Base de Datos I
Trabajo Práctico Especial
1er Cuatrimestre 2023

1. Objetivo

El objetivo de este Trabajo Práctico Especial es aplicar los conceptos de SQL Avanzado (PSM, Triggers) vistos a lo largo del curso, para implementar funcionalidades y restricciones no disponibles de forma estándar (que no pueden resolverse con Primary Keys, Foreign Keys, etc.).

2. Modalidad

El Trabajo Práctico estará disponible en el Campus a partir del 08/06/2023, indicándose allí mismo la fecha de entrega.
Se incluye junto con el enunciado el archivo: **us_births_2016_2021.csv**.
El TP deberá realizarse en grupos de 3 alumnos, a excepción de dos grupos de 4, y entregarse a través de la plataforma Campus ITBA hasta la fecha allí indicada.

3. Descripción del Trabajo

En el sitio de Kaggle se pueden encontrar distintos datasets de acceso público. En esta ocasión vamos a utilizar el dataset de *US Births by Year, State and Educational Level* que consiste en un resumen por estado, por año y por nivel educativo de la madre, de los nacimientos que sucedieron entre el 2016 y el 2021. Este dataset aporta estadísticas nacionales para analizar el comportamiento de los nacimientos a través de los años.

La información es provista en un archivo CSV (Comma Separated Values). El archivo **us_births_2016_2021.csv** contiene información de los nacimientos, que sucedieron en Estados Unidos, detallada por diferentes categorías, como ser estado, año, género del bebé y nivel educativo de la madre.
Las columnas del archivo son:

- ***State****: nombre del estado de Estados Unidos del cual provienen los nacimientos*
- ***State_Abbreviation:*** *código del nombre del estado de Estados Unidos*
- ***Year****: año calendario en el cual se sucedieron los nacimientos*
- ***Gender****: código del género de los bebés*
- ***Mother_Education_Level****: nivel de educación de las madres de los bebés*
- ***Education_Level_Code****: código del nivel de educación de las madres de los bebés*
- ***Births****: cantidad total de bebés nacidos*
- ***Mother_Average_Age****: promedio de edad de las madres de los bebés, expresado en años*
- ***Average_Birth_Weight****: promedio del peso de los bebés, expresado en gramos*

Antes de insertar el archivo en una tabla definitiva, se quiere interceptar la inserción del estado, del año y del nivel de educación de las madres, y cambiarlas por una FK a diferentes dimensiones representadas en las tablas ESTADO, ANIO y NIVEL_EDUCACION.

La finalidad de este Trabajo Práctico Especial consiste en migrar los datos del archivo CSV a una base de datos, producir un reporte y realizar algunas validaciones. Específicamente se debe hacer lo siguiente:

a) Crear las tablas de las distintas dimensiones
b) Crear la tabla definitiva donde solo se guarden los campos de género, cantidad de nacimientos y los promedios, junto a los campos identificatorios de tablas creadas en el punto a)
c) Importar los datos y cargar todas las tablas creadas en los puntos anteriores
d) Crear un reporte con información consolidada

a) Creación de las tablas de las dimensiones.

Deben crearse las tablas de las distintas dimensiones con las siguientes condiciones mínimas:

- ESTADO: debe tener un campo identificatorio y un campo con el nombre del estado
- ANIO: debe tener un campo identificatorio y un campo que indique si el año es bisiesto
- NIVEL_EDUCACION: debe tener un campo identificatorio y un campo con la descripción del nivel de educación

Se deben crear las claves y constraints apropiados.

b) Creación de la tabla definitiva.

Debe crearse una tabla definitiva que será la receptora de los datos provenientes del archivo **us_births_2016_2021.csv**. Los campos y restricciones de la tabla deben crearse en base al análisis de los datos.

Recordar que los archivos csv son archivos de texto que pueden abrirse fácilmente con cualquier editor.

Para el caso particular de los campos state, state_abbreviation, year, mother_education_level y education_level_code, se deberá cambiar su contenido para que el mismo haga referencia a la key de la tablas creadas en el punto a), antes de insertarlos en la tabla definitiva.

En base a los datos, se debe crear la clave y constraints apropiados.

c) Importación de los datos.

Utilizando el comando COPY de PostgreSQL, se deben importar TODOS los datos del archivo **csv** en la tabla creada en **b)**. El archivo **csv** provisto por la cátedra NO puede ser modificado.

