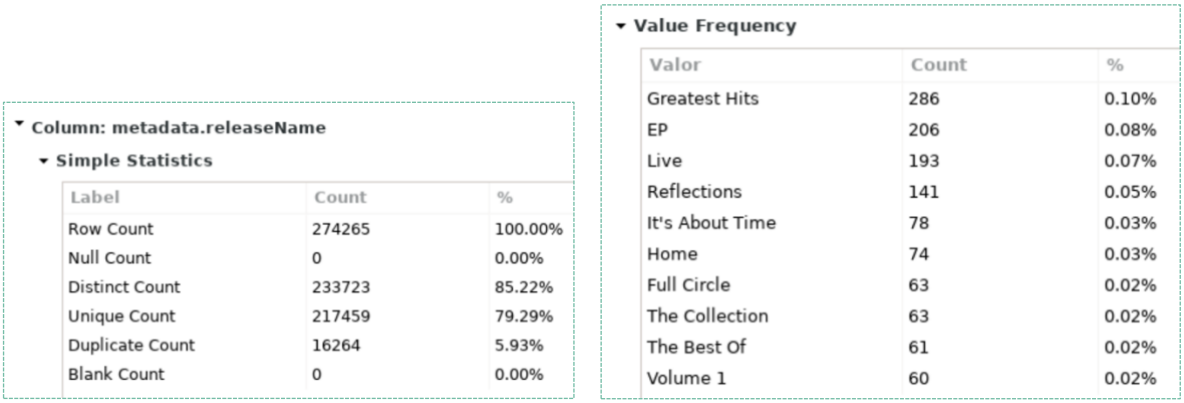
De esta forma, como primer paso, debemos tener en cuenta el tipo de conjunto de datos que manejamos. El conjunto MusicBrainz constituye un fichero de texto con campos separados por delimitadores (CSV), donde se encuentra la tabla *musicbrainz_release*, con información de los lanzamientos (*releases*). En concreto, presenta los atributos *id* (numérico), *releaseName* y *artisticName* (ambos nominales).

Posteriormente, en base a las explicaciones de las clases expositivas y el tipo de conjunto de datos, se procede a elegir los análisis que se consideran oportunos para el perfilado:

- Análisis de valores nominales *(Nominal Values Analysis)*.
- Análisis de frecuencia de patrones *(Pattern Frequency Analysis)*.
- Análisis de datos discretos *(Discrete Data Analysis)*.

ANÁLISIS DE VALORES NOMINALES *(NOMINAL VALUES ANALYSIS)*.

Se lleva a cabo este análisis sobre los atributos del nombre de lanzamiento *(releaseName)* y nombre artístico *(artisticName)*. Como indicadores configurados, se encuentran los de estadísticas básicas (recuento de filas, valores nulos, diferentes, únicos, duplicados y en blanco) y avanzadas (frecuencia de valores). Se alcanzan los siguientes resultados para el nombre de lanzamiento:



Column: metadata.releaseName

Simple Statistics

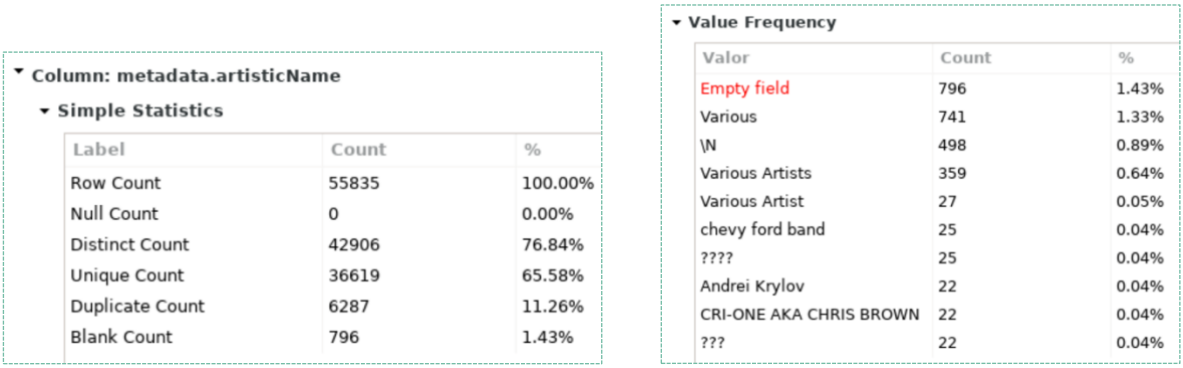
Label	Count	%
Row Count	274265	100.00%
Null Count	0	0.00%
Distinct Count	233723	85.22%
Unique Count	217459	79.29%
Duplicate Count	16264	5.93%
Blank Count	0	0.00%

