Sergio Pardo

Juan Diego Calixto

Nathalia Quiroga

# Proyecto 2 - Análisis sobre el asma

1. Identificar necesidades analíticas

   En la plantilla de Excel se encuentran los cuatro temas analíticos a desarrollar.

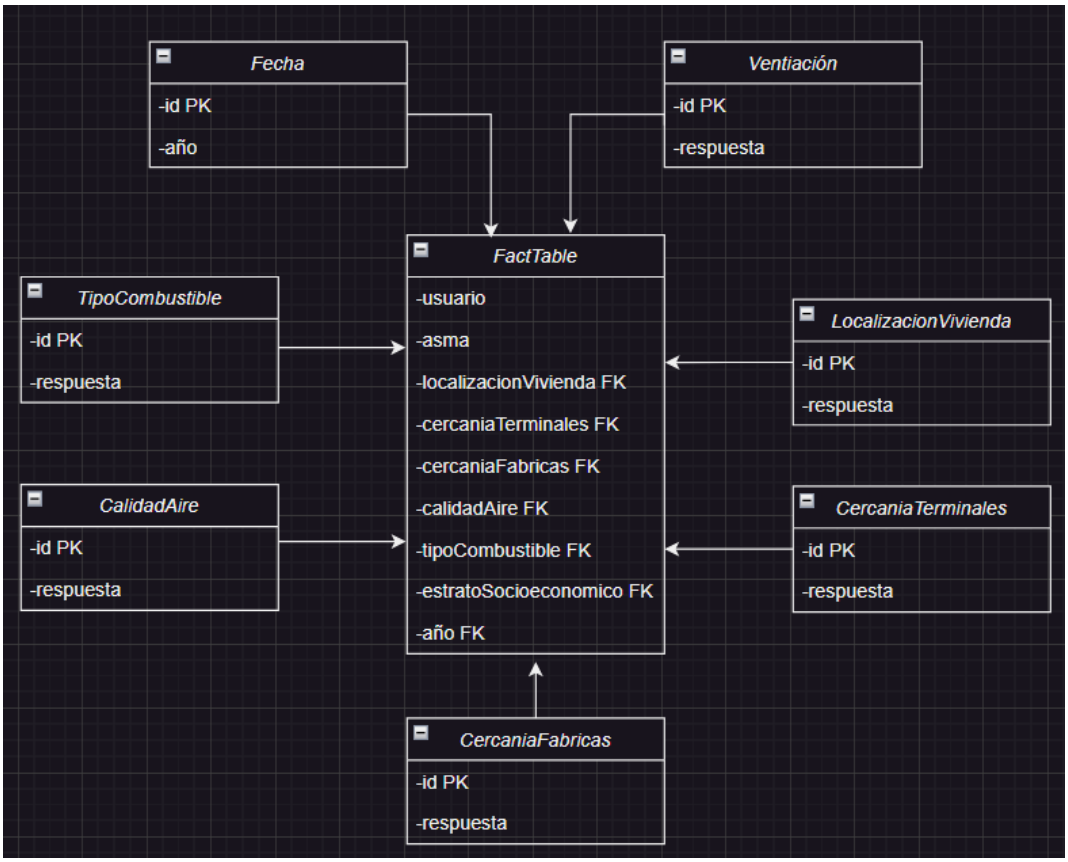
   **Justificación:**

   Durante la entrevista con los estudiantes de medicina, estos expresaron su interés en comprender cómo el tipo de combustible utilizado en los hogares de Bogotá afecta la salud en términos de asma. Si bien consideramos que esta pregunta tenía un enfoque analítico sencillo, la tomamos como un requisito específico para el análisis. A raíz de esto, decidimos ampliar nuestro enfoque y explorar otros requisitos analíticos que se relacionaran con la pregunta inicial planteada por los estudiantes de medicina. Estos requerimientos se encuentran en el Excel que se comentó anteriormente. Por último, los requerimientos analíticos no seleccionados para implementar en esta entrega se consideran sugerencias para un futuro proyecto en la misma temática.

2. Modelar *Data Marts*

   a. Modelo multidimensional:

   El modelo multidimensional se centra en el análisis de datos relacionados con el asma y factores asociados en hogares de usuarios. Este permite realizar análisis y consultas para comprender las relaciones entre el asma y factores como la localización de la vivienda, la cercanía a terminales de bus o fábricas, la calidad del aire, el tipo de combustible y la ventilación. Con este modelo, se puede explorar y evaluar la influencia de estos factores en la presencia del asma en los hogares de los usuarios.




- Dimensión de Usuario: Representada por la tabla "Usuario". Es la dimensión clave que identifica a los usuarios del modelo.
- Dimensión de Asma: Representada por la columna "Asma" en la tabla "FactTable". Es una medida que indica la presencia o ausencia de asma en los usuarios.
- Dimensión de Localización de la Vivienda: Representada por la tabla "LocalizacionVivienda". Proporciona información sobre la ubicación de la vivienda de los usuarios.
- Dimensión de Cercanía a Terminales de Bus: Representada por la tabla "CercaniaTerminales". Indica si tienen o no proximidad a terminales de autobús.
- Dimensión de Cercanía a Fábricas: Representada por la tabla "CercaniaFabricas". Indica si tienen o no proximidad de la vivienda a sectores industriales o fábricas.

