Almacén de Datos
Grupo 1 - Semestre I 2024
Synthia Pamella Gonzalez Rodriguez

Dataset:
Alzheimer_s_Disease_and_Healthy_Aging_Data

El presente dataset hace referencia a encuestas o estudios. Los datos están organizados por años y ubicaciones geográficas, y también incluyen categorías como edad o grupo étnico.

Además, cada registro tiene identificadores únicos para diferentes columnas, como RowId, LocationID, ClassID, TopicID, QuestionID, entre otros. Este dataset cuenta con un total de 284,142 registros.

Carga de datos del CSV a SQL Server:
Con la tarea "Import Flat File", se cargaron los datos desde el archivo CSV a la nueva tabla "dbo.AlzheimersDisease" en la base de datos AlzheimerStaging. Esta tabla está destinada a almacenar los registros que sean importados. En la configuración, se ajustó el tamaño de la columna "rowId" a 100 caracteres, ya que se requerirá durante el proceso de transformación de los datos.

Transformación de datos
Una vez los datos son cargado a la tabla “dbo.AlzheimersDisease” se creó la tabla “dbo.AlzheimersTransformada” con un query “Creación Tabla Transformada” en SQL SERVER dentro de la misma base teniendo en cuenta las variaciones que se van a sufrir con la transformación.

Para transformar los datos, se hace desde un archivo python “staging.py”, el cual lee un query sql. “SelectTable” consulta que hace un select a la tabla, los datos obtenido se guardan en un dataframe para proceder a realizar dicha transformación.

La transformación de datos consistió en:

Eliminar valores nulos en las columnas Data_Value, Low_Confidence_Limit, High_Confidence_Limit, StratificationCategory2, Stratification2, AgeGroup, y Geolocation, reemplazándolos con valores predeterminados como cero o 'No Data Available' según el tipo de dato de la columna.

Las columnas Data_Value y Data_Value_Alt tienen el mismo número se dejo solo una columna en la tabla transformada.

La columna Stratification1 pasa a llamarse “Age Group” en la tabla transformada, también se eliminó la palabra "years" de los registros y se transformó a un formato de "65 years or older" a ">=65".

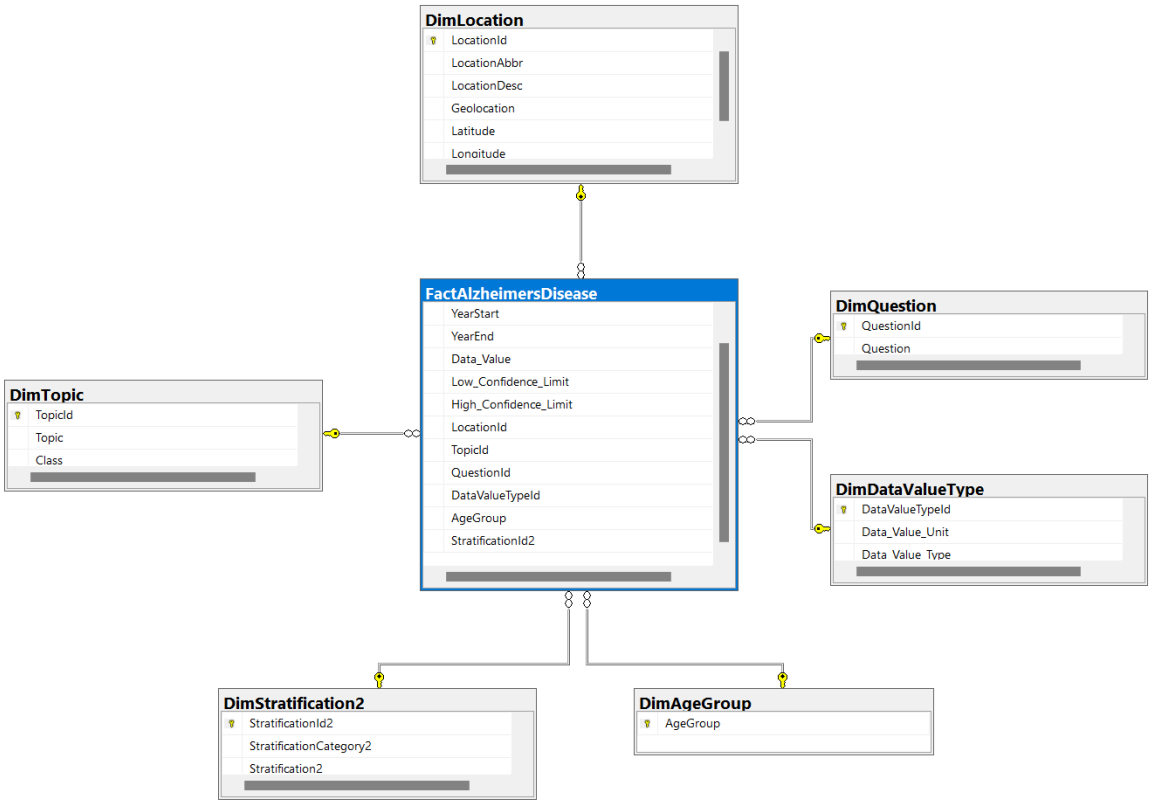
La columna “Geolocation” se separó en dos nuevas columnas, Latitude y Longitude, extrayendo y asignando las coordenadas.

La columna “RowId”, funciona como identificador único de cada registro, sin embargo, debido a la forma en que estaba estructurado se repetía pero hacía referencia a estudios diferentes por lo que se ajustó para incluir los componentes que pudieran hacerlo único y que no se repitiera, AgeGroup y Stratification2.

Insertar datos transformados a tabla “dbo.AlzheimersTransformada”

Una vez se termina de hacer la transformación de los datos, se eliminan los registros que hay en la tabla antes de realizar la carga.

Modelo Dimensional



El modelo dimensional se creó en la base de datos AlzheimerDW. Cuenta con una tabla Fact y 6 dimensiones.

DimAgeGroup contiene una sola columna que hace referencia al grupo de edades, está a la vez funciona como su primary key.

Se hizo una tabla dimensional para el topic ya que este es único y se incluye la class a la que pertenece.

Carga de datos de staging a DW

Para realizar la carga de datos a las tablas dimensionales desde el archivo python “alzheimerDW” se leyeron los archivos SQL correspondientes a cada dimensión ubicados en la carpeta "QuerySql" . Cada archivo contiene la consulta SQL para insertar los datos transformados en la tabla dimensional específica. Los archivos son DimQuestion.sql, DimLocation.sql, DimTopic.sql, DimDataValueType.sql, DimAgeGroup.sql, DimStratification2.sql y FactTable.sql. Los datos que arrojan estas consultas se almacenan en diferentes dataframes para luego insertarlos en cada tabla del modelo dimensional. Al igual que en el staging, siempre se elimina antes de insertar para evitar valores duplicados.