**Nombre: Santiago Almeida**

**CC: 1193119436**

**1. Arquitectura de Datos**

La arquitectura de datos de esta aplicación sigue el patrón ETL (Extracción, Transformación y Carga), que es un enfoque común para procesar y analizar grandes volúmenes de datos. A continuación, se describen los componentes principales de esta arquitectura:

1. **Extracción de Datos (Extract):**
   * **Fuente de Datos:** Los datos se extraen de archivos Excel que contienen información sobre películas, inventarios, alquileres, clientes y tiendas.
   * **Herramienta de Extracción:** Se utiliza PySpark para leer los archivos Excel y cargar los datos en DataFrames de Spark.
   * **Clase**[Extraccion](vscode-file://vscode-app/c:/Users/Usuario/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-sandbox/workbench/workbench.html" \o ")**:** Esta clase se encarga de leer los datos de los archivos Excel y devolverlos en forma de DataFrame de Spark.
2. **Transformación de Datos (Transform):**
   * **Limpieza de Datos:** Los datos se limpian para eliminar caracteres especiales, estandarizar formatos y manejar valores nulos.
   * **Estandarización de Datos:** Se estandarizan los datos para asegurar la consistencia en los formatos de texto y fechas.
   * **Clase**[Transformacion](vscode-file://vscode-app/c:/Users/Usuario/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-sandbox/workbench/workbench.html" \o ")**:** Esta clase se encarga de realizar las transformaciones necesarias en los DataFrames, incluyendo la limpieza y estandarización de datos.
3. Carga de Datos (Load)

* Destino de Datos: Los datos transformados se guardan en archivos CSV en un directorio de salida.
* Clase Carga: Esta clase se encarga de guardar los DataFrames transformados en archivos CSV.

**2. Arquetipo de la Aplicación**

El arquetipo de la aplicación sigue una estructura modular, donde cada módulo se encarga de una parte específica del proceso ETL. A continuación, se describe la estructura del proyecto y el flujo de trabajo de la aplicación.

**2.1. Estructura del proyecto:**

DATA\_ENGINEERING\_PROYECT/

├── data/

│ ├── originales/

│ │ └── Films\_2.xlsx

│ └── salidas/

├── ETL/

│ ├── extract.py

│ ├── transform.py

│ └── load.py

├── report/

│ └── informe.ipynb

├── main.py

└── ...