Value Frequency

Valor	Count	%
Greatest Hits	286	0.10%
EP	206	0.08%
Live	193	0.07%
Reflections	141	0.05%
It's About Time	78	0.03%
Home	74	0.03%
Full Circle	63	0.02%
The Collection	63	0.02%
The Best Of	61	0.02%
Volume 1	60	0.02%

Figura 1. Resultados del análisis de valores nominales para el nombre de lanzamiento.

Esta visión global del conjunto nos informa de que estamos tratando 274.265 filas correspondientes a nombres de lanzamientos, distinguiendo la ausencia de valores nulos, 233.723 valores diferentes y 217.459 únicos (**Figura 1**). Además, se identifican 16.264 valores duplicados, que representan un 5,93%. Por otro lado, la lista de frecuencia de valores pone de manifiesto el recuento de aquellos valores nominales que más aparecen en el conjunto. De este modo, observamos que 'Greatest Hits', 'EP' y 'Live' son los tres valores nominales con más apariciones. Con respecto al nombre artístico, se logran los resultados mostrados a continuación:



Column: metadata.artisticName

Simple Statistics

Label	Count	%
Row Count	55835	100.00%
Null Count	0	0.00%
Distinct Count	42906	76.84%
Unique Count	36619	65.58%
Duplicate Count	6287	11.26%
Blank Count	796	1.43%

Value Frequency

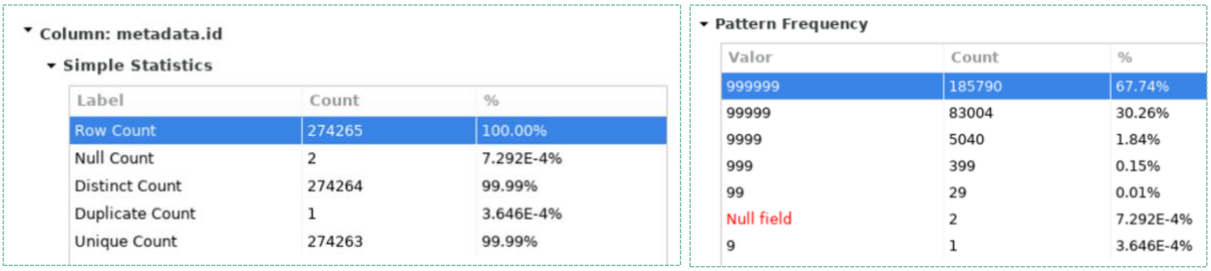
Valor	Count	%
Empty field	796	1.43%
Various	741	1.33%
\N	498	0.89%
Various Artists	359	0.64%
Various Artist	27	0.05%
chevy ford band	25	0.04%
????	25	0.04%
Andrei Krylov	22	0.04%
CRI-ONE AKA CHRIS BROWN	22	0.04%
???	22	0.04%

Figura 2. Resultados del análisis de valores nominales para el nombre artístico.

En este caso, tal y como se recoge en la **Figura 2**, se manejan 55.835 filas y no existen valores nulos. De este modo, existen 42.906 valores nominales diferentes y 36.619 únicos. Cabe destacar la existencia de 6.287 valores duplicados y 796 en blanco, que corresponden con 11,26% y 1,43% respectivamente en el conjunto de datos. En lo que respecta a la frecuencia de los valores, se determina que hay 796 registros que carecen de contenido, 741 que corresponden con lanzamientos de varios artistas ('Various') y 498 que están con registro de '\N', es decir, junto con '????' y '???', se asocian a errores que implican una problemática.

ANÁLISIS DE FRECUENCIA DE PATRONES *(PATTERN FREQUENCY ANALYSIS)*.

