



**DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**  
**Inteligencia de negocios 202220 – Proyecto 2**  
PROFESORA: Haydemar Núñez

Nombres	Apellidos	Código	Login
María Sofía	Álvarez López	201729031	ms.alvarezl
Brenda Catalina	Barahona Pinilla	201812721	bc.barahona
Álvaro Daniel	Plata Márquez	201820098	ad.plata

### Proyecto 1 – Etapa 2: Automatización analítica de textos

El objetivo de este proyecto es analizar indicadores socioeconómicos y dimensiones municipales de Colombia. El proyecto busca aprovechar los datos abiertos para generar conocimiento útil que aporte a diferentes tipos de actores, desde el ciudadano normal hasta el experto en temáticas allí trabajadas, pasando por actores como periodistas, entre otros. De igual manera, quieren ampliar la gama de análisis que se pueden observar en la página web de la iniciativa

#### 1. Identificación de necesidades analíticas

A partir de entrevistas realizadas a personas relacionadas con el proyecto que están trabajando, los datos compartidos y la retroalimentación a la consultoría realizada en semestres anteriores, se identificaron y documentaron algunos requerimientos analíticos, utilizando la matriz de requerimientos de negocio, los temas analíticos, análisis requeridos, procesos de negocio, fuentes de datos y datos requeridos siguiendo la metodología Kimball.

A continuación, se presentan los temas analíticos propuestos, con sus respectivos análisis requeridos/inferidos y las justificaciones respectivas, que se pensaron para este proyecto. Para encontrar la tabla completa, remítase al archivo EntregaAnalisisRequeridos.xlsx

#### Tema analítico: Conflicto armado.

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de control, análisis OLAP, Minería de datos	Procesos de negocio	Fuentes de datos y datos
Conflicto armado	Análisis de homicidios * año * departamento * causa * municipio * sexo	Análisis OLAP	Registro de afectaciones en el conflicto armado	Excel ConflictoArmado
	Análisis de número de personas secuestradas * año * departamento * municipio	Análisis OLAP	Registro de demografía y población	Excel DemografíaPoblación
	Visualización de las afectaciones del conflicto con base en el análisis 1 y 2	Tablero de Control	Registro de afectaciones en el conflicto armado y Registro de demografía y población	Datamart sobre el conflicto armado

Una de los temas de mayor interés en Colombia es el análisis del conflicto armado, ya que ha sido una de las crisis más grandes que ha existido en la historia de América del Sur. Siendo una guerra asimétrica, resulta conveniente analizar la cantidad de homicidios (por año - porque hay años en que el impacto del conflicto fue más fuerte -, por departamento - porque hay departamentos en que el conflicto es más fuerte y tiene mayor incidencia -, por causa - porque resulta conveniente entender las razones principales por las cuales ocurren los homicidios en el conflicto armado, con el fin de intentar mitigarlo - y por sexo - porque es fundamental entender la incidencia del conflicto en este ámbito). De manera similar, el número de personas secuestradas puede analizarse por indicadores análogos. Esto también es fundamental, porque hubo años y regiones del país en los cuales el número de secuestros alcanzó cifras altamente preocupantes.

Una de las mejores formas de visualizar esto es, por ejemplo, utilizando mapas de burbujas. Estos son una excelente herramienta de visualizar los sitios con mayor presencia de un fenómeno en particular. El mapa mostrado a continuación, por ejemplo, ilustra la proporción de surfistas en el mundo:

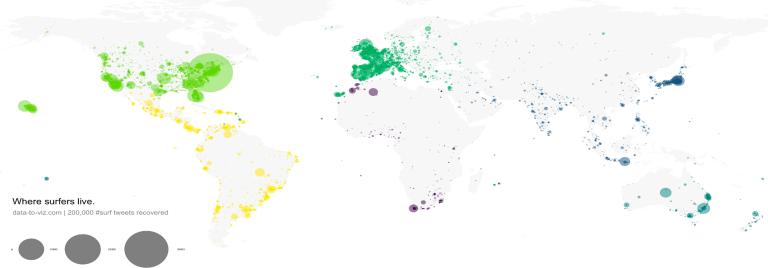


Figura 1: Número de surfistas en el mundo.

Note que para poder realizar adecuadamente estos diagramas, es necesario que los indicadores estén normalizados a la población del país.

