Proceso de ETL

Gabriela Jiménez Conde

19/05/2025

Proceso de ETL

La recopilación y preparación de datos sobre la competición hípica ha seguido las etapas de un proceso de ETL (Extract, Transform, Load), con el objetivo de obtener un conjunto de datos disponible y estructurado para ser analizado detalladamente. El resultado final del proceso es un conjunto de datos limpio y estructurado en una base de datos relacional.

Proceso de extracción

La extracción de la información se ha realizado mediante un procedimiento de web scraping de la web de la Real Federación Hípica Española (RFHE, `https://rfhe.com/competiciones/`), haciendo uso de selenium, herramienta especializada en el testeo de páginas web y web scraping. A partir de los diferentes enlaces de las diferentes disciplinas: salto de obstáculos, concurso completo, y doma clásica. En las mismas se puede encontrar un histórico de los concursos desde 2010 a nivel nacional e internacional en cada disciplina.

El pipeline parte de los concursos anuales de la disciplina seleccionada, y va accediendo a cada uno, para obtener la información general de los mismos, así como las pruebas llevadas a cabo en los mismos. Se han tenido solo en cuenta aquellos concursos que tienen resultados, sin obtener aquellos que han sido suspendidos o aplazados. Una vez se obtiene la información general así como las pruebas llevadas a cabo, a partir de la url de la primera prueba del concurso, se lanza un proceso que va obteniendo los archivos de tipo excel (xls) de cada prueba, los cuales contienen la información sobre los resultados de la prueba.

Proceso de transformación

La transformación de los datos se inicia en el proceso de extracción. La información sobre los concursos es almacenada según se ejecuta el proceso de extracción en un archivo de tipo json. La información sobre las pruebas y los resultados se obtiene mediante los excels descargados. Para obtener los resultados de las pruebas, se ha procedido a limpiar los archivos excel descargados, mediante el uso de pandas, herramienta de python especializada para el tratamiento de datos.

Dentro de cada excel se puede encontrar el nombre del concurso, el tipo de prueba, la fecha de la prueba, los resultados de los binomios, si han obtenido premio, y la puntuación realizada en la prueba realizada. Dependiendo del tipo de prueba, los resultados obtenidos tienen una estructura diferente, por lo que se han seccionado las puntuaciones y tiempos de los recorridos en diferentes columnas con el fin de utilizar esos resultados en el análisis. Una vez, el dataset de cada concurso y prueba está limpio, se procede a unir todos los diferentes datasets en uno mismo, obteniendo así todo el conjunto de datos en un mismo dataset, preparado para ser cargado.

Proceso de carga

El proceso de carga se realiza a partir del dataset completo de los datos. El almacenamiento del dataset se hace a través de una base de datos relacional de PostgreSQL (`https://github.com/Gabijc/Proyecto_final_hipica/blob/main/Documentation/ERD_BBDD_Hipica.png`). La información ha sido cargada a en el mismo orden de creación de las tablas de la base de datos. Para la carga se ha utilizado psycopg2, herramienta que permite conectar python con bases de datos de PostgreSQL.

Dificultades encontradas

La dificultad principal se encuentra en el proceso de web scraping, el cual es un proceso lento, ya que requiere de numerosas horas de extracción de información de la web de la RFHE. La extracción es de aproximadamente 300 concursos en cada año, lo cual ralentiza todo el proceso. Asimismo, la web no está completamente actualizada, ya que en cada click que se realiza, se abren nuevas pestañas.

Por otro lado, nos encontramos con la diferencia en los resultados según el tipo de prueba que se lleve a acabo, generando una serie de combinaciones a tener en cuenta a la hora de hacer la limpieza. A la hora de la carga, nos encontramos que hay concursos con el mismo nombre, generando dificultades a la hora de asignar ids adecuados a cada concurso, así como su unión con las pruebas y resultados adecuados.