

EVALUACIÓN DE HABILIDADES EN UN DESARROLLO DE UN FLUJO ETL

LUIS ARMANDO FLORES CRUZ

09 DE FEBRERO DE 2025

# Objetivo

Evaluar tus habilidades en el desarrollo de un flujo ETL, utilizando como herramientas Python, SQL y Github

## Esquematización y Diseño

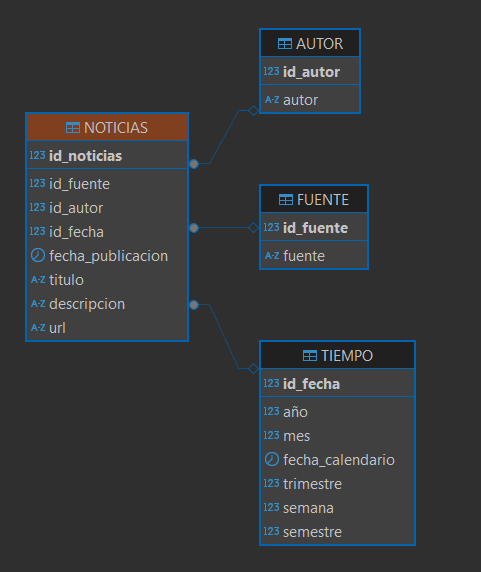
Para esta evaluación consideré utilizar railway, MySQL y Pyhton ya que las herramientas ya que las herramientas son flexibles, ergonómicas para el desarrollo de un reporte de noticias.

## Extracción de datos desde una API Publica

Opté por utilizar NewsAPI por que vi que el manejo de documentación y la construcción de la consulta es más rápida, consideré usar el servicio gratuito para cumplir con la evaluación.

## Solución

### Modelo Entidad Relación

El modelo estrella es adecuado para un reporte o indicador de noticias.

Noticias es la tabla histórica o transaccional y las constelaciones conectadas son las de autor, fuente y tiempo.

Con este modelo controlamos en una tabla de hechos los IDs y fechas y el rendimiento de la tabla será adecuado para un gran volumen de información.

Las dimensiones de autor y fuentes las construí tomándolo de la API, así aseguramos tener un catálogo de datos maestros.

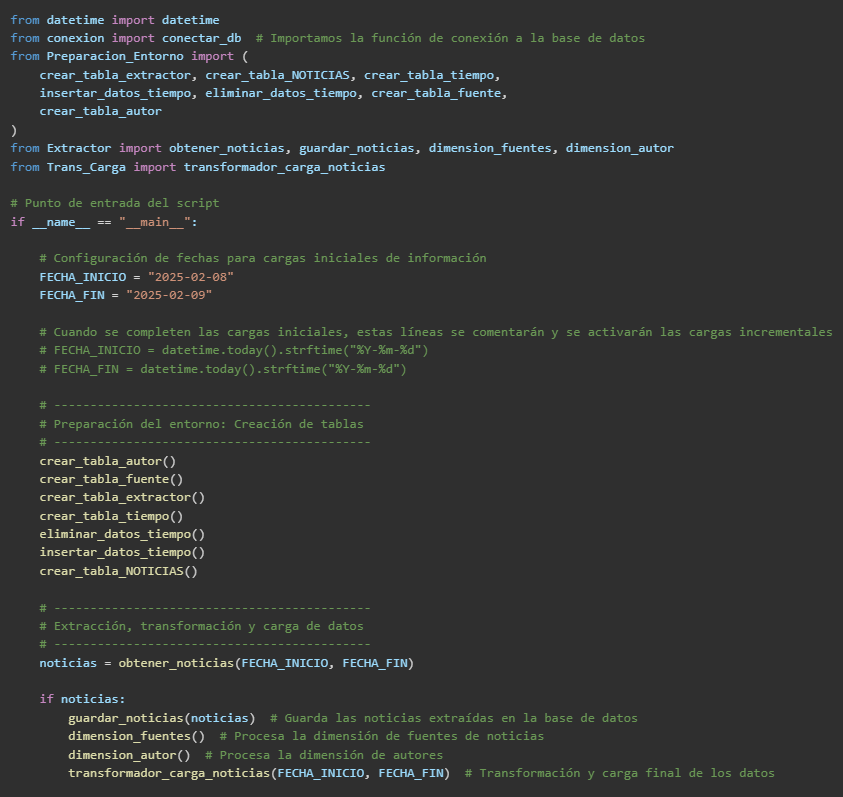
La dimensión tiempo es para poder filtrar la información por cualquier segmento de fecha (año, mes , trimestre , semana), para el diseño de un reporte o indicador es mas ergonómico incluir estos campos.