En el caso anterior se ha verificado la frecuencia de valores nominales en nombre de lanzamiento y nombre artístico. El presente análisis se centra en descubrir patrones en el conjunto de datos que se maneja. Para ello, establece una configuración de los indicadores dependiendo del tipo de atributo analizado. En consecuencia, para el identificador (*id*) se evalúan todo lo relativo a estadísticas básicas (recuento de filas, valores nulos, diferentes, únicos y duplicados) y estadísticas de la frecuencia de patrones. Talend nos devuelve los siguientes resultados:



Column: metadata.id

Simple Statistics

Label	Count	%
Row Count	274265	100.00%
Null Count	2	7.292E-4%
Distinct Count	274264	99.99%
Duplicate Count	1	3.646E-4%
Unique Count	274263	99.99%

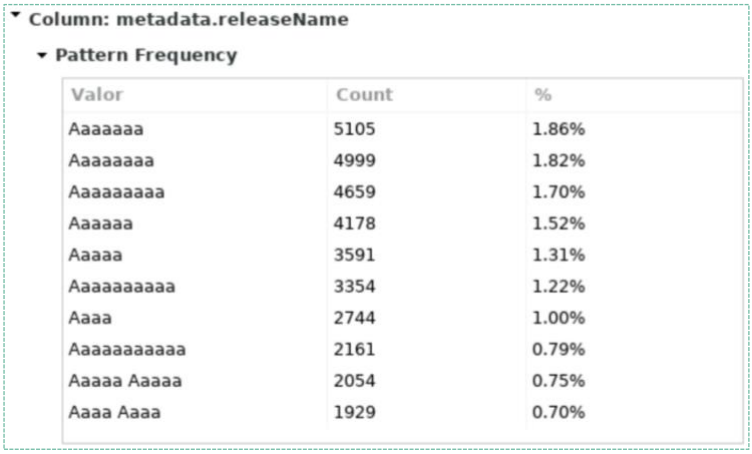
Pattern Frequency

Valor	Count	%
999999	185790	67.74%
99999	83004	30.26%
9999	5040	1.84%
999	399	0.15%
99	29	0.01%
Null field	2	7.292E-4%
9	1	3.646E-4%

Figura 3. Resultados del análisis de frecuencia de patrones para el identificador.

En este caso, tal y como refleja la **Figura 3**, se contabilizan 274.265 filas correspondientes a los identificadores, donde 2 son nulos y 1 es duplicado, de modo que hay 274.264 diferentes y 274.263 únicos. Con respecto a la frecuencia de patrones, se puede apreciar que hay un recuento de 185.790 identificadores de 6 dígitos, que representan el 67,74% del conjunto de datos. En segunda posición se encuentran los identificadores de 5 dígitos, con 83.004 contabilizados y que conforman el 30,26% del conjunto. Finalmente, cabe destacar el drástico descenso del porcentaje para el caso de identificadores con 4 dígitos o menos. De este modo, el análisis nos proporciona un enfoque relevante de la tendencia que hay en los identificadores.

A continuación, en la **Figura 4** se muestran los resultados del análisis para el nombre de lanzamiento y nombre artístico:



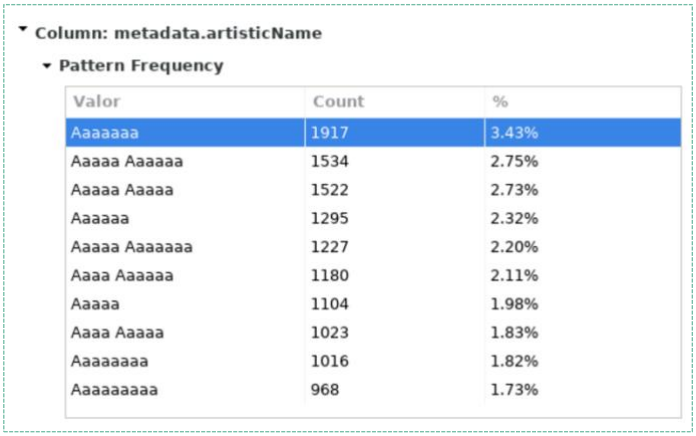
Column: metadata.releaseName

Pattern Frequency

Valor	Count	%
Aaaaaaa	5105	1.86%
Aaaaaaaa	4999	1.82%
Aaaaaaaaa	4659	1.70%
Aaaaaa	4178	1.52%
Aaaaa	3591	1.31%
Aaaaaaaaaa	3354	1.22%
Aaaa	2744	1.00%
Aaaaaaaaaaa	2161	0.79%
Aaaaa Aaaaa	2054	0.75%
Aaaa Aaaa	1929	0.70%

Figura 4. Resultados del análisis de frecuencia de patrones para el nombre de lanzamiento.

