Segunda Entrega Proyecto Ingeniería de Datos

González Laura, Mora Diryon, Rincón Laura

Jueves, 12 de mayo de 2022

1. Base de datos

Colombianos registrados en el exterior.

Como se puede apreciar, la base de datos seleccionada incluye la información de los colombianos que viven en el extranjero, los datos son totalmente anónimos e incluye información desde el 2016 hasta inicios del 2022. Sin embargo, no cuenta con la fecha en la que decidieron emigrar. Por otro lado, con el fin de mantener la información, se cuenta con el nombre del país donde viven y su respectivo código ISO, junto a la ciudad donde están registrados con su respectiva coordenada geográfica.

Se tiene la edad del colombiano, como también el área en la que trabaja y qué rama comprende esta. De manera independiente a lo anterior se incluye el nivel académico del individuo, junto a otros datos personales como lo son el género y la estatura. Para finalizar los datos enumeran la cantidad de personas que cumplen con exactamente las mismas condiciones mencionadas anteriormente.

2. Diagrama Entidad Relación

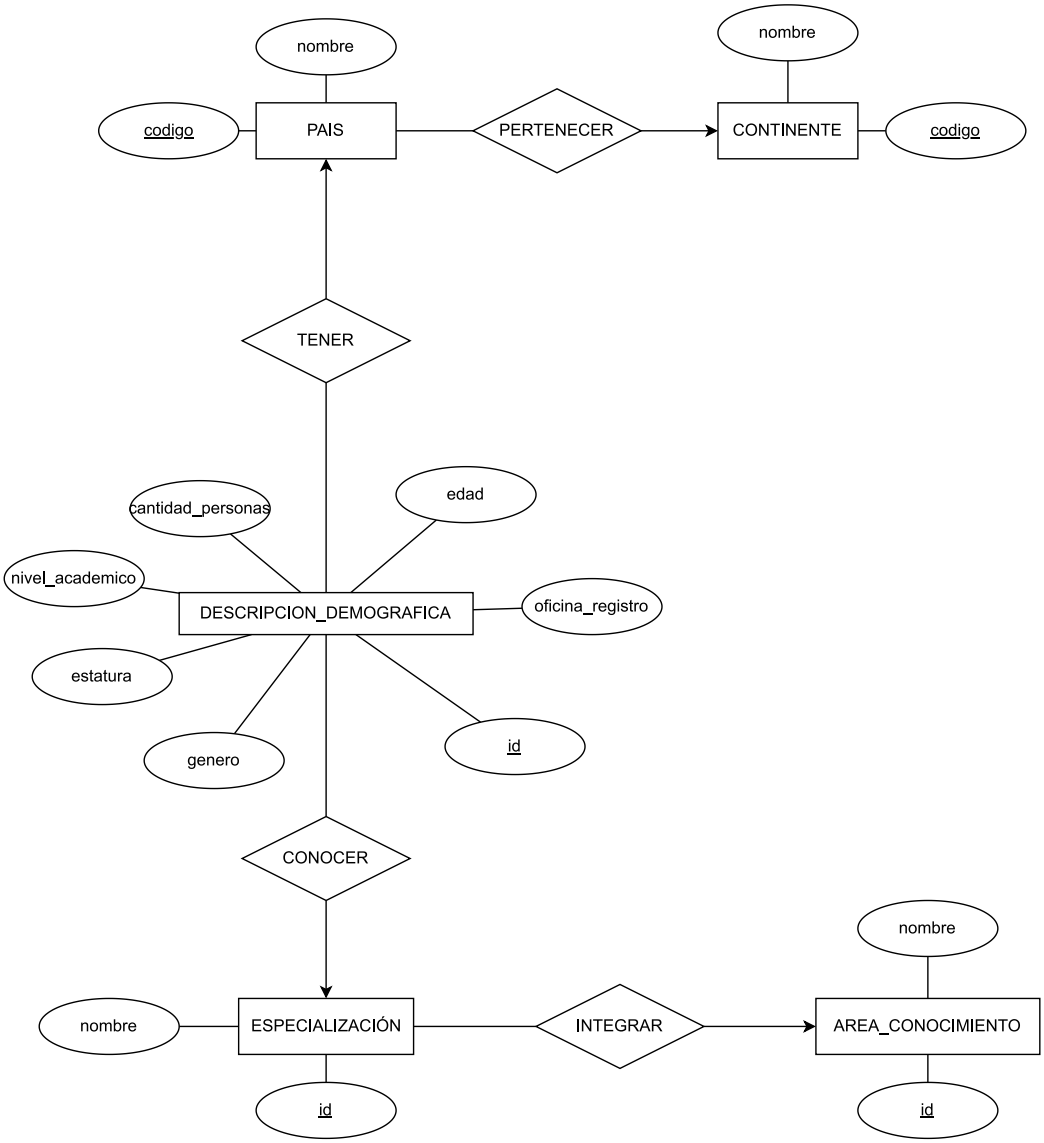

Figura 1: ER

3. Diagrama Relacional Normalizado

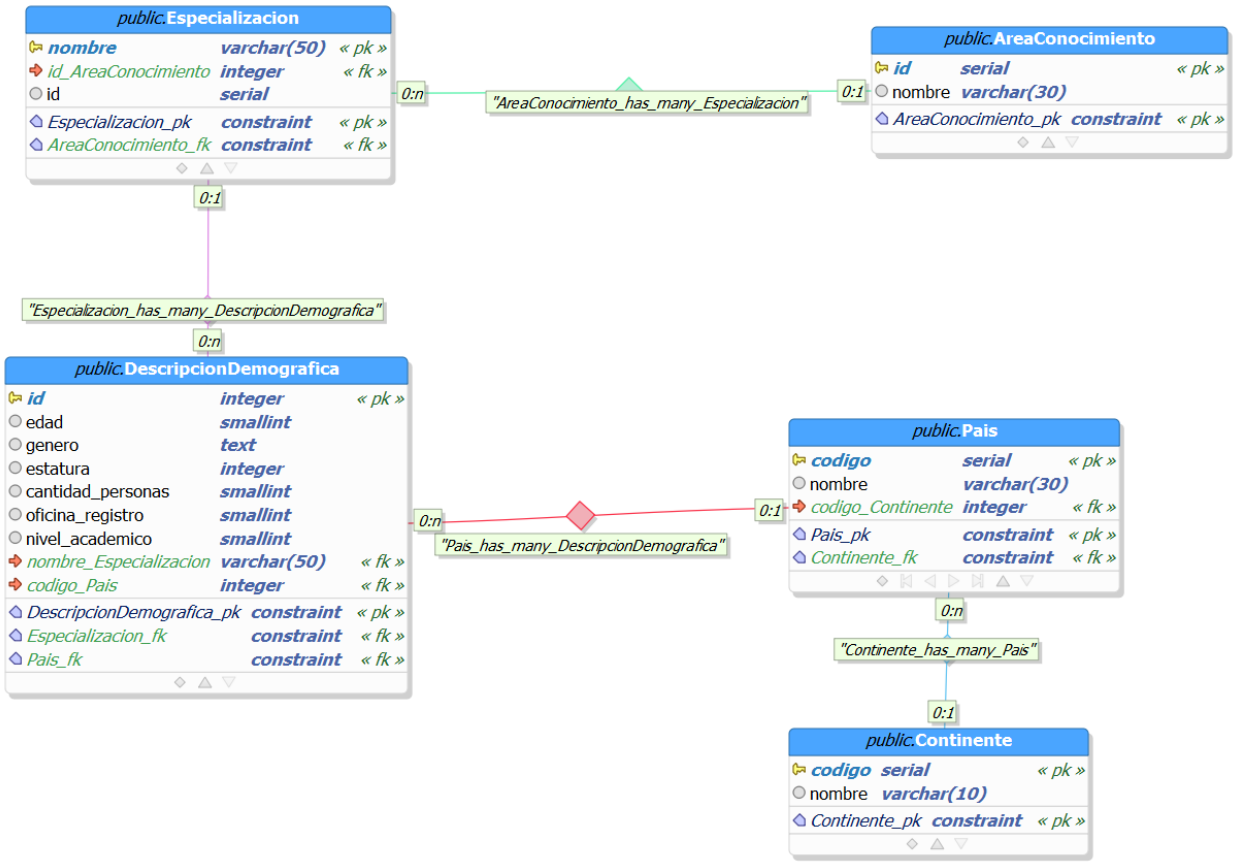

Figura 2: Diagrama Relacional Inicial

En la figura 2, se aprecia el diagrama relacional inicial. para las reglas definidas, se tienen datos atómicos, donde en cada columna de todas las tablas solo se puede ingresar un valor. Sin embargo, ocurren errores con los datos nivel_academico y oficina_registro, a la hora de insertar o