
# Sistema ETL de Web Scraping con arquitectura de lago de datos

## Resumen del proyecto

Este proyecto requiere que los estudiantes construyan un sistema ETL (Extraer, Transformar, Cargar) completo que combine los requisitos de Tarea 1 y Tarea 2. El sistema extraerá datos utilizando Scrapy, los transformará a través de canales de validación y los almacenará en una arquitectura de lago de datos de tres niveles. Además, los estudiantes crearán un panel Streamlit para visualizar los datos procesados.

## System Requirements

### Integration of Tarea 1 and Tarea 2

Debe combinar los requisitos de web scraping de la Tarea 1 con los requisitos de limpieza y transformación de datos de la Tarea 2. Esto incluye:

- Implementar un raspador web basado en Scrapy para un dominio elegido (artículos de noticias, listas de empleo o reseñas de productos).
- Extracción de datos de al menos dos fuentes diferentes.
- Configurar el rastreador para que se ejecute automáticamente cada dos días.
- Validación y limpieza adecuadas de los datos
- Almacenamiento de datos en formato JSON y en base de datos PostgreSQL

### Data Lake Architecture

Los estudiantes deben implementar un lago de datos de tres niveles con las siguientes zonas:

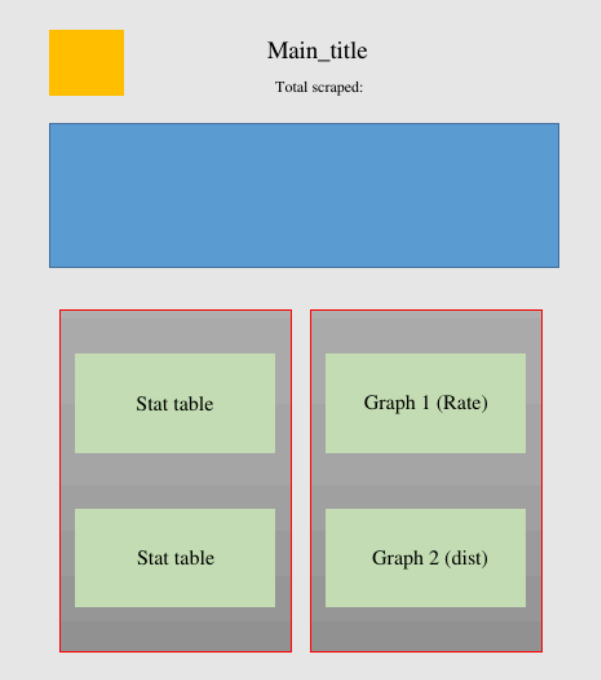
1. **Landing Zone**:
    - Contiene datos en bruto con transformaciones mínimas
    - Datos almacenados en formato JSON o CSV
    - Organizados en una estructura de carpetas dentro de «datalake/LANDING_ZONE»
2. **Refined Zone**:
    - Contiene datos depurados y transformados según los requisitos de Tarea 2
    - Datos almacenados en una base de datos relacional (PostgreSQL)
    - Implementa una lógica de validación para distintos tipos de datos

3. **Consumption Zone**:
    - Contiene datos listos para tareas de análisis o aprendizaje automático
    - Datos almacenados en una base de datos relacional (PostgreSQL)
    - Estructurados para facilitar su consulta y visualización

## Streamlit Dashboard

Los estudiantes deben desarrollar un cuadro de mando Streamlit que:

- Conecta con los datos de la Zona de Consumo y los visualiza
- Se integra con al menos una API externa elegida por el alumno.
- Proporciona visualizaciones significativas de los datos raspados



# Criterios de evaluación (50 puntos en total)

| Criterios | Descripción | Puntos máximos |
|---|---|---|
| **Scrapy Implementation** | • Aplicación correcta de la araña para el ámbito elegido (3 puntos)<br>• Extracción de datos de al menos tres fuentes (2 puntos)<br>• Ejecución automatizada cada 2 días (2 puntos)<br>• Gestión de User-Agent y cumplimiento de la política del sitio web (1 punto)<br>• Comprobación de fechas para evitar contenido duplicado (2 puntos) | **10 puntos** |
| **Data Cleaning & Transformation** | • Validación de elementos mediante elementos predefinidos de Scrapy (2 puntos)<br>• Scripts de limpieza de datos dentro de archivos pipeline (3 puntos)<br>• Validaciones específicas para diferentes tipos de datos (3 puntos)<br>• Tratamiento de errores durante la transformación (2 puntos) | **10 puntos** |
| **Data Lake Architecture** | • Aplicación correcta de la zona de aterrizaje (4 puntos)<br>  ○ Estructura de carpetas adecuada<br>  ○ Almacenamiento de datos en bruto con una transformación mínima<br>  ○ Formatos de archivo adecuados<br>• Implementación correcta de la Zona Refinada (5 puntos)<br>  ○ Conexión a base de datos PostgreSQL<br>  ○ Diseño adecuado del esquema<br>  ○ Implementación de limpieza de datos | **10 puntos** |

| | | |
|---|---|---|
| | • Implementación correcta de la Zona de Consumo (6 puntos)<br>  ○ Optimización de la base de datos PostgreSQL<br>  ○ Modelo de datos adecuado para el análisis<br>  ○ Almacenamiento y recuperación de datos eficiente | |
| **Streamlit Dashboard** | • Conexión satisfactoria a la base de datos de la Zona de Consumo (2 puntos)<br>• Integración con API externa (2 puntos)<br>• Calidad y pertinencia de la visualización de datos (3 puntos)<br>• Usabilidad y diseño del panel de control (3 puntos) | **15 puntos** |
| **Documentation and Code Quality** | • Documentación clara del proyecto (1 punto)<br>• Organización y estructura del código (1 punto)<br>• Comentarios adecuados y legibilidad del código (1 punto)<br>• Instrucciones de instalación y ejecución (1 punto)<br>• Explicación de las decisiones de diseño (1 punto) | **5 puntos** |

Fecha límite: 08/04/2025