#### Tema analítico: Educación.

<b>Educación</b>	Análisis de cobertura neta * año * mes * departamento * municipio	Análisis OLAP	Registro de datos sobre educación	Excel Educación
	Visualización del análisis indicadores educativos con base en el análisis 1	Tablero de Control	Registro de datos sobre educación	Datamart sobre la educación

El segundo tema analítico de interés es la educación. Siendo la educación uno de los principales motores de los países en vía de desarrollo para lograr la igualdad, resulta de especial interés analizar la cobertura neta de educación a nivel nacional. Es común que las áreas rurales tengan desarrollos educativos mucho más elementales y precarios que las grandes ciudades: no sólo porque acceder a estos municipios es más difícil; sino que, dado el bajo desarrollo social y económico, muy pocos profesores están dispuestos a mudarse a esas áreas. Lograr ampliar la cobertura neta de educación a nivel nacional resulta crucial puesto que es un indicador fundamental del avance económico y social del país. De manera similar al caso del conflicto armado, una visualización adecuada sería un mapa de burbujas, o un gráfico de barras corriente que evidencie el porcentaje de cobertura (normalizado a la población) para cada uno de los municipios/departamentos en cuestión.

### Tema analítico: Desempleo.

<b>Desempleo</b>	Análisis del indicador de desempleo * año * municipio	Análisis OLAP	Registro de mediciones de desempleo	Excel MediciondeDesempleoDepartamental
	Análisis del indicador de desempleo * año * departamento	Análisis OLAP	Registro de mediciones de desempleo	Excel MediciondeDesempleoMunicipal
	Visualización de análisis indicadores de desempleo municipal	Tablero de Control	Registro de mediciones de desempleo	Datamart sobre el conflicto armado
	Visualización de análisis indicadores de desempleo para departamentos	Tablero de Control	Registro de mediciones de desempleo	Datamart sobre desempleo

Gran parte de la población colombiana es bastante joven. Muchas veces, al graduarse de la Universidad o del colegio, o incluso aquellos que desertan de la educación, se ven enfrentados a una época de desempleo prolongada mientras logran conseguir un empleo. La dinámica del desempleo varía cada año (pues también depende de factores macroeconómicos externos como la migración forzada), y por municipio (dependiendo de la edad de la población, el nivel educativo, el número de plazas de empleo disponibles y la cantidad de desplazados que el municipio/departamento recibe a diario). De manera similar al caso del conflicto armado, una visualización adecuada sería un mapa de burbujas, o una tabla cuadrada de área proporcional, como la que se presenta en la figura 2.

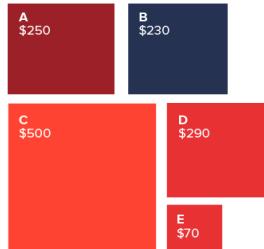


Figura 2: Tabla cuadrada de área proporcional.

Esta tabla podría mostrar (normalizado a la población) para cada municipio o departamento, la tasa de desempleo. Es interesante pues rápidamente el usuario puede notar cuál es el departamento/municipio de mayor incidencia (en el caso de la figura 2, sería la variable correspondiente al cuadrado C).

### Tema analítico: Salud.

<b>Salud</b>	Análisis del nivel de satisfacción del sistema de salud * año * departamento * municipio	Análisis OLAP	Registro de datos sobre la salud	Excel Salud
	Análisis de población afiliada a regímenes * año * departamento * municipio * tipo de régimen	Análisis OLAP	Registro de demografía y población	Excel DemografíaYPoblación
	Visualización de análisis indicadores de salud con base en el análisis 1 y 2	Tablero de Control	Registro de datos sobre la salud y Registro de demografía y población	Datamart sobre salud

Como país en vía de desarrollo, el indicador de la salud en Colombia es fundamental para entender la evolución del país hacia una cobertura igualitaria de los servicios básicos y asegurar una mayor expectativa de vida (y disminuir tasas como muertes de niños y jóvenes por enfermedades prevenibles también). Para ello, hay tres análisis que resulta conveniente analizar en este ámbito. El primero es el análisis del nivel de satisfacción del sistema de salud por año, departamento y municipio. Resulta fundamental que los usuarios sientan que están bien atendidos y que la salud en su municipio y departamento es favorable. Un buen indicador del manejo de la salud es la percepción de los ciudadanos.

Otro análisis que resulta relevante es el número de habitantes (normalizado a la población del departamento) que están afiliados a regímenes de salud, así como el tipo de régimen. Como punto básico, el gobierno debe asegurar que los usuarios tengan un régimen de salud (así sea subsidiado). No obstante, debido a las personas sin hogar, a los desplazados forzados o inmigrantes ilegales, siempre va a haber un porcentaje de la población que no esté afiliada. También puede dejar de haber personas afiliadas si los recursos no son suficientes para financiar la salud de todos los habitantes. Analizar esto es fundamental para determinar el avance económico y social del país.

Para poder visualizar esta información, es necesario utilizar tableros de control que permitan mostrar (por ejemplo, en un mapa), la satisfacción promedio de los usuarios por departamento, así como la cantidad de usuarios afiliados a un régimen (o cierto régimen de salud). Asimismo, la visualización de la tabla cuadrada de área proporcional también sería valiosa en este ámbito.