actualizar algún dato. Si bien dependen de la persona, podría darse la situación de querer otra oficina_registro y no se podría, pues falta toda la información del individuo en cuestión. De igual forma pasa con el nivel_academico, que podrían variar en países y existir más de estos datos, cosa que actualmente no se podría ver reflejado en la base de datos.

Además, en el caso de oficina_registro esta más relacionada al id del PAIS que al propio id del usuario. Haciendo que DESCRIPCION_DEMOGRAFICA esté relacionada con Oficina_Registro y esta con PAIS, generando una Dependencia Transitiva, mucha redundancia y poca facilidad para hacer cambios en el futuro.

Figura 3: Diagrama Final

Finalmente, con la creación de las tablas OFICINA_REGISTRO y NIVEL_ACADEMICO, que se solucionan los problemas de redundancia de los datos. Facilitando la inserción y actualización de los datos. Con esta modificación cada dato tiene un numero único que lo identifica por tabla, teniendo así dependencia funcional completa. Además, no existe la posibilidad de tener dependencia funcional transitiva, pues cada dato esta relacionado directamente con su PK y no con otra tabla y otro PK.

4. Escenarios de Análisis

Para este proyecto, se ha decidido plantear los siguientes escenarios para su posterior análisis y planteamiento de conclusiones:

4.1. Primero

Determinar las ciudades y los países con mayor y menor oferta laboral. Para esto se planea mostrar la cantidad de personas que se encuentran en cada zona, ayudándose de graficas de barras, circulares y de mapa.