Creación de un trigger para:

1) Determinar la FK de las distintas dimensiones

Para insertar los datos en la tabla definitiva es necesario interceptar la inserción del estado, año y nivel educativo de la madre, y luego identicar la FK a cada una de las dimensiones de las tablas creadas en el punto a).

2) Cargar los valores de las dimensiones

Además de insertar los datos del archivo, se deben poblar las distintas tablas que conforman las distintas dimensiones, siempre y cuando los valores correspondientes no existan en dichas tablas.

Por ejemplo, si partimos con la tabla ESTADO vacía y si al principio en el archivo CSV viene el estado ”Alabama” con la abreviación ”AL”, se debe insertar una tupla en la tabla ESTADO, quedando la tabla con la siguiente información:

• ESTADO: ”AL”, ”Alabama”

Sin embargo, si luego viene el mismo estado ”Alabama” con la abreviación ”AL”, se reaprovecha el registro ya cargado, quedando la tabla ESTADO igual que antes.

Si luego viene información de otro estado, ”Alaska” con la abreviación ”AK”, entonces se tiene que insertar una nueva tupla en la tabla ESTADO.

• ESTADO: ”AK”, ”Alaska”

Por ejemplo, en el caso de los años, si partimos con la tabla ANIO vacía y si al principio en el archivo CSV viene el año 2016, se debe insertar una tupla quedando la tabla con la siguiente información:

• ANIO: 2016, true

Si luego viene el año ”2018” se debe insertar una tupla en ANIO, agregando en la tabla la siguiente información:

• ANIO: 2018, false

Sin embargo, si luego viene nuevamente información de ”2018”, no se tiene que insertar ninguna tupla.

La misma lógica se aplica a la dimensión NIVEL DE EDUCACION.

Las tuplas en todas las tablas tiene que tener todos los datos bien completos.

d) Reporte de información consolidada.

El responsable del Centro Nacional de Estadísticas de Estados Unidos realiza un análisis de los nacimientos con información consolidada, agrupada por Año y por distintas categorías como ser Estado, Género del bebé y Nivel de Educación de la madre.

Se pide crear la función **ReporteConsolidado(n)** que recibe como parámetro la cantidad de años a mostrar tomando como base el primer año cargado en la tabla definitiva, la cual genere un reporte mostrando para cada año y categoría, la cantidad total de nacimientos, la edad promedio de las madres, la edad mínima, la edad máxima, el promedio de peso de los bebés, el peso mínimo y el peso máximo. Los 3 pesos expresados en kilogramos.

El reporte tendrá las siguientes características:

I. Título del reporte:

```
"CONSOLIDATED BIRTH REPORT”
```

II. Encabezado de columnas:

```
“Year Category Total AvgAge MinAge MaxAge AvgWeight MinWeight
MaxWeight”
```

III. Por cada año tiene que aparecer un renglón en el reporte, con los años ordenados de menor a mayor. La primer categoría de agrupación (State) con sus valores ordenados alfabéticamente en forma descendente y sus métricas (Total, AvgAge, MinAge, MaxAge, AvgWeight, MinWeight y MaxWeight), deben estar en el **mismo** renglón que el año. El resto de las categorías (Gender y Education Level), encolumnados a continuación en los renglones subsiguientes:

○ Para la categoría de Estados, solo interesa reportar aquellos donde haya habido más de 200.000 nacimientos

○ Para la categoría de Nivel de Educación de la madre, solo interesa reportar los niveles de educación categorizados con algún valor relevante. Es decir, no considerar cuando el nivel de educación es desconocido o no informado

IV. Al final de los renglones, tiene que aparecer el total de las métricas Total, AvgAge, MinAge, MaxAge, AvgWeight, MinWeight y MaxWeight correspondientes para ese año

En caso de que no existieran datos para los parámetros ingresados, no se debe mostrar nada (ni siquiera el encabezado del reporte).