### Tema analítico: Vivienda y servicios públicos

Vivienda y servicios públicos	Análisis del indicador de cobertura * año * departamento * municipio * tipo de servicio	Análisis OLAP	Registro de vivienda y servicios públicos	Excel ViviendayServiciosPublicos
	Análisis de penetración de banda ancha * año * departamento * municipio	Análisis OLAP	Registro de demografía y población	Excel DemografiaYPoblacion
	Visualización de análisis indicadores de servicios públicos con base en el análisis 1 y 2	Tablero de Control	Registro de vivienda y servicios públicos y Registro de demografía y población	Datamart sobre Vivienda y servicios públicos

Es importante analizar a su vez la cobertura de vivienda y servicios básicos de la población. Lo ideal es que todos los habitantes tengan acceso a estas facilidades; no obstante, debido a que aún queda un gran trecho por recorrer en este sentido, aún un gran porcentaje de la población no tiene acceso a vivienda, servicios básicos o los dos. Es importante normalizar este indicador a la población del departamento/municipio para que sea confiable.

Por otro lado, la cobertura de internet es fundamental. Es crucial que el país esté interconectado debido a la globalización a la que nos enfrentamos hoy en día. En aras de evaluar la conexión de todo el país al internet (y por ende, a un mayor y más amplio acceso a la información), la penetración de banda es un indicador fundamental. Para analizar y revisar ambos indicadores, es posible usarlo utilizando un mapa de burbujas o de calor (como el de la figura 3), dependiendo de la cobertura normalizada de vivienda y servicios públicos, o de penetración normalizada de banda ancha, para cada departamento/municipio.

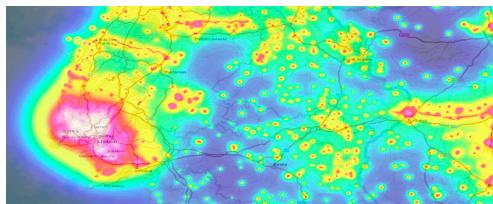


Figura 3: Ejemplo de mapa de calor.

## 2. Modelado de DataMarts

A partir de las fuentes de datos anexas a esta entrega y los requerimientos analíticos identificados en el literal anterior, se realizó un modelo dimensional que representa los tipos de requerimientos del cliente.

Con esto en mente, lo primero que se hizo fue elaborar el modelo multidimensional, con los nombres de atributos, llaves primarias, llaves foráneas y roles claramente definidos. El modelo obtenido fue el de la figura 4.

Para la creación de este modelo dimensional, seguimos el siguiente proceso:

1. **Seleccionar el proceso de negocio:** Para los requerimientos que deseamos realizar, deseamos colocar gran enfoque en los siguientes procesos de negocio: registro de Homicidios, registro de personas secuestradas, registro de la proporción de satisfacción de salud, registro de afiliados al sistema de salud y registro de la medición de desempleo departamental.
2. **Seleccionar granularidad:** En general, el nivel de granularidad de la mayoría de procesos es parecido. Escogimos este nivel de granularidad para satisfacer los requerimientos de análisis que identificamos anteriormente. Al analizar los datos nos damos cuenta que en estos podemos realizar una distinción de algunos indicadores según la ubicación y la temporalidad.
  - a. **Registro de homicidios:**  
Nivel de granularidad: a nivel de especificación de temporalidad, demografía, ubicación y causa
  - b. **Registro de personas secuestradas:**  
Nivel de granularidad: A nivel de especificación de temporalidad y de ubicación
  - c. **Registro de desempleo:**  
Nivel de granularidad: A nivel de especificación de temporalidad y de ubicación
  - d. **Registro de Satisfacción de salud:**  
Nivel de granularidad: A nivel de especificación de temporalidad, de ubicación y nivel de satisfacción
  - e. **Registro de afiliado:**  
Nivel de granularidad: A nivel de especificación de temporalidad, de ubicación y tipo de afiliación.

3. *Identificar dimensiones:* Para todos los procesos mencionados anteriormente decidimos realizar las vistas por fecha (mes y/o año) y lugar (departamento). Para los procesos de registro de homicidios, registro de la proporción de satisfacción de salud y registro de afiliados al sistema de salud se colocaron más dimensiones:

- a. Proceso de registro de homicidios: Dimensión de Causa de homicidios y Dimensión Sexo
- b. Proceso de registro de la proporción de satisfacción de salud: Dimensión nivel de satisfacción (Muy satisfecho, Muy insatisfecho, satisfecho, etc.)
- c. Proceso de registro de afiliados al sistema de salud: Dimensión tipo régimen, el cual especifica a que sistema de salud se hace referencia, por ejemplo Afiliado al régimen contributivo, afiliado al régimen subsidiado, afiliado a regímenes especiales y afiliados al SGSSS

4. *Identificar los hechos:*

- a. Registro de homicidios:

Porcentaje de homicidios (medida no aditiva). Esta es una medida no aditiva puesto que lo único que podemos obtener es un recuento de las filas, no es posible realizar cálculos adicionales sobre porcentajes.

