July Daniela Ramos Peña- ADSO 2901817

Instructor: Jesús Ariel Gonzales Bonilla

- **Conceptualización:**

- Investiga y define qué es un proceso ETL.

    - ETL es un proceso para combinar datos de diferentes fuentes y llevarlos a un almacén de datos para utilizarlo y analizarlo. ETL es extraer, transformar y cargar, lo que hace este proceso es extraer los datos de donde estén para recopilarlos, transformarlo a lo que se necesite ya sea que haya que limpiarlos o filtrarlos o otras cosas y luego se carga a su nuevo destino. Permitiendo esto organizar los datos para luego poder analizarlos o usarlos.

- Explica la importancia del ETL en proyectos de análisis de datos.

    - Es importante porque permite la recopilación, limpieza y carga de datos de diferentes lo que hace que sea más fácil de analizarlos, de llevar a cabo procesos con esos datos, que sean datos precisos y reales. Porque el ETL toma datos de diferentes fuentes y las vuelve una haciendo así mucho más fácil el hecho de recopilación y utilización de los datos.

- **Tipos de herramientas ETL:**
- Investiga y describe al menos 3 herramientas ETL (pueden ser open source o comerciales). Menciona ventajas y desventajas de cada una.

    - Talend Open Studio: plataforma que permite integrar datos, extraer, transformar y cargar datos desde diferentes fuetes, funciona con tablas y componentes reutilizables
      Ventajas:
        - Es gratis
        - Buena comunidad y documentación lo que permite la facilidad de resolver dudad o problemas.

      Desventajas:
        - Es pesado o lento en equipos de pocos recursos
        - La versión gratuita trae limitaciones

    - Microsoft SQL Server Integration Services (SSIS): Es una herramienta de Microsoft integrada en sql server. Para transformar datos. Viene con conectores integrados para extraer datos.
      Ventajas:

- Integración total con SQL Server y el ecosistema Microsoft
- Alto rendimiento en entornos Windows.

Desventajas:

- Requiere licencia de SQL server para utilizarlo
- Menos flexible cuando no es Microsoft.

- Pentaho Data Integration (PDI): es una herramienta ETL ofrecida por Hitachi. Captura datos de diversas fuentes, los limpia y los almacena en un formato uniforme y coherente.
  Ventajas:
  - Gratuita y con una comunidad activa.
  - Soporta gran variedad de fuentes y formatos de datos.

  Desventajas:
  - Interfaz un poco antigua comparada con herramientas modernas.
  - Puede ser menos eficiente con volúmenes de datos muy grandes.

3. Actividad práctica:

- Utiliza un archivo de datos de tu elección para realizar un proceso ETL básico:
- Extracción: Lee los datos desde el archivo.
- Transformación: Realiza al menos dos transformaciones (por ejemplo: limpieza de datos,
- cambio de formato, filtrado de columnas, etc.).
- Carga: Exporta el resultado a un nuevo archivo (puede ser Excel, CSV o base de datos
- simple).
- Puedes usar Python (pandas), Power BI, Talend, o la herramienta que prefieras.

4. Demostración:

- Documenta el proceso realizado (puedes usar capturas de pantalla, código fuente, o un pequeño informe).
- Explica los retos encontrados y cómo los resolviste.

- Primero importe el archivo csv
  aquí experimente mi primer reto porque quería manejar el colab en el visual studio directamente, pero al final no pude llamar al archivo correctamente entonces abrí el colab directamente en el drive y lo maneje desde ahí.

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('65k_anime_data.csv')
df.head()
```

Luego de impórtalo y leerlo me muestra a continuación la tabla con los datos del archivo

	title	title_english	title_japanese	image	aired_from
0	Cowboy Bebop	Cowboy Bebop	カウボーイビバップ	https://cdn.myanimelist.net/images/anime/4/196...	1998-04-03 00:00:00+00:00
1	Cowboy Bebop: Tengoku no Tobira	Cowboy Bebop: The Movie	カウボーイビバップ 天国の扉	https://cdn.myanimelist.net/images/anime/1439/...	2001-09-01 00:00:00+00:00
2	Trigun	Trigun	トライガン	https://cdn.myanimelist.net/images/anime/1130/...	1998-04-01 00:00:00+00:00
3	Witch Hunter Robin	Witch Hunter Robin	Witch Hunter ROBIN (ウイッチハンターロビン)	https://cdn.myanimelist.net/images/anime/10/19...	2002-07-03 00:00:00+00:00
4	Bouken Ou Beet	Beet the Vandel Buster	冒険王ビィト	https://cdn.myanimelist.net/images/anime/7/215...	2004-09-30 00:00:00+00:00

Por consiguiente implemente dos tipos de transformación:

La primera fue para eliminar datos nulos de la tabla con el dropna()

