UC | Chile

Ingenieria de datos

Trabajo final: diseño y uso de una base de datos

¡Bienvenido(a)!

- **Objetivo**: Elegir las herramientas adecuadas para el proces de análisis de datos en un caso de uso concreto.
- **Tipo de actividad**: Individual/grupal
- **Tipo de evaluación**: Sumativa (con calificación)
- **Ponderación**: 15% (Equivale al 15% de la nota final del curso)
- **Puntaje**: 60 puntos
- **Calificación**: Escala de 1 a 7, con una exigencia de 50%. La nota mínima para aprobar es 4.0.

Evaluación

Introducción

Tu trabajo anterior ha sido tan exitoso que ahora tu aplicación fue comprada por EvilCorp Inc. para incorporar tu base de datos en su negocio. Cómo siempre al hacer merge de dos empresas, hay que unir sus bases de datos. En este trabajo tendrás que restructurar tu base de datos según la especificación de nuevo esquema propuesto por el CTO del EvilCorp Inc., y cargar los datos que ya tienes a este nuevo esquema. Adicionalmente, el equipo de data science de EvilCorp Inc. ha logrado coleccionar ciertos datos sobre ciudades y países que tendrás que cruzar con tus datos existentes para satisfacer los requisitos del esquema nuevo.

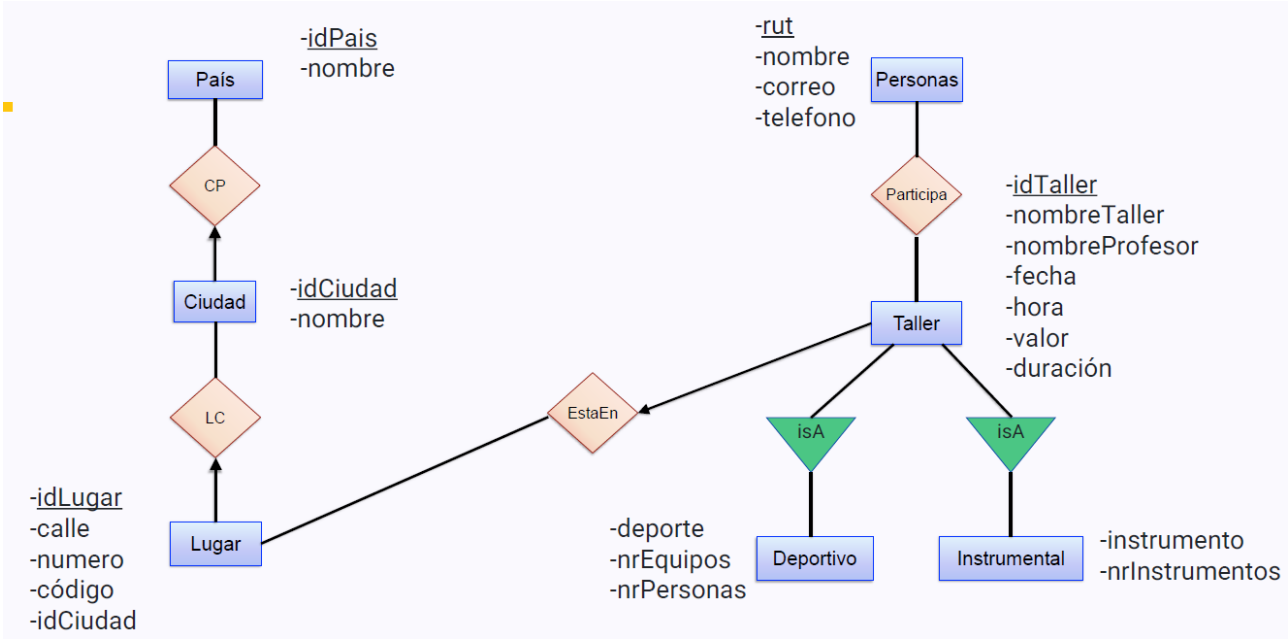
Descripción del problema

Nuestro punto de partida será el esquema y la base de datos que ya ocupamos en las tareas 2 y 3. En particular, contamos con el siguiente esquema para nuestros datos:

- Persona(rut,nombre,correo,teléfono)
- TallerDeportivo(id,nombreTaller, nombreProfesor, fecha,hora,valor,duración,deporte,nrEquipos, nrPersonas)
- TallerInstrumental(id,nombreTaller, nombreProfesor, fecha,hora,valor,duración,instrumento,nrInstrumentos)
- Lugar(id,calle,número,código,nombreCiudad,nombrePaís)
- Participa(rut,idTaller)
- EstáEn(idTaller,idLugar)

Si necesitas acordarte del significado de cada una de estas tablas, debes consultar la descripción en Tarea 2. Adicionalmente, tenemos una instancia de este esquema que te entregaremos junto con el notebook que viene con esta tarea.