A diferencia de los identificadores, para el nombre de lanzamiento, la frecuencia de patrones no es tan diferenciada. En consecuencia, podemos observar que los nombres con mayor recuento son aquellos que presentan de 7 a 9 caracteres.



Column: metadata.artisticName

Pattern Frequency

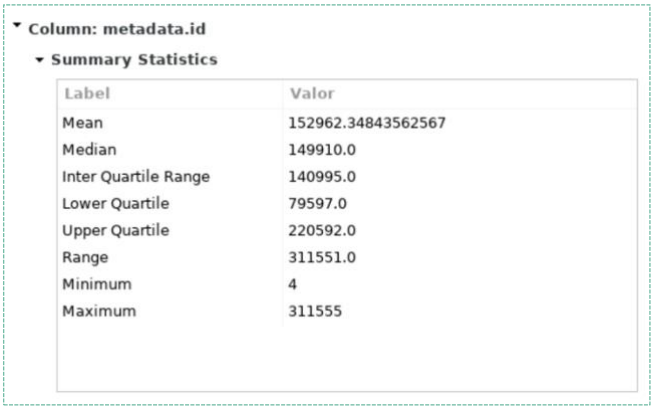
Valor	Count	%
Aaaaaaa	1917	3.43%
Aaaaa Aaaaaa	1534	2.75%
Aaaaa Aaaaa	1522	2.73%
Aaaaaa	1295	2.32%
Aaaaa Aaaaaaa	1227	2.20%
Aaaa Aaaaaa	1180	2.11%
Aaaaa	1104	1.98%
Aaaa Aaaaa	1023	1.83%
Aaaaaaaa	1016	1.82%
Aaaaaaaaa	968	1.73%

Figura 5. Resultados del análisis de frecuencia de patrones para el nombre artístico.

Por último, en lo relativo a los resultados del análisis para el nombre artístico (**Figura 5**), hay una similitud con el caso anterior, pues no hay un recuento que sean notablemente más elevado en comparación con el resto. No obstante, resulta relevante conocer aquellos nombres que poseen un mayor recuento, como sucede con los de 7 caracteres o los de 2 palabras.

ANÁLISIS DE DATOS DISCRETOS *(DISCRETE DATA ANALYSIS)*.

Se lleva a cabo este análisis para evaluar la forma/distribución general de los identificadores numéricos en el conjunto de datos. Para ello, se establecieron los indicadores de media, mediana, rango intercuartílico y rango (mínimo-máximo):



Column: metadata.id

Summary Statistics

Label	Valor
Mean	152962.34843562567
Median	149910.0
Inter Quartile Range	140995.0
Lower Quartile	79597.0
Upper Quartile	220592.0
Range	311551.0
Minimum	4
Maximum	311555

Figura 6. Resultados del análisis de datos discretos para identificadores.

El motivo de la realización de este análisis es conocer el rango de valores correspondientes a los identificadores que se manejan (**Figura 6**). En términos generales, se podría determinar por ejemplo la existencia de posibles identificadores atípicos que se añadan al conjunto y se alejen del rango, que en este caso comprende desde un valor mínimo de 4 hasta 311.555.

CONSIDERACIONES SOBRE LOS ERRORES DE FORMATO EN PERFILADO.

Cabe destacar un aspecto relevante que se detectó en la realización de los análisis anteriores para el nombre artístico. Tal y como podemos apreciar en las **Figuras 2 y 5**, se está realizando el análisis con 55.835 filas de los datos mencionados.