4.2. Segundo

Analizar las áreas y especializaciones con mayor y menor demanda. Esto se observara en las ciudades, países, continentes y el mundo (tomando la totalidad de los datos). Esto se implementara a partir de (inserte nombre de diagramas).

4.3. Tercero

Teniendo en cuenta las descripciones demográficas de los colombianos, se observara los rangos de edades más y menos comunes para emigrar. Esto considerando los años que más se repiten, los datos atípicos y partir de cuál edad se tiene un 25 %, 50 % y 75 % de personas. Todo esto a partir de diagramas de cajas y bigotes e histogramas ('box' y 'hist' en la librería pandas).

4.4. Cuarto

Examinar los niveles educativos requeridos para trabajar en una ciudad, país y continente. Considerando las personas que se encuentran en cada una de las ubicaciones. Esto se lograra por medio de tablas de frecuencia.

4.5. Quinto

Investigar si existe algún tipo de preferencia por genero en cada país y continentes. Para así poder especular sobre si existe algún tipo de machismo o feminismo significativo. Para esto se tendrá en cuenta el genero de las personas ubicadas en cada país y continente, comparando los porcentajes. Se puede apoyar en diagramas de barras o circulares.

5. Implementación

La implementación de la base de datos en PostgreSQL y los módulos de Python están disponibles en el repositorio de github. En este se encuentra un archivo README.es.md con las instrucciones de que contiene cada carpeta y la forma de guiarse. También se encuentra un archivo README.md en inglés.