```python
# Eliminar filas con valores nulos
df = df.dropna()

print("Después de eliminar filas con datos nulos:")
print(df.head())
```

```
Después de eliminar filas con datos nulos:
                  title        title_english title_japanese  \
2               Trigun              Trigun          トライガン
6   Hachimitsu to Clover     Honey and Clover     ハチミツとクローバー
9              Monster             Monster          モンスター
10              Naruto              Naruto            ナルト
12   Tennis no Oujisama  The Prince of Tennis       テニスの王子様

                                           image  \
2   https://cdn.myanimelist.net/images/anime/1130/...
6   https://cdn.myanimelist.net/images/anime/1301/...
9   https://cdn.myanimelist.net/images/anime/10/18...
10  https://cdn.myanimelist.net/images/anime/1141/...
12  https://cdn.myanimelist.net/images/anime/6/216...

                  aired_from                   aired_to  \
2   1998-04-01 00:00:00+00:00  1998-09-30 00:00:00+00:00
6   2005-04-15 00:00:00+00:00  2005-09-27 00:00:00+00:00
9   2004-04-07 00:00:00+00:00  2005-09-28 00:00:00+00:00
10  2002-10-03 00:00:00+00:00  2007-02-08 00:00:00+00:00
12  2001-10-10 00:00:00+00:00  2005-03-23 00:00:00+00:00
```

A continuación, se muestra la eliminación de columnas nulas:

Y la segunda transformación fue filtrar las tablas por dos datos que escogí título y episodios

Se ve el código y el resultado debajo de el.

```
    # Conservar solo columnas de interés (ejemplo: 'title' y 'episodes')
    df = df[['title', 'episodes']]

    print("Después de filtrar columnas:")
    print(df.head())
```

```
Después de filtrar columnas:
                    title  episodes
2                  Trigun      26.0
6   Hachimitsu to Clover       24.0
9                 Monster      74.0
10                 Naruto     220.0
12    Tennis no Oujisama      178.0
```

Por ultimo hice la carga, guarde el resultado en un nuevo archivo Excel(xlsx):

```
# Guardar el resultado en un archivo Excel
df.to_excel("nuevo.xlsx", index=False)

print("Archivo transformado guardado como 'nuevo.xlsx'")
```

```
Archivo transformado guardado como 'nuevo.xlsx'
```

5. Análisis solicitado:

- Realiza un análisis simple sobre los datos transformados (por ejemplo: estadísticas descriptivas, gráficos, tendencias, etc.).
- Expón tus conclusiones

```
import pandas as pd



# Número de animes
total_animes = len(df)

# Promedio de episodios
promedio_episodios = df["episodes"].mean()

# Anime con más episodios
anime_mas_largo = df.loc[df["episodes"].idxmax()]

# Anime con menos episodios
anime_mas_corto = df.loc[df["episodes"].idxmin()]

print(f"Total de animes: {total_animes}")
print(f"Promedio de episodios: {promedio_episodios:.2f}")
print(f"Anime con más episodios: {anime_mas_largo['title']} ({anime_mas_largo['episodes']} episodios)")
print(f"Anime con menos episodios: {anime_mas_corto['title']} ({anime_mas_corto['episodes']} episodios)")
```

```
Total de animes: 367
Promedio de episodios: 39.74
Anime con más episodios: Doraemon (1979) (1787.0 episodios)
Anime con menos episodios: I'll/CKBC (2.0 episodios)
```

En este conjunto de datos hay 367 animes.

En promedio, cada uno tiene unos 40 episodios, lo que significa que la mayoría son series cortas o de duración media.

El anime más largo es Doraemon (1979), con 1787 episodios, una serie que ha estado en emisión por muchos años.

El más corto es I'll/CKBC, con solo 2 episodios, probablemente una miniserie o especial.

Esta gran diferencia muestra que en el mundo del anime hay de todo: desde series muy largas que duran décadas, hasta historias muy cortas que se ven en un par de horas.