- Dimensión de Calidad del Aire: Representada por la tabla "CalidadAire". Proporciona información sobre la percepción de los usuarios sobre la calidad del aire en su entorno.
- Dimensión de Tipo de Combustible: Representada por la tabla "TipoCombustible". Indica el tipo de combustible utilizado en la cocina de la vivienda de los usuarios.
- Dimensión de la ventilación: Representada por la tabla "Ventilación". Proporciona información sobre la ventilación de la vivienda de los usuarios.
- Dimensión de Fecha: Representada por la tabla "Fecha". Proporciona información sobre el año específicamente.

b. Justificación del modelo

i. Usuario: Cada fila en la FactTable representa un usuario específico.

Año: Se asume que el atributo "año" en la FactTable relaciona la tabla de hechos con la dimensión de Fecha, por lo que se puede inferir que cada fila representa datos agregados para un año específico.

Por lo tanto, la granularidad de los datos en la tabla de hechos estará en el nivel de detalle de "asma de usuario por año". Esto significa que cada fila en la FactTable representará una combinación única de un usuario y un año específicos, con las medidas asociadas como el valor de asma. Además de incluir información adicional sobre las condiciones de vivienda del usuario.

ii. *Asma*: Esta medida indica la presencia o ausencia de asma en los hogares de los usuarios. La medida se vuelve semi-aditiva porque puede realizarse una agregación o sumatoria en ciertos contextos, específicamente cuando se calcula el porcentaje de usuarios que tienen asma en un conjunto de usuarios. Sin embargo, es importante tener en cuenta que, en la mayoría de los otros contextos analíticos, la medida asma se puede considerar como no aditiva ya que no tendría sentido sumar o promediar los valores de asma para diferentes usuarios o períodos de tiempo, ya que la medida solo indica si un usuario tiene o no asma,

y esto no representarían información significativa sobre la presencia de la enfermedad en el conjunto de hogares.

iii. Fecha: Este atributo representa el año y, por lo tanto, no requiere un manejo de historia de variación lenta, ya que no hay cambios lentos o evoluciones en los años. Cada año es un valor fijo y se utiliza para proporcionar contexto temporal en el análisis.

Respuesta de Usuario (para Localización de la Vivienda, Cercanía a Terminales de Bus, Cercanía a Fábricas, Calidad del Aire y Tipo de Combustible): Estos atributos si cambian lentamente y es relevante mantener un historial de los valores anteriores, se podría implementar un manejo de historia de variación lenta de tipo Tipo 2. Esto permitiría registrar los cambios en las respuestas de los usuarios a lo largo del tiempo, con registros separados para cada valor histórico junto con la fecha de inicio y fin de vigencia de cada valor. La elección de un manejo de historia de variación lenta permitiría analizar la evolución de las respuestas de los usuarios.

3. Entendimiento de los datos, creación del Datamart y proceso del ETL

Al revisar la fuente de datos, que en este caso fue la encuesta multipropósito del DANE para los años 2017 y 2021, notamos que todos los usuarios en 2017 tenían asma, mientras que en 2021 ninguno lo tenía. Sin embargo, logramos combinar estos datos con otras fuentes para los mismos años y así obtener registros tanto de usuarios con asma como sin esta. Para identificar las preguntas más relevantes relacionadas con nuestros requerimientos analíticos, utilizamos los diccionarios que adjuntaban los datos. De esta manera, pudimos determinar las columnas y tablas que utilizaríamos en nuestro modelo multidimensional.

En cuanto al proceso de ETL (Extracción, Transformación y Carga), empleamos Python, mientras que para la creación de la bodega de datos utilizamos PostgreSQL. Además, establecimos una conexión con Tableau para diseñar los tableros de control que cumplen con los cuatro requerimientos analíticos planteados inicialmente.

4. Arquitectura de solución de BI

   A continuación, se muestran los pasos que se siguieron para encontrar finalmente la solución que se implementó para los requerimientos analíticos:

   a. Fuentes de datos: La encuesta multipropósito del DANE para los años 2017 y 2021 fue la fuente principal de datos.
   b. Proceso ETL: Se utiliza Python como herramienta para realizar la extracción, transformación y carga de los datos desde las fuentes de datos (csv) hacia la capa de almacenamiento.
   c. Bodega de datos (Data Warehouse): Se implementó una estructura de almacenamiento de datos en PostgreSQL que contiene las tablas dimensionales y la tabla de hechos necesarias para el modelo multidimensional. Esta capa de almacenamiento será la base para realizar consultas y análisis en el entorno de BI.
   d. Modelo de datos: Se definió el modelo multidimensional el cual incluye las dimensiones y la tabla de hechos mencionadas.
   e. Tableros de control: Para interactuar con el modelo multidimensional y realizar análisis, se utilizó Tableau como herramienta de BI que permitió la conexión con la base de datos PostgreSQL. Con esta herramienta se crear consultas, informes y visualizaciones basadas en los datos del modelo que respondan a los requerimientos analíticos planteados inicialmente.