El nuevo esquema que propone el CTO del EvilCorp Inc. y que se ocupará desde aquí en adelante se puede visualizar con el siguiente diagrama E/R abreviado (aquí los atributos se ponen al lado para no sobrecargar la imágen):



Alternativamente, puedes asumir que tu base de datos debe tener el siguiente esquema:

- Persona(rut,nombre,correo,teléfono)
- Taller(id,nombreTaller, nombreProfesor, fecha,hora,valor,duración) ***
- Deportivo(id,deporte,nrEquipos, nrPersonas) ***
- Instrumental(id,instrumento,nrInstrumentos) ***
- LugarNuevo(id,calle,número,código,IDCiudad) ***
- Participa(rut,idTaller)
- EstáEn(idTaller,idLugar)
- Pais(id,nombrePais) ***
- Ciudad(id,NombreCiudad,idPais) ***

Dónde las relaciones marcadas con *** son cambiadas en comparación con Tareas 2 y 3.

Adicionalmente, te entregaremos 3 archivos CSV que se ocuparán para llenar las tablas Ciudad y País. Estos tres archivos son:

- Pais.csv, que contiene pares (idPais,nombrePais)
- Ciudad.csv, que contiene pares (idCiudad,nombreCiudad)
- CiudadPais.csv, que contiene pares de forma (idCiudad,idPais)

Los tres archivos se entregan junto con la tarea, pero el jupyter notebook ya viene configurado para cargarlos directamente desde un Google drive. Cómo puedes ver, el esquema nuevo tiene un solape no trivial con el esquema viejo, y estas tablas se mantiene en la nueva base de datos (por ejemplo, Persona).

Es importante decir que algunas preguntas en esta tarea tienen carácter exploratorio, y tendrán que entender algunos conceptos por su propia cuenta (por ejemplo cómo cargar datos a SQL desde un dataframe, o desde un CVS).

Pregunta 1

En esta primera pregunta, te pedimos ejecutar la primera celda del notebook para configurar el entorno de SQL, y descargar la base de datos de tareas 2/3 además de los tres archivos CSV.

Luego, te pedimos crear las tablas nuevas con los atributos adecuados (los nombres deben coincidir con los nombres de la especificación de arriba, y los tipos de atributos los debes asignar tú). Para guiar un poco tu diseño de distintos tipos de atributos, te dejamos dos ejemplos con tipos de datos especificados:

- Persona(rut VARCHAR(100) PRIMARY KEY, nombre VARCHAR(100), correo VARCHAR(100), telefono VARCHAR(100));
- Paises(pid INT PRIMARY KEY, nombre VARCHAR(100));

Para esto, debes escribir un comando del tipo CREATE TABLE… en SQL para cada tabla marcada con *** en la especificación de arriba. El punto del ejercicio es modificar tu base de datos tal que al final de la tarea contenga los datos necesarios. Quiere decir que,

algunas tablas se mantienen como están (por ejemplo Persona) y algunas se pueden borrar al final del ejercicio.

Importante: No te olvides de especificar las llaves primarias y foráneas. Por ejemplo, tenemos que IDCiudad es una llave foránea en la tabla LugarNuevo (para Ciudades), o que rut es la llave foránea en Participa.

Pregunta 2

Inserta a las tablas Taller, Deportivo e Instumental los datos que se piden según su esquema, desde las tablas TallerDeportivo y TallerInstrumental que ya existen en tu base de datos.

Para esto puedes ocupar SQL, Pandas, Python, lo que sea. Una buena alternativa se propone en el siguiente link: https://www.w3schools.com/sql/sql_insert_into_select.asp. También pueden ver el siguiente link: https://stackoverflow.com/questions/71362727/python-sqlite3-insert-data-from-for-loop (si siguen esta forma, recuerden hacer commit a la connection luego de ejecutar el INSERT INTO)

Pregunta 3

Ocupando los 3 csv entregados, llena las tablas País y Ciudad del esquema nuevo. Para esto, hay que hacer un join entre los csv. De nuevo, esto lo puedes hacer con Pandas (cargando los tres CSV a SQL) o con Python.

Aquí puedes cargar los csvs en Pandas, y luego ocupar este tutorial para cargar datos desde un dataframe a una tabla SQL: https://datacarpentry.org/python-ecology-lesson/09-working-with-sql/index.html

Alternativamente, lo más fácil sería cargar datos desde un csv directamente a sqlite, como se explica aquí: https://www.sqlitetutorial.net/sqlite-import-csv/ (pero para esto, debes tener sqlite3 instalado en tu computador).