## Explicación del Data pipeline

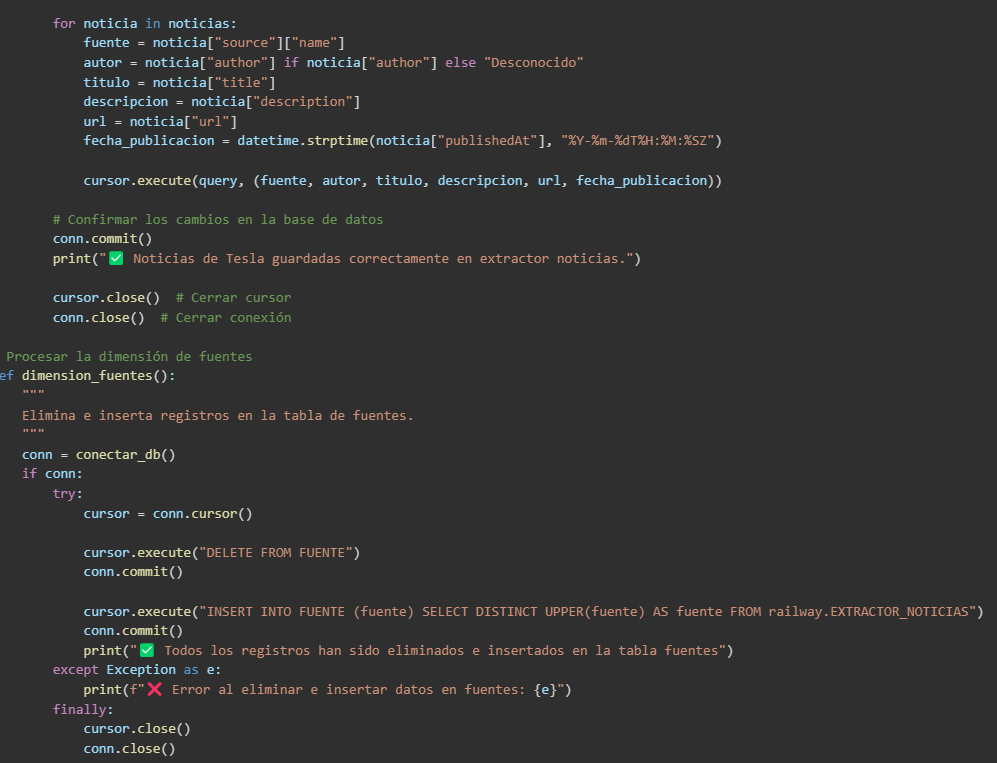
* Conexión a la Base de Datos
  + Archivo: conexion.py
  + Configura la conexión con MySQL en Railway.
* Preparación de entorno
  + Archivo: Preparacion\_Entorno.py
  + Crea las dimensiones AUTOR, FUENTE, EXTRACTOR\_NOTICIAS
  + Creación y llenado de información de la tabla TIEMPO
  + Creación de la tabla NOTICIAS
* Extracción de Datos
  + Archivo: Extractor.py
  + Obtiene noticias de la API NewsAPI sobre Tesla.
  + Guarda los datos en la tabla EXTRACTOR\_NOTICIAS.
  + También inserta datos en las tablas FUENTE y AUTOR.
* Transformación y Carga
  + Archivo: Trans\_Carga.py
  + Toma los datos de EXTRACTOR\_NOTICIAS, limpia y normaliza títulos y descripciones.
  + Carga los datos en la tabla final NOTICIAS.
* Ejecución del Proceso Completo
  + Archivo: Main.py
  + Crea las tablas necesarias y borra datos antiguos.
  + Extrae, transforma y carga los datos ejecutando todas las funciones en orden.

