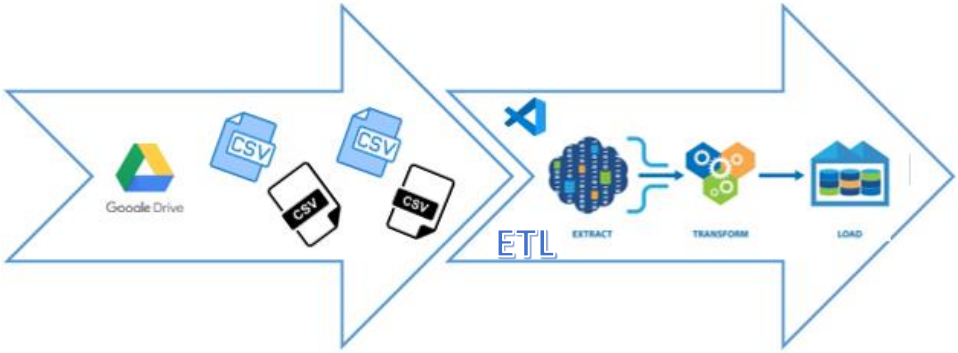
Informe de ETL



El proceso de ETL (Extract, Transform, Load) desempeña un papel crucial en la preparación y manipulación de los datos extraídos desde una fuente de datos, con el objetivo de convertirlos en un formato adecuado y coherente para su posterior análisis.

-El primer paso del proceso ETL es la extracción de datos, donde se establece una conexión con Google Drive y se selecciona el archivo CSV de interés. Posteriormente, se procede a leer el archivo CSV y extraer los datos contenidos en él.

-Una vez extraídos los datos, comienza la etapa de transformación. Durante esta fase, se aplican diversas técnicas para garantizar la calidad y coherencia de los datos. Esto incluye la eliminación de valores duplicados, la corrección de errores y la normalización de los formatos de datos.

-Tras la etapa de transformación, se realiza la carga de los datos en el formato adecuado para su posterior análisis.

En este informe, se detallará:

- La transformación de datos (técnicas utilizadas para transformar los datos extraídos, como la corrección de errores, la normalización de formatos y más)

Lenguaje utilizado:

- Python

Librerías utilizadas:

- Pandas
- Numpy
- Datetime

Giselle Pereira Nuñez

Se me entregaron archivos de plataformas de steaming:

- amazon_prime_titles.csv
- disney_plus_titles.csv
- hulu_titles.csv
- netflix_titles.csv

En la imagen se puede ver cómo eran al comienzo.

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	March 30, 2021	2014	NaN	113 min	Comedy, Drama	A small fishing village must procure a local d...
1	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	March 30, 2021	2018	13+	110 min	Drama, International	A Metro Family decides to fight a Cyber Crimin...
2	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R...	United States	March 30, 2021	2017	NaN	74 min	Action, Drama, Suspense	After a man discovers his wife is cheating on ...

Cada uno contenían los campos:

- show_id: Este campo identifica de manera única cada entrada en el conjunto de datos (alfanumérico).
- type: Indica si la entrada se refiere a una película o a una serie.
- title: El título de la película.
- director: El nombre del director de la película.
- cast: Los actores principales que participan en la película.
- country: El país de origen de la película.
- date_added: La fecha en la que la película fue añadida a la plataforma. Con formato “Mes día, año”.
- release_year: El año de lanzamiento original de la película.
- rating: La clasificación de la película.

- duration: La duración de la película (en minutos) y cuantos capítulos tiene en caso de las series.
- listed_in: Las categorías o géneros en los que se clasifica la película.
- description: Una breve descripción o sinopsis de la trama de la película.

Transformaciones:

1) Generar campo id: Cada id se compondrá de la primera letra del nombre de la plataforma, seguido del show_id ya presente en los datasets (ejemplo para títulos de Amazon = as123)

2) Los valores nulos del campo rating deberán reemplazarse por el string "G" (corresponde al maturity rating: "general for all audiences"

3) Cambiar el formato de la columna "date_added" al formato AAAA-mm-dd.

4) Pasar todos los textos a minúscula, sin excepción.

5) El campo "duration" debe convertirse en dos campos: "duration_int" y "duration_type". El primero será un integer y el segundo un string indicando la unidad de medición de duración: min (minutos) o season (temporadas).

