Universidad del Rosario

MACC Matemáticas Aplicadas y Ciencias de la Computación

Segunda entrega proyecto Manejo de Bases de Datos 2021-I

Germán David Plazas Cayachoa, David Alfonso Oviedo Salamanca, Santiago Rodríguez Morales, Jose Gabriel Álvarez Medina.

Escuela de Ingeniería, Ciencia y Tecnología, Universidad del Rosario

1. Modificaciones con respecto a la primera entrega.

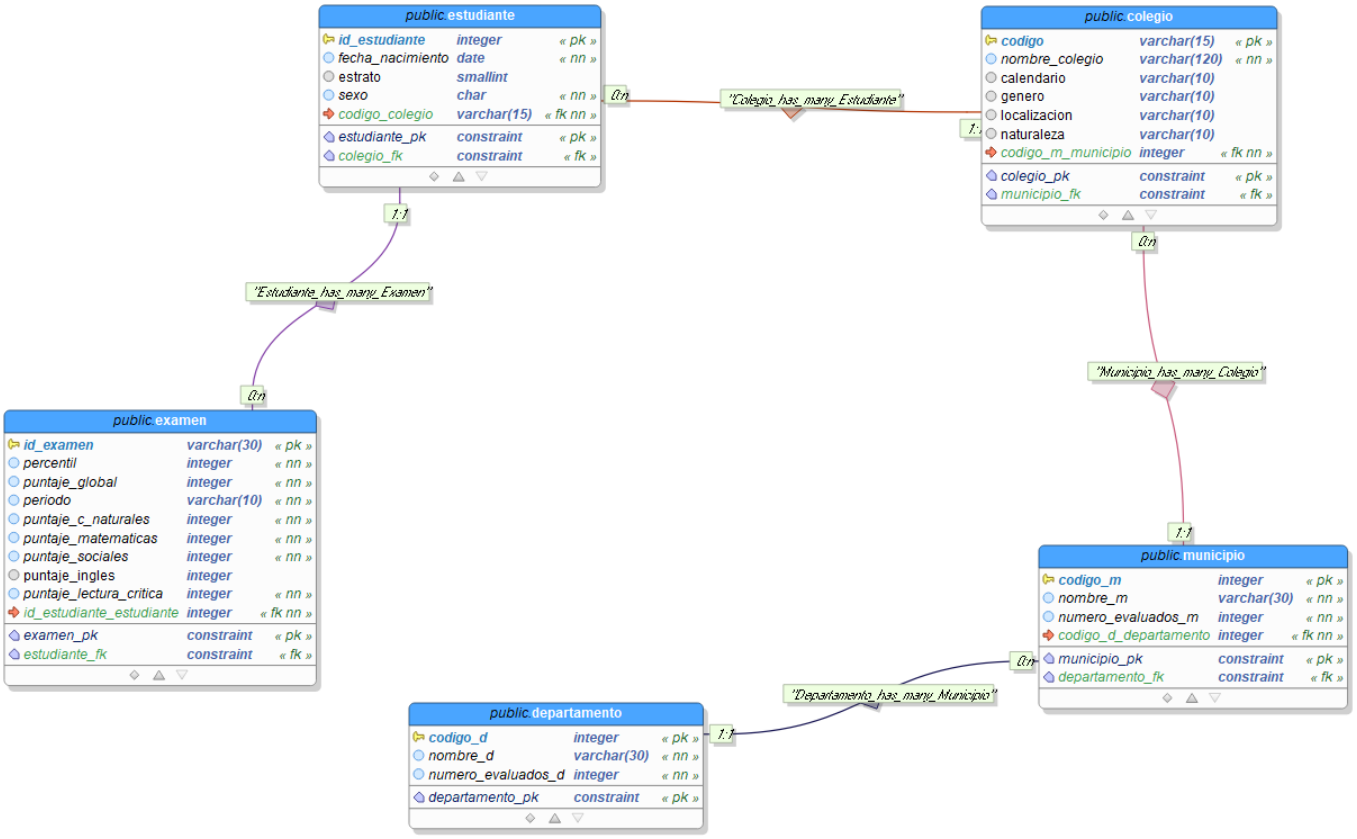
Se decide eliminar la entidad 'Sede' debido a inconsistencias en la base de datos utilizada, haciendo a una de sus llaves primarias (Id_sede) ambigua. Además, los datos que contenía la entidad 'Sede' resultaron poco relevantes para un futuro análisis, por lo que no son tenidos en cuenta.

Durante el proceso de carga de datos surgieron varios problemas debido a inconsistencias entre los tipos de datos de los atributos y sus equivalentes en la base de datos del ICFES, por lo que fue necesario realizar ciertas modificaciones:

- El atributo 'codigo' en la tabla 'colegio', pasa de ser tipo de dato entero a ser tipo de dato varchar.
- El atributo 'codigo_colegio' en la tabla 'estudiante', pasa de ser tipo de dato entero a varchar
- La restricción not null del atributo 'puntaje_ingles' de la tabla examen, se elimina.

Finalmente, las demás tablas permanecen igual a las presentadas en la primera entrega.

2. Modelo Relacional en Tercera Forma Normal



3. Proceso de carga de los datos.

Para las tablas **'departamento'** y **'municipio'** se realiza la carga de datos desde un archivo .csv a la base de datos 'proyecto' sin ningún problema.

Para la tabla **'colegio'**, se detectan varios problemas a la hora de cargar los datos desde el archivo Colegio2.csv.

- En primer lugar, la columna 'codigo' del archivo .csv contiene datos no soportados para el tipo de dato integer asignado en PostgreSQL.