- b. Registro de personas secuestradas:

Porcentaje de personas secuestradas (medida no aditiva). Esta es una medida no aditiva puesto que lo único que podemos obtener es un recuento de las filas, no es posible realizar cálculos adicionales sobre porcentajes.

- c. Registro de desempleo:

Porcentaje de personas que están desempleadas (medida no aditiva). Esta es una medida no aditiva puesto que lo único que podemos obtener es un recuento de las filas, no es posible realizar cálculos adicionales sobre porcentajes.

- d. Registro de Satisfacción de salud:

Porcentaje de personas que tienen un determinado nivel de satisfacción con el sistema de salud: (medida no aditiva). Esta es una medida no aditiva puesto que lo único que podemos obtener es un recuento de las filas, no es posible realizar cálculos adicionales sobre porcentajes.

- e. Registro de afiliado:

Porcentaje de personas afiliadas al régimen de salud (medida no aditiva). Esta es una medida no aditiva puesto que lo único que podemos obtener es un recuento de las filas, no es posible realizar cálculos adicionales sobre porcentajes.

Ahora, para cada uno de los atributos de las dimensiones planteadas, elegimos el tipo de manejo de historia, o si es de variación lenta, como se ve a continuación:

1. Dimensión tipo afiliación: El primer atributo es la llave primaria. El atributo de tipo régimen no necesita manejar historia. Basta simplemente con manejar el cambio y ya, puesto lo que importa al negocio es tener cuál es el régimen actual de salud de los individuos. El histórico de este atributo no es necesario.
2. Dimensión nivel satisfacción: El primer atributo es la llave primaria. Para el atributo de nivel satisfacción vale la pena llevar un manejo de historia tipo 3. Con ello, la entidad agrega un nuevo atributo para conocer cuál fue la anterior calificación de los usuarios y ver si, en los últimos meses, el servicio ha mejorado o empeorado.
3. Dimensión ubicación: Esta dimensión tiene una llave primaria, un código de departamento y un código de municipio. Es muy poco probable que alguno de los dos últimos atributos cambie en el tiempo. Por tanto, no se maneja historia en esta dimensión.
4. Dimensión sexo: Esta es una dimensión que tiene llave primaria y sexo. Aunque no muchas personas cambian su sexo, es posible que esto ocurra. No debe manejarse historia para este atributo, pues en ningún momento se busca hacer análisis sobre cambios de sexo.
5. Dimensión causa homicidio: Esta es una dimensión que tiene llave primaria y causa. La causa de un homicidio puede variar, dependiendo de los análisis de medicina legal, y por tanto puede ser importante llevar un récord de ello, porque puede ser información relevante para el negocio y para las investigaciones legales. Para ello, se puede usar un manejo de historia tipo 2, tal que haya un inicio y fin de una causa de homicidio determinada asignada y se pueda reconstruir la historia para análisis judiciales posteriores.
6. Dimensión Fecha: Utilizada por las otras dimensiones y la tabla de hechos. No requiere manejo de historia porque es más una dimensión que no será utilizada por sí sola ni maneja información del negocio.

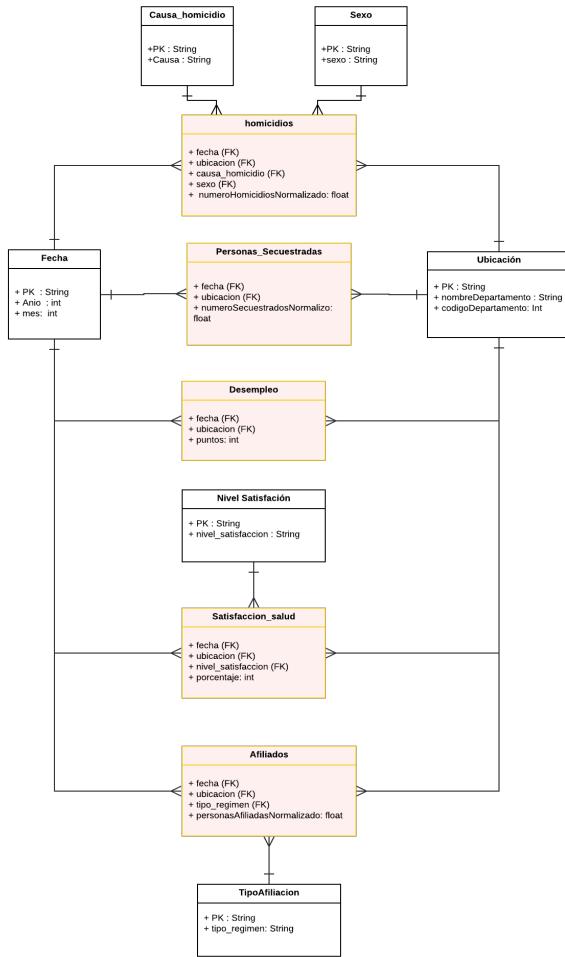


Figura 4: Modelo propuesto para el proyecto.

