upf. BARCELONA SCHOOL OF MANAGEMENT

Caso Práctico 3

Objetivos del caso práctico:

- Familiarizarse con las etapas de una ETL: Extracción, Transformación y Carga (Load)
- Implementar una etapa de extracción usando 4 fuentes de datos distintas: 3 bases de datos y 1 API
- Implementar una etapa de transformación y carga
- Responder a las preguntas planteadas usando las tablas resultantes de la etapa de transformación

Herramientas a usar:

1. **Google Colab**: Se utilizará como entorno para levantar un servidor PostgreSQL y ejecutar consultas SQL para crear la base de datos y las tablas
2. **PostgreSQL**: Sistema de gestión de bases de datos relacional que será usado en el entorno de Google Colab.
3. **Pandas**: Librería de Python usada para manipulación de datos, que se utilizará principalmente para la función `read_sql`, la cual permite lanzar consultas SQL y visualizar los resultados en formato DataFrame. También usaremos la función `to_sql` para la carga de datos y opcionalmente el alumno podrá utilizarla durante la etapa de transformación

upf. BARCELONA SCHOOL OF MANAGEMENT

Explicación del entorno:

En este caso práctico se usarán dos herramientas distintas:

1. Se hará uso de un **Google Colab** para la creación de cuatro bases de datos PostgreSQL, 3 de ellas contendrán los datos que se necesitarán extraer, y una de ellas será la base de datos de analítica donde se cargarán las tablas resultantes de la etapa de transformación
2. Se hará uso de una **API Pública** cuya documentación se facilitará en el Colab, esta será la cuarta fuente de datos desde la que tendremos que extraer datos y juntar con las demás fuentes de datos

Pasos para trabajar con el Google Colab:

1. **Acceso a Google Colab**: A través del link proporcionado (disponible en la plataforma del curso).
2. **Copia del Google Colab**: El alumno tendrá que realizar una copia del Colab en su carpeta personal para poder trabajar en él. El nombre deberá de seguir el siguiente formato: Caso Práctico 3 - ETL - [Nombre y Apellidos del alumno]
3. **Ejecución de código predefinido**: Ejecuta las celdas del código ya proporcionado para configurar el servidor PostgreSQL, crear las funciones auxiliares y cargar los datos necesarios. Los tres archivos necesarios para la carga de datos se encuentran en la plataforma de la asignatura
4. **Extracción de datos**. Se podrá utilizar `query` para la extracción de datos de las distintas bases de datos, y `api_call` para la extracción de datos de la API (esto devolverá los datos en formato JSON, el alumno deberá convertirlos en un dataframe para trabajar con ellos)
5. **Transformaciones**. El alumno podrá realizar las transformaciones necesarias usando la librería Pandas, o usando SQL con la función que se facilita llamada `sql_trans`
6. **Creación de tablas y carga de datos**: El alumno podrá utilizar la función llamada `load_data` para la carga de datos y la función `execute_ddl` si fuese necesario eliminar alguna tabla.
7. **Consultas SQL**: El alumno podrá utilizar la función `query` para la consulta de las tablas y datos cargados para responder a las preguntas planteadas

upf. BARCELONA SCHOOL OF MANAGEMENT

Enunciado del caso práctico:

En este caso práctico trabajaremos con datos de la franquicia  . El objetivo será ejecutar las tres fase de un proceso ETL:

1. **Extracción**. El alumno deberá explorar cuatro fuentes de datos distintas, tres de ellas serán bases de datos y la cuarta será una API. En este caso, de la única información que dispondremos será del nombre de las tablas en cada base de datos, será tarea del alumno descubrir lo que contiene cada una y la relación entre todas las tablas. Una vez decidido qué tablas necesitaremos, se tendrá que hacer la extracción **sólo de las tablas que se necesiten sin que sufran ningún tipo de modificación.** A continuación se listan las fuentes de datos disponibles:
    a. **pokedexdb**.
        i. *pokemon*
        ii. *estadisticas_base*
        iii. *pokemon_tipo*
        iv. *tipo*
        v. *tipo_ataque*
    b. **movimientosdb**
        i. *pokemon_movimiento_forma*
        ii. *movimiento*
        iii. *forma_aprendizaje*
        iv. *nivel_aprendizaje*
        v. *tipo_forma_aprendizaje*
        vi. *mo*
        vii. *mt*
    c. **evolucionesdb**
        i. *evoluciona_de*
        ii. *pokemon_forma_evolucion*
        iii. *forma_evolucion*
        iv. *tipo_evolucion*
        v. *nivel_evolucion*
        vi. *piedra*
        vii. *tipo_piedra*
    d. **Pokeapi API**.
        i. *Egg Groups*
2. **Transformación**. Durante esta etapa el alumno si podrá transformar las tablas extraídas durante la etapa anterior (en esta etapa no está permitido extraer ninguna tabla, si se necesita alguna adicional se tendrá que incluir en la etapa de extracción). El objetivo es crear las tablas enriquecidas (desnormalizadas) necesarias para responder a las preguntas planteadas en el caso práctico
3. **Carga (Load)**. Durante esta fase se podrán cargar en la base de datos de analítica las tablas resultantes de la transformación, y serán sobre las que lancemos las consultas

Una vez realizada la carga de los datos en la base de datos de analítica, se deberán de **contestar a una serie de preguntas** que se encuentran en el Google Colab.

upf. BARCELONA SCHOOL OF MANAGEMENT

Entrega y evaluación:

1. **Entrega del ejercicio**: La entrega consistirá en:
    a. Se tendrá que compartir el Google Colab con el profesor del grupo, incluir el enlace del Colab en el entregable del caso práctico
    b. Se tendrá que subir a la plataforma el fichero .ipnyb descargado del Colab (*Archivo* **>** *Descargar* **>** *Descargar .ipynb*)

2. **Tiempo máximo de entrega**: Antes del 19/20 de Noviembre, la fecha se podrá encontrar en la plataforma de la asignatura
3. **Criterios de evaluación (sobre 10 puntos)**:
    a. **Extracción**: Se evaluará que se extraigan únicamente las tablas necesarias para la transformación sin que hayan sufrido ninguna transformación, así como la extracción correcta de todos los datos de la API
    b. **Transformación**: Se evaluará positivamente si se hacen agregados de datos en el menor número de tablas posibles para responder las preguntas. También se valorará positivamente la justificación de las tablas de transformación resultantes. La creación del diagrama ER de estas tablas es opcional.
    c. **Carga**: Se evaluará positivamente la correcta carga de las tablas en base de datos.
    d. **Preguntas**: Se valorará positivamente el número de preguntas que se consiguen responder usando las tablas resultantes de la transformación