  Debido a que en la definición del atributo 'codigo' en la tabla 'colegio' se le asignó el tipo de dato integer, fue necesario cambiar el tipo de dato a varchar(n), donde n es la longitud de la cadena, por lo cual es la opción más viable dado que en este caso la longitud de los datos de la columna puede ser de tamaño variable.

  Para realizar el proceso descrito anteriormente, primero hubo que eliminar la restricción de que el atributo 'codigo_colegio' era llave foránea en la tabla 'estudiante', luego se eliminó la restricción de que el atributo 'codigo' era llave primaria en la tabla 'colegio'. Una vez realizado esto, se procede a cambiar el tipo de dato del atributo 'codigo' en la tabla 'colegio' a varchar(15), lo mismo ocurre para el atributo 'codigo_colegio' en la tabla 'estudiante'. Después de esto, hay que volver a crear las restricciones eliminadas anteriormente para no afectar el correcto funcionamiento de la base de datos.

- Un segundo problema se presentó con el atributo 'nombre_colegio' de la tabla 'colegio' puesto que algunos nombres de colegios exceden la longitud máxima para el tipo de dato varchar(30), luego se optó por cambiar la longitud a varchar(120).

- Se presentó un último problema con la columna 'codigo_m_municipio' del archivo .csv debido a que algunos valores no presentaban un tipo de dato numérico o en su defecto "-". Por lo cual fue necesario buscar en internet a cada colegio que presentaba el error y averiguar su municipio de ubicación. Una vez ubicado el municipio, se buscó su respectivo código y se insertó en la columna 'codigo_m_municipio'.

Por último se realiza la carga de datos para la tabla **'colegio'** correctamente.

Para la tabla **'estudiante'** surgieron algunos errores a la hora de cargar los datos.

- Inicialmente, se presentó un error en el formato de la fecha de la columna 'fecha_nacimiento' de el archivo .csv, pues la mayoría de los datos de esta

columna aparecían con el formato MM/DD/YYYY el cual no es aceptado para el tipo de dato date en PostgreSQL. Debido a esto, se intentó cambiar el formato de la columna en el propio archivo .csv pero esto únicamente alteraba algunos datos y los demás se mantenían igual a como estaban en un inicio. Después de intentar varios métodos y no obtener un resultado favorable, se decidió implementar un código en C++ para darle solución a dicho problema.

Descripción de la solución:

Se selecciona la columna 'fecha_nacimiento' del archivo .xlsx de la base de datos original del ICFES. A continuación, se crea un archivo .csv el cual contiene la columna 'fecha_nacimiento' copiada de la base de datos original. Este archivo .csv se convierte a formato .txt pues de esta manera podemos realizar los cambios correspondientes explicados en el código FixingDate.cpp (ubicado en la carpeta code en GitHub) para solucionar el problema del formato de fecha. Así, la ejecución del código genera un nuevo archivo .txt con el formato de fecha corregido y aceptado por PostgreSQL (YYYY-MM-DD). Finalmente, de este archivo .txt movemos los datos a el archivo .csv que contiene los atributos correspondientes a la tabla 'estudiante'.

- El último problema que se presentó para la tabla 'estudiante' ocurrió con la columna 'estrato'. Algunos datos de esta columna contenían tipos de datos no numéricos diferentes a smallint, tal como se declaró el atributo 'estrato', por consiguiente, se escogió un número indefinido (-1) para estos datos no numéricos que hacen referencia a los estratos cuyo valor es desconocido. Es decir, los estudiantes cuyo estrato es desconocido, se les asignará un valor por defecto de -1.

Realizadas dichas modificaciones y correcciones, se procede a realizar la carga de datos de la tabla **'estudiante'** sin mayores problemas.

Finalmente, para la tabla **'examen'** ocurren algunos problemas de naturaleza similar que con la tabla 'estudiante'.

- Primero, en la columna 'percentil' aparecen datos no numéricos por lo que se le asigna un valor de entero -1 a cada una de estas celdas y así se cumple con la condición de que todos los datos de esta columna deben ser de tipo integer.

- Segundo, en la columna 'puntaje_ingles' aparecen tipos de datos nulos ya que cierto porcentaje de la población no presenta la prueba de inglés al momento de tomar las pruebas ICFES Saber 11. En la declaración de la tabla examen se creó una restricción que impedía que los valores del atributo 'puntaje_ingles' sean nulos, luego fue necesario eliminar esta restricción.

Solucionados estos problemas, se realiza la carga de datos en la tabla **'examen'**.

4. Posibles Análisis identificados.

- Dado a la realidad de desigualdad socioeconómica en el país consideramos importante poder ver y comparar los promedios del puntaje global obtenido por los estudiantes agrupados por estrato. Cómo Hipótesis inicial, planteamos que en los estratos más bajos se obtuvieron peores puntajes que en los estratos más altos

- Lo mismo buscamos obtener comparando el puntaje global promedio por municipios.

- Ver cuáles son las diferencias de los resultados obtenidos por la naturaleza del colegio (público/privado).

- Comparar por departamentos qué tanto manejan los estudiantes el segundo idioma (en este caso inglés) dado los resultados obtenidos en la prueba de inglés.

- Contrastar en que áreas de la prueba tuvieron un mejor o peor desempeño las mujeres y los hombres.

- Rankear los mejores colegios de Bogotá dado el desempeño de sus estudiantes en las pruebas ICFES Saber 11.

5. URL GitHub.

URL del repositorio:

`https://github.com/DavPlazas/Proyecto-Manejo-Bases-de-Datos`