### 3. Entendimiento de los datos y proceso ETL

La sección a continuación describe el entendimiento y perfilamiento de los archivos de datos brindados por infraestructura visible. Un análisis más detallado se encuentra en la carpeta de perfilamiento de datos, la cual invitamos al lector a remitirse para ver los análisis realizados. A pesar de que se realizó un entendimiento de los datos para todos los archivos, como verá en el jupyter notebook de la carpeta previamente citada, aquí nos limitamos a analizar únicamente aquellos que usamos, a saber: conflicto armado, salud y demografía y población (para la normalización).

Veamos primero los datos de conflicto armado, correspondientes a indicadores (como asesinatos y secuestros) característicos de este fenómeno. Se tienen 34 códigos de departamento, correspondientes a 34 departamentos que hay en los archivos (32 departamentos de Colombia, Bogotá y un campo más correspondiente al campo 'Colombia' – con la información del país en general –). Asimismo, hay 1134 códigos distintos de entidades, pero tan solo 1044 entidades distintas, por lo que hay entidades que tienen dos códigos asociados. La columna subcategoría presenta algunos causas de homicidio, como: conflicto armado (de interés), homicidios por agresor y sexo de la víctima, seguridad ciudadana, presuntos delitos sexuales según sexo de la víctima, violencia de pareja según sexo de la víctima, percepción y violencia de género. La columna indicador tiene valores como: número de personas secuestradas y número de personas desplazadas, a lo cual responde el dato de la columna dato numérico, que fue posteriormente procesada, pero que a nivel de perfilamiento se encontraba como una variable categórica. Posteriormente, se tiene la columna unidad de medida, que representa la forma en la que deben medirse los datos de la columna dato numérico. Se tiene además la columna fuente, de donde se obtienen los datos (v.g. Instituto Nacional de Medicina Legal y Ciencias Forenses – INMLCF). Por último, se tiene una columna de dato cualitativo (que tiene un rango de años), una columna con el año del dato, y una columna con el mes del dato (12, en todas las filas del dataset).

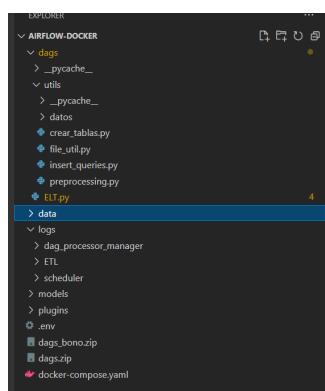
El dataset de salud tiene los mismos 34 códigos de departamento y departamentos, así como las 1044 entidades y los 1134 códigos distintos de entidades. Asimismo, hay tres categorías que pueden ser analizadas: tasas y coberturas (de salud), aseguramiento (¿cuánta población hay asegurada?) y percepción (¿cuál es la percepción dada al sistema de salud?). Por otro lado, se encuentra la tabla con

el indicador, que corresponde a lo que se está midiendo (v.g. porcentaje de nacidos vivos con bajo peso al nacer). Después, se tiene la variable de dato numérico que, de la misma forma que el de la dimensión anterior, es interpretada como una categórica (el preprocessamiento se realiza más adelante en el notebook). De la misma forma, se encuentra la columna de unidad de medida (v.g. Porcentaje (el valor está multiplicado por 100)), con el fin de poder realizar un preprocessamiento adecuado, la columna fuente (v.g. Ministerio de Salud), año (distribuido entre 1998 y 2020 con una media de 2013) y el mes del año (11 –noviembre– o 12 –diciembre–).

El dataset de población tiene la misma información de los 34 departamentos, con sus respectivos códigos, y los 1134 códigos de municipios con solo 1044 municipios. Asimismo, está la columna de subcategoría, que contiene la información del tipo de dato que se tiene (v.g. Población de mujeres), el indicador (v.g. Total registros válidos – Hogares), el dato numérico asociado a cada uno de estos indicadores, con su unidad de medida respectiva (v.g. mujeres u hombres). También se tiene la fuente de donde se obtuvo el dato (v.g. DNP), el año (entre 2010 y 2023 (dato atípico – aún estamos en el 2022 –)), con media del año 2020), y el mes (en este caso, sólo hay 5 valores diferentes: 1, 4, 6, 11 y 12. Esta tabla era crucial para el análisis normalizado de los indicadores de las tablas de conflicto armado y salud.

