Trabajo Práctico 1

Laboratorio de Datos
Facultad de Ciencias Exactas y Naturales - Universidad de Buenos Aires

Segundo Cuatrimestre 2024

El objetivo del presente trabajo práctico es realizar un modelo de datos (DER), luego construir un modelo relacional y finalmente realizar un análisis exploratorio respondiendo algunas respuestas y generando visualizaciones.

El trabajo práctico se realiza de a 3 integrantes, sin excepción.

1. Problema a modelar

Se tiene una base de datos de fútbol que contiene información de los jugadores, los equipos, las ligas y los países, además de los partidos. Los datos que modelaremos se encuentran dentro del período de tiempo 2007-2016.

Cada cierta cantidad de tiempo, a cada jugador se le asigna una serie de atributos que lo describen, como ser altura, peso, velocidad, resistencia, etc.

A su vez, cada jugador pertenece a un equipo y juega un conjunto de partidos. Cada partido se juega en una liga y en un país. Tiene una fecha, un resultado (ganado, perdido o empatado) y los goles a favor de cada equipo (local y visitante).

En cada temporada, un equipo tiene una serie de jugadores asociados que son quienes juegan los partidos, llamaremos a esta agrupación *plantel*. Importante: no importa qué jugador jugó cada partido, sino a qué equipo pertenecía.

Cada equipo solamente juega en su liga, y su liga pertence a un país. Quedan fuera del alcance de este trabajo práctico los torneos internacionales.

Realizar un DER que contemple la información descripta. Seleccionar las entidades principales, las entidades débiles y las relaciones que consideren necesarias. Establecer la cardinalidad y participación de las relaciones y justificar las decisiones tomadas.

Entregable: Diagrama Entidad-Relación, con sus justificaciones y decisiones explicadas. Máximo 2 carillas.

2. Modelo Relacional

Una vez realizado el DER, se debe construir un modelo relacional que contemple las tablas necesarias para almacenar la información. Justificar los criterios para la elección de las claves primarias y foráneas.

Con un primer diseño inicial, el que surje de la traducción del DER al modelo relacional, realizar un análisis de las tablas y sus atributos. En caso de encontrar redundancia o falta de normalización, realizar las modificaciones necesarias para poder llevarlo a la 3FN.

Entregable: Modelo Relacional. Análisis de las tablas y sus atributos. Justificación de las decisiones tomadas. Máximo 2 carillas.

3. Los datos

Los datos que se utilizan en este trabajo práctico se encuentran en el campus de la materia. Se cuenta con los siguientes archivos csv:

- `enunciado_equipos.csv`: contiene datos de los equipos.

- `enunciado_jugadores.csv`: contiene datos de los jugadores.
- `enunciado_jugadores_atributos.csv`: contiene métricas de los jugadores.
- `enunciado_liga.csv`: contiene los nombres de las ligas de fútbol.
- `enunciado_paises.csv`: contiene los nombres de los países.
- `enunciado_partidos.csv`: contiene los datos de los partidos.

Estos archivos son una adaptación de un dataset público[1] obtenido mediante la técnica de Scrapping y se encuentran en formato csv.

Algunas aclaraciones, es posible consultar los atributos de los jugadores en una página web, los mismos fueron recolectados de la página `https://sofifa.com/`, y es la recopilación de los datos de los juegos de la saga FIFA. El atributo `player_fifa_api_id` es el que puede ser remplazado en la url `https://sofifa.com/player/player_fifa_api_id` para obtener la información de un jugador en particular. Por ejemplo: `https://sofifa.com/player/158023` corresponde a Lionel Messi.

4. Creación de tablas

En el campus de la materia cuentan con seis archivos `.csv` que representan fuentes de información de la cual deberán extraer los datos para cargar las tablas del modelo relacional de acuerdo al diseño que hayan decidido.