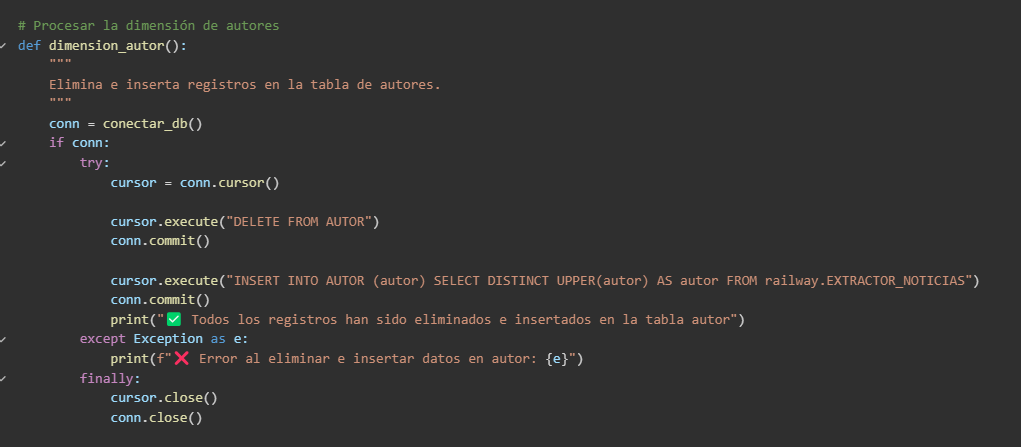
Scripts incluyendo comentarios:

Main.py:

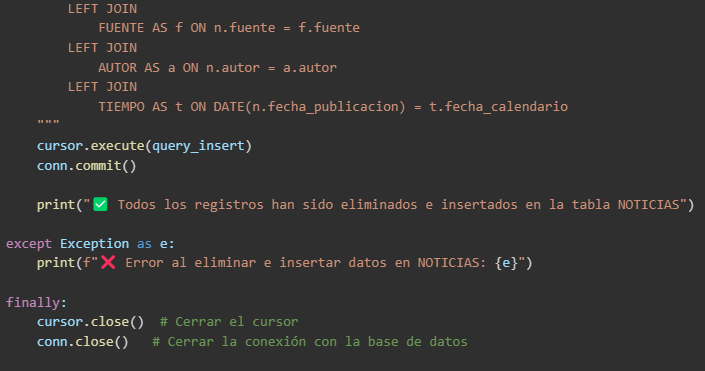
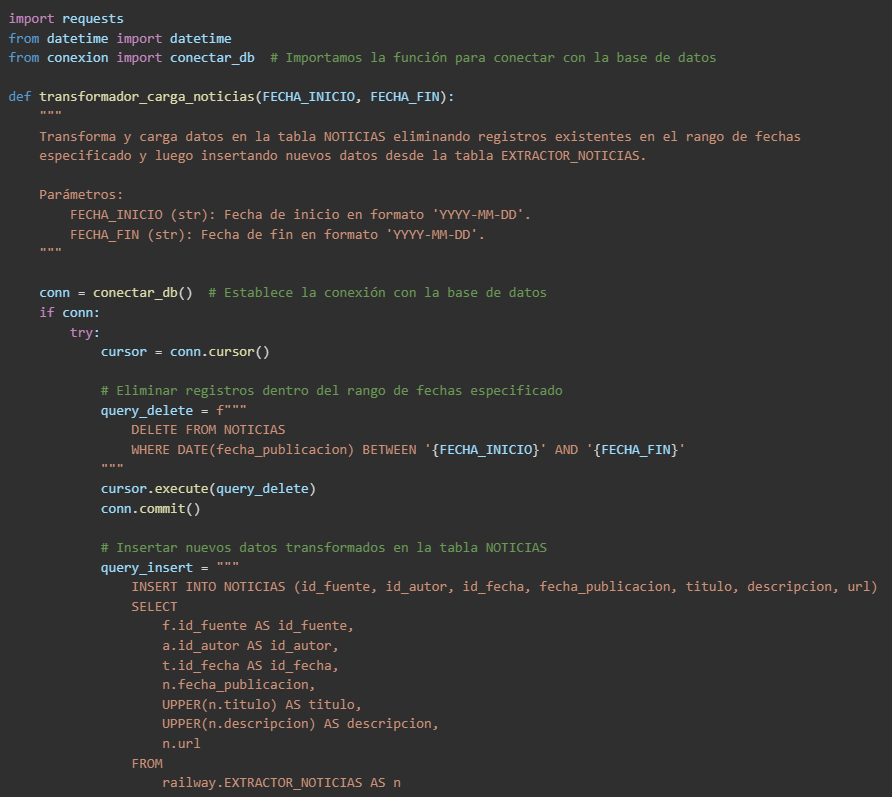