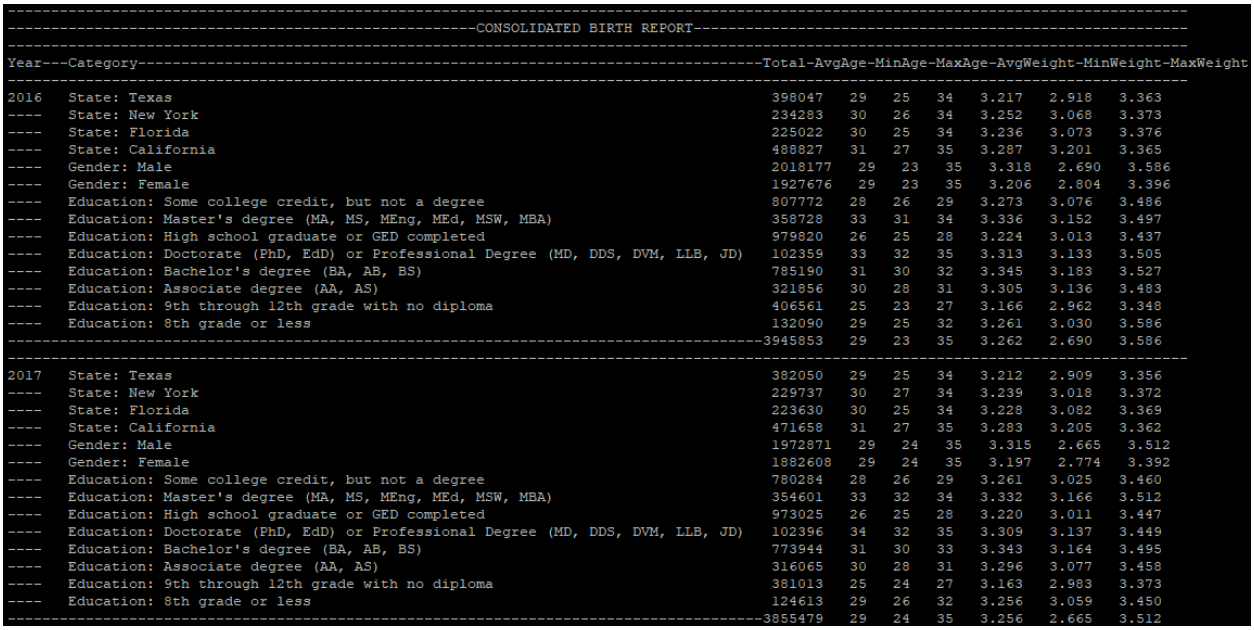
La función debe manejar los posibles errores.

Por ejemplo,

- si invocamos **ReporteConsolidado(1)** se debe obtener información del año 2016:

```
-------------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------CONSOLIDATED BIRTH REPORT-------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
Year---Category-------------------------------------------------------------------------Total-AvgAge-MinAge-MaxAge-AvgWeight-MinWeight-MaxWeight
-------------------------------------------------------------------------------------------------------------------------------------
2016   State: Texas                                                                     398047   29   25   34   3.217   2.918   3.363
----   State: New York                                                                  234283   30   26   34   3.252   3.068   3.373
----   State: Florida                                                                   225022   30   25   34   3.236   3.073   3.376
----   State: California                                                                488827   31   27   35   3.287   3.201   3.365
----   Gender: Male                                                                    2018177   29   23   35   3.318   2.690   3.586
----   Gender: Female                                                                  1927676   29   23   35   3.206   2.804   3.396
----   Education: Some college credit, but not a degree                                 807772   28   26   29   3.273   3.076   3.486
----   Education: Master's degree (MA, MS, MEng, MEd, MSW, MBA)                         358728   33   31   34   3.336   3.152   3.497
----   Education: High school graduate or GED completed                                 979820   26   25   28   3.224   3.013   3.437
----   Education: Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD)   102359   33   32   35   3.313   3.133   3.505
----   Education: Bachelor's degree (BA, AB, BS)                                        785190   31   30   32   3.345   3.183   3.527
----   Education: Associate degree (AA, AS)                                             321856   30   28   31   3.305   3.136   3.483
----   Education: 9th through 12th grade with no diploma                                406561   25   23   27   3.166   2.962   3.348
----   Education: 8th grade or less                                                     132090   29   25   32   3.261   3.030   3.586
-------------------------------------------------------------------------------------3945853   29   23   35   3.262   2.690   3.586
```

- Si invocamos **ReporteConsolidado(2)** se debe obtener información del año 2016 y 2017