Para cargar un archivo csv en una tabla de una base de datos, se puede utilizar el siguiente código en Python:

```
import pandas as pd

jugadores_crudo = pd.read_csv('enunciado_jugadores.csv')
```

Luego deberán poder, mediante consultas SQL, generar las tablas que se correspondan con su modelo relacional. Esto pueden realizarlo con un código con esta estructura:

```
import duckdb

# Generar la tabla que queremos
jugadores = duckdb.sql('''SELECT Nombre, Apellido, Potencia FROM jugadores_crudo''')

# Guardar la tabla en un archivo csv
jugadores.to_csv('jugadores.csv', index=False)
```

Para el análisis posterior y las consultas a la base podrán usar directamente los archivos generados en este punto, sin necesidad de volver a cargar y procesar los datos iniciales.

Entregable: Los archivos `.csv` con los datos de cada tabla que generen. Archivo `generar_tablas.py` que contenga el código para generar las tablas a partir de los archivos csv originales.

5. Consultas y visualizaciones

Para este ítem se les asignará un país y una liga en particular por grupo. Deberán realizar un análisis exploratorio de los datos, respondiendo las siguientes preguntas y generando las visualizaciones que consideren pertinentes.

El intervalo de tiempo a considerar deben definirlo entre ustedes en base a los datos que tengan disponibles. Se espera que elijan al menos 4 años consecutivos y que definan el criterio para hacerlo (ej. comparten el 80 % de los equipos en dichos años; tienen cantidad similar de partidos; etc.)

[1] `https://github.com/hugomathien/football-data-collection/tree/master`

Consultas SQL

Realizar al menos las siguientes consultas SQL y dar su resultado.

- Sobre equipos del país y la liga asignada:
  - ¿Cuál es el equipo con mayor cantidad de partidos ganados?
  - ¿Cuál es el equipo con mayor cantidad de partidos perdidos de cada año?
  - ¿Cuál es el equipo con mayor cantidad de partidos empatados en el último año?
  - ¿Cuál es el equipo con mayor cantidad de goles a favor?
  - ¿Cuál es el equipo con mayor diferencia de goles?
  - ¿Cuántos jugadores tuvo durante el período de tiempo seleccionado cada equipo en su plantel?
- Sobre jugadores del país y la liga asignada:
  - ¿Cuál es el jugador con mayor cantidad de goles?
  - ¿Cuáles son los jugadores que más partidos ganó su equipo?
  - ¿Cuál es el jugador que estuvo en más equipos?
  - ¿Cuál es el jugador que menor variación de potencia ha tenido a lo largo de los años? (medida en valor absoluto)
- Joineate algo (optativa, ver preguntas propias a continuación):
  - ¿Hay algún equipo que haya sido a la vez el más goleador y el que tenga mayor valor de alguno de los atributos (considerando la suma de todos los jugadores)?

En esta etapa exploratoria será tomado como positivo la formulación de preguntas propias que les generen interés. Las preguntas de más valor, por supuesto, son las que necesitan de consultas a más de una tabla.

Visualizaciones

Realizar las siguientes visualizaciones, justificando en cada una de ellas qué visualización eligieron y dando una breve explicación de lo que observan en la misma:

- Graficar la cantidad de goles a favor y en contra de cada equipo a lo largo de los años que elijan.
- Graficar el promedio de gol de los equipos a lo largo de los años que elijan.
- Graficar la diferencia de goles convertidos jugando de local vs visitante a lo largo del tiempo.
- Graficar el número de goles convertidos por cada equipo en función de la suma de todos sus atributos.

Entregable: Archivo `analisis.py` que contenga el código para realizar el análisis exploratorio y las visualizaciones. Informe con las respuestas a las preguntas y las visualizaciones generadas.

Entrega

Se debe entregar un informe en formato PDF en el que se incluya:

- Carátula con título y nombres de los integrantes.
- Diagrama Entidad-Relación, con sus justificaciones y decisiones explicadas
- Modelo Relacional.
- Análisis exploratorio.

Además se deberán incluir los siguientes archivos `.py`:

- `generar_tablas.py`: script que dada la fuente de datos genera los archivos `.csv` que se corresponden con las tablas planificadas en el modelo relacional.
- `analisis.py`: script que del análisis exploratorio y de las visualizaciones.

Además deberán entregar un archivo `.csv` por cada tabla que utilicen.