Extractor.py:

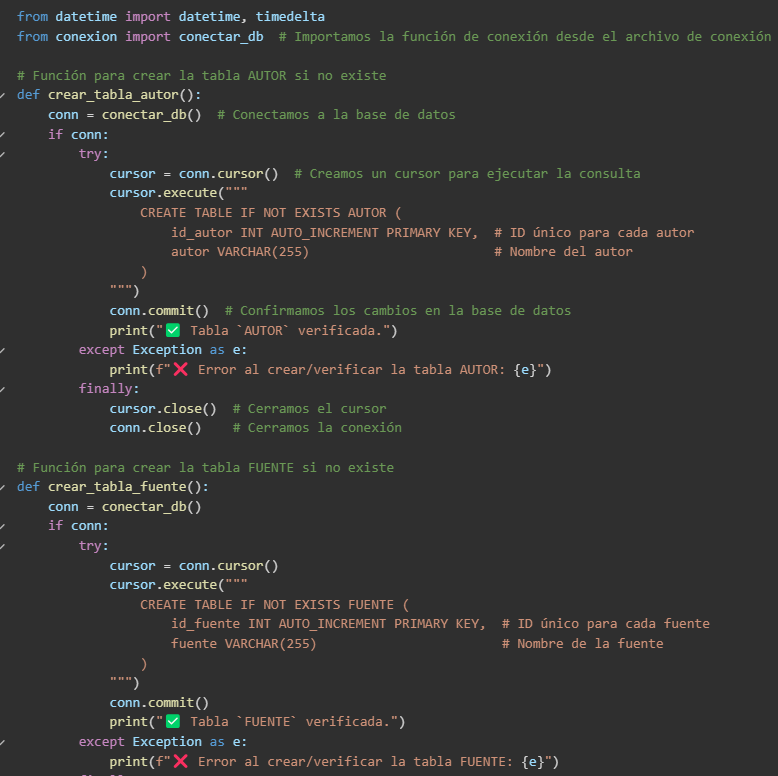


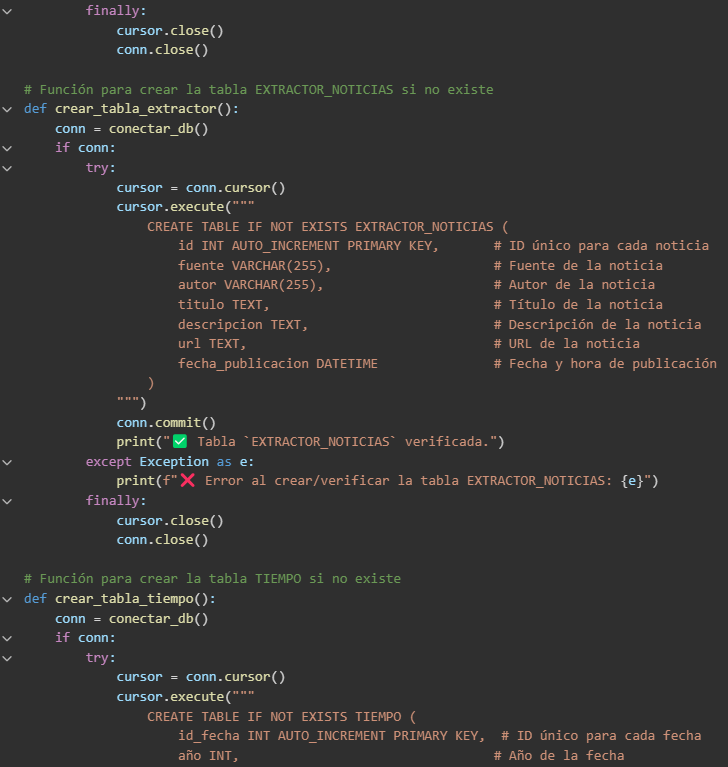


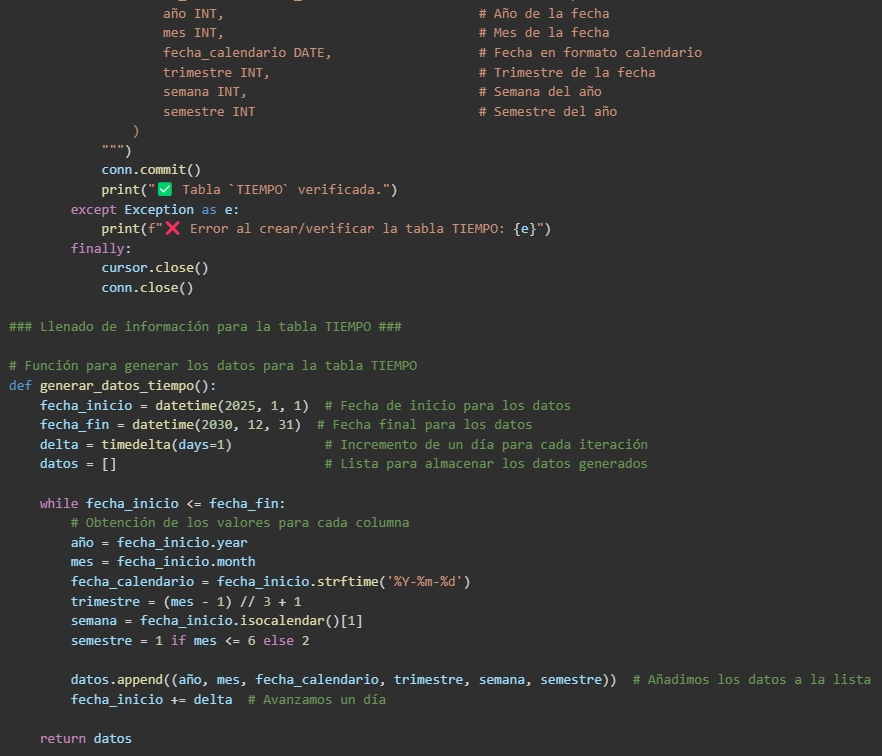
Trans\_Carga.py

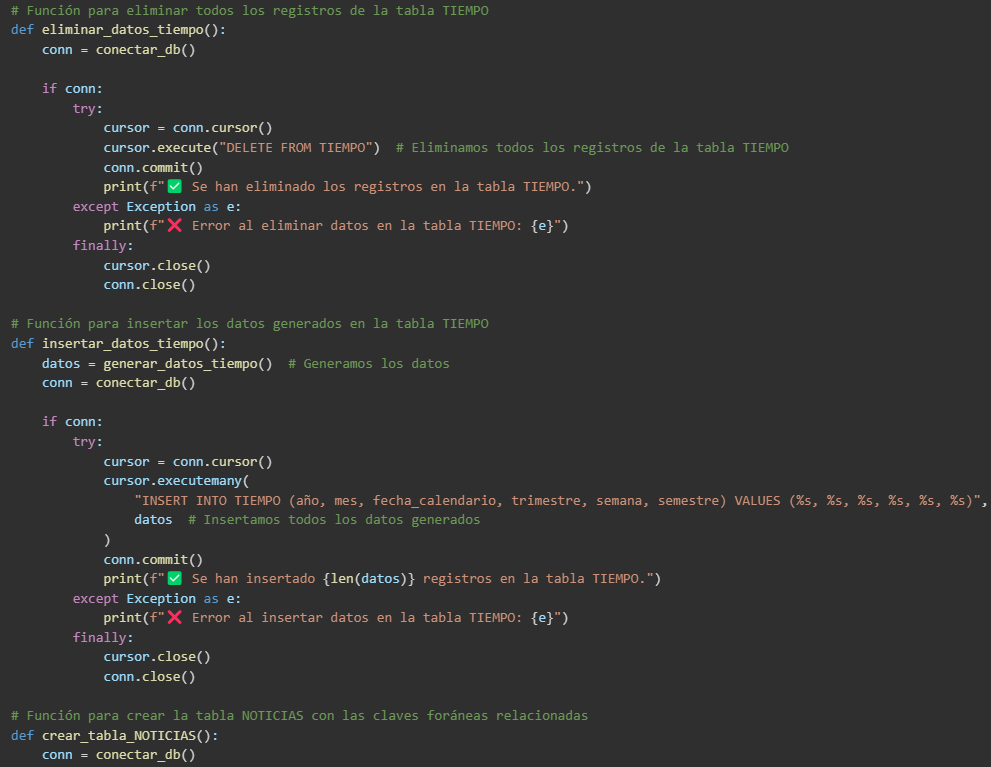