Para la creación del proceso ETL en el proyecto seguimos los siguientes pasos:

- Creación de la estructura del proyecto en la máquina virtual asignada:** Creamos la estructura del proyecto basándonos en la misma estructura del laboratorio 5. La estructura es la siguiente:

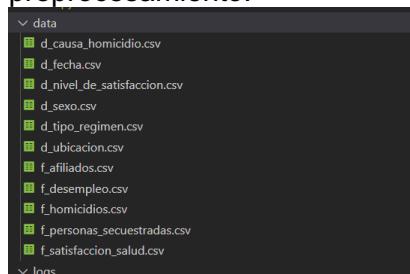


Los archivos en esta estructura cumplen las mismas funciones que en el laboratorio 5. Explicaremos más adelante el contenido y función de los más relevantes.

- Creamos el servidor de Airflow con la misma imagen de Docker utilizada en el laboratorio 5. Este servidor se crea con la ejecución de los siguientes comandos desde la carpeta raíz del proyecto:

- docker-compose up airflow-init
- docker-compose up

- Agregamos en la carpeta /data del proyecto los archivos .csv resultados del preprocessamiento.



- Creamos el archivo crear-tablas.py, que se encargará de ser el script de creación de tablas del modelo multidimensional

- e. Creamos el archivo file\_util.py, donde definimos las funciones de escritura y lectura de csv:

```
dados = pd.read_csv('airflow/dados/iris.csv', sep=',', encoding='utf-8', encoding_errors='ignore', index_col=False)
```

- f. Creamos el archivo `insert_queries.py`, donde definimos las funciones que transformarán el contenido de archivos csv a queries de inserción en SQL

```

from utilitarios.util import cargar_datos

def main():
    # Carga de datos
    insert_query_fechal = """INSERT INTO Fecha (Fecha_Key,Año,Mes) VALUES """
    insertQuery = " "
    insertQuery += insert_query_fechal
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{})".format(1, 2018)
    insertQuery += "({},{})".format(2, 2018)
    insertQuery += "({},{})".format(3, 2018)
    insertQuery += "({},{})".format(4, 2018)
    insertQuery += "({},{})".format(5, 2018)
    insertQuery += "({},{})".format(6, 2018)
    insertQuery += "({},{})".format(7, 2018)
    insertQuery += "({},{})".format(8, 2018)
    insertQuery += "({},{})".format(9, 2018)
    insertQuery += "({},{})".format(10, 2018)
    insertQuery += "({},{})".format(11, 2018)
    insertQuery += "({},{})".format(12, 2018)
    insertQuery += ";\n"

    insert_query_ubicacion = """INSERT INTO Ubicacion (Ubicacion_Key,Departamento,Codigo_Entidad) VALUES """
    insertQuery += insert_query_ubicacion
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{},{})".format(1, 1, 1)
    insertQuery += "({},{},{})".format(2, 1, 2)
    insertQuery += "({},{},{})".format(3, 1, 3)
    insertQuery += "({},{},{})".format(4, 1, 4)
    insertQuery += "({},{},{})".format(5, 1, 5)
    insertQuery += "({},{},{})".format(6, 1, 6)
    insertQuery += "({},{},{})".format(7, 1, 7)
    insertQuery += "({},{},{})".format(8, 1, 8)
    insertQuery += "({},{},{})".format(9, 1, 9)
    insertQuery += "({},{},{})".format(10, 1, 10)
    insertQuery += "({},{},{})".format(11, 1, 11)
    insertQuery += "({},{},{})".format(12, 1, 12)
    insertQuery += ";\n"

    insert_query_genereo = """INSERT INTO Genero (Genero_Key,Genero) VALUES """
    insertQuery += insert_query_genereo
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{})".format(1, "Masculino")
    insertQuery += "({},{})".format(2, "Femenino")
    insertQuery += ";\n"

    insert_query_sexo = """INSERT INTO Sexo (Sexo_Key,Sexo) VALUES """
    insertQuery += insert_query_sexo
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{})".format(1, "Masculino")
    insertQuery += "({},{})".format(2, "Femenino")
    insertQuery += ";\n"

    insert_query_homicidio = """INSERT INTO Causa_Homicidio (Causa_Homicidio_Key,Causa) VALUES """
    insertQuery += insert_query_homicidio
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{})".format(1, "Asesinato")
    insertQuery += "({},{})".format(2, "Homicidio")
    insertQuery += ";\n"

    insert_query_segustradas = """INSERT INTO Personas_Secuestradas (Personas_Secuestradas_Key,Fecha_Key,Ubicacion_Key,Causa_Homicidio_Key,Numeros_Secuestrados) VALUES """
    insertQuery += insert_query_segustradas
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{},{},{},{})".format(1, 1, 1, 1, 1)
    insertQuery += "({},{},{},{},{})".format(2, 2, 2, 2, 2)
    insertQuery += "({},{},{},{},{})".format(3, 3, 3, 3, 3)
    insertQuery += "({},{},{},{},{})".format(4, 4, 4, 4, 4)
    insertQuery += "({},{},{},{},{})".format(5, 5, 5, 5, 5)
    insertQuery += "({},{},{},{},{})".format(6, 6, 6, 6, 6)
    insertQuery += "({},{},{},{},{})".format(7, 7, 7, 7, 7)
    insertQuery += "({},{},{},{},{})".format(8, 8, 8, 8, 8)
    insertQuery += "({},{},{},{},{})".format(9, 9, 9, 9, 9)
    insertQuery += "({},{},{},{},{})".format(10, 10, 10, 10, 10)
    insertQuery += "({},{},{},{},{})".format(11, 11, 11, 11, 11)
    insertQuery += "({},{},{},{},{})".format(12, 12, 12, 12, 12)
    insertQuery += ";\n"

    insert_query_personas_sequestradas = """INSERT INTO Personas_Secuestradas (Personas_Secuestradas_Key,Fecha_Key,Ubicacion_Key,Numeros_Secuestrados) VALUES """
    insertQuery += insert_query_personas_sequestradas
    insertQuery += "\n"
    insertQuery += "VALUES\n"
    insertQuery += "({},{},{},{})".format(1, 1, 1, 1)
    insertQuery += "({},{},{},{})".format(2, 2, 2, 2)
    insertQuery += "({},{},{},{})".format(3, 3, 3, 3)
    insertQuery += "({},{},{},{})".format(4, 4, 4, 4)
    insertQuery += "({},{},{},{})".format(5, 5, 5, 5)
    insertQuery += "({},{},{},{})".format(6, 6, 6, 6)
    insertQuery += "({},{},{},{})".format(7, 7, 7, 7)
    insertQuery += "({},{},{},{})".format(8, 8, 8, 8)
    insertQuery += "({},{},{},{})".format(9, 9, 9, 9)
    insertQuery += "({},{},{},{})".format(10, 10, 10, 10)
    insertQuery += "({},{},{},{})".format(11, 11, 11, 11)
    insertQuery += "({},{},{},{})".format(12, 12, 12, 12)
    insertQuery += ";\n"

