IMDb Data Warehouse

Bases de Datos Avanzadas

Antonio Manjavacas, Miguel Millán, Desiré Saponaro

2 de marzo de 2020

Resumen

Los almacenes de datos, también denominados *Data Warehouses*, son bases de datos analíticas destinadas a la toma de decisiones así como a ofrecer perspectivas de alto nivel sobre las diferentes dimensiones de negocio. El principal objetivo de este trabajo es el desarrollo de un almacén de datos orientado al dominio de la industria cinematográfica, tomando como referencia los datos correspondientes a 5000 películas extraídos de *IMDb* y ofrecidos por la plataforma *Kaggle*.

Índice

1. Introducción

Definimos almacenes de datos (*Data Warehouses*) como bases de datos analíticas, integradas, no volátiles y variables en el tiempo orientadas a la toma de decisiones en un determinado ámbito. Generalmente, los almacenes de datos reúnen una gran cantidad de información empleada desde la perspectiva de la analítica de datos y la ingeniería de negocio. Estos datos se encuentran orientados a temas y son fruto de la unión de múltiples subconjuntos de sistemas de información de más bajo nivel denominados *Data marts*.

Una vez definido el concepto de *Data Warehouse*, así como sus aplicaciones, el objetivo de este proyecto será llevar la elaboración de nuestro propio almacén de datos. El área de aplicación elegida ha sido la industria cinematográfica, empleando los datos proporcionados por la plataforma *Kaggle* referentes a 5000 películas presentes en *IMDb*. El conjunto de datos, así como una definición detallada del mismo, puede consultarse en el siguiente enlace: `https://www.kaggle.com/tmdb/tmdb-movie-metadata`.

Así, los resultados esperados tras el desarrollo de este proyecto incluirán tanto la limpieza, transformación y carga de datos en el sistema *cloud* no relacional *MongoDB Atlas*, como la elaboración de las diferentes dimensiones del almacén de datos.

Los principales medios tecnológicos empleados han sido: el entorno de programación Python, *Google Colab*, orientado al desarrollo software colaborativo; *MongoDB* y, concretamente, su servicio en la nube *MongoDB Atlas*, para el almacenamiento de datos y LaTeXpara la elaboración de la memoria del proyecto.

2. Desarrollo

En las siguientes subsecciones se detallará el proceso de elaboración de nuestro almacén de datos, desde el diseño del mismo hasta la limpieza y carga de los datos en sus diferentes dimensiones.

2.1. Diseño y dimensiones del almacén de datos

El esquema en estrella es un modelo de datos consistente en una tabla central (tabla de hechos) que contiene el conjunto de registros a analizar, así como una serie de tablas de dimensiones que la rodean. Dichas dimensiones ofrecen información más concreta y facilitan las labores de análisis sobre el conjunto de datos.

En este caso, el almacén de datos es una base de datos sobre películas, así que como tabla central tenemos la tabla con los datos de todas las películas de las que disponemos (Figura 1).