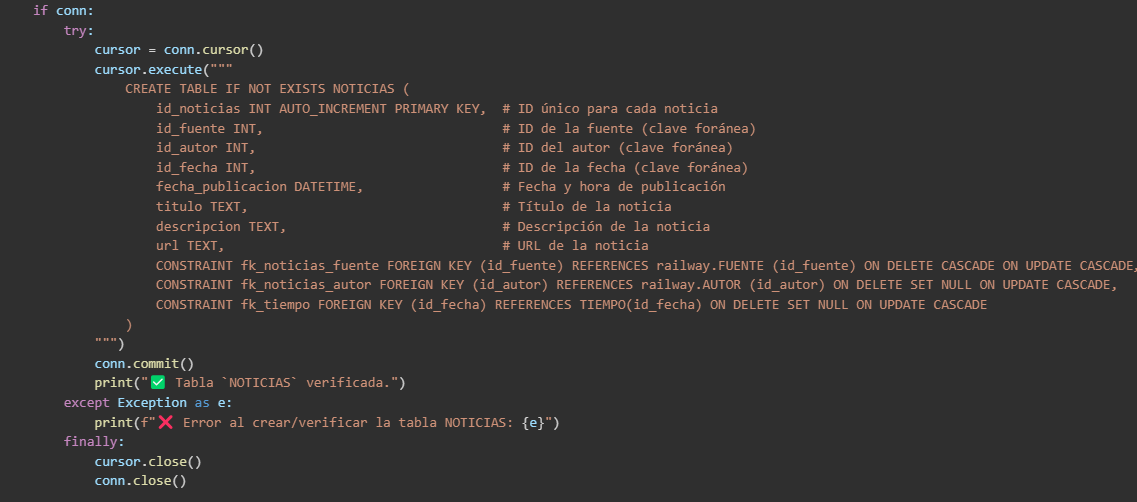
Preparacion\_entorno.py



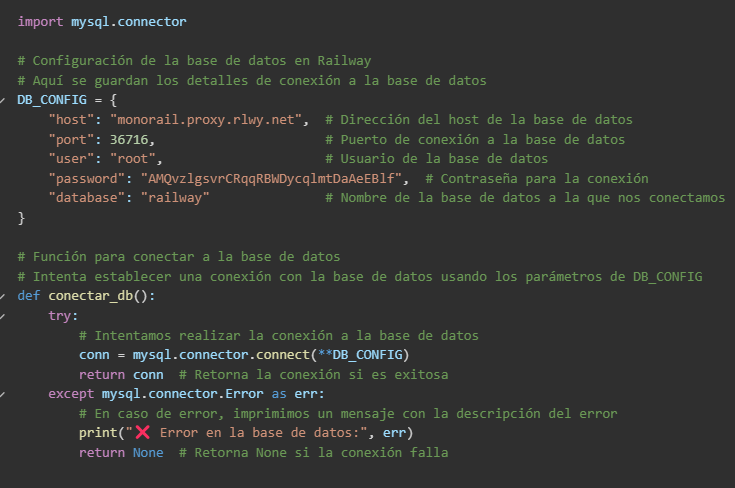








Conexión.py



## README

# Proyecto de Base de Datos con MySQL

Este proyecto maneja la creación y gestión de tablas en una base de datos MySQL para almacenar información relacionada con autores, fuentes, noticias, y tiempos. Está diseñado para ejecutarse con Python, utilizando la biblioteca `mysql-connector` para interactuar con MySQL.

