Segundo entregable – ETL

Introducción

Este entregable tiene como objetivo explicar la construcción de un proceso ETL (Extracción, Transformación y Carga) completo sobre productos de la web de Sephora. Se parte desde la obtención de los datos directamente desde la página web, pasando por una mínima transformación necesaria para mantener la calidad del dato, hasta su almacenamiento en una base de datos relacional que permita su posterior análisis, incluyendo el seguimiento temporal de ciertos atributos como el precio y las valoraciones.

Descripción del Proceso ETL

- **Extracción**

La fase de extracción se ha llevado a cabo mediante técnicas de web scraping utilizando Python. El proceso principal consiste en: Obtener las URLs de todos los productos de maquillaje disponibles en la web de Sephora España e ir iterando por cada URL individual para extraer información como: nombre del producto, marca, precio, número de valoraciones, valoración media, descripción, número de variaciones del producto. Además, se ha implementado un sistema para extraer la información de los filtros (como cobertura, texturas, acabados, formulaciones, etc.) mediante una lógica alternativa: se accede a las URLs correspondientes a cada subcategoría dentro de cada filtro (por ejemplo, “cobertura ligera”, “cobertura media”, etc.), se obtienen los productos que aparecen en cada uno, y se comparan las URLs con las previamente extraídas. Esta lógica permite determinar de forma indirecta a qué categorías pertenece cada producto.

El proceso incluye también operaciones como scroll automático en la página, manejo de excepciones y verificación de estructuras HTML, lo cual ha requerido un trabajo técnico importante dada la variabilidad del contenido web.

- **Transformación**

En este caso, la transformación ha sido mínima y se ha incorporado directamente durante la fase de extracción. Las únicas transformaciones realizadas han sido: conversión y limpieza del tipo de dato de ciertas columnas (por ejemplo, convertir valores de precio a float, valoraciones a int, etc.), y normalización de valores textuales en lo posible.

No se han imputado nulos ni realizado transformaciones adicionales ya que, tras análisis, los valores ausentes eran representaciones correctas del dato (por ejemplo, ausencia de cierto filtro en un producto).

- **Carga**

La carga se ha diseñado cuidadosamente para permitir un proceso de actualización incremental. Se ha implementado una lógica que: en primer lugar, verifica si un producto ya existe en la base de datos (identificado de forma única mediante la URL del producto). Si el producto no existe, se inserta junto con sus atributos asociados, incluyendo sus relaciones con marcas, categorías, subcategorías y filtros (a través de tablas intermedias). Si el producto ya existe, se registra únicamente una nueva entrada en la tabla histórico con los valores que puedan variar en el tiempo (precio, número de valoraciones, etc.).

De esta forma, se puede realizar un análisis temporal de la evolución de cada producto cuando se disponga de un número considerable de scrapeos.

Esquema de la Base de Datos

La base de datos relacional diseñada sigue una estructura altamente normalizada. Se destacan la tabla de productos (tabla principal que contiene los productos). Las tablas auxiliares como marcas, categorías y subcategorías con claves foráneas en productos. La tabla de histórico que permite guardar distintas extracciones temporales de cada producto para seguimiento de cambios. Y, por último, las tablas de filtros, donde para cada filtro (acabado, textura, cobertura, formulación, etc.) se ha creado una tabla de categorías y otra tabla intermedia que vincula productos con sus correspondientes atributos múltiples. Esto permite manejar relaciones muchos a muchos.

El uso de claves foráneas y la política de ON DELETE CASCADE permite mantener la integridad referencial.

Problemas Encontrados y Soluciones

Durante el desarrollo del proyecto, se han enfrentado diversos desafíos técnicos. A continuación, se detallan algunos de ellos:

- Desafíos propios del scraping: Implementación del scroll automático y manejo dinámico del contenido web, dificultades para localizar ciertos datos dentro del HTML debido a clases variables o estructura inconsistente, ajustes necesarios para detectar pequeñas excepciones en la estructura HTML (por ejemplo, precios en clases distintas para algunos productos).
- Problemas descartados: Se intentó extraer la información de los ingredientes de cada producto, pero resultó inviable debido a la enorme variabilidad en el formato de presentación (comas, bullet points, texto entre ingredientes, etc.). Se concluyó que requeriría un tratamiento de NLP o reglas muy complejas para ser abordado adecuadamente.

- Problemas con identificadores: Inicialmente se usó el nombre del producto como identificador, pero se detectaron productos con nombres idénticos. Se corrigió utilizando la URL del producto como identificador único, lo que resolvió los problemas de duplicidad.
- Errores en la carga del histórico: Se detectó que, tras el primer scraping completo, aparecían 7 duplicados en la tabla histórico, lo cual no debería ocurrir en la primera ejecución. El error no ha podido ser identificado aún, por lo que se deja como pendiente para una próxima revisión.

Conclusiones

El proyecto ha permitido construir un proceso ETL completo, funcional y adaptable a futuros scrapeos. Se ha diseñado una base de datos relacional sólida y extensible, se han enfrentado desafíos reales propios del scraping y de la heterogeneidad del contenido web, y se ha priorizado la calidad y consistencia de los datos cargados.

Aunque quedan detalles por pulir (como la causa de ciertos duplicados en el histórico), el sistema está operativo y preparado para análisis exploratorios y temporales.