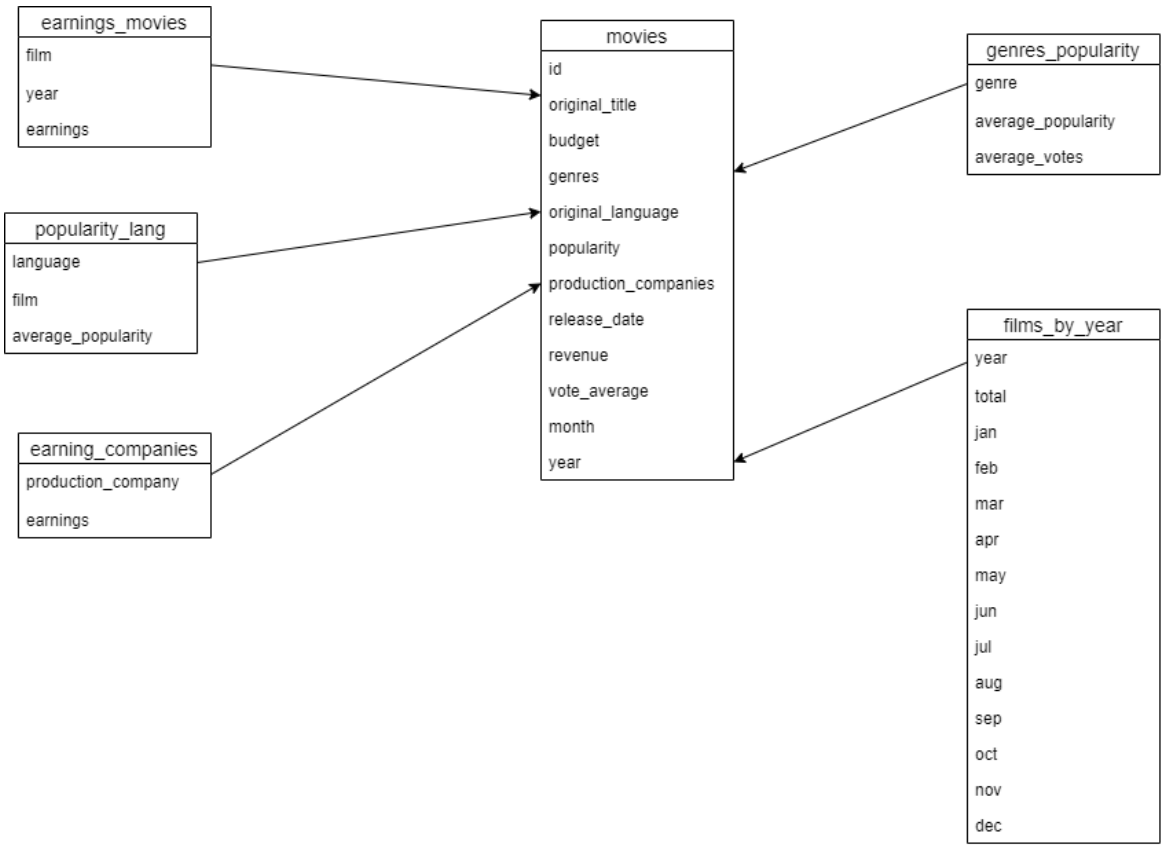

Figura 1: Diseño del almacén de datos

Por otro lado, contamos con las siguientes dimensiones:

- **Dimensión 1**. La primera dimensión está compuesta por las películas que más dinero generaron por año, calculada basándonos en la diferencia entre presupuesto de la película y las ganancias totales. Véase la Figura 2.

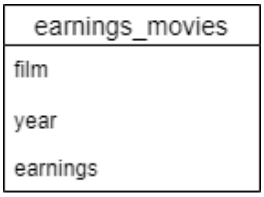

Figura 2: Dimensión earnings_movies

- **Dimensión 2**. Muestra la película más popular para cada idioma. Las consultas a esta tabla pueden ser de utilidad para estudiar que tipo de película gusta más en función del idioma que hablan los espectadores. Véase la Figura 3.

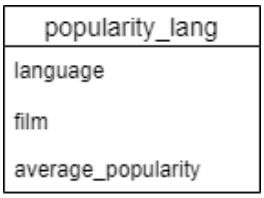

Figura 3: Dimensión popularity_lang

- **Dimensión 3**. Representa las ganancias totales de las diferentes compañías productoras. Estos datos permiten conocer qué productoras han generado mayores beneficios. Véase la Figura 4.

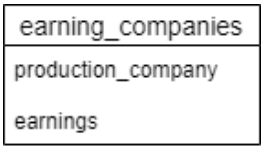

Figura 4: Dimensión earning_companies

- **Dimensión 4**. Compuesta por la popularidad y nota medias entre las películas de cada género. Con dicha información es posible conocer que géneros atraen más al público. Véase la Figura 5.

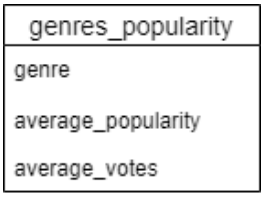

Figura 5: Dimensión genres_popularity

- **Dimensión 5**. Muestra el número de películas realizadas por año, además del número de películas por mes dentro de dicho año. Su utilidad reside en poder conocer en qué año se estrenaron más películas o en qué suelen producirse más estrenos. Véase la Figura 6.

2.2. Limpieza y carga de datos

Una vez definido el diseño y dimensiones de nuestro *Data Warehouse*, procedemos a explicar cómo se llevó a cabo la limpieza y transformación del conjunto de datos inicial de cara a su posterior carga en *MongoDB Atlas*. Las diferentes labores del proceso de *data cleaning* se definen en los siguientes puntos:

- **Extracción de las características de interés**. Las columnas extraídas del conjunto de datos original fueron:
  - Identificador de la película.
  - Título original.
  - Presupuesto.
  - Géneros.
  - Idioma original.

Figura 6: Dimensión films_by_year

- Popularidad.
- Compañías productoras.
- Fecha de lanzamiento.
- Ganancias.
- Puntuación media.