A pesar de que Talend no mostró una ventana emergente en el transcurso de los análisis, la posterior repetición de los análisis habilitó un aviso como el siguiente (se podría asociar a algún complemento que Talend pedía que se instalase y se activó posteriormente):

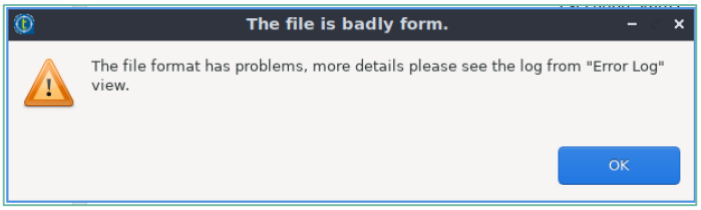


Figura 7. Cuadro emergente sobre el error de formato que ofrece Talend.

Acudiendo al log de errores, Talend indicó que la columna correspondiente al nombre artístico en la fila 55.836 carecía de datos. Por tanto, acudimos al fichero CSV y examinamos la fila mencionada:

```
55835   63653;From the Inside Out;Das The Insider
55836   63654;"Do it Movin U Beezy"" Mixtape    Da Stooie Bros. 2009-06-18 01:27:00.715271+00   2009-06-18 01:27:00.715271+00   1   0   0
        689076846547    CD Baby id:dastooiebros
```

Figura 8. Captura correspondiente a la fila 55.836 del fichero *musicbrainz_release* original.

Podemos apreciar que la fila presenta unas fechas en los campos y tabulaciones que dificultan determinar el correspondiente nombre del lanzamiento y nombre artístico (**Figura 8**). En consecuencia, ante esa falta de claridad en cada atributo, se tomó la decisión de proceder a su eliminación manualmente. A continuación, se conectó el fichero modificado en Talend y se efectuó un nuevo análisis. No obstante, Talend nos advirtió de un mismo problema en la fila 73.096, que presentaba un aspecto similar al caso anterior para los campos. Se optó por la misma solución aplicada anteriormente y se repitieron los análisis para el nombre artístico sobre la versión definitiva del archivo modificado.

El análisis de valores nominales para el nombre artístico ofrece los siguientes resultados:

Column: metadata.artisticName

Simple Statistics

Label	Count	%
Row Count	274263	100.00%
Null Count	0	0.00%
Distinct Count	181219	66.07%
Unique Count	143262	52.24%
Duplicate Count	37957	13.84%
Blank Count	4077	1.49%

Column: metadata.artisticName

Simple Statistics

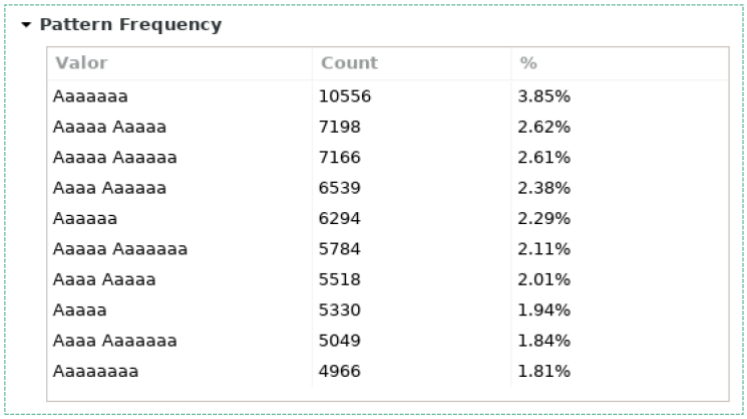
Value Frequency

Valor	Count	%
Various	4641	1.69%
Empty field	4077	1.49%
\N	3465	1.26%
Various Artists	2141	0.78%
????	194	0.07%
Wolfgang Amadeus Mozart	172	0.06%
Various Artist	133	0.05%
Beethoven	133	0.05%
Johann Sebastian Bach	131	0.05%
Mozart	127	0.05%

Figura 9. Resultados del análisis de valores nominales en nombre artístico del fichero modificado.

En general, si lo comparamos con el análisis previamente realizado en el archivo sin modificar, se manejan 274.263 filas y no existen valores nulos. Existen 181.219 valores diferentes, con 143.262 únicos. Por otro lado, se determinan 37.957 valores duplicados, junto con 4.077 valores en blanco, siendo este número notablemente inferior si lo comparamos con el resto (**Figura 9**). En lo referente a la frecuencia de los valores analizada, se puede observar que 4.641 hacen referencia a los lanzamientos de varios artistas ('Various') y que 3.465 y 194 se ubican con registro de '\N' o '????' respectivamente, lo que corresponde con una problemática para tener en cuenta.

