**TRABAJO PRÁCTICO 02: Procesos ETL**

# Introducción

En este trabajo práctico se abordan las cuestiones relacionadas con los procesos de ETL Se plantean ejercicios y datasets cuyas resoluciones serán realizadas utilizando Pandas en Python y Apache Hop[1], en el repositorio de la materia se encuentra esta herramienta dockerizada[2].

# Consignas

Utilizando el dataset de Education Data[3] del World Bank Group se requiere lo siguiente.

1. **CSV de Países con la siguiente estructura:**
    a. id auto incremental
    b. Nombre del País
    c. Codigo del pais
2. **CSV de Preguntas con la siguiente estructura:**
    a. id auto incremental
    b. Pregunta (se requiere remover el contenido entre paréntesis)
3. **CSV de Educación primaria en años ordenado por duración de mayor a menor según el año 2023**

Resolver el ejercicio utilizando Apache Hop y realizar el mismo proceso a través de en una Notebook de Colab o Jupyter. Comentar los resultados obtenidos en cada una de las aproximaciones y comparar el proceso. Para esto genere un breve informe y envíelo en formato PDF.

1 https://github.com/bdm-unlu/2024/tree/main/dockers/hop
2 https://hop.apache.org/
3 https://data.worldbank.org/topic/education?view=chart

Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

2° Cuatrimestre 2024
Universidad Nacional de Luján