- **Eliminación de registros con valores nulos**. Se decidió optar por la eliminación de aquellas películas con información incompleta, ya que podrían perjudicar a los resultados finales.
- **Conversión del formato de la columna *géneros***. Se modificó el formato de esta columna, inicialmente expresada como JSON, y se convirtió en una lista de géneros, eliminando su identificador.
- **Conversión del formato de la columna *compañías productoras***. El procedimiento fue similar al caso anterior: se eliminaron identificadores y las compañías productoras quedaron expresadas como listas.
- **Conversión del formato de las fechas de lanzamiento**. Las fechas, inicialmente expresadas en formato AA/MM/DD, fueron convertidas al formato DD/MM/AA.
- **Creación de las columnas *month* y *year***. Estas columnas facilitarían el trabajo en la elaboración de la mayoría de las dimensiones.

Una vez procesados los datos, contamos con un total de 4802 registros que fueron cargados en la base de datos.

3. Resultados y conclusiones

Una vez las diferentes dimensiones fueron cargadas en la base de datos, se procedió a la recuperación de información ya ubicada en la base de datos en la nube. Los resultados correspondientes a dichas consultas son los que se muestran en las Figuras 7, 8, 9, 10 y 11.

	year	film	earnings
0	1916	Intolerance	8.008844e+06
1	1925	The Big Parade	2.175500e+07
2	1927	Metropolis	-9.196958e+07
3	1929	The Broadway Melody	3.979000e+06
4	1930	Hell's Angels	4.050000e+06
...	...	...	...
85	2013	Frozen	1.124219e+09
86	2014	Transformers: Age of Extinction	8.814051e+08
87	2015	Jurassic World	1.363529e+09
88	2016	Captain America: Civil War	9.033045e+08
89	2017	Growing Up Smith	0.000000e+00

Figura 7: Consulta a la base de datos: dimensión *earnings movies*

	year	total	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
0	1916	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
1	1925	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	1927	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	1929	2.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	1930	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	2013	231.0	19.0	18.0	17.0	15.0	16.0	15.0	26.0	18.0	26.0	23.0	14.0	24.0
86	2014	238.0	22.0	11.0	18.0	18.0	19.0	20.0	16.0	24.0	24.0	28.0	15.0	23.0
87	2015	216.0	13.0	19.0	17.0	17.0	10.0	14.0	19.0	22.0	26.0	27.0	17.0	15.0
88	2016	104.0	17.0	14.0	13.0	11.0	13.0	12.0	15.0	4.0	4.0	1.0	0.0	0.0
89	2017	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figura 8: Consulta a la base de datos: dimensión *films by year*

	genre	average_popularity	average_votes
0	Action	30.940382	5.989515
1	Adventure	39.268042	6.156962
7	Animation	38.813439	6.341453
10	Comedy	18.221001	5.945587
4	Crime	22.853274	6.274138
17	Documentary	3.945724	6.238182
5	Drama	17.764853	6.388594
8	Family	27.832849	6.029630
2	Fantasy	36.387043	6.096698
18	Foreign	0.686787	6.352941
14	History	17.444839	6.719797
12	Horror	18.295915	5.626590
16	Music	13.101512	6.355676
13	Mystery	24.586827	6.183908

Figura 9: Consulta a la base de datos: dimensión *genres popularity*

	production_company	earnings
1536	"DIA" Productions GmbH & Co. KG	8350926.0
4808	1.85 Films	-401983.0
4325	10 West Studios	0.0
2258	100 Bares	33965843.0
4362	1019 Entertainment	801508.0
...	...	...
2250	verture Films	0.0
1350	warner bross Turkey	280170008.0
3069	winchester films	0.0
4577	África Filmes	0.0
2927	Österreichischer Rundfunk (ORF)	73841160.0

Figura 10: Consulta a la base de datos: dimensión *earnings companies*

	language	film	average_popularity
24	af	Tsotsi	2.504169
27	ar	The Square	4.892203
10	cn	葉問3	19.167377
19	cs	Obsluhoval jsem anglického krále	2.387463
16	da	What Happens in Vegas	38.100488
5	de	Das Leben der Anderen	34.938177
36	el	Κυνόδοντας	28.858238
0	en	Minions	875.581305
4	es	El laberinto del fauno	90.809408
32	fa	جدایی نادر از سیمین	12.049373
2	fr	Le fabuleux destin d'Amélie Poulain	73.720244
26	he	Vals Im Bashir	14.082510
6	hi	Dabba	14.017809

Figura 11: Consulta a la base de datos: dimensión *popularity lang*

Los resultados de este proyecto han sido satisfactorios y nos han permitido comprender el proceso ETL (extracción, transformación y carga de datos) de un almacén de datos de forma práctica. Toda la información referente al código empleado puede consultarse en `https://github.com/manjavacas/BBDD-Avanzadas`.