Por su parte, los resultados del análisis de frecuencias de patrones en nombre artístico sobre el archivo modificado poseen el siguiente aspecto:



Pattern Frequency

Valor	Count	%
Aaaaaaa	10556	3.85%
Aaaaa Aaaaa	7198	2.62%
Aaaaa Aaaaaa	7166	2.61%
Aaaa Aaaaaa	6539	2.38%
Aaaaaa	6294	2.29%
Aaaaa Aaaaaaa	5784	2.11%
Aaaa Aaaaa	5518	2.01%
Aaaaa	5330	1.94%
Aaaa Aaaaaaa	5049	1.84%
Aaaaaaaa	4966	1.81%

Figura 10. Resultados del análisis de frecuencia de patrones en nombre artístico del fichero modificado.

Al igual que en el caso del análisis realizado anteriormente, se produce una tendencia similar, donde no hay un recuento que sea notablemente elevado si se compara con el resto (**Figura 10**). De nuevo, cabe destacar que este análisis es relevante para conocer aquellos nombres con un mayor recuento, como es el caso de aquellos con 7 caracteres (3,85%) o separados por 2 palabras (2,62%).

ANÁLISIS DE CALIDAD DE DATOS SOBRE EL *JOIN* ENTRE MUSICBRAINZ_RELEASE Y BESTALBUMS3000.

CONFIGURACIÓN DEL ANÁLISIS.

Para el *join* entre musicbrainz_release y bestalbums300, que será la fuente de datos correspondiente, se considera llevar a cabo el análisis de redundancia *(Redundancy Analysis)*. En consecuencia, se exploran los datos asociados a los artistas para dos conjuntos de columnas correspondientes a cada tabla almacenada en la base de datos creada en MySQL. De este modo, podemos alcanzar una visión en torno a la correspondencia existente entre ambas tablas para el atributo de artista.

ANÁLISIS DE REDUNDANCIA *(REDUNDANCY ANALYSIS)*.

Se procede a realizar este análisis sobre el atributo de artista, mediante el cual se produce el *join*. Como resultado, Talend nos devuelve la siguiente información:

	bestalbums3000	musicbrainz_release
%Match	83.64%	3.92%
%NotMatch	16.36%	96.08%
#Match	2209	10742
#NotMatch	432	263518
#Rows	2641	274260

Figura 11. Resultados del análisis de redundancia en el *join* solicitado.

Se produce una notable correspondencia en lo que respecta a los artistas presentes en la tabla de bestalbums3000 con la de musicbrainz_release, de modo que el porcentaje es del 83,64% (**Figura 11**). Por el contrario, para los artistas de la tabla musicbrainz_release se produce un porcentaje de correspondencia muy bajo, que se ubica en un 3,92%. Cabe destacar la importancia de este análisis para identificar unos porcentajes de correspondencia muy elevados en ambos casos, que sería necesario tener en cuenta de cara a la calidad.

PARTE 2: INTEGRACIÓN DE DATOS

INTRODUCCIÓN.

La integración de datos se trata de un proceso de combinación de los datos heterogéneos procedentes de diversas fuentes que presentan variaciones en forma y estructura de una aplicación. Como objetivo principal, se persigue la fusión de diferentes datos (matrices, tablas…) para una mejora en los análisis de conjuntos de datos para múltiples sectores. Debemos tener en cuenta la fuente de datos, volumen de estos y factores como usuarios, es decir, la orientación al proceso donde se aplicará el análisis correspondiente. Como consecuencia, distinguimos arquitecturas de integración como sistemas operacionales y sistemas informacionales que varían en operaciones, datos de operación, concurrencia y actualización. Las grandes compañías han optado por una de ellas para la extracción de información, su transformación y carga posterior, lo que pone de manifiesto la notable inversión económica y de recursos a nivel mundial en torno a la integración de datos.

OBJETIVOS.