```

- g. Creamos un DAG que consta de 2 pasos que utilizará el operador de Airflow llamado "PostgresOperator" el cual es el encargado de manejar conexiones y procesos relacionados con bases de datos PostgreSQL. Estas tareas consisten en crear las tablas de la base de datos y posteriormente poblarlas, todo esto utilizando los scripts descritos anteriormente.

- h.** Para realizar el proceso ETL, definimos el archivo "ELT.py" que contiene el código correspondiente a la implementación del DAG:

#### 4. Propuesta de la arquitectura de solución:

A continuación, se presenta la arquitectura de solución planteada para este problema. Se siguió la metodología del grupo KIMBALL:

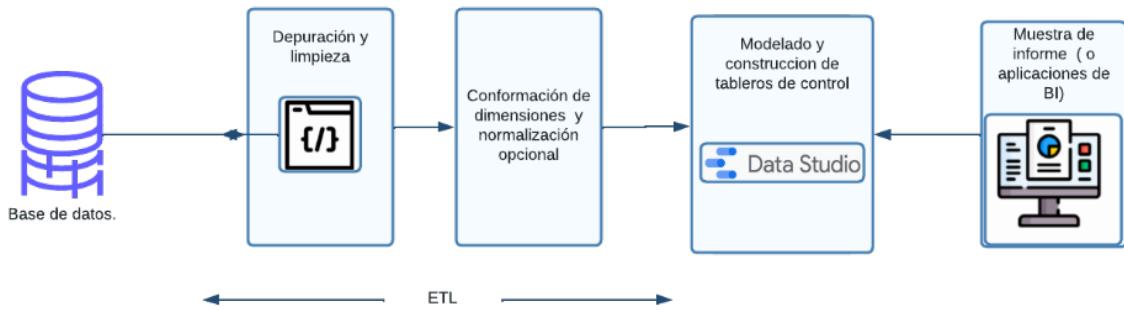


Figura: Arquitectura de solución propuesta para este proyecto.

La arquitectura es sencilla. Primero, partimos de las fuentes de datos: los archivos .csv proporcionados por infraestructura visible. A continuación, estos datos pasan por un proceso que prepara los datos para subirlos a la base de datos, denominado ETL, el cual fue implementado en *Python* utilizando Airflow, como se describió en la sección anterior. A continuación, hay una capa de presentación, en donde los datos estan disponibles para realizar una consulta. En este caso, la capa de presentación corresponde la base de datos PostgreSQL, que es el paso del tercer cuadrito de la figura 5. Por último, se encuentra el modelado y construcción de los tableros de control, los cuales pueden ser accedidos a partir del servicio Data Studio de Google.