6) Eliminar las columnas "show_id" y "duration", ya que fueron reemplazadas en los puntos 1 y 5.

7) Unir los datasets, obteniendo un resultado total de 22.998 (veintidós mil novecientos noventa y ocho) filas y 13 (trece) columnas.

8) En el campo "duration_type" se encuentran dos formas de identidicar a las series (seasons y season), estas son todas pasadas a solo 'season'.

Giselle Pereira Nuñez

En la próxima imagen vera los tres primeros renglones, luego de las transformaciones.

	id	type	title	director	cast	country	date_added	release_year	rating	listed_in	description	duration_int	duration_type
0	as1	movie	the grand seduction	don mckellar	brendan gleeson, taylor kitsch, gordon pinsent	canada	2021-03-30	2014	g	comedy, drama	a small fishing village must procure a local d...	113	min
1	as2	movie	take care good night	girish joshi	mahesh manjrekar, abhay mahajan, sachin khedekar	india	2021-03-30	2018	13+	drama, international	a metro family decides to fight a cyber crimin...	110	min
2	as3	movie	secrets of deception	josh webber	tom sizemore, lorenzo lamas, robert lasardo, r	united states	2021-03-30	2017	g	action, drama, suspense	after a man discovers his wife is cheating on ...	74	min

Se me entregaron archivos de reseñas:

En la imagen se puede ver cómo eran al comienzo.

	userId	rating	timestamp	movieId
0	1	1.0	1425941529	as680
1	1	4.5	1425942435	ns2186
2	1	5.0	1425941523	hs2381
3	1	5.0	1425941546	ns3663

Son 8 archivos, en donde cada uno contiene:

- userId: Representa el identificador único del usuario que ha proporcionado una calificación para una película.
- rating: Indica la calificación dada por el usuario a una película en particular.
- timestamp: Representa la marca de tiempo en la que se registró la calificación dada por el usuario. Es un valor numérico que generalmente se presenta en segundos.
- movieId: Es el identificador único de la película a la que se le ha dado una calificación por parte del usuario. El movieId se utiliza para asociar

la calificación del usuario con una película específica en un conjunto de datos o sistema de recomendación.

Transformaciones:

1) Se concatenan los 8 archivos, resultando un único dataset de 11.024.289 (once millones veinticuatro mil doscientos ochenta y nueve) filas y 4 (cuatro) columnas.

2) Renombrar campo “rating” (ya que este ya se encuentra en el dataset de plataformas) a “score”.

3) Convertir el campo “timestamp” que se encontraba en formato de segundos, al formato adecuado: AAAA-mm-dd.

4) Unión de los datasets donde están contenidas las plataformas y reseñas, en base al campo “id” en plataformas y al campo “movieId” en reseñas. Quedando un único dataset llamado ‘Proveedores_Reseñas’ con 11.024.289 (once millones veinticuatro mil doscientos ochenta y nueve) filas y 18 (dieciocho) columnas.

5) Crear un nuevo campo llamado “plataform”, donde se encontrarán los nombres de las plataformas.

En la próxima imagen vera los tres primeros renglones, luego de las transformaciones (se ve la misma película calificada por 3 diferentes usuarios).

	id	type	title	director	cast	country	date_added	release_year	rating	listed_in	description	duration_int	duration_type	userId	score	timestamp	movieId	plataform
0	as1	movie	the grand seduction	don mckellar	brendan gleeson, taylor kitsch, gordon pinsent	canada	2021-03-30	2014	g	comedy, drama	a small fishing village must procure a local d...	113	min	543	5.0	2003-07-30	as1	amazon
1	as1	movie	the grand seduction	don mckellar	brendan gleeson, taylor kitsch, gordon pinsent	canada	2021-03-30	2014	g	comedy, drama	a small fishing village must procure a local d...	113	min	595	3.0	1996-08-13	as1	amazon
2	as1	movie	the grand seduction	don mckellar	brendan gleeson, taylor kitsch, gordon pinsent	canada	2021-03-30	2014	g	comedy, drama	a small fishing village must procure a local d...	113	min	611	3.0	2001-01-03	as1	amazon

Los datos finales son guardados en un CSV, llamado “archivo.csv”.

Este archivo resultante es utilizado para realizar el EDA (Exploratory Data Analysis), cuyos pasos sin explicados en el Informe_EDA.