El objetivo es crear un flujo de trabajo para cargar en MySQL todos los lanzamientos (*releases*), con todas las canciones de los artistas que:

- Tengan al menos un álbum entre los 100 mejores de todos los tiempos.
- Sean el artista favorito de al menos 15 usuarios de Last.fm, es decir, que se les haya aplicado la *tag 'Favourite'* o *'Favourites'* o cualquiera de sus variaciones al menos 15 veces (en total entre todas las variaciones).

El esquema de la tabla destino debe constar de estos campos:

- ***artist:*** nombre del artista.
- ***album:*** nombre del álbum.
- ***trackname:*** nombre de la canción.
- ***track_position:*** posición de la canción en el álbum.

CONFIGURACIÓN DEL FLUJO UTILIZADO.

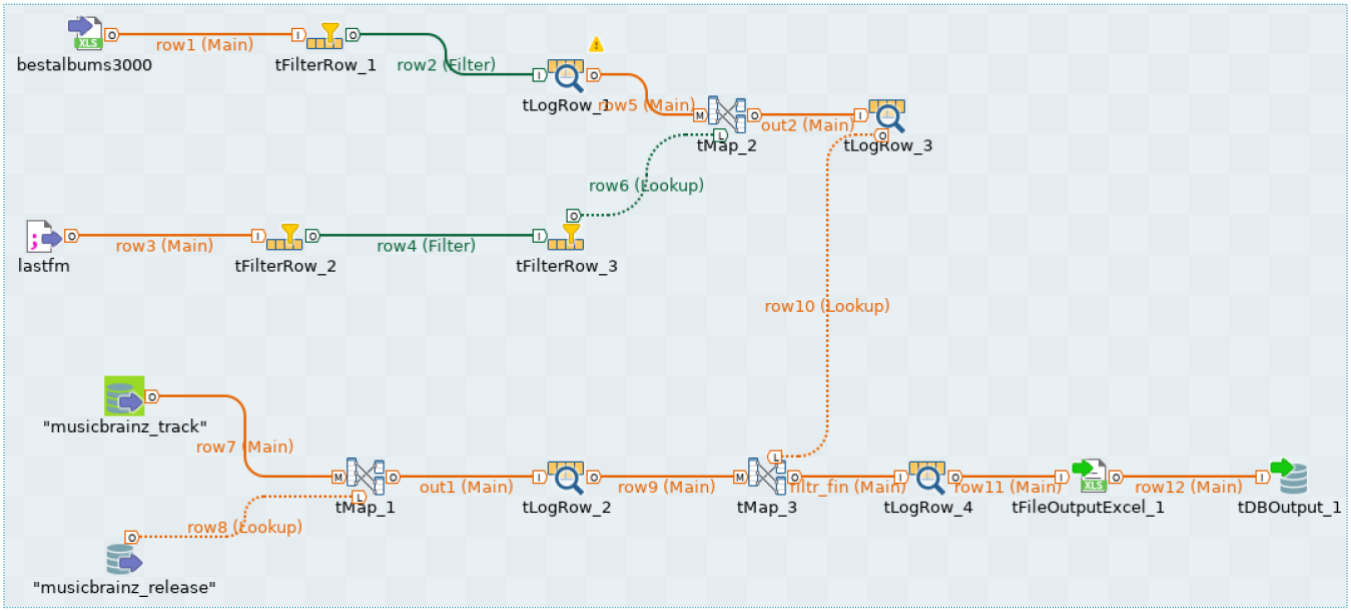


Figura 12. Flujo de trabajo general planteado en *Data Integration* de Talend para la resolución.

Para alcanzar el objetivo propuesto, se ha planteado un flujo de trabajo basado en un rendimiento adecuado, por medio de la verificación progresiva en la ejecución del desarrollo. En primer lugar, con respecto a la lectura de los ficheros delimitados y los datos de la conexión en la base de datos de MySQL, se han importado siguiendo un procedimiento similar al de la parte 1 de la práctica. Posteriormente, se integraron en el entorno de trabajo arrastrándolos desde el panel lateral hacia la región izquierda, tal y como se aprecia en la **Figura 12**.