----------CONSOLIDATED BIRTH REPORT----------

Year	Category	Total	AvgAge	MinAge	MaxAge	AvgWeight	MinWeight	MaxWeight
2016	State: Texas	398047	29	25	34	3.217	2.918	3.363
----	State: New York	234283	30	26	34	3.252	3.068	3.373
----	State: Florida	225022	30	25	34	3.236	3.073	3.376
----	State: California	488827	31	27	35	3.287	3.201	3.365
----	Gender: Male	2018177	29	23	35	3.318	2.690	3.586
----	Gender: Female	1927676	29	23	35	3.206	2.804	3.396
----	Education: Some college credit, but not a degree	807772	28	26	29	3.273	3.076	3.486
----	Education: Master's degree (MA, MS, MEng, MEd, MSW, MBA)	358728	33	31	34	3.336	3.152	3.497
----	Education: High school graduate or GED completed	979820	26	25	28	3.224	3.013	3.437
----	Education: Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD)	102359	33	32	35	3.313	3.133	3.505
----	Education: Bachelor's degree (BA, AB, BS)	785190	31	30	32	3.345	3.183	3.527
----	Education: Associate degree (AA, AS)	321856	30	28	31	3.305	3.136	3.483
----	Education: 9th through 12th grade with no diploma	406561	25	23	27	3.166	2.962	3.348
----	Education: 8th grade or less	132090	29	25	32	3.261	3.030	3.586
		3945853	29	23	35	3.262	2.690	3.586
2017	State: Texas	382050	29	25	34	3.212	2.909	3.356
----	State: New York	229737	30	27	34	3.239	3.018	3.372
----	State: Florida	223630	30	25	34	3.228	3.082	3.369
----	State: California	471658	31	27	35	3.283	3.205	3.362
----	Gender: Male	1972871	29	24	35	3.315	2.665	3.512
----	Gender: Female	1882608	29	24	35	3.197	2.774	3.392
----	Education: Some college credit, but not a degree	780284	28	26	29	3.261	3.025	3.460
----	Education: Master's degree (MA, MS, MEng, MEd, MSW, MBA)	354601	33	32	34	3.332	3.166	3.512
----	Education: High school graduate or GED completed	973025	26	25	28	3.220	3.011	3.447
----	Education: Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD)	102396	34	32	35	3.309	3.137	3.449
----	Education: Bachelor's degree (BA, AB, BS)	773944	31	30	33	3.343	3.164	3.495
----	Education: Associate degree (AA, AS)	316065	30	28	31	3.296	3.077	3.458
----	Education: 9th through 12th grade with no diploma	381013	25	24	27	3.163	2.983	3.373
----	Education: 8th grade or less	124613	29	26	32	3.256	3.059	3.450
		3855479	29	24	35	3.256	2.665	3.512

- Si invocamos **ReporteConsolidado(7)** se debe obtener lo mismo que invocando **ReporteConsolidado(6)**, es decir información del año 2016, 2017, 2018, 2019, 2020 y 2021

- Si invocamos **ReporteConsolidado(0)** no se obtiene nada

4. Entregables

Los alumnos deberán entregar los siguientes documentos:

- El script sql **funciones.sql** con el código necesario para crear las tablas, las funciones y los triggers
- Un informe que debe contener:
    - El rol de cada uno de los participantes del grupo. Si bien en el TP deben estar involucrados todos los integrantes, se debe asignar un rol de supervisión de cada una de las tareas. Mínimamente los roles son: encargado del informe, encargado de las funciones, encargado del trigger, encargado del funcionamiento global del proyecto y encargado de investigación. Pueden asignarse más roles en caso de requerirse
    - Todo lo investigado para realizar el TP
    - Las dificultades encontradas y cómo se resolvieron
    - También se debe detallar aquí el proceso de importación de los datos realizado
    - El informe debe tener como máximo 3 páginas

5. Evaluación

La evaluación del trabajo se llevará a cabo utilizando los parámetros establecidos en la rúbrica asociada a la actividad en el Campus.
Se tendrá en cuenta que las consultas, más allá del funcionamiento (lo cual es fundamental), sean genéricas.
Los docentes ejecutarán el proceso usando el conjunto de datos entregado pero podrán también hacer pruebas con otros conjuntos de datos de similares características para evaluar el funcionamiento en distintos escenarios.
El informe deberá estar completo y sin faltas de ortografía.
En caso de que el trabajo no cumpliera los requisitos básicos para ser aprobado, los alumnos serán citados en la fecha de recuperatorio para defenderlo y corregir los errores detectados.