## Requisitos previos

Antes de comenzar, asegúrate de tener instalado lo siguiente en tu entorno:

- \*\*Python 3.x\*\*: Asegúrate de tener una versión reciente de Python instalada. Puedes verificar tu versión ejecutando el siguiente comando en tu terminal o consola:

```bash

python --version

###MySQL Server: Debes tener una instancia de MySQL en ejecución. Si estás utilizando servicios en la nube, asegúrate de tener las credenciales correctas para acceder a la base de datos.

Instalación

Clona el repositorio (o descarga los archivos del proyecto). Si estás usando Git, ejecuta:

bash

Copiar

Editar

git clone <URL\_DEL\_REPOSITORIO>

cd <nombre\_del\_repositorio>

Crea un entorno virtual (opcional pero recomendado): Si no tienes un entorno virtual, puedes crear uno con los siguientes comandos:

bash

Copiar

Editar

python -m venv env

Activa el entorno virtual:

En Windows:

bash

Copiar

Editar

.\env\Scripts\activate

En macOS/Linux:

bash

Copiar

Editar

source env/bin/activate

Instala las dependencias: Asegúrate de tener el archivo requirements.txt en el proyecto (si no lo tienes, puedes crear uno con las dependencias necesarias). Ejecuta el siguiente comando para instalar las bibliotecas necesarias:

bash

Copiar

Editar

pip install -r requirements.txt

Si no tienes el archivo requirements.txt, instala mysql-connector directamente con:

bash

Copiar

Editar

pip install mysql-connector

Configuración de la Base de Datos

Conexión a la Base de Datos: Asegúrate de modificar las credenciales de conexión a la base de datos en el archivo conexion.py (o donde se gestionen las configuraciones de conexión). Aquí es donde se especifican los datos como el host, usuario, contraseña y base de datos de MySQL.

python

Copiar

Editar

DB\_CONFIG = {

"host": "tu\_host",

"port": 3306, # Cambia si usas otro puerto

"user": "tu\_usuario",

"password": "tu\_contraseña",

"database": "tu\_base\_de\_datos"

}

Ejecución del Proyecto

solo ejecuta el archivo Main.py , dicho archivo hará todo el procedimiento de :

\*creación de tablas

\*Extracción tranformación y carga de datos

Estructura de la Base de Datos

El proyecto creará las siguientes tablas en la base de datos:

AUTOR: Almacena información sobre los autores.

id\_autor: Identificador único del autor (clave primaria).

autor: Nombre del autor.

FUENTE: Almacena información sobre las fuentes.

id\_fuente: Identificador único de la fuente (clave primaria).

fuente: Nombre de la fuente.

EXTRACTOR\_NOTICIAS: Almacena las noticias extraídas.

id: Identificador único de la noticia (clave primaria).

fuente: Fuente de la noticia.

autor: Autor de la noticia.

titulo: Título de la noticia.

descripcion: Descripción de la noticia.

url: URL de la noticia.

fecha\_publicacion: Fecha de publicación de la noticia.

TIEMPO: Almacena información sobre fechas y tiempos.

id\_fecha: Identificador único de la fecha (clave primaria).

año: Año de la fecha.

mes: Mes de la fecha.

fecha\_calendario: Fecha en formato calendario (YYYY-MM-DD).

trimestre: Trimestre de la fecha.

semana: Semana del año.

semestre: Semestre del año.

NOTICIAS: Relaciona noticias con fuentes, autores y fechas.

id\_noticias: Identificador único de la noticia (clave primaria).

id\_fuente: Clave foránea que referencia a FUENTE.

id\_autor: Clave foránea que referencia a AUTOR.

id\_fecha: Clave foránea que referencia a TIEMPO.

fecha\_publicacion: Fecha y hora de publicación de la noticia.

titulo: Título de la noticia.

descripcion: Descripción de la noticia.

url: URL de la noticia.

Errores Comunes

Error de conexión: Si hay un error en la conexión a la base de datos, asegúrate de que las credenciales de la base de datos son correctas y que el servidor de MySQL está en ejecución.

Tablas ya existentes: Si las tablas ya existen en la base de datos, el código las verificará y no las recreará.