Con el fin de filtrar al comienzo los datos por cuestión de eficiencia, se empleó el componente tFilterRow sobre los ficheros bestalbums3000 y lastfm. En el primer caso, se estableció la condición 'PLACE' ≤ 100 en el cuadro de configuración, pues uno de los requisitos es que el álbum se ubicase entre los 100 mejores de todos los tiempos. Por otro lado, con respecto al fichero de lastfm, el filtrado se realizó en dos fases, donde la primera contenía 'TAG' = Favourite OR 'TAG' = Favourites para seleccionar aquellos artistas que presentaban esas tag concretas. A continuación, la siguiente fase fue para lograr aquellos artistas favoritos de al menos 15 usuarios de lastfm, de modo que la condición consistió en 'tag' ≥ 15.

El siguiente paso fue proceder al *join* por el componente tMap entre el filtrado de bestalbums3000 y lastfm realizados. Se estableció que el primero actuase como *Main* y el otro como *Lookup*, de modo que se controla lo relativo a memoria y, por tanto, rendimiento. Con respecto a la configuración, cabe destacar que se estableció el enlace a través del atributo *artist* en la ventana de tMap. Se verificó la salida correspondiente, conformada por los atributos *artist, place* y *album*, con el correspondiente componente tLogRow_3.

Una vez que se alcanzó la estructura anterior, se procedió al *join* entre musicbrainz_release y musicbrainz_track en tMap por el atributo correspondiente al identificador (*id*). En este caso, la salida está constituída por los atributos del nombre artístico (*artisticName*), nombre de la canción

(*trackName*) y la posición de esta en el álbum (*trackNumber*). De nuevo, se evaluó la correspondiente salida por el componente tLogRow_2. Posteriormente, se llevó a cabo el *join* entre dicha salida y la obtenida por el otro proceso explicado anteriormente. Para ello, se empleó el componente tMap, reflejado como tMap_3 en la **Figura 12**, seleccionando el atributo de artista y visualizando la salida por el componente tLogRow_4.

Finalmente, con esta salida se cumplen los requisitos planteados en el enunciado de la práctica, por lo que se procedió a generar una salida en formato de fichero Excel (tFileOutputExcel_1). Esta se volcó a una tabla creada en la configuración del componente tDBOutput_1 y denotada como *resultado* en MySQL, la cual contiene 161 tuplas y los campos solicitados del nombre de artista, nombre del álbum, nombre de la canción y la posición de la canción en el álbum:

1	Led Zeppelin	Physical Graffiti	Ol' Riley	14
2	Led Zeppelin	Physical Graffiti	Take This Hammer	16
3	Led Zeppelin	Physical Graffiti	Corn Bread Rough	19
4	Led Zeppelin	Physical Graffiti	John Henry	20
5	Radiohead	The Bends	Mephisto Waltz No.1	1
6	The Beatles	Rubber Soul	Lady fish and chips	1
7	U2	Achtung Baby	Are You Going Home For Christma…	1
8	Led Zeppelin	Physical Graffiti	La Bamba	20
9	Led Zeppelin	Physical Graffiti	Another Rock and Roll Christmas	21
10	Led Zeppelin	Physical Graffiti	Weihnacht	12
11	The Beatles	Rubber Soul	Chor und Orchester Konrad Plaickn…	14
12	Radiohead	The Bends	arawak ée karayib	1
13	Radiohead	The Bends	Bailantas Costeiras	4
14	Led Zeppelin	Physical Graffiti	???? ?? ??? ?? ??? GURU	9
15	Pink Floyd	The Piper at …	Luna	3
16	Pink Floyd	The Piper at …	Circulos	5
17	Pink Floyd	The Piper at …	CD 2 - Happiness Now	10
18	The Beatles	Rubber Soul	Come Alone	1
19	U2	Achtung Baby	Oat Bran	9

Figura 13. Muestra de las 19 primeras tuplas para la tabla *resultado* en MySQL con los campos del nombre de artista, nombre del álbum, nombre de la canción y la posición de la canción en el álbum.

Cabe destacar en términos de eficiencia, que el tiempo aproximado del flujo de trabajo fue de 1 minuto y 10 segundos, lográndose el mayor tiempo en los filtrados y *join* realizados. De este modo, como se mencionó anteriormente, se ha atendido siempre a la eficiencia para construir un flujo de trabajo con este planteamiento, sin necesidad de valorar otras posibles opciones de optimización por coste.