#### Implementación de los tableros de control :

La idea de esta parte del proyecto era utilizar GoogleDataStudio conectándose a la base de datos donde se alojan los datos que representan el modelo multidimensional propuesto, para la elaboración de los tableros de control.

Puede encontrar los tableros de control desarrollados en el siguiente link <https://datastudio.google.com/reporting/c775cee7-acb5-427e-98e3-51ef65a4c058>

El tablero de control se realizó en GoogleDataStudio. En este puede ver en la parte izquierda que estan las opciones para realizar los filtros correspondientes. En sí,aca hay dos tableros de control, en el primero se muestran los hechos del proceso de registro del conflicto armado y en el segundo se muestran los hechos del proceso de registro de Salud. A continuación, se muestra una imagen del tablero de control. Para pasar al segundo tablero de control, debe darle clic en el número 2 de la barra vertical izquierda.



Figura 6: Implementación del tablero de control.

El usuario puede elegir uno de los indicadores que desea visualizar, haciendo click en la barra del indicador, como se muestra en la figura a continuación:



Como se observa en la figura 6, también puede discriminar por departamento y fecha, para limitar su análisis dependiendo de ciertos factores y/o características.

Recién se ingresa al tablero, la tabla de pie muestra indicadores de homicidios (Según causa) para todo el territorio nacional. La gráfica de abajo presenta la evolución poblacional y el mapa de la derecha permite una visualización, por departamento, donde el tamaño de la burbuja representa la incidencia del conflicto en dicho territorio.

El usuario puede discriminar por departamento también si hace click en alguna de las burbujas, con lo que obtendrá los resultados específicos para cada departamento, como se muestra en la figura a continuación para Vichada:



Por defecto, se presentan los valores en todos los años si el usuario no selecciona un año o rango de años en particular.

El tablero de salud es bastante similar: en la tabla de Pie, se presentan el porcentaje nacional de afiliados o satisfacción (dependiendo de la métrica elegida por el usuario en la opción de la izquierda), tal que los departamentos de mayores afiliados y satisfacción tendrán una mayor cobertura en el pie:



De manera similar, los usuarios pueden discriminar por departamento, indicador y fecha en el tablero de salud. Para explorar cada una de estas opciones, remítase al link de arriba.

## 8. Descripción de las actividades realizadas:

- Identificación de las necesidades analíticas: Brenda, Alvaro, Sofía
- Modelado del datamarts: Brenda, Alvaro, Sofía
- Entender las fuentes de datos: Brenda, Alvaro, Sofía (5 horas)
- Diseñar e implementar el proceso de ETL: Brenda, Alvaro, Sofía (2 horas)
- Proponer arquitectura de decisión: Brenda, Alvaro (1 hora)
- Implementar los tableros de control: Brenda, Alvaro (6 horas)

- Preparacion del video: Brenda, Sofia, Alvaro (1:15 horas)
- Realización del video: Sofia (1 hora)
- Realización del informe: Brenda, Sofia, Alvaro (2 horas)

**- Repartición de puntos:**

Si tuviéramos que repartir 100 puntos entre los integrantes del grupo, repartiríamos 33.33 a cada uno.

**9. Presentación y video**

Se encuentran en el mismo repositorio de GitHub de este proyecto.

**10. Avance bono estadística**

Después de la anterior entrega, tuvimos una segunda reunión con el grupo interdisciplinario de estadística. Nos comentaron que la interfaz gráfica les pareció muy interesante; y, más aún, que el hecho de que no sólo de la información de la enfermedad que el modelo arrojó, sino la probabilidad de que dicha enfermedad haya sido la elegida, es bastante conveniente.

Tener la probabilidad de una enfermedad le da un buen indicio al médico de si el clasificador está teniendo un buen desempeño o no. Si la probabilidad de clasificación de una enfermedad determinada es bastante alta con respecto a la de otras enfermedades, es más posible que el médico confíe en el diagnóstico. No obstante si la probabilidad de clasificación de una enfermedad es similar a otra, el médico puede entrar a revisar, basado en su criterio, cuál de las dos categorías responde mejor a los síntomas del paciente (pero descartando las otras con probabilidades más bajas), o si su enfermedad cae en dos categorías (como sucede con algunas enfermedades multisistémicas).

**Referencias:**

[1] Imagen de surfistas en mapa de burbuja. Recuperado de: [https://www.data-to-viz.com/graph/IMG/Surfer\\_bubble.png](https://www.data-to-viz.com/graph/IMG/Surfer_bubble.png)