Otro link que puede ser de utilidad es el siguiente:
https://kontext.tech/article/633/pandas-save-dataframe-to-sqlite

Pregunta 4

Llena la tabla LugarNuevo de tu esquema con los datos que se piden. Para esto necesitas ocupar la tabla Lugar (que viene en la base de datos anterior) y las tablas Ciudad y País (o los CSVs si prefieres).

Debes tener cuidado al cruzar la información de las ciudades proveniente de la base de datos anterior con las ciudades que vienen de los archivos nuevos, ya que no todas las ciudades tendrán lugares asociados en la base de datos anterior. La idea es que no queden nulos en tu base de datos.

Quizás te puede servir este link https://datatofish.com/sql-to-pandas-dataframe/

Pregunta 5

Ahora que terminamos la carga de datos en las tablas nuevas, es tiempo de limpiar nuestra base de datos. Para esto te pedimos borrar las tablas: Lugar, TallerInstrumental y TallerDeportivo de la base de datos antigua.

Pregunta 6

Con nuestra base de datos lista, podemos hacer un poco de análisis de los datos. En esta pregunta, te pedimos escribir una consulta SQL que ordene los talleres deportivos según la participación. Quiere decir, debes devolver **id del taller, nombre del taller, nombre del deporte y la cantidad de gente *distinta* que participa en este taller**, *ordenado de forma descendente*.

Pregunta 7

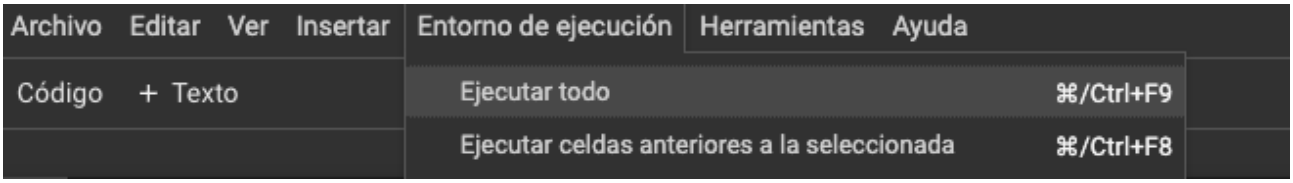
En esta pregunta debes escribir una consulta SQL que entregue el **nombre y país de la ciudad** donde se organiza la máxima cantidad de talleres instrumentales. Quiere decir, hay que calcular cuántos talleres instrumentales hay en cada ciudad, y devolver la(s) ciudad(es) dónde se alcanza el número máximo.

Pregunta 8

En esta pregunta queremos saber los nombres de los países con la cantidad máxima de lugares con al menos dos talleres deportivos (2 o más) organizados en estos lugares. Quiere decir, para cada lugar hay que ver si hay al menos 2 talleres deportivos en él, y se necesita devolver **el país con la cantidad máxima de lugares** con esta propiedad.

IMPORTANTE: Debes entregar tu notebook (archivo con extensión ipynb) ejecutado según tus últimos cambios ya que se revisarán los resultados en base a lo que salga ejecutado en el notebook**. De no cumplir con lo anterior, se realizará un descuento de puntaje.**

En el menú superior de google colab > Entorno de ejecución > Ejecutar todo, tienes la opción de ejecutar todas las celdas del notebook. Puedes realizar esta acción antes de entregar tu tarea para asegurarte de que tus últimos cambios en el código sean los ejecutados.



Referencias

- Ramakrishnan, R., Gehrke, J., Database Management Systems, 3rd edition, McGraw-Hill, 2002.
- Soto, A., Bases de Datos, https://github.com/alanezz/Syllabus-2019-1, 2019.
- Cuerpo docente UC, Materiales del curso Bases de Datos, https://github.com/IIC2413/Syllabus-2021-1, 2021.

Instrucciones

1. Realiza esta evaluacion cuando hayas terminado las clases de la semana 8
2. Debes entregar el formulario .ipynb
    - Haz clic en el botón para agregar entrega. Se abrirá una nueva ventana que permite arrastrar el archivo y subirlo.
    - Comprueba que el archivo arrastrado es el correcto y presiona el botón para guardar cambios. El documento quedará guardado en la plataforma.

Aspectos formales

Debes entregar un archivo .ipynb

¡Mucho éxito!

Importante: la fecha de entrega está indicada en el calendario del curso. Cuidar la redacción y la ortografía. Si tienes alguna duda sobre los contenidos o sobre cómo realizar esta actividad, puedes utilizar la herramienta "Mensajes" y enviar tu pregunta. Recibirás la respuesta de su tutor con las orientaciones correspondientes